Electrical Power and Energy Systems 119 (2020) 105928

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

A deep reinforcement learning method for managing wind farm uncertainties through energy storage system control and external reserve purchasing

J.J. Yang[a,b], M. Yang[a,*], M.X. Wang[a], P.J. Du[a], Y.X. Yu[a]

[a] Key Laboratory of Power Intelligent Dispatch and Control (Shandong University), Ministry of Education, Jinan 250061, China
[b] State Grid Shaanxi Electric Power Research Institute, Xi'an, 710054, China

ARTICLE INFO

Keywords:
Deep reinforcement learning
Energy storage system
Optimal controller
Rainbow
Reserve
Wind power producer

ABSTRACT

In deregulated environment, the wind power producers (WPPs) will face the challenge of how to increase their revenues under uncertainties of wind generation and electricity price. This paper proposes a method based on deep reinforcement learning (DRL) to address this issue. A data-driven controller that directly maps the input observations, i.e., the forecasted wind generation and electricity price, to the control actions of the wind farm, i.e., the charge/discharge schedule of the relevant energy storage system (ESS) and the reserve purchase schedule, is trained according to the method. By the well-trained controller, the influence of the uncertainties of wind power and electricity price on the revenue can be automatically involved and an expected optimal decision can be obtained. Furthermore, a targeted DRL algorithm, i.e., the Rainbow algorithm, is implemented to improve the effectiveness of the controller. Especially, the algorithm can overcome the limitation of the conventional reinforcement learning algorithms that the input states must be discrete, and thus the validity of the control strategy can be significantly improved. Simulation results illustrate that the proposed method can effectively cope with the uncertainties and bring high revenues to the WPPs.

1. Introduction

In recent years, as the most economically promising new energy generation form, wind power accounts for an increasing proportion of the total power generation [1]. Meanwhile, with the deregulation of the electricity sector, wind power producers (WPPs) must pursue their maximum profit in the market. However, the uncertainties of wind power and electricity price pose challenges for WPPs, because 1) the uncertainty of wind power may cause the deviation between the scheduled and actual wind generation, which may entail extra penalties on the WPPs [2]; and 2) the uncertainty of electricity price makes it not economic enough for the energy storage system (ESS) equipped in the wind farm to adjust the generation, reducing the revenues of the WPPs [3]. Therefore, it is crucial for WPPs to optimize their control actions of the wind farms and reduce the losses caused by the uncertainties.

In order to optimize the control actions of wind farms, [4] characterized the uncertainty of wind power in terms of a representative set of weighted scenarios, and proposed a two-stage stochastic programming framework to decide the reserve purchase plan of the wind farm. In [5], the pumped-hydro-storage plant was controlled to alleviate the uncertainty of the integrated wind power, and a chance-constrained optimization model for controlling the pumped-hydro-storage plant was presented. In the model, the wind power forecast error was characterized by the Gaussian distribution. [6] treated wind power as a stochastic variable depended on the normally distributed wind speed, and presented a dynamic programming algorithm to control the ESS in the wind farm to ensure the power generation schedule of the wind farm. [7] assumed that the wind power forecast error obeyed a joint normal distribution, and optimized the operation of the pumped-storage power plant based on the sampled scenarios to minimize the penalty costs caused by wind power forecast errors. The research works in [4–7] belong to the stochastic optimal control method, which generally assumed that the distribution of wind power was pre-known. In [8], a risk measure based robust optimization was proposed to optimize the bidding strategy of the wind farm with an ESS to optimize its revenue via electricity sales. However, in robust optimization, the uncertainty of wind power might be considered conservatively and thus lead to uneconomic decision results.

In previous studies, the uncertainty of wind power is approximately assumed to follow a certain distribution or is conservatively

* Corresponding author. Tel.: +86 15898916098.
E-mail address: myang@sdu.edu.cn (M. Yang).

https://doi.org/10.1016/j.ijepes.2020.105928
Received 11 April 2019; Received in revised form 2 October 2019; Accepted 6 February 2020
Available online 12 February 2020
0142-0615/

Nomenclature

ESS	Energy storage system
WPP	Wind power producer
$P_{w,t}$	Forecasted wind generation
$P_{w,t}^{act}$	Actual wind generation
$\Delta P_{w,t}$	Wind power forecast error
λ_t	Forecasted electricity price
λ_t^{act}	Actual electricity price released by market
P_t	Scheduled generation of the wind farm
P_t^{act}	Actual generation of the wind farm
ΔP_t	Deviation between P_t^{act} and P_t
S_t^{up}/S_t^d	Amount of the purchased up/down reserve
μ_t^{up}/μ_t^d	Price of the up/down reserve
$P_{ESS,t}$	Scheduled charge/discharge power of the ESS
$P_{ESS,t}^{act}$	Actual charge/discharge power of the ESS
r_t	Revenue generated by dispatch each time
C_t	Penalty fee charged by market due to ΔP_t
M_t	Maintenance cost of ESS
T	Number of time interval in an ESS control cycle
$P_{pen,t}$	Penalty price for power deviation
$P_{men,t}$	Price of the basic maintenance
$P_{extra\text{ - }men,t}$	Price of the extra maintenance
$P_{ESS,t}^{ch}/P_{ESS,t}^{dis}$	Charge/discharge power of the ESS
$\eta_{ESS}^{ch}/\eta_{ESS}^{dis}$	Battery charge/discharge efficiency
$P_{ESS,\max}^{ch}$	Maximum charge power of the ESS
$P_{ESS,\max}^{dis}$	Maximum discharge power of the ESS
E_t	Electricity stored in ESS
$E_{\max}/E_{\min}$	Maximum/minimum ESS capacity
E_{exp}	Expected electricity for the ESS to enter a new control cycle
N_{ESS}	Maximum state transition number in a control cycle of the ESS

characterized. These assumptions are rarely met for a variety of reasons in practice, which makes the proposed control methods not as effective as expected in actual scenarios. Different from previous studies, we do not try to describe the uncertainty laws but directly consider the impact of the uncertainties on the objective function, i.e., the WPP's revenue, through deep reinforcement learning (DRL). The complete uncertainty laws reflected in the historical dataset can be fully exploited and utilized by such a data mining technology. DRL has been widely used in the optimal control of power systems [9–13]. This paper investigates the use of DRL, in particular, the Rainbow algorithm, on coping with multiple uncertainties of wind power and electricity price, as well as to increase the WPP's long-term revenues. Furtherly, both ESS control and reserve purchase are used as control actions in the proposed approach, which is also rarely considered at the same time in prior researches.

In this paper, a DRL model about wind farm generation control is proposed to maximize WPP's revenue. The merits of the model are twofold. Firstly, the influence of the uncertainties of wind power and electricity price on the WPP's revenue can be automatically involved, without any assumption on the probability distribution rules during the optimization of the control strategy. Secondly, in addition to the ESS control, purchasing reserves from other market participates is also considered as an effective means of the operational control, which makes the model more comprehensive.

2. Overviews of the proposed method

The ESS control and reserve purchase are two effective means for a WPP to deal with wind generation uncertainties. In this paper, the optimal operation of a wind farm equipped with an ESS is investigated, whose schematic diagram is shown in Fig. 1. The charge/discharge of the ESS and the up/down reserve purchase are considered as control actions that should be optimized to increase the revenue of the WPP.

2.1. Cooperation rule between the ESS and reserve response

The ESS and reserve play different roles in accommodating wind generation uncertainties, and the cooperation rule between them is illustrated in Fig. 2.

In the figure, $\Delta P_{w,t}$ is the wind power forecast error at time interval t, $P_{r,t}$ is the probability of $\Delta P_{w,t}$, $P_{ESS,t}$ is the scheduled charge/discharge power of the ESS, $P_{ESS,t}^{act}$ is the actual charge/discharge power of the ESS, and S_t^{up}/S_t^{dn} is the amount of the purchased up/down reserve. Here, a positive value of $P_{ESS,t}^{act}$ or $P_{ESS,t}$ indicates that the ESS is in the state of discharge, and vice versa.

