IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 17, 2021, accepted May 21, 2021, date of publication June 3, 2021, date of current version June 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3086002

Research on Bidding Strategy of Thermal Power Companies in Electricity Market Based on Multi-Agent Deep Deterministic Policy Gradient

DUNNAN LIU, YUAN GAO, WEIYE WANG, AND ZHIXIN DONG
School of Economics and Management, North China Electric Power University, Beijing 102206, China

Corresponding author: Yuan Gao (gaoyuanhd@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 72001078, in part by the Power System State Key Laboratory under Grant SKLD20M12, and in part by the Central China Electric Power Dispatching and Control Sub-Center of State Grid under Grant 52140019000W.

ABSTRACT With the continuous improvement of new energy penetration in the power system, the price of the spot market of power frequently fluctuates greatly, which damages the income of a large number of thermal power enterprises. In order to lock in the profit, thermal power enterprises should turn the main target of profit to the medium and long-term power market. With the continuous advancement of the reform in China's power system, major changes have taken place in the medium and long-term power transactions, including the transaction target, organization method, clearing method and so on, so it is urgent to explore the quotation strategy of thermal power enterprises under the medium and long term market changes. Based on the theory of game equilibrium, this paper establishes non-cooperative game and cooperative game models between thermal power companies. Considering that the traditional reinforcement learning method is difficult to solve the multi-agent incomplete information game model, this paper uses the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm to solve the above model. Finally, the validity of the proposed model is proved by a numerical example. The results show that, compared with other reinforcement learning algorithms, when solving the multi-agent incomplete information game model, the quotation obtained by MADDPG is more accurate, the revenue is increased by 5.2%, and the convergence time is reduced by 50%.In addition, this paper finds that in the medium and long-term power market, thermal power companies are more inclined to use physical retention methods to make profits. The greater the market power of thermal power companies, the greater the probability of physical retention. When low-cost thermal power companies retain more power, they will increase market clearing electricity prices and harm market efficiency. Regulators should focus on the market behavior of such thermal power companies.

INDEX TERMS Electricity market, bidding strategy, multi agent reinforcement learning, multi agent deep deterministic policy gradient algorithm.

I. INTRODUCTION

With the gradual advancement of China's electric power market reform, some problems have also been exposed, the typical one is the large-scale loss of thermal power enterprises [1]. During the trial operation of the power spot market, power generation enterprises have quoted zero price for startup for many times. In an environment of loose supply and demand for electricity, this phenomenon further depressed spot prices, and some thermal power companies even failed to repay loans and declared bankruptcy. It is difficult for thermal power companies to make sustainable profits in the spot market [2]. Therefore, they should turn the profit target to the medium and long-term market. With the continuous reform of China's power system, the scale of medium and long-term transactions is growing, and the contract coverage is gradually improving. Under the influence of the power spot market, the granularity of the trading time is further subdivided, and the object of transaction is gradually transferred to the time period. Medium and long-term transactions have undergone great changes in the object of transaction, the way of organization and clearing methods [3]. In order to give

The associate editor coordinating the review of this manuscript and approving it for publication was Zhouyang Ren.

full play to the role of medium and long-term transactions in stabilizing the income of thermal power companies, and considering the great changes in the medium and long-term market, the bidding strategy of thermal power companies in the medium and long-term market should be explored [4].

This paper mainly studies the bidding strategy of thermal power companies from two aspects, including the determination of the optimal declared price of thermal power companies and the evaluation of power market operation efficiency. Firstly, it considers the impact of multi-dimensional environmental parameters to find the equilibrium point of the game that can maximize the benefits of thermal power companies. Secondly, the paper evaluates the influence of the behaviors of game players on the market efficiency under the equilibrium state, and puts forward corrective suggestions on the inappropriate bidding behaviors existing in the market. At present, scholars at home and abroad divide the main research methods into the following categories in terms of bidding strategies of power suppliers: the first is the cost analysis method, which takes the production cost plus certain profit of the power producer as the quotation of the power producer [5], [6]; the second is the clearing price forecasting method, in which the generators predict the electricity market price and quote within the appropriate range of the forecasting [7]–[9]; the third is the competitor quotation analysis method, which uses probability statistics or fuzzy mathematics to estimate the competitor's quotation to establish an optimal bidding model, and finally obtains the optimal bidding strategy by solving the model [10]; the fourth type is game theory analysis method, in which the thermal power companies build game theory model based on the trading situation, and solve the game equilibrium solution as their optimal bidding strategy [11]–[13]; the fifth is the intelligent optimization algorithm analysis method, which solves the optimal bidding of power suppliers by such methods as evolutionary algorithm [14], [15], fuzzy adaptive algorithm [16], reinforcement learning algorithm [17]–[19].

In the above methods, the cost analysis method, which is the basis of the bidding strategy, does not consider the changes of market supply and demand and the decisions of other power suppliers, so it is difficult to maximize its own interests; the second and third methods require a large amount of historical data to support the calculation. But the new round of power market system reform has just started, the data is not yet sufficient. Moreover, the market structure and its rules are constantly changing, it is difficult to make accurate predictions on market prices; the game theory methods have good results for two-player game problems and perfect information game problems, but the effects are not ideal in terms of multiplayer games and incomplete information games; in the process of solving game problems with traditional intelligent optimization algorithms, it is necessary to realize the optimization process through feedback iterations of thermal power companies benefits under unknown environmental parameters. The optimization process mainly uses traditional reinforcement learning methods such as Q-learning [20], Policy Gradient and other algorithms [21]–[26], and it is implemented by setting action-reward incentives. Due to the uncertainty of the game results under bounded rationality, in the case of multi-agent interaction, Q-learning, Policy Gradient and other algorithms will have problems such as unstable environment and increased variance, which makes power generation companies unable to achieve precise and continuous decision-making, thus limiting the search for the best offer.

In order to solve the optimal quotation problem of thermal power companies under the multi-agent incomplete information game, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm based on the multi-agent reinforcement learning method was proposed [27]–[30]. The neural network parameters are updated to simulate the bounded rational process of the game to ensure that the game process is close to reality.

This paper studies the bidding strategy of thermal power companies in the medium and long-term power market under the new power reform. Under the premise of maximizing their own interests, the non-cooperative game and cooperative game models between thermal power companies are established respectively based on the game equilibrium theory. In order to make up for the limitations of game theory and traditional reinforcement learning methods to solve the multi-agent incomplete information game problem, this paper uses the MADDPG method to solve the game model, and on the basis of obtaining the optimal market quotation of thermal power companies, indirectly observes the behavior of market entities through market efficiency. Furthermore, by identifying loopholes in the electricity market mechanisms and rules, corresponding normative policy recommendations are put forward.

II. PRINCIPLES OF MULTI-AGENT REINFORCEMENT LEARNING

A. MULTI-AGENT REINFORCEMENT LEARNING THEORY

Reinforcement learning is an important part of machine learning, which is widely used in solving decision-making problems. And it is an interactive learning method based on the Markov decision process. The agent interacts with the environment to maximize the long-term return and take certain actions, then the environment returns to the agent for a certain reward, and the above process is repeated until the agent achieves a certain goal. The cyclic interaction process is shown in Figure 1. At time t, the Agent starting from a certain state $\mathrm{s_t}$, uses strategy $\pi(a\,|s)$ to choose action $\mathrm{a_t}$ to interact with the environment, and then obtains the immediate return $\mathrm{r_{t+1}}$ of the environment, and transfers to the new state st+1 according to the state transition probability $P(\mathrm{s}_{t+1}\,|s_t\ , a_t)$.

In single-agent reinforcement learning, the environment that the agent faces is fixed. However, most of the problems in real life are complex adaptive system problems. The behaviors of subjects will influence each other, and subjects can adjust their behavior rules through continuous observation

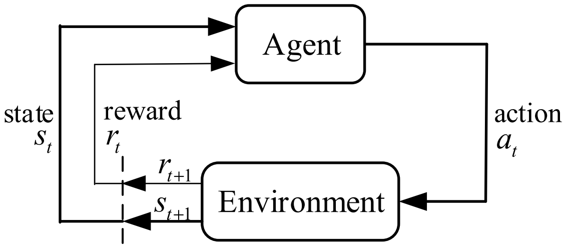

FIGURE 1. The interaction process between agent and environment in reinforcement learning.

and learning to better adapt to the environment. When studying complex multi-agent system problems, it is necessary to use multi-agent reinforcement learning [31]–[34], that is, each agent learns to improve its own strategy by interacting with the environment to obtain reward values, so as to obtain the optimal strategy of the system. However, in a multi-agent system, the learning process of the agent will become complicated. Firstly, the number of action combinations of each Agent increases exponentially with the number of agents, the dimensionality of the solution is large, and the calculation difficulty is high. Secondly, the environment is dynamic. Each Agent is learning and optimizing its own strategy at the same time, so the change of one Agent's strategy will affect the strategy of other Agents, and then affect the convergence of the algorithm. Finally, the tasks of each Agent may be different, and they influence each other, which complicates the reward design and directly affects the quality of the learning strategy.

Since the Agents in the multi-agent system may involve cooperation and competition, on the basis of single-agent reinforcement learning, multi-agent reinforcement learning introduces the concept of game and combines game theory with reinforcement learning, which is conducive to solving complex problems of multi-agent system. The basic algorithms of multi-agent reinforcement learning include gradient ascent (descent) algorithm, Q-Learning, policy hill climbing algorithm, etc., at the same time, new algorithms are also emerging in an endless stream, most of which are continuous improvements to the basic algorithms.

