A Reinforcement-learning-based Bidding Strategy for Power Suppliers with Limited Information

Qiangang Jia, Yiyan Li, Zheng Yan, Chengke Xu, and Sijie Chen

***Abstract*—The power market is a typical imperfectly competitive market where power suppliers gain higher profits through strategic bidding behaviors. Most existing studies assume that a power supplier is accessible to the sufficient market information to derive an optimal bidding strategy. However, this assumption may not be true in reality, particularly when a power market is newly launched. To help power suppliers bid with the limited information, a modified continuous action reinforcement learning automata algorithm is proposed. This algorithm introduces the discretization and Dyna structure into continuous action reinforcement learning automata algorithm for easy implementation in a repeated game. Simulation results verify the effectiveness of the proposed learning algorithm.**

***Index Terms*—Power market, bidding strategy, limited information, repeated game, continuous action reinforcement learning automata.**

I. Introduction

ELECTRICITY market reforms are gradually occurring around the world, particularly in China. A electricity market typically includes power suppliers, independent system operators (ISOs), and power consumers. Power suppliers bid in the market to satisfy the electricity demand of power consumers, while an ISO is responsible for the operation and maintenance of the market. Due to the limited number of power suppliers, the power supply-side market is typically considered as an oligopoly market, where the profit of one supplier will be affected by both the power system operation condition and the bidding actions of the other suppliers. Thus, all suppliers are incentivized to bid strategically [1] to increase their profits.

The most common approach to developing supplier bidding strategies is to establish a game-theoretical model [2], [3]. The most widely used game-theoretical methods are based on the Karush-Kuhn-Tucker (KKT) conditions. It formulates the problem as an equilibrium problem with equilibrium constraints (EPEC) [4]. To build a game-theoretical model, one supplier should have a global view of the system and its opponents, such as the locational marginal prices (LMPs) of the other nodes, the bidding actions and the cost functions of its opponents. For one supplier, this information is defined as its "external information". However, the external information available to a power supplier is often limited, particularly in emerging markets (e.g., a power spot market has just launched, and there is less historical information about the bids of members in the market), making analytical methods impractical.

Under such circumstances, the reinforcement learning [5] becomes a powerful tool for the power suppliers to optimize the bids. Reinforcement learning is a field of machine learning that emphasizes how to act based on the feedback from the environment to maximize expected benefits. Although the convergence of certain reinforcement learning algorithms may not be proven theoretically [6], they also achieve successful applications in engineering [7]-[9] due to their ability to explore with limited information. Many reinforcement learning methods have been studied in market bidding. Reference [10] uses Q learning to assist power suppliers in strategical bidding to gain higher profits. Reference [11] proposes a Markov reinforcement learning approach for multi-agent bidding in an electricity market. Reference [12] forms a stochastic game to model market bidding and proposes a reinforcement learning solution. Recently, deep-learning-based algorithms have also emerged. Reference [13] proposes a deep reinforcement learning method combined with a prioritized experience replay strategy to optimize supplier bids. Reference [14] uses the deep reinforcement learning algorithm for optimal bidding and pricing policies. Reference [15] proposes a deep reinforcement learning algorithm to help wind power companies formulate bidding strategies in energy markets and capacity markets jointly. Reference [16] uses a value function approximation method to obtain the optimal bidding strategy in the power market.

However, existing studies typically formulate the supplier bidding and market clearing process as a Markov (stochastic) game [17], which is questionable. Briefly, in this process, power suppliers and consumers submit their bids, and the ISO solves an economic dispatch problem to calculate the dispatched power generation quantities, LMP, etc., completing a round of the market clearing process. A Markov game assumes that the current state of a system is associated with both its past state and the joint actions of all players.

Manuscript received: July 20, 2020; revised: December 1, 2020; accepted: May 19, 2021. Date of CrossCheck: May 19, 2021. Date of online publication: November 24, 2021.

This work was supported by the National Natural Science Foundation of China (No. U1866206).

Q. Jia, Z. Yan, C. Xu, and S. Chen (corresponding author) are all with Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jiaqiangang@sjtu.edu.cn; yanz@sjtu.edu.cn; xck032599@sjtu.edu.cn; sijie.chen@sjtu.edu.cn).

Y. Li is with the Department of Electrical and Computer Science, North Carolina State University, Raleigh 27695, USA (e-mail: yli257@ncsu.edu).

DOI: 10.35833/MPCE.2020.000495

This assumption is developed in certain scenarios. If renewable power plants constitute the majority of power plants in the system and the ramping ability of the thermal units is insufficient, the current state of the market (e.g., LMPs) is typically related to past LMPs. This scenario is more suitable for the Markov game due to the strong relationship between adjacent time slices. In this paper, thermal generators are focused on and their ramping ability is assumed to be sufficient. Current LMPs are related only to the current bidding actions of all power suppliers and consumers but not to past LMPs. This scenario is more suitable for repeated games due to the weak relationship between adjacent time slices.

Additionally, most studies still assume that suppliers can obtain their rivals' historical bidding information. This assumption may not apply, particularly in the early stages of a market, where the power supplier only has access to its own historical bidding information. Few studies have discussed the case where suppliers have to bid with little external information. Besides, the efficiency of the algorithms has not been paid enough attention, leading to inefficient learning process.

The contributions of this paper are outlined below:

1) This paper defines the bidding procedure of power suppliers with thermal power units as a repeated game [18] rather than the widely used Markov games, avoiding the stringent requirements on state transition.

