Hindawi
Journal of Electrical and Computer Engineering
Volume 2019, Article ID 9406072, 16 pages
https://doi.org/10.1155/2019/9406072

Research Article

Continuous Reinforcement Algorithm and Robust Economic Dispatching-Based Spot Electricity Market Modeling considering Strategic Behaviors of Wind Power Producers and Other Participants

Zhenyu Zhao,[1,2] Shuguang Yuan,[1,2] Qingyun Nie,[1,2] and Weishang Guo[1,2]

[1]*Beijing Key Laboratory of New Energy and Low-Carbon Development (North China Electric Power University), Changping District, Beijing 102206, China*
[2]*School of Economics and Management, North China Electric Power University, Beijing 102206, China*

Correspondence should be addressed to Shuguang Yuan; shuguangyuan@ncepu.edu.cn

Received 17 June 2018; Revised 24 October 2018; Accepted 2 December 2018; Published 3 March 2019

Academic Editor: Daniele Menniti

In a spot wholesale electricity market containing strategic bidding interactions among wind power producers and other participants such as fossil generation companies and distribution companies, the randomly fluctuating natures of wind power hinders not only the modeling and simulating of the dynamic bidding process and equilibrium of the electricity market but also the effectiveness about keeping economy and reliability in market clearing (economic dispatching) corresponding to the independent system operator. Because the gradient descent continuous actor-critic algorithm is demonstrated as an effective method in dealing with Markov's decision-making problems with continuous state and action spaces and the robust economic dispatch model can optimize the permitted real-time wind power deviation intervals based on wind power producers' bidding power output, in this paper, considering bidding interactions among wind power producers and other participants, we propose a gradient descent continuous actor-critic algorithm-based hour-ahead electricity market modeling approach with the robust economic dispatch model embedded. Simulations are implemented on the IEEE 30-bus test system, which, to some extent, verifies the market operation economy and the robustness against wind power fluctuations by using our proposed modeling approach.

1. Introduction

Wind power is one of the fastest growing renewable power resources [1]. In the spot electricity market (EM) with wind power penetration, the fluctuating and random nature of this intermittent resource hinders the integration of wind power into EM and operation of power systems. Moreover, the strategic interactions among wind power producers (WPPs) and other market participants such as fossil generation companies (GenCOs) and distribution companies (DisCOs) have increased the complexity of EM modeling which is a necessary tool for market analysis, design, bidding decision-making, and every market modification [2].

The objectives of all participants bidding in EM are maximizing their own profits. Wind power and some other renewable power resources often participate in spot EM as "price takers" because of their low marginal costs. Therefore, the only bidding parameter a WPP needs to determine is its production level [3]. On the one hand, the limited predictability nature of wind power makes WPPs usually not meet the production level they bid, which increases the probability of system imbalances [4]. Relevant regulators in many countries have designed various penalty mechanisms to financially punish WPPs for their deviations of real-time productions from their bidding ones. Hence, if neglecting the marginal costs of wind power [5], maximizing a WPP's profit means minimizing the deviation cost and maximizing the bidding revenue simultaneously. On the other hand, the fluctuating and random nature of wind power makes other EM participants to bid in this stochastically fluctuating EM environment in order to maximizing their own profits, which in turn affects the bidding revenues of WPPs mainly

through locational marginal prices (LMPs) clearing by the independent system operator (ISO). Therefore, in this more complicated situation, developing fast and reliable market modeling approaches which contain bidding interactions among all kinds of participants has become considerably more important than before. One aim of this paper is to apply a new reinforcement learning algorithm based on the gradient descent continuous actor-critic (GDCAC) algorithm for solving double-side hour-ahead EM modeling containing strategic bidding interactions among WPPs and other market participants such as GenCOs and DisCOs.

Generally speaking, literatures relevant to our research can be divided into two categories: optimal wind power (or other renewable power) bidding in EM with wind penetration and EM modeling considering (or not considering) wind and some other renewable power penetrations. In the aspect of optimal wind power bidding in EM, methods for finding the optimal bidding strategy for a WPP have been introduced by many researchers. Vilim and Botterud [3] proposed two stochastic bidding models based on kernel density estimation (KDE) for a WPP to obtain the optimal day-ahead bidding strategy. Ravnaas et al. [6] proposed a seasonal autoregressive integrated moving average (SARIMA) algorithm for a WPP to obtain the optimal day-ahead bidding strategy. Sharma et al. [5] studied the behaviors of strategic WPPs in markets dominated by wind generators using the Cournot game model. In [7], Matevosyan et al. proposed an imbalance cost minimization bidding strategy for a WPP by forecasting the wind power probability distribution functions. Li and Shi [8] proposed a stochastic bidding model for a WPP based on the Roth–Erev reinforcement learning algorithm. Laia et al. [9] considered the uncertainty on the electricity price through a set of exogenous scenarios and solved the bidding problem of a price-taker thermal-wind power producer by using a stochastic mixed-integer linear programming approach. In [10], Chaves-Ávila et al. analyzed the impact of different balancing rules (penalty mechanism) on wind power short-term bidding strategies through a stochastic optimization model. Based on the Stackelberg game model, Xiao et al. [11] put forward a closed analysis on WPP's optimal bidding strategy in day-ahead EM involving large-scale wind power. Lei et al. [12] studied, using a stochastic bilevel model, the optimal bidding decision for a WPP participating in a day-ahead EM that employs stochastic market clearing and energy and reserver cooptimization, in which only the wind generation uncertainty is considered. Similar researches on the optimal bidding strategy of a WPP can also be seen in [11, 13–18].

However, authors in [3, 5–18] only studied how to find the optimal bidding strategy for a WPP within EM environment, and the modeling methods of those literatures are either static game models (Cournot and Stackelberg game models) or bilevel stochastic optimization model which cannot simulate the impact of wind power on dynamic bidding process of other participants (GenCOs and DisCOs) in a spot EM considering wind power penetration.

In order to overcome those deficiencies listed above, researches on spot EM modeling methods considering or not considering wind and some other renewable power penetration have been proposed in many literatures.

In general, the main purpose of EM modeling approaches is to regard the EM as a whole system, in which the interactions among all market participants are investigated, and the bidding process or the equilibrium result is simulated. EM modeling approaches mostly lie within twofold [2]: game-based models and agent-based models. In [2], Salehizadeh and Soltaniyan have summarized that game-based EM models are inferior to agent-based models, and the reasons are as follows: (1) some game-based models often result in a set of nonlinear equations which cannot be easily solved or might yield no solution; (2) some game-based models need to repeatedly solve the multilevel mathematical programming approaches so as to depict the dynamic bidding process in EM, while the computational complexity limits the ability to simulate large EM systems with a game-based model; and (3) almost all game-based models are based on an assumption which is to take the known probability distribution function of the market clearing price (MCP) or other competitors' bidding strategies as common knowledge, and the abovementioned assumption is not more applicable in a realistic situation [19]. Hence, many researches about the application of agent-based methods for EM modeling have been proposed recently. Rahimiyan and Rajabi Mashhadi [19] modeled and simulated the EM bidding process using the multiagent Q-learning algorithm considering discrete state and action sets and the game-based approach, respectively. Comparison of the agent-based model with the game-based model in [19] confirms the superiority of the agent-based model in this issue. Santos et al. [20] proposed an agent-based wholesale EM test bed (called MASCEM: multiagent simulator of competitive electricity markets) in which the variant Roth–Erev reinforcement learning (VRERL) algorithm was used to model the bidding behavior of the GenCOs agents. Similar researches on agent-based EM modeling can also be seen in [21–28], but none of researches in [19–28] is involved in considering wind and some other renewable power penetrations.

Shafie-khah et al. [29] proposed a multiagent EM model based on a heuristic dynamic algorithm to help analyzing the market powers of GenCOs in EM considering wind power uncertainty. Dallinger and Wietschel [30], based on an agent-based EM equilibrium model, have studied the impact of plug-in electric vehicle on EM with renewable power penetration. Reeg et al. [31] studied the policy design problem to foster the integration of renewable energy sources into EM by using an agent-based approach. Zamani-Dehkordi et al. [32] studied the impact of a proposed wind farm project on wholesale and retail electricity prices by using EM models based on nonparametric regression algorithms. In [33], by using the Q-learning algorithm, Haring et al. proposed a multiagent EM approach to analyze the effects of renewable power uncertainty on the spot EM bidding progress. Salehizadeh and Soltaniyan [2] modified the multiagent EM approach through the fuzzy Q-learning algorithm, by which the effects of renewable power uncertainty on the spot EM bidding progress was also studied within a continuous market state (wind power) space, but discrete action spaces. Paschen [34] analyzed the dynamic behavior of day-ahead EM prices in Germany due to

structural shocks in wind and solar power by using a dynamic structural vector autoregressive model. Similar studies can also be seen in[35, 36], but researches in [29–36] regard the wind power or other renewable powers as an exogenous random variable so that strategic bidding behaviors of wind or other renewable power producers as well as impact of the EM bidding process on WPPs are neglected in those literatures.

So far as we know, there is no relevant research containing the following three points simultaneously:

(1) To construct a multiagent-based EM model which contains not only the impact of WPPs' uncertain output on strategic bidding behaviors of other market participants but also the impact of the EM bidding process on WPPs' bidding decision-making

(2) To construct a multiagent-based EM model in which both the EM environment state space and bidding strategy (action) spaces of all kinds of market participants such as WPPs, GenCOs, and DisCOs are continuous

(3) To construct a multiagent-based EM model in which the market clearing model of ISO is propitious to promote the wind power accommodation capacity of the power system, which is another aim of this paper

This paper applies a new modified reinforcement learning algorithm, namely, GDCAC algorithm, for hour-ahead EM modeling. In our proposed EM approach, all kinds of participants such as WPPs, GenCOs, and DisCOs are regarded as interactively strategic bidding agents who, during the bidding process, must select their optimal bidding strategies from their continuous strategy spaces based on the EM environment state they learned within a continuous state space, respectively, and without causing troubles of "curse of dimensionality." The market clearing model of ISO in our approach is a robust economic dispatch model (REDM) [37] which can optimize the permitted real-time wind power deviation intervals based on WPPs' bidding power output. By using our proposed approach, the dynamic interactions among all kinds of participants as well as the Nash equilibrium (NE) results of EM can be simulated and obtained. On the one hand, our proposed approach can provide a bidding decision-making tool for WPPs, GenCOs, and DisCOs to get more profits in EM. On the other hand, our proposed approach can also provide an economic and operational analysis tool for promoting the development of renewable resources. Moreover, in our simulation, the proposed approach is implemented on the IEEE 30-bus test system. Other than testing and verifying the feasibility and rationality of our proposed approach such as reaching NE results after enough iterations and being superior to other agent-based approaches, comparison of our proposed market clearing model with that in [12] under the same bidding approach based on the GDCAC algorithm is also implemented, which indicates the necessity of adopting the REDM for promoting wind power accommodation in EM.

The rest of this paper is organized as follows: in Section 2, the multiagent double-side hour-ahead EM modeling containing strategic bidding interactions among WPPs, GenCOs, and DisCOs are explained. Sections 3 and 4 describe the detailed procedure of applying the GDCAC algorithm for EM modeling. Section 5 conducts the simulations and comparisons. Section 6 concludes the paper.

2. Multiagent Hour-Ahead EM Modeling

2.1. Participants' Bidding Models. In our proposed double-side hour-ahead wholesale EM model, we consider every WPP, GenCO, and DisCO as an agent. An agent has the ability of learning through its bidding experiences in order to maximize its own profit. For the sake of simplicity and without the loss of generality, we assume that every WPP and GenCO has only one generation unit. In each hour, every GenCO and DisCO solves its own bidding problem and sends its price-quantity bid curve for the next hour to the ISO. Moreover, every WPP, because of its "price taker" role in EM, solves its own bidding problem and sends its bidding power output to the ISO. ISO, after receiving all bid curves from GenCOs and DisCOs as well as all bidding power outputs from WPPs, performs the process of robust economic dispatch management and sends the scheduled power results as well as LMPs to all market participants (WPPs, GenCOs, and DisCOs).

