

# Scheduled power tracking control of the wind-storage hybrid system based on the reinforcement learning theory

**Ze Li**

International Education Institute, North China Electric Power University (Baoding), China.

E-mail: LiZe\_NCEPU@163.com.

**Abstract.** In allusion to the intermittency and uncertainty of the wind electricity, energy storage and wind generator are combined into a hybrid system to improve the controllability of the output power. A scheduled power tracking control method is proposed based on the reinforcement learning theory and Q-learning algorithm. In this method, the state space of the environment is formed with two key factors, i.e. the state of charge of the energy storage and the difference value between the actual wind power and scheduled power, the feasible action is the output power of the energy storage, and the corresponding immediate rewarding function is designed to reflect the rationality of the control action. By interacting with the environment and learning from the immediate reward, the optimal control strategy is gradually formed. After that, it could be applied to the scheduled power tracking control of the hybrid system. Finally, the rationality and validity of the method are verified through simulation examples.

## 1 Introduction

Wind power offers several advantages such as matured technology, low cost, and zero carbon emissions. However, wind power is intermittent due to the wind speed fluctuations, which has brought enormous challenges to the power grid's operation and dispatch [1, 2]. In recent years, many researches have become reducing such adverse impact from the perspective of improving the self-discipline control of wind power plant.

Energy storage (ES) is generally rechargeable and can provide a fast response. A combination of ES and wind power can make the hybrid system's output more controllable, stable and planned. So, the cooperative and self-discipline control of ES and intermittent power has aroused extensive attention [3-7]. While, the wind power is extremely uncertain, and the regulation of ES is subjected to multiple constraints such as storage capacity, power output limit, and so on. Accordingly, the scheduled power tracking control of the hybrid system is particularly complex, which can hardly be addressed using the conventional analytical optimization methods.

In reinforcement learning theory [8, 9], the learning system interacts with the environment, during which it can acquire the feedback information continuously and thus develop the decision-making capacity regarding the issue. Recent years, it has been widely used in multiple fields of power system and played an active role in decision making for uncertainty problems. In [10], Q-learning algorithm was used to solve the problem of power flow control with constraints and the performance of the algorithm was analyzed in detail. A practical voltage control method in regional power grid was proposed in [11], in which the prior knowledge and reinforcement learning theory were combined to improve the control effect. In [12], the  $Q(\lambda)$  learning method with multi-step forecasting capability was applied for the multi-objective optimal power flow, yielding improved convergence speed. In [13],



the reinforcement learning theory was used in micro-grid energy control, in which charging and discharging of the ES in future period was decided based on the wind power forecast. In [14], based on Q-learning method, the forecasted wind power and the current state of charge (SOC) of ES were used to determine the future planned power and reserve capacity, and the applicability of reinforcement learning theory to stochastic problems was verified.

Based on the above studies, the wind generator and ES were combined to form a wind-storage hybrid system (WSHS) in this paper, which participated in the electricity market as an individual unit. The Q-learning algorithm is adopted to enable the hybrid system to continuously improve its self-control ability through its interaction with the environment. It is shown that the method proposed could make good use of the fast charging/discharging characteristics of the ES and take into account the relevant constraints and control requirements. As a result, the WSHS could follow the scheduled power as close as possible, reducing the reserve cost and penalty cost.

The paper is organized as follows. Section 2 introduces the cooperation mechanism and mathematical model of the WSHS in electricity market. Next, the basic principle of the reinforcement learning theory and Q-learning algorithm are described in Section 3. In section 4, a WSHS control strategy training method is established based on the Q-learning algorithm. Finally in Section 5, the applicability of the proposed method for the planned power tracking control of the WSHS is verified through numerical simulation.

## 2 Cooperation Model of WSHS

### 2.1 Cooperation mechanism of WSHS

As shown in Fig.1, the wind power and ES are combined to form a WSHS, which can participate in the electricity market as an individual. The WSHS declares its generation scheduling and required reserve capacity to the power grid in advance. During the operation, the WSHS tries to track the pre-set generation scheduling as much as possible through self-regulation. However, due to the uncertainty of the wind power, deviation between the actual power and scheduled power will appear inevitably. Therefore, the power grid should provide the reserve and the WSHS has to pay for it. If the deviation is so great that it exceeds the declared reserve capacity, the corresponding penalty cost should be paid.

