Wind and Storage Cooperative Scheduling Strategy Based on Deep Reinforcement Learning Algorithm

QIN Jingtao\textsuperscript{1, a}, HAN Xueshan\textsuperscript{2, b}, LIU Guojing\textsuperscript{3}, WANG Shang\textsuperscript{4}, LI Wenbo\textsuperscript{5}, JIANG Zhe\textsuperscript{6}

\textsuperscript{1}Key Laboratory of Power System Intelligent Dispatch and control of Ministry of Education (Shandong university) Jinan 250061, Shandong Province, China +8617865190501

\textsuperscript{2}Key Laboratory of Power System Intelligent Dispatch and control of Ministry of Education (Shandong university) Jinan 250061, Shandong Province, China

\textsuperscript{3}State Grid Jiangsu Electric Power Company Economic Research Institute Nanjing 210008, Jiangsu Province, China

\textsuperscript{4}Tianjin Chengdong Power Supply Company Hedong District, Tianjin 300250, China

\textsuperscript{5}State Grid Shandong Electric Power Research Institute Jinan 250003, Shandong Province, China

\textsuperscript{6}State Grid Shandong Electric Power Research Institute Jinan 250003, Shandong Province, China

\textsuperscript{a}1476393346@qq.com \textsuperscript{b}xshan@sdu.edu.cn

ABSTRACT: Wind energy has become one of the most promising new energy sources in the context of the global energy interconnection. However, the inherent uncertainty of wind power and the volatility of its output make it difficult to join the internet for scheduling. The uncertainty of the output must be reduced or even eliminated by certain methods. In this paper, wind power and energy storage are coordinated to eliminate the uncertainty of wind power output, and to reduce the burden on the grid while ensuring the long-term operation of wind farms. This paper first introduces Q-learning in reinforcement learning as a controller. Through a large number of historical wind power data training, the controller has good decision-making ability, so as to reduce the punishment caused by wind power uncertainty; then the Q-learning is improved aiming at the maximum average income of the stage. Finally, the Q-value network is established with conventional Q-learning and improved Q-learning. The DQN algorithm in the deep reinforcement learning algorithm is introduced for deep training and decision-making and the three algorithms are verified. The result proves that the deep reinforcement learning algorithm can achieve better control effect than Q-learning.

1. INTRODUCTION
As a clean, pollution-free and renewable green energy source, wind energy is widely optimistic in the context of the global energy interconnection, and its development prospects are broad. Although the
development of wind power is imperative, wind power itself has natural uncertainties and fluctuates greatly in time and space, and free connection to grid will bring difficulties to grid dispatching and control[1]. The cooperation of wind power and energy storage is one of the methods to solve the uncertainty of wind power output. Many scholars at home and abroad have made great research in this aspect. Literature [2] establishes a joint scheduling model for wind power and storage, with the goal of maximizing the long-term income of wind farm operations; literature [3] studies the cooperation between pumped storage and wind power, and considered various constraints of the turbine group; literature [4] divides wind power output into high frequency and low frequency and is stabilized by different devices; literature [5] performs random optimization in two stages of the cooperation of wind power and energy storage, and considers the uncertainty of electricity price and wind power; A cooperation of wind power and energy storage model based on dynamic programming is proposed in literature [6] to further compensate the wind power forecasting error in the electricity market environment; A multi-objective model for the cooperation of wind power and energy storage is established in literature [7] by minimizing the loss of load caused by wind power as one of the objective functions.

The above-mentioned literatures are all the ideas of the control of traditional power grid based on the traditional mathematical optimization, and separate the prediction from the scheduling, so it is difficult to obtain a good control effect. In recent years, AI algorithms, including deep learning and reinforcement learning have emerged, and have achieved good application results in the fields of image recognition, autonomous driving, intelligent robots and medical health. The reinforcement learning algorithm has great advantages in dealing with the uncertainty problem among them and has been successfully applied to the optimal power flow calculation[8], AGC control[9], power producer bidding[10][11] and other fields. Literature [12] uses the Q-learning algorithm in the reinforcement learning theory to study the regulation and configuration of the wind farm self-discipline; literature [13] further uses the Q-learning algorithm to empirically learn and optimize the cooperation of wind power and energy storage. Both of them have achieved good control results. In recent years, deep reinforcement learning has become the latest research hotspot in the field of artificial intelligence[14], and the most successful application is the case of Alpha Dog. Literature [15] introduces its application in video games. The power grid emergency control strategy is studied in literature [16] based on the deep reinforcement learning algorithm.

