Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/158/2025.TJ23293

Reinforcement Learning in Practice: Muti-applications of Basic Theory

Hanyuan Su

International School, Jinan University, Guangzhou, China
hysu0508@stu2023.jnu.edu.cn

Abstract: Reinforcement Learning (RL), as an essential branch of machine learning, has garnered significant attention due to its outstanding performance in complex decision-making tasks and dynamic environments, exemplified by notable achievements such as AlphaGo and DeepSeek in recent years. To provide researchers with a comprehensive perspective of current developments in this area, this paper systematically reviews RL from four distinct perspectives: fundamental theoretical frameworks, practical applications, existing limitations, and future outlook. Initially, core concepts and classical algorithms including Q-learning, SARSA, and Temporal Difference (TD) learning are clearly introduced, establishing a solid theoretical foundation. Then, practical implementations of RL are elaborated within three representative fields: path planning, voltage control and game theory, highlighting their significance and effectiveness through literature analysis. Moreover, this paper critically analyzes the limitations of current RL techniques, such as low sample efficiency, unstable training processes, and insufficient generalization. Finally, potential research directions, including hybrid learning paradigms and multi-agent collaborations, are proposed to inspire future advancements. The insights provided in this review aim to stimulate further theoretical innovation and practical breakthroughs in reinforcement learning.

Keywords: Reinforcement Learning, Path Planning, Voltage Control, Game Theory, Limitations.

1. Introduction

With the quick advancement of artificial intelligence technologies in recent years, reinforcement learning, a crucial area in the machine learning, has progressively emerged as a prominent domain of inquiry and application. In particular, in milestone events such as AlphaGo defeating the human Go champion (2016), OpenAI Five defeating the human team in the DOTA2 game (2018-2019), Chinese AI lab DeepSeek releasing the R1 inference model (2025), reinforcement learning has demonstrated excellent decision-making and optimization capabilities, which has greatly promoted the attention of academia and industry to its theory and practice. Reinforcement learning learns the optimal strategy through the interaction between the agent and the environment in the process of trial and error, providing a powerful tool for solving complex problems with high dimensions, dynamics, and uncertainty.

Reinforcement learning not only shows well performance in traditional fields such as game theory, robot control, and autonomous driving, but also has achieved initial application results in emerging fields such as financial transactions, personalized recommendations, and bioinformatics. However,

DOI: 10.54254/2755-2721/158/2025.TJ23293

reinforcement learning still faces challenges such as low sample efficiency, unstable training, and weak generalization ability. How to build a more stable bridge between theory and practice is still an important topic of current research.

Most of the existing review papers on reinforcement learning focus on the latest cutting-edge theories or are limited to a certain field, lacking the introduction of basic theories and the diversity of the field. This article aims to systematically review the current status and development trends of reinforcement learning and focus on the following four aspects: first, an overview of the basic theoretical framework and main algorithms of reinforcement learning; second, an exploration of the application practice of reinforcement learning in the real world, including path planning, voltage control and game theory; third, an analysis of the key challenges and limitations of current reinforcement learning research and applications; and fourth, an outlook on the future development direction and potential breakthroughs of reinforcement learning. Through the above analysis, this article aims to offer readers a comprehensive and systematic perspective on reinforcement learning research, promoting the integrated development of theory and application.

2. Reinforcement learning

2.1. The basic of reinforcement learning

Reinforcement learning is the process of learning the mapping from environmental states to actions. The objective of reinforcement learning is to discover a behavior strategy that maximizes a numerical reward signal. There are currently two primary approaches to solving problems in reinforcement learning. The first approach involves searching for the optimal behavior within the agent's behavior space, typically achieved using genetic algorithms and other search techniques. In the second way, the utility function value of an activity in a certain environmental condition is estimated using statistical approaches and dynamic programming methods.The agent's interactions with the environment may be explained by the conventional reinforcement-learning paradigm depicted in Figure 1. Through its behaviors and the environment's input, an agent in the model engages with its surroundings. The agent will receive input i, or some indicator of the present condition of the environment, at each stage of the interaction. Based on a few techniques, the agent will then select an action, which will produce the outcome. Through a value-state function or other transitions, the action will alter the state of the environment, and the environment will respond with feedback in the form of a scalar signal reinforcement signal r. The action that tends to raise the total value of the reinforcement signal over time will be selected by the agent from the collection of behaviors B. With the help of a number of algorithms, the agent may repeat this phase through trial and error.