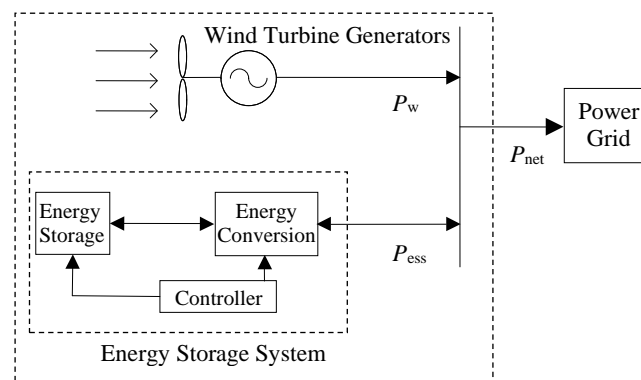


Figure 1: Schematic of WSHS.

In order to track the generation scheduling as closely as possible and to reduce the reserve and penalty cost, the WSHS must make full use of the fast power regulation feature of the ES and control the charging/discharging power of the ES reasonably. So that the power output deviation should be restricted within a reasonable limit. By this, not only the reserve capacity cost can be reduced, but also the negative impact on the power grid can be relieved.

### 2.2 Mathematical model of the WSHS in electricity market

Mathematical model and market trading model of the WSHS are established in this section. The key factor of WSHS controlling is the control of the ES. For time period  $t$ , the scheduled power declared to

the power grid by WSHS is  $P_{net,t}^{plan}$ , and the reserve capacity is  $R_t$ . In order to achieve maximum utilization of the wind power generation, maximum power tracking strategy is still used in the wind generator. That is, the wind power output is determined entirely by wind speed and uncontrollable. Therefore, the actual power output will deviate from the scheduled power inevitably.

$$P_{net,t}^{cha} = P_{w,t}^{real} - P_{net,t}^{plan} \quad (1)$$

In equation (1),  $P_{net,t}^{cha}$  is the uncertainty that need to be eliminated or reduced. The greater its absolute value, the more reserve cost is required. With the wind-storage cooperation adopted, the power output deviation can be compensated by controlling the charging/discharging power  $P_{ess,t}$  of ES:

$$\Delta P_{net,t} = P_{ess,t} + P_{w,t}^{real} - P_{net,t}^{plan} \quad (2)$$

Where:  $P_{ess,t}$  is the output power of the ES,  $\Delta P_{net,t}$  is the output power deviation of the WSHS.

After the determination of  $P_{ess,t}$  according to control strategy and the state of the environment, the actual income of the WSHS in time interval  $t$  can be calculated according to the actual operation of the WSHS:

$$I(t) = B_{pw}(t) - P_{unb}(t) - P_{res}(t) \quad (3)$$

In equation (3),  $I(t)$  is the actual total income of the WSHS,  $B_{pw}(t)$  is the electricity selling income,  $P_{unb}(t)$  is the reserve cost due to the deviation between the hybrid system's actual output and the scheduled output, including reserve cost and penalty cost,  $P_{res}(t)$  is the cost of the pre-ordered reserve capacity by the WSHS. Since the pre-ordered reserve capacity is determined before the controlling procedure, this value is fixed and unchangeable with the control variables.

The individual components in equation (3) may be expressed as follows:

$$B_{pw}(t) = \lambda_{ele,t} (P_{w,t}^{real} + P_{ess,t}) \cdot \Delta t \quad (4)$$

$$P_{unb}(t) = (\lambda_{resE,t} \cdot \Delta P_{net,t}^2 + C_{res}) \cdot \Delta t \quad (5)$$

$$P_{res}(t) = R_t \cdot \lambda_{resC,t} \cdot \Delta t \quad (6)$$

Where:  $\lambda_{ele,t}$  is the power purchase price of the power grid from the hybrid system in period  $t$ ,  $\lambda_{resE,t}$  is the reserve power price coefficient in period  $t$ ,  $C_{res}$  is the penalty cost to the WSHS by power grid in case the output power exceeds the declared reserve capacity,  $R_t$  is the reserve capacity pre-ordered by WSHS for period  $t$ ,  $\lambda_{resC,t}$  is the reserve capacity price in period  $t$ .

$C_{res}$  in equation (5) can be represented as:

$$C_{res} = \begin{cases} (|\Delta P_{net,t}| - R_t) \cdot \lambda_{Cres} & (|\Delta P_{net,t}| > R_t) \\ 0 & (|\Delta P_{net,t}| \leq R_t) \end{cases} \quad (7)$$

Where:  $\lambda_{Cres}$  is the reserve penalty coefficient.