2) This paper proposes a modified continuous action reinforcement learning automata (M-CARLA) algorithm to enable power suppliers to bid with limited information in the repeated game. This algorithm combines the discretization and Dyna structure [19], making it more applicable and efficient.

The remainder of this paper is organized as follows. Section II presents the market structure and the repeated game. Section III details the proposed M-CARLA algorithm. A case study is performed in Section IV. Section V concludes the paper.

II. Market Structure and Repeated Game

A. Market Structure

A power market typically includes three major parts: the power suppliers, the power consumers, and the market operator.

1) Power Suppliers

The supply function model [20] is a typical mathematical model that describes the bidding behaviors of a power supplier with thermal generation units. In this paper, the thermal generation units all have the flexible ramping ability. The cost function of power supplier i is given as:

$$C_i = \frac{1}{2} a_i g_i^2 + b_i g_i \tag{1}$$

where i is the index of the power supplier; C_i is the cost function; a_i and b_i are the coefficients of the secondary and primary terms, respectively; and g_i is the dispatched power output.

Before each round of market clearing, supplier i submits the cost function to the market operator. The power supply-side market is imperfectly competitive, motivating power supplier i to bid strategically to gain a higher profit. The strategic factor can be the slope or the intercept of the supply function and it is assigned as the slope in this paper.

Based on this assumption, the submitted cost function will become:

$$C_{i,\text{submit}} = \frac{1}{2} a_{i,\text{strategic}} g_i^2 + b_i g_i \tag{2}$$

where $C_{i,\text{submit}}$ is the submitted cost function; and $a_{i,\text{strategic}}$ is the strategic slope.

After each round of market clearing, supplier i obtains the dispatched power output g_i and LMP λ_i of the node where it is located.

The objective of supplier i is to maximize its profit q_i:

$$q_i = \lambda_i g_i - C_i \tag{3}$$

2) Power Consumers

The utility function [21] of consumer j can still be written in a quadratic form as:

$$U_j = c_j l_j - \frac{1}{2} d_j l_j^2 \tag{4}$$

where j is the index of the power consumer; U_j is the utility function; c_j and d_j are the coefficients of the primary and secondary terms, respectively; and l_j is the load demand.

Before each round of market clearing, consumer j submits the true utility function to the market operator.

After each round of market clearing, consumer j obtains the dispatched power demand l_j and the LMP λ_j of the node where it is located.

3) Market Operator

The market operator gathers the bids of all power suppliers and consumers and then runs the economic dispatch algorithm. The objective function is:

$$\min\left(\sum_{i\in I} C_{i,\text{submit}} - \sum_{j\in J} U_j\right) \tag{5}$$

where I is the set of suppliers; and J is the set of consumers.

The objective is to maximize social welfare. The equality constraint of the optimization problem is the balance of power generation and consumption:

$$\sum_{i\in I} g_i = \sum_{j\in J} l_j \tag{6}$$

The inequality constraints include the power flow constraints of transmission lines, the generation limits of suppliers, and the demand limits of consumers:

$$-F_{y,\max} \le F_y \le F_{y,\max} \quad y \in Y \tag{7}$$

$$g_{i,\min} \le g_i \le g_{i,\max} \quad i \in I \tag{8}$$

$$l_{j,\min} \le l_j \le l_{j,\max} \quad j \in J \tag{9}$$

where F_y is the power flow of the transmission line y; $F_{y,\max}$ is the upper limit of F_y; Y is the set of transmission lines; $g_{i,\min}$ and $g_{i,\max}$ are the lower and upper limits of the power output of suppliers, respectively; and $l_{j,\min}$ and $l_{j,\max}$ are the lower and upper limits of the power demand, respectively.

The power flow of each transmission line can be calculated based on [22]:

$$\boldsymbol{F} = \boldsymbol{T}(\boldsymbol{G} - \boldsymbol{L}) \tag{10}$$

where $\boldsymbol{F}$ is the power flow matrix; $\boldsymbol{T}$ is the power transfer distribution factor (PTDF) matrix; and $\boldsymbol{G}$ and $\boldsymbol{L}$ are the power output and the load consumption matrices, respectively.

B. Repeated Game

This paper focuses on optimizing a single time period bidding strategy in a real-time market (RTM). In Fig. 1, the power market has a bi-level framework. Assume that the consumers bid based on their real utility functions, and suppliers bid strategically to increase profits. After the market is cleared, the power output of each supplier, the load demand of each consumer, and the LMP of each node can be calculated and returned to the corresponding market participants. This type of gaming process is called a "stage game". A "repeated game" is a game in which the same "stage game" is played repeatedly over several discrete periods. During repeated interactions with the market, suppliers can gradually understand the market and derive their optimal bidding strategies.

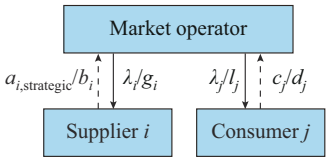

Fig. 1. Market structure based on supply function model.

The market in a repeated game can be defined as stationary or nonstationary to a single strategic supplier. If other suppliers bid their true marginal cost function, the environment is stationary; if other suppliers also bid strategically, the environment is nonstationary.

III. M-CARLA Algorithm

The continuous action reinforcement learning automata (CARLA) algorithm [23] is considered useful due to its low requirements for external information. The CARLA algorithm uses a nonparametric probabilistic model to update the probability density function (PDF) over the action space. The core part of the algorithm is to reinforce the probability of better actions being chosen through a Gaussian neighborhood function by interacting with the environment. After several interactions, a stable action PDF centered around the optimal actions is obtained.

