energies

MDPI

Article

Application of a Gradient Descent Continuous Actor-Critic Algorithm for Double-Side Day-Ahead Electricity Market Modeling

Huiru Zhao, Yuwei Wang, Sen Guo *, Mingrui Zhao and Chao Zhang

School of Economics and Management, North China Electric Power University, Beijing 102206, China; zhaohuiru@ncepu.edu.cn (H.Z.); wangyuwei2010@126.com (Y.W.); 1142206037@ncepu.edu.cn (M.Z.); superzhang_hd@ncepu.edu.cn (C.Z.)

* Correspondence: guosen@ncepu.edu.cn; Tel.: +86-158-1142-4568

Academic Editor: Robert Lundmark

Received: 1 June 2016; Accepted: 30 August 2016; Published: 9 September 2016

Abstract: An important goal of China's electric power system reform is to create a double-side day-ahead wholesale electricity market in the future, where the suppliers (represented by GenCOs) and demanders (represented by DisCOs) compete simultaneously with each other in one market. Therefore, modeling and simulating the dynamic bidding process and the equilibrium in the double-side day-ahead electricity market scientifically is not only important to some developed countries, but also to China to provide a bidding decision-making tool to help GenCOs and DisCOs obtain more profits in market competition. Meanwhile, it can also provide an economic analysis tool to help government officials design the proper market mechanisms and policies. The traditional dynamic game model and table-based reinforcement learning algorithm have already been employed in the day-ahead electricity market modeling. However, those models are based on some assumptions, such as taking the probability distribution function of market clearing price (*MCP*) and each rival's bidding strategy as common knowledge (in dynamic game market models), and assuming the discrete state and action sets of every agent (in table-based reinforcement learning market models), which are no longer applicable in a realistic situation. In this paper, a modified reinforcement learning method, called gradient descent continuous Actor-Critic (GDCAC) algorithm was employed in the double-side day-ahead electricity market modeling and simulation. This algorithm can not only get rid of the abovementioned unrealistic assumptions, but also cope with the Markov decision-making process with continuous state and action sets just like the real electricity market. Meanwhile, the time complexity of our proposed model is only O(n). The simulation result of employing the proposed model in the double-side day-ahead electricity market shows the superiority of our approach in terms of participant's profit or social welfare compared with traditional reinforcement learning methods.

Keywords: bidding strategy; double-side day-ahead electricity market; gradient descent continuous Actor-Critic (GDCAC) algorithm; reinforcement learning; market clearing price (*MCP*)

1. Introduction

1.1. Background and Motivation

In China, with the development of the economy and society, electricity consumption has increased rapidly in recent years [1]. In order to meet the economic and social development need of an effective power supply, besides the continuous power system construction, the electricity industry in China has undergone a series of restructuring and changes in the last decades, similar to many other countries around the world. The direct objective of the electricity market restructuring in many countries, including China, is to enhance the competition and improve the operational efficiency [2]. Before 2015,

there have already been many regulatory reforms in China's electricity sector, which mainly include the Investment Decentralization in 1986, the first unbundling reform (the unbundling of Government Admin and Business Operation) in 1997, and the second unbundling reform (the unbundling of Electricity Generation and Transmission) in 2002 [1,3].

However, the first two reforms only provided weak incentives because most power plants were still dominated with state ownership participation so that they only face soft budget constraints. Furthermore, the entire electricity industry was still operated vertically by the State Power Corporation (SPC) which contains sectors of generation, transmission and distribution. In 2002, SPC was separated into 11 new corporations including five power generation corporation groups, two power grid corporations, and four auxiliary corporations [2], which is known as the third reform aiming at increasing competition in the electricity industry in China. Although some issues such as mandatory plan systems, the integration of government administration with enterprises, and the integration of plants with the power grid have been basically solved, there are still many unsolved issues which seriously hinder the efficiency of the electricity industry in China: firstly, a transaction mechanism, which plays the decisive role of the market in the difficult to realize allocation of resources is missing; secondly, the pricing relationship doesn't line up, which is the consequence of market pricing mechanism deficiencies; thirdly, the transformation of the government function is not in place so that many planning and coordination works concerning the electricity market are hard to implement; fourthly, the development mechanism is unsound, making the development and utilization of renewable energy very difficult; finally, the legislative work is lagging, which hinders the deregulation of the electricity industry.

To get over those problems listed above, in March 2015 the Communist Party of China (CPC) Central Committee and the State Council issued the policy paper "Several opinions on further deepening the reform of the electric power system" (which is also known as Chinese State Council (2015) No. 9 Policy Paper in China) in which one of the general lines of the future reform of the electricity industry is described as that based on further improving the decoupling of government administration of enterprises, of plants with the power grid and the main-auxiliary separation, freeing the consumption side option, and establishing effective electricity markets with double-side competitive transaction mechanisms in many regions.

Just like many developed countries around the world which have already experienced restructuring of their power systems, in the future the electricity markets which will be established in China based on the general reform lines abstracted from the Chinese State Council (2015) No. 9 Policy Paper and Reference [4] can be classified according to many standards. For example, considering different traders, a market existing between the generating companies and distribution companies, retailers or large consumers is called the wholesale marketplace. A market existing between retailers, distribution companies and end users is called the retail marketplace. Considering different durations of the transaction, the electricity market can be classified as forward market, spot market and auxiliary market [4–6].