For a wind farm, the purchased reserve is preferentially used to balance the wind power forecast error, since the reserve has been paid beforehand. Meanwhile, if the reserve capacity is insufficient, the scheduled charge/discharge power of the ESS $P_{ESS,t}$ will be adjusted to further balance the error. According to this rule, the actual charge/discharge power $P_{ESS,t}^{act}$ can be expressed by

$$P_{ESS,t}^{act} = \begin{cases} P_{ESS,t} - (\Delta P_{w,t} - S_t^{up}), & \Delta P_{w,t} > S_t^{up} \\ P_{ESS,t}, & -S_t^{dn} \leqslant \Delta P_{w,t} \leqslant S_t^{up} \\ P_{ESS,t} - (\Delta P_{w,t} + S_t^{dn}), & \Delta P_{w,t} < -S_t^{dn} \end{cases} \quad (1)$$

Moreover, if $P_{ESS,t}^{act}$ calculated by (1) cannot be satisfied because of the physical constraints of the ESS, the wind power forecast error cannot be balanced completely and then the WPP will face a high penalty fee.

Here the scheduled generation of the ESS-integrated wind farm that should be submitted to the market before observing the real wind generation can be expressed as

$$P_t = P_{w,t} + P_{ESS,t} \quad (2)$$

where P_t is the scheduled generation of the wind farm, and $P_{w,t}$ is the forecasted wind generation.

And the actual wind farm generation can be calculated by

$$P_t^{act} = P_{w,t}^{act} + P_{ESS,t}^{act} \quad (3)$$

where P_t^{act} is the actual generation of the wind farm, and $P_{w,t}^{act}$ is the actual wind generation.

Under this cooperation rule, the WPP needs to make control decisions to maximize its revenue via optimizing the ESS control and reserve purchase. A corresponding DRL model will be introduced in the next subsection to help the WPP fulfill this purpose.

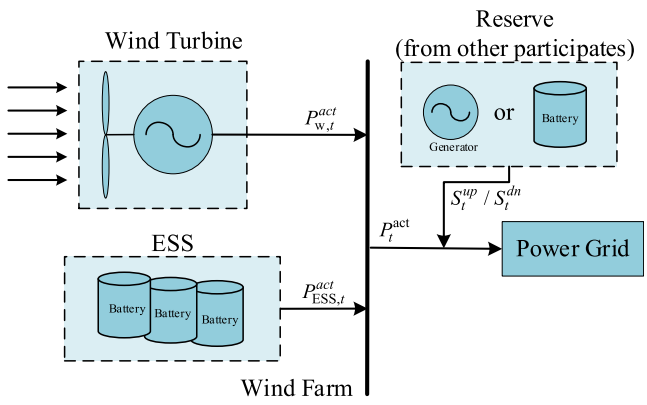


Fig. 1. A wind farm with ESS.

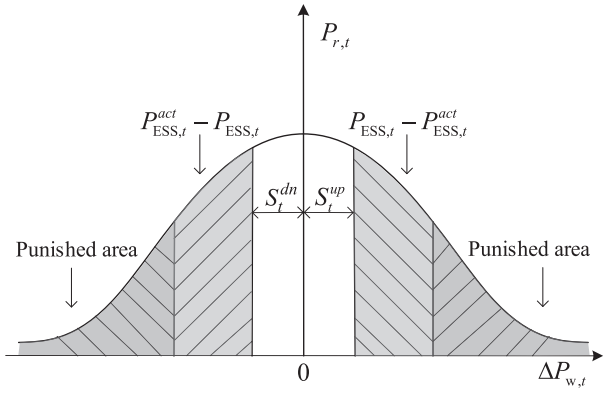


Fig. 2. Cooperation rule between ESS and reserve.

2.2. Framework of the DRL model

The DRL model consists of two processes: a training process and an application process. In the offline training process, the controller is trained so that the action selected can bring higher revenue to the WPP. In the application process, the well-trained controller is used to make online control decisions.

The training process consists of three modules: 1) the virtual decision module, 2) the electricity market simulation module, and 3) the learning module, as illustrated in Fig. 3.

In the virtual decision module, the WPP will make a virtual decision on the ESS charge/discharge and reserve purchase. The input state of this module includes the forecasted wind power $P_{w,t}$, the forecasted electricity price λ_t, the up and down reserve price μ_t^{up} and μ_t^{dn}, as well as the electricity quantity E_t stored in the ESS. For each input state, the controller should first evaluate the Q values for all feasible control actions, where the Q value is a measurement that quantifies the effect of an action on the immediate and delayed revenues of the WPP. Then according to the stochastic action selection policy, a group of control actions, i.e., the charge/discharge power and the purchased reserve, is selected. These actions will be submitted to the market simulation module as inputs.

The electricity market simulation module is responsible for estimating the revenue of the WPP under the selected control actions, which mimics the external environment in the DRL. By using the historically recorded actual electricity price λ_t^{act} and actual wind generation $P^{act}_{w,t}$, the revenue can be calculated. Then the input state, the selected control actions, and the revenue will be stored as a sample for the learning module.

The learning module is used to calculate the Q values based on the available samples and then updates the parameters of the evaluation network in the virtual decision module to form the new map between the input state and the Q values.

The three modules are executed sequentially until the parameters of the evaluation network reach convergence, so that a well-trained controller is obtained.

In the application process, a deterministic action selection policy is adopted to the well-trained controller. When a new input state comes, the algorithm will select the action with the highest Q value, following the greedy strategy. The decision process of the controller in the application process is the forward propagation of the well-trained neural network whose calculation time is almost negligible.

As the key techniques in the training process, the DRL algorithm and the external environment formulations will be detailedly discussed in Section 3 and Section 4, respectively. And the implementation of the training process and the application process will be presented in Section 5.

3. Rainbow algorithm

The basic idea of RL/DRL is to strengthen or encourage the agent (controller) to generate the behavior with higher reward in interaction with the external environment. In essence, the agent is a map from the state space S to the action space A, whose target is to find a sequence of actions that can maximize the expected accumulated discounted reward [14,15].

In the context of wind farm control, the input state is composed of several continuous variables, e.g., the forecasted wind power and price. However, for the conventional RL, it can hardly handle the continuous input variables directly and must discretize them to fit the algorithm [16], which may cause unnecessary information loss. For this reason, DRL is adopted in this paper to overcome the shortage of RL. In DRL, the deep neural network technique is integrated into RL to tackle intractable continuous input variables. The DRL algorithms used to design the learning module are introduced in the following subsections.

3.1. Deep Q network (2015)

The deep Q network (DQN) developed by the DeepMind team in 2015 is one of the most classic DRL algorithms [17]. DQN takes advantage of the deep neural network (evaluation network) to build the map between the continuous state space and the Q values. In the optimization process, DQN updates the parameters in the evaluation network iteratively to obtain the optimal mapping rules. The update of the DQN follows

$$Q(s_t, a_t; \theta_t) \leftarrow Q(s_t, a_t; \theta_t) + \alpha \begin{bmatrix} r_t + \\ \gamma \max\limits_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) \\ - Q(s_t, a_t; \theta_t) \end{bmatrix} \tag{4}$$

where $Q(s_t, a_t; \theta_t)$ is the Q value of the action a_t with respect to the state s_t, θ_t means the parameters of the evaluation network. In addition to the evaluation network, a target network is set, whose parameters are represented as θ^-, to estimate the new Q value more accurately. $0 < \alpha < 1$ is the learning rate for the evaluation network update, r_t is the immediate reward from external environment and $0 < \gamma < 1$ is the discount factor that influences the current value of the future rewards.

In (4), $r_t + \gamma \max\limits_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) - Q(s_t, a_t; \theta_t)$ is called the time difference error (TD-Error) which is the change value of the Q value during the iteration. At each iteration, the DQN calculates the TD-Error based on the samples extracted from a replay buffer and then updates the evaluation network parameters θ_t to minimize the loss function $L(\theta_t) = TD\text{ - }Error^2$. The replay buffer is a sample memory to store and extract the sample for the neural network update.

Although the DQN has been successful in many fields, it still suffers many shortcomings in terms of Q-value estimation accuracy, convergence, and convergence speed.