From equation (2), it is known that if the power and capacity of ES are unlimited, the actual output power can strictly track the planned curve, that is  $\Delta P_{net,t} = 0$ . However, restricted by economic and technical conditions, the ES is still a scarce resource and wind farms cannot configure ES unlimitedly to ensure fully tracking of the scheduled power. So the operation of ES must meet the operation constraint requirements. In the process of charging and discharging, it needs to satisfy the constraint of the energy balance, that is:

$$\begin{cases} E_{ess,t+1} = E_{ess,t} - P_{ess,t} \cdot \eta_{ch} \cdot \Delta t & P_{ess,t} \geq 0 \\ E_{ess,t+1} = E_{ess,t} - P_{ess,t} / \eta_{dis} \cdot \Delta t & P_{ess,t} < 0 \end{cases} \quad (8)$$

Where:  $E_{ess,t}$  is the electricity stored in the ES at the beginning of the time period  $t$ ,  $E_{ess,t+1}$  is the electricity stored in the ES at end of the time period  $t$ .  $P_{ess,t}$  is the active power of ES in time period  $t$ , in which positive values represent discharging power and negative values represent absorbing power.  $\eta_{ch}$  and  $\eta_{dis}$  are charging and discharging efficiency of ES respectively,  $\Delta t$  is the duration of the time period  $t$ .

In addition, the ES is required to meet the constraints of the power and capacity:

$$P_{ess\min} \leq P_{ess,t} \leq P_{ess\max} \quad (9)$$

$$E_{ess\min} \leq E_{ess,t} \leq E_{ess\max} \quad (10)$$

Where:  $P_{ess\max}$  and  $P_{ess\min}$  is the maximum and minimum active power of ES, respectively,  $E_{ess\max}$  and  $E_{ess\min}$  is the maximum and minimum capacity of ES, respectively.

Therefore, optimizing the operation strategy of the ES according to actual operation state and determining the appropriate output power  $P_{ess,t}$  to maximize these utility are the key points to improve the economic operation quality of the WSHS. As can be seen from the model, the complexity of the problems involved in this paper lay mainly in the uncertainty of wind generator power output, complexity of the ES model and the time period relevance caused by the capacity constraints of the ES. The interaction effects of these issues poses great difficulties to determine the optimal control strategy of the WSHS, which cannot be effectively solved by traditional optimization methods.

### 3 Principle of the Reinforcement Learning Theory and Q-learning Algorithm

In a reinforcement learning algorithm, through the interaction between the learning agent and its environment, the learning agent continually receives feedback and gradually upgrades its action strategy with the passage of time. After a specific period of time, the agent could achieve a decision-making capacity for a given problem. The learning process of the reinforcement learning only requires its own experience, thus showing certain advantages for solving control problems with uncertainty.

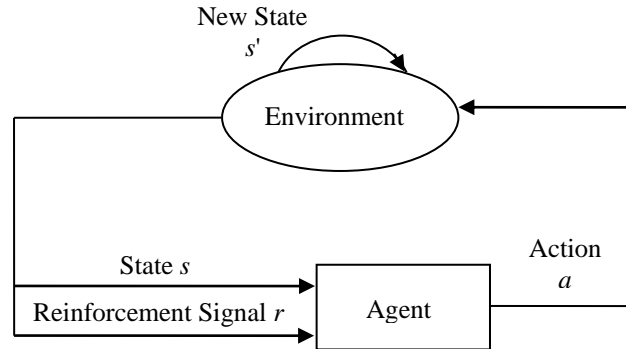


Figure 2: Reinforcement learning model.

Figure 2 shows a standard reinforcement learning model. The reinforcement learning agent receives the environmental state input  $s$ , and generates the corresponding action  $a$  according to internal mechanism. Under the action  $a$ , the environment transits to a new state  $s'$ , meanwhile produces a reinforcement signal  $r$  back to the agent. According to the reinforcement signal  $r$ , the agent modifies its internal mechanism and generates the next step action according to the new state  $s'$ . The ultimate goal is to make the agent achieve the maximal cumulative reward  $R_t$ , which is the accumulation of all immediate reward values  $r$  during a long time period.

$$R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \quad (11)$$

Where:  $\gamma$  is the discount factor.

Q-learning is a model-free reinforcement learning algorithm. During the iteration, a value function of the state-action pairs  $Q(s, a)$  is taken as the evaluation function. The specific meaning of  $Q(s, a)$  can be expressed as: under the state  $s$  and after the selection of action  $a$ , the expected cumulative reward under strategy  $\pi$ .

$$Q^\pi(s, a) = E_\pi \left\{ \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \mid s_t = s, a_t = a \right\} \quad (12)$$

The iteration process of Q-learning is as follows:

$$Q^{k+1}(s_k, a_k) = Q^k(s_k, a_k) + \alpha \begin{bmatrix} r_{k+1}(s_k, a_k, s_{k+1}) + \\ \gamma \max_{a' \in A} Q^k(s_{k+1}, a') \\ - Q^k(s_k, a_k) \end{bmatrix} \quad (13)$$

Where:  $s_k$  is the state of the environment during the  $k$ th iteration,  $a_k$  is the action the agent selects during the  $k$ th iteration,  $Q^k(s_k, a_k)$  is the state-action value function,  $\alpha$  is the learning factor.