However, this algorithm is difficult to use due to large symbolic and integration operations in continuous action space. As iterations continue, computation costs are high, and calculations may be unsolvable [24]. To use this algorithm, a discretization method is proposed to modify the CARLA algorithm, making it more computationally tractable. Additionally, to accelerate the learning process, a virtual experience generation process is introduced. Compared with the original CARLA algorithm, the proposed algorithm only discretizes the PDF and introduces a virtual experience generation process. Therefore, the convergence of the proposed M-CARLA algorithm is in line with the CARLA algorithm, which has been proven in [23].

The M-CARLA algorithm contains four steps that repeat from *Step 2* through *Step 4*.

Step 1: initialize the PDF.

The bidding action (i.e., the strategic slope) and the action PDF at the n^{th} iteration of the supplier are denoted by $a(n)$ and $f(a,n)$, respectively.

Because suppliers have little prior knowledge about the market, $f(a,0)$ will be initialized as a uniform distribution:

$$f(a,0)\sim U(a_{\min},a_{\max}) \tag{11}$$

where $a_{\max}$ and $a_{\min}$ are the upper and lower limits of the slope a, respectively.

Step 2: choose actions.

The action space is divided equally into x equal subintervals with the endpoints as $\{a_0, a_1, ..., a_x\}$, where each segment length is d. The continuous PDF is then replaced by discrete values at different endpoints. At the n^{th} iteration, the discrete PDF is represented by the set $\{f(a_0,n), f(a_1,n), ..., f(a_x,n)\}$.

Based on the trapezoidal rule [25], the area of subinterval m can be written as:

$$s_m(n)=\frac{d}{2}\left(f(a_{m-1},n)+f(a_m,n)\right) \tag{12}$$

After calculating the areas of all subintervals, the cumulative probability of the action at endpoint m can be calculated by:

$$S_m(n)=\sum_{u=1}^{m}s_u(n) \tag{13}$$

Before an action is selected, a random variable $z(n)$ is generated from the uniform distribution over [0,1]. The subinterval t is determined according to $z(n)$ based on the cumulative probability, then $a(n)$ can be written as:

$$a(n)=a_{t-1}+2\frac{z(n)-S_{t-1}}{f(a_{t-1},n)+f(a_t,n)} \tag{14}$$

This process preserves the continuity of the selected action, which is different from the finite action learning automata (FALA) algorithm [26], [27].

The following example in Fig. 2 shows the "choose actions" process. It is hard to select an action value $a(n)$ in the continuous action PDF as Fig. 2(a). In discretized action PDF in Fig. 2(b), $a(n)$ can be selected easily if the left shaded area of $a(n)$ is $z(n)$.

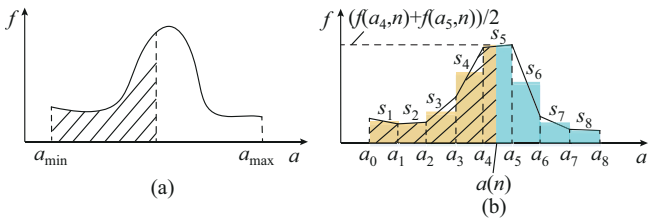

Fig. 2. "Choose actions" process. (a) Continuous action PDF. (b) Discretized action PDF.

Assume that the action space is divided into 8 subintervals. The continuous action PDF in Fig. 2(a) can then be discretized, as shown in Fig. 2(b). The area of each interval can be calculated by (12) as $\{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$, as indicated by the rectangles in Fig. 2(b). The cumulative probability

can also be calculated by (13) as $\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8\}$. If the random variable $z(n)$ (the area of the orange-rectangular parts in Fig. 2(b)) is between S_4 and S_5, action $a(n)$ can be chosen by:

$$a(n)=a_4+2\frac{z(n)-S_4}{f(a_4,n)+f(a_5,n)} \tag{15}$$

Step 3: generate reinforcement signals.

After the market is cleared in the current round, the power supplier will calculate the profit $q(n)$ based on (3) to obtain the real experience $(a(n), q(n))$. Then, the strategic supplier will evaluate the reinforcement signal $\beta(n)$ as:

$$\beta(n)=\max\left\{0,\frac{q(n)-q_{\text{med}}}{q_{\max}-q_{\text{med}}}\right\} \tag{16}$$

where $q_{\max}$ and q_{med} are the maximum and the median values in data buffer 1, respectively.

Data buffer 1 provides the historical profit data for evaluating the reinforcement signal; the initial value in data buffer 1 is 0. A larger $\beta(n)$ indicates a stronger reward signal, while a smaller $\beta(n)$ indicates a stronger punishment signal. The supplier saves $q(n)$ into data buffer 1 after this evaluation. To avoid storage overflow, only the latest L rounds of $q(n)$ are saved.

However, solely relying on interactions with the real world is sometimes inefficient. Inspired by the Dyna structure [28], the historical experience may be able to provide more guidance for learning by generating virtual experience. The Dyna structure is further explained in Appendix A.

