# Q-Learning: Applications and Convergence Rate Optimization

Peiyi Liu*

China University of Petroleum (Beijing) at Karamay, Karamay, China

*Corresponding author: 2020015650@st.cupk.edu.cn

**Abstract.** As an important algorithm of artificial intelligence technology, Q-learning algorithm plays a significant part in a number of fields, such as driverless technology, industrial automation, health care, intelligent search, game, etc. As a classical learning algorithm in reinforcement learning, with the help of an experienced action sequence in a Markov environment, an agent can learn to select the best course of action using the model-free learning technique known as Q-learning. This paper mainly discusses the addition of received signal strength (RSS) to the Q-learning algorithm to navigate unmanned aerial vehicle (UAV), summarizes the main content and results of the neural Q-learning algorithm helping UAV avoid obstacles, adaptive and random exploration (ARE) method is proposed to address the issue in UAV planning a route tasks, summarizes the content and results of route designing of moving robot using obstacle characteristics as Q-learning states and actions, the Q-learning algorithm employs a novel exploration technique that combines ε-rapacious exploring with Boltzmann theory. to help mobile robot to plan its path, and analyzes the convergence speed of the algorithm for planning a route of stage Q-learning and the path planning algorithm of traditional Q-learning. When there are many states and actions, the operation efficiency of the Q-learning algorithm will be greatly reduced, so it is necessary to study in depth how to reduce the operation time of the algorithm for Q-learning and increase the convergence velocity of algorithm for Q-learning.

**Keywords:** Q-learning, Markov environment, UAV, mobile robot, convergence speed.

## 1. Introduction

Since 1956, 13 scientists including Marvin Minsky, John McCarthy, and Claude Elwood Shannon have met at Dartmouth College in the United States, marking the birth of the discipline of artificial intelligence. Since the advent of artificial intelligence, both theory and technology have advanced and have had a significant role in a number of sectors, such as fingerprint recognition, face recognition, intelligent search, games, etc. Today's artificial intelligence is still developing rapidly, Today's artificial intelligence is still developing rapidly, reinforcement learning is a kind of machine learning, which is used to solve the maximum reward value of the agent during the connection with the surroundings through the learning strategy to achieve a specific goal, reinforcement learning differs from supervised learning and unsupervised learning in that it doesn't require previously provided necessary information, but by receiving the environment to the action reward to obtain learning information and update model parameters, the reinforcement learning system (RLS) must learn from its own experience in order to gather knowledge in the context of action-evaluation and improve the course of action in order to adapt to the environment because the external environment does not supply much information. Q-learning, a traditional reinforcement learning algorithm, is a model-free learning approach, and does not have many parameters and can run offline, it gives intelligent systems the capacity to learn to choose the best course of action based on the sequence of actions they have already taken in a Markov environment. It is mathematically possible to simulate sequential decision-making using the Markov decision process, and its theoretical basis is the Markov chain, which mimics the randomness strategy and rewards attained by agents in the environment of the Markov characteristics system state. There are some variants of the Markov decision-making process, include Markov decision-making techniques that are only partially observable, restricted, and fuzzy. Markov decision-making process is widely used in dynamic programming, random sampling, automatic control, recommendation systems, etc. A fundamental tenet of Q-learning is that the agent's contacts with the surroundings can be thought of as a Markov decision-making process, where the agent's existing state and selected action, which acquires via action-value function, could finally give the

requested action based on the present and ideal strategic plan. The goal of Q-learning is to find a strategy that maximizes the rewards obtained.

The second part of this paper mainly explains the addition of received signal strength (RSS) to the Q-learning algorithm to navigate unmanned aerial vehicles (UAVs), helps drones avoid obstacles by studying neural Q-learning algorithms, and proposes an ARE method to solve the problems in UAV path planning tasks. The third section of this essay focuses on how robotic systems plan their paths. employing obstacle features as the state and action of Q-learning, and it also suggests a fresh investigation method that blends ε-rapacious exploring with Boltzmann theory in order to improve the algorithm for Q-learning. The focus of the fourth section of this study is a approach to routing based on stage Q-learning that converges more quickly than the conventional algorithm for Q-learning, enabling the agent to discover the best collision-free path more quickly.