For WPP i ($i = 1, 2, \ldots, N_w$), the only bidding parameter for hour t is its planed (bidding) power output $P_{wi,t}$ ($P_{wi,t} \in [P_{wi,\min}, P_{wi,\max}]$). WPP i can adjust its bid by changing this parameter. In power systems of many countries, wind power is given priority to be scheduled by ISO comparing with other nonrenewable resources [37], which is to say prior-scheduled wind power for hour t, namely, $P^*_{wi,t}$, is equal to $P_{wi,t}$. However, because of the high variability and random nature of this intermittent resource, the (predicted) real-time output power of WPP i for hour t, namely, $P^{(r)}_{wi,t}$ ($P^{(r)}_{wi,t} \in [P_{wi,\min}, P_{wi,\max}]$), which is actually a random variable [12], usually tends to deviate from the scheduled one, which is harmful to the secure operation of the power system and tends to cause system imbalance. Hence, penalty mechanisms to financially punish WPPs for their deviations of real-time productions from their bidding ones must be involved. Taking the penalty method of [12] into consideration, the expected profit of WPP i for hour t can be described as follows:

$$r_{wi,t} = \sum_{\varepsilon \in S} p_\varepsilon \left[\mathrm{LMP}_{wi,t} P^{(r,\varepsilon)}_{wi,t} - \rho^{(\varepsilon)}_{wi,t} \left| P^*_{wi,t} - P^{(r,\varepsilon)}_{wi,t} \right| \right], \quad (1)$$

where $\mathrm{LMP}_{wi,t}$ represents the hour-ahead nodal price (LMP) for hour t at the bus connecting WPP i. ε is a random variable, which is used to describe the scenarios of wind power uncertainty. S represents the envelope space of wind power scenarios. p_ε represents the probability of occurrence of the scenario ε. $P^{(r,\varepsilon)}_{wi,t}$ and $\rho^{(\varepsilon)}_{wi,t}$ represent the (predicted) real-time power output and penalty price of WPP i for hour t in scenario ε, respectively. In this paper, we involve that the penalty price of WPP i is related to the (predicted) real-time LMP at the bus connecting WPP i [12].

Moreover, there is a difference between the (predicted) real-time power output and the (predicted) natural power

output (namely, $P_{wi,t}^{(\mathrm{na})}$ and $P_{wi,t}^{(\mathrm{na})} \in [P_{wi,\min}, P_{wi,\max}]$) of WPP i in hour t. WPP i can determine whether its (predicted) real-time power output is equal to the natural one by conducting pitch control or using storage equipment [37]. The functional relationship between these two random variables can be formulated as follows [37]:

$$P_{wi,t}^{(r)} = \begin{cases} P_{wi,t}^{\mathrm{lb}}, & P_{wi,t}^{(\mathrm{na})} \le P_{wi,t}^{\mathrm{lb}}, \\ P_{wi,t}^{(\mathrm{na})}, & P_{wi,t}^{\mathrm{lb}} < P_{wi,t}^{(\mathrm{na})} < P_{wi,t}^{\mathrm{ub}}, \\ P_{wi,t}^{\mathrm{ub}}, & P_{wi,t}^{(\mathrm{na})} \ge P_{wi,t}^{\mathrm{ub}}, \end{cases} \tag{2}$$

where $P_{wi,t}^{\mathrm{ub}}$ and $P_{wi,t}^{\mathrm{lb}}$ ($P_{wi,t}^{\mathrm{ub}}$ and $P_{wi,t}^{\mathrm{lb}} \in [P_{wi,\min}, P_{wi,\max}]$) represent the permitted upper and lower bounds of power output of WPP i that the can be accepted by system for hour t. In this paper, we consider the (predicted) real-time natural wind power outputs of all WPPs as common knowledge.

For GenCO$_j$ ($j = 1, 2, \ldots, N_g$), the formulation of its bid curve for the next hour t is a supply function based on its real marginal cost function [28]:

$$\mathrm{SF}_{j,t}\left(P_{gj,t}, k_{gj,t}\right) = k_{gj,t}\left(a_j P_{gj,t} + b_j\right), \quad P_{gj,t} \in \left[P_{gj,\min}, P_{gj,\max}\right], \tag{3}$$

where $P_{gj,t}$ and $k_{gj,t}$ represent the power production (MW) and bidding strategy ratio of GenCO$_j$ for hour t, respectively. GenCO$_j$ can adjust its bid curve by changing its parameter $k_{gj,t}$.

The marginal cost function of GenCO$_j$ is

$$\mathrm{MC}_j\left(P_{gj,t}\right) = a_j P_{gj,t} + b_j, \tag{4}$$

where a_j and b_j represent the slope and intercept parameters of GenCO$_j$'s marginal cost function, respectively.

Moreover, we assume every GenCO is an AGC (automatic generation control [37]) unit which can automatically undertake the real-time power imbalance of system with a certain proportion (namely, α). Therefore, the expected profit of GenCO$_j$ can be described as

$$r_{gj,t} = \sum_{\varepsilon \in S} p_\varepsilon \left\{ \mathrm{LMP}_{gj,t} P_{gj,t}^{*} - \rho_{gj,t}^{(\varepsilon)} \alpha_{j,t} \sum_{i=1}^{N_w} \left(P_{wi,t}^{(r,\varepsilon)} - P_{wi,t}^{*}\right) - \left\{ \frac{1}{2} a_j \left[P_{gj,t}^{*} - \alpha_{j,t} \sum_{i=1}^{N_w} \left(P_{wi,t}^{(r,\varepsilon)} - P_{wi,t}^{*}\right) \right]^2 + b_j \left[P_{gj,t}^{*} - \alpha_{j,t} \sum_{i=1}^{N_w} \left(P_{wi,t}^{(r,\varepsilon)} - P_{wi,t}^{*}\right) \right] \right\} \right\}, \tag{5}$$

where $\mathrm{LMP}_{gj,t}$ represents the hour-ahead nodal price (LMP) for hour t at the bus connecting GenCO$_j$, $\rho_{gj,t}^{(\varepsilon)}$ represents the (predicted) real-time nodal price (LMP) for hour t at the bus connecting GenCO$_j$ in scenario ε, and $P_{gj,t}^{*}$ represents the GenCO$_j$'s hour-ahead scheduled output power result for hour t.

For DisCO$_m$ ($m = 1, 2, \ldots, N_d$), the formulation of its bid curve for the next hour t is a demand function based on its real marginal revenue function [28]:

$$\mathrm{DF}_{m,t}\left(P_{dm,t}, k_{dm,t}\right) = k_{dm,t}\left(-c_m P_{dm,t} + d_m\right), \quad P_{dm,t} \in \left[P_{dm,\min}, P_{dm,\max}\right], \tag{6}$$

where $P_{dm,t}$ and $k_{dm,t}$ represent the power demand (MW) and bidding strategy ratio of DisCO$_m$ for hour t, respectively. DisCO$_m$ can adjust its bid curve by changing its parameter $k_{dm,t}$.

The marginal revenue function of DisCO$_m$ is

$$\mathrm{MD}_m\left(P_{dm,t}\right) = -c_m P_{dm,t} + d_m, \tag{7}$$

where $-c_m$ and d_m represent the slope and intercept parameters of DisCO$_m$'s marginal revenue function, respectively.

Profit of DisCO$_m$ can be described as

$$r_{dm,t} = \left(-\frac{1}{2} c_m P_{dm,t}^{*2} + d_m P_{dm,t}^{*}\right) - \mathrm{LMP}_{dm,t} P_{dm,t}^{*}, \tag{8}$$

where $\mathrm{LMP}_{dm,t}$ is the hour-ahead nodal price (LMP) for hour t at the bus connecting DisCO$_m$ and $P_{dm,t}^{*}$ represents the DisCO$_m$'s hour-ahead scheduled power demand (load) result for hour t.

2.2. ISO's Market Clearing Model. In the traditional dispatching mode considering wind power penetration, ISO sends the scheduled values of wind power to WPPs and WPPs are required to strictly follow the scheduled values in the case of their generation capacities. This traditional mode has the following two obvious defects [37]:

(1) In the case of low precision of wind power prediction, the traditional dispatching mode is not conducive to the wind power accommodation. It can lead to extreme operating conditions, which may seriously threaten the system security when the wind power violently fluctuates.

(2) It may lead to frequent pitch control when wind turbines strictly track the scheduled values of output power, which would affect the lives of the wind turbines.

The main reason for those two defects listed above is that in the traditional dispatch mode, the uncertainty of wind power is not taken into account. Hence, ISO does not know the maximum permitted wind power output fluctuation range in the premise of ensuring system security and cannot optimize wind power accommodation capacity of the power grid. Therefore, nowadays, more and more attentions have been paid to the REDM [37] which aims to promote the wind power accommodation in considering wind power uncertainty. According to [37], the robust hour-ahead economic dispatch model for hour t can be mathematically described as follows:

$$\begin{aligned}\underset{P^*_{gj,t},\forall j;P^*_{dm,t},\forall m;\ P^{ub}_{wi,t},P^{lb}_{wi,t},\forall i}{\text{Max}} & \left\{ \sum_{m=1}^{N_d}\left[k_{dm,t}\left(-\frac{1}{2}c_m P^{*2}_{dm,t}+d_m P^*_{dm,t}\right)\right] \right. \\ & \left. - \sum_{j=1}^{N_g}\left[k_{gj,t}\left(\frac{1}{2}a_j P^{*2}_{gj,t}+b_j P^*_{gj,t}\right)\right]\right\} \\ & - \sum_{i=1}^{N_w} M_i^{\text{ub}}\left(P_{wi,\max}-P^{\text{ub}}_{wi,t}\right)^2 \\ & - \sum_{i=1}^{N_w} M_i^{\text{lb}}\left(P^{\text{lb}}_{wi,t}-P_{wi,\min}\right)^2, \end{aligned} \tag{9}$$

$$\text{s.t.} \quad \sum_{i=1}^{N_w} P^*_{wi,t}+\sum_{j=1}^{N_g} P^*_{gj,t}-\sum_{m=1}^{N_d} P^*_{dm,t}=0, \tag{10}$$

$$P_{l,\min}\le \sum_{z=1}^{Z} P^*_{\text{G}z,t}\times sf_{l,\text{G}z}-\sum_{z=1}^{Z} P^*_{\text{D}z,t}\times sf_{l,\text{D}z}\le P_{l,\max}, \quad \forall l, \tag{11}$$

$$P^*_{\text{G}z,t}=\sum_{i'\in \text{BUS}_z} P^*_{wi',t}+\sum_{j'\in \text{BUS}_z} P^*_{gj',t}, \tag{12}$$

$$P^*_{\text{D}z,t}=\sum_{m'\in \text{BUS}_z} P^*_{dm',t}, \tag{13}$$

$$P^*_{dm,t}\in\left[P_{dm,\min},P_{dm,\max}\right], \quad \forall m, \tag{14}$$

$$P^*_{gj,t}\in\left[P_{gj,\min},P_{gj,\max}\right], \quad \forall j, \tag{15}$$

$$P^{(r)}_{gj,t}=P^*_{gj,t}-\alpha_{j,t}\sum_{i=1}^{N_w}\left(\widetilde{\widetilde{P}}^{(r)}_{wi,t}-P^*_{wi,t}\right)\in\left[P_{gj,\min},P_{gj,\max}\right], \quad \forall j, \tag{16}$$

$$P_{l,\min}\le \sum_{z=1}^{Z} P^{(r)}_{\text{G}z,t}\times sf_{l,\text{G}z}-\sum_{z=1}^{Z} P^*_{\text{D}z,t}\times sf_{l,\text{D}z}\le P_{l,\max}, \quad \forall l, \tag{17}$$

$$P^{(r)}_{\text{G}z,t}=\sum_{i'\in \text{BUS}_z} P^{(r)}_{wi',t}+\sum_{j'\in \text{BUS}_z} P^{(r)}_{gj',t}, \tag{18}$$

$$P_{wi,\min}\le P^{\text{lb}}_{wi,t}\le P^*_{wi,t}\le P^{\text{ub}}_{wi,t}\le P_{wi,\max}, \quad \forall i, \tag{19}$$

where M_i^{ub} and M_i^{lb} (M_i^{ub} and $M_i^{\text{lb}}>0$) in equation (8) represent the deviation penalty coefficients of permitted upper and lower bounds of the wind power output of WPP i, and equations (9)–(15) represent the hour-ahead system constraints including power balance constraint (equation (9)), DC power flow constraints in each transmission line l (equations (11)–(13)), and load and power production of every DisCO and GenCO (equations (14) and (15)). The hour-ahead LMPs of system can be calculated by using dual variables of equations (9)–(13). Formulations for hour-ahead LMP are in Appendix A. Equations (16)–(19) represent the (predicted) real-time system constraints including power balance constraint (equation (16)), DC power flow constraints in each transmission line l (equations (17) and (18)), and power production of every WPP (equation (19)).