Under each state the agent shall select an action according to the Q value table. If actions corresponding to the maximum Q value are always selected, it is called greedy strategy and expressed as  $\pi^*$ :

$$\pi^* = \arg \max_{a \in A} Q(s, a) \quad (14)$$

If greedy strategy is used for each action selection, it means that current knowledge of the value of the actions is being exploited. While, if non-greedy actions are select, which is called exploring, this enables us to improve the estimate of the non-greedy actions' value. Exploitation can help us maximize the expected reward on the one way, but exploration may produce the greater total reward in the long run. So exploration and exploitation must be balanced for particular problems [9].

## 4 Scheduled Power Tracking Control of the WSHS Based on Q-learning

### 4.1 State set of the environment

In this paper,  $E_{ess,t}$ , the energy stored in the ES at beginning of period  $t$ , and  $\Delta P_{net,t}$ , the difference between the actual wind power and hybrid system's planned power, are taken as the state variables. These two variables are divided in form of intervals, respectively. The Cartesian product of these two discrete variables forms the state set  $S$  of the environment.

The capacity of the ES is divided into  $m$  equal intervals, each of which has a length of  $\Delta E_{ess}$ :

$$\Delta E_{ess} = \frac{E_{ess \max} - E_{ess \min}}{m} \quad (15)$$

The divided  $m$  intervals of the ES include:  $[E_{ess \min}, E_{ess \min} + \Delta E_{ess})$ ,  $[E_{ess \min} + \Delta E_{ess}, E_{ess \min} + 2\Delta E_{ess})$ , ...,  $[E_{ess \max} - \Delta E_{ess}, E_{ess \max}]$ .

$P_{net,t}^{cha}$ , the difference between the actual wind power and scheduled power, is also divided into intervals with the length of  $\Delta P_{net}^{cha}$ , which is set according to the requirements. Thus the  $n$  intervals of  $P_{net,t}^{cha}$  are  $(-\infty, -n \cdot \Delta P_{net}^{cha})$ ,  $[-(n/2-1) \cdot \Delta P_{net}^{cha}, -(n/2-2) \cdot \Delta P_{net}^{cha})$ , ...,  $[n/2 \cdot \Delta P_{net}^{cha}, +\infty)$ . The  $n$  is an even value.

Hence, the environmental state of the WSHS is divided into  $m \times n$  states:

$$S = \{s_1, s_2, \dots, s_{m \times n}\} \quad (16)$$

The smaller the state is divided, the more accurate the state description of the environment will be. However, it will make the number of elements in the state set so large that the learning process becomes too long to be applicable for online control. Therefore, an appropriate value should be determined according to actual requirements and experience.

#### 4.2 Feasible action set of the agent

The ES output power  $P_{ess,t}^a$  is taken as the action of the control system.  $P_{ess,t}^a$  is divided into  $b$  feasible actions with interval length of  $\Delta P_{ess} : P_{ess \min}, P_{ess \min} + \Delta P_{ess}, \dots, P_{ess \min} + (b-2)\Delta P_{ess}, P_{ess \max}$ , which compose the feasible action set  $A$ :

$$A = \{a_1, a_2, \dots, a_{2b+1}\} \quad (17)$$

In actual operation, due to the ES capacity constraint, the output power obtained according to the action selection strategy may not be able to be performed. To determine the actual output power of ES, the ES capacity limit needs to be verified to determine the actual output power of ES.

If  $P_{ess,t}^a > 0$ , it should be checked that whether the capacity lower limit of the ES will be exceeded after the ES output the power  $P_{ess,t}^a$ . If the limit is exceeded, the actual output power of ES should be adjusted:

$$P_{ess,t} = \begin{cases} P_{ess,t}^a & \left( E_{ess,t} - \frac{P_{ess,t}^a}{\eta_{dis}} \Delta t \geq E_{ess \min} \right) \\ \frac{E_{ess,t} - E_{ess \min}}{\Delta t} \cdot \eta_{dis} & \left( E_{ess,t} - \frac{P_{ess,t}^a}{\eta_{dis}} \Delta t < E_{ess \min} \right) \end{cases} \quad (18)$$