The day-ahead wholesale electricity market is one of the most common forms of spot market in many countries. In a double-side day-ahead electricity market, the sellers (i.e., generating companies, which we call GenCOs for the sake of description convenienc,) and the buyers (i.e., distribution companies, retailers or large consumers, for the sake of description convenience and without loss of generality, we indiscriminately call all of them DisCOs) are required to submit bids for selling and buying energy to an independent system operator (ISO) for every time interval of the next day based on the supply and demand curves, respectively. After receiving all time interval biddings for the next day from GenCOs and DisCOs, the aggregated timely supply and demand bidding curves can be then constructed by the ISO to determine the market clearing price (*MCP*). Meanwhile, the corresponding supply and demand schedules for every time interval in the next day can also be determined by ISO, in which the constraints such as the security and stability of transmission network and the power balance of power system must be taken into account [7]. GenCOs and DisCOs get paid in accordance

with the *MCP* and their accepted schedules, or according to their bid (pay as bid, (PAB)). In this paper, we take the *MCP* mechanism into account.

Generally, the restructured day-ahead wholesale electricity market can be defined as an imperfect competitive market or more accurately an oligopoly market, due to the limited number of power producers, long period of power plant construction, large scale of capital investment, transmission constraints, and transmission losses [7,8]. This imperfect competitive or oligopoly nature of the electricity industry makes GenCOs and DisCOs bid strategically in a day-ahead market to obtain more profits. For example, due to the oligopolistic nature, a GenCO has the market power to bid at a higher price than its marginal cost, which is defined as the bidding strategy of the GenCO. DisCOs have the market power to bid at a lower price than their marginal revenue, which is defined as the bidding strategy of the DisCO. Hence one can see that different bidding strategies of GenCOs and DisCOs determine different shapes of their supply and demand curves which in turn affect *MCP*, the schedules of market, the profits of all GenCOs and DisCOs, and even the welfare of society. Therefore, to scientifically model and simulate the dynamic bidding process and market equilibrium in the double-side day-ahead electricity market is not only of importance to some developed countries, but also to China, so that for participants (GenCOs or DisCOs), it can provide a bidding decision-making tool to obtain more profits in market competition, and for the government, it can provide an economic analysis tool to help design proper market mechanisms and policies.

1.2. Literature Review and Main Contributions

There are many papers related to modeling and simulating the dynamic bidding process or equilibrium of the day-ahead electricity market, which generally can be divided into two kinds: single-side studies and double-side studies. In single-side studies, researching generation-side bidding strategies and equilibrium represents the main consideration. A supply function equilibrium (SFE) game model for modeling GenCO's strategic bids was presented by Al-Agtash et al. [8], where the competition among all GenCOs with imperfect information about their rivals and the transmission constraints were taken into consideration. In reference [9], Damoun et al. proposed a direct SFE-based approach to compute the robust Nash strategies for GenCOs in the spot electricity market without taking transmission constraints into consideration. In the study performed by Alberto et al. [10], the Nash equilibrium of the single-side day-ahead market was analyzed with a static game model considering the transmission constraints. Gao, et al. [11] researched on how to find the optimal bidding strategy of a GenCO in the single-side day-ahead electricity market, based on the parametric linear programming method and with the assumption that all GenCOs in the day-ahead market pursue profit maximization. In the papers by Kumar et al. [12] and Wang [13], every GenCO in the single-side market optimizes its bidding strategy by evaluating the strategy probability distributions of its rivals with the information about their cost functions (complete information) and their strategies from last game iteration (but imperfect information). The dynamic evolution process of GenCOs' bidding strategy was simulated by shuffled frog leaping algorithm (SFLA) [12] and genetic algorithm (GA) [13], respectively. Liu et al. [14] reported an incentive bidding mechanism in which the semi-randomized approach is applied to model the information disturbance in the electricity auction markets. Nojavan et al. [15], used the information gap decision theory to model the market price with severe uncertainty so that an optimal bidding strategy can be determined for the day-ahead market. In the study performed by Wen and David [16], a GenCO estimated other rivals' bidding strategies using Monte Carlo simulation, and the stochastic optimization model of GenCOs for strategic bids pursuing profit maximization was established. In the study performed by Kumar et al. [17], from a GenCO's point of view, all of other participants' bidding strategy variables were taken as random variables which obey a Gaussian distribution, and the dynamic game process in the single-side day ahead electricity market was solved by the fuzzy adaptive gravitational search algorithm. All the methods listed above are actually based on game theory. Azadeh et al. [18] simulated the dynamic adjustment process of GenCOs in day-ahead market through multi-agent-based method. In the literature from Rahimiyan et al. [19],

a GenCO's optimal bidding strategy problem was modeled and simulated by the Q-learning algorithm considering discrete state as well as action sets and the game model-based approach, respectively. Comparison of those two methods confirms the superiority of Q-learning in this issue.

In the aspect of double-side studies, research on simultaneous generation-side and consumption-side bidding strategies and the market equilibrium represent the main consideration. Shivaie et al. [20] proposed an environmental/techno-economic game approach for bidding strategies of GenCOs and DisCOs in a security-constrained day-ahead electricity market, and the dynamic process of bidding adjustment was simulated by a bi-level harmony search algorithm. In Reference [20], every GenCO and DisCO was assumed to have imperfect information about ongoing strategies of its rivals, but complete information about historical ones so that the parameters in optimization model of every GenCO or DisCO were estimated as the historical strategies of its rivals. In the study by Menniti et al. [21], an evolutionary game model only to simulate the behaviors of the generation-side was proposed, and the modeling approaches of this paper can also be extended to the consumption-side issue. The classical evolutionary game theory can only solve problems with a discrete strategy set, which is not in line with the actual situation in the day-ahead electricity market. Ladjici et al. [22,23] proposed a stochastic optimization model for GenCOs, which is also suitable for DisCOs in the day-ahead electricity market, and the evolutionary processes of strategies within continuous intervals were simulated by using competitive co-evolutionary algorithms drawn lessons from the classical evolutionary game theory mentioned above. However, the assumption of these two papers is that the strategy probability distribution functions of every participant are taken as common knowledge in the marketing game.