From equations (16)–(18), it is obvious that (predicted) real-time DC power flow in each transmission line l is the linear function of (predicted) real-time power output by every WPP. From equation (2), (predicted) real-time power output of WPP i ($i=1, 2, \ldots, N_w$) must satisfy

$$P^{(r)}_{wi,t}\in\left[P^{\text{lb}}_{wi,t},P^{\text{ub}}_{wi,t}\right], \quad \forall i, \tag{20}$$

to say we can solve the abovementioned REDM by replacing $P^{(r)}_{wi,t}$ with $P^{\text{ub}}_{wi,t}$ and $P^{\text{lb}}_{wi,t}$, respectively (Appendix B) [37] and generating new (predicted) real-time balancing and transmission constraints as follows:

$$P^{(r1)}_{gj,t}=P^*_{gj,t}-\alpha_{j,t}\sum_{i=1}^{N_w}\left(P^{\text{ub}}_{wi,t}-P^*_{wi,t}\right)\in\left[P_{gj,\min},P_{gj,\max}\right], \quad \forall j, \tag{21}$$

$$P_{l,\min}\le \sum_{z=1}^{Z} P^{(r1)}_{\text{G}z,t}\times sf_{l,\text{G}z}-\sum_{z=1}^{Z} P^*_{\text{D}z,t}\times sf_{l,\text{D}z}\le P_{l,\max}, \quad \forall l, \tag{22}$$

$$P^{(r1)}_{\text{G}z,t}=\sum_{i'\in \text{BUS}_z} P^{\text{ub}}_{wi',t}+\sum_{j'\in \text{BUS}_z} P^{(r1)}_{gj',t}, \tag{23}$$

$$P^{(r2)}_{gj,t}=P^*_{gj,t}-\alpha_{j,t}\sum_{i=1}^{N_w}\left(P^{\text{lb}}_{wi,t}-P^*_{wi,t}\right)\in\left[P_{gj,\min},P_{gj,\max}\right], \quad \forall j, \tag{24}$$

$$P_{l,\min}\le \sum_{z=1}^{Z} P^{(r2)}_{\text{G}z,t}\times sf_{l,\text{G}z}-\sum_{z=1}^{Z} P^*_{\text{D}z,t}\times sf_{l,\text{D}z}\le P_{l,\max}, \quad \forall l, \tag{25}$$

$$P^{(r2)}_{\text{G}z,t}=\sum_{i'\in \text{BUS}_z} P^{\text{lb}}_{wi',t}+\sum_{j'\in \text{BUS}_z} P^{(r2)}_{gj',t}. \tag{26}$$

The (predicted) real-time LMPs (RTLMP_1s) of system when (predicted) real-time power output of every WPP increases to its (scheduled) permitted upper bound can be calculated by using dual variables of equations (9) and (21)–(23), and the (predicted) real-time LMPs (RTLMP_2s) of system when (predicted) real-time power output of every WPP decreases to its (scheduled) permitted lower bound can be calculated by using dual variables of equations (9) and (24)–(26). Therefore, RTLMP_1s and RTLMP_2s represent 2 extreme real-time dispatching results caused by real-time wind power deviations of all WPPs. For the sake of simplicity and without loss of generality, we approximately consider the mean value of RTLMP_1 and RTLMP_2 at bus z as the (predicted) real-time

LMP at bus **z** and neglect the impact of different $\varepsilon \in S$ on (predicted) real-time LMPs.

3. Agent-Learning Mechanism

For an agent in our proposed approach, all the other agents together constitute the EM environment it faces. Therefore, interactions between an agent and all the other agents are equivalent to interactions between this agent and the EM environment it faces. An agent has the ability of learning through repeated interactions with the EM environment for finding its optimal action (bidding strategy or bidding power output), which can maximize its (expected) profit in face of whatever the EM environment state is. In this paper, in order to clearly describe our proposed approach, we use the definitions which are organized as follows:

(1) *Iteration*. Since the market is assumed to be cleared in hour-ahead basis, we define each market round as an iteration.

(2) *State Variable*. For WPP_i and in iteration t, the hour-ahead and (predicted) real-time LMPs at the bus connecting WPP i calculated in iteration $t-1$, namely, $\text{LMP}_{wi,t-1}$, $\rho_{wi,t-1}$, are defined as the EM environment state variables; for GenCOj, the hour-ahead and (predicted) real-time LMPs at the bus connecting GenCOj calculated in iteration $t-1$, namely, $\text{LMP}_{gj,t-1}$ and $\rho_{gj,t-1}$, are defined as the EM environment state variable. For DisCO_m, the hour-ahead LMP at the bus connecting DisCO_m calculated in iteration $t-1$, namely, $\text{LMP}_{dm,t-1}$, is defined as the EM environment state variable. Hence, the state vectors and scalar for WPP_i, GenCO_j, and DisCO_m can be formulated as follows [28]:

$$\begin{aligned}\mathbf{x}_{wi,t} &= \left(\text{LMP}_{wi,t-1}, \rho_{wi,t-1}\right) \in \mathbf{X}_{wi},\\ \mathbf{x}_{gj,t} &= \left(\text{LMP}_{gj,t-1}, \rho_{gj,t-1}\right) \in \mathbf{X}_{gj},\\ \mathbf{x}_{dm,t} &= \text{LMP}_{dm,t-1} \in \mathbf{X}_{dm},\end{aligned} \tag{27}$$

where $\mathbf{X}_{wi}$, $\mathbf{X}_{gj}$, and $\mathbf{X}_{dm}$ are continuous, closed, and bounded state spaces for WPP_i, GenCO_j, and DisCO_m, respectively.

(3) *Action Variable*. For WPP_i, the hour-ahead bidding power output, namely, $P_{wi,t}$ ($P_{wi,t} \in [P_{wi,\min}, P_{wi,\max}]$), is defined as the action variable of this agent in iteration t. For GenCO_j or DisCO_m, the hour-ahead bidding strategy rate, namely, $k_{gj,t}$ or $k_{dm,t}$, is defined as the action variable of GenCO_j or DisCO_m in iteration t. Hence, the action scalars for WPP_i, GenCO_j, and DisCO_m can be formulated as follows:

$$u_{wi,t} = P_{wi,t} \in \left[P_{wi,\min}, P_{wi,\max}\right], \tag{28}$$

$$u_{gj,t} = k_{gj,t} \in \left[k_{gj,\min}, k_{gj,\max}\right], \tag{29}$$

$$u_{dm,t} = k_{dm,t} \in \left[k_{dm,\min}, k_{dm,\max}\right]. \tag{30}$$

Obviously, from equations (28)–(30), we can see that the action spaces for WPP_i, GenCO_j, and DisCO_m are continuous, closed, and bounded intervals.

(4) *Reward*. In iteration t, similar to what was mentioned in [28], every agent learns from the state of the EM environment ($\mathbf{x}_{wi,t}, \mathbf{x}_{gj,t}$, and $\mathbf{x}_{dm,t}$) and then selects its action which in turn forms its bidding power output or curve for sending to the ISO. After receiving all bidding outputs and curves, hour-ahead LMPs permitted upper and lower bounds of (predicted) real-time power outputs by WPPs, as well as hour-ahead power supply and demand schedules are determined by ISO with our REDM represented by equations (8)–(19). Rewards of WPP_i, GenCO_j, and DisCO_m can be depicted as equations (1), (5), and (8), respectively.

Based on experiencing these received rewards over enough iterations, an agent in EM can gradually learn to know how to take the corresponding optimal hour-ahead action

$$\begin{aligned}u_{wi,t}^{(\text{opt})}\left(\mathbf{x}_{wi,t}\right) &\in \left[P_{wi,\min}, P_{wi,\max}\right], \quad i = 1, 2, \ldots, N_w\left(\frac{\pi}{2} - \theta\right),\\ u_{gj,t}^{(\text{opt})}\left(\mathbf{x}_{gj,t}\right) &\in \left[k_{gj,\min}, k_{gj,\max}\right], \quad j = 1, 2, \ldots, N_g,\\ u_{dm,t}^{(\text{opt})}\left(\mathbf{x}_{dm,t}\right) &\in \left[k_{dm,\min}, k_{dm,\max}\right], \quad m = 1, 2, \ldots, N_d,\end{aligned} \tag{31}$$

which brings the most profit in face of any state ($\mathbf{x}_{wi,t}, \mathbf{x}_{gj,t}$, and $\mathbf{x}_{dm,t}$) of the EM environment. Hence, $u_{wi,t}, u_{gj,t}$, and $u_{dm,t}$ and $\mathbf{x}_{wi,t}, \mathbf{x}_{gj,t}$, and $\mathbf{x}_{dm,t}$ ($i = 1, 2, \ldots, N_w$; $j = 1, 2, \ldots, N_g$; $m = 1, 2, \ldots, N_d$) are changing dynamically over iterations, which may be or not be constant after enough iterations.

4. Methodology

Inspired by the studies in [19–26], the dynamic bidding process in spot EM can be realized via table-based reinforcement learning algorithms (TBRLAs) such as Q-learning, fuzzy Q-learning, Roth–Erev learning, and SARSA algorithms. As mentioned in [28, 38], TBRLAs can only rapidly solve the Markov decision-making problems with discrete state and action spaces. When one of the state and action spaces becomes continuous, the problem called "curse of dimensionality" will be caused, and the learning speed of TBRLAs becomes so slow that the agent cannot find its optimal action under any given state of environment over iterations.

As mentioned in Section 3, actually both the state and action spaces of every agent in EM are continuous, closed, and bounded space or interval, which guarantees the process of global optimization. Therefore, it is improper to model and simulate the dynamic bidding process in our proposed hour-ahead EM containing strategic bidding interactions among WPPs, GenCOs, and DisCOs by using TBRLAs. Method in this paper is to apply a modified reinforcement learning algorithm, called the GDCAC algorithm [28, 38], for modeling and simulating our proposed EM.

Because the mathematical principle and pseudocode of the GDCAC algorithm have been described in [28], we only propose the step-by-step procedure of implementing the GDCAC algorithm for hour-ahead EM modeling containing strategic bidding interactions among WPPs, GenCOs, and DisCOs as follows:

(1) *Input*. For the whole EM, input common knowledge is such as every WPP's reduced (predicted) real-time wind power output scenarios (WPOSs) with corresponding probabilities and all WPP's joint real-time WPOSs with corresponding probabilities. For WPPi ($i = 1, 2, ..., N_w$), input the basic function $\overrightarrow{\boldsymbol{\phi}}_{wi}: \mathbf{X}_{wi} \longrightarrow \mathbf{R}^n$ for formulating its value function $\widehat{V}_{wi,t}(\mathbf{x}_{wi,t}) = \sum_{h=1}^{n} \phi_{wi,h}(\mathbf{x}_{wi,t})\theta_{wih,t} = \overrightarrow{\boldsymbol{\phi}}_{wi}(\mathbf{x}_{wi,t})^T \boldsymbol{\theta}_{wi,t}$, $\mathbf{x}_{wi,t} \in \mathbf{X}_{wi}$, and its optimal policy function $u_{wi,t}^{(\mathrm{opt})}(\mathbf{x}_{wi,t}) = \overrightarrow{\boldsymbol{\phi}}_{wi}(\mathbf{x}_{wi,t})^T \boldsymbol{\omega}_{wi,t}, \mathbf{x}_{wi,t} \in \mathbf{X}_{wi}$, time step length parameter series $\left\{\alpha_t^{(w)}\right\}_{t=1}^{\infty}$ and $\left\{\beta_t^{(w)}\right\}_{t=1}^{\infty}$ where $\sum_{t=1}^{\infty} \alpha_t^{(w)} = \infty$ and $\sum_{t=1}^{\infty} (\alpha_t^{(w)})^2 < \infty$ and $\sum_{t=1}^{\infty} \beta_t^{(w)} = \infty$ and $\sum_{t=1}^{\infty} (\beta_t^{(w)})^2 < \infty$. For GenCO$_j$ ($j = 1, 2, ..., N_g$), input the basic function $\overrightarrow{\boldsymbol{\phi}}_{gj}: \mathbf{X}_{gj} \longrightarrow \mathbf{R}^n$ for formulating its value function $\widehat{V}_{gj,t}(\mathbf{x}_{gj,t}) = \sum_{h=1}^{n} \phi_{gj,h}(\mathbf{x}_{gj,t})\theta_{gjh,t} = \overrightarrow{\boldsymbol{\phi}}_{gj}(\mathbf{x}_{gj,t})^T \boldsymbol{\theta}_{gj,t}$, $\mathbf{x}_{gj,t} \in \mathbf{X}_{gj}$ and its optimal policy function $u_{gj,t}^{(\mathrm{opt})}(\mathbf{x}_{gj,t}) = \overrightarrow{\boldsymbol{\phi}}_{gj}(\mathbf{x}_{gj,t})^T \boldsymbol{\omega}_{gj,t}, \mathbf{x}_{gj,t} \in \mathbf{X}_{gj}$, time step length parameter series $\left\{\alpha_t^{(g)}\right\}_{t=1}^{\infty}$ and $\left\{\beta_t^{(g)}\right\}_{t=1}^{\infty}$, where $\sum_{t=1}^{\infty} \alpha_t^{(g)} = \infty$ and $\sum_{t=1}^{\infty} (\alpha_t^{(g)})^2 < \infty$ and $\sum_{t=1}^{\infty} \beta_t^{(g)} = \infty$ and $\sum_{t=1}^{\infty} (\beta_t^{(g)})^2 < \infty$. For DisCO$_m$, ($m = 1, 2, ..., N_d$) input the basic function $\overrightarrow{\boldsymbol{\phi}}_{dm}: \mathbf{X}_{dm} \longrightarrow \mathbf{R}^n$ for formulating its value function $\widehat{V}_{dm,t}(\mathbf{x}_{dm,t}) = \sum_{h=1}^{n} \phi_{dm,h}(\mathbf{x}_{dm,t})\theta_{dmh,t} = \overrightarrow{\boldsymbol{\phi}}_{dm}(\mathbf{x}_{dm,t})^T \boldsymbol{\theta}_{dm,t}, \mathbf{x}_{dm,t} \in \mathbf{X}_{dm}$ and its optimal policy function $u_{dm,t}^{(\mathrm{opt})}(\mathbf{x}_{dm,t}) = \overrightarrow{\boldsymbol{\phi}}_{dm}(\mathbf{x}_{dm,t})^T \boldsymbol{\omega}_{dm,t}, \mathbf{x}_{dm,t} \in \mathbf{X}_{dm}$, time step length parameter series $\left\{\alpha_t^{(d)}\right\}_{t=1}^{\infty}$ and $\left\{\beta_t^{(d)}\right\}_{t=1}^{\infty}$ where $\sum_{t=1}^{\infty} \alpha_t^{(d)} = \infty$ and $\sum_{t=1}^{\infty} (\alpha_t^{(d)})^2 < \infty$ and $\sum_{t=1}^{\infty} \beta_t^{(d)} = \infty$ and $\sum_{t=1}^{\infty} (\beta_t^{(d)})^2 < \infty$. Moreover, input the discount, standard deviation, as well as the maximum training and decision-making iterations parameters, namely, $0 \le \gamma \le 1$, $\sigma_{wi}(\sigma_{gj}, \sigma_{dm}) > 0$ and T_1 and T_2, for every WPP, GenCO, and DisCO.

(2) $t = 0$.

(3) Initialize the linear parameter vectors $\boldsymbol{\theta}_{wi,0}$ and $\boldsymbol{\omega}_{wi,0}$ for WPP$_i$, linear parameter vectors $\boldsymbol{\theta}_{gj,0}$ and $\boldsymbol{\omega}_{gj,0}$ for GenCO$_j$, and linear parameter vectors $\boldsymbol{\theta}_{dm,0}$ and $\boldsymbol{\omega}_{dm,0}$ for DisCO$_m$.

(4) If $t < T_1$, then in iteration t, WPPi selects and implements an action $u_{wi,t} \sim N(\overrightarrow{\boldsymbol{\phi}}_{wi}(\mathbf{x}_{wi,t})^T \boldsymbol{\omega}_{wi,t}, \sigma_{wi}^2)$ ($u_{wi,t} \in [P_{wi,\min}, P_{wi,\max}]$) from state $\mathbf{x}_{wi,t}$, GenCO$_j$ selects and implements an action $u_{gj,t} \sim N(\overrightarrow{\boldsymbol{\phi}}_{gj}(\mathbf{x}_{gj,t})^T \boldsymbol{\omega}_{gj,t}, \sigma_{gi}^2)$ ($u_{gj,t} \in [k_{gj,\min}, k_{gj,\max}]$) from state $\mathbf{x}_{gj,t}$, and DisCO$_m$ selects and implements an action $u_{dm,t} \sim N(\overrightarrow{\boldsymbol{\phi}}_{dm}(\mathbf{x}_{dm,t})^T \boldsymbol{\omega}_{dm,t}, \sigma_{dm}^2)$ ($u_{dm,t} \in [k_{dm,\min}, k_{dm,\max}]$) from state $\mathbf{x}_{dm,t}$. If $T_1 < t < T_1 + T_2$, then in iteration t, WPP$_i$ selects and implements an action $u_{wi,t} = \overrightarrow{\boldsymbol{\phi}}_{wi}(\mathbf{x}_{wi,t})^T \boldsymbol{\omega}_{wi,t}$ ($u_{wi,t} \in [P_{wi,\min}, P_{wi,\max}]$) from state $\mathbf{x}_{wi,t}$, GenCO$_j$ selects and implements an action $u_{gj,t} = \overrightarrow{\boldsymbol{\phi}}_{gj}(\mathbf{x}_{gj,t})^T \boldsymbol{\omega}_{gj,t}$ ($u_{gj,t} \in [k_{gj,\min}, k_{gj,\max}]$) from state $\mathbf{x}_{gj,t}$, and DisCO$_m$ selects and implements an action $u_{dm,t} = \overrightarrow{\boldsymbol{\phi}}_{dm}(\mathbf{x}_{dm,t})^T \boldsymbol{\omega}_{dm,t}$ ($u_{dm,t} \in [k_{dm,\min}, k_{dm,\max}]$) from state $\mathbf{x}_{dm,t}$. After action selecting and sending it to ISO by every agent, ISO implements the REDM represented by equations (8)–(19) by which the EM environment state vector variables are updated from $\mathbf{x}_{wi,t}$, $\mathbf{x}_{gj,t}$, and $\mathbf{x}_{dm,t}$ to $\mathbf{x}_{wi,t+1}, \mathbf{x}_{gj,t+1}$, and $\mathbf{x}_{dm,t+1}$ and the immediate reward $r_{wi,t}$, $r_{gj,t}$, and $r_{dm,t}$ are generated.

(5) WPP i observes the immediate reward $r_{wi,t}$ by using equation (1) and the new EM environment state $\mathbf{x}_{wi,t+1}$; GenCOj observes the immediate reward $r_{gj,t}$ by using equation (5) and the new EM environment state $\mathbf{x}_{gj,t+1}$; and DisCO$_m$ observes the immediate reward $r_{dm,t}$ by using equation (8) and the new EM environment state $\mathbf{x}_{dm,t+1}$.

(6) *Learning*. In this step, $\boldsymbol{\theta}_{wi,t}$ and $\boldsymbol{\omega}_{wi,t}$ for WPP i, $\boldsymbol{\theta}_{gj,t}$ and $\boldsymbol{\omega}_{gj,t}$ for GenCO$_j$, and $\boldsymbol{\theta}_{dm,t}$ and $\boldsymbol{\omega}_{dm,t}$ for DisCO$_m$ are updated by using the *TD* (0) error (namely, $\delta_{wi,t}$, $\delta_{gj,t}$, and $\delta_{dm,t}$) and gradient descent method.

WPP i:

$$\begin{aligned}
\delta_{wi,t} &= r_{wi,t} + \gamma \overrightarrow{\boldsymbol{\phi}}_{wi}\left(\mathbf{x}_{wi,t+1}\right)^T \boldsymbol{\theta}_{wi,t} - \overrightarrow{\boldsymbol{\phi}}_{wi}\left(\mathbf{x}_{wi,t}\right)^T \boldsymbol{\theta}_{wi,t}, \\
\boldsymbol{\theta}_{wi,t+1} &= \boldsymbol{\theta}_{wi,t} + \alpha_t^{(w)} \delta_{wi,t} \overrightarrow{\boldsymbol{\phi}}_{wi}\left(\mathbf{x}_{wi,t}\right), \\
\boldsymbol{\omega}_{wi,t+1} &= \boldsymbol{\omega}_{wi,t} + \beta_t^{(w)} \frac{1}{1 + e^{-m\delta_{wi,t}}} \\
&\quad \cdot \left(u_{wi,t} - \overrightarrow{\boldsymbol{\phi}}_{wi}\left(\mathbf{x}_{wi,t}\right)^T \boldsymbol{\omega}_{wi,t}\right) \overrightarrow{\boldsymbol{\phi}}_{wi}\left(\mathbf{x}_{wi,t}\right).
\end{aligned} \tag{32}$$

GenCO$_j$:

$$\begin{aligned}
\delta_{gj,t} &= r_{gj,t} + \gamma \overrightarrow{\boldsymbol{\phi}}_{gj}\left(\mathbf{x}_{gj,t+1}\right)^T \boldsymbol{\theta}_{gj,t} - \overrightarrow{\boldsymbol{\phi}}_{gj}\left(\mathbf{x}_{gj,t}\right)^T \boldsymbol{\theta}_{gj,t}, \\
\boldsymbol{\theta}_{gj,t+1} &= \boldsymbol{\theta}_{gj,t} + \alpha_t^{(g)} \delta_{gj,t} \overrightarrow{\boldsymbol{\phi}}_{wi}\left(\mathbf{x}_{gj,t}\right), \\
\boldsymbol{\omega}_{gj,t+1} &= \boldsymbol{\omega}_{gj,t} + \beta_t^{(g)} \frac{1}{1 + e^{-m\delta_{gj,t}}} \left(u_{gj,t} - \overrightarrow{\boldsymbol{\phi}}_{gj}\left(\mathbf{x}_{gj,t}\right)^T \boldsymbol{\omega}_{gj,t}\right) \\
&\quad \cdot \overrightarrow{\boldsymbol{\phi}}_{gj}\left(\mathbf{x}_{gj,t}\right).
\end{aligned} \tag{33}$$

DisCO$_m$:

$$\begin{aligned}
\delta_{dm,t} &= r_{dm,t} + \gamma \overrightarrow{\boldsymbol{\phi}}_{dm}\left(\mathbf{x}_{dm,t+1}\right)^T \boldsymbol{\theta}_{dm,t} - \overrightarrow{\boldsymbol{\phi}}_{dm}\left(\mathbf{x}_{dm,t}\right)^T \boldsymbol{\theta}_{dm,t}, \\
\boldsymbol{\theta}_{dm,t+1} &= \boldsymbol{\theta}_{dm,t} + \alpha_t^{(d)} \delta_{dm,t} \overrightarrow{\boldsymbol{\phi}}_{dm}\left(\mathbf{x}_{dm,t}\right), \\
\boldsymbol{\omega}_{dm,t+1} &= \boldsymbol{\omega}_{dm,t} + \beta_t^{(d)} \frac{1}{1 + e^{-m\delta_{dm,t}}} \\
&\quad \cdot \left(u_{dm,t} - \overrightarrow{\boldsymbol{\phi}}_{dm}\left(\mathbf{x}_{dm,t}\right)^T \boldsymbol{\omega}_{dm,t}\right) \overrightarrow{\boldsymbol{\phi}}_{dm}\left(\mathbf{x}_{dm,t}\right).
\end{aligned} \tag{34}$$

(7) $t = t + 1$.

(8) If $t < T_1 + T_2$, return to step (4).