In a similar way, if  $P_{ess,t}^a < 0$ , it should be checked that whether the capacity upper limit of the ES will be exceeded after the ES absorbs the power  $P_{ess,t}^a$ . If the limit is exceeded, the actual output power of ES should be adjusted:

$$P_{ess,t} = \begin{cases} P_{ess,t}^a & (E_{ess,t} - P_{ess,t}^a \eta_{ch} \Delta t \leq E_{ess \max}) \\ \frac{E_{ess \max} - E_{ess,t}}{\eta_{ch} \cdot \Delta t} & (E_{ess,t} - P_{ess,t}^a \eta_{ch} \Delta t > E_{ess \max}) \end{cases} \quad (19)$$

#### 4.3 Immediate reward

In this paper, the main purpose of the ES in the WSHS is to reduce the reserve cost. Further considering the ES operating condition, control cost and so on, the immediate reward of the WSHS at period  $t$  is defined as follows:

$$r_{t+1}(s_t, a_t) = P_{umb}(t) + C_a(t) \quad (20)$$

$C_a(t)$  is the penalty in time period  $t$  if the hybrid system fails to perform the selected action due to the capacity limitation of ES. Considering the penalty within the immediate reward can improve the feasibility of the selected action.

$$C_a(t) = \begin{cases} k_{Ca} & (P_{ess,t}^a \neq P_{ess,t}) \\ 0 & (P_{ess,t}^a = P_{ess,t}) \end{cases} \quad (21)$$

Where:  $k_{Ca}$  is the penalty constant.

#### 4.4 Action selection strategy

In the process of learning, the control action  $a_t$  should be selected based on the current state  $s_t$  and the value function  $Q(s, a)$ . If each selection is based on the action corresponding to the maximal Q-value, i.e. the greedy strategy is conducted, then the feasible action set cannot be fully explored and the learning may converge to local solution. If the randomness of the selection is too strong, then the learning speed will be too slow to form the final control strategy. Hence, a gradient action selection strategy is adopted in this paper.

- 1) Generate a random number  $R_a$ ,
- 2) Calculate the probability threshold  $T_H$ ,



$$T_H = \begin{cases} 1 & (N_n \leq 0.1N_t) \\ \varepsilon & (0.1N_t \leq N_n \leq 0.9N_t) \\ 0 & (0.9N_t \leq N_n \leq N_t) \end{cases} \quad (22)$$

Where:  $N_n$  is the number of action selections which have been performed,  $N_t$  is the total number of action selections during the training process,  $\varepsilon$  is a constant between 0 and 1.

3) Select the action.

If  $R_a < T_H$ , select an action from the feasible action set randomly with equal probability. If  $R_a \geq T_H$ , then select the action according to the greedy strategy, which is:

$$a_t = \arg \max_{a_t \in A} Q^t(s_t, a_t) \quad (23)$$

According to action selection strategy, at the beginning of the learning, the action is selected in random strategy. Then,  $\varepsilon$ -greedy strategy is used to perform the greedy strategy with certain probability. As soon as the action selection number reaches  $0.9N_t$ , greedy strategy will be conducted. The value of the  $N_t$  needs to be set according to the actual training effect and experience.

## 5 Simulation Example

### 5.1 Model of the simulation example

A WSHS composed of 55 MW wind turbine generators and 10MW ES is used for simulating and verifying the method proposed in this paper. The maximum charging and discharging power of the ES are both 15 MW, while the lower and upper limits of the capacity are 2MWh and 12MWh, respectively. The charging and discharging efficiency are both 0.9 and the initial SOC of the ES is 7MWh. The reserve capacity cost coefficient of the hybrid system is 150 Yuan/MWh<sup>2</sup>, and the reserve capacity penalty coefficient is 1000 Yuan/MWh. The WSHS declares its scheduled power and reserve capacity in accordance with 24 hours a day.

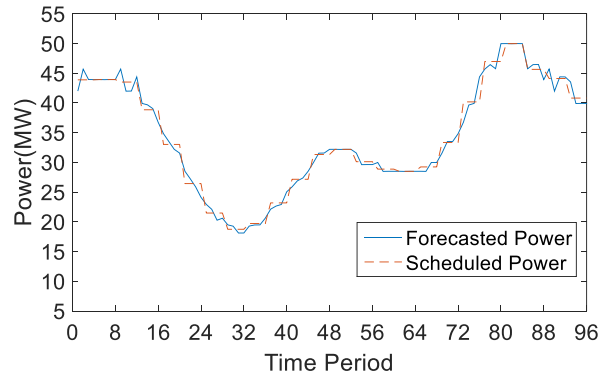


Figure 3. Forecasted power and scheduled Power

Based on the wind power forecasting data of every 15 minutes within a day, the average value of the 4 prediction data points within each hour is taken as the scheduled power of the WSHS for that hour. The reserve capacity is 6MW during the whole day. The wind power forecasting data and the scheduled output power within a day are shown in Fig.3. Assuming that the wind power prediction values satisfy the normal distribution with the predicted value as the average value and 15% of the predicted value as the standard deviation, then the actual wind power output in each period can be simulated.