B. MADDPG ALGORITHM

This paper mainly introduces the MADDPG algorithm based on multi-agent reinforcement learning. The MADDPG algorithm is a natural extension of the DDPG algorithm under the multi-agent system. It belongs to centralized training and has an algorithm framework for decentralized execution. Its improvement is that, in order to solve the environmental non-stationary problem of the multi-agent system, in the modeling process of the Q-value function, the input of other agents' current strategy sampling actions are introduced as additional information. The MADDPG algorithm has two major advantages. One is that in the training phase, the actor network of each agent makes decisions based on local information (that is, the agent's own actions and states). The other is that the algorithm does not require the input of environmental change information, nor does it require the contact relationship between agents. Therefore, the algorithm is not only applicable to a cooperative environment, but also applicable to a competitive environment.

The implementation framework of the MADDPG algorithm is shown in Figure 2.

For the DDPG algorithm of a single agent in Figure 2, the implementation process is as follows:

1. Random initialization $\theta, \omega, \omega' = \omega, \theta' = \theta$. Clear the set of experience replays D,

2. Iterate from the first step of the first round,

A) Initialize as the first state S of the current state sequence and get its eigenvector $\phi(S)$.

B) The current network in Actor $A = \pi_\theta(\phi(S)) + N$ gets an action based on the state S,

C) Perform actions A, get new status S', reward R, terminate status or not,

D) Store the quintuple $\{\phi(S), A, R, \phi(S'), is\_end\}$ into the experience playback set D,

E) $S = S'$,

F) Sample m samples from the experience playback set D and calculate the current target Q value y_j:

$$y_j = \begin{cases} R_j & is\_end_j \ is\ true \\ R_j + \gamma Q(\phi(S_{j+1}), A_{j+1}, \omega) & is\_end_j \ is\ false \end{cases} \tag{1}$$

G) Use the mean square error loss function $\frac{1}{m}\sum_{j=1}^{m}\left(y_j - Q(\phi(S_j), A_j, \omega)\right)^2$, to update all parameters of the current network through gradient back propagation of the neural network,

H) Use $J(\theta) = -\frac{1}{m}\sum_{j=1}^{m} Q(s_i, a_i, \theta)$ to update all parameters of the Actor's current network through gradient back propagation of the neural network.

I) If $T\%C = 1$, then update the target network and Actor target network parameters:

$$\omega' \leftarrow \tau\omega + (1-\tau)\omega' \tag{2}$$

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta' \tag{3}$$

J) If S_{j+1} is the termination state, the current round iteration is completed, otherwise go to step B).

In the MADDPG algorithm, $\theta = [\theta_1, \ldots, \theta_n]$ represents the parameters of the strategy of n agents, and $\pi = [\pi_1, \ldots, \pi_n]$ represents the strategy of n agents. For the cumulative expected reward $J(\theta_i) = \mathrm{E}_{s\sim\rho^\pi, a_i\sim\pi_i}[\sum_{t=0}^{\infty}\gamma^t r_{i,t}]$ of the i-th agent, considering a random strategy, its strategy gradient is:

$$\nabla_{\theta_i} J(\theta_i) = \mathrm{E}_{s\sim\rho^\pi, a_i\sim\pi_i}[\nabla_{\theta_i}\log\pi_i(a_i|o_i)Q_i^\pi(\mathrm{x}, \mathrm{a}_1, \ldots, \mathrm{a}_n)] \tag{4}$$

Among them, o_i represents the observation of the i-th agent, $x = [\mathrm{o}_1, \ldots, \mathrm{o}_n]$ represents the observation vector, and state $\mathrm{Q}_i^\pi(\mathrm{x}, \mathrm{a}_1, \ldots, \mathrm{a}_n)$ represents the centralized state-action function of the i-th agent. Since each agent can learn its own

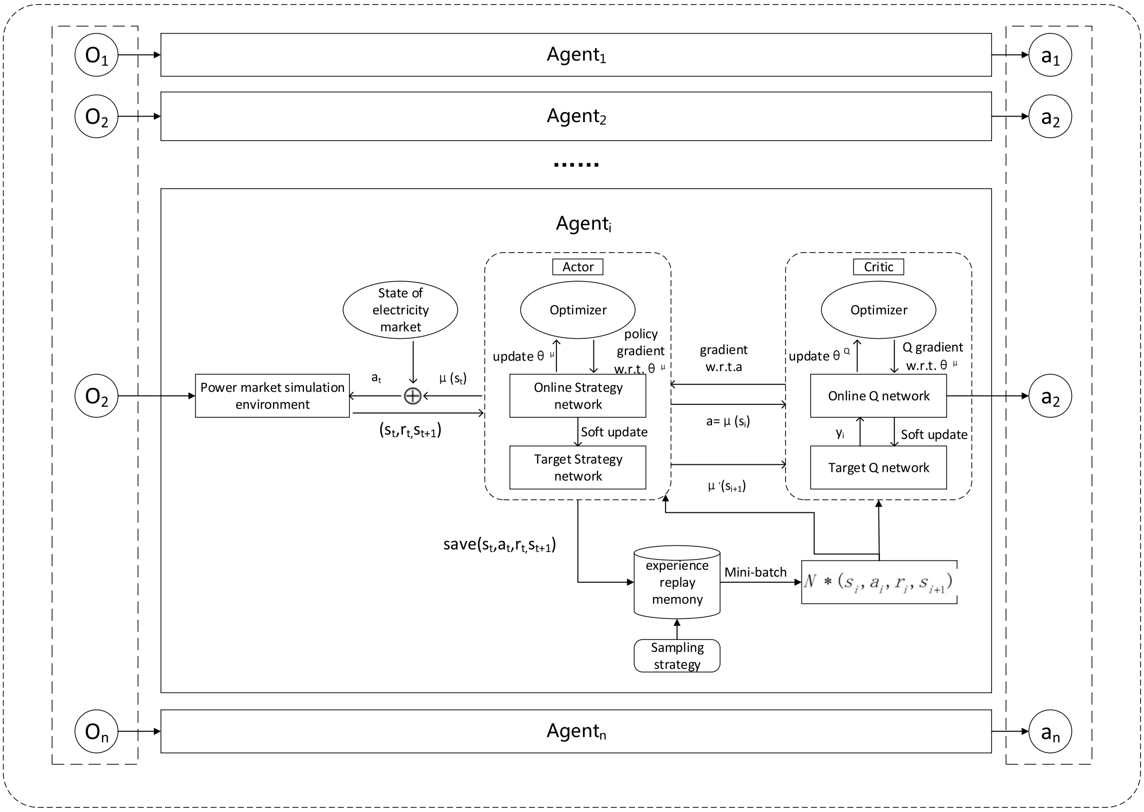

FIGURE 2. The implementation framework of MADDPG algorithm.

function Q_{i}^{π} independently and has its own reward function, it can complete cooperation or competition tasks.

For deterministic strategy $\mu_{\theta_{\mathrm{i}}}$, the gradient formula is:

$$\nabla_{\theta_{\mathrm{i}}} J(\mu_i) = \mathrm{E}_{\mathrm{x,a}\sim D}[\nabla_{\theta_{\mathrm{i}}} \mu_i(a_i|o_i)\nabla_{a_{\mathrm{i}}} Q_i^{\mu}(\mathrm{x,a}_1,\ldots,\mathrm{a}_n)\big|_{a_{\mathrm{i}}=\mu_{\mathrm{i}}(o_{\mathrm{i}})}] \tag{5}$$

In the formula, D is an experience store, and the element composition is $(x, x', a_1, \ldots, a_n, r_1, \ldots, r_n)$.

The centralized critic update method draws on the idea of Temporal-Difference and target network in DQN:

$$L(\theta_i) = \mathrm{E}_{\mathrm{x,a,r,x'}}[(Q_i^{\mu}(\mathrm{x,a}_1,\ldots,\mathrm{a}_n)\text{-y})^2] \tag{6}$$

$$y = r_i + \gamma \overline{Q}_i^{\mu'}(x', a_1', \ldots, a_{\mathrm{n}}')\big|_{a_j'=\mu_j'(o_j)} \tag{7}$$

$\overline{Q}_i^{\mu'}$ represents the target network, $\mu' = [\mu_1', \ldots, \mu_{\mathrm{n}}']$ is the parameter θ_j' of the target strategy that has a lagging update. The strategies of other agents can be obtained by fitting approximation, without the need for communication and interaction between agents. The critic borrows global information for learning, while the actor only uses local observation information.