Formally, the model is composed of a discrete set of agent actions A, a discrete set of environment states S, and a set of scalar reinforcement signals, usually {0,1} or other real values [1].

Finding a method that directs the agent in choosing the best course of action to optimize the reward obtained from the environment is the aim of reinforcement learning. The rewards in the majority of puzzles, however, consist of both immediate and delayed reward values. Therefore, in order to estimate the ideal behavior in the future, an objective function must be defined. This objective function is typically expressed using the state value function or state-action value function. The function forms are as follows.

$$V^{\pi}(s_t) = \sum_{t=0}^{\infty} \gamma^t r_t, \quad 0 < \gamma \leq 1 \tag{1}$$

$$V^{\pi}(s_t) = \sum_{t=0}^{h} r_t \tag{2}$$

$$V^{\pi}(s_t) = \lim_{h \to \infty} \left(\frac{1}{h} \sum_{t=0}^{h} r_t \right) \tag{3}$$

DOI: 10.54254/2755-2721/158/2025.TJ23293

Where γ is the discount factor, r_t is the reward values received from the environment after translating the agent's state from s_t to s_{t+1}. Formula (1) is infinite-horizon discounted model, which takes long-term reward of agent into account accumulates reward with discounted factor in the infinite steps. Formula (2) is finite-horizon model, which only considers the reward in the next h steps. Formula (3) is average-reward model, which takes average reward in the long-term into consideration. Obviously, if the objective function can be fixed, then the agents can determine the optimal behavior strategy under this formula:

$$\pi^* = \arg_\pi \max V^\pi(s_t), \quad \forall s \in S \tag{4}$$

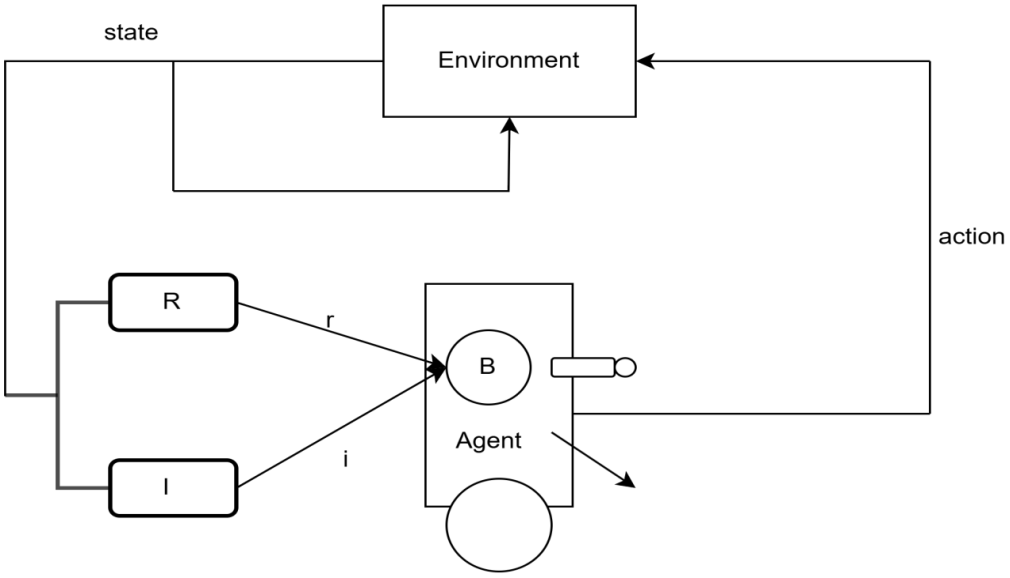

Figure 1: The standard reinforcement-learning model (picture credit: original)

Problems with delayed reinforcement are typically formalized using the framework of Markov Decision Processes (MDPs) [1,2]. A five-tuple $M= \langle S, A, R, T, \gamma \rangle$, where S is the finite set of states, A is the finite set of actions, $R : S \times A \rightarrow$ Real number is the reward function, $T : A \times S \rightarrow P(S)$ is the transition function, P(S) is a probability distribution over the set S and γ is the discount factor, can be used to describe MDPs. $R(s, a)$ is the reward function that denotes the value of the immediate reward after the agent takes the action an in the state s. $T(s, a, s')$ is the probability transfer function, which represents the probability that the state s translates to the state s' when agent takes the action a.

