ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Grid stability based on asynchronous modeling of reinforced Q-networks

Zhenchao Zhang[1,*], *Xianlong Ma*[1] *and Ruobing Wu*[1]

[1] Information Center of Yunnan Power Grid Co., LTD., Kunming, Yunnan, 650000, China

SUMMARY: *This paper investigates the application of deep reinforcement learning in the composite energy management problem of microgrid, establishes the DQN algorithm model for the characteristics of microgrid, and reduces the system operation cost and improves the renewable energy consumption by arranging the time-sequential charging and discharging states of the energy storage system. On this basis, a controller based on Q learning algorithm is designed, and the Q learning algorithm is utilized to dynamically correct the sag parameter, coordinate multiple distributed power sources of the grid for frequency restoration control, and verify the stability of the grid. The results show that the multi-source coordinated frequency control method proposed in this paper can fully exploit the economic optimization potential of demand-side response and use the energy optimization allocation capability of the energy storage system. It effectively improves the load-side power consumption, enhances the stability and reliability of the system, and reduces the system power cost. It is verified that the sag control at the primary control layer has the effect of reasonably allocating the system output power and stabilizing the output voltage and frequency, and the Q-learning frequency and voltage secondary controllers based on reinforcement learning can effectively improve the frequency and voltage deviation caused by the primary sag control, and improve the quality of the grid output power.*

KEYWORDS: *DQN algorithm; deep reinforcement learning; composite energy management; grid stability; multi-source coordinated frequency control approach*

1 Introduction

With the electrification of the grid and the access of a large number of emerging loads as well as the gradual increase in the proportion of renewable energy sources, the complexity of today's grid is gradually increasing, the analysis of the security and stability of the traditional power system is becoming more and more difficult, and the safe and stable operation of the system is bound to face a huge test [1-3]. The rapid changes in the traditional grid structure and power plant start-up and the continuous access to new energy sources have resulted in the complex operation of the grid as a whole and significant changes in the current level. On the one hand, due to the mixing of AC and DC of UHV in the pattern of large power grids, the mechanism analysis of power grids is becoming more and more difficult, which makes the stability characteristics of the system as well as the operation and control of the system intricate and complicated, and the previous inverting operation, system control and some scheduling experience can no longer satisfy the demand of real-time control strategy nowadays [4-8]. On the other hand, large-scale power outages caused by factors such as extreme weather and sudden accidents will cause immeasurable economic losses to society and people, as well as significant

*W5689741232025@163.com
https://doi.org/10.65102/is2026702
Published: 30 April, 2026

impacts on public order and people's daily lives, such as the Ukraine blackout in 2015, and the U.S. blackout in 2019 and 2021 [9-11]. The main characteristics of current power grid accidents are: wide range of impact, high accident impact, and high fault complexity [12, 13]. Although the number of power outages has decreased in recent years, the extent of its impact is still huge.

Currently, the monitoring and control of grid operation stability is mainly based on the analysis of annual calculations in summer and winter offline modes, as well as rule- and model-based predictive control methods. However, this annual analysis method has some underconsidered security constraints, and its analytical calculations for each scenario may be too conservative and optimistic [14]. In addition, the effectiveness, accuracy, adaptability, and real-time performance of these methods will gradually decrease as the degree of grid coupling changes [15]. Therefore, there is an urgent need to develop an automated and intelligent grid stability study method to improve the early warning capability of grid workers. While Reinforcement Learning (RL) is a trial-and-error based machine learning method in which an intelligent body learns the optimal behavioral strategies by continuously interacting with the environment. It is suitable for scenarios where the environment is uncertain and long-term planning is required. In control systems, RL is widely used in the fields of optimal control, adaptive control and intelligent control.

In grid stability control, the intelligences and environment of RL represent the controller and the grid reality, respectively, in order to carry out control tasks in frequency, voltage, and distributed energy sources.Rocchetta et al. created an RL framework of artificial neural networks to maximize the expected benefits through predictive health information, which demonstrated a comprehensive performance better than the expert scheme in cases containing renewable energy sources for optimal grid operation and maintenance management [16]. Feng et al. formulated an RL approach with Lyapunov stability constraints to design voltage controllers via monotonic neural networks, which significantly reduces the control cost and voltage restoration time and improves the reliability of real-time voltage control while guaranteeing system stability [17]. Wang et al. proposed a robust and secure RL-based distribution network resilience enhancement framework, which effectively improves system resilience and operational stability under extreme weather by aggregating distributed energy flexibility and integrating frequency security constraints [18]. Similarly, Su et al. proposed a secure RL-based emergency control method for transient stability of isolated microgrids, which uses a deep sigma point process to model the stability probability constraints and optimizes an efficient learning control strategy through a reward constraint strategy, effectively improving the dynamic stability and training efficiency of the system [19]. Wang et al. combined a deep deterministic strategy gradient with a deep RL optimization of RMSprop framework to achieve smart grid adaptive stability control through a two-layer control structure, which outperforms traditional proportional-integral-derivative controllers and other RL methods in terms of interference immunity accuracy and convergence speed [20]. Rustamovich Esanov and Gyoon Lim proposed an RL strategy based on a soft-actor-critic algorithm combined with an auto-regulation mechanism to optimize demand response for a smart grid in a photovoltaic- energy storage system demand response, significantly reducing energy cost and improving grid operation stability [21]. The RL-based grid stability control solves the problems of insufficient real-time and poor adaptive capability of the traditional methods, and improves the accuracy benefit of the control.

In addition, Deep Q-Network (DQN) is one of the important algorithms in the field of RL, which achieves the estimation of the optimal policy through the approximation of the Q-value function, thus achieving good performance in decision-making environments with high-dimensional state spaces. Zhang et al. formulated a DQN approach for load shedding, which utilizes a convolutional long- and short-term memory network model to capture dynamic

features and incorporates a customized reward function to achieve fast and accurate stability control and voltage restoration under complex fault scenarios in the Southern China Power Grid [22]. Quakernack et al. used a multi-intelligence dual DQN to autonomously control a low-voltage grid containing PV and energy storage, etc., which effectively suppressed voltage fluctuations and reverse power, and improved grid stability and equipment utilization [23]. Zhao et al. applied convolutional neural network to improve DQN and set up an online dynamic multi-microgrid construction scheme to deal with flexible switching actions, which effectively enhances the topology reconfiguration capability and operational stability of the power system under real-time perturbations [24]. Behara and Saha designed a hybrid adaptive DQN controller to improve the transient response efficiency and grid stability of a wind energy conversion system by optimizing the rotor current of a doubly-fed induction generator, which significantly reduces the amount of overshooting, harmonic distortion and stabilization time in response to wind speed fluctuations [25]. Liu et al. reported a dual DQN-based frequency regulation optimization strategy for wind storage systems, combining virtual inertia and sag control, and optimizing the fuzzy logic controller parameters using dual DQNs, which effectively improves the frequency stability and dynamic response speed of low-inertia power grids [26]. Li et al. used DQN for dynamic power system optimization decision-making, which has stronger real-time adaptability and robustness than traditional planning methods, and significantly improves the stability of power grid operation and anti-interference ability [27]. Xiao et al. used a kriging agent to enhance GRU-TCN to construct an equivalent environment model and improved DQN by incorporating a k-cross-sampling strategy so as to optimize the decision-making of multi-energy microgrids, which effectively improved the convergence, stability and energy management efficiency of the system operation [28]. Dong et al. implemented dynamic resilience decision making with DQN in a 118-node distribution network by integrating renewable energy and energy storage systems to significantly reduce outage events and improve system stability and economics [29]. Zhang and Liu fused DQN and particle swarm optimization algorithms to establish an adaptive scheduling architecture, which significantly improves the clean energy utilization rate and power supply reliability of off-grid wind-photovoltaic-diesel-storage microgrids by dynamically adjusting the parameters with multi-strategy switching [30]. It can be seen that the DQN-based grid stability method can realize the real-time parallel analysis of grid stability, which can effectively adapt to the rapidly changing and very large-scale grid, and can also support the effective operation of the system and the high-speed operation of the scheduling capability while significantly improving the level of grid management and emergency response.