In formula (6), the methods of fitting and approximating the strategies of other agents are as follows: Each agent maintains n-1 strategy approximation functions, and $\hat{\mu}_{\phi_i^j}$ represents the function approximation of the i-th agent to the j-th agent's strategy μ_{j}. This approximate strategy learns by maximizing the logarithmic probability of the action of agent j and adding an entropy regularization term:

$$L(\phi_{\mathrm{i}}^j) = -E_{o_j,a_j}[\log \hat{\mu}_{\phi_{\mathrm{i}}^j}(a_j\,|o_j) + \lambda H(\hat{\mu}_{\phi_{\mathrm{i}}^j})] \tag{8}$$

As long as (8) is minimized, an approximate estimate of the strategies of other agents can be obtained. Therefore, the formula (7) can be replaced by the following function:

$$y = r_i + \gamma \overline{Q}_i^{\mu'}(x', \hat{\mu}^1_{\phi_{\mathrm{i}}^j}(\mathrm{o}_1), \cdots, \hat{\mu}^n_{\phi_{\mathrm{i}}^j}(\mathrm{o}_n)) \tag{9}$$

Before update Q_{i}^{μ}, using a sample batch of experience playback to update $\hat{\mu}_{\phi_{\mathrm{i}}^j}$.

Since the strategy of each agent is being updated and iterated, the environment is unstable for a specific agent. Therefore, in a competitive environment, an agent often overfits a strong strategy to its competitors. But this strong strategy is very fragile. This is because that with the updating of competitors' strategies, it is difficult for this strong strategy to adapt to the new opponent's strategy.

In order to better solve the above problems, MADDPG proposed a strategy set idea. Strategy μ_{i} of the i-th agent consists of a set of K sub-strategies, and only one sub-strategy $\mu_{\mathrm{i}}^{(k)}$ is used in each training episode. For each agent, it is necessary to maximize the overall reward $J_{\mathrm{e}}(\mu_i) = \mathrm{E}_{\mathrm{k}\sim unif(1,K), s\sim\rho^{\mu}, a\sim\mu_{\mathrm{i}}^{k}}[\sum_{t=0}^{\infty}\gamma^t r_{i,t}]$ of its strategy set, and to construct a memory storage $D_{\mathrm{i}}^{(k)}$ for each sub-strategy k.

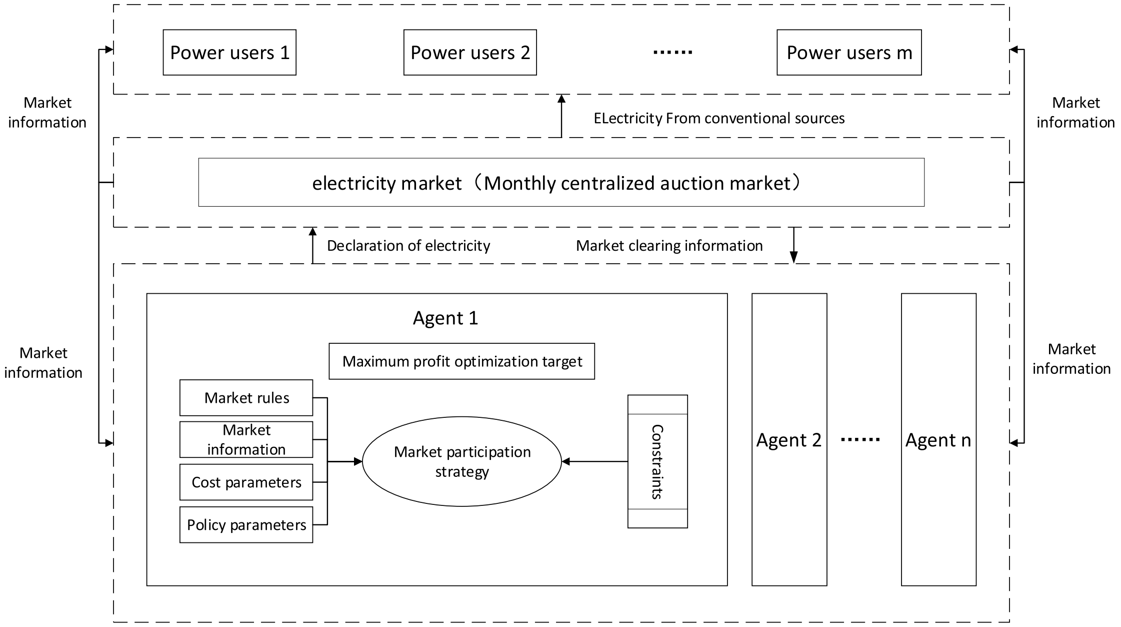

FIGURE 3. Schematic diagram of scenario assumptions.

Taking into account the overall effect of the optimization strategy set, the update gradient for each sub-strategy is:

$$\nabla_{\theta_i^{(k)}} J_e(\mu_i) = \frac{1}{K}\mathrm{E}_{x,a\sim D_i^{(k)}}[\nabla_{\theta_i^{(k)}}\mu_i^{(k)}(a_i \times |o_i)\nabla_{ai}Q^{\mu_i}(\mathrm{x},\mathrm{a}_1,\ldots,\mathrm{a}_n)\Big|_{a_i=\mu_i^{(k)}(o_i)}] \quad (10)$$

III. BIDDING MODEL OF GENERATION COMPANIES IN MEDIUM AND LONG-TERM ELECTRICITY MARKET BASED ON MADDPG ALGORITHM

A. SCENARIO ASSUMPTIONS

This paper assumes that the research scenario is shown in Figure 3. In the electricity market, thermal power companies formulate their trading strategies based on their own operating data and incomplete market information, with the goal of maximizing profits. Thermal power companies sell electricity in the electricity market, and electricity users buy demand electricity on the grid. According to their own cost situation and incomplete market information, each thermal power company decides its own declared price. Based on the cooperation of thermal power companies with different costs, this paper sets up the following two types of scenarios, and constructs a non-cooperative game model and a cooperative game model between thermal power companies.

Scenario 1: Thermal power companies with different costs do not cooperate, and thermal power companies makes quotations with the goal of maximizing their own interests.

Scenario 2: Thermal power companies with different costs can cooperate, and each thermal power company makes quotations with the goal of maximizing overall benefits.

In the scenario, the market clearing mechanism of thermal power companies transactions adopts uniform marginal price clearing. This paper does not consider the physical constraints of the power system network. The market clearing process is as follows.

(1) the demander with the highest priority (the largest declared price) and the supplier with the highest priority (the smallest declared price) are matched first, and then the demander and supplier with the second highest priority are matched, and so on;

(2) When the declared prices of the demand side and the supply side are equal, or all the demand side and the supply side have completed the matching, the liquidation is ended and the transaction is completed;

(3)The clearing price is the average price declared by the last matched demander and supplier.

B. NON-COOPERATIVE GAME MODEL

1) PROFIT CALCULATION MODEL

The medium and long-term market selected in this paper is the monthly centralized bidding market, and the profit of thermal power companies is the profit from the sale of electricity minus the cost of power generation. In the monthly centralized bidding transaction, the reference cost of the thermal power supplier's quotation decision is the thermal power supplier's marginal power generation cost, as shown below:

$$C_G^T(P_G) = a_G \times (P_G)^2 + b_G \times P_G + c_G \quad (11)$$

$$C_G^{B,Mon} = a_G Q_G^{B,Mon}\frac{2Q_G^{B,Year,D} + Q_G^{B,Mon}}{T^{Mon}} + b_G Q_G^{B,Mon} \quad (12)$$

$C_G^T(P_G)$ is the production cost of thermal power companies; a_G, b_G, c_G is the quadratic, primary and constant coefficients of the production costs of thermal power companies;

P_G is the output of thermal power units; $C_G^{B,Mon}$ is the power generation cost of thermal power companies in monthly centralized bidding transactions; $Q_G^{B,Year,D}$ is the monthly decomposition power of the thermal power supplier's annual contract power; $Q_G^{B,Mon}$ is the transaction electricity volume of thermal power companies in monthly centralized bidding transactions; T^{Mon} is the number of monthly hours;

The non-cooperative profit model of thermal power companies can be expressed as:

$$R_G^{B,Mon} = Q_G^{B,Mon} p_{mc} - (a_G Q_G^{B,Mon} \frac{2Q_G^{B,Year,D} + Q_G^{B,Mon}}{T^{Mon}} + b_G Q_G^{B,Mon}) \tag{13}$$

$R_G^{B,Mon}$ is the profit of thermal power companies in monthly centralized bidding transactions; p_{mc} is the market clearing price of the monthly centralized bidding transaction.

2) DECISION MODEL

By using their own market power and adopting a certain bidding strategy, power producers are called market holding behavior, including physical holding and economic holding. Physical retention refers to that the power generation companies hold the power generation quantity to reduce the market supply and make the power generation companies with higher declared electricity price become the marginal power generation companies in the market; economic retention refers to that the power generation companies declare electricity price higher than their marginal generation cost and make themselves become the marginal power generation companies in the market, so as to improve the market clearing price.

In monthly centralized bidding transactions, the declared electricity quantity and declared electricity price of thermal power companies are as follows.

$$Q_G^{R,Mon} = \alpha_G Q_G^{S,Mon} = \alpha_G (Q_G^{Max,Mon} - Q_G^{B,Year,D}) \tag{14}$$

$$p_G^{R,Mon} = \beta_G C_G^{M,Mon} \tag{15}$$

$Q_G^{R,Mon}$ is the declared electricity quantity of thermal power companies in monthly centralized bidding transactions; α_G is the decision coefficient of the declared electricity quantity in monthly centralized bidding transactions of the thermal power companies; $Q_G^{S,Mon}$ is the maximum monthly remaining power generation capacity of thermal power companies; $Q_G^{Max,Mon}$ is the monthly maximum power generation capacity of thermal power companies; $Q_G^{B,Year,D}$ is the monthly decomposition power of the thermal power supplier's annual contract power; $p_G^{R,Mon}$ is the declared electricity price of the thermal power supplier in monthly centralized bidding transactions; β_G is the decision coefficient of the declared electricity price in monthly centralized bidding transactions of the thermal power supplier; $C_G^{M,Mon}$ is the monthly average marginal power generation cost of thermal power companies.