This paper is organized as follows. The second part first gives the mathematical model of the cooperation of wind power and energy storage. The third part establishes the improved Q-learning algorithm aiming at maximizing the periodical average income of the wind farm. The pre-learning model is trained under a large amount of historical wind power data, and then is input to online learning. The fourth part introduces the DQN algorithm in deep reinforcement learning as a controller, uses the MATLAB neural network toolbox to create a BP network and uses the same data for training. The fifth part is the analysis and comparison of the three algorithms. The sixth part is the conclusion.

2. THE COOPERATION OF WIND POWER AND ENERGY STORAGE SYSTEM

2.1 The realization mechanism of the cooperation
The realization mechanism of the cooperation is shown in figure 1. The wind power and energy storage system (ESS) form a hybrid system for coordinated scheduling. Specifically, when the actual value of the wind power is lower than the planned output value, the energy storage discharge, and on the contrary the energy storage charge.

2.2 The Principle of dealing with power uncertainty
The wind power output has strong randomness. Generally, the wind power prediction error $P_{w_t}$ can be assumed to follow a normal distribution, where $P_{w_t}$ is the difference between the actual value of the
wind power $P_{w,t}^{\text{real}}$ and the predicted value of the wind power $P_{w,t}^{\text{pred}}$ in the $t$-time period as shown in equation (1).

$$P_{w,t} = P_{w,t}^{\text{real}} - P_{w,t}^{\text{pred}}$$  \hspace{1cm} (1)$$

Since the energy storage system can compensate for wind power prediction errors by charging and discharging, it can reduce the wind power reserve capacity that needs to be purchased. Assume that the up-and-down reserve capacity from the grid during the $t$-time period that the hybrid system purchases are $R_{u,t}^m$ and $R_{d,t}^m$. If the wind power prediction error exceeds the standby range, the energy storage system will compensate as shown in figure 2. Where $(P_{w,t})$ is the probability density function for wind power prediction error in $t$-time period, $\sigma_{w,t}$ is the standard deviation of the probability distribution of wind power prediction error in $t$ period; $\alpha$ and $\beta$ are the confidence coefficients when the true value of wind power is higher or lower than the predicted value of wind power.

As can be seen from figure 2, the cooperation of wind power and energy storage can reduce the spare capacity that needs to be purchased. In the process of the cooperation, the actual charging and discharging power of the energy storage system can be expressed as equation (2).

$$P_{w,t}^{\text{es}} = \begin{cases} 
P_{w,t}^{\text{es}} &= P_{w,t}^{\text{es}} = P_{w,t}^{\text{es}} 
&\quad \text{if } P_{w,t} > R_{u,t}^m \\
&= P_{w,t}^{\text{es}} = P_{w,t}^{\text{es}} 
&\quad \text{if } -R_{d,t}^m \leq P_{w,t} \leq R_{u,t}^m \\
&= P_{w,t}^{\text{es}} = P_{w,t}^{\text{es}} 
&\quad \text{if } P_{w,t} < -R_{d,t}^m
\end{cases}$$  \hspace{1cm} (2)$$

2.3 Mathematical model of the cooperation

As mentioned earlier, the objective function of the cooperation is to maximize the operating income of wind farms, which can be expressed as equation (3).

$$\max \sum_{t=1}^{T} (\lambda_l P_{w,t}^{\text{es}} - \mu^u R_{u,t}^m - \mu^d R_{d,t}^m - C)$$  \hspace{1cm} (3)$$

Where $T$ is the number of time periods divided by the total operation period; $\lambda$ is the price of the electricity during the $t$-time period; $P_{w,t}^{\text{es}}$ is the actual output power of the hybrid system in the $t$-time period; $R_{u,t}^m$ and $R_{d,t}^m$ are the up-and-down reserve capacity from the grid during the $t$-time period that the hybrid system purchases; $\mu^u$ and $\mu^d$ are the price of the up-down reserve capacity for the $t$-time period; $c$ is the penalty fee of the hybrid system when the desired control effect is not achieved.