## 2. Q-Learning Applied in UAV

### 2.1. UAV Navigation Based on Q-Learning

UAV need to make their own decisions to respond to changes in the environment because knowing beforehand of the environment is limited during mission performance, and the environment could alter over time or in response to the target model [1]. If the Q-learning algorithm is added to the UAV navigation, it can help the UAV to learn in unknown actions according to the current state, so as to overcome the uncertainty of the environment and obtain the maximum return value. The Q-learning algorithm was given RSS by M. M. U. Chowdhury et al., enabling the UAV to be guided toward the source signal [1]. RSS depends on both the source signal and the location of the UAV, unlike the original position Q-learning technique. M. M. U. Chowdhury et al. used RSS to define the state and reward values of the algorithm.

The trajectory of the UAV obtained by the RSS-based Q-learning algorithm is shown in Fig. 1. In both cases, because the trajectory attempts to prevent areas with lower RSS values and because the reward is defined as the disparity among RSS values in a continuous state, the UAV shows a tendency to have higher RSS values in the next step instead of locating the source signal location with the shortest path. Similarly, if the UAV is simulated from a different starting position, it will have basically the same effect, sensing a path with a higher RSS value, and ultimately directing the UAV to the source signal location. The UAV may effectively go to its destination despite not knowing where the originating signal is located.

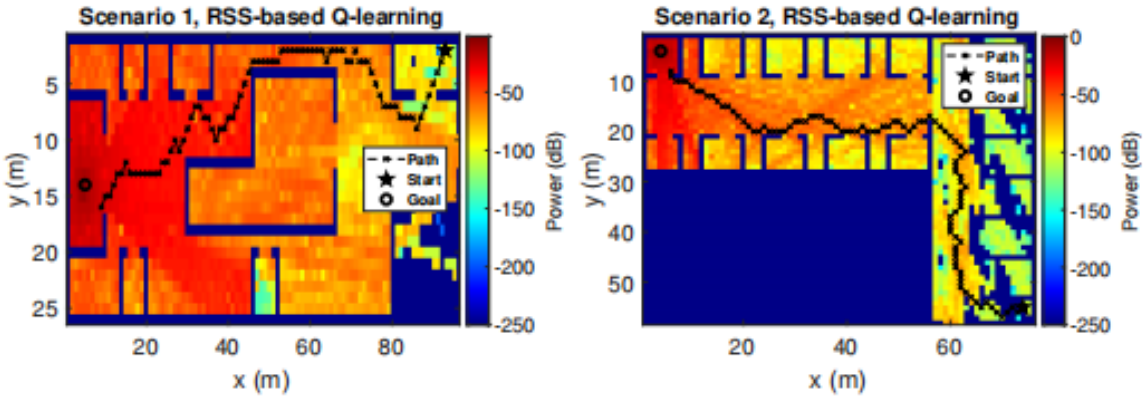


**Fig. 1** UAV trajectories RSS-based Q-learning [1]

### 2.2. UAV Obstacle Avoidance Based on Q-Learning

UAV obstacle avoidance is very important in performing tasks. Navigation may be challenging for the UAV during missions since it may not be aware of the precise environment and obstacle information. Therefore, Benchun Zhou et al. helped the UAV avoid obstacles by studying the neural Q-learning method [2]. They assumed that the environment in which the UAV was located was

unknown, and when there were obstacles within the execution range, In order to address the quality issue with the Q-table, maintain the route planning process, and take into account the UAV's physical limitations, they integrated BP neural networks to Q-learning.

In Fig. 2 and Fig. 3, there are three static barriers in the environment, When approaching the target initially, the UAV uses a fast approach tactic, and when the UAV detects an obstacle for the first time, neural Q-learning (NQL) is applied to put the status information back into the propagation neural network, thus obtaining the possibility between actions. In order to earn incentives and update the neural network, the UAV chose a fantastic opportunity to turn right. When the UAV runs into barriers once more, again, continuing to apply NQL. The UAV must take 72 steps to complete the journey in Fig. 2. However, with additional practice, the UAV can discover a different path to the target, which in Fig. 3 only calls for 69 steps.