3.2. Rainbow

For the shortcomings of the basic DQN, various extensions of the

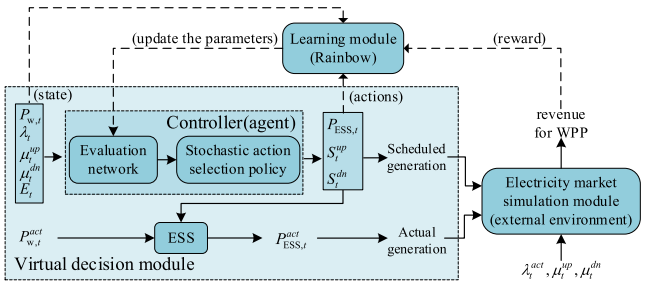


Fig. 3. Block diagram of the training process in the DRL model.

DQN were proposed in [18–21]. However, the improvement for the DQN in each extension is different from each other. In October 2017, the Rainbow proposed in [22] provides the possibility of integrating a variety of complementary extensions, which means that users can add multiple specific extensions to the basic DQN to solve their own problems more effectively.

Inspired by the excellent performance of the Rainbow on Atari games, we construct a specific Rainbow algorithm in the context of wind farm control. As shown in Fig. 4, the constructed Rainbow algorithm bases on the basic DQN and integrates three extensions, which are the prioritized replay buffer, double Q-learning and the dueling network. Moreover, the dropout layer technology is also applied to the deep neural network training. The details of each extension or technology are introduced below.

1) Prioritized Replay Buffer:

The replay buffer in the basic DQN randomly selects the samples for the update of the evaluation network. Such a sampling mechanism suffers from low training efficiency since many samples in the buffer is useless for the evaluation network update. Instead, the prioritized replay buffer proposed in [18] selects the samples based on the TD-Error values to enhance the training efficiency, shown as

$$P_{\text{sam},t} \propto |r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) - Q(s_t, a_t; \theta_t)|^{\omega} \tag{5}$$

where $P_{\text{sam},t}$ is the sampling probability, and ω determines the degree of influence of TD-Error on the sampling probability.

The sample with a larger TD-Error is extracted more frequently in the prioritized replay buffer, which can obviously accelerate the convergence and reduce the sample requirement.

2) Double Q-Learning:

The Q-value update rule of the basic DQN is called Q-learning, which may overestimate the Q value because the maximum of the Q value is always used to calculate the TD-Error. To overcome this defect, a double Q-learning algorithm is proposed in [19], which completes the estimation of the Q value by using two different neural networks. The TD-error in double Q-learning is improved as

$$r_t + \gamma Q(s_{t+1}, \underset{a_{t+1}}{argmax}\, Q(s_{t+1}, a_{t+1}; \theta_t); \theta^-) - Q(s_t, a_t; \theta_t) \tag{6}$$

Double Q-learning allows the Q value to more accurately quantify the effect of the actions on the average cumulative rewards, thus further increasing the revenue of WPPs.

3) Dueling Network:

Dueling network changes the architecture of the evaluation network in the basic DQN, to allow the Q value being represented in a more detailed form [20]. The improved-structure evaluation network has two output branches which output the current state value and the state-dependent action advantage values, respectively. Then, the final Q values can be calculated by

$$Q(s_t, a_t) = v(s_t) + \left(\mathscr{A}(s_t, a_t) - \frac{1}{|A|} \sum_{a_{t+1}} \mathscr{A}(s_t, a_{t+1}) \right) \tag{7}$$

where $|A|$ is the number of groups of the control actions in action space, $v(s_t)$ is the state value and A $\mathscr{A}(s_t, a_t)$ is the state-dependent action advantage values.

The dueling network comprehensively improves the performance of the Rainbow algorithm, including learning convergence, convergence speed and Q-value estimation accuracy. Here, an important reason for adding this extension is that it can increase WPP's revenue by enhancing the Q-value estimation accuracy, similar to the double Q-learning.

4) Dropout Layer [23]:

In the application process, the obtained real-time wind power and electricity price might be different from all historical observations, so it is necessary to enhance the generalization capability of the controller. Therefore, the dropout layer technology is applied to forming the evaluation network to avoid overfitting the training data. The basic idea of the technology is to randomly cut off certain neural links between adjacent layers during each neural network update, where the probability that a neural link is cut off is called the dropout rate.

3.3. Action selection policy

In the training process, the controller consists of the evaluation network and a stochastic action selection policy. Here we choose the ε - greedy policy to select the final action based on the Q values. The ε - greedy policy shown in (8) calculates the probability π that the feasible action in the action space is selected.

$$\pi(a, s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|A|}, & a = \text{argmax}_a Q(s, a) \\ \frac{\varepsilon}{|A|}, & a \neq \text{argmax}_a Q(s, a) \end{cases} \tag{8}$$

where $\varepsilon(\neq 0)$ is used to determine the probability of selecting the action.

Moreover, in (8), if $\varepsilon = 0$, the policy will become the greedy policy. The greedy policy guides the controller to always choose the action with the highest Q value. The greedy policy is used as the deterministic action selection policy in the application process.

4. External environment formulations

In the DRL model, the external environment simulation module simulates the settlement scheme of the electricity market and returns WPP's revenue for each time interval to the learning module as the immediate reward.

The model-free Rainbow algorithm based learning module only focuses on the reward values returned by the external environment, so that the construction of the external environment is not limited by the algorithm. Therefore, the purpose of this section is not to model a specific external environment in detail, but rather to provide a case framework for the simulation.

4.1. Immediate reward and Long-term revenue object

The revenue generated at each time interval returned as the immediate reward is calculated by (9).

$$r_t = \lambda_t^{act} P_t^{act} \Delta t - \mu_t^{up} S_t^{up} \Delta t - \mu_t^{dn} S_t^{dn} \Delta t - C_t - M_t \tag{9}$$

where C_t is the penalty fee charged by the market to the WPP when the actual generation does not match the scheduled generation, M_t is the maintenance cost for the ESS.

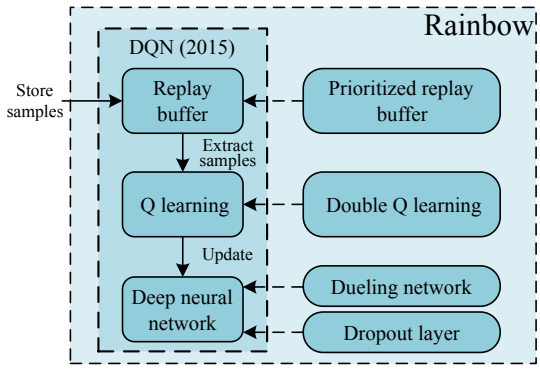


Fig. 4. Rainbow architecture adopted in this paper.

Maximizing WPP's cumulative revenue is the controller's object which can be expressed as:

$$\max \sum_{i=1}^{\infty} \sum_{t=1}^{T} r_t \tag{10}$$

where T is the number of time intervals in a control cycle of the ESS, and i is the number of control cycles experienced by the controller.

4.2. Operational constraints of the ESS

Battery pack is chosen to form the ESS in this study. The scheduled charge/discharge power of the ESS can be expressed as:

$$P_{ESS,t} = U^{dis}_{ESS,t} P^{dis}_{ESS,t} - U^{ch}_{ESS,t} P^{ch}_{ESS,t} \tag{11}$$

$$U^{dis}_{ESS,t} + U^{ch}_{ESS,t} \leqslant 1 \tag{12}$$

where $P^{ch}_{\mathrm{ESS},t}/P^{dis}_{ESS,t}$ is the charge/discharge power, and $U^{ch}_{\mathrm{ESS},t}/U^{dis}_{\mathrm{ESS},t}$ is the charge/discharge state variable of the ESS respectively which means yes when the value is 1 and no when the value is 0. Eq. (12) indicates that the charge state and the discharge state cannot coexist at the same time.

In addition, the electricity quantity stored in the ESS at the interval t is related to the electricity quantity stored at the interval t-1, shown in (13).

$$E_t = E_{t-1} + (U^{ch}_{\mathrm{ESS},t} P^{ch}_{ESS,t} \eta^{ch}_{ESS} - U^{dis}_{ESS,t} P^{dis}_{ESS,t} / \eta^{dis}_{ESS}) \Delta t \tag{13}$$

where $\eta^{ch}_{\mathrm{ESS}}/\eta^{dis}_{\mathrm{ESS}}$ is the battery charge/discharge efficiency.

The battery pack is subject to the following constraints.