In essence, MDPs is a decision process in the probability and reward of the transition from the current state to the next state only depend on the current state and the action chosen in the decision. It is not related to the historical states and actions. If it is possible to clearly determine the probability transfer function T and the reward function R, then the optimal strategy can be found by Dynamic Programming (DP). The two main ideas of DP, policy iteration and value iteration, are proposed by Ronald A Howard and Richard Bellman respectively. However, reinforcement learning focuses on the case that the function T and function R is not clear. In this case, the usual method is to use iteration techniques to adjust the estimated value of the value function in the current state and next step [3].

Consider objective function (1), the definition of optimal value function is:

$$V^* = \max_a \left(R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V^*(S^*) \right), \forall s \in S \tag{5}$$

The equation:

DOI: 10.54254/2755-2721/158/2025.TJ23293

$$V(s) = (1-\alpha)V(s) + \alpha\big(R(s,a) + \gamma V(s')\big) \tag{6}$$

is a basic equation that many reinforcement algorithms will modify appropriately based on learning rate and convergence speed [3].

2.2. The typical reinforcement learning algorithm

If the agent does not need to learn the reward function and probability transfer function in the MDP model, this method is called model-free. On the contrary, if it is necessary to learn the knowledge of a model, it is called model-based.

2.2.1. TD algorithm

The simplest method in reinforcement learning is called temporal difference (TD) learning. It makes use of dynamic programming concepts and Monte Carlo techniques. On the one hand, it doesn't require a model and can learn from the agent's expertise. However, like dynamic programming, it makes adjustments to estimations depending in part on other estimates [4].

The simplest TD algorithm is algorithm, which was proposed by Sutton in 1988. It uses the update rule:

$$V(s) = V(s) + \alpha\big(r + \gamma V(s') - V(s)\big) \tag{7}$$

Where α is a constant learning rate and it is guaranteed to make optimal value function converge. The key idea of $TD(0)$ is that it updates the estimated value of the adjacent state at each step by iteration. The $TD(0)$ algorithm's slow convergence speed, however, is a drawback because it only looks forward in one step when modifying value estimates. The more efficient way is looking arbitrary steps back when the agent gets an immediate reward, which is $TD(\lambda)$ algorithm. The $TD(\lambda)$ rule mentioned above is comparable to the broader TD (0) rule:

$$V(s) = V(s) + \alpha\big(r + \gamma V(s') - V(s)\big)e(s) \tag{8}$$

Where $e(s)$ is the eligibility of each state s. It can be calculated by follow rules:

$$e(s) = \sum_{k=1}^{t} (\lambda\gamma)^{t-k}\delta_{s,s_k}, \text{where } \delta_{s,s_k} = \begin{cases} 1 \text{ if } s = s_k \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

$$e(s) = \begin{cases} \gamma\lambda e(s) + 1 \text{ if } s = \text{current stete} \\ \gamma\lambda e(s) \text{ otherwise} \end{cases} \tag{10}$$

From equation (9), if a state s is visited for many times, then the eligibility of s will be more and its contribution for the reward is greater. Equation (10) is a better version.

2.2.2. Q-learning

In 1992, Watkins presented the model-free reinforcement learning method known as Q-learning. It is also called off-policy TD. It uses rewards for state-action pairs and $Q^*(s,a)$ instead of status reward $V(s)$ in TD learning. $Q^*(s,a)$ can be written recursively:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \max_{a'} Q^*(s',a') \tag{11}$$

Where $Q^*(s,a)$ denotes the optimal sum of discount reward when the agent takes action in the state s. In the Q-learning, the agent will choose optimal action by using $\varepsilon - \text{greedy strategy}$[4] and get a series of tuples like $\langle s,a,s',a' \rangle$ as experience knowledge. Then the agent will modify the Q value based on this rule:

DOI: 10.54254/2755-2721/158/2025.TJ23293

$$Q(s,a) = Q(s,a) + \alpha\left(r' + \gamma \max_{a'} Q(s',a') - Q(s,a)\right) \quad (12)$$

Like the TD(λ) algorithm, Q-Learning can also be expanded to update more steps than previously.