The difference between the actual wind power and scheduled power is divided into thirty two states with 1 MW steps, which are  $(-\infty, -15]$ ,  $(-15, -14]$ , ...,  $(14, 15]$  and  $(15, +\infty)$ . The SOC of the ES is divided into ten states with 1 MW steps, which are  $(2, 3]$ ,  $(3, 4]$ , ...,  $(10, 11]$  and  $(11, 12]$ . Therefore, the number of the environment state is  $32 \times 10 = 320$ .

Thirty one feasible actions are defined in the feasible action set, and the output power of the ES could be -15, -14, -13, ..., -1, 0, 1, ..., 14, 15.

### 5.2 Simulation results

The learning factor is set as 0.01 and the discount factor is set as 0.5. The operation process within one day is simulated with 300 iterations. The random action selection strategy is used for the first thirty iterations, the  $\epsilon$ -greedy strategy with probability threshold of 0.1 is used for the 31<sup>th</sup>-269<sup>th</sup> iterations, and the greedy strategy is applied for the last thirty iterations.

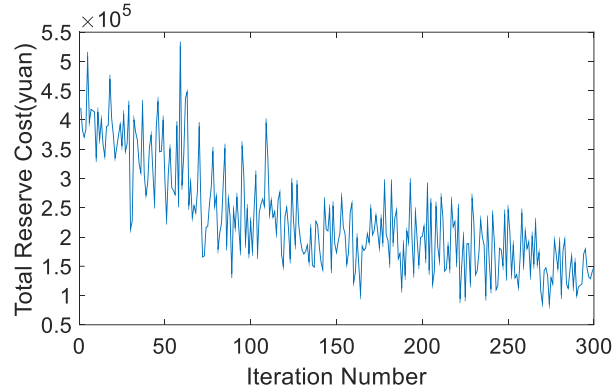


Figure 4: Total reserve cost of the WSHS during the learning process

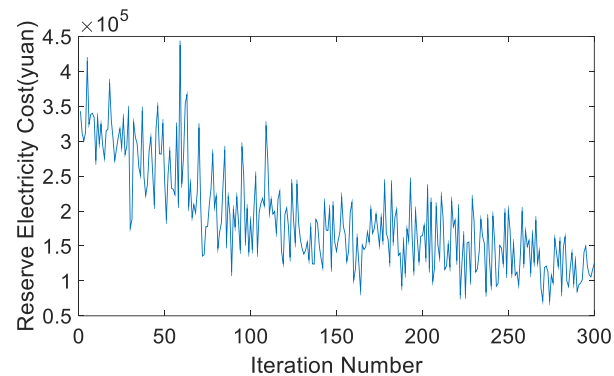


Figure 5: Reserve electricity cost of the WSHS during the learning process

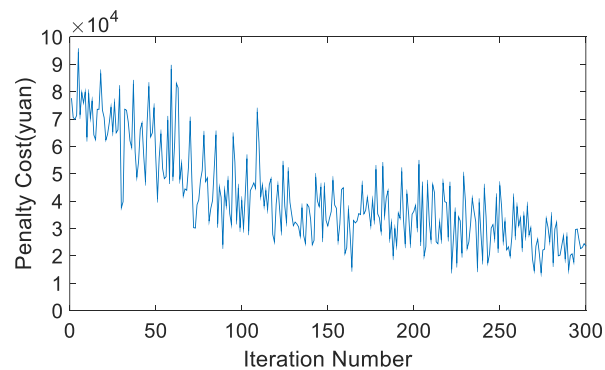


Figure 6: Penalty cost of the WSHS during the learning process

Fig.4, Fig.5 and Fig.6 are the changing curves of the total reserve cost, reserve electricity cost and penalty cost in the process of simulation, respectively. The horizontal axis represents the number of iterations, which is the number of simulations of the WSHS operation process within one day. The vertical axis is the actual cost of the WSHS within the day.



From Fig.4-6, it can be seen that, although all the reserve costs cannot be fixed at a certain optimal value due to the uncertainty of the wind power, the costs present an overall downtrend along the advance of the learning process under the effect of reinforcement learning.