This paper first introduces the basic principles of reinforcement learning and deep Q-network algorithms to provide a theoretical foundation for the next deep reinforcement learning algorithms to be used in microgrid optimization solving. Then for the composite energy management problem of microgrid, considering the impacts of different seasons, weather, and moments on the energy system, the real-time optimization control of the charging and discharging states of batteries and hydrogen-based energy storage is carried out in order to improve the consumption of renewable energy sources and the efficiency of the system. After that, in order to improve the grid frequency anti-interference, a reinforcement learning-based grid multi-source coordinated frequency control method is proposed, which dynamically adjusts the droop control parameters of multiple distributed power sources to change their output power, and realizes the coordinated active frequency control of multiple sources in the grid. Finally, the effectiveness and adaptability of this paper's method are simulated and analyzed in terms of grid energy management and voltage control.

2 Grid stability control method based on reinforced Q learning algorithm

2.1 Deep Q-network algorithm based on reinforcement learning

2.1.1 Enhanced learning

(1) Markov Decision Process

The reinforcement learning process can be described in terms of a Markov decision process. In a stochastic process, a future state is said to be Markovian if it is related only to the current state and not to any of the previous states.

A Markov decision process can be represented by a quaternion (S, A, P, R), where:

S is the state space, $S = \{s_1, s_2, ..., s_n\}$, s_i denotes the state at i time step.

A is the action space, $A = \{a_1, a_2, ..., a_n\}$, a_i denotes the action at i time step.

P is the state transfer probability, $p(s'|s,a)$ denotes the probability of state s transferring to state s' after choosing a action in the current state s.

R is the reward function, and $r(s,a)$ denotes the reward obtained by transferring state s to state s' after choosing a action in the current state s.

The intelligent performs the action a_t in state s_t, transfers to the next state s_{t+1} with probability P and obtains an immediate reward r_t. The payoff of the whole Markov process is:

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i) \tag{1}$$

where R_t denotes the cumulative return from the state at the current moment t to the termination state; γ is the attenuation coefficient, the further away from the moment t the smaller the weighting of the immediate return; T is the termination time.

(2) Value function

The goal of reinforcement learning is to find an optimal strategy to maximize the long-term cumulative return, from equation (1), it can be seen that the cumulative return R_t obtained by the state s_t is not only related to the current state s_t, but also related to the return obtained by all future states, however, the action of the intelligent body is different each time, and the return obtained by the future state is uncertain, so the state s_t the cumulative reward R_t obtained is also uncertain. In order to evaluate the reward that an intelligent body can obtain in a given state, reinforcement learning defines a value function to evaluate the reward value of state s_t.

Under the policy π, the value function $V_\pi(s)$ is used to denote the expected return obtained by executing the policy π from state s:

$$V_\pi(s) = E_\pi\left[R_t \middle| s_t = s\right] \tag{2}$$

where $E_\pi[\cdot]$ denotes the expectation of the random variable. The function $V_\pi(s)$ that evaluates only the state s is referred to as the state-value function of the policy π.

The value function $Q^{\pi}(s,a)$ denotes the expected payoff obtained by executing the action a from the state s and employing the strategy π, and this function $Q^{\pi}(s,a)$ that evaluates both the state s and the action a is referred to as the state-action-value function of the strategy π:

$$Q^{\pi}(s,a) = E_{\pi}\left[R_t \middle| s_t = s, a_t = a\right] \tag{3}$$

The optimal value function can be obtained by iterating over the following Bellman equation with the recursion property:

$$Q^{*}(s_t,a_t) = E_{\pi}\left[r_t + \max_{a_{t+1}} Q^{*}(s_{t+1},a_{t+1})\right] \tag{4}$$

The policy corresponding to the optimal value function is the optimal policy π^{*}:

$$\pi^{*} \in \arg\max_{a\in A} Q^{*}(s,a) \tag{5}$$

(3) Dynamic Programming

Dynamic programming decomposes a multi-stage problem into multiple single-stage problems and obtains the solution to the original problem by solving each single-stage problem. The dynamic programming process is shown in Figure 1. Assuming that the problem can be divided into T time periods, the immediate gain of time period t is $J_t(S_t,x_t)$, $S_t = f(S_{t-1},x_{t-1})$ is the equation of state transfer from the state S_{t-1} in time period $t-1$ to the state S_t in time period t after taking the decision x_{t-1}, $V_t(S_t,x_t)$ is the value function corresponding to the state S_t at time period t, i.e., the optimal value of the cumulative gain obtained from the beginning of time period t to time period T.

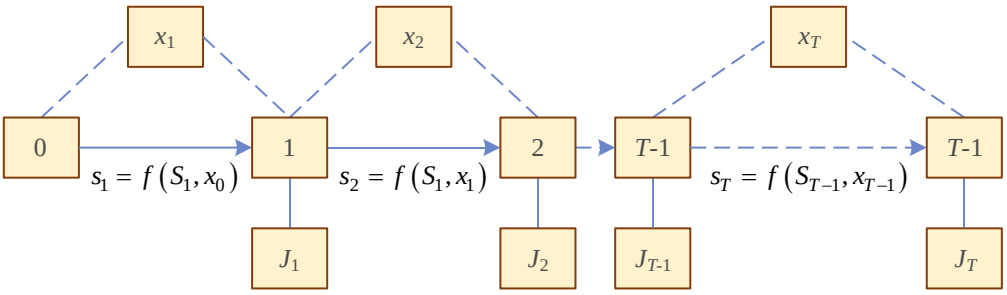

Figure 1: Dynamic programming process

1) Starting from the last time period T, make its value function $V_T(S_T,x_T)=0$;

2) Calculate the state value function $V_T(S_{T-1},x_{T-1})$ for time period $T-1$ with the following formula:

$$\begin{aligned} &V_{T-1}(S_{T-1},x_{T-1}) = \\ &\max\left\{J_{T-1}(S_{T-1},x_{T-1}) + \gamma V(S_T,x_T)\middle| S_T = f(S_{T-1},x_{T-1})\right\} \end{aligned} \tag{6}$$

3) Similarly, compute the state-value function $V_{T-t}(S_{T-t}, x_{T-t})$ for each time period from back to front:

$$V_t(S_t, x_t) = \max\left\{ J_t(S_t, x_t) + \gamma V(S_{t+1}, x_{t+1}) \middle| S_{t+1} = f(S_t, x_t) \right\} \tag{7}$$

After calculating the state value function for each time period, the initial state S_0 is substituted into the value function to introduce the optimal decision for each time period from front to back.

2.1.2 Deep Q-network algorithm

The deep Q network algorithm combines deep learning with the Q-Learning algorithm, using a deep neural network to fit a state-action value function, using the state as an input to the neural network, and outputting the Q value corresponding to each action, to select the action with a $\varepsilon - greedy$ strategy:

$$a = \begin{cases} \arg\max_{a \in A} Q(s, a; \omega), & rand(\ ) > \varepsilon \\ \text{Randomly select an action}, & rand(\ ) < \varepsilon \end{cases} \tag{8}$$

where $\varepsilon \in (0,1)$ and ω are the weights of the neural network.

In order to reduce the correlation between samples, the DQN algorithm introduces an experience playback mechanism, where $\{s_t, a_t, r_t, s_{t+1}\}$ is stored in the experience pool at each training step, and a certain number of samples are randomly selected to update the weights of the neural network during the training, which improves the stability of training.

In addition, in order to prevent the training of the same network, the Q value is constantly changing and make the training unstable, the DQN algorithm introduces two neural networks: the target network and the estimation network, at first the target network and the estimation network have the same structure and parameters, the estimation network with the training of real-time update of the weight parameters, and the target network freezes the parameters, and the estimation network will copy its own parameters to the target network every certain number of steps. The loss function of the DQN algorithm is:

$$J(\omega) = E\left[(y_t - Q(s_t, a_t, \omega))^2 \right] \tag{9}$$

$$y_t = r_t + \gamma \arg\max_a Q'(s_{t+1}, a, \omega') \tag{10}$$

where ω is the weight of the estimated network and ω' is the weight of the target network.