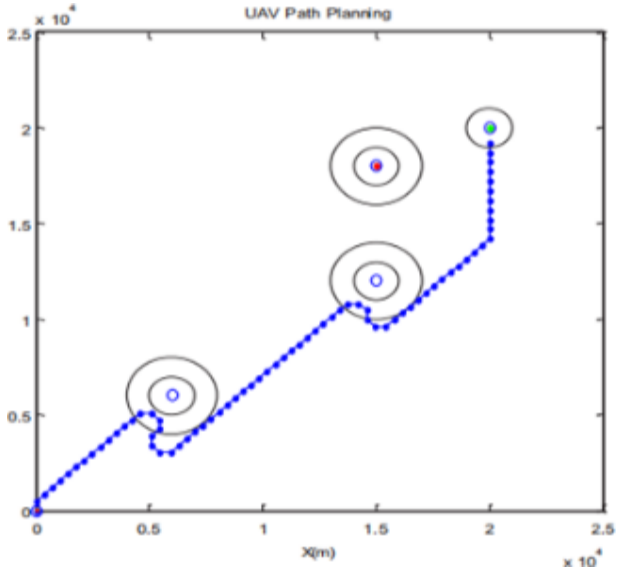


**Fig. 2** Path planning based NQL [2]

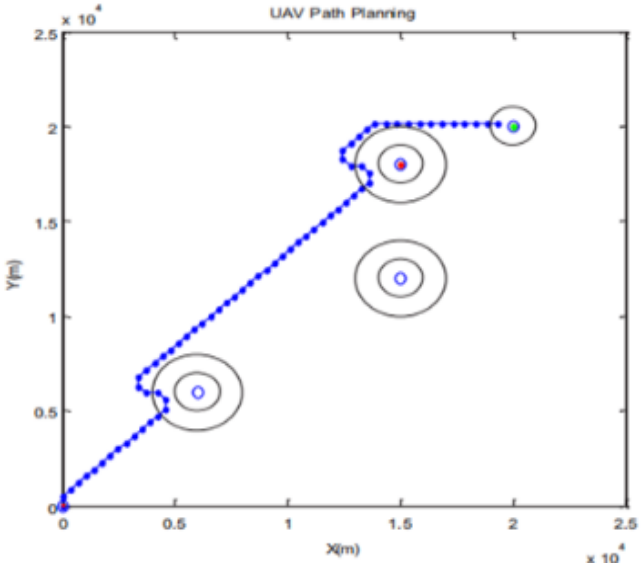


**Fig. 3** Another path planning based NQL [2]

When static and moving obstacles are present at the same time, the drone will collide with something in the preceding period, but with enough practice, it will be able to avoid it and eventually find a collision-free route to the intended location. The entire procedure took 72 moments, but after many trainings, the drone only needed 66 steps to find another shorter path, not only avoiding obstacles, but also saving time [2]. The NQL algorithm proposed by Benchun Zhou et al. can reach the target position regardless of whether the target and obstacle are moving, it demonstrates the NQL algorithm's adaptability to many contexts.

Zhao Yijing et al. propose an ARE to solve the problem in UAV route arranging tasks [3]. The core principle of ARE is for UAV to explore the environment and act on current assessments (Fig. 4). In addition, whenever it approaches an obstacle, it is corrected to a safe path through a random mechanism. This method balances the convergent random search combined with an adaptive method of self-learning, so that the UAV cannot only have the skill to find the general direction, but also get rid of the dilemma caused by learning errors.

The three components that make up ARE are the snare-escape component, the operation component, and the learning component. The learning component and the snare-escape component provide instructions to the operation component during path planning, which then authorizes the UAV to perform appropriate action. The learning component can teach action choosing tactics based on the current status of the UAV and historical operation data sequences. The UAV can be guided by the snare-escape component to a safe path by using a random tree search method to help it escape a hazardous scenario.