2.2.3. SARSA

In 1994, Rummery and Niranjan proposed the model-based SARSA algorithm. Another name for it is on-policy TD learning. It iterates using the Q value, just like Q-learning. The rules determine how it updates:

$$Q(s,a) = Q(s,a) + \alpha\left(r' + \gamma Q(s',a') - Q(s,a)\right) \quad (13)$$

The difference is that SARSA uses a practical Q value instead of the max value of the value function. At each step, the agent determines the action based on the current Q value, which means that it is an on-policy TD learning.

3. Reinforcement learning application

3.1. Path planning

For autonomous mobile devices like robots, unmanned vehicles, and drones, path planning is the process of determining the best or most efficient route from a starting point to a destination within a given environment. It is essential to domains like disaster robotics, military mission planning, and self-driving car navigation.

Minh et al. used a deep neural network to perform value-based reinforcement learning function estimation, which is called a deep Q-network (DQN). DQN first solved the key problems between Q-learning and neural networks and made reinforcement learning suitable for path learning [5]. Ji et al. utilized a Neural-Network-Driven Prediction model (NDR), which employed its guidelines and regional predictions as prior knowledge to assist Q-learning in achieving higher accuracy and faster convergence speed. It provides an effective method for mobile robot path planning [6]. The NDR-QL method diagram is shown in Figure 2. Pan et al. proposed an improving Q-learning method by combining the Radial Basis Function Neural Network (RBF) with Q-learning, which is called the RBF-Q network [7]. The structure of the RBF-Q network is shown in Figure 3. It not only uses the RBF network to approximate the action-value function of the Q-learning algorithm, but also expands action sets of robots based on dynamically adjusting. It helps agents to find a shorter path in fewer training rounds.

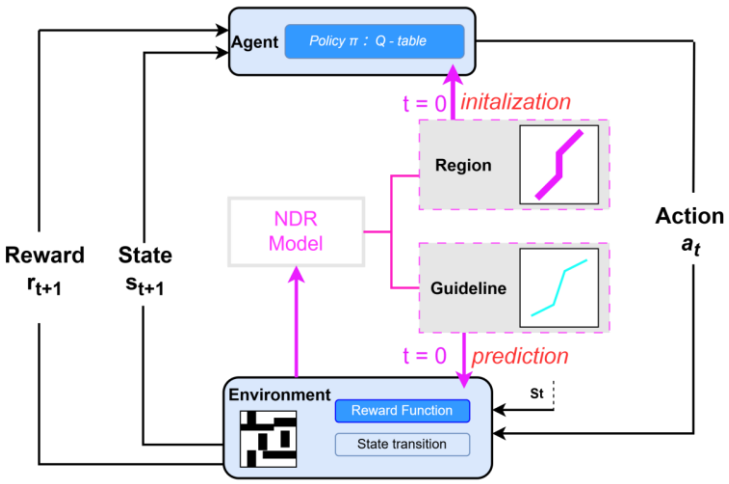

Figure 2: The diagram of the NDR-QL method [8]

DOI: 10.54254/2755-2721/158/2025.TJ23293

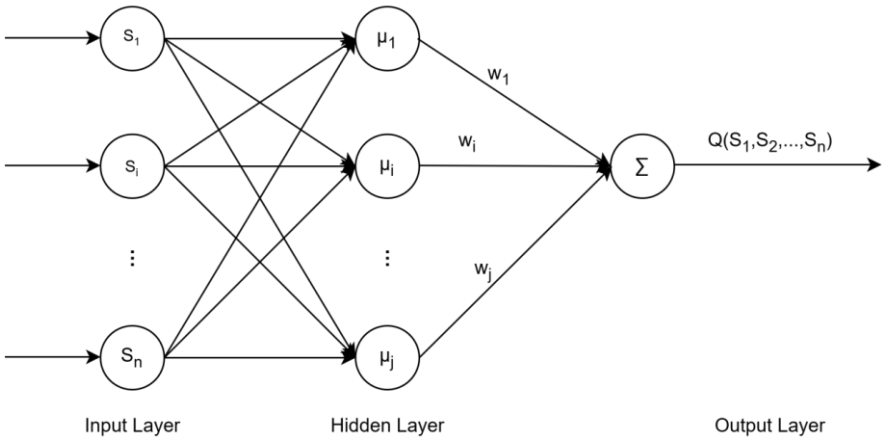

Figure 3: The structure of the RBF-Q network [9]