2.2 DQN-based asynchronous energy management method for power grids

2.2.1 Composite energy management model

The AC side considers the photovoltaic power generation system and loads, and the DC side includes batteries and a hydrogen-based energy storage system consisting of an electrolyzer, a hydrogen storage device, and a fuel cell, and the AC side and the DC side are connected by an AC/DC converter, whose main part compositions include:

(1) photovoltaic power generation system. In the grid system, the final energy source is photovoltaic power generation, which is characterized by the law:

$$P_{pv,t} = P_{\mathrm{st},t} \times \eta_{pv} \tag{11}$$

where $P_{\mathrm{pv},t}$ is the maximum power output from PV, $P_{\mathrm{st},t}$ is the real-time solar radiation power received by the PV panels, and η_{pv} is the conversion efficiency of PV.

(2) Battery. In this system, the battery is mainly used as a short-term energy storage device with high charging and discharging efficiency and high power, mainly used to maintain real-time supply and demand balance. Its characteristic law:

$$S_{bat,t+1} = S_{bat,t} + \int P_{bat,t}^{cha} dt \times \eta_{bat}^{cha} - \int P_{bat,t}^{disch} dt / \eta_{bat}^{disch} \tag{12}$$

where $S_{bat,t}$ denotes the battery capacity at moment t, P_t^{cha}, P_t^{disch} are the charging and discharging power, $\eta_{bat}^{cha}, \eta_{bat}^{disch}$ are the battery charging and discharging efficiency, respectively.

(3) Hydrogen-based energy storage. In this system, hydrogen-based energy storage is mainly used as a long-term energy storage device, with lower charging and discharging efficiency and smaller peak power, but it can store energy for a long time through electrolysis, and it is mainly used to balance the energy imbalance between seasons. Namely:

$$S_{H_2,t+1} = S_{H_2,t} + \int P_{el,t} dt \times \eta_{el} - \int P_{fc,t} dt / \eta_{fc} \tag{13}$$

where $S_{H_2,t}$ denotes the capacity of the hydrogen storage device at the moment t, $P_{el,t}, P_{fc,t}$ are the power of the electrolyzer and the fuel cell, respectively, and η_{el}, η_{fc} are the efficiency of the electrolyzer and the fuel cell, respectively. .

(4) Consumer load. The part of this system that consumes electric energy, because the total amount of load in the grid is small, so it is affected by random factors, and the load curve has obvious volatility with time.

In the above grid composite energy management system model, the following necessary constraints also need to be satisfied:

1) Current balance constraints. At any moment in the grid, the total input power and total output power are balanced in real time, i.e., the following relationship is satisfied:

$$\begin{aligned} &P_{pv,t} + P_{wt,t} + P_{bat,t}^{disch} \cdot y_{bat}^{disch} + P_{fc,t} \cdot y_{fc} + P_{un,t} \\ &= P_{load,t} + P_{bat,t}^{cha} \cdot y_{bat}^{cha} + P_{el,t} \cdot y_{el} + P_{ex,t} \end{aligned} \tag{14}$$

where $y_{bat}^{disch}, y_{bat}^{cha}, y_{fc}, y_{el}$ are binary data indicating the state of the device, where 1 means running and 0 means stopped.

2) Power constraints. Due to the characteristics of the energy storage system itself, the charging and discharging power are all limited, then there is the following relationship:

$$0 \le P_{bat,t}^{cha} \le P_{bat,\max}^{cha} \tag{15}$$

$$0 \le P_{bat,t}^{disch} \le P_{bat,\max}^{disch} \tag{16}$$

$$0 \le P_{fc,t} \le P_{fc,\max} \tag{17}$$

$$0 \le P_{el,t} \le P_{el,\max} \tag{18}$$

3) Capacity constraint. Due to constraints such as construction and operating costs, all energy storage systems have capacity constraints, which are then related as follows:

$$0 \le S_{bat,t} \le S_{bat,\max} \tag{19}$$

$$0 \le S_{H_2,t} \le S_{H_2,\max} \tag{20}$$

2.2.2 Deep reinforcement learning methods

(1) Algorithm structure and framework

In this section, the DQN algorithm in deep reinforcement learning is used to solve the above grid composite energy management.The structure of the DQN algorithm to solve the grid composite energy management problem consists of the process of data flow and information interaction, which can be mainly divided into the generation, caching, and sampling of data, the design and calculation of neural network, the gradient computation and parameter updating, and the generation and selection of actions.

(2) Space and reward function

In the DQN algorithm to solve the grid composite energy management problem, the state space parameters include real-time battery power $S_{\mathrm{bat},t}$, real-time capacity of hydrogen storage equipment $S_{\mathrm{H}_2,t}$, real-time load power $P_{\mathrm{load},t}$, real-time photovoltaic (PV) generation power $P_{\mathrm{pv},t}$, and forecasts of generation power and load at the next moment $P_{\mathrm{pv},t}^{\mathrm{pre}}$, and $P_{\mathrm{load},t}^{\mathrm{pre}}$, then:

$$S:\left\{S_{bat,t},S_{H_2,t},P_{load,t},P_{pv,t},P_{pv,t}^{pre},P_{load,t}^{pre}\right\} \tag{21}$$

The action space parameters include the battery real-time action $a_{\mathrm{bat},t}$ and the hydrogen-based energy storage device real-time action $a_{\mathrm{H}_2,t}$, and each action specifically includes three states of charging, discharging, and no-operation, so there are only nine states to choose from in the action space, then:

$$A:\{a_{bat,t},a_{H_2,t}\} \tag{22}$$

Consider the objective function when designing the reward function, and since it is in the form of a cumulative function of state and time, then the reward function can be designed as:

$$r_t = -C(s,t) \tag{23}$$

where $C(s,t)$ is the state- and time-dependent immediate use cost, which is preceded by a negative sign since the goal is to reduce costs. $C(s,t)$ can be further denoted as $C(s,t)=C_{uti,pv}(s,t)+C_{uti,bat}(s,t)+C_{uti,H_2}(s,t)+C_{ex}(s,t)+C_{un}(s,t)$, where $C_{\mathrm{uti,pv}}(s,t)$, $C_{\mathrm{uti,bat}}(s,t), C_{\mathrm{uti,H_2}}(s,t)$ denotes the immediate use cost of the PV power system, the battery, and the hydrogen-based energy storage, respectively, and in particular, it should be noted that

$C_{ex}(s,t), C_{un}(s,t)$ is the cost of abandoned power and the cost of lost power, respectively, which is related to the lost power and the abandoned power due to the difference in the impacts of the abandoned power and the lost power on the system: $C_{ex}(s,t)=\kappa_{ex}P_{ex,t}$, $C_{un}(s,t)=\kappa_{un}P_{un,t}$, where κ_{ex},κ_{un} is the penalization coefficient, which is $\kappa_{ex}<\kappa_{un}$ in general.

(3) Parameters and learning process

The input of the neural network is the data tuple (s_t,a_t,r_t,s_{t+1}), where $P_{pv,t}^{pre}$ and $P_{load,t}^{pre}$ in s_t are computed by the LSTM, and the input dimensions of the LSTM layer are 24 , Tanh is used for activation function, mean square error (MSE) is used for loss function, RMSprop is used for optimization method, a fully connected layer is connected afterward, and the number of neurons in the output layer is 1, which indicates the forecast value of load or generation at the next moment, and the training process is iterated 100 times.

The structure of the Q network is shown in Fig. 2.The Q network is in the form of a convolutional neural network combined with a fully connected network. The input layer data is 2×24, which represents the previous 24h load and PV power generation data, the first layer is 8 2×1 convolution kernels, and the second layer is 16 2×2 convolution kernels, the pooling layer adopts maximum pooling with the size of 1×1, and the step size to the right and down is 1, and the data is spliced into a one-dimensional vector with the length of 352 after the convolution of the 2 layers, and is inputted to the fully-connected layer along with other state parameters; the fully-connected layer consists of 2 hidden layers, which are connected with the other state parameters for 100 times. The fully connected layer consists of 2 hidden layers with 60 and 30 neurons respectively, and the number of neurons in the output layer is 9×1, representing the Q values obtained from 9 different combinations of actions, and the activation function of the output layer is Tanh.