To generate a virtual experience at the n^{th} iteration, a new data buffer (data buffer 2) $\{(a_1, q_1), ..., (a_w, q_w), ..., (a_W, q_W)\}$ is introduced to save the latest W real historical action-profit pairs from the real market environment. A virtual action from data buffer 2 is chosen when the length of data buffer 2 is above E (the threshold of data buffer 2):

$$a_{\text{v}}(n)=a_{\text{rand}}+\theta \tag{17}$$

where $a_{\text{v}}(n)$ is the virtual action at the n^{th} iteration; a_{rand} is the historical action randomly selected from data buffer 2; and θ is the additive random white Gaussian noise.

The mapping from the virtual action to the corresponding profit is a regression problem. The K-nearest neighbor (KNN) method [29] is chosen as the regression tool due to its low requirements on prior knowledge. K nearest neighbors of $a_{\text{v}}(n)$ in data buffer 2 are chosen to formulate a new data set $\{(a_1, q_1), ..., (a_k, q_k), ..., (a_K, q_K)\}$. The distance between $a_{\text{v}}(n)$ and its neighbors is measured by Euclid distance.

The corresponding virtual profit $q_{\text{v}}(n)$ can be generated as:

$$q_{\text{v}}(n)=\frac{1}{K}\sum_{k=1}^{K}q_k \tag{18}$$

Then, the virtual reinforcement signal $\beta_{\text{v}}(n)$ at the n^{th} iteration can be calculated based on data buffer 1 by (16).

Step 4: update the PDF.

Two Gaussian neighborhood functions $h_1(n)$ and $h_2(n)$ are defined at the n^{th} iteration as (19) and (20). Then they are discretized as $\{h_1(a_0, n), h_1(a_1, n), ..., h_1(a_x, n)\}$ and $\{h_2(a_0, n), h_2(a_1, n), ..., h_2(a_x, n)\}$ as the update signal:

$$h_1(n)=\eta\exp\left(-\frac{(a-a(n))^2}{2\sigma^2}\right) \tag{19}$$

$$h_2(n)=\eta\exp\left(-\frac{(a-a_{\text{v}}(n))^2}{2\sigma^2}\right) \tag{20}$$

where η and σ are the height and width of the update signal, respectively.

At the n^{th} iteration, the update of the action PDF can be expressed as the linear combination of the discrete old action PDF and the discrete Gaussian neighborhood function:

$$\begin{aligned}f(a_e,n+1)=&(1-\delta)\alpha(f(a_e,n)+\beta(n)h_1(a_e,n))+\\&\delta\alpha_{\text{v}}(f(a_e,n)+\beta_{\text{v}}(n)h_2(a_e,n))\quad e=0,1,...,x\end{aligned} \tag{21}$$

where δ is a weight factor to describe the importance between the real and virtual experiences. δ is larger in a stationary environment compared with a nonstationary environment since the usable experience is changing all the time as learning progresses in the nonstationary environment. α and α_{v} can be calculated based on the composite trapezoidal rule [30], guaranteeing that the integration of the PDF is 1 over the action space.

An example shows the "update PDF" process, as shown in Fig. 3 (if δ is 0).

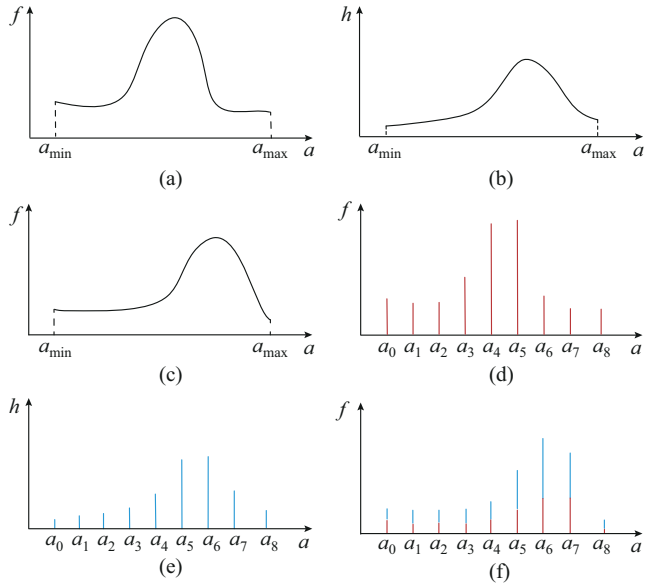

Fig. 3. "Update PDF" process. (a) Continuous old action PDF. (b) Continuous update signal. (c) Continuous new action PDF. (d) Discretized old action PDF. (e) Discretized update signal. (f) Discretized new action PDF.

Assume that the action space remains divided into 8 subintervals. The modification transforms the symbolic operation Fig. 3(a) - (c) to the linear operation of the discrete values Fig. 3(d) - (f), which reduces the complexity significantly. Note that the two different colors in Fig. 3(f) represent old PDF and update signal.

IV. Case Study

Simulations are run in MATLAB R2020a. The primary objective lies in validating the effectiveness of the M-CARLA algorithm.

The topology of the 8-bus testing system is based on [31], and some modifications are added. The topology of the system is shown in Fig. 4. The parameters of the system are given in Table I.

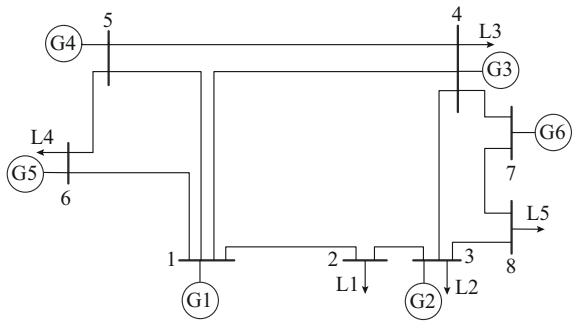

Fig. 4. Topology of 8-bus system.