The decision-making models of various thermal power companies are as follows:

$$\max R_G^{B,Mon} = Q_G^{B,Mon} p_{mc} - (a_G Q_G^{B,Mon} \frac{2Q_G^{B,Year,D} + Q_G^{B,Mon}}{T^{Mon}} + b_G Q_G^{B,Mon}) \tag{16}$$

$$b_{n,t} P_{n,t}^{\min} \leq P_{n,t} \leq b_{n,t} P_{n,t}^{\max} \tag{17}$$

$$P_{n,t} - P_{n,t-1} \leq \Delta P_n^U b_{n,t-1} + P_{n,t}^{\min}(b_{n,t} - b_{n,t-1}) + P_{n,t}^{\max}(1 - b_{n,t}) \tag{18}$$

$$P_{n,t} - P_{n,t-1} \leq \Delta P_n^D b_{n,t} + P_{n,t}^{\min}(b_{n,t} - b_{n,t-1}) + P_{n,t}^{\max}(1 - b_{n,t-1}) \tag{19}$$

$$T_{n,t}^D - (b_{n,t} - b_{n,t-1}) T_D \geq 0 \tag{20}$$

$$T_{n,t}^U - (b_{n,t} - b_{n,t-1}) T_U \geq 0 \tag{21}$$

$$\sum_t \eta_{n,t} \leq \eta_{n,\max} \tag{22}$$

$$\sum_t \gamma_{n,t} \leq \gamma_{n,\max} \tag{23}$$

$$Q_G^{R,Mon} \leq Q_G^{\max} \tag{24}$$

$$p_G^{min} \leq p_G^{R,Mon} \leq p_G^{\max} \tag{25}$$

Equation (17) is the thermal power output constraint; Equations (18) and (19) are thermal power climbing constraints; Equations (20) and (21) are the minimum continuous start stop time constraints of thermal power plants; Equations (22) and (23) are the constraints of maximum startup and shutdown times of thermal power plants; The formula (24) is the quantity restriction of thermal power enterprises, and the formula (25) is the quotation restriction of thermal power enterprises.

Where: $P_{n,t}^{\min}$ and $P_{n,t}^{\max}$ are the minimum and maximum output of fire motor group n respectively; $b_{n,t}$ is the 0-1 variable of the start-up and stop state of unit n in period T, $b_{n,t} = 1$ is the start-up state and $b_{n,t} = 0$ is the shutdown state; ΔP_n^U and ΔP_n^D are the maximum ascent and descent rates of unit n, respectively; T_U and T_D are the minimum continuous start-up time and the minimum continuous shutdown time respectively; $T_{n,t}^U$ and $T_{n,t}^D$ are the continuous start-up time and continuous shutdown time of unit n at time t respectively; $\eta_{n,t}$ and $\gamma_{n,t}$ are the switching variables of start-up and shutdown respectively. $\eta_{n,t}$ indicates whether unit n switches to start-up state in time t, and $\gamma_{n,t}$ indicates whether unit n switches to shutdown state in time t; $\eta_{n,\max}$ and $\gamma_{n,\max}$ are the maximum start-up and shutdown times of unit n in t period respectively.

C. COOPERATIVE GAME MODEL

1) PROFIT CALCULATION MODEL

Various thermal power companies cooperate to form an alliance, and they make decisions with the goal of maximizing the overall profits of the alliance. The profit function is:

$$\begin{aligned} R_G^T &= \sum_{i=1}^{n} R_{Gi}^{B,Mon} \\ &= Q_{Gi}^{B,Mon} p_{mc} - (a_{Gi} Q_{Gi}^{B,Mon} \frac{2Q_{Gi}^{B,Year,D} + Q_{Gi}^{B,Mon}}{T^{Mon}} \\ &\quad + b_{Gi} Q_{Gi}^{B,Mon}) \end{aligned} \tag{26}$$

R_G^T is the total profits function of the alliance.

The cooperative game involves the issue of how to distribute the benefits among multiple entities. Therefore, how the cooperative benefits are distributed among the thermal power companies will directly affect the achievement of cooperation. In this paper, the Shapley value method is used to study the cooperative income distribution among thermal power companies. The core idea of this method is to distribute the benefits of participants according to their contributions to the alliance. The more contributions, the more benefits.

We assume that subset $S \subseteq M$ of set $M = d\{1, 2, \ldots, M\}$ of any non-empty participant is called a coalition. The Shapley value can be used to calculate the profit φ distributed by participant i, as shown in formulas: (27) and (28).

$$\varphi_{\mathrm{i}} = \sum_{s} \omega(|S|)(v(S) - v(S - \{i\})) \tag{27}$$

$$\omega(|S|) = \frac{(m - |S|)!(|S| - 1)!}{m!} \tag{28}$$

In the above formulas: $|S|$ is the number of participants in subset S; $v(S)$ is the profit of alliance cooperation including participant i; $v(S - \{i\})$ is the alliance cooperation profit that does not include participant i; $\omega(|S|)$ is the weighting factor; $m!$ is the number of all possible arrangements of participants in the cooperative game.

If each thermal power supplier is denoted as 1, 2,…, n, then m = n. $v(1)$, $v(2)$, , $v(n)$ respectively represents the non-cooperative profit of thermal power companies 1 to n, and $v(T)$ represents the total profits of the cooperation between the two parties. According to formulas (27) and (28), the respective benefits allocated to wind farms and electric vehicle aggregators can be obtained.

2) DECISION MODEL

When various thermal power companies cooperate, the decision model can be described as follows:

$$\begin{aligned} \max R_G^T &= \sum_{i=1}^{n} R_{Gi}^{B,Mon} \\ &= Q_{Gi}^{B,Mon} p_{mc} - (a_{Gi} Q_{Gi}^{B,Mon} \frac{2Q_{Gi}^{B,Year,D} + Q_{Gi}^{B,Mon}}{T^{Mon}} \\ &\quad + b_{Gi} Q_{Gi}^{B,Mon}) \end{aligned} \tag{29}$$

Its thermal power output constraints, quotation constraints, and volume constraints are the same as the decision-making model of non-cooperative games.

IV. GAME MODEL SOLVING BASED ON MADDPG ALGORITHM

In the electricity market, based on their own operating data and incomplete market information, each thermal power company formulates its bidding strategy with the goal of maximizing profits.

A. ACTOR STRATEGY NETWORK MODEL

The goal of the Actor strategy network is to learn and optimize strategies to make performance of the strategy better and better. Therefore, the input of the Actor strategy network model is the state feature, and the output is the action distribution. According to scenarios 1 and 2, the status features include the power market status and the thermal power supplier's own status, and the action distribution is the thermal power supplier's decision-making behavior on the declared electricity quantity and the declared electricity price.

1) STATE

Before the start of monthly centralized bidding transactions, regulators will disclose market information. On the one hand, based on the public market information released by power trading institutions, thermal power companies can judge the supply and demand of the power market this month; on the other hand, combined with private information such as the maximum remaining power generation capacity, thermal power companies can judge their share and market power in the electricity market this month. Therefore, the status features include two parts: the status of the electricity market, and the status of thermal power companies themselves.

a: STATE OF THE ELECTRICITY MARKET

The state of the electricity market is expressed by the "market supply-demand ratio", which is the ratio of market electricity supply to market electricity demand.

$$R = \frac{Q_S}{Q_D} \tag{30}$$

R is the market supply-demand ratio, Q_S is market supply, Q_D is market demand.

b: THE STATUS OF THERMAL POWER COMPANIES THEMSELVES

Thermal power companies' own status is expressed by "thermal power companies' market share", which is the ratio of the remaining maximum power generation of thermal power companies to market power demand.

$$S_G = \frac{\sum Q_G^{S,Mon}}{Q_D} \tag{31}$$

S_G is the market share of thermal power companies; $Q_G^{S,Mon}$ is the maximum monthly remaining power generation capacity of thermal power companies.

Based on the market public information of the monthly centralized bidding transaction and its own private information, thermal power companies make decisions on the declared electricity quantity and the declared electricity price. The quantity and price have been shown in the decision model in the previous section.

The action distribution in the Actor strategy network model can be expressed as a two-dimensional vector $[\alpha_G, \beta_G]$ of the declared power decision coefficient and the declared power price decision coefficient. Among them, in the monthly centralized bidding transaction, the declared electric power of the thermal power supplier cannot be less than zero and not greater than the maximum remaining monthly power generation. Therefore, the range of the declared electric power decision coefficient α_G is [0,1]; At the same time, in the monthly centralized bidding transaction, the declared electricity price of the thermal power supplier should be greater than or equal to its monthly average marginal power generation cost, and considering the monthly centralized bidding transaction regulations, thermal power companies are not allowed to make profiteering quotations. Therefore, the range of the declared electricity price decision coefficient β_G is [1,1.2].