After obtaining the Q values, the intelligent body adopts the $\varepsilon-$greedy strategy, i.e., there is a $1-\varepsilon$ probability of choosing $\max_a Q(s_{t+1},a_{t+1})$, and at the same time, generates a randomized action with a probability of ε, and updates the state space, and continues to calculate the new Q value until the end of the experiment.

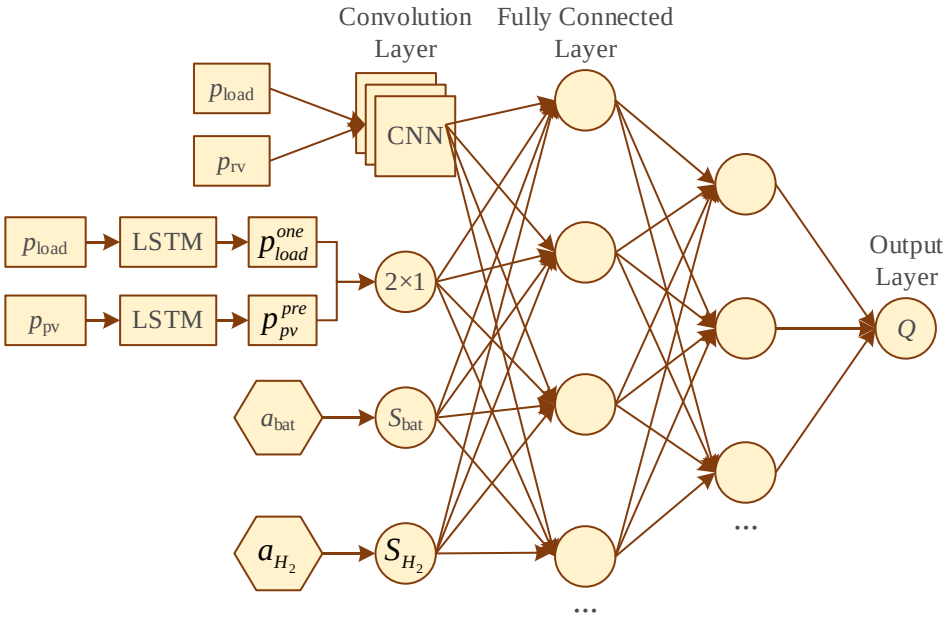

Figure 2: Q-net work structure diagram

2.3 Reinforcement Learning Based Coordinated Frequency Stability Control for Power Grid with Multiple Sources

2.3.1 Introduction to Q-learning algorithms

Q learning is one of the more maturely developed algorithms in reinforcement learning theory, which has experienced a long period of development and application since it was proposed.Q can be obtained through training and is continuously updated during the application process, in which the Q value is updated in the following way:

$$Q_{k+1}(s_i,a_i)=Q_k(s_i,a_i)+\alpha\left[r(s_i,a_i)+\gamma\max Q_k(s_{i+1},a_{i+1})-Q_k(s_i,a_i)\right] \tag{24}$$

where: s_i and a_i denote the system state and action taken at i moment, respectively; $Q_k(s_i,a_i)$ denotes the Q table at the k th iteration; $r(s_i,a_i)$ denotes the value of the reward for taking the action a_i in state s_i; $\max Q_k(s_{i+1},a_{i+1})$ denotes the maximum Q value that can be selected in s_i state after taking the action a_i, the maximum value of Q that can be selected at the next moment in the system state s_{i+1}; α and γ are fixed parameters, the smaller α means the more importance to the previous training results, and the larger γ means the more importance to the utility of the next action.

For some scenarios, the system state s_{i+1} at the next moment can be obtained from s_i and a_i, and the value of $\max Q_k(s_{i+1},a_{i+1})$ can be obtained, and the $Q_k(s_i,a_i)$ can be updated directly at the i moment. However, for the grid frequency restoration scenario discussed in this paper, it is difficult to accurately obtain the system state s_{i+1} at the next moment through s_i and a_i, considering that the number of distributed power sources connected to the grid is different at different moments and the parameters of each distributed power source may be different.

Firstly, the Q-table is updated through a large number of training to make it converge gradually. After obtaining the trained Q-table, it is then applied to the grid frequency restoration control. A greedy strategy is adopted to regulate the sag control parameters, i.e., only the current control optimization is pursued when each control action is selected, and the local optimal solution is pursued without considering the optimization on the whole. Therefore, the action that corresponds to the largest Q value in the Q table at this time is selected in each control step:

$$a^*=\arg\max Q^*(s,a) \tag{25}$$

where: a^* denotes the selected optimal action; Q^* denotes the trained Q-table. While the frequency is regulated according to the greedy strategy, the Q-table is also updated according to the result of each control.

2.3.2 Coordinated control of grid frequency based on Q-learning algorithm

(1) Frequency control

Sag control is a commonly used control method for grid-connected distributed power supply, which controls the output power by controlling the frequency and voltage amplitude.

Take distributed power supply 1 as an example, its output active power P is:

$$P \approx \frac{EU_1}{X}\delta_1 \tag{26}$$

where: δ_1 is the voltage power angle of distributed power supply 1; U_1 is the output voltage amplitude of distributed power supply 1; X is the line impedance; E is the voltage amplitude of the common point.

Therefore, it can be taken to control the frequency and voltage amplitude of the way to control the output power, the frequency - active power control method is as follows:

$$f = f_0 - k_P(P - P^*) \tag{27}$$

where: f is the desired frequency of sag control; f_0 is the rated frequency; P^* is the rated active power; k_P is the active sag coefficient.

For the grid operating frequency deviation from the desired value, you can adjust the sag parameter for frequency control, so that the grid frequency back to the desired value, the control mode is as follows:

$$f = f_0 - k_P(P - P^* - \Delta P^*) \tag{28}$$

where ΔP^* is the correction of P^*, obtained by reinforcement learning algorithm.

(2) State space and action set definition

In this paper, the state of the grid is defined based on the frequency deviation of the system. The frequency deviation of 0.01Hz is taken as the cut-off point, and the frequency deviation between 0.01~0.15Hz requires certain frequency adjustment, which is used as the standard for grid state division, and according to the value of grid frequency deviation, a total of 7 grid states are divided into $F_1 - F_7$, and the state space s is defined as $\{(-\infty,-0.2),[-0.2,-0.15),[-0.15,-0.01),[-0.01,0.01],(0.01,0.15],(0.15,0.2],(0.2,+\infty)\}$.

Then the action set A is designed. Each element in the action set A corresponds to a scheme to correct the sag parameter P^*, and the purpose of balancing the power supply and demand of the grid and frequency restoration control is achieved by correcting the sag parameter P^*.

(3) Controller design based on Q-learning algorithm

The reward function $r(s_i, a_i)$ denotes the reward value given for taking action a_i under state s_i, defined as:

$$r(s_i, a_i) = \sum r_j(s_i, a_i) \tag{29}$$

where $r_j(s_i, a_i)$ denotes the reward value of the j th controllable power supply in the form:

$$r_j(s_i, a_i) = \begin{cases} 0, & |\Delta f| \leqslant 0.01 \\ -\alpha_1 |\Delta f|, & 0.01 < |\Delta f| \leqslant 0.15 \\ -\alpha_2 |\Delta f| - \beta_1, & 0.15 < |\Delta f| \leqslant 0.20 \\ -\alpha_3 |\Delta f| - \beta_2, & |\Delta f| > 0.20 \end{cases} \tag{30}$$

where: Δf denotes the difference between the actual frequency of the system and the specified frequency of 50Hz; α_1, α_2, α_3, β_1 and β_2 are fixed control parameters designed for different frequency intervals.