(9) *Output*. For WPP_i, $\boldsymbol{\theta}^*_{wi} = \boldsymbol{\theta}_{wi,T_1+T_2}$ and $\boldsymbol{\omega}^*_{wi} = \boldsymbol{\omega}_{wi,T_1+T_2}$ and $\widehat{V}^*_{wi}(\mathbf{x}) = \widehat{V}_{wi,T_1+T_2}(\mathbf{x})$ and $u^{(\text{opt})*}_{wi}(\mathbf{x}) = u^{(\text{opt})}_{wi,T_1+T_2}(\mathbf{x})$. For GenCO_j, $\boldsymbol{\theta}^*_{gi} = \boldsymbol{\theta}_{gj,T_1+T_2}$ and $\boldsymbol{\omega}^*_{gj} = \boldsymbol{\omega}_{gj,T_1+T_2}$ and $\widehat{V}^*_{gj}(\mathbf{x}) = \widehat{V}_{gj,T_1+T_2}(\mathbf{x})$ and $u^{(\text{opt})*}_{gj}(\mathbf{x}) = u^{(\text{opt})}_{gj,T_1+T_2}(\mathbf{x})$. For DisCO_m, $\boldsymbol{\theta}^*_{dm} = \boldsymbol{\theta}_{dm,T_1+T_2}$ and $\boldsymbol{\omega}^*_{dm} = \boldsymbol{\omega}_{dm,T_1+T_2}$ and $\widehat{V}^*_{dm}(\mathbf{x}) = \widehat{V}_{dm,T_1+T_2}(\mathbf{x})$ and $u^{(\text{opt})*}_{dm}(\mathbf{x}) = u^{(\text{opt})}_{dm,T_1+T_2}(\mathbf{x})$.

According to [28, 38], we choose Gaussian radial basis function as $\overrightarrow{\boldsymbol{\phi}}_{wi}(\mathbf{x})$, $\overrightarrow{\boldsymbol{\phi}}_{gj}(\mathbf{x})$, and $\overrightarrow{\boldsymbol{\phi}}_{dm}(\mathbf{x})$.

5. Simulation Results and Discussions

5.1. Data and Assumptions. In this section, our proposed approach is implemented on the IEEE 30-bus test system with 2 WPPs, 6 GenCOs, and 20 DisCOs [2]. The schematic structure of this test system is shown in Figure 1. The output power of the WPP connected to bus 7 (marked as WPP 1) and 10 (marked as WPP 2) lies within the ranges of [0 80] MW and [0 50] MW, respectively. According to [39, 40], we assume both of the real-time wind power outputs of these two WPPs follow the Weibull distribution independently and respectively. Then, the (predicted) real-time WPOSs of these two WPPs can be generated by using the Monte Carlo method, and method of real-time WPOS reduction is referred to [39, 40]. Table 1 shows the reduced 10 (predicted) real-time WPOSs and their corresponding probabilities of these two WPPs which can be used as exogenous parameters in our proposed approach.

Based on Table 1, the number of joint WPOSs corresponding to combinations of (predicted) real-time power outputs generated by WPP1 and WPP2 is still 100 (10×10) which is too many for the subsequent calculations. Hence, in this paper, the 100 joint WPOSs are further reduced to 10 by using the tabu search algorithm proposed in [40]. Table 2 shows the reduced 10 (predicted) real-time joint WPOSs and their corresponding probabilities.

Moreover, parameters of GenCOs' and DisCOs' bid functions are shown in Tables 3 and 4 [2], respectively.

In order to verify the 3 points, which are as follows: (1) our proposed EM approach can reach dynamic stability and Nash equilibrium (NE) after enough training and decision-making iterations, (2) the superiority of our proposed EM approach comparing with approaches based on TBRL algorithms (e.g., Q-learning algorithm) in terms of participants' (expected) profits and expected social welfare (SW) can be calculated as the sum of (expected) profits of all participants [2], and (3) the impact of different market clearing methods (e.g., REDM and stochastic economic dispatch model (SEDM) [12]) on bidding stability results considering strategic interactions among WPPs and other participants, 3 corresponding simulations conducted by using Matlab R2014a software are carried out one by one as follows.

5.2. Testing the Ability of Our Proposed EM Approach to Reach Dynamic Stability and NE. In this section, we assume that every WPP, GenCO, and DisCO in the market are the GDCAC-based agents with continuous state and action spaces, and dynamic interactions among all GDCAC-based agents actually constitute our proposed GDCAC-based EM approach. The related parameters of the GDCAC algorithm are listed in Table 5.

In our simulation and comparisons (the same as the subsequent sections), every agent will go through a process of training with 3000 iterations in which all agents' action selecting policies consider the balance of exploration and exploitation [28]. After the training process, decision-making process with 500 iterations will be implemented by all agents, in which only the greedy policy will be adopted when selecting actions in face of any state of the market [28]. Moreover, in the beginning of the first training iteration, because every agent has no experience in strategy selecting, we randomly set hour-ahead bidding outputs of WPPs and bidding strategies of GenCOs and DisCOs within their respective intervals.

During the decision-making process, the dynamic adjustment of the EM environment state and bidding strategy (output) of every agent may be constant which means the market reaches the dynamic stability. Testing and verifying whether our proposed GDCAC-based approach reaches to dynamic stability after 3000 training iterations can be shown in Figures 2–4, respectively.

From Figures 2–4, we can see that the adjusting processes of hour-ahead LMPs, (expected) profit of every agent, and (predicted) real-time LMPs connecting WPPs (penalty prices charging from WPPs) in our proposed GDCAC-based approach keep constant during 500 decision-making iterations. It has been verified in [28] that other adjusting processes in EM such as that of expected SW and every agent's bidding strategy would reach constant while the adjusting process of LMPs keeps constant. Therefore, reaching the dynamic stability of our proposed GDCAC-based approach after 3000 training iterations is concluded in this paper. However, dynamic stability is not equivalent to NE. Hence, in order to examine whether the obtained bidding strategies of all agents after 3000 iterations of the training process and 500 iterations of decision-making process reach NE, we observe each agent's (expected) profit by changing its bidding strategy but fixing the other agents' bidding strategies after 3500 iterations. A combination of the obtained bidding strategies of all agents represents NE when there is no agent that can increase its (expected) profit in case of other agents' bidding strategies unchanged. We define a Nash index [2] which is equal to 1 when the NE is reached and otherwise is equal to 0. Figure 5 demonstrates the adjusting process of Nash indices during 3500 iterations in our proposed GDCAC-based approach.

It is known to us from Figure 5 that our proposed GDCAC-based EM approach is able to successfully generalize agents' experiences in face of any state point from the adjacent state points to reach NE after enough training and decision-making iterations. Moreover, by using the same method, the ability to reach the dynamic stability and NE of the comparative Q-learning-based approach, which will be

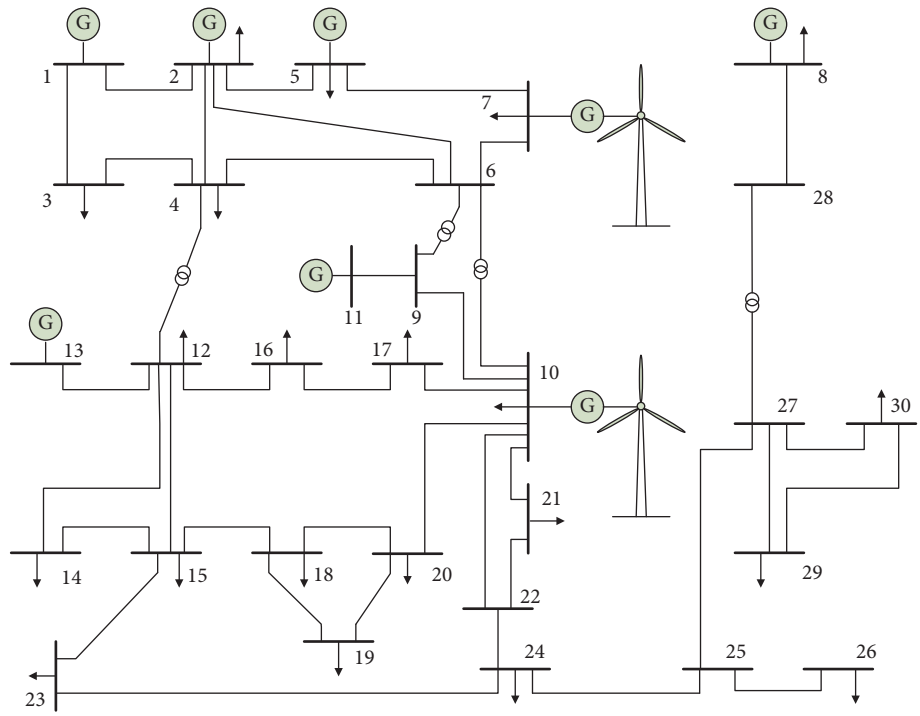

Figure 1: Diagram of the test system. *Note*. Figure 1 is reproduced from [41] (under the Creative Commons Attribution License/public domain). For the sake of simplicity, here it is assumed that the minimum and maximum congestion constraints in all transmission lines are ± 40 MW.

Table 1: Reduced 10 real-time WPOSs and their corresponding probabilities of these two WPPs.

WPOS index	WPP_1		WPP_2	
	Output power (MW)	Probability	Output power (MW)	Probability
1	58.0508	0.012	35.46167	0.082
2	48.3364	0.17	29.52738	0.21
3	80	0.129	48.86979	0.079
4	26.7776	0.265	16.35773	0.245
5	71.0479	0.052	43.40124	0.027
6	59.4902	0.036	36.34092	0.021
7	54.2876	0.062	33.16279	0.065
8	65.3884	0.05	39.94397	0.047
9	14.2553	0.083	8.708177	0.097
10	10.2223	0.141	6.244564	0.127

Table 2: Reduced joint real-time WPOSs and their corresponding probabilities.

Joint WPOS	WPP_1 output (MW)	WPP_2 output (MW)	Joint probability
1	80.0000	48.8698	0.0529
2	14.2553	39.9440	0.0202
3	14.2553	48.8698	0.0340
4	48.3364	6.2446	0.1121
5	48.3364	16.3577	0.2162
6	10.2223	36.3409	0.0154
7	26.7776	6.2446	0.1747
8	80.0000	29.5274	0.1406
9	10.2223	16.3577	0.1793
10	14.2553	6.2446	0.0547

Table 3: Parameters of GenCOs' bid functions.

Bus	NRGenCO	a_i (10^3 \$/MW^{2}h)	b_i (10^3 \$/MWh)	$P_{gi,\min}$ (MW)	$P_{gi,\max}$ (MW)
1	$GenCO_1$	0.2	20	0	80
2	$GenCO_2$	0.175	17.5	0	80
13	$GenCO_3$	0.625	10	0	50
22	$GenCO_4$	0.0834	32.5	0	55
23	$GenCO_5$	0.25	30	0	30
27	$GenCO_6$	0.25	30	0	30

mentioned in Section 5.3, after the same iterations can also be verified, which will not be demonstrated here due to the length of the article.

The obtained hour-ahead LMPs, $RTLMP_1$s, and $RTLMP_2$s of 30 buses after 3500 iterations in our GDCAC-based EM approach are depicted in Figure 6.

It can be seen in Figure 6 that hour-ahead LMPs of 30 buses are equal to each other after 3500 iterations, which is to say the hour-ahead dispatched results causes no congestion in any transmission line of this test system. In addition, there exist differences among $RTLMP_1$s and $RTLMP_2$s of 30 buses no matter with respect to the permitted upper or lower bound of power outputs by WPPs. Explanations of the above simulation results given by this paper can be expressed as when deviations between the (predicted) real-time outputs of WPPs and their hour-ahead scheduled ones exist, the power output of each generator connected bus and the power flow on each transmission line in this system are redistributed, in order for the system to tolerate the (predicted) real-time wind power deviations to a certain degree,

Table 4: Parameters of DisCOs' bid functions.