1) Charge/discharge Power Constraints:

$$0 \leqslant P^{dis}_{\mathrm{ESS},t} \leqslant P^{dis}_{\mathrm{ESS,max}} U^{dis}_{\mathrm{ESS},t} \tag{14}$$

$$0 \leqslant P^{ch}_{\mathrm{ESS},t} \leqslant P^{ch}_{\mathrm{ESS,max}} U^{ch}_{\mathrm{ESS},t} \tag{15}$$

where $P^{ch}_{\mathrm{ESS,max}}/P^{dis}_{\mathrm{ESS,max}}$ is the maximum charge/discharge power of the ESS.

2) Capacity Constraint:

$$E_{\min} \leqslant E_t \leqslant E_{\max} \tag{16}$$

where $E_{\max}/E_{\min}$ is the maximum/minimum capacity of the battery pack.

4.3. Penalty fee calculation

Penalties are divided into the following two types.

1) Violation of the Charge/discharge Power Constraints:

The deviation between scheduled generation and actual generation is calculated as:

$$\Delta P^{(1)}_t = U^{dis}_{\mathrm{ESS}}(P^{dis}_{ESS,t} - P^{dis}_{ESS,\max}) + U^{ch}_{\mathrm{ESS}}(P^{ch}_{\mathrm{ESS},t} - P^{ch}_{\mathrm{ESS,max}}) \tag{17}$$

2) Violation of the Capacity Constraint:

The deviation between scheduled generation and actual generation is calculated as:

$$\Delta P^{(2)}_t = \begin{cases} P^{dis}_{\mathrm{ESS},t} - \frac{E_{t-1} - E_{\min}}{\Delta t} \eta^{dis}_{\mathrm{ESS}}, & U^{dis}_{\mathrm{ESS},t} = 1 \\ P^{ch}_{\mathrm{ESS},t} - \frac{E_{\max} - E_{t-1}}{\Delta t \eta^{ch}_{\mathrm{ESS}}}, & U^{ch}_{\mathrm{ESS},t} = 1 \end{cases} \tag{18}$$

The penalty fee passed to the WPP finally is the larger of two types:

$$C_t = \max(\Delta P^{(1)}_t, \Delta P^{(2)}_t) P_{\mathrm{pen},t} \Delta t \tag{19}$$

where $P_{\mathrm{pen},t}$ is the penalty price for the power deviation.

4.4. Maintenance cost calculation

Maintenance costs are divided into the following three types.

1) Basic Cost of Maintenance and Operation:

The charge and discharge behavior of ESS will result in inevitable maintenance costs which are related to $P^{act}_{ESS,t}$:

$$M^{(1)}_t = |P^{act}_{ESS,t}| \, \Delta t P_{\mathrm{men},t} \tag{20}$$

where $P_{\mathrm{men},t}$ is the price of the basic maintenance.

2) Violation of the Electricity Requirement for the ESS to Enter a New Cycle:

In order to avoid the loss of the adjustment ability of the ESS, the electricity quantity stored in the ESS at the end of each cycle should be close to an expected electricity value for ESS to enter a new cycle. During the training process, an extra fee is charged to enable the ESS to maintain the adjustment ability:

$$M^{(2)}_t = \begin{cases} 0, & t \neq T \\ |E_{\exp} - E_T| \, P^{(1)}_{\text{extra - men},t}, & t = T \end{cases} \tag{21}$$

where E_T is the electricity quantity stored in the ESS at the end of the cycle, $E_{\exp}$ is the expected electricity required for ESS to enter a new cycle, $P^{(1)}_{\text{extra - men},t}$ is the price of the extra fee as well as determines the degree of influence of Eq. (21) on the control strategy. This price is gradually reduced to zero in the training process and always equal to zero in the application process.

3) Violation of the Number Constraint of State Transition:

The number constraint of state transition is as follow:

$$\sum_{t=2}^{T} Y_{ESS,t} \leqslant N_{\mathrm{ESS}} \tag{22}$$

where $Y_{ESS,t}$ is the state transition variable of the ESS which means the state of charge or discharge of the ESS has changed when the value is 1 and unchanged when the value is 0, N_{ESS} is the maximum state transition number in a control cycle of the ESS.

For the violation of the number constraint of state transition, WPP should pay the extra maintenance cost based on the deviation between $\sum_{t=2}^{T} Y_{\mathrm{ESS},t}$ and N_{ESS}:

$$M^{(3)}_t = \begin{cases} 0, & t \neq T \\ (\sum_{t=1}^{T} Y_{\mathrm{ESS},t} - N_{\mathrm{ESS}}) P^{(2)}_{\text{extra - men},t}, & t = T \end{cases} \tag{23}$$

And the extra maintenance cost finally paid by WPP is

$$M_t = M^{(1)}_t + M^{(2)}_t + M^{(3)}_t \tag{24}$$

5. DRL model implementation

After introducing the learning module (Rainbow) and the electricity market simulation module (external environment) separately, this section describes in detail the implementation of the whole DRL model.

5.1. State space and action space

The state space shown in (25) consists of the forecasted wind generation, the forecasted electricity price, the reserve price, and the electricity quantity stored in the ESS.

$$S = \{s|s_t = (P_{w,t}, \lambda_t, \mu_t^{up}, \mu_t^{dn}, E_t)\} \quad (25)$$

The scheduled charge/discharge power of the ESS, the amount of up reserve and the amount of down reserve are used, from which the discretized action space is generated as follows:

$$A = \{a|a_{t,j,k,l} = (P_{ESS,t,j}, S_{t,k}^{up}, S_{t,l}^{dn}),$$

$$j \in [1, 2, ...,M], k \in [1, 2, ...,N], l \in [1, 2, ...,O]\} \quad (26)$$

where M, N and O are the numbers of equally divided segments in the ranges of $P_{ESS,t}$, S_t^{up} and S_t^{dn}, respectively.

5.2. Training process implementation

In the training process, enough samples are accumulated by operating the virtual decision module and electricity market simulation module, and then the evaluation network is updated in the Rainbow based learning module. The flowchart of the training process is shown in Fig. 5 and the details are explained as follows.

1) Build Networks and Initial the Parameters:

Two deep neural networks with same structure and equal initialization parameters are constructed as the evaluation network and the target network, respectively. The values of their initialization parameters are randomly generated. The number of neurons in the input layer of the networks is equal to the dimension of the state space, and the number of the neurons in the output layer is the same as the number of the actions ($M \times N \times O$) contained in the action space.

2) Sample Accumulation:

Before operating the learning module, a certain number of samples need to be generated and stored into the prioritized replay buffer. The operation of the virtual decision module and electricity market simulation module for each time interval can generate one sample. The specific flow of the sample generation is as follow. The evaluation network takes the current state s_t as the input to obtain the Q values of all actions in the action space. Then, an action a_t is selected through the ε - greedy policy. After this dispatch, a new state s_{t+1} is observed, and the immediate reward r_t is calculated by (9). Finally, (s_t, a_t, s_{t+1}, r_t) is stored as a sample in the prioritized replay buffer. If the buffer is full, the learning module can be done.

3) Learning module:

The prioritized replay buffer determines the sampling probability of each sample according to (5) firstly. Then, a minibatch of samples are randomly selected from the buffer to calculate the TD-Error by (6), and an Adam optimizer [24] is used to complete the parameter update of the evaluation network. For the target network, its parameters are untrainable. Whenever the parameters of the evaluation network have been updated N times, the parameters of the evaluation network are copied to the target network.

If WPP's revenue has increased to a stable, the parameters of the evaluation network are converged. If not, the sample accumulation and network update should be continued. The sample accumulation at this time is to renew the samples in the prioritized replay buffer.

5.3. Application process implementation

The well-trained controller in the application process consists of the converged evaluation network from the training process and the greedy policy. By using the well-trained controller, the optimal action for each input state is determined. The flowchart of the application process is shown in Fig. 6.

In the application process, the parameters of the evaluation network are fixed and untrainable. The ε - greedy policy which is a stochastic policy is replaced with the greedy policy, which is a deterministic policy, to always select the action with the highest Q value. The well-trained controller focuses on maximizing WPP's revenue, instead of improving it.

Dividing the method into two independent processes is to make the well-trained controller suitable for online application. The training process is an offline training process whose optimization time does not affect the decision process of the controller in the application process. The application process no longer optimizes the controller and the decision-making process can be completed in an instant.