TABLE I
PARAMETERS OF 8-BUS SYSTEM

Line No.	Start node	End node	Reactance (p.u.)	Limit (MW)
1	1	2	0.1	1000
2	1	4	0.1	1000
3	1	5	0.1	100
4	1	6	0.1	1000
5	2	3	0.1	1000
6	3	4	0.1	1000
7	3	8	0.1	100
8	4	5	0.1	1000
9	4	7	0.1	1000
10	5	6	0.1	100
11	7	8	0.1	1000

There are 6 power suppliers in buses 1, 3, 4, 5, 6, and 7. The parameters of all power suppliers are shown in Table II.

TABLE II
PARAMETERS OF POWER SUPPLIERS IN 8-BUS SYSTEM

Supplier	a (\$/MW^{2}h)	b (\$/MWh)	g_{min} (MW)	g_{max} (MW)
1	0.030	1	0	1000
2	0.020	10	0	1000
3	0.025	5	0	1000
4	0.035	10	0	1000
5	0.020	20	0	1000
6	0.015	4	0	1000

There are 5 power consumers in buses 2, 3, 4, 6, and 8. The parameters of power consumers in 8-bus system are shown in Table III.

TABLE III
PARAMETERS OF POWER CONSUMERS IN 8-BUS SYSTEM

Consumer	c (\$/MW^{2}h)	d (\$/MWh)	l_{min} (MW)	l_{max}(MW)
1, 2, 3, 4, 5	100	0.06	0	500

The DC power flow model is used, and the reactance of each transmission line is set to be 0.1 p.u.. The capacities of transmission lines 3, 7, and 10 are set as 100 MW to cause congestion.

To better describe the superiority of the M-CARLA algorithm, the existing algorithms in the repeated game environment are compared. The results of the qualitative analysis are shown in Table IV.

TABLE IV
COMPARISON BETWEEN EXISTING ALGORITHMS AND PROPOSED ALGORITHM

Algorithm	Information requirement
Reference [32]	System parameters, historical bids of opponents, market clearing model
References [31], [33]	System parameters, opponents' estimated bids
Reference [20]	Load level, opponents' estimated bids
Proposed algorithm	Self-historical bids

From this comparison, it can be found that the proposed algorithm has much lower information requirements. Therefore, the proposed algorithm is more suitable for use within limited-information environments.

The Nash equilibrium calculated by analytical methods [4] in the complete information environment is taken as a reference to evaluate the learning results of the proposed algorithm, where the action resolution is 0.1 \$/MW^{2}h. The analytical Nash equilibrium is shown in Table V.

TABLE V
NASH EQUILIBRIUM

Supplier	Profit (\$/h)	Action value (\$/MW^{2}h)	LMP (\$/MWh)
1	20015	0.04	36.8
2	21425	0.05	21.6
3	11684	0.04	31.1
4	537	0.09	17.7
5	10487	0.16	69.2
6	19882	0.09	48.1

To provide a numerical index to evaluate its effectiveness, the accuracy A between the learning solution S_L and the analytical solution S_A is defined as:

$$A = 1 - \frac{\left| S_L - S_A \right|}{S_A} \times 100\% \tag{22}$$

Because the action is chosen based on the action PDF, randomness is inevitable. To eliminate random factors, the same simulation is run 10 times to take an average in both stationary and nonstationary environments.

The learning parameters of all suppliers are shown in Table VI, where M is the iteration threshold.

1) Stationary Environment

A stationary environment indicates that except for the strategic supplier (the learner using the M-CARLA algorithm), the others are assumed to use the fixed strategies.

Six scenarios are investigated: each supplier is chosen as the learner in turn, and when a supplier is chosen, others fix their strategies as the Nash equilibrium. δ is set to be 0.3 in

these stationary environments.

TABLE VI
LEARNING PARAMETERS OF ALL SUPPLIERS

Supplier	M	W	E	K	L	η	σ	$a_{\min}$	$a_{\max}$	x
1	600	30	3	3	10	0.1	0.002	0	0.2	200
2	600	30	3	3	10	0.1	0.002	0	0.2	200
3	600	30	3	3	10	0.1	0.002	0	0.2	200
4	600	30	3	3	10	0.1	0.003	0	0.3	300
5	600	30	3	3	10	0.1	0.005	0	0.5	500
6	600	30	3	3	10	0.1	0.003	0	0.3	300

The learning results of power suppliers in different scenarios in stationary environment are shown in Table VII.

TABLE VII
LEARNING RESULTS OF POWER SUPPLIERS IN DIFFERENT SCENARIOS IN STATIONARY ENVIRONMENT

Scenario	Supplier	Profit ($/h)	Action value ($/MW^{2}h)	LMP ($/MWh)	Accuracy of actions (%)
1	1	20146	0.042	37.2	95.0
2	2	21433	0.052	21.8	96.0
3	3	11682	0.039	30.9	97.5
4	4	537	0.088	17.7	97.8
5	5	10547	0.220	69.8	100.0
6	6	19947	0.088	47.9	97.8

The performance bound of the proposed algorithm in the stationary environment is 95%-100%. The learning process becomes stable after 100-200 iterations. The bid curves of the different power suppliers in the stationary environment are shown in Fig. 5.

2) Nonstationary Environment

A nonstationary environment indicates that all suppliers use the M-CARLA algorithm to bid. δ is set to be 0.1 in this nonstationary environment.