B. CRITIC VALUE NETWORK MODEL

The goal of the Critic value network model is to learn and evaluate the value function, so that the value function can more accurately evaluate the pros and cons of the strategy. Its input is a state feature, and its output is a state value function. The state characteristics of the Critic value network model are the same as the state characteristics of the Actor strategy network model, and the state value function is the discounted sum of the profit of different thermal power companies.

C. THE PROCESS OF MODEL SOLUTION

The quotation game problem of thermal power companies in the medium and long-term power market studied in this paper is a complex multi-agent system problem, which is suitable for multi-agent reinforcement learning methods. In this paper, the MADDPG algorithm is used to solve the non-cooperative game and cooperative game models of thermal power companies with different costs. The model solution framework is shown in Figure 4. There are n thermal power companies in the multi-agent system, and each agent has a strategy network.

The MADDPG algorithm uses centralized training and distributed execution. Firstly, in the training process, n agents use joint strategy $\vec{\pi} = (\vec{\pi_1}, \vec{\pi_2}, \ldots, \vec{\pi_n})$ to interact with the environment. At the same time, the joint behavior value function $Q_i(o_1, a_1, o_2, a_2, \ldots, o_n, a_n)$ of each agent i is evaluated, and the strategy of each agent is updated according to the gradient of the joint behavior value function to the strategy parameter. The policy input of each agent i is local observation o_i, and the output is action a_i of agent i. Secondly, the input of agent i in the execution phase is local observation o_i, and the output is action a_i of agent i.

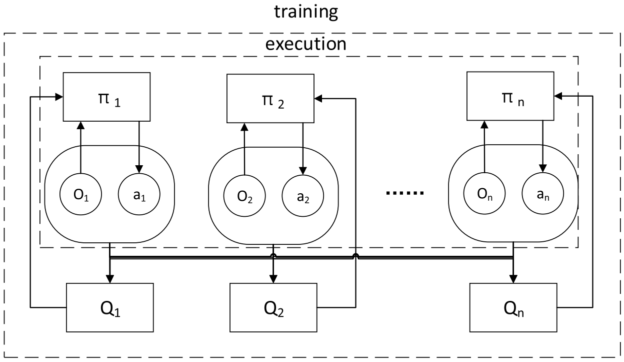

FIGURE 4. The solution framework of the MADDPG algorithm model.

In this paper, the input of the strategy network of the MADDPG algorithm is the electricity market and its own state characteristics, and the output is the quotation behavior of thermal power companies. The input of the value network is the same as that of the strategy network, and the output is the value function of the quotation behavior. The strategy network selects behaviors based on the probability distribution (that is, the declared electricity quantity and the declared electricity price). The value network first judges the quality of the behavior of the strategy network, and then the strategy network adjusts the probability distribution of the behavior according to the evaluation value of the value network. The environmental model is a medium and long-term electricity market clearing model. The input of this model is the action taken by the agent, and the output is the reward obtained by the agent and the state of the next month. The process of using the MADDPG algorithm to solve the model is shown in Table 1.

V. EXAMPLE ANALYSIS

A. PARAMETER SETTING

Based on the medium and long-term electricity market transaction data of a province in China, the basic parameters of medium and long-term electricity market simulation boundary conditions are set. The power generation type of power generation companies is coal-fired thermal power, and the rated capacity and power generation cost parameters are shown in Table 2.

The monthly basic generation plan, monthly maximum generation, monthly residual generation, market demand of medium and long-term electricity market, supply-demand ratio, HHI and other market situation data of each power producer are shown in the appendix.

Market clearing: unified marginal price clearing is adopted, and all the electricity transactions of power producers are settled according to the system marginal price.

Security check: at present, there is no monthly centralized bidding transaction, and the market clearing power is reduced due to the failure of security check, so this paper does not

TABLE 1. MADDPG algorithm flow.

```
Procedure of method:
1、 For episode=1 to M do
2、 Initiating a random process N for action exploration
3、 Accepting the initial state x
4、 For t=1 to max-episode-length do
5、 For each agent i, sampling action $a_i = \mu_{\theta_i}(o_i) + N_t$
    according to the strategy network plus noise
6、 Performing joint action $\vec{a} = (a_1, \dots, a_n)$, and
    observing the return $\vec{r} = (r_1, \dots, r_n)$ and new state $x'$
    of each agent
7、 Storing the joint state, joint action, and joint return
    $(\vec{x}, \vec{a}, \vec{r}, \vec{x}')$ in buffer D
8、     Pushing one step forward, that is $x = x'$
9、     For agent i=1 to N do
10、        Collecting S mini batches of data
    $(\vec{x}^j, \vec{a}^j, \vec{r}^j, \vec{x}'^j)$ from experience playback D, and
    j=a,⋯S
11、        Setting the TD target of the joint behavior value
    function：$y^j = r_i^j + \gamma Q_i^{\mu'}(x'^j, a_1, a_2)|a_k' = \mu_k'(o_k^i)$
12、        Updating the critic by minimizing the loss
    function $L(\theta_i) = \frac{1}{S}\sum_j (y^j - Q_i^u(\vec{x}^j, a_1^j, a_2^j))2$
13、        Calculating the deterministic strategy
    gradient：$\nabla_{\theta_i} J =$
    $\frac{1}{S}\sum_j \nabla_{\theta_i}\mu_i(a_i|o_i)\nabla_{a_i} Q_i^{\vec{\mu}}(x, a_1, \dots, a_n)|a_i = \mu_i(o_i)$
14、      End for
15、      Updating the target network $\theta_i' \leftarrow \tau\theta_i + (1 - \tau)\theta_i'$ of the joint reward function of agent i
     16、  end for
     17、end for
```

consider the physical constraints and security check of power system.

Deviation assessment: unless there are random factors such as unexpected unit failure, it is generally considered that the power generation companies strictly implement the power generation plan arranged by the power trading agency and the power dispatching agency, and the deviation of contract power is basically ignored.

There are many power selling companies and users on the demand side. By fitting the electricity demand curve in the monthly centralized bidding transaction, we find that the electricity demand curve in the monthly centralized bidding transaction meets the power function of index 4, and the demand curve fitted by historical data is as follows:

TABLE 2. Generator parameters.

Power producer	Rated capacity /MW	a	b	c
1	300	0.088	305	1150
2	500	0.057	285	1280
3	600	0.052	274	420
4	650	0.050	267	1480
5	850	0.045	260	1650
6	1000	0.036	252	1800

$$p = 0.36 - 0.05 \times \left(\frac{q}{q_{\max}}\right)^4 \tag{32}$$

p is electricity price; q is electricity demand; $q_{\max}$ is monthly maximum electricity demand.

In this paper, the MADDPG algorithm of multi-agent reinforcement learning method is adopted to solve the power supplier quotation game model. Simulation training of non-cooperative and cooperative scenarios is carried out on the above data respectively. The models train 2500episodes respectively, the batch size is 32, and the number of hidden nodes of neural network is 500. Discount rate: 0.9; Policy cross entropy weight: 0.01; Actor policy network learning rate: 0.0001; Number of hidden layer nodes of actor policy network: 200; Actor policy network activation function: relu; Critic value network learning rate: 0.01; Number of hidden layer nodes of critical value network: 200; Critical value network activation function: softmax. The simulation environment is Intel Core i5-9400@2.90GHz, 6 cores and 12 threads, the memory is 16GB, the software is configured with Python3.7.1 and Tensorflow2.2.0.

B. ANALYSIS ON BIDDING STRATEGY OF POWER SUPPLIERS UNDER COMPETITION

All the generators are agent generators, and the bidding model based on maddpg algorithm is used for bidding.

With the goal of maximizing the overall profits of generators, the DQN algorithm, DDPG algorithm and MADDPG algorithm are used to carry out the simulation of generator quotes in the medium and long-term power market. The convergence results of the scalar, declared volume, quotation, and profit of each power generation company are shown in Table 3, and the overall profit convergence process of the power generation company is shown in Figure 5. It can be seen from the figure that the MADDPG algorithm starts to converge in about 350 rounds, and the DDPG algorithm and the DQN algorithm start to converge after about 1000 rounds. The convergence speed of the MADDPG algorithm is about 2.8 times that of the other two algorithms. In addition, it can be clearly observed from the figure that the convergence result of the overall profit of the generator in the MADDPG algorithm is significantly greater than the convergence result

TABLE 3. **Monthly basic generation plan of power suppliers.**

Power producer	January	February	March	April	May	June	July	August	September	October	November	December
1	12686	11690	12686	12360	12686	12360	12686	7210	12360	12686	12360	12686
2	20902	5798	4375	20293	20902	20293	20902	20902	20293	20902	20293	20902
3	29090	26707	29090	28306	29090	28306	29090	29090	28306	29090	5361	29090
4	28607	26361	28607	27871	28607	27871	28607	28607	27871	28607	27871	16257
5	34852	32024	34852	33922	34852	33922	34852	34852	33922	10188	33922	34852
6	34012	43475	47244	46007	47244	32584	47244	47244	46007	47244	46007	47244

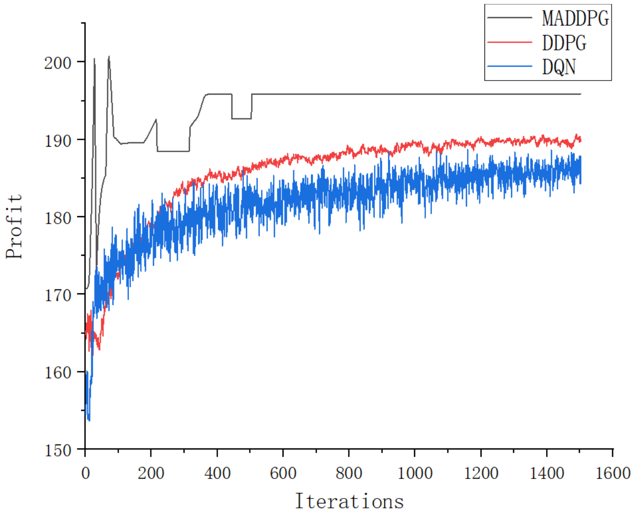

FIGURE 5. **Comparison chart between algorithms.**

of the DDPG algorithm and the DQN algorithm. The reason is that in the case of limited information, the MADDPG algorithm adopts the fitting approximation method to obtain the strategies of other thermal power companies, and forms a multi-agent through different neural networks. Through the output layer of the neural network, the agent's continuous behavior decision is realized, and a better quotation decision is obtained.

