Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/93/20240942

Exploring the efficacy of Multi-Armed Bandit Algorithms in dynamic decision-making

Lai Fu

Department of Computer Science, Hangzhou Dianzi University, Hangzhou, China

22050938@hdu.edu.cn

Abstract. Originating from the scenario of gambling machines in casinos, the Multi-Armed Bandit problem aims to optimize decision-making processes under limited resources to achieve maximum returns. This article delves into the principles, classifications, and practical applications of this problem. Researchers have proposed various algorithms to address this issue, including ε-greedy, Upper Confidence Bound, and Thompson Sampling, which have demonstrated good performance across different scenarios. The article further elaborates on the fundamental principles of Multi-Armed Bandit algorithms, encompassing the trade-off between exploration and exploitation, and provides a detailed classification of algorithms based on probability (e.g., ε-greedy) and value (e.g., UCB). These algorithms not only provide a framework for addressing real-world problems such as advertisement placement and resource allocation, but also possess significant theoretical value in the fields of machine learning and reinforcement learning. By balancing exploration and exploitation, Multi-Armed Bandit algorithms offer effective tools for making optimal decisions in uncertain environments, thus driving the development of related fields.

Keywords: Multi-Armed Bandit, ε-greedy Strategy, reinforcement learning, machine learning.

1. Introduction

The rapid growth of information and the finiteness of resources have made decision-making more difficult and complex recently. In this context, the Multi-Armed Bandit problem, as a classic decision optimization problem, has gained significant importance [1,2]. Its origin can be traced back to the slot machines in casinos. Imagine standing in front of a row of slot machines, each with a different probability of winning. The player needs to decide how to allocate their funds to maximize their earnings. This problem is actually a decision optimization problem, requiring the player to find an optimal decision-making scheme through continuous trials and adjustments under limited resources [3].

Currently, research on the Multi-Armed Bandit problem has made significant progress. Researchers have proposed various algorithms to address this issue, including ε-greedy, UCB, and Thompson Sampling [4]. These algorithms have demonstrated good performance in different scenarios, providing effective tools for practical applications [5]. The Multi-Armed Bandit problem plays a crucial role in areas such as advertising placement, resource allocation, and online learning. These problems often involve making decisions with incomplete information, and the Multi-Armed Bandit problem provides a framework and approach to solve such issues [6].

DOI: 10.54254/2755-2721/93/20240942

From a theoretical standpoint, the Multi-Armed Bandit problem constitutes a pivotal research avenue within both machine learning and reinforcement learning domains. It involves decision-making in uncertain environments, balancing exploration and exploitation, and finding optimal solutions within limited timeframes [7,8]. Research on these issues not only contributes to the development of related theories but also provides strong support for practical applications. Moreover, the Multi-Armed Bandit problem has spurred advancements in related domains like machine learning and reinforcement learning. [9.10]. It has driven research in algorithm design, theoretical analysis, experimental validation, and other aspects, injecting new vitality into these fields [11].

However, despite the significant progress made in the Multi-Armed Bandit problem, there are still some challenges and unresolved issues. For example, in the era of big data, improving the computational efficiency and scalability of algorithms while maintaining their performance is an important challenge [12]. Additionally, the diversity and dynamic changes in practical scenarios pose challenges to the performance of algorithms, necessitating the design of more robust and adaptive algorithms. Privacy and security are also indispensable considerations. In the process of collecting and analyzing user data, it is crucial to prioritize protecting user privacy and security in algorithm design [13].

In summary, the Multi-Armed Bandit problem not only possesses significant practical implications but also holds vital theoretical research value. With the continuous advancement of technology and the expanding application scenarios, it is believed that Multi-Armed Bandit algorithms will play an increasingly important role in the future and provide new impetus for research and development in related fields.

2. Multi-Armed Bandit Algorithm Principles and Classification

2.1. Basic Principles of Multi-Armed Bandit Algorithms

The Multi-Armed Bandit algorithm is a classic algorithm for sequential decision-making in uncertain environments, aiming to find the optimal strategy to maximize long-term gains through continuous experimentation and observation [14]. It simulates the scenario of slot machines in casinos, where each machine (or "arm") represents a possible decision or action, and the probability of winning or the payout of each arm represents the potential return of that decision or action.