3. THE PRINCIPLE OF IMPROVED Q-LEARNING ALGORITHM

3.1 The principle of Q-learning algorithm

The Q-learning algorithm is an online learning algorithm based on action value functions, which uses action value functions to evaluate selected actions in a certain state. The iterative process of Q-learning can be expressed by equation (4).

$$Q^{(t+1)}(s,a) = Q^{(t)}(s,a) + \alpha(Q^{(t)}(s,a) + \gamma \max_a Q^{(t)}(s_{t+1},a) - Q^{(t)}(s,a))$$

$$Q^{(t+1)}(s,a) = Q^{(t)}(s,a), \quad \forall (s,a) \neq (s_t,a_t)$$  \hspace{1cm} (4)$$

Where $\alpha$ is the step factor. When the value of $\alpha$ is large, the convergence of the learning process can be accelerated. When the value of $\alpha$ is small, the sufficiency of the exploration can be guaranteed. The larger the discount factor $\gamma$, the greater the impact of the future reward value on the current decision.

This paper uses the greedy coefficient $\beta$ to establish an action selection strategy, and calculates the probability that each action is selected according to equation (5), which can balance the greed and exploration degree of action selection.
\[ P_s^k(a) = \beta + \frac{Q'(s,a) - \min_{a'} Q'(s,a') + \delta}{\sum_{a'} [Q'(s,a') - \min_{a''} Q'(s,a'') + \delta]} (1 - \beta) \]
\[ P_s^k(a) = \frac{Q'(s,a) - \min_{a'} Q'(s,a') + \delta}{\sum_{a'} [Q'(s,a') - \min_{a''} Q'(s,a'') + \delta]} (1 - \beta), \quad a \neq a' \quad \text{(5)} \]

Where the range of $\beta$ is $0 \leq \beta \leq 1$, the closer the $\beta$ is to one, the greater the degree of greed. $P_s^k(a)$ indicates the probability that the greedy strategy is selected in the state $s$ at the $k$th iteration, and $P_s^k(a)$ indicates the probability that the non-greedy strategy $a$ is selected in the state $s$ at the $k$th iteration. Introduce a very small positive number $\delta$ to ensure that the probability that each action is selected will not be zero.

3.2 Improved Q-learning algorithm

As mentioned earlier, the operational goal of wind farms is to maximize long-term rewards, so this constraint should be added to the decision. In the pre-learning phase of conventional Q-learning, the control period can only be advanced in one direction, and one control period can only be performed once, which is consistent with the reality. However, in computer programs, this limitation can be completely broken, and multiple backtracking learning is performed in one control cycle. Therefore, this section establishes an improved Q-learning algorithm with the goal of maximizing the periodical average income of the wind farm to apply to the pre-learning stage. At the same time, referring to relevant literature, the greedy degree coefficient is increased as the learning process progresses.

The specific steps to improve Q-learning are as follows.

1) Read historical wind power data, set parameters, initialize matrix $Q^t = 0$, $T = 1$, $\beta = \beta_{in}$.
2) Temporarily store the predicted historical wind power data at the current stage in the queue, and set Temporary $Q$ matrix $Q_{\text{temp}} = Q$, $\beta = \beta_{in} - \frac{T \times (\beta_{in} - \beta_{out})}{\sum_{T}}$.
3) Initialize the number of loops $n_{\text{exploration}} = 1$.
4) Initialize the period in $T$ stage $t = 1$.
5) According to the current energy storage value $E_t$ and the predicted wind power value $P^\text{pred}_{w,t+1}$ in the next period, the current state $s_t$ is determined, and then the corresponding action strategy is selected according to the equation (5), thereby obtaining the planned charging and discharging power of the energy storage system and up-and-down reserve capacity purchasing from the power grid;
6) After the next period, the actual wind power value $P^\text{act}_{w,t+1}$ is obtained. According to the selected action strategy, the actual charging and discharging power of the energy storage system $P^\text{act}_{es,t+1}$ is determined by equation (2);
7) According to the electricity price and the standby price of the $t+1$ period, combined with the actual output of the energy storage system and the actual output of the wind power, the immediate reward value is calculated according to the following equation (6);
\[ r_{t+1} = \lambda (P^\text{act}_{es,t+1} + P^\text{act}_{w,t+1}) - \mu^\text{up}_{t+1} R^\text{up}_{w,t+1} - \mu^\text{dn}_{t+1} R^\text{dn}_{w,t+1} \quad \text{(6)} \]
8) Update $Q_{\text{temp}}$ according to equation (4);
9) Determine whether stage $T$ is over, and if it is finished, calculate the average income of the stage at the $n_{\text{exploration}}$ loop. Otherwise $n_{\text{exploration}} = n_{\text{exploration}} + 1$ and goes to step 5;
10) If the average income of the stage $\text{avrr}_{\text{exploration}}(n_{\text{exploration}})$ is greater than the average income of the $T-1$ stage $\text{avrr}(T-1)$, then set the current stage average income $\text{avrr}(T) = \text{avrr}_{\text{exploration}}(n_{\text{exploration}})$, $Q = Q_{\text{temp}}$, end the loop and set stage $T = T + 1$, go to step 3, otherwise, go to step 5;
11) Determine whether the maximum number of loops is reached. If so, set $\text{avrr}(T)$ equal to the maximum value that appears in the loop, set the matrix $Q$ equal to $Q_{\text{temp}}(n_{\text{exploration}})$ when the maximum value occurred.
4. DECISION-MAKING OF THE COOPERATION BASED ON DQN ALGORITHM