From the experiences in some developed countries, common characteristics of GenCOs and DisCOs in the deregulated day-ahead electricity market include:

(1) Every participant (GenCO or DisCO) has no idea about what the cost and revenue functions of all its rivals are;
(2) Every participant has no idea about what the ongoing and historical strategies of all its rivals and those strategies' real probability distribution functions are in the day-ahead market every day;
(3) The common information published by ISO after the completion of bidding and market clearing every day is only the *MCP* of every time intervals of the next day. One participant can only be noticed by ISO its own producing or consuming schedules in every time interval of the next day;
(4) Every participant can adjust its bidding strategy within a continuous interval of values, and the *MCP* also changes within a continuous interval of values over time.

Considering the above provisions from Equations (1)–(3), the modeling approaches of all literatures introduced above except for reference [19] are not quite suitable for the actual situation of the day-ahead electricity market. That is because every participant in market neither has information about the cost and revenue functions of all its rivals (then do not know their profits) nor has information about the ongoing and historical strategies of all rivals, or even the probability distribution functions of strategy choosing by rivals. The modeling and simulating approach in [19] does not require that information, and an agent representing a participant learns the best strategy while meeting with a certain state of the market (the *MCP* formed in last iteration) through its experiences of the past. In the literature from Salehizadeh [24], an agent-based fuzzy Q-learning algorithm was used for modeling the dynamic bidding strategy adjustment of GenCOs in a spot electricity market by considering renewable power penetration, in which the fuzzy rule was used to define the continuously changing states of renewable power production. In the literature from Thanhquy [25], the participants' dynamic behaviors in single-side day-ahead electricity markets were modeled by Q-learning with greedy, ε-greedy, and Boltzmann ε-greedy action decision method, respectively. The comparison result shows that with Boltzmann ε-greedy decision method, the participants in the day-ahead market can receive more profits after sufficient learning iterations, because the value of the temperature variable in the Boltzmann action choosing probability distribution function of every agent (participant) can be

adjusted as the iteration proceeds. Similar studies in electricity market simulation and other areas can also be found in [26–36].

Taking provisions (1)–(4) into consideration, the methods both in [8–18,20–23] and in [19,24–33] are not quite suitable for modeling and simulating the practical day-ahead electricity market exactly, the reason of which is that in the real day-ahead electricity market, every participant can adjust its bidding strategy within a continuous interval of values, but the literatures suitable for provisions (1)–(3) have assumed that the sets for alternative actions (e.g., bidding strategies) or potential states (e.g., historical *MCP*) are discrete.

As far as this paper can tell, there is no literature in this area which proposes a feasible method which can model and simulate the double-side day-ahead electricity market in accordance with all four provisions listed above simultaneously. Therefore, the objective of this paper is to establish a suitable and feasible method that can model and simulate the dynamic bidding adjustment process and equilibrium of a double-side day-ahead electricity market scientifically, in which every participant has the ability to gradually learn the decision-making conditions adaptively through imperfect and incomplete information (i.e., satisfying provisions (1)–(4) simultaneously) and in its repeated bidding process. The participants in the electricity market can use this decision-making tool to obtain more profits in a competitive environment. Meanwhile, the government can use the simulation result to test the effects of diverse policies implemented in electricity market.

The rest of the paper is organized as follows: in Section 2, the agent-based double-side day-ahead electricity market model is established mathematically. Participants' bidding and market clearing mechanisms are also discussed in this section. In Section 3, the mathematical principles of the gradient descent continuous Actor-Critic (GDCAC) algorithm [37], which can model and simulate the dynamic bidding strategy adjustment process of GenCOs or DisCOs in a double-side day-ahead electricity market while simultaneously conforming to provisions (1)–(4), is introduced in details. Then, our proposed methodology for modeling and simulating the dynamic bidding process of GenCOs or DisCOs in a double-side day-ahead electricity market conforming to provisions (1)–(4) simultaneously is established based on the GDCAC algorithm. In Section 4, a simulation is performed, and the results shows the superiority of our proposed method in participant's profit and social welfare compared with traditional reinforcement learning methods. Section 5 concludes the paper.

2. Agent-Based Double-Side Day-Ahead Electricity Market Model

2.1. Participants' Bidding Model

In a double-side day-ahead electricity market, all GenCOs and DisCOs have the ability of learning by doing in order to maximize their own profits through their experiences in the competitive bidding procedure. Therefore, each of the GenCOs and DisCOs can be considered as an agent [19,24–33]. In this paper, without loss of generality, it is assumed that each GenCO has only one generation unit and submits one bid curve for each time interval of the next day, and considers the profit obtained through bidding in corresponding time interval of the next day, the same as each DisCO.

For GenCO i ($i = 1, 2, \ldots, N_g$), the formulation of its bid curve for time interval t ($t = 1, 2, \ldots, 24$) of the next day is a supply function based on its real marginal cost function:

$$SF_{i,t}(P_{gi,t}, k_{gi,t}) = k_{gi,t}(a_i P_{gi,t} + b_i), \quad P_{gi,t} \in [P_{gi,\min}, P_{gi,\max}] \tag{1}$$

where, $P_{gi,t}$, $k_{gi,t}$ represent the power production (MW) and bidding strategy ratio of GenCO i in time interval t, respectively.

The marginal cost function of GenCO i is:

$$MC_i(P_{gi,t}) = a_i P_{gi,t} + b_i \tag{2}$$

where, a_i, b_i represent the slope and intercept parameter of GenCO i's marginal cost function, respectively.

Because of the market power of GenCO i, it can bid with a supply function higher than its real marginal cost, so the bidding strategy ratio variable $k_{gi,t}$ satisfies $k_{gi,t} \in [1, k_{i,\max}]$.

For DisCO j ($j = 1, 2, \ldots, N_d$), the formulation of its bidding curve for time interval t ($t = 1, 2, \ldots, 24$) of the next day is a demand function based on its real marginal benefit function:

$$DF_{j,t}(P_{dj,t}, k_{dj,t}) = k_{dj,t}(-c_j P_{dj,t} + d_j), \quad P_{dj,t} \in [P_{dj,\min}, P_{dj,\max}] \tag{3}$$

where $P_{dj,t}$, $k_{dj,t}$ represent the power demand (MW) and bidding strategy ratio of DisCO j in time interval t, respectively.