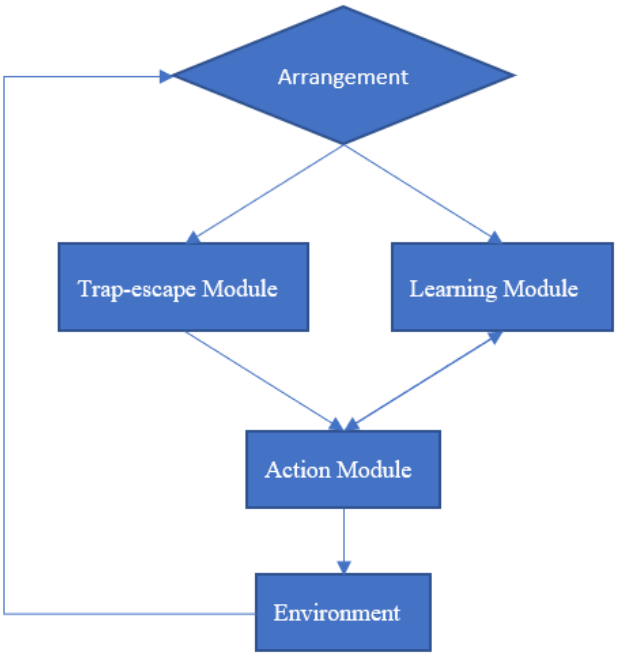


**Fig. 4** Structure chart of the ARE approach [3]

## 3. Q-Learning Applied in Mobile Robot

### 3.1. Mobile Robot Path Planning Based on Q-Learning

Path planning for moving robots involves continuously determining whether a movement from the origin to the destination position is optimal or suboptimal. Since there is no model feature in Q-learning, the system can find the ideal route from the starting place to the destination position simply by designing the status, behavior, and benefit functions, but the size of the Q-table may have a large impact on the calculation time [4,5].

Jing-Kai Lin et al. will apply them to dynamic environments based on robotic operating systems (ROS) and leverage the features of barriers as algorithm for Q-learning states and actions to reduce computation time [6-8]. Consequently, the moving robot can successfully develop and track the path from the initial location to the target location. Jing-Kai Lin et al. set all obstacles as rectangles, specify the each obstacle has four perimeter markers, and the area of the perimeter marker is a rectangular region made up of the boundary point of an obstacle, according to the corner points of the rectangle and the largest size of the moving robot. Jing-Kai Lin et al. simulated the process of moving a mobile robot from the starting position to the desired position in the surroundings of kinetic obstacles and stationary barriers. The dynamic path planning system initially begins planning a route, and when a period of time has passed, the location of the moving obstruction changes, and a fresh route is created via the route planning system that is dynamic, and so on until the moving robot successfully reaches the destination.

The traditional Q-learning algorithm occupies a large amount of search space and time, and is simple to settle for a local optimum, Therefore, In their enhanced Q-learning algorithm, Siding Li et al. provide a novel exploration approach that combines ε-rapacious exploring with Boltzmann theory [9]. They also offer a heuristic search technique to condense the area of the search, restrict a variety of direction angle modifications, etc., and lessen the number of learning iterations required. The mobile robot route planning problem's MDPs model is illustrated below [9]:

$$R(s,a) = \sum_{s' \in S} p(s,a,s') R(s,a,s') \quad (1)$$

$$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} p(s,a,s') \sum_{a' \in A} Q(s',a') \quad (2)$$

An MDP is described as a 4-ary function with the elements (S,A,P,R), where S is a set of robot states and A denotes all of its actions. P is the likelihood of a choice, and R denotes the benefit function. The robot is in the status s at the moment, and the state following action a in the s state is $s'$. The state-action with the relevant reward value is $Q(s,a)$. The results show that this method has more time advantages than traditional Q-learning and other path planning methods such as A-magnitude in different dynamic experimental environments. Finally, they smoothed the path predicted by the algorithm using the B-spline curve [9]. The smoothed path is more likely to be used for accurate navigation in autonomous vehicles.

## 4. Q-Learning Algorithm Convergence Rate Optimization

An essential reinforcement learning algorithm, the traditional algorithm for Q-learning sets the Q-value to an equal or random value during the initialization process, that is, learning without prior knowledge, this will cause the algorithm's convergence pace to be slower [10]. Xu Xiaosu et al. introduced the gravitational potential field in the artificial potential field method during the Q-value initialization process [11]. It quickens the algorithm's convergence rate. Duan Jianmin et al accelerates the algorithm's rate of convergence by introducing the ambient potential energy value as search heuristic information to initialize the Q-value [12]. Therefore, the research on accelerating the convergence velocity of Q-learning algorithms is ongoing.

A technique for route designing based on stage Q-learning was introduced by Yang Xiuxia et al., which determined the exploration step size of each stage through the environment scale and expected exploration information [13]. Then, the reward pool and reward threshold were set according to the exploration step of each stage to ensure that the exploration path of each stage is the optimal path of that stage. Then the variable factor was reset so that this stage continued to explore based on the optimal path of the previous stage. Finally, the optimal path of each stage was determined, and the optimal path of each stage was combined into a global optimal path to accomplishing the robot's path planning from its initial position to its destination. By comparing the simulation's outcomes in Table 1, it can be seen that under the same simulation environment, stage Q-learning is superior than the traditional algorithm for Q-learning in terms of algorithm convergence time and path planning time indicators. The 202s route mapping time of the stage Q-learning algorithm is 49% inferior to the 398s designing a route time of the conventional algorithm for Q-learning.