In real life, many individuals encounter situations reminiscent of the multi-armed bandit problem, where decisions must be made under conditions of uncertainty [15]. These decisions may involve areas such as advertisement placement, product recommendations, network routing choices, and more. The Multi-Armed Bandit algorithm provides a framework for solving such problems, helping us to make optimal decisions with limited information and time. In the Multi-Armed Bandit problem, the typical assumption is that the rewards of each arm adhere to some undisclosed distribution (e.g., Bernoulli, Gaussian, etc.). The objective of the algorithm is to gauge the genuine reward distribution of each arm via ongoing experimentation and observation, aiming to identify the arm yielding the highest reward [16]. This process requires balancing exploration and exploitation: exploration means trying arms that are not yet well understood to gain more information about their rewards; while exploitation refers to choosing the arm with the currently known highest reward to maximize immediate gains.

DOI: 10.54254/2755-2721/93/20240942

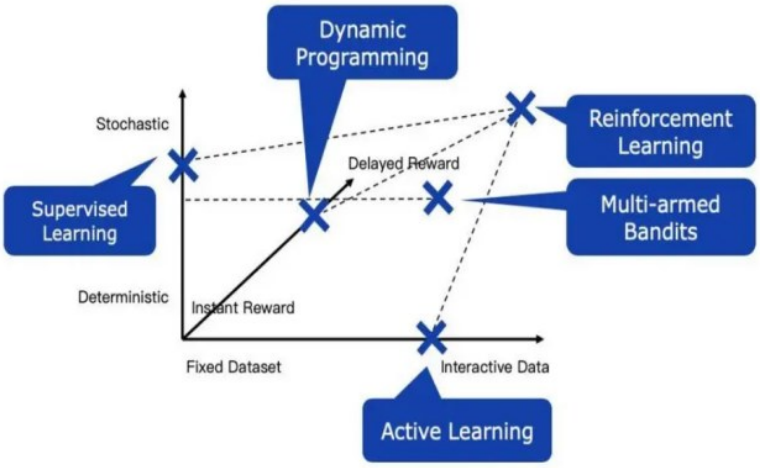

Figure 1. Operating diagram of multi-armed.

Figure 1 shows he fundamental concept of the Multi-Armed Bandit algorithm, which revolves the exploration-exploitation dilemma. In the initial stage of the algorithm, due to limited knowledge of the reward distribution of each arm, there is a need for more exploration to collect enough information to estimate the true rewards of each arm. Over time, the algorithm accumulates data on the rewards of each arm and updates its estimates accordingly. At this point, the algorithm can shift towards exploitation, choosing arms that have performed better to maximize long-term gains. However, completely abandoning exploration may lead the algorithm to converge to a locally optimal solution and miss out on better arms. Hence, within the Multi-Armed Bandit algorithm, managing the balance between exploration and exploitation remains an ongoing endeavor. The algorithm needs to dynamically adjust the ratio of exploration to exploitation based on current information and historical data to make optimal decisions in uncertain environment.

2.2. Classification and Strategies of Multi-Armed Bandit Algorithms

The classification of Multi-Armed Bandit algorithms primarily depends on their approach to managing the equilibrium between "exploration" and "exploitation". The following are some of the main classifications. A common probability-based method in the context of the Multi-Armed Bandit problem is the ε-greedy strategy. With this strategy, exploration occurs through random selection of an arm with a probability ε, while exploitation happens by opting for the currently identified best arm (i.e., the one with the highest average reward) with a probability of 1-ε[17]. In the ε-greedy strategy, an exploration rate ε, usually ranging from 0 to 1, is specified. During each step of action selection, a random number within the range of 0 to 1 is generated. If this number is smaller than ε, a random action is selected; otherwise, if it is equal to or greater than ε, the optimal action is chosen based on the accumulated knowledge. This strategy allows the agent to extensively explore unknown states and actions in the early stages of learning and gradually reduce the exploration rate, increasing the likelihood of exploiting the learned knowledge. The ε-greedy strategy is widely utilized across various reinforcement learning algorithms, including Q-learning, where it plays a crucial role. In the Q-learning algorithm, this strategy is commonly employed to choose the next action, effectively balancing the agent's learning process with the exploitation of acquired knowledge. Adjusting the value of ε allows for control over the trade-off between exploration and exploitation of the agent, consequently influencing the learning speed and eventual performance.