The learning results of all suppliers in the nonstationary environment are shown in Table VIII.

The performance bound of the proposed algorithm in the nonstationary environment is 90.0%-97.8%. The learning process becomes stable after 200-300 iterations. The bid curves of the different power suppliers in the nonstationary environment are shown in Fig. 6.

The accuracy of actions and the learning efficiency in the stationary environment are higher than those in the nonstationary environment since the nonstationary environment introduces more randomness and uncertainty in the learning process. The computational complexity of this algorithm is described in Appendix B.

The parameters of the demand curves in this case study are constant. If the load fluctuates, a day can be divided into different periods with given load levels. The gaming process of the same period on different days can be considered to be a repeated game. The M-CARLA algorithm can be used in different repeated games to optimize the bidding strategy.

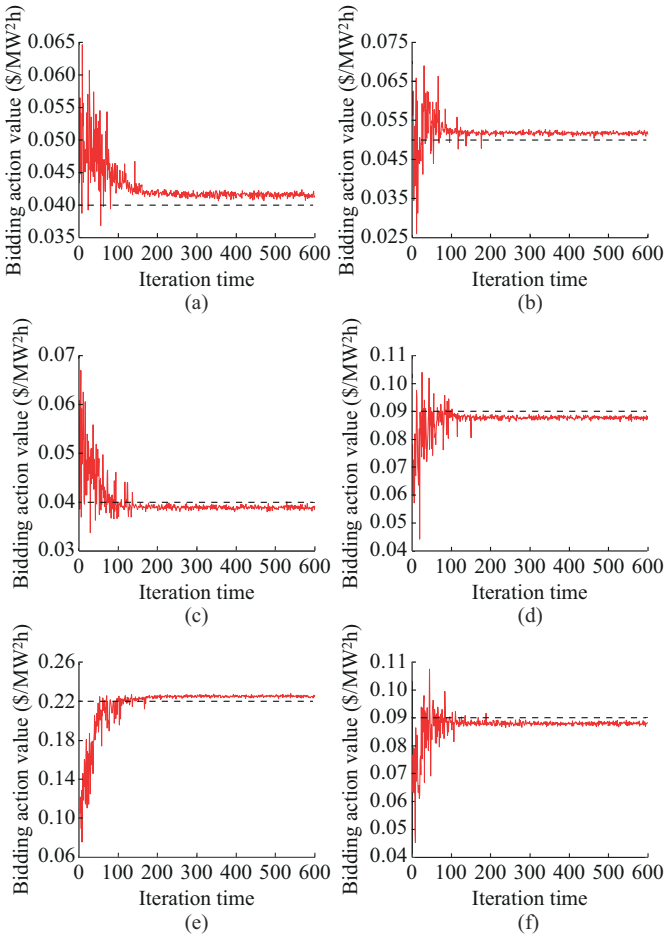

Fig. 5. Bid curves of different power suppliers in stationary environment. (a) Power supplier 1. (b) Power supplier 2. (c) Power supplier 3. (d) Power supplier 4. (e) Power supplier 5. (f) Power supplier 6.

TABLE VIII
LEARNING RESULTS OF POWER SUPPLIERS IN NONSTATIONARY ENVIRONMENT

Supplier	Profit ($/h)	Action value ($/MW^{2}h)	LMP ($/MWh)	Accuracy of actions (%)
1	19661	0.044	37.3	90.0
2	24238	0.055	22.6	90.0
3	12553	0.039	31.8	97.5
4	718	0.093	19.0	96.7
5	11131	0.200	68.4	90.0
6	19942	0.088	47.9	97.8

V. CONCLUSION

This paper proposes a practical bidding strategy for power suppliers with limited information. Firstly, the modeling method of the gaming process is proposed. The gaming process of thermal power suppliers that can provide flexible ramping is modeled as a repeated game based on the supply function model. Then, an M-CARLA algorithm is proposed to enable suppliers to bid based on only personal data. Finally, the proposed algorithm is tested in an 8-bus system to demonstrate its effectiveness in both stationary and nonstationary environments.

However, there are still certain limitations in this study:

the virtual experience is not always reliable in a nonstationary environment, and the scalability of the proposed algorithm in a more complex and variable environment must be further validated. In future work, we plan to focus on how to use the historical experience to accelerate learning in a nonstationary environment and extend the algorithm to manage a fluctuating load profile.