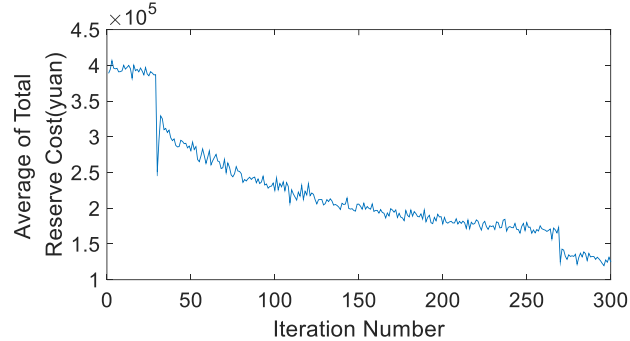


Figure 7: Average of the total reserve cost during the learning process.

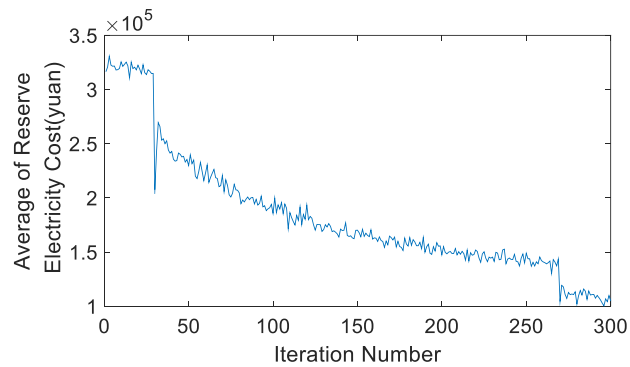


Figure 8: Average of the reserve electricity cost during the learning process.

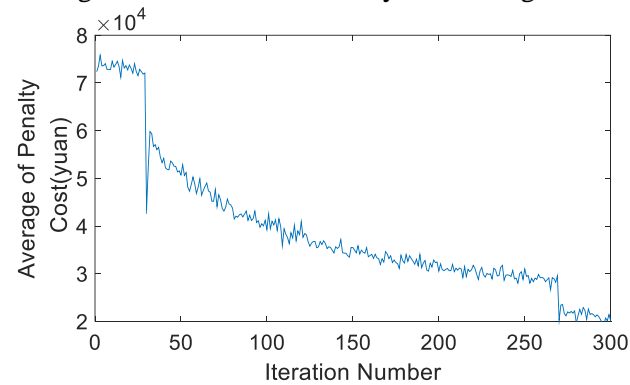


Figure 9: Average of the penalty cost during the learning process

To eliminate the disturbance of the wind power randomness to the training effect evaluation, the training process is repeated for a hundred times, and then, the average data in the corresponding simulation phase is achieved, which provides the training effect of the reinforcement learning. The simulation results are shown in Fig.7-9. In the first thirty iterations of the reinforcement learning, all kinds of costs show no change as the completely random strategy is adopted. Then, when the  $\epsilon$ -greedy strategy is used, all kinds of reserve costs present an obvious descending trend as the training progresses. So the effectiveness of the reinforcement learning algorithm is verified. After that, once the training reaches a certain degree, the greedy strategy is used for the simulation, which makes the reserve cost stay at a low level constantly. As can be seen, with the gradual accumulation of the learning experience, a good control strategy for the WSHS is formed gradually, enabling the system to track the predetermined scheduled power and reducing the reserve electricity cost and penalty cost greatly.

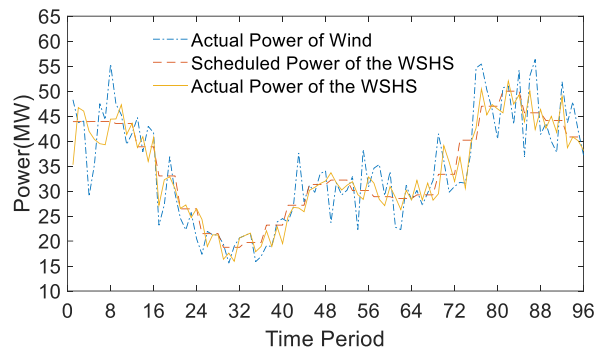


Figure 10: Control effect curve of the WSHS

Simulate the operation process of the WSHS within a day according to greedy strategy and the Q-value function achieved from learning. The comparison curves for the actual wind power, scheduled power and the actual power of the WSHS are shown in Figure 10. It can be seen that the ES can effectively reduce the deviation between the actual wind power and the scheduled power of the WSHS, making the actual power of the WSHS track the scheduled power as close as possible.