The Upper Confidence Bound (UCB) is a value-based strategy. Its core idea is to assign an upper bound to each arm (or option) based on historical information when selecting an arm, and then choose

DOI: 10.54254/2755-2721/93/20240942

the arm which has the highest upper bound for exploration or exploitation. In the UCB (Upper Confidence Bound) algorithm, the upper bound, characterized by the confidence interval, assumes a pivotal role in reconciling the exploration-exploitation trade-off. Its primary objective is to maximize cumulative rewards [18].

In other words, the UCB algorithm employs an "explore-exploit" strategy. In the initial stage, due to the limited knowledge of the reward distribution of each arm, the algorithm tends to explore more arms to collect information. As data accumulates, the algorithm obtains more accurate estimates of the average reward and uncertainty for each arm. At this point, the algorithm becomes more inclined to select arms with high average rewards and low uncertainty to maximize the total reward.

The most prevalent characterization of the confidence interval in the UCB algorithm is embodied in the UCB1 algorithm. It calculates the upper confidence bound by combining the average reward of each arm and the estimated uncertainty (often a standard deviation or similar measure).

Thompson Sampling is a heuristic strategy used to solve online decision-making problems, particularly adept at handling the explore-exploit dilemma. Its core idea is to describe uncertainty in the form of probabilities, based on Bayesian probability principles, and to balance exploration and exploitation probabilistically when selecting actions. In Thompson Sampling, each choice or action corresponds to a probability model that describes the distribution of rewards or returns that the choice may yield. At each decision point, the algorithm draws a sample from the probability model of each choice and selects the action with the higher sample return. In this way, the algorithm may choose actions that are known to perform well (exploitation) and also actions that have less current information but potentially high returns (exploration)[19]. Thompson Sampling is a powerful and flexible online decision-making strategy that effectively balances exploration and exploitation in uncertain environments by combining Bayesian probability and sampling methods, thereby finding the optimal decision strategy.

3. Advantages and Limitations of Various Algorithms

Allowing an agent to make random exploratory decisions with a certain probability ε, the ε-greedy strategy stands as a commonly used exploration-exploitation trade-off strategy in reinforcement learning. This approach enables the discovery of potentially superior solutions, while exploiting the currently known best strategy with a probability of 1-ε [20]. Here is a detailed analysis of the pros and cons of the ε-greedy strategy:

There are some advantages of the ε-greedy strategy. For example, the ε-greedy strategy balances exploration and exploitation by introducing a probability ε. In the early stages of learning, the agent can explore to discover new and potentially better strategies as learning progresses, the agent gradually reduces exploration and makes decisions based on the knowledge it has learned. This trade-off helps the agent quickly adapt in uncertain environments and find the optimal strategy. In addition, the ε-greedy strategy is simple to implement, computationally efficient, and does not require complex parameter tuning or optimization processes. This makes it easy to implement and deploy in practical applications. Thirdly, the ε-greedy strategy is suitable for various reinforcement learning scenarios, such as advertising placement, recommendation systems, game AI, and more. In these scenarios, the agent needs to continuously adjust its strategy based on environmental feedback, and the ε-greedy strategy effectively helps the agent find a balance between exploration and exploitation.

However, in the ε-greedy strategy, the exploration rate ε is fixed, which may lead to under-exploration or over-exploration in certain situations. For example, when the environment becomes complex or unstable, a fixed exploration rate may not adapt to changes, resulting in the agent failing to find a better strategy in a timely manner. The ε-greedy strategy does not consider the different needs of the agent in different states. In some states, the agent may require more exploration to discover new strategies; in other states, the agent may need to make decisions based on known information. A fixed exploration rate cannot be adjusted according to the actual needs of the agent. Since the ε-greedy strategy involves a certain probability of random exploration in each decision, this may cause the agent to miss

DOI: 10.54254/2755-2721/93/20240942

the global optimal solution and converge to a local optimum in some cases. This problem can be more severe, especially in sparse reward or complex environments.

To overcome the limitations of the ε-greedy strategy, researchers have proposed some improvement methods. For example, an adaptive exploration rate can be used to dynamically adjust the proportion of exploration and exploitation based on environmental changes and the learning progress of the agent; or other exploration strategies (such as uncertainty-based exploration, entropy-based exploration, etc.) can be combined to enhance the agent's exploration capabilities. Additionally, consider combining the ε-greedy strategy with other reinforcement learning algorithms (such as Q-learning, Policy Gradient, etc.) to fully utilize their respective advantages and improve overall performance.