Seven states are defined according to different operating frequencies of the system. In each Q-learning process, the operating frequency of the system at this time is first measured to determine the system state F_i that the system is currently in, and the sag parameter correction P_i is selected; after that, the operating frequency of the system at the next moment is measured to obtain the system state F_{i+1} at the next moment, and the reward function $r(F_i, P_i)$ is calculated; finally the Q table is updated at the $i+1$ moment the Q-table is updated. At this time, the Q table is updated in the following way:

$$Q_{k+1}(F_i, P_i) = Q_k(F_i, P_i) + \alpha\left[r(F_i, P_i) + \gamma \max Q_k(F_{i+1}, P_{i+1}) - Q_{k+1}(F_i, P_i)\right] \tag{31}$$

where: F_i denotes the system state at i moment; P_i denotes the P^* correction selected at i moment.

After obtaining the trained Q-table, each step of the grid frequency restoration process is controlled according to the greedy strategy, i.e., the optimal P^* correction amount ΔP^* is selected according to the system state F_i of the grid at each moment, which is selected as:

$$\Delta P^* = \arg\max Q^*(F_i, P_i) \tag{32}$$

3 Stability simulation analysis based on asynchronous model of reinforced Q-network

3.1 Analysis of the effects of grid energy management

3.1.1 Analysis of the optimization effect of the actual control of energy

In order to demonstrate the actual control optimization effect of the energy management scheme proposed in this paper under the influence of multiple uncertainties of grid sources and loads, and to reflect the optimization scheduling effect under complex environment, this section selects two days with large data differences from 360 sets of real-time data of the grid for the optimization simulation of the energy management, and investigates the generation power and dynamic loads of distributed power sources of the 190 scheduling cycles of these two days.

The power and load output are shown in Fig. 3; the real-time tariff data of the distribution network is shown in Fig. 4. In the following, the effect of energy management will be analyzed from two actions: energy storage system and demand side response.

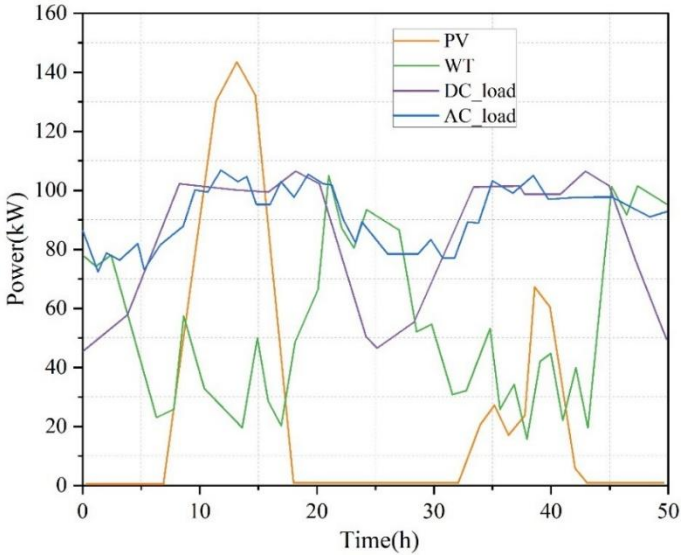

Figure 3: Power Supply and Load Output

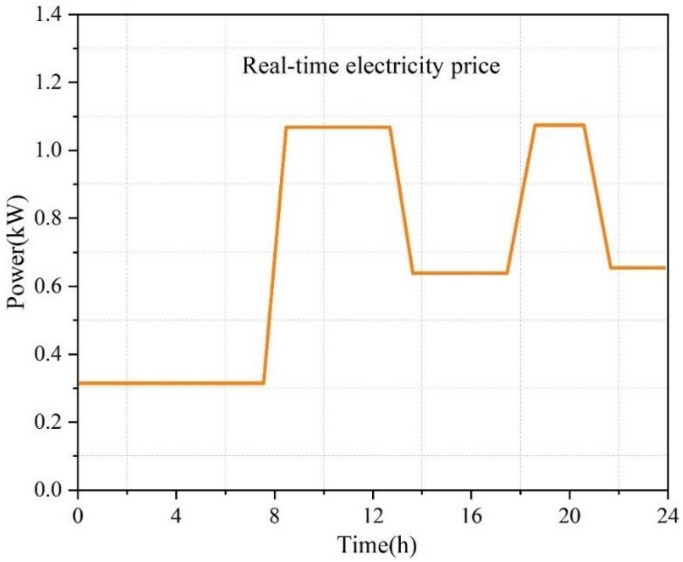

Figure 4: Real-time electricity price data of distribution network

3.1.2 Analysis of energy storage system scheduling results

The energy management optimization action is shown in Fig. 5. Combined with the real-time status data of the grid and the real-time price of electricity in the distribution network, it can be seen that when the distributed power supply is larger, the power of the grid has a surplus, and at this time, the energy storage system is in the state of charging to balance the system power, and at the same time, the demand-side response is in the state of accepting the load leveling, which is able to maximize the consumption of renewable energy sources, i.e., the storage system maximizes the economic benefits under the circumstance of meeting the power demand. The energy storage system maximizes the economic benefits while meeting the power demand.

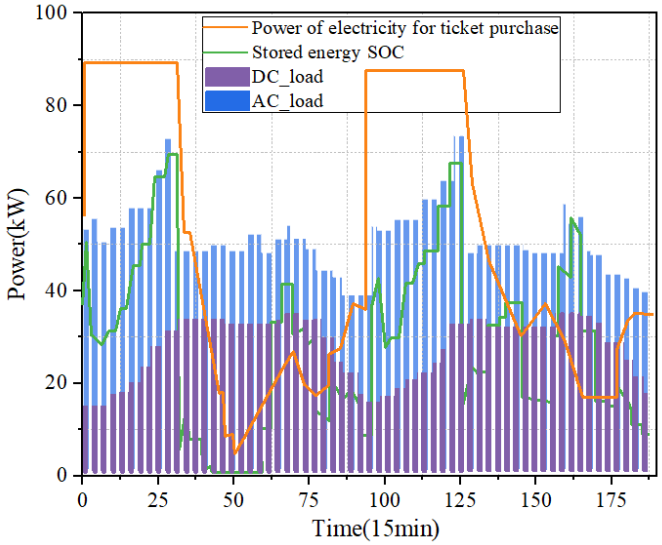

Figure 5: Energy management optimization actions

The energy storage system also responds significantly to the change of electricity price, and the result of the energy storage system action is shown in Fig. 6. When the electricity price is in the lowest zone the energy storage continues to charge and maintains a full state, and when the electricity price is high it continues to discharge, which alleviates the power shortage problem during the peak period and cuts down the power purchase cost of the grid. When the price of electricity is in the middle region, the power generation and load of the grid fluctuates greatly, and then the energy management model optimizes the power allocation through fine control of the energy storage system to maximize the economic benefits in real time.

It is worth noting that the comparison of the actions of the energy storage system before and after the demand-side response shows that the demand-side response has a limited impact on the decision-making actions of the energy storage system at low tariff moments, and in the middle and high tariff intervals, where the control logic is more complex, the demand-side response has a more pronounced impact on the energy storage. This shows that the two control parameters affect each other when the grid is under multi-objective control, and the multi-objective model solving capability of the algorithm in this paper will eventually be reflected in the cost optimization effect of the energy management model.