Based on python, the multi-agent medium and long-term electricity market simulation is carried out. Taking the income of each power producer as its reward function, the iteration and convergence process of scalar, declared quantity, quotation and income of each power producer is shown in the figure. By analyzing the clearing results of medium and long-term electricity market, the bidding strategy of power producers and the operation of electricity market are explained.

Under the market clearing mechanism of unified marginal price clearing, there is no decisive and inevitable relationship between the system marginal price and the declared price of other generators except the market marginal generator. Therefore, there is a certain game behavior in the bidding strategy of power generation companies, which is essentially that power generation companies use their own market power to influence the market clearing price, and then enhance their own income.

Through the simulation results of multi-agent medium and long-term electricity market simulation scenarios, we can find the complex game behaviors of power producers with different cost and price in the market.

1) HIGH GENERATION COST

This type of generator is mainly generator 1. According to Figure 6, it is difficult for power producer 1 to obtain profits through market trading for what kind of market holding behavior or bidding strategy it adopts. It can be seen that when generator 1 tries to raise the marginal clearing price by increasing the quoted price and reducing the declared quantity, it is out of the market marginal generator because of its inferior generation cost, and the scalar value is always 0.

2) THE COST OF ELECTRICITY GENERATION IN CHINA

This type of generators are mainly generators 2 and 3. If the price is quoted according to the ''Declaration of total electricity marginal cost'', the marginal generation cost of 2,3 generation companies will be higher than the market clearing price, resulting in the risk of failure to win the bid. It can be seen from the figure that at the beginning of the iteration, generator 2 tries to take physical retention behavior to reduce the declared electricity quantity and its marginal generation cost, so as to win the bid and obtain the corresponding market benefits. However, the scalar of generator 2 is almost zero, and it is observed that the marginal clearing price is slightly higher than the marginal cost, so generator 2 uses the form of small profit but quick turnover to reduce the quotation and increase the declared quantity, so as to become a marginal generator. As can be seen from Figure 6, its strategy is more successful. Power producer 3 is a opportunist. At the beginning of the iteration, it raised the bid twice for economic holding, but failed.

3) LOW GENERATION COST

This type of generators are mainly generators 4, 5 and 6, which have advantages in market power and power generation cost. However, unless the market supply exceeds

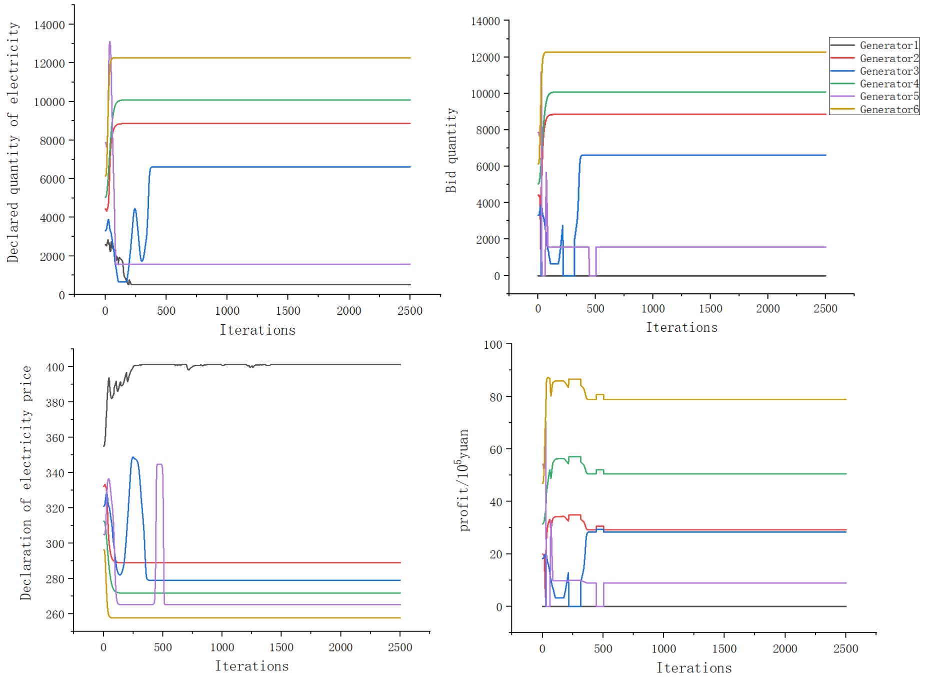

FIGURE 6. Iterative process of clearing indexes in competitive game.

the demand, the extra profit brought by raising the market clearing price is often less than the loss of retained electricity. Therefore, this type of generator often adopts a relatively safe bidding strategy based on marginal cost pricing. As can be seen from Figure 6, the bidding of low-cost power suppliers 4 and 6 is based on their respective marginal costs, and the bidding is low, so they can win all the bidding; while power supplier 5 is a opportunist. At the beginning of the iteration, power supplier 5 tries to carry out economic retention by increasing the bidding and declaration quantity, but the bidding is too high, which exceeds the marginal clearing price, resulting in a sharp decrease in its scalar After the quotation of other power suppliers is stable, power supplier 5 carries out economic retention by raising the quotation, but it is not successful.

As can be seen from Figure 7, in the initial stage of iteration, due to speculators' economic or physical holding, the market clearing price is relatively high and the clearing power is relatively low. In the overall downward trend of market clearing price, there are four rises, corresponding to the two market holding behaviors of 3 and 5 respectively.

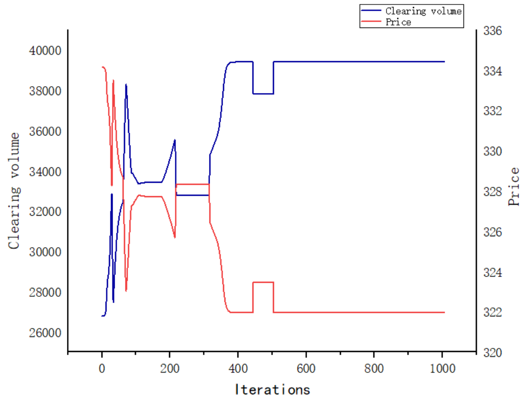

FIGURE 7. Changes of market clearing price and total clearing volume.

By comparing the declared capacity and the maximum capacity of power generation companies, the vast majority of power generation companies will adopt the physical retention behavior, but whether the power generation companies with different market power and generation cost levels will adopt the physical retention behavior, and the size of the retained capacity are different. At the same time, on the basis of physical retention, some high market power and low-cost power producers will also take certain economic retention behavior. The bidding strategy and game behavior of power suppliers will be more complex.

The residual supply rate is used to describe the market power of power generation companies, and the deviation degree of electricity price is used to describe the generation cost level of power generation companies. The relationship between the residual supply rate of power generation companies, the deviation degree of electricity price and the proportion of retained electricity are shown in figures 8 and 9. The following conclusions can be drawn from the figure: 1) under the multi-agent decision-making, compared with the generators with larger surplus supply rate, the generators with smaller surplus supply rate will be more inclined to take the market speculation of physical retention; 2) except for the

TABLE 4. Maximum monthly power generation of power producers.

Power producer	January	February	March	April	May	June	July	August	September	October	November	December
1	17856	16128	17856	17280	17856	17280	17856	9216	17280	17856	17280	17856
2	29760	7680	5760	28800	29760	28800	29760	29760	28800	29760	28800	29760
3	35712	32256	35712	34560	35712	34560	35712	35712	34560	35712	5760	35712
4	38688	34944	38688	37440	38688	37440	38688	38688	37440	38688	37440	19968
5	50592	45696	50592	48960	50592	48960	50592	50592	48960	13056	48960	50592
6	40320	53760	59520	57600	59520	38400	59520	59520	57600	59520	57600	59520

TABLE 5. Monthly surplus generation of power producers.