The marginal revenue function of DisCO j is:

$$MD_j(P_{dj,t}) = -c_j P_{dj,t} + d_j \tag{4}$$

where $-c_j$ and d_j represent the slope and intercept parameter of DisCO j's marginal revenue function, respectively.

Because of the market power of DisCO j, it can bid with a demand function lower than its real marginal revenue, so the bidding strategy ratio variable $k_{dj,t}$ satisfies $k_{dj,t} \in (0, 1]$.

The profit of GenCO i after the completion of market for time interval t of the next day is:

$$R_{gi,t} = MCP_t \cdot Ps_{gi,t} - (\frac{1}{2} a_i \cdot {Ps_{gi,t}}^2 + b_i \cdot Ps_{gi,t}) \tag{5}$$

where, MCP_i represents the MCP in time interval t; and $Ps_{gi,t}$ represents the scheduled power production of GenCO i in time interval t. For the sake of simplicity and without loss of generality, it is assumed that the fixed cost of GenCO i is not considered in this paper.

The profit of DisCO j after the completion of market for time interval t of the next day is:

$$R_{dj} = (-\frac{1}{2} c_j \cdot {Pd_{dj,t}}^2 + d_j \cdot Pd_{dj,t}) - MCP_t \cdot Pd_{dj,t} \tag{6}$$

where $Pd_{dj,t}$ represents the scheduled and dispatched power consumption of DisCO j in time interval t. For the sake of simplicity and without loss of generality, it is assumed that the fixed benefit of DisCO j is not considered in this paper.

Taking [8,12,13,20–22] as references, we only consider one (negotiation) time interval for the next day. Hence, maximizing Equations (5) and (6) are the objectives of GenCO i and DisCO j in the double-side day-ahead electricity market, respectively.

2.2. Market Clearing Model

After receiving all bids for a certain time interval of the next day from GenCOs and DisCOs, the aggregated supply and demand bidding curves can then be constructed by the ISO to determine the MCP as well as the corresponding supply and demand schedules for the corresponding time interval in the next day. The ISO's market clearing management model for t can be described as follows:

$$\mathrm{Max}_{P_{gi,t}, \forall i, P_{dj,t}, \forall j} \sum_{j=1}^{N_j} [k_{dj,t}(-\frac{1}{2} c_j {P_{dj,t}}^2 + d_i P_{dj,t})] - \sum_{i=1}^{N_i} [k_{gi,t}(\frac{1}{2} a_i {P_{gi,t}}^2 + b_i P_{gi,t})] \tag{7}$$

$$\text{s.t.} \sum_{j=1}^{N_j} P_{dj,t} = \sum_{i=1}^{N_i} P_{gi,t} \tag{8}$$

$$LF_l = lf_l(P_{d1,t}, \ldots, P_{dN_j,t}, P_{g1,t}, \ldots P_{gN_i,t}), \forall l \tag{9}$$

$$|LF_l| \leq LF_{l,\max}, \forall l \tag{10}$$

$$P_{dj,t} \in [P_{dj,\min}, P_{dj,\max}], \forall j \tag{11}$$

$$P_{gi,t} \in [P_{gi,\min}, P_{gi,\max}], \forall i \tag{12}$$

where Equation (8) represents the power balance constraint, Equation (9) represents the power flow function in a transmission line l, and Equation (10) represents the system security constraints. The concrete formations of Equations (9) and (10) can be seen in [31]. By solving this optimization problem represented by Equations (7)–(12), the optimal scheduled power volumes of every GenCO and DisCO in time interval t corresponding to the maximal social welfare can be obtained. If we take the system security constraints into account, then the locational marginal prices (LMPs) of the whole system in time interval t can be calculated based on the dual variables of Equation (9), otherwise, the *MCP* of whole system in time interval t can be calculated based on the dual variable of Equation (8).

2.3. Agent Learning Mechanism

In a real double-side day-ahead electricity market, the rivals of a GenCO are the rest of GenCOs and all DisCOs in the same market, and the rivals of a DisCO are the rest of DisCOs and all GenCOs in the same market. As listed in Section 1.2, every participant (a GenCO or a DisCO) has no idea about its rivals' strategies historically and currently, what it knows is the information about historical *MCP*s. Literatures [19,25,26,29,33] have proposed that an agent-based GenCO (or DisCO) learns from the *MCP* (or LMP) of the last round market competition calculated and published by the ISO to decide which bidding strategy can be used in current market bidding competition in order to pursue its own profit maximization.

Based on the viewpoints expressed in [19,25,26,29,33], this paper proposes that an agent-based GenCO or DisCO participating in a double-side day-ahead electricity market learns from the historical *MCP*, calculated and published by the ISO yesterday (for today), to decide which bidding strategy can be applied for the next day in order to pursue its own profit maximization. Hence, there are some definitions described in this paper as follows:

(1) Transaction day: In a transaction day T ($T = 1, 2, \dots$), since the market is assumed to be cleared in a day-ahead single (negotiation) time interval basis, every GenCO or DisCO bids only one supply or demand function for the single time interval corresponding to the next day by use of *MCP* information calculated and published by the ISO in transaction day $T-1$ (for transaction day T).

(2) State variable: Historical *MCP* information calculated and published by the ISO in transaction day $T-1$ constitutes a value of the state variable in transaction day T.

$$x_T = MCP^{(T-1)} \tag{13}$$

(3) Action variable: In a transaction day T, the GenCO i or DisCO j's bidding strategy constitutes a value of its action variable. Hence, the action variable for GenCO i and DisCO j can be respectively described as follows:

$$u_{gi,T} = k_{gi}^{(T)} \tag{14}$$

$$u_{dj,T} = k_{dj}^{(T)} \tag{15}$$

(4) Iteration: We consider each transaction day as one iteration.