The Upper Confidence Bound (UCB) strategy is employed for exploration in both reinforcement learning and multi-armed bandit problems. Its main idea is to make decisions based on the average reward and uncertainty (confidence interval) of each action. Below is a detailed analysis of the advantages and disadvantages of the UCB strategy [21]:

Advantages: The UCB strategy is theoretically proven to converge to the optimal solution. It selects actions with the highest upper confidence bound, which is calculated by adding a term related to uncertainty to the average reward. This approach balances exploration and exploitation, ensuring that the agent explores new actions while also making full use of known information. Compared to the ε-greedy strategy, the UCB strategy does not require a preset fixed exploration rate. Instead, it dynamically adjusts the ratio of exploration to exploitation based on the historical data and uncertainty of each action. When the uncertainty of a certain action is high, the UCB strategy tends to explore it; whereas, when an action has a high average reward and low uncertainty, the UCB strategy is more inclined to exploit it. This adaptive exploration rate allows the UCB strategy to better adapt to different environments and tasks. In environments with sparse rewards, where most actions yield low or zero rewards, and only a few actions provide significant rewards, the UCB strategy can still effectively explore. By calculating the uncertainty of each action, it can balance exploration and exploitation, avoiding local optima or excessive exploration of low-reward actions.

Disadvantages: The UCB strategy requires calculating the upper confidence bound for each action, which typically involves estimating the average reward and uncertainty of each action. In some cases, when the action space is large or there is a significant amount of historical data, calculating the upper confidence bound can become relatively complex and time-consuming. This may limit the practical application of the UCB strategy due to computational resource constraints. The performance of the UCB strategy can be affected by parameter settings. For example, when calculating the upper confidence bound, it is often necessary to determine a coefficient that balances exploration and exploitation. The value of this coefficient needs to be adjusted according to specific tasks and environments, and improper settings may lead to performance degradation. Additionally, the UCB strategy also requires selecting appropriate confidence levels or confidence interval widths, which can also influence the strategy's exploration and exploitation behavior. The UCB strategy assumes that the reward distribution of actions is static or changes slowly. However In real-world scenarios, the dynamic and non-stationary nature of the environment could impact the efficacy of the UCB strategy. When the reward distribution of the environment changes rapidly, the UCB strategy may not be able to adapt in time and find the optimal strategy.

It's worth noting that the advantages and disadvantages of the UCB strategy are not absolute, and they depend on the specific application scenarios and task requirements. In practical applications, selecting the appropriate exploration strategy is crucial, tailored to the specific situation and coupled with other reinforcement learning algorithms and techniques to enhance performance.

Thompson Sampling is a strategy commonly used in multi-armed bandit problems to balance exploration and exploitation. Its core idea is to use Bayesian inference to maintain estimates of the reward distributions for each action and to select actions based on these estimates [22]. Here is a detailed analysis of the pros and cons of the Thompson Sampling strategy:

Advantages: Thompson Sampling can adaptively adjust the proportion of exploration and exploitation based on historical data and estimates of the reward distributions for each action. When the

DOI: 10.54254/2755-2721/93/20240942

uncertainty of an action is high, the strategy tends to explore that action to obtain more information; when the average reward of an action is high and the uncertainty is low, it tends to exploit that action to maximize gains. This adaptability allows Thompson Sampling to perform well in different environments and tasks. In environments where rewards are sparse, meaning most actions have low or zero rewards with only a few actions providing significant gains, Thompson Sampling can maintain effective exploration. By continuously updating estimates of the reward distributions, it can find a balance between exploration and exploitation, thus avoiding local optima or over-exploring low-reward actions. Thompson Sampling has solid theoretical support. It is based on Bayesian inference to estimate the reward distributions for each action, which is probabilistically reasonable. Additionally, the strategy has been combined with other reinforcement learning algorithms to form a series of extended and improved versions, further enhancing its theoretical reliability.