4.1 The principle of DQN algorithm
The dimension of the value function of Q-learning algorithm is the product of the state space dimension and the action space dimension. After the pre-learning, the Q-value table is stored in the form of a table, but when the number of states of the problem is big, the size of the Q-table will be too large to be stored.

One way to solve the above problem is to approximate the value function. The approximation of the value function actually uses a function to estimate the value function. The input is the state $s$, and the output is the value corresponding to the state $s$. That is $Q(s,a) = f(s,a,w)$, where $w$ represents the introduced parameter. Linear combinations, neural networks, or other methods can be used as approximation functions. This paper uses BP neural network as an approximation function.

4.2 The process of DQN algorithm
Neural network training is an optimization problem that optimizes a loss function, which is the deviation of the label and network output. The goal is to minimize the loss function. To do this, we need samples and a large amount of tagged data, and then use the gradient descent method to update the parameters of the neural network by backpropagation. Therefore, to train the Q-network, it is necessary to provide labeled samples for the Q-network, and the Q-learning algorithm can provide labeled samples. The loss function of Q-network training is the following equation (7) and the calculation process is shown in table 1.

$$L(w) = E[(r + \gamma \max_{a'} Q(s',a',w) - Q(s,a,w))^2]$$  \hspace{1cm} (7)

5. CASE ANALYSIS

5.1 Case parameters
This paper analyzes and verifies the hybrid system composed of a wind farm and energy storage system in Shaanxi. The total capacity of the wind farm is 45MW. The parameters of the energy storage system are shown in table 2.

In this paper, the stored energy of the energy storage system is divided into 20 discrete spaces at intervals of 2MJ, and the wind power is divided into three discrete spaces at intervals of 15MW. The control period includes 24 time periods in one day, so the whole state space contains 1920 states.

The planned charging and discharging power is discretized into 7 fixed values at intervals of 2.5MW, namely {-7.5, -5, -2.5, 0, 2.5, 5, 7.5}. It is assumed that the standard deviation of the wind power prediction error distribution is 5% of the predicted value, that is $\sigma_w = 5\% \cdot \hat{P}_{w,t}$, the confidence coefficients $\alpha, \beta$ are taken as 2.0. The up-down reserve capacity purchased from the grid is discretized into two fixed values, namely {0, $2\sigma_w$}, so the action strategy space contains a total of 28 actions.

5.2 Analysis of improved Q-learning algorithm
Set the parameter $\alpha=0.01, \beta=0.5, \gamma=0.8$. In the 12000h’s pre-learning phase, the average income per 240h of conventional Q-learning and improved Q-learning are shown in figure 3 below.

Apply the Q-value table trained in the pre-learning phase to the online learning, select the wind power data of 1920h, the parameter setting is unchanged, and get the average income of the conventional Q-learning and improved Q-learning every 240h as shown in table 3. It can be seen that the average return of the improved Q-learning in each period under the two parameter settings is significantly higher than the conventional Q-learning, and the gain in the whole operating loop is maximized.
5.3 Analysis of DQN algorithm

In the training phase, the same 12000h wind power historical data is used, and the trained Q-value network is put into online learning. The new historical wind power data of 1200h is used for verification, and the average income per 240h is obtained as shown in table 4. It can be seen that the DQN algorithm based on the improved Q-network further improves the average income in the phase of wind-storage coordinated scheduling.