Besides Q-learning, there are also other reinforcement learning applications. Tai et al. focused on developing a mapless motion planner based on deep deterministic policy gradients (DDPG) which is expanded into an asynchronous version. Compared with traditional DDPG, asynchronous deep deterministic policy gradients (ADDPG) decrease the iteration steps to get higher Q values and show better robustness in extremely complicated environments [8]. Zhang et al. proposed improving dueling double deep Q network (D3QN) for AUV path planning in the ocean [9]. They combined D3QN with the LSTM neural network and N-step temporal difference, which is called LN-D3QN. From their experiment, the LN-D3QN algorithm performs well in the presence of ocean current disturbances and can successfully avoid obstacles. It has excellent adaptability in dealing with complex ocean current environment path planning problems.

3.2. Voltage control

Voltage control is integral to daily electrical needs and safety. Reinforcement learning has been extensively and effectively applied in the field of voltage control for power systems. It is particularly adept at solving complex, nonlinear, and dynamic uncertainty problems that traditional voltage control methods struggle to address. However, with the increase in electrical needs, more reasonable power distribution and more precise power control require more effective algorithms [10].

Duan et al. proposed autonomous voltage control strategies (AVC) based on the deep Q-network (DQN) and deep deterministic policy gradients (DDPG). Different from DQN, DDPG is trained with an actor-critic-based approach [11]. It was found that DDPG performs well in controlling capacitor banks and transformer taps of active distribution networks to achieve dynamic voltage control. However, it needs more computation sources since it searches a wider ranging environment at the beginning. Zhang et al. proposed a data-driven distributed voltage control method using the advanced multi-agent deep reinforcement learning (MADRL) algorithm and spectrum clustering [12]. The technique divided the distribution system into many sub-networks according to the sensitivity of voltage and reactive power using unsupervised clustering. Every sub-network is represented as an adaptive agent in the context of a Markov game. Deep neural networks are used to approximate the policy and value functions in the training process using an improved Multi-Agent Deep Reinforcement Learning (MADRL) method. While choices are carried out in a distributed method utilizing just local knowledge, all agents receive centralized training to learn the best coordinated voltage regulation tactics. This approach has a reduced computation time, controls voltage variation, and has strong scalability and adaptation to large-scale systems.

DOI: 10.54254/2755-2721/158/2025.TJ23293

3.3. Game theory

A "game theory problem" refers to a decision-making scenario involving multiple decision-makers—also known as participants or agents—whose actions impact one another. In such a problem, each participant aims to maximize their own benefits by selecting an optimal strategy, and their benefits frequently hinge on the choices made by others, thereby creating a strategic interactive relationship.

In early times, reinforcement learning is applied for zero-sum game problems between two people, two teams, or multiple people. Great progress has been made in areas such as Go, games, Texas Hold'em, and Mahjong. For example, systems such as AlphaGo [13], OpenAI Five [14, 15], AlphaStar [16], DeepStack, and Suphx have reached or exceeded the level of human experts in these areas. Table 1 shows these applications. However, most of these applications focused on zero-sum game problems involving two people, two teams, or multiple people, while research on mixed game problems lacks substantial progress and breakthroughs [17]. In recent years, reinforcement learning has been applied to more game types in a wide range of fields. Jin et al. suggested a bi-level motion planning system for intersections that is based on heuristic reinforcement learning and a reasoning game theory scheme. The main idea is to apply a recurrent neural network in the upper layer, while interactive games with many kinds of agents are made possible by Q networks that are selectively coupled in the lower layer [18]. Then, the self-agent can gradually update its estimations and derive corresponding actions from the historical joint state. It is found that this two-layer controller improves collision avoidance performance and reduces the passing time of the vehicles, which is helpful to vehicles to pass intersections effectively in the self-driving fields. Sun et al. proposed Curriculum Learning Multi-Agent Deep Deterministic Policy Gradient (CL-MADDPG), dividing the large-scale Unmanned Aerial Vehicle (UAV) cluster game confrontation task into several small-scale confrontation tasks [19]. It gradually increases the number of confrontational UAVs, so that the confrontation strategy of the UAV cluster is gradually improved on the original basis. It was found that the average reward value and the winning rate of the adversarial strategy trained using the CL-MADDPG algorithm are significantly better than those of the unimproved multi-agent reinforcement learning algorithm. This method provides a new direction for the UAV training process which can help UAV to make a precise decision.