6. Case study

A wind farm with a 50 MW installed capacity located in Jiangsu Province, China is selected as the case wind farm in this paper. NaS battery pack is integrated into the wind farm as the ESS, whose technical parameters are illustrated in Table 1. In present work, the ESS has a control cycle length of one day that includes 24 time intervals, and the maximum state transition number of the ESS is 18. The price of the basic maintenance cost is fixed as 20 ¥/MWh, the penalty prices for power deviation and extra maintenance are assumed to be 1000 ¥/MWh.

The case wind farm is equipped with a prediction system that is based on the recurrent neural network (RNN) technology to obtain the forecasts of wind generation and electricity price. Mean absolute percentage error (MAPE) is used to evaluate the forecast level. The MAPE on wind power is calculated to be 10.7% and the MAPE on electric price is 14.9%. The prices of the reserve are listed in Table 2.

In the action space, the scheduled charge/discharge power of the ESS is discretized into {−7.5, −5.0, −2.5, 0, 2.5, 5.0, 7.5}, both the amounts of up and down reserves to be purchased are discretized into {2.5, 5.0}. The discretized action space in the case contains a total of 28 actions.

We designed the Rainbow by following the literatures [17–21,22] and implement it in TensorFlow [25].

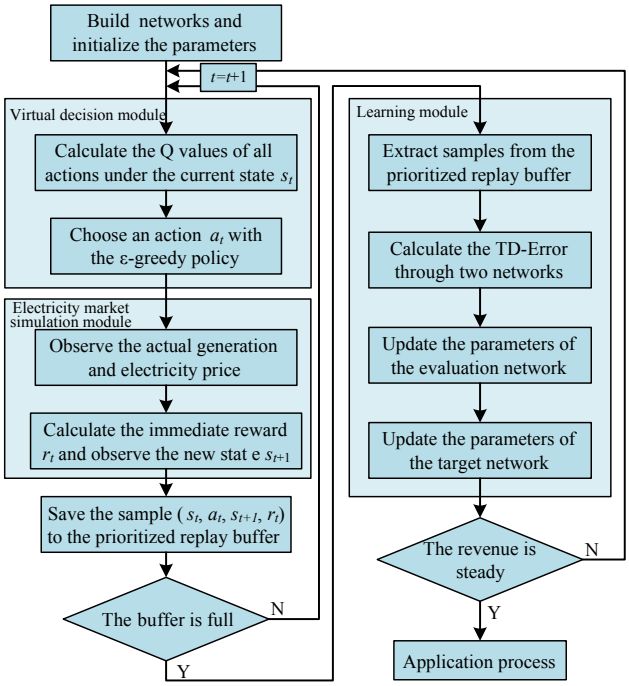


Fig. 5. Flowchart of the training process to obtain the converged evaluation network.

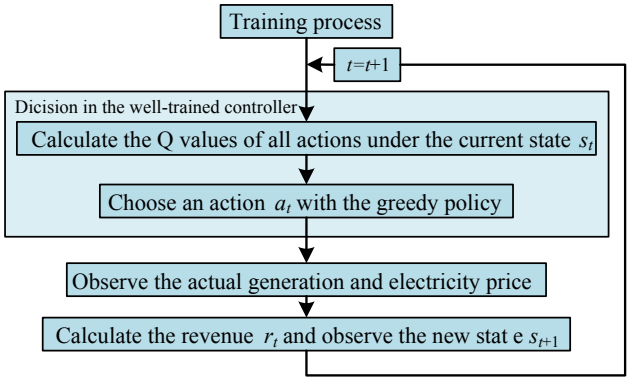


Fig. 6. Flowchart of the application process.

Table 1
Technical parameters of the NaS battery pack.

$P^{ch}_{ESS,\max}$ (MW)	$P^{dis}_{ESS,\max}$ (MW)	$E_{\max}$ (MWh)	$E_{\min}$ (MWh)	η^{ch}_{ESS}	η^{dis}_{ESS}	E_{exp} (MWh)
7.5	7.5	45	5	0.85	0.95	25

Table 2
Reserve price in each time interval.

Interval index t	1	2	3	4	5	6
Reserve price (¥)	205	195	185	185	185	190
Interval index t	7	8	9	10	11	12
Reserve price (¥)	195	200	205	210	215	220
Interval index t	13	14	15	16	17	18
Reserve price (¥)	225	230	235	240	245	250
Interval index t	19	20	21	22	23	24
Reserve price (¥)	255	255	245	235	225	245

6.1. Training process

The structures of the evaluation network and the target network in the Rainbow are both finalized as a fully connected neural network with two hidden layers and a dueling layer from the dueling network. The structure of the deep neural network and the number of neurons in each layer are shown in Fig. 7. The input of the neural network is a vector of length 5 and the output of the neural network is a vector of length 28, which represents the Q values of all actions in the action space. The activation function of each neuron in the hidden layers and the dueling layer is Rectified Linear Unit (ReLU). In the dueling layer, the outputs of the first 28 neurons are the state-dependent action advantage values $\mathscr{A}(s_t, a_t)$, and the 29th neuron outputs the state-based value $v(s_t)$. The calculation rule from the dueling layer to the output layer is expressed by (7).

In the training process, the ε in the ε - greedy policy is fixed as 0.1. In each evaluation network update, a minibatch of 32 samples are randomly selected from the prioritized replay buffer that can hold up to 3000 samples, to compute the TD-Error. Whenever the evaluation network has been trained 300 times, the parameters of the target network are updated once. The details on the hyperparameters of the Rainbow in the training process are listed in Table 3.

The curves of the revenue, extra cost and reserve cost in the training process are shown as follows to illustrate that the proposed DRL model can effectively optimize the control strategy in the controller.

1) Change of WPP's Revenue in the Training Process:

Fig. 8 shows WPP's revenues returned by the external environment in the training process. It can be seen that the strong volatilities of wind power and electricity cause the WPP's revenue to fluctuate. On the other hand, WPP's revenue tends to increase in the fluctuations with the increasing number of the samples experienced by the controller. In the sample-accumulation stage of the training process (0–3000 samples), the learning module does not work because of the insufficient samples in the prioritized replay buffer. The revenues in this stage are low and non-increased. In the subsequent revenue growth stage (3000–9600 samples), the control strategy in the controller is optimized and the revenue increases. WPP's revenue becomes stable after the controller experiencing approximately 9,600 samples, which means that the parameters of the evaluation network have converged. The controller at this time can obtain the average revenue of 4859.9¥ for the WPP in each time interval.

2) Change of WPP's Extra Cost:

Extra cost is the sum of the penalty fee and the extra maintenance cost. Fig. 9 shows the extra costs paid by the WPP for each control cycle in the training process. The extra cost also decreases in the fluctuations with the increasing number of samples experienced by the controller, and then becomes stable after the controller experiencing about 9,600 samples.

3) Change of WPP's Reserve Cost:

Fig. 10 shows the costs of the reserve for each control cycle in the training process, which can also reflect the optimization directions of the control strategy in the controller at different stages of the training process. In the sample-accumulation stage (0–3000 samples), the immature controller randomly purchases amounts of the reserve. And then huge penalty fees inspire the controller to purchase a large amount of reserve to realize the scheduled generation as much as possible (experiencing 3000–4800 samples). After reducing the penalty fee to a certain level, the controller learns how to avoid over-purchasing the reserve to reduce the reserve cost (experiencing 4800–9600 samples).

6.2. Application process

Based on a series of new wind power data and electricity price data, WPP's average revenue in the application process is evaluated. And the performance of the well-trained controller on tracking/realizing generation schedule is shown. The impacts of the forecast level and prediction time window on earning revenue are also analyzed.