An agent-based participant has the ability of learning-by-doing or learning from its own experience so that when it has experienced sufficient iterations, the participant can take the optimal action (bidding strategy) which produces the most profit in face of any given state (x_T) of the environment (market). Hence, from a viewpoint of long period of time (many iterations), the values of $u_{gi,T}$, $u_{dj,T}$ ($i = 1, 2, \dots, N_g$; $j = 1, 2, \dots, N_d$) and x_T can be adjusted dynamically with the iterations,

which may be or not be constant after enough iterations, just as defined in [11–13] as Nash equilibrium of the market.

Now, the issue that we need to tackle with is as follows: practically, not only x_T, but also $u_{gi,T}$ and $u_{dj,T}$ ($i = 1, 2, \ldots, N_g$; $j = 1, 2, \ldots, N_d$) vary within a topologically continuous, bounded and closed set included in **R** respectively. Therefore, we need to find an appropriate method to model and simulate the dynamic process of strategy adjusting of every GenCO and DisCO in the incomplete and imperfect informational double-side day-ahead electricity market (satisfying provisions (1)–(4) simultaneously).

3. Methodology

In order to solve the issue mentioned in the last paragraph of Section 2, we proposed a modified reinforcement learning algorithm, namely the GDCAC algorithm.

Classic table-based reinforcement learning algorithms (e.g., SARSA algorithm, Q-learning algorithm et al.) can rapidly solve the Markov Decision Process (MDP) problems with discrete state and action spaces. For example, every GenCO's and DisCO's bidding strategy and the *MCP* of the market are assumed to vary within both two discrete and finite sets [19,25,26,29,33]. However, as mentioned above, the assumption of discrete sets of strategy (action) and *MCP* (state) isn't suitable for the actual situation of double-side day-ahead electricity market. Therefore, when using the classic reinforcement learning algorithm which uses a lookup table to store the state or state-action value information to model and simulate actual day-ahead electricity market bidding issue, the problem of "curse of dimensionality" will be generated, which challenges the classic table-based reinforcement learning algorithms on both memory space and learning efficiency. A common solution is to combine the classic reinforcement learning algorithms with function approximation methods in order to enhance the abstraction ability and generalization ability on state space and action space [37,38].

In this paper, the Actor-Critic method is used as basic structure of the agent-based participants' learning model in which the state value function corresponding to the Critic and the strategic/optimal action selecting policy function corresponding to the Actor are both approximated by the linear function model. The temporal difference (TD) error-based method is used to learn the parameters of the state function on line. The sigmoid function of TD error is used to construct a mean squared error (*MSE*) about policy parameters which is learned on line by a so-called gradient descent method [37,38]. After enough iterations of GDCAC algorithm, the parameters of the state value function and the strategic/optimal action selecting policy function are approximated optimally. Meanwhile, the agent can tell the optimal action in face of whatever state it met within continuous state space.

3.1. Policy Search

The reinforcement learning methods can be divided into three kinds, namely value iteration, policy iteration and policy search. Value iteration methods calculate the optimal value function in an iterative way. After reaching the convergence of the optimal value function, the optimal policy to select the best action in face of any state is determined by the optimal value function, and the typical value iteration algorithm is Q-learning algorithm. In policy iteration methods, the agent selects the actions according to an initial policy and interacts with the environment. During the process of interaction, the agent assesses the value function of the initial policy, and after reaching the convergence of the value function, the agent can obtain a better policy using the greedy method according to the value function. Then, the agent will take this obtained better policy as a new initial policy to repeat the aforementioned process, and finally get an optimal or near optimal policy. A typical policy iteration method is the SARSA algorithm [38]. Both the value iteration and policy iteration method are based on a lookup table to assess the value function, and their major defect has been briefly described above. Classic policy search methods are also based on a lookup table to store the value function information. In the process of interaction with the environment, the agent uses the immediate reward which is fed back by the environment to adjust its policy, which makes the probability of choosing the better action increase and the probability of the bad action decrease. Because of the explicit representation feature

of policy, the modification work of improving policy search method to become suitable for solving agent-based reinforcement learning problems with continuous state and action spaces is easier than that of the other two methods. Therefore, the modified policy search method is commonly used under the situations of continuous state and action spaces [37,38].

3.2. Introduction of the Gradient Descent Continuous Actor-Critic Algorithm

The Actor-Critic method consists of two parts, namely the actor and the critic. The actor part represents a clear policy which gives the probability of each action being selected at each state, and the critic part represents a value function which is the value function of the policy for the maintenance of the actors. The agent complies with the policy maintained by the actor to generate an action. In applying the action on the environment, the critic is responsible for receiving environmental immediate feedback reward and then updates the value function. At the same time, the critic calculates the corresponding TD error which is given back to the actor who adjusts the policy according to TD error in order to increase the probability of selecting the better action and decrease the probability of selecting the worse action. The basic structure of the Actor-Critic method is shown in Figure 1.

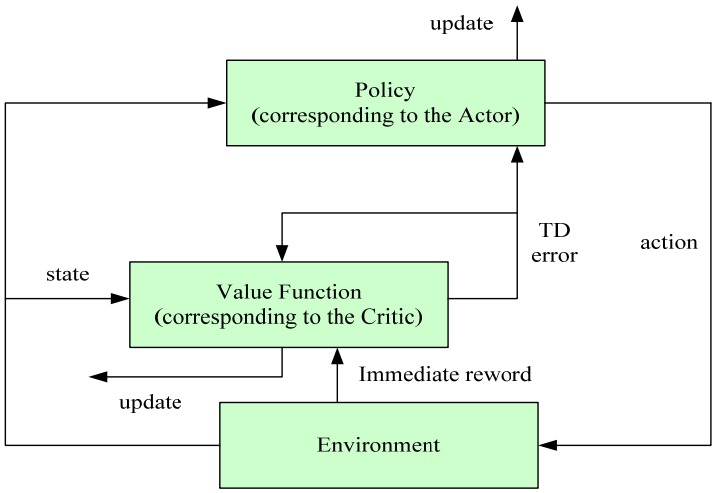

Figure 1. The diagram of an Actor-Critic reinforcement algorithm. TD: temporal difference.