Table 1: The reinforcement learning applications in the game

System/Method	Application Scenario	Reinforcement Learning Method	Game Type
AlphaGo/AlphaZero	Go, Chess	Monte Carlo Tree Search + Policy/Value Network	Zero-sum game
OpenAI Five	Dota2	Proximal Policy Optimization (PPO)	Muti-agent confrontation game
AlphaStar	StarCraftII	MARL + PPO	Muti-agent complex game
DeepStack	Heads-Up No-limit Texas Hold’em	Deep Reinforcement Learning (DRL)	Zero-sum in complete information game
Suphx	Chinese Mahjong	Deep Q network (DQN)/Actor-Critic	Muti-player game (incomplete information)

DOI: 10.54254/2755-2721/158/2025.TJ23293

4. Limitation and future outlook

While reinforcement learning has achieved some success in research and application, it remains fundamentally limited to ideal, highly structured experimental data within simulated environments. Reinforcement learning models perform well in the training environment but often perform poorly and generalize poorly to slightly different environments. For instance, in path planning, a trained strategy might only be effective under a specific map or a particular obstacle layout; in voltage control, the model may not be able to adapt to various load fluctuation scenarios; and in gaming, a highly confrontational strategy can often be easily overcome by new tactics. Moreover, other limitations such as high training costs, low sample efficiency, poor explanatory power, and limited real-world deployment also persist. To advance reinforcement learning towards general artificial intelligence, research and application in reinforcement learning should focus on the following areas:

The first is focusing on transfer learning and meta-learning to enhance reinforcement learning adaptability in new environments. For instance, in path planning, researchers can leverage prior knowledge to train models that quickly adapt to new tasks, subsequently developing general strategies applicable to various maps and dynamic obstacles.

The second is combining the hybrid training mechanism of the simulation environment (simulator) with real data, the fusion method of model predictive control (MPC) and reinforcement learning, as well as model-based RL. It can help agents improve sample utilization and decrease the training interactions with the environment, which improves training efficiency. In tasks such as voltage control, workers can obtain historical operating data from sensors, and then train a power grid environment model based on this data. Finally, they train reinforcement learning strategies offline in the model and deploy them to the actual system.

The third is introducing the explainable artificial intelligence (XAI) method and safe constraint reinforcement learning (Safe RL). They can provide visualization and decision logic tracking for the reinforcement learning model and improve stability and credibility in key scenarios like game problems.

The last is closely integrating with traditional optimization algorithms, such as linear programming and control theory, to form a "hybrid intelligent strategy" that leverages the advantages of both. For instance, in voltage control, the RL strategy can be coupled with a traditional PID controller. In path planning, RL is responsible for the global strategy and the traditional algorithm is used for local refinement.

5. Conclusion

Generally, despite the significant advancements reinforcement learning has made in areas such as path planning, voltage control, and game theory, it still encounters numerous challenges before it can be widely applied in practice. Future research is likely to concentrate on enhancing generalization, improving sample efficiency, boosting model interpretability and security, and fostering integration with traditional optimization techniques, as well as extending the capabilities for multi-agent systems and real-world deployment. As techniques develop, computing power increases, and cross-domain integration progresses, reinforcement learning is hopeful to demonstrate greater adaptability and decision-making prowess in increasingly complex and dynamic real-world settings.