1) Analysis of WPP's Revenue:

In the application process, the average revenue obtained in each time interval is calculated as 5331.6¥ (> 4859.9¥). The reason why the revenue obtained by the controller in the application process becomes significantly higher than that in the training process is that the greedy policy selects the action with the highest Q value every time, instead of

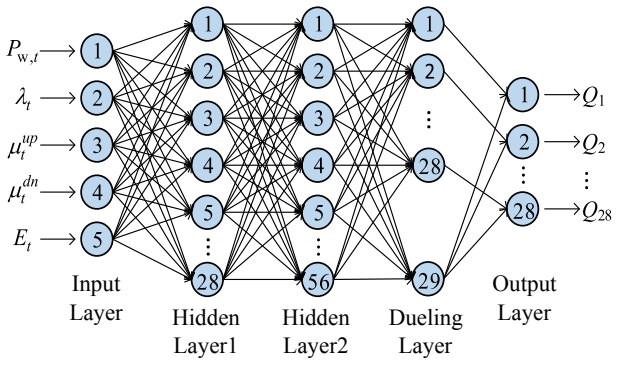


Fig. 7. Structure of the evaluation network.

Table 3
Hyperparameters of the rainbow.

Structures	Hyperparameters	Values
Evaluation network	Learning rate: α	10^{-3}
	Discount factor: γ	0.9
	Minibatch size	32
	Dropout rate	0.7
	Activation function	ReLU
	Optimizer	Adam
	Dropout rate	0.5
Prioritized replay buffer	Buffer size	3000
	ω	0.6
Target network	N	300

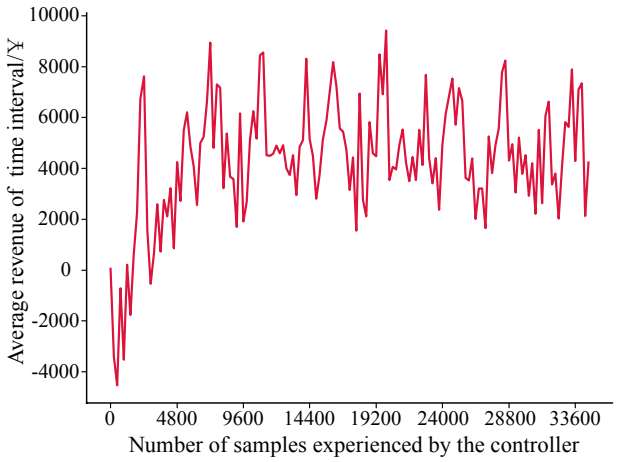


Fig. 8. Average revenues in the training process.

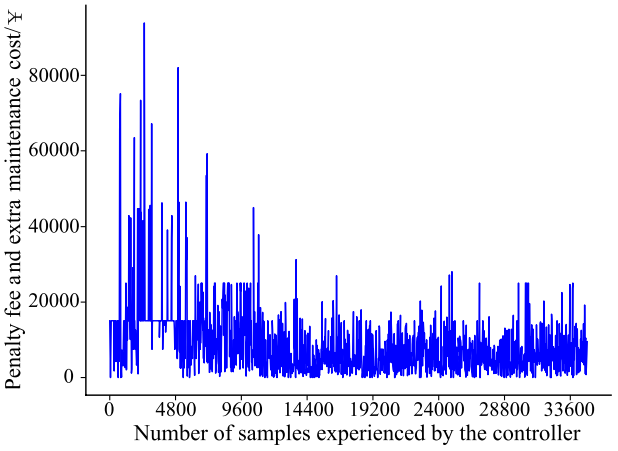


Fig. 9. Extra cost for each ESS control cycle in the training process.

randomly trying other actions sometimes like the ε - greedy policy.

2) Electricity Change Curve in the ESS:

Fig. 11 shows the change curve of the electricity stored in the ESS. In the application process, the electricity stored in the ESS is always within the allowable range ([5, 45] MWh), and no upper and lower limits are reached. It should be noted that when the electricity reaches the maximum/minimum capacity, ESS is prone to lose the ability to adjust, causing revenue losses.

3) Generation Schedule Tracking:

While gaining high revenues for the WPP, the controller can effectively track/realize the generation schedule with the uncertainty of wind power, as shown in Fig. 12. The result shows that the controller can reduce the deviation between the actual generation and the scheduled generation into the range $[-2.5, 2.5]$ ($[-0.05, 0.05]$ in the figure) which can be further eliminated by purchasing a small amount of the reserve in advance.

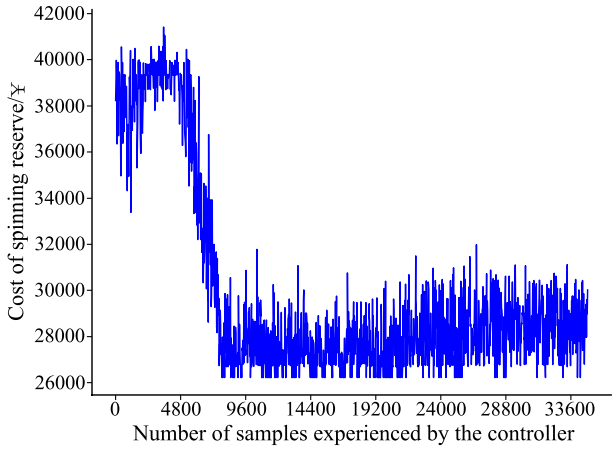


Fig. 10. Reserve cost for each ESS control cycle in the training process.

4) Impacts of Forecast Level and Corresponding Solution:

The forecast level affects WPP's revenue and the performance on tracking the generation schedule [26]. Fig. 13 shows the performance of the well-trained controller on tracking generation schedule when the MAPE on wind power is calculated to be 15.3%. The result shows that the generation schedule cannot be realized/tracked sometimes, even if the largest amount of the reserve ($[-5, 5]$) is purchased in advance. Table 4 compares the average revenues that are affected by the forecast level and the maximum purchase amount of the reserve. The results suggest that enlarging the feasible amount range of the reserve purchased in advance can alleviate the decline in WPP's revenue due to the poor prediction system, which means that a new controller with a larger action space should be trained.

5) Impacts of Prediction Time Window:

Maximizing WPP's long-term revenues is the focus of this paper. Therefore, the trends of the price and wind power over time are also important decision-making information that has big impacts on the

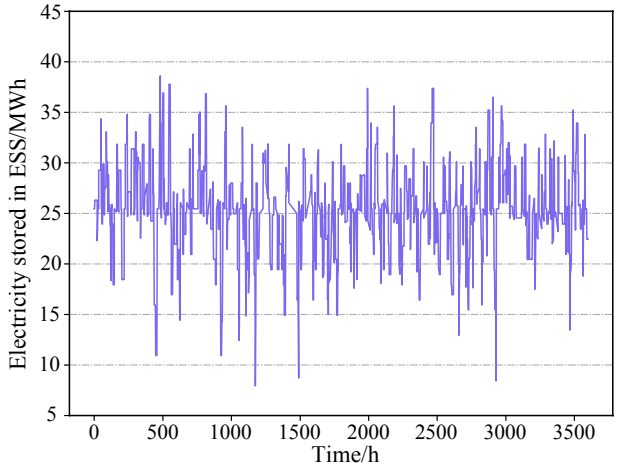
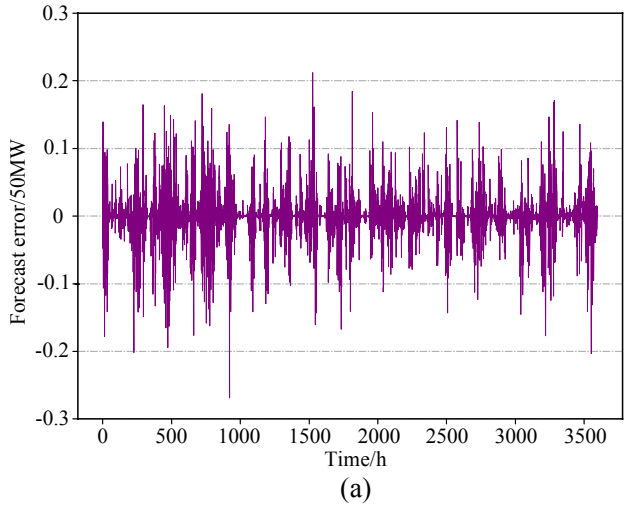


Fig. 11. Electricity stored in the ESS.