Compared to some simple exploration strategies (such as ε-greedy), Thompson Sampling may have higher computational complexity. It requires estimating and updating the reward distributions for each action, which involves a large number of numerical calculations and sampling operations. When the action space is large or there is a significant amount of historical data, this computational complexity may become a bottleneck for its application. The performance of Thompson Sampling may be influenced by parameter settings. For example, when choosing prior distributions and likelihood functions, adjustments need to be made based on specific tasks and environments. Improper parameter settings may lead to a decrease in performance. Additionally, the strategy requires determining parameters such as the sampling strategy and update frequency, which may also impact performance. In some cases, Thompson Sampling may have a slower convergence rate. Since it relies on Bayesian inference to update estimates of the reward distributions, it may require more exploration in the initial phase to accumulate sufficient data. This may result in suboptimal performance in the early stages, which gradually improves over time.

In practical applications, the analysis of an algorithm's advantages and limitations is relative and depends on the specific application scenario and task requirements. Appropriate exploration strategies should be chosen based on specific circumstances, and combining them with other reinforcement learning algorithms and techniques can improve performance. Furthermore, as research continues, new improvements and optimization methods may overcome the limitations of Thompson Sampling, enhancing its performance in practical applications.

4. Limitations of Various Algorithms Research Prospects

With the advent of the big data era, the amount of data that algorithms need to process is growing exponentially [23]. Finding the optimal solution quickly and accurately among massive data is a significant challenge for Multi-Armed Bandit algorithms. Future research needs to focus on improving the computational efficiency and scalability of algorithms while ensuring their performance. Secondly, practical scenarios often involve various complex factors, such as the diversity of user behaviors and dynamic changes in the environment. These factors may lead to decreased or failed algorithm performance. Therefore, designing more robust and adaptive Multi-Armed Bandit algorithms to meet the needs of different scenarios is an important research direction in the future. Finally, privacy and security issues are also non-negligible aspects in the research of Multi-Armed Bandit algorithms. During the process of collecting and analyzing user data, protecting user privacy and security and preventing data leakage and misuse are key considerations in algorithm design.

In addition, with the continuous development of technologies such as deep learning and reinforcement learning, combining Multi-Armed Bandit algorithms with these advanced technologies to further improve algorithm performance and generalization capabilities is also a worthwhile direction to explore. For example, deep learning can be utilized to extract latent features from data, or reinforcement learning can be employed to optimize the exploration and exploitation strategies of the algorithm.

DOI: 10.54254/2755-2721/93/20240942

5. Conclusions

In conclusion, the Multi-Armed Bandit problem is a classic decision-making problem that has gained significant importance in recent years due to the rapid growth of information and the finiteness of resources. This problem originated from the scenario of gambling machines in casinos, where decisions must be made under conditions of uncertainty to achieve maximum returns. Researchers have proposed various algorithms, such as ε-greedy, Upper Confidence Bound, and Thompson Sampling, to address this issue. These algorithms have demonstrated good performance across different scenarios and have significant theoretical value in the fields of machine learning and reinforcement learning.

The fundamental principles of Multi-Armed Bandit algorithms revolve around the exploration-exploitation dilemma, which highlights the importance of balancing exploration and exploitation to make optimal decisions in uncertain environments. The classification of algorithms is based on probability and value, and they not only provide a framework for addressing real-world problems such as advertisement placement and resource allocation but also possess significant theoretical value.

However, there are still some challenges and unresolved issues in this field. Improving the computational efficiency and scalability of algorithms, designing more robust and adaptive algorithms, and addressing privacy and security issues are some of the key areas that require further research. With the continuous advancement of technology and the expanding application scenarios, Multi-Armed Bandit algorithms are expected to play an increasingly important role in the future and provide new impetus for research and development in related fields.

In summary, the research on Multi-Armed Bandit algorithms still has broad prospects and challenges, and future research needs to focus on computational efficiency, robustness, adaptability, privacy, and security, among other issues, to promote the application and development of this algorithm in more fields. By balancing exploration and exploitation, Multi-Armed Bandit algorithms offer effective tools for making optimal decisions in uncertain environments, thus driving the development of related fields.