Figure 4 show the fluctuations of the grid-connected power of the wind farm within 1920 hours before and after the implementation of the wind-storage cooperation, which are also the deviation between the planned power of the wind farm and the actual power of the wind farm. The comparison shows that the frequency and amplitude of frequency fluctuations have been significantly reduced after the implementation of cooperation. The deviation between planned power and actual power is 0 in some certain periods, which reduces the burden of grid regulation.

The following is a further description of the changes in the storage capacity of the energy storage system within 1920 hours after the implementation of cooperation. As can be seen from figure 5, the energy stored in the energy storage system can be maintained within the allowable range for most of the time period. When the energy storage reaches the upper or lower limit, the hybrid system will be punished.

6. CONCLUSION

1) Wind power and energy storage can constitute a hybrid system to participate in power trading. Through the cooperation, the influence of wind power uncertainty can be mitigated or eliminated, and the schedulability of wind power can be enhanced.

2) The improved Q-learning algorithm adds the operational objectives of the wind farm to the training process, and introduces the idea of simulated annealing algorithm. And each stage is backtracked. After comparing with the conventional Q-learning, the stage income are obviously improved, the number of punishments and the cost become much less.

3) The DQN algorithm uses BP neural network for training. Under certain parameter settings, it can achieve better control effects than conventional Q-learning and improved Q-learning and can solve the problem of excessive status.

| Table 1. Deep Q-learning with experience replay. |
|-----------------------------------------------|
| Initialize replay memory \(D\) to capacity \(N\) |
| Initialize action-value function \(Q\) with random weights \(\theta\) |
| Initialize target action-value function \(\hat{Q}\) with weights \(\theta'=\theta\) |
| For episode = 1,2…\(M\), do |
| Initialize sequence \(s_i = \{x_i\}\) and preprocessed sequence \(\phi_i=\phi(s_i)\) |
| For \(t=1,2,…,T\), do |
| With probability \(\varepsilon\) select a random action \(a_i\) |
| Otherwise select \(a_i = \text{arg max}_a Q(\phi(s_i),a;\theta)\) |
| Execute action \(a_i\) in emulator and observe reward \(r_i\) and image \(x_{i+1}\) |
| Set \(s_{i+1} = s_i, a_i, x_{i+1}\) and preprocess \(\phi_{i+1} = \phi(s_{i+1})\) |
| Store transition \((\phi_i,a_i,r_i,\phi_{i+1})\) in \(D\) |
| Sample random minibatch of transitions \((\phi, a, r, \phi)\) |
| Set \(y_j = \begin{cases} r_i & \text{if episode terminates at step } i+1 \\ r_i + \gamma \max_a \hat{Q}(\phi_i,a;\theta') & \text{otherwise} \end{cases} \) |
| Perform a gradient descent step on \((y_i - Q(\phi_i,a_i;\theta))^2\) |
with respect to the network parameters $\theta$.

Every $C$ steps reset $\hat{Q} = Q$.

\textbf{End for}

\textbf{End for}

Table 2. Energy storage system parameters.

| Max power | Max capacity | Min capacity | Charging efficiency | Discharging efficiency |
|-----------|--------------|--------------|---------------------|------------------------|
| $7.5 \, MW$ | $50 \, MW \cdot h$ | $10 \, MW \cdot h$ | $0.85$ | $1.0$ |

Table 3. Average income per 240h (unit: $).$

| $\beta$ | 0.5 | 1.0 |
|---------|-----|-----|
| Algorithm | Conventional Q | Improved Q | Conventional Q | Improved Q |
| Period 1 | 3056.2 | 3705.8 | 4130.2 | 4267.9 |
| Period 2 | 2447.8 | 3239.1 | 3809.6 | 4067.6 |
| Period 3 | 3140.2 | 3480.0 | 3855.3 | 3554.3 |
| Period 4 | 3603.1 | 4048.1 | 4051.2 | 4432.1 |
| Period 5 | 2589.2 | 3576.7 | 3463.1 | 3811.7 |
| Period 6 | 2474.5 | 3450.9 | 2930.2 | 3740.3 |
| Period 7 | 2808.0 | 3118.7 | 3564.9 | 4170.8 |
| Period 8 | 3408.8 | 3936.2 | 3728.6 | 4061.4 |