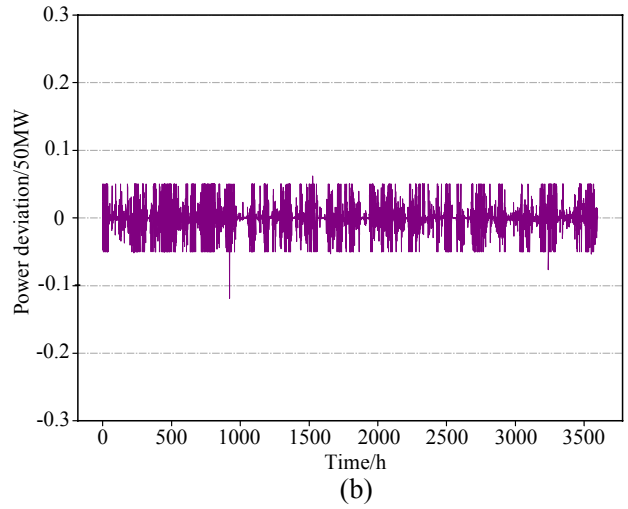
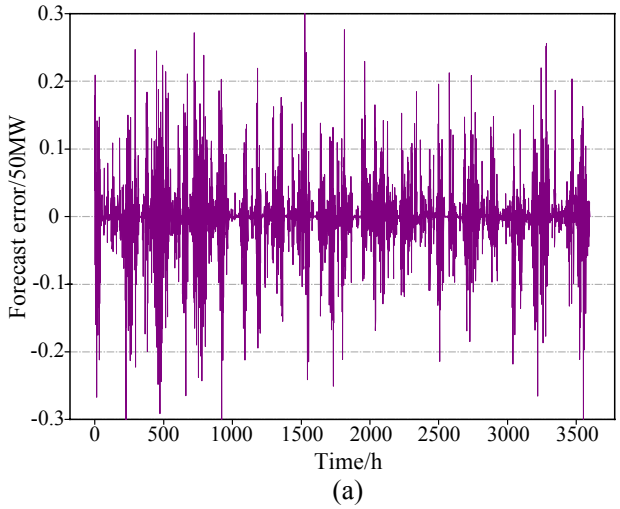


Fig. 12. Generation schedule tracking in the application process. (a) wind power forecast error, (b) power deviation between the actual generation and the scheduled generation.



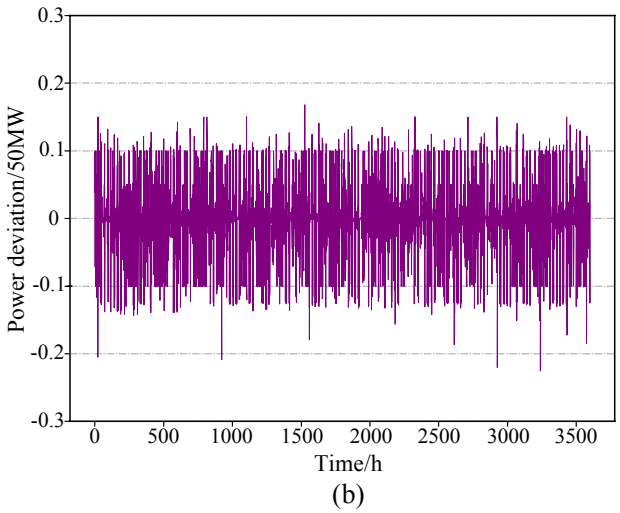


Fig. 13. Generation schedule tracking under a poor wind power prediction system in the application process. (a) wind power forecast error, (b) power deviation between the actual generation and the scheduled generation.

revenue. A case is that when the wind power time series [$P_{w,t}$, $P_{w,t+2}$, $P_{w,t+4}$] is taken as part of the input state at time t, the average revenue is enhanced to 5468.1¥.

However, it is not recommended to select the time window containing lots of forecasted values. An excess of forecasted values with uncertainties obviously reduce the convergence performance of the DRL algorithm and increase the computational cost required in the training process.

6.3. Sensitivity analysis on learning rate

The performance of the DRL algorithm is sensitive to many hyperparameters, which makes DRL related results reproduction seldom straightforward. This section just integrates the sensitivity analysis on the learning rate which is directly related to the calculation cost. And literature [27] is recommended to be referred to investigate problems with reproducibility in DRL.

Fig. 14 depicts the sensitivities of the average revenue obtained and the minimum number of the samples required for convergence to the learning rate α(discount factor is fixed as 0.9). Conclusions are as follows: 1) The Rainbow with small learning rate can bring high revenue for WPP in the application process. It is noted that the revenues become stable when the learning rate decreases to a certain level. 2) A small learning rate will also cause a high calculation cost. For a WPP, a good learning rate can maximize the revenue to the stable level with the smallest number of samples.

Table 4
Impact of forecast level on average revenue.

MAPE on wind power	Feasible reserve range (MW)	Average revenue in the application process (¥)
10.7%	[−5.0, 5.0]	5351.6
15.3%	[−5.0, 5.0]	4669.3
15.3%	**[−7.5, 7.5]**	4827.4

6.4. Comparative studies on different optimization solutions

Scenario-based stochastic programming (SSP) and robust programming (RP) are the existing most popular optimization methods for coping with uncertainty. To further demonstrate the advantages of the proposed DRL-based wind farm control method, it is compared with the two solutions above in the condition with the same data, as shown in Table 5.

In the scenario-based stochastic programming method, the wind power and electricity price forecast errors are assumed to obey the normal distribution $N(\mu, \sigma^2)$ [28,29]. And a large number of scenarios are required to cope with the multiple uncertainties in wind power and

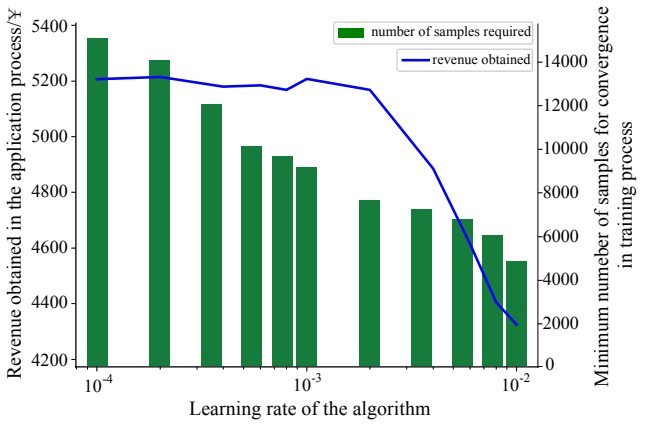


Fig. 14. Sensitivity analysis on learning rate of the Rainbow.

electricity price, which challenges the memory capacity and computing speed in computer. The optimization result shows that the average revenue obtained by the SSP method is significantly lower than that obtained by the DRL method because of: 1) the difference between the assumed distributions and the actual uncertainty laws and 2) the limited number of the scenarios due to computational cost.

The robust programming method just considers the extremes of the uncertainty while ignoring the probability information, which can give a reference to avoid risks. However, robust optimization methods usually obtain conservative control strategies that make it difficult to maximize WPP's revenue in electricity market.

In addition, the mathematical optimization solutions impose requirements on the construction of the external environment. The actual objective function and constraints need to be approximated or translated into required mathematic expressions sometimes, which makes these control methods weak in solving the practical problems. As mentioned in Section 4, the model-free DRL focuses on the reward values returned by the external environment instead of its formulations. Therefore, the control method based on DRL has better application prospects.

6.5. Comparative studies on RL/DRL algorithms

1) Comparative Studies on RL Algorithm:

Q-learning is chosen as the case RL algorithm. Different from the Rainbow, the state space of the Q-learning is discrete: the forecasted wind generation is divided into 5 discrete intervals at each interval of 10 MW, the forecasted electrical price is divided into 3 discrete intervals, and the electricity quantity stored in the ESS is divided into 10 discrete intervals at each interval of 4 MWh. Considering that one control cycle contains 24 time intervals with different reserve prices, the state space in the Q-learning is finally discretized into a set consisting of 3600 discrete input states. The comparison results are listed in Table 6.

In the training process, due to the discrete state space, the Q-learning obviously has faster convergence speed, and the number of samples required for the training process is smaller. However, the discrete state space destroys the uncertainty laws, leading a lower average revenue than the Rainbow's. In the application process, there is still a large gap between their average revenues.