Power producer	January	February	March	April	May	June	July	August	September	October	November	December
1	5170	4438	5170	4920	5170	4920	5170	2006	4920	5170	4920	5170
2	8858	1882	1385	8507	8858	8507	8858	8858	8507	8858	8507	8858
3	6622	5549	6622	6254	6622	6254	6622	6622	6254	6622	399	6622
4	10081	8583	10081	9569	10081	9569	10081	10081	9569	10081	9569	3711
5	15740	13672	15740	15038	15740	15038	15740	15740	15038	2868	15038	15740
6	6308	10285	12276	11593	12276	5816	12276	12276	11593	12276	11593	12276

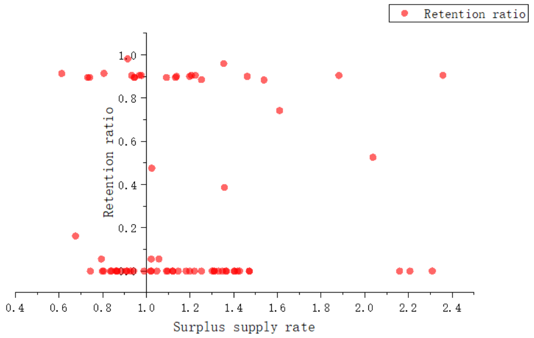

FIGURE 8. Relationship between surplus supply rate and electricity retention ratio.

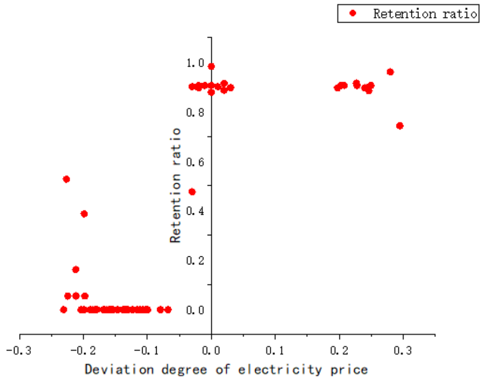

FIGURE 9. Relationship between electricity price deviation and electricity retention ratio.

generators whose bidding price is too high due to too high generation cost, the closer the deviation degree of electricity price is to 0.0%, the larger the proportion of retained electricity will be.

C. ANALYSIS ON BIDDING STRATEGY OF POWER SUPPLIERS UNDER COOPERATIVE GAME

Based on python, the multi-agent medium and long-term electricity market simulation is carried out. Taking the maximization of the overall income of power generation companies as its reward function, the iteration and convergence process of scalar, declared quantity, quotation and income of each power generation company is shown in the figure. By analyzing the clearing results of medium and long-term electricity market, the bidding strategy of power generation companies and the operation of electricity market are explained.

From Figure 10, we can see that when the total profits maximization of power generation companies is regarded as the reward function, power generation companies 4, 5 and 6 win a large number of bids by offering high quantity and low price; power generation companies 2 and 3 bid too much, so it is difficult to win in the market; power generation company 1 with medium price wins a small number of bids. In the initial stage of the iteration, each power producer makes a tentative offer. After a period of learning, when their own cost is known, the unit with lower generation cost will win the bid, so as to achieve the goal of maximizing the overall profits of the power producer.

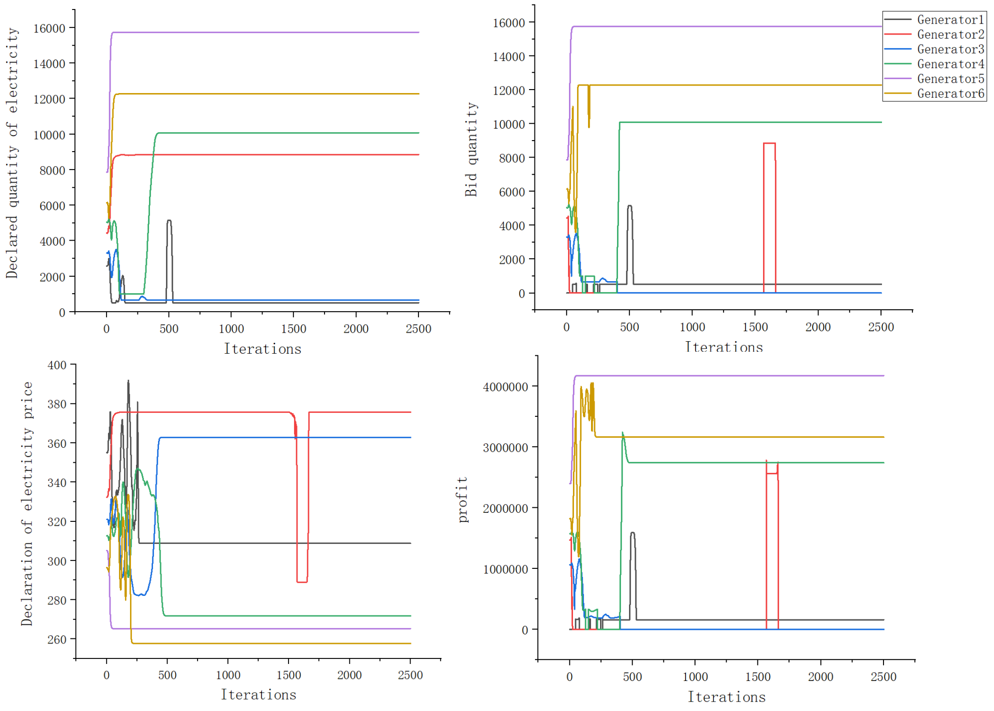

FIGURE 10. Iterative process of clearing indexes in cooperative game.

TABLE 6. Market situation.

	January	February	March	April	May	June	July	August	September	October	November	December
Market electricity demand	44051.9	42057.6	63920.9	56578.1	51789.2	60031.8	32471.2	42799	48304.8	52261.9	56972.8	50017.9
Supply demand ratio	1.29	1 .19	0.96	1 .11	1.22	1.01	1.65	1.34	1.24	1.06	1.05	1 .18
HHI	1097.8	1164.3	1262.8	1159.7	1146.2	1100.9	1116.4	1172.1	1118.6	1103.5	1185.5	1135.7

TABLE 7. Comparison of convergence results in different algorithms' declaration volume and quotation.

	Declared volume						Declared price					
algorithm	Power producer 1	Power producer 2	Power producer 3	Power producer 4	Power producer 5	Power producer 5	Power producer 1	Power producer 2	Power producer 3	Power producer 4	Power producer 5	Power producer 6
MADDPG	517	8858	6622	10081	1574	12276	401.27	289.00	279.08	271.81	265.27	257.71
DDPG	540	8345	6316	9803	1543	11573	408.96	292.15	284.35	274.98	271.23	263.78
DQN	508	8457	6812	9201	1487	12017	411.25	293.41	285.62	277.54	274.35	266.23

By comparing the market clearing price and clearing volume in competitive game and cooperative game, we can get the following conclusions. When the iteration is stable, the total market clearing in the competitive game is greater than that in the cooperative game, and the market clearing price in the cooperative game is also greater than that in the cooperative game. The ultimate goal of market operation is to optimize the allocation of resources and improve the efficiency of comprehensive utilization of resources. The most intuitive performance in the market is the reduction of

TABLE 8. Comparison of convergence results between bid-winning power and income in different algorithms.

	Bid-winning power						Income					
algorithm	Power producer 1	Power producer 2	Power producer 3	Power producer 4	Power producer 5	Power producer 5	Power producer 1	Power producer 2	Power producer 3	Power producer 4	Power producer 5	Power producer 6
MADDPG	0	8858	6622	10081	1574	12276	0.000	29.185	28.387	50.550	8.922	78.856
DDPG	0	8345	6316	9803	1543	11573	0.000	27.226	27.018	48.127	6.987	77.015
DQN	0	8457	6812	9201	1487	12017	0.000	27.124	26.879	47.938	6.689	76.874

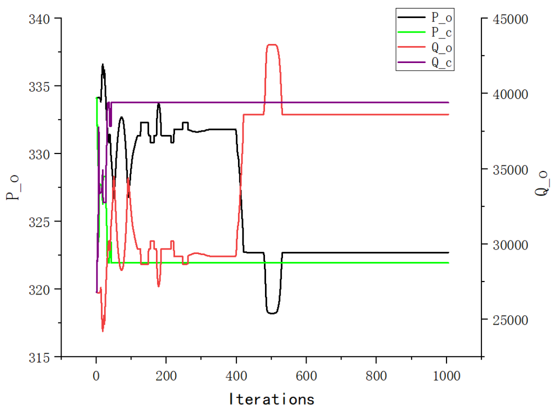

FIGURE 11. Changes of market clearing price and total clearing volume under cooperative game.

market clearing price. It can be seen from Figure 13 that the market efficiency of competitive game is higher than that of cooperative game.

According to the previous paper, almost all the winning units in the cooperative game are low-cost power producers. It can be found in Figure 11 that when the low-cost power producers hold a large amount of electricity, the market clearing price will increase and the market efficiency will be low. When the generation cost is slightly higher than the marginal price, the market clearing price will be reduced and the market efficiency will be improved.

VI. CONCLUSION

In this paper, a bidding model based on MADDPG reinforcement learning algorithm for medium and long-term electricity market is constructed, and the simulation of medium and long-term electricity market is carried out based on multi-agent. The main research conclusions are as follows.

A. BIDDING STRATEGY OF POWER SUPPLIERS

1) The bidding strategies of generation companies in medium and long term electricity market mainly include physical retention and economic retention. The bidding behavior of economic holding has high prediction accuracy for market clearing price and great market risk, so power generation companies prefer to take the bidding behavior of physical holding.

2) When the market supply and demand are relatively large and the power generation companies have certain market power, the power generation companies will have a greater probability to take the bidding behavior of physical holding. At the same time, the closer the deviation degree of electricity price is to 0%, the larger the proportion of electricity holding will be.