In order to tackle with the issues of continuous state and action spaces, references [37,38] have proposed a method of using linear function to model the state value function and policy. The state value function model corresponding to the critic based on linear function can be described as follows:

$$\hat{V}(x) = \sum_{i=1}^{n} \phi_i(x)\theta_i = \vec{\boldsymbol{\phi}}(x)^T\theta \tag{16}$$

where, $\phi_i : \boldsymbol{X} \to \mathbf{R}\,(i = 1, 2, \ldots, n)$ represents the ith basis function of state $x \in \boldsymbol{X}$.

Then, a fixed basis function vector of state $x \in \boldsymbol{X}$ can be described as:

$$\vec{\boldsymbol{\phi}}(x) = (\phi_1(x), \phi_2(x), ..., \phi_n(x))^T \tag{17}$$

The linear parameter vector is:

$$\theta = (\theta_1, \theta_2, ..., \theta_n)^T \in \mathbf{R}^n \tag{18}$$

Then, we define a linear function $A : \boldsymbol{X} \rightarrow \boldsymbol{U}$ as the optimal policy model corresponding to the actor where the functional relationship between the optimal action $u_{opt}(x) \in \boldsymbol{U}$ and state $x \in \boldsymbol{X}$ is as follows:

$$u_{opt}(x) = A(x) = \vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\omega} \tag{19}$$

where, the linear parameter vector is:

$$\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_n)^T \in \mathbf{R}^n \tag{20}$$

In order to balance the exploration and exploitation in the reinforcement learning process, the policy by which the action is generated in face of every state must have the ability of exploration which is to select the sub-optimal action with a certain probability at each choice of action. This paper employs a Gaussian distribution function as the action generating model (policy) corresponding to the actor:

$$\rho(x, u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(u - \vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\omega})^2\right\} \tag{21}$$

where, $\sigma > 0$ is a standard deviation parameter which represents the exploring ability of the algorithm. Equation (21) indicates that when facing the state x, the probability of selecting the optimal action $\vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\omega}$ is the largest.

Therefore, the *MSE* function of parameter θ corresponding to the critic is:

$$MSE(\theta) = \frac{1}{2} \int_{x \in X} P^{(\rho)}(x)[V^{(\rho)}(x) - \vec{\boldsymbol{\phi}}(x)^T \theta]^2 dx \tag{22}$$

where, $P^{(\rho)}(x)$ is the probability distribution function of the state under policy ρ. The ideal goal is to find the global optimal parameter θ^* which satisfies:

$$MSE(\theta^*) \leq MSE(\theta) \tag{23}$$

Equation (23) indicates the generalization error of Equation (16) is minimized. However, we have no priori-knowledge about the real value function $V^{(\rho)}(x)$. Therefore, minimizing Equation (22) directly is impossible.

What we all know is that the gradient of a function represents the fastest increasing direction of the function value, and the negative gradient is the fastest decreasing direction of the function value. We can calculate the approximate formation of the gradient of $MSE(\theta)$:

$$grad(MSE(\theta)) = -\int_{x \in X} P^{(\rho)}(x)[V^{(\rho)}(x) - \vec{\boldsymbol{\phi}}(x)^T \theta]\vec{\boldsymbol{\phi}}(x) dx \tag{24}$$

As mentioned above, because we have no priori-knowledge about $V^{(\rho)}(x)$ and $P^{(\rho)}(x)$, [38] has used TD error to approximately replace $V^{(\rho)}(x) - \vec{\boldsymbol{\phi}}(x)^T \theta$. Assuming that at time step (iteration) T, the agent implements action u_T in the state of environment x_T and receives the immediate reword r_T, then the state of the environment shifts to x_{T+1}. The TD error at time step T is:

$$\delta_T = r_T + \gamma \vec{\boldsymbol{\phi}}(x_{T+1})^T \theta_T - \vec{\boldsymbol{\phi}}(x_T)^T \theta_T \tag{25}$$

where, $0 \leq \gamma \leq 1$ is a discount factor, θ_T is the estimated value of linear parameter vector θ at time step T. Based on the gradient descent method, the updating formula of parameter vector θ is:

$$\theta_{T+1} = \theta_T + \alpha_T \delta_T \vec{\boldsymbol{\phi}}(x_T) = \theta_T + \alpha_T [r_T + \gamma \vec{\boldsymbol{\phi}}(x_{T+1})^T \theta_T - \vec{\boldsymbol{\phi}}(x_T)^T \theta_T] \vec{\boldsymbol{\phi}}(x_T) \tag{26}$$

where, $\alpha_T > 0$ is the step length parameter which satisfies the following mathematical conditions:

$$\sum_{T=1}^{\infty} \alpha_T = \infty \text{ and } \sum_{T=1}^{\infty} (\alpha_T)^2 < \infty \tag{27}$$

Then, the problem of parameter vector $\boldsymbol{\omega}$ updating corresponding to the actor is analyzed. Assuming that in the state of environment x, the agent respectively implements action u_1 and u_2 ($u_1 \neq u_2$), and then the state of environment x will shift to state x_1 and x_2 correspondingly, and the corresponding immediate rewords are r_1 and r_2, respectively. Therefore, the two TD errors relevant to u_1 and u_2 are:

$$\delta(x, u_1) = r_1 + \gamma \vec{\boldsymbol{\phi}}(x_1)^T \boldsymbol{\theta} - \vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\theta} \text{ and } \delta(x, u_2) = r_2 + \gamma \vec{\boldsymbol{\phi}}(x_2)^T \boldsymbol{\theta} - \vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\theta} \tag{28}$$