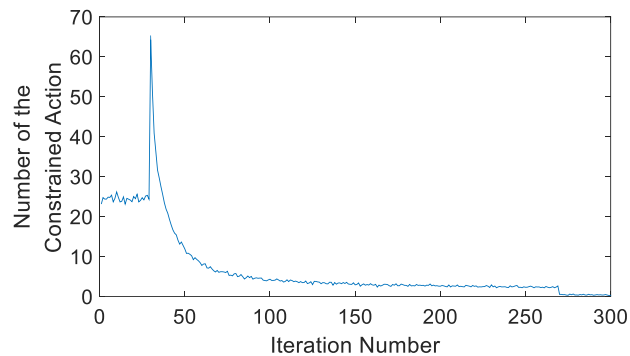


Figure 11: Average number of the constrained action during the learning process

Figure 11 shows the average number of the control action constrained by the ES capacity within a day. It can be seen that the changing trend is similar to that of the reserve costs. In the first thirty iterations, due to the use of the random strategy, the number is almost constant. After that, due to the adoption of  $\epsilon$ -greedy strategy, the number of constraining decreases gradually with the iteration process. In the last thirty iterations, due to the usage of greedy strategy, the number of constraining keeps at a relatively low value stably. This indicates that by considering the relative indicators in the immediate reward, effective adjustment of control strategy can be realized.

## 6 Conclusion

In this paper, a WSHS was established by combing the ES with the wind turbine generator. With the scheduled power and reserve capacity determined, using ES as the regulation equipment, a training method of the ES control strategy was established based on reinforcement learning theory and Q-learning algorithm. The feasibility and effectiveness of the method is verified by a simulation example. In addition, considering the variety of problems such as ES in operation and wind power prediction error that faced in actual operation, further researches on various constraints, indexes and training methods are required to make the proposed method more practical.

## References

- [1] Li Haibo, Lu Zongxiang, Qiao Ying, et al. "Assessment on operational flexibility of power grid with grid-connected large-scale wind farms", *Power System Technology*, **39**(6), pp. 1672-1678, (2015).
- [2] Lin Li, Sun Caixin, Wang Yongping, et al. "Calculation analysis and control strategy for voltage stability of power grid with large capacity wind farm interconnected", *Power System*

- Technology*, **32(3)**, pp. 41-46, (2008).
- [3] Li Bei, GuoJianbo. "A control strategy for battery energy storage system to level wind power output", *Power System Technology*, **36(8)**, pp. 38-43, (2012).
  - [4] Zhao Shuqiang, Liu Chenliang, Wang Mingyu, et al. "Chance-constrained programming based day-ahead optimal scheduling of energy storage", *Power System Technology*, **37(11)**, pp. 3056-3059, (2013).
  - [5] Jiang Zhe, Han Xueshan, Li Zhimin, et al. "Dependent Chance Goal Programming of Wind-EVBSS Considering its Multiple Benefits", *Power System Technology*, **40(4)**, pp. 1134-1139, (2016).
  - [6] Wu Xiong, Wang Xiuli, Li Jun, et al. "A Joint Operation Model and Solution for Hybrid Wind Energy Storage Systems", *Proceedings of the CSEE*, **33(13)**, pp.10-17, (2013).
  - [7] Maria Dicorato, Giuseppe Forte, Mariagiovanna Pisani, et al. "Planning and Operation Combined Wind-Storage System in Electricity Market", *IEEE Transactions on Sustainable Energy*, **3(2)**, pp. 209-217, (2012).
  - [8] L P Kaelbling, M L Littman, and A W Moore, "Reinforcement learning: A survey", *Journal of Artificial Intelligence Research*, **4**, pp.237-285, (1996).
  - [9] R S Sutton, A G Barto, "Reinforcement learning: An introduction", *Adaptive Computations and Machine Learning*, (1998).
  - [10] John G Vlachogiannis, Nikos D Hatziargyriou. "Reinforcement Learning for Reactive Power Control", *IEEE Transactions on Power Systems*, **19(3)**, pp. 1317-1325, (2004).
  - [11] Diao Haoran, Yang Ming, Chen Fang, et al. "Reactive power and voltage optimization control approach of the regional power grid based on reinforcement learning theory", *Transactions of China Electrotechnical Society*, **30(12)**, pp.408-414, (2015).
  - [12] Yu Tao, Hu Xibing, Liu Jing. "Multi-Objective Optimal Power Flow Calculation Based on Multi-Step  $Q(\lambda)$  Learning Algorithm", *Journal of South China University of Technology (Natural Science Edition)*, **38(10)**, pp. 139-145, (2010).
  - [13] Elizaveta Kuznetsova, Yan-Fu Li, et al. "Reinforcement learning for microrid energy management", *Energy*, **59**, pp. 133-146, (2013).
  - [14] 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", *Power System Technology*, **40(9)**, pp. 2729-2736, (2016).