3) Physical retention will cause cost and substitution effects on market efficiency: when more marginal generators hold more electricity and the cost effect is greater than the substitution effect, the market clearing price will decrease and the market efficiency will increase; when more low-cost generators hold more electricity and the substitution effect is greater than the cost effect, the clearing price will increase and the market efficiency will decrease

B. SUGGESTIONS ON ELECTRICITY MARKET SUPERVISION

1) In the medium and long-term electricity market, compared with the economic retention behavior, the regulatory authorities should pay more attention to the supervision of the physical retention behavior of the power generation companies. By comparing the declared power of the power generation companies with their installed capacity and residual power generation, the physical retention behavior of the power generation companies can be found.

2) Different market forces and different generation costs have different impacts on the market clearing price and market efficiency. As the power producers with large market power and low generation cost adopt physical holding behavior, the market clearing price will be increased and the market efficiency will be reduced. Therefore, the regulatory authorities should pay more attention to the supervision and punishment of this type of power producers.

APPENDIX

See Tables 3–8.

REFERENCES

[1] J. Sun, M. Chen, H. Liu, Q. Yang, and Z. Yang, "Workload transfer strategy of urban neighboring data centers with market power in local electricity market," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3083–3094, Jul. 2020.

[2] I. Khamlich, K. Zeng, G. Flamant, J. Baeyens, C. Zou, J. Li, X. Yang, X. He, Q. Liu, H. Yang, Q. Yang, and H. Chen, "Technical and economic assessment of thermal energy storage in concentrated solar power plants within a spot electricity market," *Renew. Sustain. Energy Rev.*, vol. 139, Apr. 2021, Art. no. 110583.

[3] Y. He, P. Liu, L. Zhou, Y. Zhang, and Y. Liu, "Competitive model of pumped storage power plants participating in electricity spot market—In case of China," *Renew. Energy*, vol. 173, pp. 164–176, Aug. 2021.

[4] C. Ramos and C.-C. Liu, "AI in power systems and energy markets," *IEEE Intell. Syst.*, vol. 26, no. 2, pp. 5–8, Mar. 2011.

[5] W. Xiong, H.-Y. Liu, Y.-L. Jiang, P.-F. Li, D.-N. Liu, and F. Peng, "Bidding strategy of power generation enterprises in hunan medium and long-term trading market," in *Proc. 2nd IEEE Conf. Energy Internet Energy Syst. Integr. (EI)*, Oct. 2018, pp. 1–9.

[6] A. C. Tellidou and A. G. Bakirtzis, "Agent-based analysis of capacity withholding and tacit collusion in electricity markets," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1735–1742, Nov. 2007.

[7] F. A. Campos, A. M. S. Roque, E. F. Sanchez-Ubeda, and J. P. Gonzalez, "Strategic bidding in secondary reserve markets," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2847–2856, Jul. 2016.

[8] G. B. Shrestha and S. Qiao, "Market oriented bidding strategy considering market price dynamics and generation cost characteristics," *IET Gener. Transmiss. Distrib.*, vol. 4, no. 2, pp. 150–161, 2010.

[9] M. P. Muñoz, C. Corchero, and F.-J. Heredia, "Improving electricity market price forecasting with factor models for the optimal generation bid," *Int. Stat. Rev.*, vol. 81, no. 2, pp. 289–306, Aug. 2013.

[10] V. Nanduri and T. K. Das, "A reinforcement learning model to assess market power under auction-based energy pricing," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 85–95, Feb. 2007.

[11] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[12] L. Xiaotong, L. Yimei, Z. Xiaoli, and Z. Ming, "Generation and transmission expansion planning based on game theory in power engineering," *Syst. Eng. Procedia*, vol. 4, pp. 79–86, Jan. 2012.

[13] Y. Ze and P. Junda, "Research on A-J effect and strategy of power generation enterprise investment under incomplete information dynamic game," *J. Changsha Univ. Technol. (Social Sci. Ed.)*, vol. 26, no. 3, pp. 55–60, 2011.

[14] A. Tiguercha, A. A. Ladjici, and M. Boudour, "Competitive co-evolutionary approach to stochastic modeling in deregulated electricity market," in *Proc. IEEE Int. Energy Conf. (ENERGYCON)*, May 2014, pp. 514–519.

[15] A. Tiguercha, A. A. Ladjici, and M. Boudour, "Suppliers' optimal biding strategies in day-ahead electricity market using competitive coevolutionary algorithms," in *Proc. 3rd Int. Conf. Syst. Control*, Oct. 2013, pp. 821–826.

[16] J. V. Kumar, D. M. V. Kumar, and K. Edukondalu, "Strategic bidding using fuzzy adaptive gravitational search algorithm in a pool based electricity market," *Appl. Soft Comput.*, vol. 13, no. 5, pp. 2445–2455, May 2013.

[17] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.

[18] A. L. C. Bazzan, "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control," *Auton. Agents Multi-Agent Syst.*, vol. 18, no. 3, p. 342, Sep. 2008.

[19] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to reinforcement learning," in *Deep Reinforcement Learning Fundamentals, Research and Applications: Fundamentals, Research and Applications*. 2020.

[20] M. Rahimiyan and H. R. Mashhadi, "Supplier's optimal bidding strategy in electricity pay-as-bid auction: Comparison of the Q-learning and a model-based approach," *Electric Power Syst. Res.*, vol. 78, no. 1, pp. 165–175, Jan. 2008.

[21] M. Liu, L. Jiaxing, Z. Hu, J. Liu, and X. Nie, "A dynamic bidding strategy based on model-free reinforcement learning in display advertising," *IEEE Access*, vol. 8, pp. 213587–213601, 2020.

[22] W.-Y. Shih, Y.-S. Lu, H.-P. Tsai, and J.-L. Huang, "An expected win rate-based real time bidding strategy for branding campaign by the model-free reinforcement learning model," *IEEE Access*, vol. 8, pp. 151952–151967, 2020.

[23] D. Xiao and H. Chen, "Stochastic up to congestion bidding strategy in the nodal electricity markets considering risk management," *IEEE Access*, vol. 8, pp. 202428–202438, 2020.

[24] Q. Yu, Y. Liu, D. Xia, and L. Martinez, "The strategy evolution in double auction based on the experience-weighted attraction learning model," *IEEE Access*, vol. 7, pp. 16730–16738, 2019.

[25] D. Chen, Z. Jing, and H. Tan, "Optimal siting and sizing of used battery energy storage based on accelerating benders decomposition," *IEEE Access*, vol. 7, pp. 42993–43003, 2019.

[26] W. Tang and H. T. Yang, "Optimal operation and bidding strategy of a virtual power plant integrated with energy storage systems and elasticity demand response," *IEEE Access*, vol. 7, pp. 79798–79809, 2019.

[27] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents," *IEEE Trans. Autom. Control*, early access, Jan. 5, 2021, doi: 10.1109/TAC.2021.3049345.

[28] A. Shahmohammadi, R. Sioshansi, A. J. Conejo, and S. Afsharnia, "Market equilibria and interactions between strategic generation, wind, and storage," *Appl. Energy*, vol. 220, pp. 876–892, Jun. 2018.

[29] Y. Ye, D. Qiu, J. Li, and G. Strbac, "Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 130515–130529, 2019.

[30] L. Cheng, J. Zhang, L. Yin, Y. Chen, J. Wang, G. Liu, X. Wang, and D. Zhang, "General three-population multi-strategy evolutionary games for long-term on-grid bidding of generation-side electricity market," *IEEE Access*, vol. 9, pp. 5177–5198, 2020.

[31] X. Gan, H. Guo, and Z. Li, "A new multi-agent reinforcement learning method based on evolving dynamic correlation matrix," *IEEE Access*, vol. 7, pp. 162127–162138, 2019.

[32] Y. J. Park, Y. J. Lee, and S. B. Kim, "Cooperative multi-agent reinforcement learning with approximate model learning," *IEEE Access*, vol. 8, pp. 125389–125400, 2020.

[33] M.-L. Li, S. F. Chen, and J. Chen, "Adaptive learning: A new decentralized reinforcement learning approach for cooperative multiagent systems," *IEEE Access*, vol. 8, pp. 99404–99421, 2020.

[34] S. Y. Luis, D. G. Reina, and S. Marin, "A multiagent deep reinforcement learning approach for path planning in autonomous surface vehicles: The Ypacaraí lake patrolling case," *IEEE Access*, vol. 9, pp. 17084–17099, 2021.

DUNNAN LIU received the B.E. and Ph.D. degrees in electrical engineering from Tsinghua University, China. He is currently an Associate Professor with the School of Economics and Management, North China Electric Power University (NCEPU), China. His research interests include risk management and operation of electricity market.

YUAN GAO received the bachelor's degree, in 2019. He is currently pursuing the master's degree with the School of Industrial Engineering, NCEPU. His research interests include electricity market and demand side management.

WEIYE WANG received the bachelor's degree, in 2019. He is currently pursuing the master's degree with the School of Economics and Management, NCEPU. His research interest includes electricity market.

ZHIXIN DONG received the bachelor's degree, in 2020. He is currently pursuing the master's degree with the School of Economics and Management, NCEPU. His research interest includes electricity market.