If $\delta(x, u_1) > \delta(x, u_2)$, which means $r_1 + \gamma \vec{\boldsymbol{\phi}}(x_1)^T \boldsymbol{\theta} > r_2 + \gamma \vec{\boldsymbol{\phi}}(x_2)^T \boldsymbol{\theta}$, then the action u_1 is better than u_2 in the state of environment x, which is to say the parameter vector $\boldsymbol{\omega}$ needs to be adjusted/updated so as to make $A(x)$ closer to u_1 than u_2. In this state of environment x, the probability of selecting action u_1 needs to be larger than that of u_2. On the contrary, if $\delta(x, u_1) < \delta(x, u_2)$, (i.e., $r_1 + \gamma \vec{\boldsymbol{\phi}}(x_1)^T \boldsymbol{\theta} < r_2 + \gamma \vec{\boldsymbol{\phi}}(x_2)^T \boldsymbol{\theta}$), then the action u_2 is better than u_1 in the state of environment x, which is to say the parameter vector $\boldsymbol{\omega}$ needs to be adjusted/updated so as to make $A(x)$ closer to u_2 than u_1. In this state of environment x, the probability of selecting action u_2 needs larger than u_1.

Therefore, the *MSE* function of parameter $\boldsymbol{\omega}$ corresponding to the actor is:

$$MSE(\boldsymbol{\omega}) = \frac{1}{2} \int_{x \in X} P^{(\rho)}(x) \int_{u \in U} sig[\delta(x, u)] [\vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\omega} - u]^2 \, du dx \tag{29}$$

where, $sig\,[\delta(x, u)]$ is the sigmoid function of TD error $\delta(x, u)$. Reference [38] gives its formulation as follows:

$$sig[\delta(x, u)] = \frac{1}{1 + e^{-m\delta(x,u)}} \tag{30}$$

where, $m > 0$ is an adjustable parameter. From Equation (30), it is easy to know that $sig\,[\delta(x, u)]$ is a monotonically increasing function of $\delta(x, u)$, and $sig\,[\delta(x, u)] \in (0, 1]$.

If we minimize $sig\,[\delta(x, u)]\,[\vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\omega} - u]^2$, T then in the state of environment x, the larger the value of TD error $\delta(x, u)$, the higher the probability of selecting action u. The approximate formation of the gradient of $MSE(\boldsymbol{\omega})$ is:

$$grad[MSE(\boldsymbol{\omega})] = \int_{x \in X} P^{(\rho)}(x) \int_{u \in U} \frac{1}{1 + e^{-m\delta(x,u)}} [\vec{\boldsymbol{\phi}}(x)^T \boldsymbol{\omega} - u] \vec{\boldsymbol{\phi}}(x) \, du dx \tag{31}$$

Similar to the value function parameter $\boldsymbol{\theta}$ updating method, assuming that at time step T, the agent implements action u_T in the state of environment x_T and receives immediate reword r_T, then the state of the environment shifts to x_{T+1}, and the TD error is $\delta(x_T, u_T) = \delta_T$. Based on the gradient descent method, the updating formula of parameter vector $\boldsymbol{\omega}$ is:

$$\boldsymbol{\omega}_{T+1} = \boldsymbol{\omega}_T + \beta_T \frac{1}{1 + e^{-m\delta_T}} (u_T - \vec{\boldsymbol{\phi}}(x_T)^T \boldsymbol{\omega}_T) \vec{\boldsymbol{\phi}}(x_T) \tag{32}$$

where, $\beta_T > 0$ is the step length parameter which satisfies the mathematical conditions as follows:

$$\sum_{T=1}^{\infty} \beta_T = \infty, \text{ and } \sum_{T=1}^{\infty} (\beta_T)^2 < \infty \tag{33}$$

The pseudo-code of GDCAC algorithm is as follows:

(1) *Input*: the feature extraction function $\vec{\boldsymbol{\phi}} : \boldsymbol{X} \to \boldsymbol{R}^n$, discount factor , $0 \le \gamma \le 1$, step length parameter series $\{\alpha_T\}_{T=0}^{\infty}$, $\{\beta_T\}_{T=0}^{\infty}$, and parameters σ, m.

(2) Initialize linear parameter vectors θ_0 and $\boldsymbol{\omega}_0$.

(3) Repeat (for each episode)

Initialize state x_0 randomly

Repeat (for each time step $T = 0, 1, 2, \ldots$ in the episode)

Choose and implement an action $u_T \sim N(\vec{\boldsymbol{\phi}}(x_T)^T \boldsymbol{\omega}_T, \sigma^2)$ from state x_T, then observe immediate reword r_T and the next state x_{T+1};

$\delta_T = r_T + \gamma \vec{\boldsymbol{\phi}}(x_{T+1})^T \theta_T - \vec{\boldsymbol{\phi}}(x_T)^T \theta_T$;

$\theta_{T+1} = \theta_T + \alpha_T \delta_T \vec{\boldsymbol{\phi}}(x_T)$;

$\boldsymbol{\omega}_{T+1} = \boldsymbol{\omega}_T + \beta_T \frac{1}{1+e^{-m\delta_T}} (u_T - \vec{\boldsymbol{\phi}}(x_T)^T \boldsymbol{\omega}_T) \vec{\boldsymbol{\phi}}(x_T)$;

Until x_{T+1} is terminal

Until the desired number of episodes has been searched.

(4) *Output*: $\theta^* = \theta_{T+1}, \boldsymbol{\omega}^* = \boldsymbol{\omega}_{T+1}$ and $V^*(x)$, $A^*(x)$.