**Table 1.** Comparison of simulation data of two algorithms [13]

| Parameter | Path Planning Time | Convergent Turns |
|---|---|---|
| Traditional Q-learning | 398 | 55 |
| Stage Q-learning | 202 | 34 |

## 5. Conclusion

This paper reviews the addition of RSS to the algorithm for Q-learning to navigate UAV, the application of the brain-based Q-learning algorithm to assist UAV to avoid obstacles, and an ARE method to make the issue disappear in UAV route planning tasks. This paper also summarizes the route designing of moving robots using the characteristics of barriers as the state and action of Q-learning, a fresh approach to exploring combining ε-rapacious exploring with Boltzmann theory in the improved Q-learning algorithm is elaborated, and analyzes the algorithm's rate of convergence for planning a route of stage Q-learning and the planning algorithm for path of traditional Q-learning. Q-learning requires a Q-table for the choice of agent as a reference. In the case of many states, Q-table scale will be very large, so finding and storing need to consume a lot of time and space, which will greatly reduce the operating efficiency of the algorithm for Q-learning. Therefore, the need for

in-depth research direction is to reduce the operation time of algorithm for Q-learning, raise the convergence rate of algorithm for Q-learning to boost the algorithm's overall operational effectiveness.

## References

[1] M. M. U. Chowdhury, F. Erden and I. Guvenc, "RSS-Based Q-Learning for Indoor UAV Navigation," MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM), Norfolk, VA, USA, 2019, pp. 121-126

[2] B. Zhou, W. Wang, Z. Wang and B. Ding, "Neural Q-Learning Algorithm based UAV Obstacle Avoidance," 2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC), Xiamen, China, 2018, pp. 1-6

[3] Z. Yijing, Z. Zheng, Z. Xiaoyi and L. Yang, "Q-learning algorithm based UAV path learning and obstacle avoidence approach," 2017 36th Chinese Control Conference (CCC), Dalian, China, 2017, pp. 3397-3402

[4] C. Watkins and P. Dayan, "Q-learning", Machine Learning, vol. 8, pp. 257-277, 1992.

[5] E.S. Low, P. Ong and K.C. Cheah, "Solving the optimal path planning of a mobile robot using improved Q-learning, " Robot. Auton. Syst., 115, pp. 143-161, 2019.

[6] J. -K. Lin, S. -L. Ho, K. -Y. Chou and Y. -P. Chen, "Q-learning based Collision-free and Optimal Path Planning for Mobile Robot in Dynamic Environment," 2022 IEEE International Conference on Consumer Electronics - Taiwan, Taipei, Taiwan, 2022, pp. 427-428

[7] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, Leibs and Ng, "ROS: an open-source Robot Operating System", in ICRA workshop on open source software, vol. 3, no. 3.2, pp. 5, 2009.

[8] A. Koubaa, "Robot Operating System (ROS): The Complete Reference", vol. 1, Springer, 2016.

[9] S. Li, X. Xu and L. Zuo, "Dynamic path planning of a mobile robot with improved Q-learning algorithm," 2015 IEEE International Conference on Information and Automation, Lijiang, China, 2015, pp. 409-414

[10] Bernstein AV, Burnaev E V, Kachan O N. Reinforcement learning for computer vision and robot navigation[M]/ / Perner P. Machine Learning and Data Mining in Pattern Recognition. Cham: Springer International Publishing, 2018: 258 - 272.

[11] Xu Xiaosu,Yuan Jie. Path Planning Method for Mobile Robots Based on Improved Reinforcement Learning [J]. Journal of China Inertial Technology, 2019, 27 (03) : 314 - 320. Xu X S,Yuan J. Mobile robot path planning method based on improved reinforcement learning[J]. Journal of Chinese Inertial Technology,2019,27(03) : 314- 320.

[12] Duan Jianmin, Chen Qianglong. Research on Q-Learning Path Programming Algorithm Using Prior Knowledge [J]. Electro-Optics & Control, 2019, 26(09): 29 - 33. Duan J M,Chen Q L. Research on Q-Learning path planning algorithm using prior knowledge[J]. Electro-optical and control,2019,26 (09) : 29 - 33.

[13] Yang Xiuxia, Gao Hengjie, Liu Wei, et al. Robot path planning based on stage Q-learning algorithm [J]. Journal of Ordnance Equipment Engineering, 2022, 43(05):197-203.