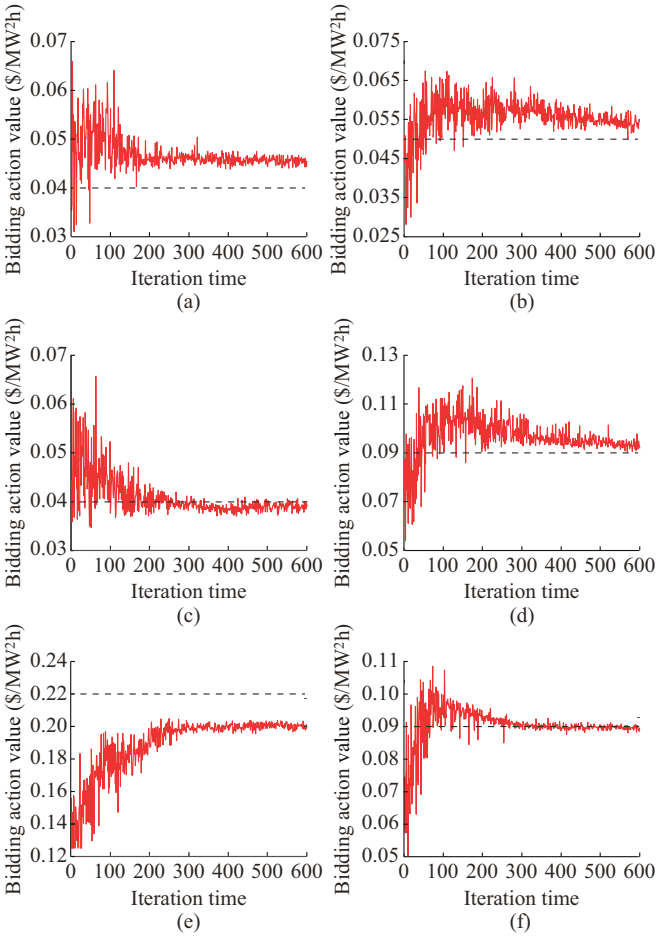

Fig. 6. Bid curves of power suppliers in nonstationary environment. (a) Power supplier 1. (b) Power supplier 2. (c) Power supplier 3. (d) Power supplier 4. (e) Power supplier 5. (f) Power supplier 6.

APPENDIX A

The Dyna structure combines model-free learning with a virtual model. The virtual model in the Dyna structure can generate virtual experiences to feed model-free learning. The general form of the Dyna structure is shown in Fig. A1.

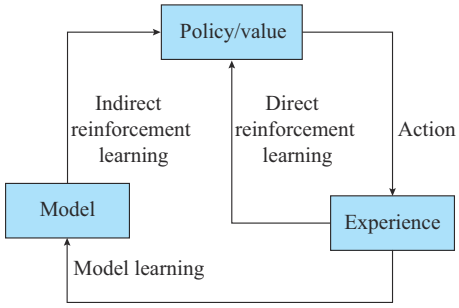

Fig. A1. General form of Dyna structure.

The Dyna structure is extended from reinforcement learning and includes policy learning and model learning. In the interaction process, the structure integrates real experiences and virtual experiences. The real experiences are for learning policy/value, i.e., direct reinforcement learning (RL), and for learning the model concurrently. The simulated experiences produced by the model can be used to update the policy, i.e., indirect RL.

APPENDIX B

The number of subintervals markedly affects the computational complexity of the algorithm. Therefore, the computational complexity is analyzed from the perspective of a single supplier with different subinterval magnitudes, as shown in Table BI. The time is recorded which takes to update the action PDF and select an action in each round. To eliminate interference from other power suppliers, the environment is assumed to be stationary. All simulations are run on a computer with an Intel Core i7™ CPU and 16 GB RAM.

TABLE BI
COMPUTATIONAL COMPLEXITY ANALYSIS

x	Time (s)
10	0.00006
10^2	0.00024
10^3	0.00139

The computational complexity increases linearly as the magnitude of the subintervals increases.

REFERENCES

[1] M. Mallaki, M. S. Naderi, M. Abedi *et al.*, "Strategic bidding in distribution network electricity market focusing on competition modeling and uncertainties," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 3, pp. 561-572, May 2021.
[2] B. F. Hobbs, C. B. Metzler, and J. Pang, "Strategic gaming analysis for electric power systems: an MPEC approach," *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 638-645, May 2000.
[3] Q. Jia, Y. Li, Z. Yan *et al.*, "Reactive power market design for distribution networks with high photovoltaic penetration," *IEEE Transactions on Smart Grid*, doi: 10.1109/TSG.2022.3186338
[4] M. Rayati, A. Sheikhi, A. M. Ranjbar *et al.*, "Optimal equilibrium selection of price-maker agents in performance-based regulation market," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 1, pp. 204-212, Jan. 2022.
[5] C. Huang, H. Zhang, L. Wang *et al.*, "Mixed deep reinforcement learning considering discrete-continuous hybrid action space for smart home energy management," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 743-754, May 2022.
[6] H. M. Schwartz, *Multi-agent Machine: A Reinforcement Learning Approach*. New Jersey: John Wiley & Sons, 2014.
[7] Y. Zhou, W. -J. Lee, R. Diao *et al.*, "Deep reinforcement learning based real-time ac optimal power flow considering uncertainties," *Journal of Modern Power Systems and Clean Energy*, doi: 10.35833/MPCE.2020.000885
[8] S. Wu, W. Hu, Z. Lu *et al.*, "Power system flow adjustment and sample generation based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1115-1127, Nov. 2020.
[9] D. Cao, W. Hu, X. Xu *et al.*, "Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1101-1110, Sept. 2021.