Bus	DisCO	c_j (10^3 \$/MW^{2}h)	d_j (10^3 \$/MWh)	$P_{dj,\min}$ (MW)	$P_{dj,\max}$ (MW)
2	$DisCO_1$	−0.5	60	16.7	35.7
3	$DisCO_2$	−0.5	50	0	16.4
4	$DisCO_3$	−0.5	48	2.6	21.6
7	$DisCO_4$	−0.5	80	17.8	36.8
8	$DisCO_5$	−0.5	50	25	44
10	$DisCO_6$	−0.5	45	0.8	19.8
12	$DisCO_7$	−0.5	60	6.2	25.2
14	$DisCO_8$	−0.5	50	1.2	20.2
15	$DisCO_9$	−0.5	52	3.2	22.2
16	$DisCO_{10}$	−0.5	40	0	17.5
17	$DisCO_{11}$	−0.5	53	4	23
18	$DisCO_{12}$	−0.5	55	0	17.2
19	$DisCO_{13}$	−0.5	44	4.5	23.5
20	$DisCO_{14}$	−0.5	60	0	16.2
21	$DisCO_{15}$	−0.5	45	12.5	31.5
23	$DisCO_{16}$	−0.5	45	0	17.2
24	$DisCO_{17}$	−0.5	42	3.7	22.7
26	$DisCO_{18}$	−0.5	57	0	17.5
29	$DisCO_{19}$	−0.5	44	0	16.4
30	$DisCO_{20}$	−0.5	50	5.6	24.6

Note: in order to ensure that all DisCOs do not lose in competition because of their obvious deference in revenue parameters from other DisCOs and to ensure the general balance between the sum of maximum outputs and demands in the market, a small part of parameters in the 4th and 6th column are slightly adjusted from [2].

and it is necessary for REDM in hour ahead to not only make each GenCO maintain a certain value of reserve capacity but also to reserve for each transmission line some additional transmission capacity to deal with the (predicted) real-time power flow changes.

5.3. Comparison of Our Proposed Approach and TBRL-Based Approach. In this section, for the purpose of approaches comparisons, our proposed GDCAC-based EM approach and the Q-learning-based EM approach are implemented on this test system, respectively. There are 3 learning scenarios (LSNs) which are set in this paper for simulation and comparisons. LSN.1 assumes that every WPP, GenCO, and DisCO in the market are the GDCAC-based agents with the continuous state and action spaces, which is the same as our proposed GDCAC-based approach mentioned in Section 5.2. LSN.2 assumes that WPP1 is a Q-learning-based agent with discrete state and action spaces, while other agents are the same as that in LSN.1, and LSN.3 assumes that every WPP, GenCO, and DisCO in the market is a Q-learning-based agent with discrete state and action spaces, which means the comparative Q-learning-based EM approach. Table 6 presents the related information while taking LSN.2 and LSN.3 into account, respectively. The parameters of the comparison of the Q-learning algorithm [19, 28] which use ε – greedy policy to balance exploration and exploitation in 3000 training iterations and greedy policy in 500 decision-making iterations are also listed in Table 6.

After 3500 iterations, (expected) profits of all agents and expected SWs in 3 LSNs are listed in Table 7.

From Table 7, the following can be inferred:

(1) After the same number of iterations, WPP_1's (expected) profit in LSN.1 is higher than that in LSN.2. This, to some extent, indicates one can get more profit by using our proposed GDCAC-based method to bid in EM than using the Q-learning based one within the same condition (namely, the same parameters values, number of iterations, and adaptive learning mechanism of other agents).

(2) After the same number of iterations, the expected SW in LSN.1 is higher than that in LSN.2 and the expected SW in LSN.2 is higher than that in LSN.3. This, to some extent, indicates that, with the increase in the number of agents using our proposed GDCAC-based method to bid in EM, the expected SW can be improved.

In conclusion, regarding to the (expected) profit of a specific agent and expected SW, it is obvious that our proposed GDCAC-based approach is better than the comparative Q-learning based one. The main reasons of this result are as follows: (1) the state and action spaces in the comparative Q-learning approach are discrete; otherwise, it will cause the curse of dimensionality, which is not the same as all continuous state and action spaces in the GDCAC-based approach; (2) the phenomenon of discrete state and action spaces makes it harder to find the globally optimal action solution in face of any given state than the continuous ones [28].

5.4. Comparison of Different Market Clearing Models in Our Proposed EM Approach. In this section two market clearing models embedded in our proposed GDCAC-based EM approach are compared in this test system, respectively. One is the REDM mentioned in Section 2.2 The other is the SEDM mentioned in [12]. Under SEDM, we still assume that it gives priority to the scheduling of the hour-ahead bidding outputs of WPPs in the system. Moreover, with respect to SEDM, it, *based on 10 joint real-time WPOSs listed in* Table 2, takes maximizing the expected SW as the objective function [12], and simultaneously considers hour ahead and (predicted) real-time transmission constraints etc. in order to obtain the optimal hour-ahead scheduled power output and demand results of all GenCOs and DisCOs. In this paper, we adopt expected SW, bidding power outputs, and permitted (predicted) real-time upper and lower bounds of power outputs of WPP1 and WPP2 obtained after 3500 iterations for comparison. The calculating results of these indices by using different dispatch models in our proposed EM approach are listed in Table 8.

From Table 8, the following can be inferred:

(1) After 3500 iterations, hour-ahead bidding outputs of WPPs within the REDM-embedded EM approach are significantly more than those within the SEDM-embedded EM approach. Explanations of this simulation result given by this paper can be expressed as follows: although both of the two dispatch models have endogenous penalty mechanism for wind

Table 5: Related information about the GDCAC algorithm.

Agents	EM state set ($/MWh) $\mathbf{X}_1$	EM state set ($/MWh) $\mathbf{X}_2$	Action set	ε	γ	α	β	σ_1	σ_2	m
WPP1	[0, 100]	[−10, 100]	[0, 80] MW	—	0.5	0.1	0.1	4	6	1
WPP2	[0, 100]	[−10,100]	[0, 50] MW	—	0.5	0.1	0.1	4	6	1
All GenCOs	[0, 100]	[−10, 100]	[1, 3]	—	0.5	0.1	0.1	4	6	1
All DisCOs	[0, 100]	[−10, 100]	(0, 1]	—	0.5	0.1	0.1	4	6	1
Central point sets in the Gauss radial basis function corresponding to $\mathbf{X}_1$ and $\mathbf{X}_2$		$\mathbf{X}_1$				{0, 5, 10, 15, . . ., 100}				
		$\mathbf{X}_2$				{−10, −5, 0, 5, . . ., 100}				

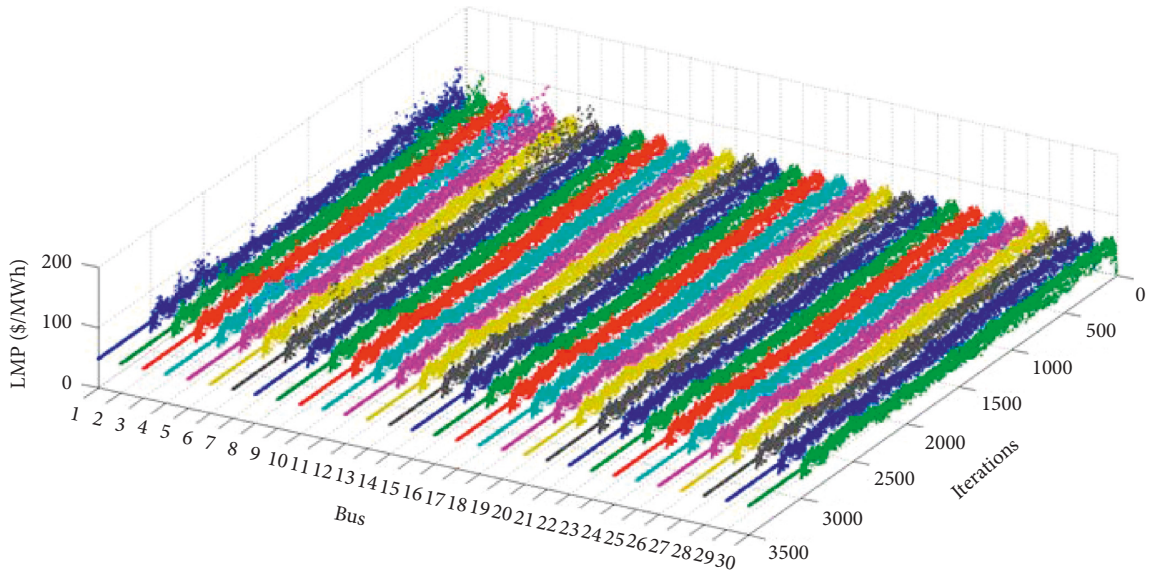

Figure 2: Dynamic adjusting process of hour-ahead LMPs in the GDCAC-based approach.

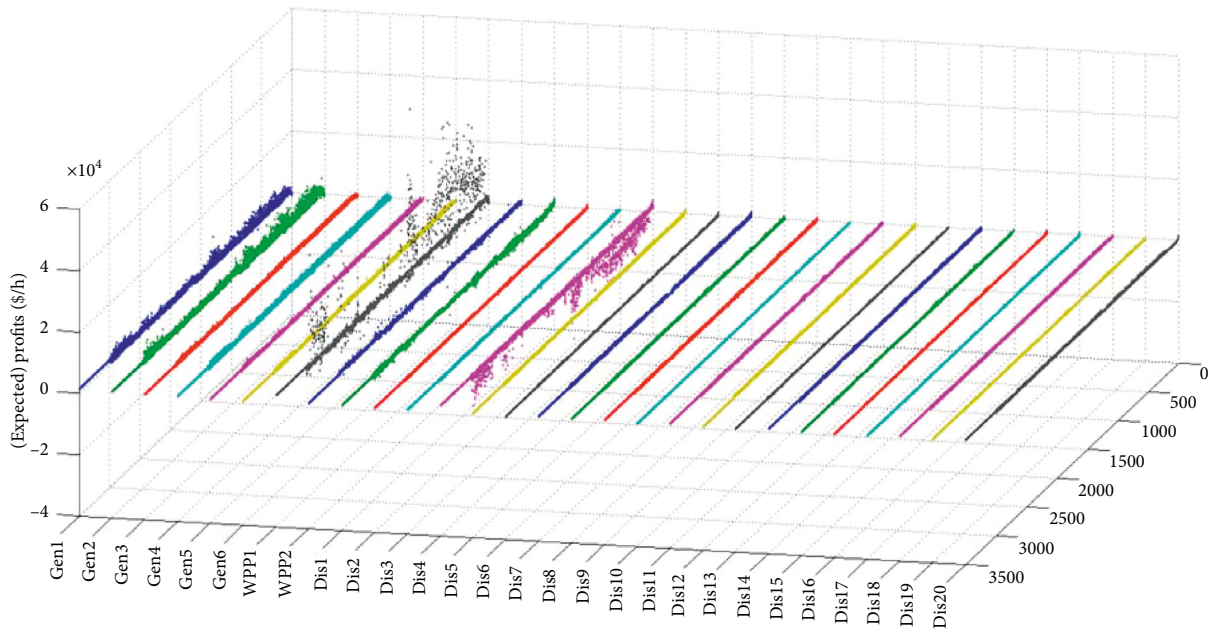

Figure 3: Dynamic adjusting process of (expected) profit of every agent in the GDCAC-based approach.

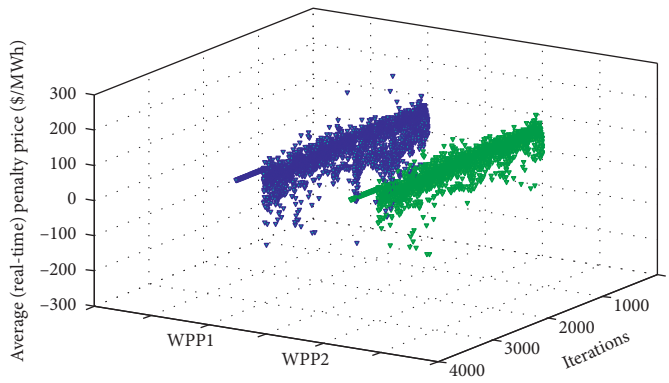

Figure 4: The dynamic adjusting process of (predicted) real-time LMPs connecting WPP1 and WPP2 (penalty prices charging from WPP1 and WPP2) in the GDCAC-based approach.