This paper has systematically reviewed reinforcement learning from multiple perspectives, including fundamental theoretical frameworks, related applications, existing limitations, and potential future directions. Initially, we elucidated core concepts and classical algorithms in reinforcement learning, such as Q-learning, SARSA, and Temporal Difference (TD) learning, establishing a theoretical foundation for subsequent discussions. Furthermore, practical applications of reinforcement learning are thoroughly examined across three representative domains: path planning,

DOI: 10.54254/2755-2721/158/2025.TJ23293

voltage control, and game theory. Extensive literature reviews illustrated the capability of reinforcement learning to address complex decision-making tasks within uncertain and dynamic environments.

Despite significant progress, reinforcement learning still encounters critical limitations, notably low sample efficiency, unstable convergence during training, and limited generalization ability. This review identifies these drawbacks and suggests prospective solutions, emphasizing promising research directions including the integration of reinforcement learning with supervised learning, meta-learning, and multi-agent cooperation techniques.

In summary, this comprehensive overview serves as a reference for understanding current advancements and challenges in reinforcement learning, aiming to stimulate deeper theoretical investigations and innovative practical applications.

References

[1] *Ghallab, M., & Laruelle, H. (1994). Representation and control in IxTeT, a temporal planner. Journal of Artificial Intelligence Research, 1, 61–113. https://doi.org/10.1613/jair.9*

[2] *Zhang, X., Wang, H., Zhang, J., & Zhao, Y. (2020). Hybrid path planning algorithm based on improved artificial potential field method and reinforcement learning. arXiv preprint. https://arxiv.org/abs/2006.15085v1*

[3] *Zheng, G. Q., & Zhu, S. Q. (2004). An improved artificial potential field approach for path planning. Acta Automatica Sinica, 30(1), 89–93.*

[4] *Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.*

[5] *Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.*

[6] *Ji, Y., Yun, K., Liu, Y., Xie, Z., & Liu, H. (2024). Neural-network-driven reward prediction as a heuristic: Advancing Q-learning for mobile robot path planning. arXiv preprint arXiv:2412.12650.*

[7] *Pan, Q., Zhao, Y., & Gan, Y. (2025). Robot path planning based on improved Q-learning algorithm. Internet of Things Technologies, 2025(3), 23–30.*

[8] *Tai, L., Paolo, G., & Liu, M. (2017). Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 31–36.*

[9] *Zhang, Y., & Gao, B. (2025). Research on AUV path planning based on deep reinforcement learning. Journal of Northeast Normal University (Natural Science Edition), 57(1), 53–62.*

[10] *Kuznetsova, E., Li, Y. F., Ruiz, C., & Zio, E. (2013). Reinforcement learning for microgrid energy management.Energy, 59, 133–146.*

[11] *Duan, J., Shi, D., Diao, R., Wang, H., Zhang, Z., & Bian, D. (2020). Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations.IEEE Transactions on Power Systems, 35(1), 814–817.*

[12] *Zhang, Y., Xu, Y., Hu, W., & Wen, S. (2020). Distributed Voltage Regulation of Active DistributionNetworks Based on Enhanced Multi-Agent Deep Reinforcement Learning.IEEE Transactions on Power Systems, 35(6), 4964–4976.*

[13] *OpenAI. (2019). OpenAI Five. https://openai.com/research/openai-five*

[14] *Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.*

[15] *Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140–1144.*

[16] *Vinyals, O., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning.Nature, 575, 350–354.*

[17] *Dong, S., Li, C., Yang, G., Ge, Z., Cao, H., Chen, W., Yang, S., Chen, X., Li, W., & Gao, Y. (2025). A survey on solving and applications of hybrid game problems. Journal of Software, 36(1), 107–151.*

[18] *Jin, X., Li, K., Jia, Q. S., Xia, H., Bai, Y., & Ren, D. (2020, November). A game-theoretic reinforcement learning approach for adaptive interaction at intersections. In 2020 Chinese Automation Congress (CAC) (pp. 4451-4456). IEEE.*

[19] *Sun, S., & Wang, Q. (2025). Large-scale UAV swarm game confrontation based on improved multi-agent reinforcement learning. Journal of Southeast University (Natural Science Edition), 55(2), 123–134.*