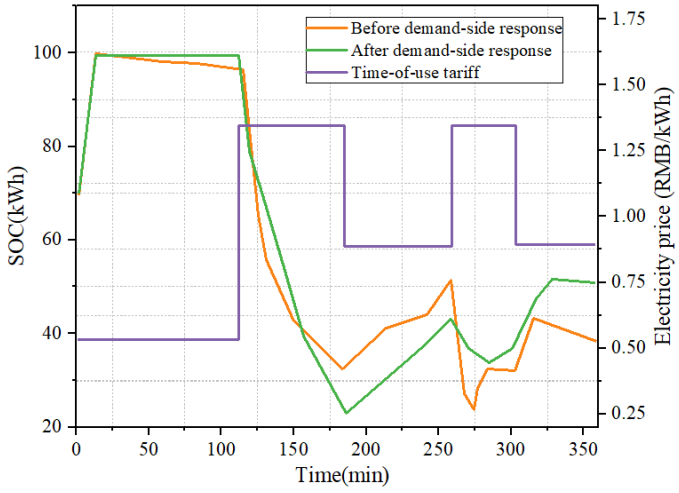

Figure 6: Action result of energy storage system

3.1.3 Demand-side response analysis

The demand side response results are shown in Figure 7. It can be seen that: when the electricity price is in the lowest range the total load of the system is also low, at this time the demand-side response action is to accept a large number of levelizable loads; and when the electricity price is in the high range and the load of the system is also high, the demand-side response will cut down the levelizable loads to reduce the power supply pressure of the grid. Since the maximum load leveling of the grid is limited to 50% of the total load in this paper, the load growth and curtailment in the demand-side response actions should not exceed 50%. In the simulation, it shows that the demand-side response is very responsive to the change of electricity price, with 48.6% of the maximum load increase and 47.5% of the maximum load curtailment, which indicates that the energy management model proposed in this paper can give full play to the demand-side response potential of the grid.

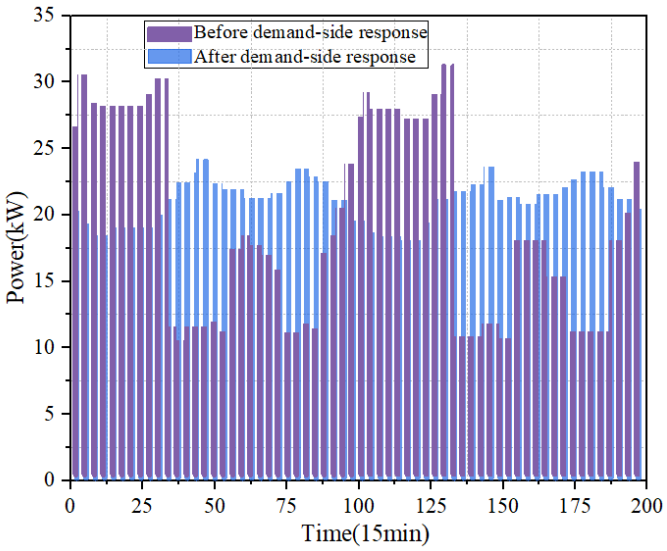

Figure 7: Demand side response result

The specific cost comparison of demand side response in the grid is shown in Table 1. It can be seen that the total load after the addition of demand side response is concentrated towards the low tariff moments, while the average change in the energy storage system action is 3.98%. This results in a 57.35% reduction in the power purchase cost of the grid compared to no demand-side response, and a smaller fluctuation in the O&M cost and the depreciation cost of the energy storage (0.76%), resulting in a 48.63% reduction in the total cost.

The above results reflect that the optimal scheduling method proposed in this paper can fully utilize the economic optimization potential of demand-side response and use the energy optimization allocation capability of the energy storage system. It effectively improves the load-side power consumption, improves the stability and reliability of the system, and reduces the system cost of power consumption.

Table 1: Comparison of the Specific Cost of Demand Side Response in Power Grid

Model	Operation and management cost (RMB)	Energy storage depreciation cost (RMB)	Power purchase and sale cost/(yuan)	Total cost (RMB)
Not have DR	93.33	55.57	779.73	921.01
Have DR	92.62	53.36	332.58	473.12

3.2 Analysis of grid voltage control results

3.2.1 Analysis of the simulation results of the secondary controller

Both the learning rate α and the discount factor γ are used to weigh the knowledge previously acquired by the intelligent body against the knowledge acquired by the current learning. In this paper, the learning rate α and the discount factor γ are set to 0.8, so that the intelligent body can update the Q value as fast as possible, in order to learn the optimal action strategy in the shortest possible time, and at the same time take into account the convergence of the Q learning algorithm. In this paper, the ε parameter of 0.8 and 0.2 are firstly selected for the pre-learning simulation of the secondary controller, to compare the effects of the two parameter values on the control performance of the controller, and then further optimize the parameters. Taking the frequency sub-controller based on Q-learning algorithm as an example, the initial state of the distributed power supply intelligent body is set to a frequency deviation of 1Hz, and the sub-controller is accessed so that the intelligent body starts training.

The pre-learning simulation results with $\varepsilon = 0.8$ are shown in Fig. 8, where the shades of the colors are related to the number of times the intelligent body learns, and the lighter the color, the more times the intelligent body learns. In the total number of 3000 episodes of the learning process, each time the intelligent body reaches the target state that is the frequency recovery to the rated value of the number of learning times experienced. During the learning process of 3000 episodes, the least number of learning times of the intelligent body in a single episode is 1 time, and the most number of learning times of the intelligent body is up to 2975 times. The number of learning times of the intelligences is mainly concentrated in the range of 0~700 times, and most of the rest of them can reach the target state within 700~1000 times of learning.

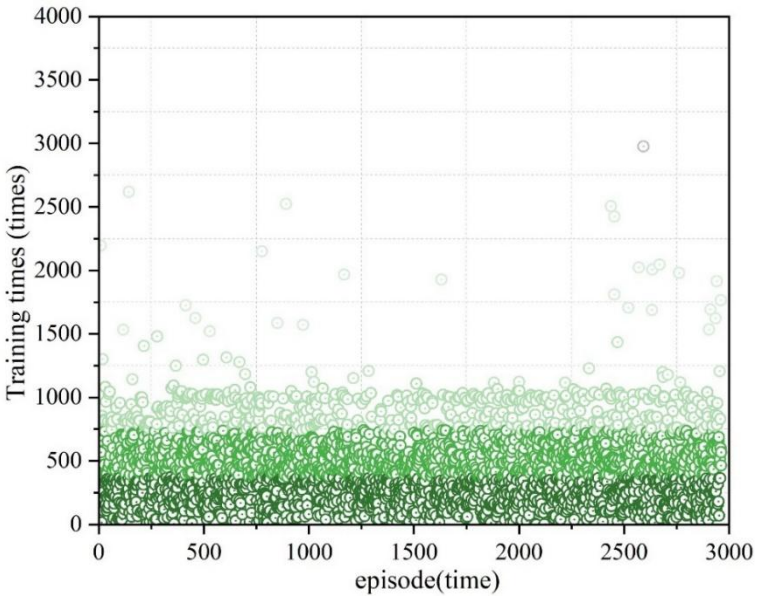

Figure 8: Pre-learning simulation results

3.2.2 Pre-learning results for secondary controllers

Then the pre-learning simulation was conducted for the secondary controller using $\varepsilon = 0.2$, with the same initial environment settings as in the previous section and a frequency deviation of 1Hz, and the results of the pre-learning simulation with $\varepsilon = 0.2$ are shown in Fig. 9. During the learning process of 3000 episodes, the minimum number of intelligences learned in a single

episode is 5 times, and the maximum number of intelligences learned is 88 times. The number of times the intelligent body learns is mainly concentrated in 8 times, and there are 1143 times more than half of the episodes learning utilizes the optimal action strategy.