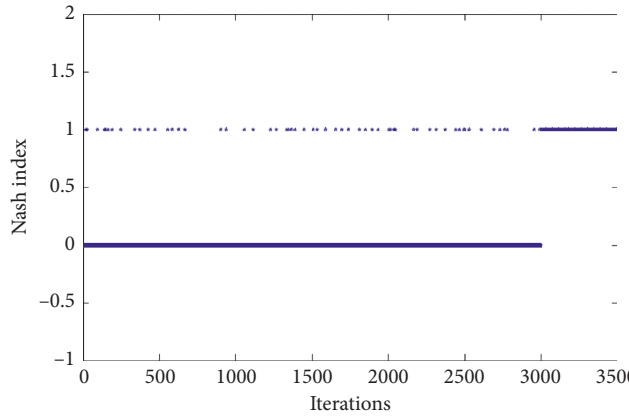

Figure 5: Adjusting process of Nash indices during 3500 iterations in our proposed GDCAC-based approach.

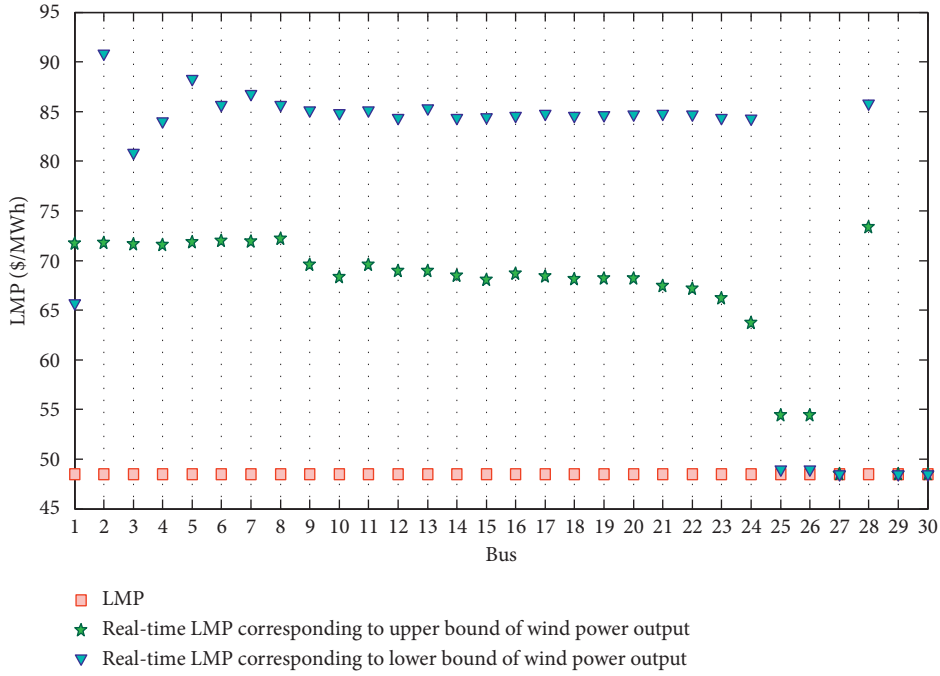

Figure 6: The obtained hour-ahead LMPs, $RTLMP_1$s, and $RTLMP_2$s of 30 buses after 3500 iterations in our GDCAC-based EM approach.

Table 6: Related information about LSN.2 and LSN.3.

Agents	EM state set ($/MWh) $\mathbf{X}_1$	EM state set ($/MWh) $\mathbf{X}_2$	Action set	ε	γ	α	β	σ_1	σ_2	m
LNS.2										
WPP 1	{0, 5, 10, . . ., 100}	{−10, −5,. . .,100}	{0, 5, 10, . . ., 80} MW	0.1	0.5	0.1	0.1	—	—	—
WPP 2	[0, 100]	[−10,100]	[0, 50] MW	—	0.5	0.1	0.1	5	5	1
All GenCOs	[0, 100]	[−10, 100]	[1, 3]	—	0.5	0.1	0.1	5	5	1
All DisCOs	[0, 100]	[−10, 100]	(0, 1]	—	0.5	0.1	0.1	5	5	1
Central point sets in the Gauss radial basis function corresponding to $\mathbf{X}_1$ and $\mathbf{X}_2$		$\mathbf{X}_1$		{0, 5, 10, 15, . . ., 100}						
		$\mathbf{X}_2$		{−10, −5, 0, 5, . . ., 100}						
LNS.3										
WPP_1	{0, 5, 10, . . ., 100}	{−10, −5, . . ., 100}	{0, 5, 10, . . ., 80} MW	0.1	0.5	0.1	0.1	—	—	—
WPP_2	{0, 5, 10, . . ., 100}	{−10, −5, . . ., 100}	{0, 5, . . ., 50} MW	0.1	0.5	0.1	0.1	—	—	—
All GenCOs	{0, 5, 10, . . ., 100}	{−10, −5, . . ., 100}	{1, 1.1, 1.2, . . ., 3}	0.1	0.5	0.1	0.1	—	—	—
All DisCOs	{0, 5, 10, . . ., 100}	{−10, −5, . . ., 100}	{0.1, 0.2, . . ., 1}	0.1	0.5	0.1	0.1	—	—	—

Table 7: (Expected) profit of every agent and expected SW results of 3 LSNs.

Agent	LSN.1		LSN.2		LSN.3	
	(Expected) profit ($/h)	Expected SW ($/h)	(Expected) profit ($/h)	Expected SW ($/h)	(Expected) profit ($/h)	Expected SW ($/h)
WPP_1	2769.989	17295.695	2195.839	16398.689	2110.501	14528.399
WPP_2	868.618		946.147		1018.178	
$GenCO_1$	1332.376		2464.719		1924.365	
$GenCO_2$	1049.794		1344.667		245.864	
$GenCO_3$	773.92611		877.657		883.918	
$GenCO_4$	716.12599		809.011		798.424	
$GenCO_5$	404.6452		213.969		293.889	
$GenCO_6$	380.873		455.478		296.238	
$DisCO_1$	624.266		143.059		157.595	
$DisCO_2$	368.282		218.333		202.086	
$DisCO_3$	392.174		250.374		245.412	
$DisCO_4$	2294.708		1978.318		2231.968	
$DisCO_5$	507.655		464.095		499.326	
$DisCO_6$	55.96419		43.599		49.346	
$DisCO_7$	623.75		518.597		551.826	
$DisCO_8$	336.628		267.704		38.679	
$DisCO_9$	532.938		428.904		457.055	
$DisCO_{10}$	139.583		82.187		97.215	
$DisCO_{11}$	582.043		505.382		504.764	
$DisCO_{12}$	118.287		102.379		93.173	
$DisCO_{13}$	112.231		87.021		98.471	
$DisCO_{14}$	607.609		540.468		544.891	
$DisCO_{15}$	199.147		162.117		113.276	
$DisCO_{16}$	174.468		137.546		123.667	
$DisCO_{17}$	50.435		35.279		41.438	
$DisCO_{18}$	571.921		491.382		545.126	
$DisCO_{19}$	205.266		181.887		168.852	
$DisCO_{20}$	501.993		452.571		192.856	

Table 8: Calculating results of expected SW, bidding power outputs, and permitted (predicted) real-time upper and lower bounds of power outputs considering different dispatch models.

Iterm	Expected SW ($/h)	Bidding power output (MW)		Permitted upper and lower bounds of power output (MW)			
		WPP1	WPP2	WPP1		WPP2	
REDM	17295.695	51.3587	38.4874	65.2964	42.4196	47.4874	25.8710
SEDM	13587.68	34.5695	17.36686	Meeting transmission constraints under the 10 joint real-time WPOSs listed in Table 2			

power output deviations, which can affect the dynamic adjustment process of bidding power outputs of WPPs, the permitted upper and lower bounds are dynamically adjusted to adapt for the hour-ahead bidding power output of each WPP within the REDM-embedded EM approach while in each iteration of the SEDM-embedded EM approach, the hour-ahead bidding power output of each WPP is required to meet the (predicted) real-time transmission constraints corresponding to 10 WPOSs listed in Table 2. Therefore, WPPs in REDM-embedded EM approach can adjust their bidding power outputs to relatively high levels while those in SEDM embedded one are more inclined to adjust their bidding power outputs to the average level of the 10 WPOSs listed in Table 2 in order to avoid the risks of (expected) profit decline caused by larger power deviations.

(2) After 3500 iterations, expected SW obtained from the REDM-embedded EM approach is significantly more than that obtained from the SEDM embedded one. Explanations of this simulation result given by this paper can be expressed as follows: in order to meet all (predicted) real-time transmission constraints corresponding to 10 obviously different WPOSs listed in Table 2, more reserve transmission capacity in each transmission line are required by using SEDM, which may force out more scheduled power outputs and demands by GenCOs and DisCOs than REDM under the same bidding power outputs of WPPs.

(3) Moreover, other than the scheduled hour-ahead power outputs and demands of all GenCOs and DisCOs, it can also be scheduled by REDM the permitted upper and lower bounds of (predicted) real-time power output of each WPP. If a WPP's (predicted) natural power output exceeds its permitted power output interval which is defined by its scheduled permitted upper and lower bounds, its (predicted) real-time power output can be adjusted equal to the adjacent bound by conducting pitch control or using storage equipment [37]. This characteristic of REDM means continuous arbitrary changes of the real-time power output within the corresponding permitted power output interval by each WPP would not cause congestion in any transmission line in the system. However, by using SEDM, only the hour-ahead power outputs and demands of all GenCOs and DisCOs can be scheduled. Although a scheduled output by REDM can meet all (predicted) real-time transmission constraints corresponding to 10 WPOSs listed in Table 2, it cannot be guaranteed that real-time power outputs of WPPs other than any of the WPOS listed in Table 2 also would not cause congestion in any transmission line in the system, and WPPs would not know their corresponding permitted power deviation intervals according to which they can adjust their natural power outputs by conducting pitch control or using storage equipment. Hence, the SEDM-embedded EM approach is less conducive to the wind power accommodation than the REDM embedded one.

Therefore, no matter with respect of economy or reliability, REDM has a lot of advantages over SEDM when being embedded in the EM modeling approach.

6. Conclusion

In this paper, considering strategic interactions among WPPs, GenCOs, and DisCOs, we have proposed a GDCAC-based EM modeling approach with REDM embedded. Simulation results have verified the feasibility and the scientific nature of our proposed approach, and some conclusions can be drawn as follows:

(1) With our proposed GDCAC-based EM approach, the simulated bidding process after enough training and decision-making iterations can reach dynamic stability which has been tested and verified as the NE result.

(2) Our simulation on the IEEE 30 bus test system with 28 participants takes only 1.17 minutes to reach the final result. That is to say, the time complexity of our proposed approach is relatively low so that we can extend it to the modeling and simulation of more realistic and more complex EM system.

(3) Our proposed GDCAC-based EM approach is superior to the TBRL- (Q-learning-) based approach in terms of increasing the profit of a specific agent and expected SW. The main reason is that only TBRL algorithm can be used to analyze Markov decision-making problems with discrete state and action spaces.

(4) The obtained bidding results also reveal that in, the premise of maintaining relatively high wind power accommodation ability of the system, the overall SW can be improved by using REDM as the market clearing model when comparing with SEDM. This, to some extent, has verified the robustness against wind power fluctuations, the reliability about scheduling results, and the market operation economy of our proposed EM approach with REDM embedded.

Moreover, on the one hand, our proposed approach can provide a bidding decision-making tool for WPPs, GenCOs, and DisCOs to get more profits in EM. On the other hand, our proposed approach can also provide an economic and operational analysis tool for promoting the development of renewable resources.

Appendix

A. Formulations for Hour-Ahead LMP

The hour-ahead LMP for energy credit and load payment at bus Gz (or Dz) can be calculated as

$$\mathrm{LMP}_{\mathrm{Gz}} = \frac{\partial L}{\partial P^{*}_{\mathrm{Gz},t}} = \lambda - \sum_{l} sf_{l,\mathrm{Gz}}\left(\eta_{l1} - \eta_{l2}\right), \tag{A.1}$$

where λ, η_{l1}, and η_{l2} represent the dual variables of equations (9), (12), and (13), respectively. L represents the generalized Lagrange function of model (equations (8)–(19)).