Table 4. DQN algorithm average income per 240h (unit: $).$

| Network type | Conventional Q | Improved Q |
|--------------|----------------|------------|
| Period 1     | 4141.3         | 4274.9     |
| Period 2     | 3510.8         | 4350.2     |
| Period 3     | 3957.5         | 5294.8     |
| Period 4     | 4347.3         | 4967.8     |
| Period 5     | 3793.1         | 4880.0     |
| Period 6     | 3389.4         | 4980.4     |
| Period 7     | 3585.9         | 4761.5     |
| Period 8     | 3718.7         | 4627.4     |

Figure 1. The realization mechanism of the cooperation
Figure 2. ESS compensates for wind power prediction error

Figure 3. (a) conventional Q  (b) improved Q

Figure 4. (a) before the cooperation  (b) after the cooperation

Figure 5. Stored energy curve in energy storage system

ACKNOWLEDGMENTS
This work are supported by the Science and Technology Foundation of SGCC (SGSDDK00KJJS1600061).

REFERENCES
[1] GE Ju, WANG Fei, ZHANG Lizi. Spanning Reserve Model in the Wind Power Intergrated Power System[J]. Automation of Electric Power Systems, 2010, 34(06): 32-36.
[2] WU Xiong, WANG Xiuli, LI Jun,et al. A Joint Operation Model and Solution for Hybrid Wind Energy Storage Systems[J]. Proceedings of the CSEE, 2013, 33(13): 10-17.
[3] HU Zechun, DING Huajie, KONG TAO. A Joint Daily Operational Optimization Model for Wind Power and Pumped-storage Plant[J]. Automation of Electric Power Systems, 2012, 36(2): 36-41.
[4] XIONG Meng, GAO Feng, ZHANG Haifeng, et al. Hybrid Energy Storage Based Generation Scheduling of Wind Farms[J]. *Power System Technology*, 2014, 38(5): 1242-1249.

[5] Garcia-Gonzalez J, DelM R M R, Santos L M, et al. Stochastic Joint Optimization of Wind Generation and Pumped-storage Units in an Electricity Market[J]. *IEEE Transactions on Power Systems*, 2008, 23(2): 460-468.

[6] Dicorato M, Forte G, Pisani M, et al. Planning and Operating Combined Wind-storage System in Electricity Market[J]. *IEEE Transactions on Sustainable Energy*, 2012, 3(2): 209-217.

[7] Dukpa A, Duggal I, Venkatesh B, et al. Optimal Participation and Risk Mitigation of Wind Generators in an Electricity Market[J]. *IET Renewable Power Generation*, 2010, 4(2): 165-175.

[8] YU Tao, HU Xibing, LIU Jing. Multi-objective Optimal Power Flow Calculation Based on Multi-step Q(4) Learning Algorithm[J]. *Journal of South China University of Technology (Natural Science Edition)*, 38(10): 139-145.

[9] LI Hongmei, YAN Zheng. Application of Q-learning Approach with Prior Knowledge to Nonlinear AGC System[J]. *Automation of Electric Power Systems*, 2009, 32(23): 36-40.

[10] LU Li. Strategic Bidding of Electricity Supplier in Competitive Market Based on Improved Q-learning[D]. *Shanghai Jiao Tong University*, 2007.

[11] GAO Zhan. A Supplier Bidding Strategy Based on Q-learning Model[D]. *Shanghai Jiao Tong University*, 2008.

[12] WANG Shang. Studies on Regulation and Storage Allocation Strategies of Wind Farm Self-discipline[D]. *Shandong University*, 2015.

[13] LIU Guojing, HAN Xueshan, WANG Shang, et al. Optimal Decision-Making in the Cooperation of Wind Power and Energy Storage Based on Reinforcement Learning Algorithm[J]. *Power System Technology*, 2017, 40(9): 2729-2736.

[14] LIU Quan, ZHAI Jianwei, ZHANG Zongchang, et al. A Survey on Deep Reinforcement Learning[J]. *Chinese Journal of Computers*, 2017, 40(1): 1-28.

[15] QIU Liwei. Application of Deep Reinforcement Learning in Video Game Playing[D]. *South China University of Technology*, 2015.

[16] SHEN Yifan, ZHAO Ning, XIA Mengjue, et al. A Deep Q-Learning Network for Ship Stowage Planning Problem[J]. *Scientific Research Academy of Shanghai Maritime University*, 2017.