[10] N. Yu, C. C. Liu, and L. Tesfatsion, "Modeling of suppliers' learning behaviors in an electricity market environment," in *Proceedings of 2007 International Conference on Intelligent Systems Applications to Power Systems*, Toki Messe, Niigata, Nov. 2007, pp. 1-6.
[11] N. Rashedi, M. A. Tajeddini, and H. Kebriaei, "Markov game approach for multi-agent competitive bidding strategies in electricity market," *IET Generation, Transmission & Distribution*, vol. 10, no. 15, pp. 3756-3763, Nov. 2016.
[12] R. Ragupathi and T. K. Das, "A stochastic game approach for modeling wholesale energy bidding in deregulated power markets," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 849-856, May 2004.
[13] Y. Ye, D. Qiu, M. Sun *et al*., "Deep reinforcement learning for strategic bidding in electricity markets," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343-1355, Mar. 2020.
[14] H. Xu, H. Sun, D. Nikovski *et al.*, "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6366-6375, Nov. 2019.
[15] D. Cao, W. Hu, and X. Xu, "Bidding strategy for trading wind energy and purchasing reserve of wind power producer – a DRL based approach," *International Journal of Electrical Power & Energy Systems*, vol. 117, pp. 1-10, May 2020.
[16] H. K. Nunna, A. Sesetti, and A. K. Rathore, "Multiagent-based energy trading platform for energy storage systems in distribution systems with interconnected microgrids," *IEEE Transactions on Industry Applications*, vol. 56, no. 3, pp. 3207-3217, May 2020.
[17] V. Hakami and M. Dehghan, "Learning stationary correlated equilibria in constrained general-sum stochastic games," *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1640-1654, Jul. 2016.
[18] L. Li, C. Langbort, and J. Shamma, "An LP approach for solving two-player zero-sum repeated Bayesian games," *IEEE Transactions on Automatic Control*, vol. 64, no. 9, pp. 3716-3731, Sept. 2019.
[19] K. Hwang, W. Jiang, Y. Chen *et al*., "Model-based indirect learning method based on Dyna-Q architecture," in *Proceedings of 2013 IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK, Oct. 2013, pp. 2540-2544.
[20] K. Dehghanpour, M. H. Nehrir, J. W. Sheppard *et al*., "Agent-based modeling in electrical energy markets using dynamic Bayesian networks," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4744-4754, Nov. 2016.
[21] L. B. Cunningham, R. Baldick, and M. L. Baughman, "An empirical study of applied game theory: transmission constrained Cournot behavior," *IEEE Transactions on Power Systems*, vol. 17, no. 1, pp. 166-172, Feb. 2002.
[22] Y. Wang, "The calculation of nodal price based on optimal power flow," M.S. thesis, Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, 2014.
[23] M. N. Howell, G. P. Frost, T. J. Gordon *et al*., "Continuous action reinforcement learning applied to vehicle suspension control," *Mechatronics*, vol. 7, no. 3, pp. 263-276, Apr. 1997.
[24] X. Liu and N. Mao, "New continuous action-set learning automata," *Journal of Data Acquisition and Processing*, vol. 30, no. 6, pp. 1310-1317, Nov. 2015.
[25] Q. I. Rahman and G. Schmeisser, "Characterization of the speed of convergence of the trapezoidal rule," *Numerische Mathematik*, vol. 57, no. 1, pp. 123-138, Dec. 1990.
[26] T. Tao, *Learning Automata and Its Application in Stochastic Point Location Problem*. Shanghai, China: Shanghai Jiao Tong University, 2007.
[27] K. S. Narendra and M. A. *Thathachar, Learning Automata: An Introduction.* New York: Dover Publications, 1989.
[28] H. Shi, "A sample aggregation approach to experiences replay of Dyna-*Q* learning," *IEEE Access*, vol. 6, pp. 37173-37184, Apr. 2018.
[29] J. Song, J. Zhao, F. Dong *et al*., "A novel regression modeling method for PMSLM structural design optimization using a distance-weighted KNN algorithm," *IEEE Transactions on Industry Applications*, vol. 54, no. 5, pp. 4198-4206, Sept.-Oct. 2018.
[30] D. Cruz-Uribe and C. J. Neugebauer, "Sharp error bounds for the trapezoidal rule and Simpson's rule," *Journal of Inequalities in Pure and Applied Mathematics*, vol. 3, no. 4, pp. 1-22, Apr. 2002.
[31] T. Li and M. Shahidehpour, "Strategic bidding of transmission-constrained GENCOs with incomplete information," *IEEE Transactions on Power Systems*, vol. 20, no. 1, pp. 437-447, Feb. 2005.
[32] F. Wen and A. K. David, "Optimal bidding strategies and modeling of imperfect information among competitive generators," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 15-21, Feb. 2001.
[33] R. W. Ferrero, J. F. Rivera, and S. M. Shahidehpour, "Application of games with incomplete information for pricing electricity in deregulated power pools," *IEEE Transactions on Power Systems*, vol. 13, no. 1, pp. 184-189, Feb. 1998.

Qiangang Jia received the B.S. degree in electrical engineering from Shandong University, Jinan, China, in 2018. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include electricity market, machine learning, energy blockchain, and demand response.

Yiyan Li received the B.S. and Ph.D. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014 and 2019, respectively. He is currently a Postdoctoral Researcher in the Department of Electrical and Computer Science, North Carolina State University, Raleigh, USA. His research interests include machine learning and data analytics in power systems, load forecasting, photovoltaic forecasting, and distribution system operation.

Zheng Yan received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1987 and 1991, respectively. He is currently a Professor of electrical engineering with Shanghai Jiao Tong University. His research interests include application of optimization theory to power systems, power market, and dynamic security assessment.

Chengke Xu received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently pursuing the M.S. degree in the Department of Electrical Engineering, Shanghai Jiao Tong University. His research interests include power market and machine learning.

Sijie Chen received the B.E. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2009 and 2014, respectively. He was an Assistant Research Professor with the Department of Electrical Engineering and Computer Science, Washington State University, Puuman, USA, from 2014 to 2016. He is currently a Tenure Track Associate Professor of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include energy blockchain, demand response, transactive energy system, and electricity market.