B. Discussion on the Reformulation of Constraints (16–18) to (21–25)

From equations (16)–(18), it is obvious that $\sum_{z=1}^{Z} P^{(r)}_{\mathrm{Gz},t} \times sf_{l,\mathrm{Gz}} - \sum_{z=1}^{Z} P^{*}_{Dz,t} \times sf_{l,Dz}$ increases with the increase of $P^{(r)}_{wi',t} \in [P^{\mathrm{lb}}_{wi',t}, P^{\mathrm{ub}}_{wi',t}]$ ($i' \in \mathrm{BUS}_z$) and decreases with the decrease of $P^{(r)}_{wi',t} \in [P^{\mathrm{lb}}_{wi',t}, P^{\mathrm{ub}}_{wi',t}]$ ($i' \in \mathrm{BUS}_z$). This is to say the violation of real-time constraints is most likely to happen when $P^{(r)}_{wi',t} = P^{\mathrm{lb}}_{wi',t}$ or $P^{(r)}_{wi',t} = P^{\mathrm{ub}}_{wi',t}$. Hence, for the purpose of maintaining robustness, we can solve the abovementioned REDM by replacing $P^{(r)}_{wi,t}$ with $P^{\mathrm{ub}}_{wi,t}$ and $P^{\mathrm{lb}}_{wi,t}$, respectively, and generating new (predicted) real-time balancing and transmission constraints as follows:

$$P^{(r1)}_{gj,t} = P^{*}_{gj,t} - \alpha_{j,t} \sum_{i=1}^{N_w} \left(P^{\mathrm{ub}}_{wi,t} - P^{*}_{wi,t}\right) \in \cdot \left[P_{gj,\min}, P_{gj,\max}\right], \quad \forall j, \tag{B.1}$$

$$P_{l,\min} \le \sum_{z=1}^{Z} P^{(r1)}_{\mathrm{Gz},t} \times sf_{l,Gz} - \sum_{z=1}^{Z} P^{*}_{Dz,t} \times sf_{l,Dz} \le P_{l,\max}, \forall l, \tag{B.2}$$

$$P^{(r1)}_{\mathrm{Gz},t} = \sum_{i' \in \mathrm{BUS}_z} P^{\mathrm{ub}}_{wi',t} + \sum_{j' \in \mathrm{BUS}_z} P^{(r1)}_{gj',t}, \tag{B.3}$$

$$P^{(r2)}_{gj,t} = P^{*}_{gj,t} - \alpha_{j,t} \sum_{i=1}^{N_w} \left(P^{\mathrm{lb}}_{wi,t} - P^{*}_{wi,t}\right) \in \cdot \left[P_{gj,\min}, P_{gj,\max}\right], \quad \forall j, \tag{B.4}$$

$$P_{l,\min} \le \sum_{z=1}^{Z} P^{(r2)}_{\mathrm{Gz},t} \times sf_{l,\mathrm{Gz}} - \sum_{z=1}^{Z} P^{*}_{\mathrm{Dz},t} \times sf_{l,\mathrm{Dz}} \le P_{l,\max}, \quad \forall l, \tag{B.5}$$

$$P^{(r2)}_{\mathrm{Gz},t} = \sum_{i' \in \mathrm{BUS}_z} P^{\mathrm{lb}}_{wi',t} + \sum_{j' \in \mathrm{BUS}_z} P^{(r2)}_{gj',t}. \tag{B.6}$$

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the fund project of the Central University of North China Electric Power University under 2017XS113.

References

[1] J. Aghaei, M. A. Akbari, A. Roosta, M. Gitizadeh, and T. Niknam, "Integrated renewable-conventional generation expansion planning using multiobjective framework," *IET Generation, Transmission & Distribution*, vol. 6, no. 8, pp. 773–784, 2012.

[2] M. R. Salehizadeh and S. Soltaniyan, "Application of fuzzy Q-learning for electricity market modeling by considering renewable power penetration," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 1172–1181, 2016.

[3] M. Vilim and A. Botterud, "Wind power bidding in electricity markets with high wind penetration," *Applied Energy*, vol. 118, pp. 141–155, 2014.

[4] M. Bueno-Lorenzo, M. Á. Moreno, and J. Usaola, "Analysis of the imbalance price scheme in the Spanish electricity market: a wind power test case," *Energy Policy*, vol. 62, pp. 1010–1019, 2013.

[5] K. C. Sharma, R. Bhakar, and H. P. Tiwari, "Strategic bidding for wind power producers in electricity markets," *Energy Conversion and Management*, vol. 86, pp. 259–267, 2014.

[6] K. W. Ravnaas, G. Doorman, and H. Farahmand, "Optimal wind farm bids under different balancing market arrangements," in *Proceedings of 2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems*, pp. 30–35, Singapore, June 2010.

[7] J. Matevosyan, M. Olsson, and L. Söder, "Hydropower planning coordinated with wind power in areas with congestion problems for trading on the spot and the regulating market," *Electric Power Systems Research*, vol. 79, no. 1, pp. 39–48, 2009.

[8] G. Li and J. Shi, "Agent-based modeling for trading wind power with uncertainty in the day-ahead wholesale electricity markets of single-sided auctions," *Applied Energy*, vol. 99, pp. 13–22, 2012.

[9] R. Laia, H. M. I. Pousinho, R. Melíco, and V. M. F. Mendes, "Bidding strategy of wind-thermal energy producers," *Renewable Energy*, vol. 99, pp. 673–681, 2016.

[10] J. P. Chaves-Ávila, R. A. Hakvoort, and A. Ramos, "The impact of European balancing rules on wind power economics and on short-term bidding strategies," *Energy Policy*, vol. 68, pp. 383–393, 2014.

[11] Y. Xiao, X. Wang, X. Wang, C. Dang, and M. Lu, "Behavior analysis of wind power producer in electricity market," *Applied Energy*, vol. 171, pp. 325–335, 2016.

[12] M. Lei, J. Zhang, X. Dong, and J. J. Ye, "Modeling the bids of wind power producers in the day-ahead market with stochastic market clearing," *Sustainable Energy Technologies and Assessments*, vol. 16, pp. 151–161, 2016.

[13] T. Dai and W. Qiao, "Trading wind power in a competitive electricity market using stochastic programing and game theory," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 3, pp. 805–815, 2013.

[14] J. Geng, K. Zhang, Z. Yang, C. Chen, and Y. Zheng, "Proposal of block bidding for large-scale wind power energy," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 2776–2781, 2014.

[15] M. Huang, X. Wang, and S. Zhang, "Analysis of an electricity market equilibrium model with penalties for wind power's

bidding deviation," in *Proceedings of 2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*, pp. 35–40, Changsha, China, November 2015.

[16] K. Bhaskar and S. N. Singh, "Wind power bidding strategy in a day-ahead electricity market," in *Proceedings of 2014 IEEE PES General Meeting Conference & Exposition*, pp. 1–5, National Harbor, MD, USA, July 2014.

[17] A. Giannitrapani, S. Paoletti, A. Vicino, and D. Zarrilli, "Wind power bidding in a soft penalty market," in *Proceedings of 52nd IEEE Conference on Decision and Control*, pp. 1013–1018, Florence, Italy, December 2013.

[18] Q. Wang, J. Wang, and Y. Guan, "Wind power bidding based on chance-constrained optimization," in *Proceedings of 2011 IEEE Power and Energy Society General Meeting*, pp. 1-2, Detroit, MI, USA, July 2011.

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

[20] G. Santos, R. Fernandes, T. Pinto, I. Praça, Z. Vale, and H. Morais, "MASCEM: EPEX SPOT day-ahead market integration and simulation," in *Proceedings of 2015 18th International Conference on Intelligent System Application to Power Systems (ISAP)*, pp. 1–5, Porto, Portugal, September 2015.

[21] A. Y. F. Lau, D. Srinivasan, and T. Reindl, "A reinforcement learning algorithm developed to model GenCo strategic bidding behavior in multidimensional and continuous state and action spaces," in *Proceedings of 2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 116–123, Singapore, April 2013.

[22] H. Li and L. Tesfatsion, "The AMES wholesale power market test bed: a computational laboratory for research, teaching, and training," in *Proceedings of 2009 IEEE Power & Energy Society General Meeting*, pp. 1–8, Calgary, Canada, July 2009.

[23] G. Conzelmann, G. Boyd, V. Koritarov, and T. Veselka, "Multi-agent power market simulation using EMCAS," in *Proceedings of IEEE Power Engineering Society General Meeting*, vol. 3, pp. 2829–2834, San Francisco, CA, USA, January 2005.

[24] M. Mahvi and M. M. Ardehali, "Optimal bidding strategy in a competitive electricity market based on agent-based approach and numerical sensitivity analysis," *Energy*, vol. 36, no. 11, pp. 6367–6374, 2011.

[25] Z. Liu, J. Yan, Y. Shi, K. Zhu, and G. Pu, "Multi-agent based experimental analysis on bidding mechanism in electricity auction markets," *International Journal of Electrical Power & Energy Systems*, vol. 43, no. 1, pp. 696–702, 2012.

[26] T. Bach, "Using reinforcement learning to study the features of the participants' behavior in wholesale power market," Doctor's Degree, Hunan University, Changsha, China, 2013.

[27] N. P. Ziogos and A. C. Tellidou, "An agent-based FTR auction simulator," *Electric Power Systems Research*, vol. 81, no. 7, pp. 1239–1246, 2011.

[28] H. Zhao, Y. Wang, S. Guo, M. Zhao, and C. Zhang, "Application of gradient descent continuous actor-critic algorithm for double-side day-ahead electricity market modeling," *Energies*, vol. 9, no. 9, p. 725, 2016.

[29] M. Shafie-khah, M. Parsa Moghaddam, and M. K. Sheikh-El-Eslami, "Development of a virtual power market model to investigate strategic and collusive behavior of market players," *Energy Policy*, vol. 61, pp. 717–728, 2013.

[30] D. Dallinger and M. Wietschel, "Grid integration of intermittent renewable energy sources using price-responsive plug-in electric vehicles," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 3370–3382, 2012.

[31] M. Reeg, W. Hauser, S. Wassermann et al., "AMIRIS: an agent-based simulation model for the analysis of different support schemes and their effects on actors involved in the integration of renewable energies into energy markets," in *Proceedings of 2012 23rd International Workshop on Database and Expert Systems Applications*, pp. 339–344, Vienna, Austria, September 2012.

[32] P. Zamani-Dehkordi, L. Rakai, and H. Zareipour, "Deciding on the support schemes for upcoming wind farms in competitive electricity markets," *Energy*, vol. 116, pp. 8–19, 2016.

[33] T. Haring, G. Andersson, and J. Lygeros, "Evaluating market designs in power systems with high wind penetration," in *Proceedings of the 2012 9th internatioanal conference on the European energy market (EEM)*, pp. 1–8, Florence, Italy, May 2012.

[34] M. Paschen, "Dynamic analysis of the German day-ahead electricity spot market," *Energy Economics*, vol. 59, pp. 118–128, 2016.

[35] T. Soares, G. Santos, T. Pinto, H. Morais, P. Pinson, and Z. Vale, "Analysis of strategic wind power participation in energy market using MASCEM simulator," in *Proceedings of the 2015 18th International Conference on Intelligent System Application to Power Systems (ISAP)*, pp. 1–6, Porto, Portugal, September 2015.

[36] J. Abrell and F. Kunz, "Integrating intermittent renewable wind generation-a stochastic multi-market electricity model for the European electricity market," *Networks and Spatial Economics*, vol. 15, no. 1, pp. 117–147, 2013.

[37] Z. Li, W. Wu, and B. Zhang, "A robust interval economic dispatch method accommodating; large-scale wind power generation: part one dispatch scheme and mathematical model," *Automation of Electric Power System*, vol. 38, pp. 33–39, 2014.

[38] G. Chen, "Research on value function approximation methods in reinforcement learning," Master's Degree, Soochow University, Suzhou, China, 2014.

[39] K. C. Sharma, P. Jain, and R. Bhakar, "Wind power scenario generation and reduction in stochastic programming framework," *Electric Power Components and Systems*, vol. 41, no. 3, pp. 271–285, 2013.

[40] J. Li, H. Sun, J. Wen et al., "A two-dimensional optima technology for constructing wind power time series scenarios," *Proceedings of the CSEE*, vol. 34, pp. 2544–2551, 2014.

[41] H. Zhao, Y. Wang, M. Zhao, C. Sun, and Q. Tan, "Application of gradient descent continuous actor-critic algorithm for bilateral spot electricity market modeling considering renewable power penetration," *Algorithms*, vol. 10, no. 2, p. 53, 2017.