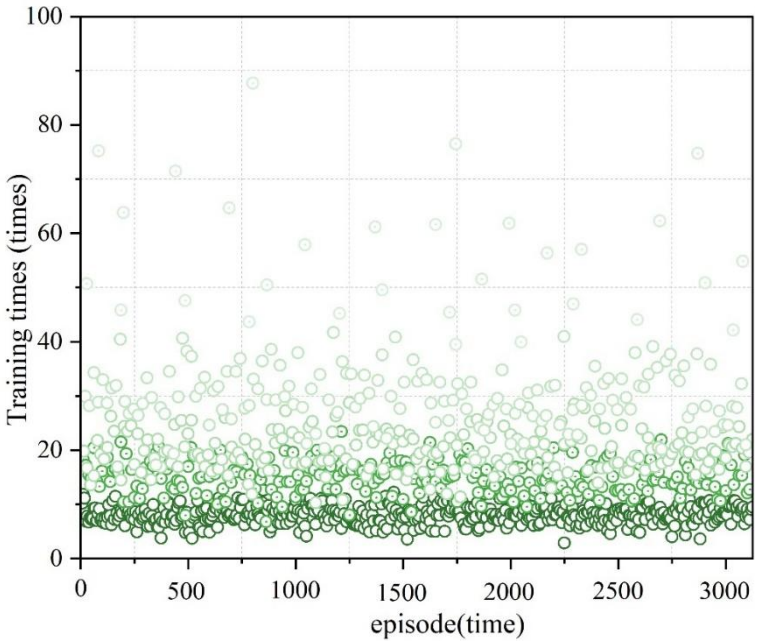

Figure 9: Pre-learning simulation results

According to the results of the above experimental statistics, when $\varepsilon = 0.8$ and $\varepsilon = 0.2$, with the increase of the number of episodes experienced by the intelligent body, the change of the average number of learning times of the intelligent body is shown in Figure 10. At the early stage of learning, because the intelligent body is in an unknown state about the environment, the intelligent body cannot directly utilize the "previous experience" to perform the optimal action with the highest value of Q, i.e., regardless of the value of the probability of exploration, ε, the intelligent body will carry out a period of time for the process of exploration. When $\varepsilon = 0.2$, after the initial learning process, once the intelligent body obtains the optimal action strategy, the intelligent body will immediately change from the state of exploring strategy to the state of utilizing strategy, and will select the optimal action strategy to execute with 80% probability. As the total number of learning times increases, the intelligence will tend to utilize only the optimized action strategy for execution and ignore the exploration process. When $\varepsilon = 0.8$, the intelligent body is in the exploration state most of the time during the learning process of a total of 3000 episodes, resulting in the intelligent body possibly obtaining the optimal action strategy in a shorter time.

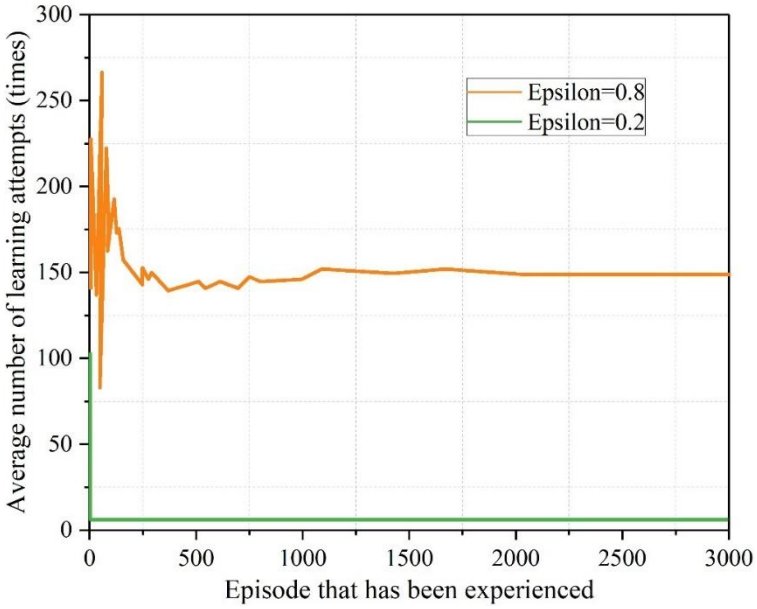

Figure 10: Results of average learning times for agents

The pre-learning results of $\varepsilon = 0.4$ for the intelligentsia are shown in Fig. 11. In the frequency and voltage secondary controller designed in this paper, there are two optimization objectives for the recovery of frequency and voltage in the grid system: first, the value of frequency and voltage is restored to the rated value; second, the rate of recovery. Combined with the analysis of the above results, it can be seen that: if $\varepsilon = 0.8$, ε value is too large, the intelligent body exploration time is too long, in the learning of the distributed power supply intelligent body of the actual grid system, too random to the distributed power supply active power, reactive power output to regulate the possibility of frequency and voltage deviation values continue to increase, although after the Although a high number of learning processes can theoretically restore the frequency and voltage to the rated value, the continuous deviation of the frequency and voltage from the rated value may cause great harm to the power grid system, and even lead to the paralysis of the entire power grid system.

If $\varepsilon = 0.2$, the value of ε is small, the utilization time of the intelligent body is too long, resulting in that with the increase of the total number of times of learning, once the intelligent body searches for the relatively optimal control strategy, it is likely to stop the exploration process, and a large amount of time is spent in recycling the optimal control strategy that has already been obtained, and ultimately fall into the situation of local optimality.

After the above analysis, combined with the comprehensive consideration of the simulation parameters of the grid system designed in this paper, this paper sets the exploration probability ε to 0.4, i.e., the probability of utilizing the optimal action strategy is relatively large, occupying 60%, while the exploration probability of randomly selecting actions occupies 40%. When $\varepsilon = 0.4$, the number of learning times of the intelligent body in 3000 episodes of learning is not concentrated in the optimal strategy obtained by pre-exploration, but is uniformly distributed in the interval of each learning times, and finally realizes the balance between utilizing the optimal strategy and exploring the strategy. The method in this paper can minimize the number of learning times for each episode of the intelligent body, reduce the multiple continuous regulation of active and reactive power output of the distributed power supply in the grid system, and recover the frequency and voltage deviation as quickly as possible under the condition of ensuring the safe and stable operation of the grid system.

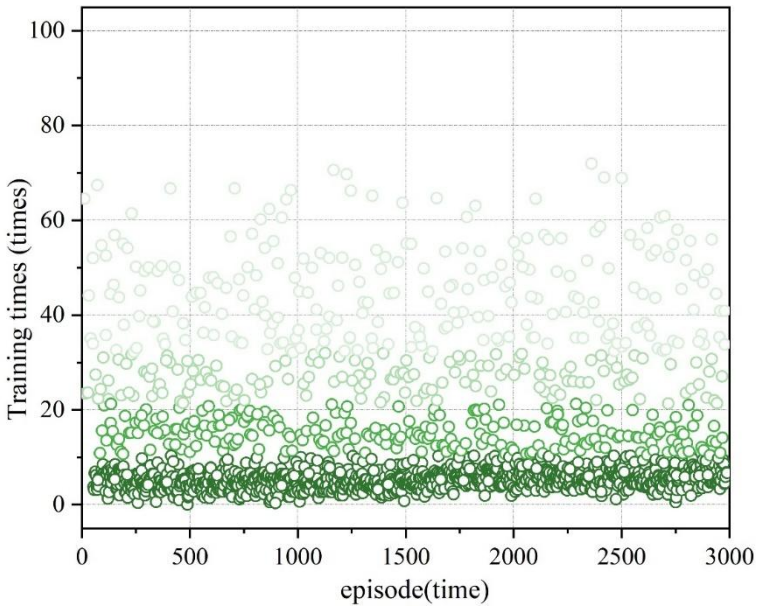

Figure 11: Agent pre-training results at $\varepsilon = 0.4$

4 Conclusion

In this paper, we propose a multi-source coordinated frequency control method for power grid based on reinforcement learning, and simulate and analyze the stability of the results of power grid energy management and voltage control from the aspects of power grid energy management and power grid voltage control. The main conclusions are as follows:

(1) The energy management optimization strategy proposed in this paper has high effectiveness and self-consistency, and it can significantly reduce the power purchase cost and total operation cost of the grid by 57.35% and 48.63% through the demand-side response and the regulation and optimization of the energy storage system.

(2) The energy management strategy of the power grid presents the control logic of purchasing a large amount of electricity at a low price and charging with the energy storage system, and purchasing a small amount of electricity at a high price and discharging with the energy storage system, which effectively reduces the operating cost of the power grid by playing the regulating role of the energy storage system.

(3) The energy management strategy of the power grid in demand-side response shows the control logic of increasing the leveling load when the price of electricity is low and the output of the power source is low, and reducing the leveling load when the price of electricity is high and the output of the power source is high. The demand-side response effectively alleviates the problem of electricity tension during the peak period and significantly improves the economic efficiency of the power grid.

(4) The frequency and voltage sub-controller based on reinforcement learning is able to quickly recover the frequency and voltage offsets generated by the sag control, effectively improving the control accuracy of the sag control as well as the output power quality of the grid system.