2) Impacts of the extensions/technology in this Rainbow:

Fig. 15 shows the impact of the prioritized replay buffer on the revenue curve in the initial stage of the training process. It can be seen that the Rainbow with prioritized replay buffer allows the controller to obtain higher revenues after controller experiencing the same number of samples, which obviously accelerates algorithm's convergence speed and reduces the required sample number. The impacts of double Q-learning, dueling network and dropout layer technology on the revenue addition are listed in Table 7. The results show that applying these extensions and technology can improve the revenue obtained by the controller.

7. Conclusion and future work

In electricity market, the multiple uncertainties of the wind power and electricity price cause a huge revenue loss to the WPPs. Utilizing the concept of DRL allows the uncertainties not to be assumed and formulated during the optimization, improving the effectiveness of the control strategy. Three main contributions of present work are summarized as follows:

1) DRL has been applied to train the wind farm controller against uncertainties. During the optimization, the influence of the uncertainties of wind power and electricity price on the WPP's revenue can be automatically considered, without any assumption.
2) Both managing ESS and purchasing reserve are regarded as the available measures to maximize WPP's cumulative revenues.
3) The proposed DRL method and the effectiveness of the Rainbow algorithm have been validated by simulation results for an ESS-integrated wind farm located in Jiangsu Province, China.

A number of open directions are suggested by the present work. Firstly, Rainbow is one of the DRL algorithms based on value function, whose action space must be discretized. We hope to apply some policy gradient algorithms, such as Deep Deterministic Policy Gradient (DDPG) [30], to output the continuous action instructions. However, whether a good or optimal control strategy can be learned by the DRL based on policy gradient is unknown and needs further investigation.

Secondly, wind farm's measurement data (including wind speed, wind direction, air pressure, etc.) are considered to be able to provide more information (including wind power uncertainty) for optimization than wind power data. In future, we plan to study on training an end-to-end controller whose input states are wind farm's measurement data instead of forecasted wind power and to see if it can bring higher revenue to WPPs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 5
Comparison of different optimization methods.

Optimization method	Need prior distribution of uncertainty	Computational cost	Average revenue (¥)
SSP	**Yes**	**High**	4824.6
RP	No	Low	**4399.1**
DRL	No	Low	5351.6

Table 6
Comparison between Q-learning and rainbow.

Algorithms	Training process		Application process	
	Average revenue in each interval (¥)	Number of the samples for convergence	Average revenue in each interval (¥)	Amount of the reserve purchased
Q-learning	3845.9	7100	4383.1	Large
Rainbow	4859.9	9600	5351.6	Appropriate

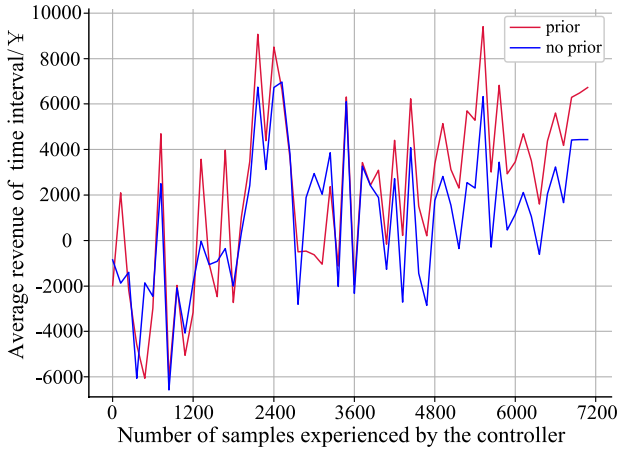


Fig. 15. The impact of the prioritized replay buffer on the revenue in the initial stage of the training process.

Table 7
Impacts of double Q-learning and dueling network and dropout layer on revenue addition.

Algorithm	Average revenue in the application process (¥)
Rainbow	5351.6
Rainbow without double Q-learning	5121.7
Rainbow without dueling network	5291.9
Rainbow without dropout layer	5110.0

Acknowledgment

This work was supported by the State Grid Corporation of China (52060018000X).

References

[1] Zhang X, Yuan Y, Hua L, Cao Y, Qian K. On generation schedule tracking of wind farms with battery energy storage systems. IEEE Trans Sustainable Energy 2017;8(1):341–53.
[2] Du E, et al. Managing wind power uncertainty through strategic reserve purchasing. IEEE Trans Power Syst 2017;PP(99). 1–1.
[3] Luo F, Meng K, Dong ZY, Zheng Y, Chen Y, Wong KP. Coordinated operational planning for wind farm with battery energy storage system. IEEE Trans Sustainable Energy 2015;6(1):253–62.
[4] Papavasiliou A, Oren SS, O'Neill RP. Reserve requirements for wind power integration: A scenario-based stochastic programming framework. IEEE Trans Power Syst 2011;26(4):2197–206.
[5] Ding H, Hu Z, Song Y. Stochastic optimization of the daily operation of wind farm and pumped-hydro-storage plant. Renewable Energy 2012;48(6):571–8.
[6] Korpaas M, Holen AT, Hildrum R. Operation and sizing of energy storage for wind power plants in a market system. Int J Electr Power Energy Syst 2003;25(8):599–606.
[7] Duque ÁJ, Castronuovo ED, Sánchez I, Usaola J. Optimal operation of a pumped-storage hydro plant that compensates the imbalances of a wind power producer. Electr Power Syst Res 2011;81(9):1767–77.
[8] Thatte AA, Xie L, Viassolo DE, Singh S. Risk measure based robust bidding strategy for arbitrage using a wind farm and energy storage. IEEE Trans Smart Grid 2013;4(4):2191–9.
[9] Xiao L, Xiao X, Dai C, Pengy M, Wang L, Poor HV. Reinforcement learning-based energy trading for microgrids. 2018.
[10] Wang B, Zhou M, Xin B, Zhao X, Watada J. Analysis of operation cost and wind curtailment using multi-objective unit commitment with battery energy storage. Energy 2019;178:101–14.
[11] Shang XY, Li ZG, Zheng JH, Wu QH. Equivalent modeling of active distribution network considering the spatial uncertainty of renewable energy resources. Int J Electr Power Energy Syst Nov 2019;112:83–91.
[12] Sun J, et al. An integrated critic-actor neural network for reinforcement learning with application of DERs control in grid frequency regulation. Int J Electr Power Energy Syst Oct 2019;111:286–99.
[13] Xiong R, Cao JY, Yu QQ. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. Appl Energy Feb 2018;211:538–48.
[14] Watkins CJ, Dayan P. Q-learning. Machine Learn 1992;8(3–4):279–92.
[15] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Machine Learn 1992;8(3–4):225–7.
[16] Wu YK, Tan HC, Peng JK, Zhang HL, He HW. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus. Appl Energy 2019;247:454–66.
[17] Mnih V, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529.
[18] Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. Comput Sci 2015.
[19] Hasselt HV, Guez A, Silver D. Deep reinforcement learning with double Q-learning. Comput Sci 2015.
[20] Wang Z, Schaul T, Hessel M, Van Hasselt H, Lanctot M, De Freitas N. Dueling network architectures for deep reinforcement learning. 2015. p. 1995–2003.
[21] Fortunato M, et al. Noisy networks for exploration. 2017.
[22] Hessel M, et al. Rainbow: Combining improvements in deep reinforcement learning. 2017.
[23] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580, 2012.
[24] Kingma DP, Ba J. Adam: A method for stochastic optimization. Comput Sci 2014.
[25] Abadi M., et al. TensorFlow: a system for large-scale machine learning. 2016.
[26] Liu M, Quilumba FL, Lee WJ. Dispatch scheduling for a wind farm with hybrid energy storage based on wind and LMP forecasting. IEEE Trans Ind Appl 2015;51(3):1970–7.
[27] Henderson P, Islam R, Bachman P, Pineau J, Meger D. Deep reinforcement learning that matters. 2017.
[28] Yao F, Dong ZY, Meng K, Xu Z, Iu HC, Wong KP. Quantum-Inspired particle swarm optimization for power system operations CONSIDERING wind power uncertainty and carbon tax in Australia. IEEE Trans Ind Inf 2012;8(4):880–8.
[29] Ma Z, Hao C, Chai Y. Analysis of voltage stability uncertainty using stochastic response surface method related to wind farm correlation. Protect Control Modern Power Syst 2017;2(1):13.
[30] Lillicrap TP, et al. Continuous control with deep reinforcement learning. Comput Sci 2015;8(6):A187.