References

[1] Kumar A B ,Hrishikesh D ,Subir B .2023.Federated Multi-Armed Bandit Learning for Caching in UAV-aided Content Dissemination[J].Ad Hoc Networks,151

[2] Hiba D ,Raphaël F ,Nadège V , et al.2023.Massive multi-player multi-armed bandits for IoT networks: An application on LoRa networks[J].Ad Hoc Networks, 151

[3] Jian D ,Xuan L ,Shuang Q , et al.2023.Spectrum Allocation and User Scheduling Based on Combinatorial Multi-Armed Bandit for 5G Massive MIMO.[J].Sensors (Basel, Switzerland), 23(17)

[4] Runqi W ,Linlin Y ,Hanlin C , et al. 2023.Anti-Bandit for Neural Architecture Search[J].International Journal of Computer Vision, 131(10):2682-2698

[5] Linxiong H ,Bin S ,Shizheng L , et al. 2023.Portfolio allocation strategy for active learning Kriging-based structural reliability analysis[J].Computer Methods in Applied Mechanics and Engineering, 412

[6] Nicolas C ,Guillaume B ,Serge H , et al. 2020.Entangled N-photon states for fair and optimal social decision making[J].Scientific Reports, 10(1):20420-20420.

[7] Roberto G C . 2021.A multi-armed bandit algorithm speeds up the evolution of cooperation[J].Ecological Modelling, 439109348-.

[8] Carrascosa M ,Bellalta B . 2019.Decentralized AP selection using Multi-Armed Bandits: Opportunistic ε-Greedy with Stickiness.[J].CoRR, abs/1903.00281

[9] Kojima F ,Usami T ,Duong T N . 2007.Identification of Stress Corrosion Cracking Profiles Using ε-Greedy Search Inverse Analysis in Eddy Current Testing[J].Proceedings of the ISCIE International Symposium on Stochastic Systems Theory and its Applications, 2007(0):118-123.

[10] Wijith M ,Sven S ,Anthony T , et al. 2016.Effect of veliparib (ABT-888) on cardiac repolarization in patients with advanced solid tumors: a randomized, placebo-controlled crossover study.[J].Cancer chemotherapy and pharmacology, 78(5):1003-1011.

DOI: 10.54254/2755-2721/93/20240942

[11] Liu Y ,Tsuruoka Y . 2016.Modification of improved upper confidence bounds for regulating exploration in Monte-Carlo tree search[J].Theoretical Computer Science, 64492-105.
[12] Honghan Y ,Xiaochen X ,C. R J C , et al. 2023.Online nonparametric monitoring of heterogeneous data streams with partial observations based on Thompson sampling[J].IISE Transactions, 55(4):392-404.
[13] Darak J S ,Zhang H ,Palicot J , et al. 2017.Decision making policy for RF energy harvesting enabled cognitive radios in decentralized wireless networks[J].Digital Signal Processing, 6033-45.
[14] Sengupta S . 2015.Comparisons of sampling strategies for estimating finite population proportions in direct and randomized response surveys under a super population model[J].Model Assisted Statistics and Applications, 10(4):385-390.
[15] Ortega A P ,Braun A D . 2014.Generalized Thompson sampling for sequential decision-making and causal inference[J].Complex Adaptive Systems Modeling, 2(1):1-23.
[16] Alagumani S ,Natarajan M U . 2024.Q-learning and fuzzy logic multi-tier multi-access edge clustering for 5g v2x communication.[J].Network (Bristol, England), 21-24.
[17] Zhang L ,Wu J ,Liu C , et al. 2024.A novel control strategy of automatic parallel parking system based on Q-learning[J].Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 238(4):661-673.
[18] Hu Y ,Li W ,Luo Q . 2024.Real-Time Adjustment Method for Metro Systems with Train Delays Based on Improved Q-Learning[J].Applied Sciences, 14(4):
[19] Ma H ,Zhang H ,Tian D , et al. 2024.Optimal demand response based dynamic pricing strategy via Multi-Agent Federated Twin Delayed Deep Deterministic policy gradient algorithm[J].Engineering Applications of Artificial Intelligence, 133(PA):108012-.
[20] Lyu B ,Yang Y ,Cao Y , et al. 2024.Efficient multi-objective neural architecture search framework via policy gradient algorithm[J].Information Sciences, 661120186-.
[21] Multi-Objective Optimization of Vehicle-Following Control for Connected Electric Vehicles Based on Deep Deterministic Policy Gradient[J].SAE International Journal of Electrified Vehicles,2023,13(1):
[22] David D ,Javier G ,Á. M S . 2023.Twin-delayed deep deterministic policy gradient algorithm for the energy management of microgrids[J].Engineering Applications of Artificial Intelligence, 125
[23] Xiaohui H ,Xiong Z ,Jiahao L , et al. 2023.Effective credit assignment deep policy gradient multi-agent reinforcement learning for vehicle dispatch[J].Applied Intelligence, 53(20):23457-23469.