About the Author

Zhenchao Zhang (1991.11-), male, Han ethnicity, from Honghe, Yunnan Province. Master's

degree holder and engineer. Research focuses on: digital technology support systems for power generation operations, as well as big data applications and analysis in the power industry.

Xianlong Ma (1990.03-), male, Hui ethnicity, from Tengchong, Yunnan Province, holds a master's degree and works as an engineer. His research focuses on intelligent technologies for primary power grid equipment and their application in technical supervision and related fields.

Ruobing Wu (1990.08-), male, Han ethnicity, from Yuxi, Yunnan Province, holds a bachelor's degree and works as an engineer. His research focuses on digital technology support systems for power marketing, as well as big data applications and analysis in the power industry.

References

[1] Soares, J., Pinto, T., Lezama, F., & Morais, H. (2018). Survey on complex optimization and simulation for the new power systems paradigm. Complexity, 2018(1), 2340628.

[2] Oladeji, I., Makolo, P., Abdillah, M., Shi, J., & Zamora, R. (2021). Security impacts assessment of active distribution network on the modern grid operation—A review. Electronics, 10(16), 2040.

[3] Smith, O., Cattell, O., Farcot, E., O'Dea, R. D., & Hopcraft, K. I. (2022). The effect of renewable energy incorporation on power grid stability and resilience. Science advances, 8(9), eabj6734.

[4] Pabbuleti, B., & Somlal, J. (2020). A review on hybrid AC/DC microgrids: Optimal sizing, stability control and energy management approaches. Journal of Critical Reviews, 7, 376-381.

[5] Wen, Y., Lu, Y., Gou, J., Liu, F., Tang, Q., & Wang, R. (2022). Robust transmission expansion planning of ultrahigh-voltage AC–DC hybrid grids. IEEE Transactions on Industry Applications, 58(3), 3294-3302.

[6] Liu, T., Song, Y., Zhu, L., & Hill, D. J. (2022). Stability and control of power grids. Annual Review of Control, Robotics, and Autonomous Systems, 5(1), 689-716.

[7] Han, F., Zhang, X., Li, M., Li, F., & Zhao, W. (2023). Stability control for grid-connected inverters based on hybrid-mode of grid-following and grid-forming. IEEE Transactions on Industrial Electronics, 71(9), 10750-10760.

[8] Khosravi, N., Çelik, D., Bevrani, H., & Echalih, S. (2025). Microgrid stability: A comprehensive review of challenges, trends, and emerging solutions. Electrical Power and Energy Systems, 170(11082), 9.

[9] Haes Alhelou, H., Hamedani-Golshan, M. E., Njenda, T. C., & Siano, P. (2019). A survey on power system blackout and cascading events: Research motivations and challenges. Energies, 12(4), 682.

[10] Xia, J., Xu, F., & Huang, G. (2020). Research on power grid resilience and power supply restoration during disasters-A review. Flood Impact Mitigation and Resilience Enhancement.

[11] Hawker, G., Bell, K., Bialek, J., & MacIver, C. (2024). Management of extreme weather

impacts on electricity grids: an international review. Progress in Energy, 6(3), 032005.

[12] Yi, Z. H. A. N. G., Feng, Z. H. A. N. G., Youchun, L. I., Lv, T. A. N. G., Xuehu, P. E. N. G., & Jianguo, M. O. (2021, February). Analysis Of Typical Power Grid Blackout Accidents And Suggestions For Countermeasures. In Journal of Physics: Conference Series (Vol. 1754, No. 1, p. 012023). IOP Publishing.

[13] Sun, Y., Wu, J., Zhang, J., Xiong, Y., Liu, X., & Bai, Y. (2024). Scenario construction and vulnerability assessment of natural hazards-triggered power grid accidents. Journal of Safety Science and Resilience, 5(4), 498-511.

[14] Arzamasov, V., Böhm, K., & Jochem, P. (2018, October). Towards concise models of grid stability. In 2018 IEEE international conference on communications, control, and computing technologies for smart grids (SmartGridComm) (pp. 1-6). IEEE.

[15] Zhang, H., Mehrabankhomartash, M., Saeedifard, M., Meng, Y., Wang, X., & Wang, X. (2021). Stability analysis of a grid-tied interlinking converter system with the hybrid AC/DC admittance model and determinant-based GNC. IEEE Transactions on Power Delivery, 37(2), 798-812.

[16] Rocchetta, R., Bellani, L., Compare, M., Zio, E., & Patelli, E. (2019). A reinforcement learning framework for optimal operation and maintenance of power grids. Applied energy, 241, 291-301.

[17] Feng, J., Shi, Y., Qu, G., Low, S. H., Anandkumar, A., & Wierman, A. (2023). Stability constrained reinforcement learning for decentralized real-time voltage control. IEEE Transactions on Control of Network Systems, 11(3), 1370-1381.

[18] Wang, X., Ke, J., Wu, H., Dong, Y., Jiang, C., Huang, W., & Zhang, S. (2025). A Robust Safe Reinforcement Learning Approach for Power Grid Resilience Enhancement against Typhoons via DER Flexibility Aggregation. IEEE Transactions on Industry Applications.

[19] Su, T., Zhao, J., Yao, Y., Selim, A., & Ding, F. (2025). Safe Reinforcement Learning-Based Transient Stability Control for Islanded Microgrids With Topology Reconfiguration. IEEE Transactions on Smart Grid.

[20] Wang, B., Baziar, A., & Askari, M. (2025). A Deep Reinforcement Learning Framework for Adaptive Resiliency Enhancement in Smart Power Grids. IEEE Access.

[21] Rustamovich Esanov, A., & Gyoon Lim, C. (2026). Enhancing grid stability and renewable energy integration with reinforcement learning for optimized demand response. Energy Exploration & Exploitation, 44(1), 596-613.

[22] Zhang, J., Luo, Y., Wang, B., Lu, C., Si, J., & Song, J. (2021). Deep reinforcement learning for load shedding against short-term voltage instability in large power systems. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 4249-4260.

[23] Quakernack, L., Kelker, M., & Haubrock, J. (2022, October). Deep reinforcement learning for autonomous control of low voltage grids with focus on grid stability in future power grids. In 2022 IEEE PES innovative smart grid technologies conference Europe (ISGT-Europe) (pp. 1-5). IEEE.

[24] Zhao, J., Li, F., Mukherjee, S., & Sticht, C. (2022). Deep reinforcement learning-based model-free on-line dynamic multi-microgrid formation to enhance resilience. IEEE Transactions on Smart Grid, 13(4), 2557-2567.

[25] Behara, R. K., & Saha, A. K. (2024). Deep Q-Network Reinforcement Learning Based Rotor Side Control System of A Grid Integrated DFIG Wind Energy System Under Variable Wind Speed Conditions. IEEE Access.

[26] Liu, X., Zou, P., You, J., Wang, Y., Wu, J., Zheng, Z., ... & Hao, W. (2024). Advanced Primary Frequency Regulation Optimization in Wind Storage Systems with DC Integration Using Double Deep Q-Networks. Electronics, 13(12), 2249.

[27] Li, D., Zheng, H., & Pan, T. (2025). Resilience and flexibility optimization in solar integrated power systems via deep Q network under extreme weather. Scientific Reports, 15(1), 37374.

[28] Xiao, H., Pu, X., Pei, W., Ma, L., & Ma, T. (2023). A novel energy management method for networked multi-energy microgrids based on improved DQN. IEEE Transactions on Smart Grid, 14(6), 4912-4926.

[29] Dong, S., Wang, C., Zhang, Y., Wang, Y., Zhang, X., Guo, L., & Ju, L. (2025). Hierarchical deep Q-network-based optimization of resilient grids under multi-dimensional uncertainties from extreme weather. Scientific Reports, 15(1), 24927.

[30] Zhang, B., & Liu, B. (2025). An Adaptive Scheduling Method for Standalone Microgrids Based on Deep Q-Network and Particle Swarm Optimization. Energies, 18(8), 2133.