3.3. The Proposed Market Procedure

The procedure of implementing the GDCAC algorithm for electricity market modeling by considering continuous state (*MCP*) and action (bidding strategy) sets is as follows:

(1) *Input*: for GenCO i $\vec{\boldsymbol{\phi}}_g : \boldsymbol{MCP} \to \mathbf{R}^n$ and for DisCO j $\vec{\boldsymbol{\phi}}_d : \boldsymbol{MCP} \to \mathbf{R}^n$, discount factor $0 \le \gamma \le 1$, step length parameter series $\{\alpha_T{}^{(g)}\}_{T=0}^{\infty}$ and $\{\beta_T{}^{(g)}\}_{T=0}^{\infty}$ for GenCO i, step length parameter series $\{\alpha_T{}^{(d)}\}_{T=0}^{\infty}$ and $\{\beta_T{}^{(d)}\}_{T=0}^{\infty}$ for DisCO j, and parameters σ, m.

(2) Initialize the linear parameter vectors $\theta_0{}^{(gi)}$ and $\boldsymbol{\omega}_0{}^{(gi)}$ for GenCO i, linear parameter vectors $\theta_0{}^{(dj)}$ and $\boldsymbol{\omega}_0{}^{(dj)}$ for DisCO j.

(3) $T = 0$.

(4) Initialize $k_{gi,0} \in [1, k_{i,max}]$ for GenCO i and $k_{dj,0} \in (0, 1]$ for DisCO j randomly, and calculate $x_1 = MCP_0$ through Equations (1), (3), and (7)–(12).

(5) Repeat (for each time step $T = 1, 2, \ldots,$ TN).

GenCO i chooses and implements an action $u_T{}^{(gi)} = k_{gi,T} \sim N(\vec{\boldsymbol{\phi}}_g(x_T)^T \boldsymbol{\omega}_T{}^{(gi)}, \sigma^2)$ from state $x_T = MCP_{T-1}$, then observes the immediate reword $r_{gi,T}$ using Equation (5) and the next state x_{T+1} generated by Equations (1), (3), and (7)–(12);

DisCO j chooses and implements an action $u_T{}^{(dj)} = k_{dj,T} \sim N(\vec{\boldsymbol{\phi}}_d(x_T)^T \boldsymbol{\omega}_T{}^{(dj)}, \sigma^2)$ from state $x_T = MCP_{T-1}$, and then observes the immediate reword $r_{dj,T}$ using Equation (6) and the next state x_{T+1} generated by Equations (1), (3), and (7)–(12);

GenCO i updates: $\delta_{gi,T} = r_{gi,T} + \gamma \vec{\boldsymbol{\phi}}_g(x_{T+1})^T \theta_T{}^{(gi)} - \vec{\boldsymbol{\phi}}_g(x_T)^T \theta_T{}^{(gi)}$;

$\theta_{T+1}{}^{(gi)} = \theta_T{}^{(gi)} + \alpha_T{}^{(g)} \delta_{gi,T} \vec{\boldsymbol{\phi}}_g(x_T)$;

$\boldsymbol{\omega}_{T+1}{}^{(gi)} = \boldsymbol{\omega}_T{}^{(gi)} + \beta_T{}^{(g)} \frac{1}{1+e^{-m\delta_{gi,T}}} (u_T{}^{(gi)} - \vec{\boldsymbol{\phi}}_g(x_T)^T \boldsymbol{\omega}_T{}^{(gi)}) \vec{\boldsymbol{\phi}}_g(x_T)$;

DisCO j updates: $\delta_{dj,T} = r_{dj,T} + \gamma \vec{\boldsymbol{\phi}}_d(x_{T+1})^T \theta_T{}^{(dj)} - \vec{\boldsymbol{\phi}}_d(x_T)^T \theta_T{}^{(dj)}$;

$\theta_{T+1}{}^{(dj)} = \theta_T{}^{(dj)} + \alpha_T{}^{(d)} \delta_{dj,T} \vec{\boldsymbol{\phi}}_d(x_T)$;

$\boldsymbol{\omega}_{T+1}{}^{(dj)} = \boldsymbol{\omega}_T{}^{(dj)} + \beta_T{}^{(d)}$;

(6) *Output*: for GenCO i: $\theta_{gi}{}^* = \theta_{T+1}{}^{(gi)}, \boldsymbol{\omega}_{gi}{}^* = \boldsymbol{\omega}_{T+1}{}^{(gi)}$ and $V_{gi}{}^*(x)$, $A_{gi}{}^*(x)$; for GenCO i: $\theta_{dj}{}^* = \theta_{T+1}{}^{(dj)}, \boldsymbol{\omega}_{dj}{}^* = \boldsymbol{\omega}_{T+1}{}^{(dj)}$ and $V_{dj}{}^*(x)$, $A_{dj}{}^*(x)$.

From the step-by-step procedure listed in this subsection, it is easy to know that the time complexity of our proposed GDCAC-based electricity market model is O(n). According to Reference [38], we choose Gaussian radial basis function as $\vec{\Phi}_g(x)$ and $\vec{\Phi}_d(x)$.

4. Simulation and Discussions

Because in China the double-side day-ahead electricity market has not yet been established in any region (it is clear that one of the development directions in China's power restructuring in coming days is to establish double-side spot electricity markets in many regions and levels—province, city or district etc. [4]), the proposed GDCAC approach is implemented on a double-side day-ahead electricity market test system containing six GenCOs and five DisCOs [20], but without taking network constraints into consideration [9]. Considering that in a newly established electricity market, all participants are initially short of bidding experience and historical market data, they must firstly go through a repeated process of exploration and trial and error, accumulating experiences gradually, and then reach making more rational bidding decisions in the face of any market environment state. Hence, in the first iteration of market competition ($T = 0$), we assume every participant chooses its bidding strategy randomly because of lack of experience [19,25], and in iteration T ($T > 0$), we assume every participant chooses its bidding strategy by considering the historic *MCP* generated from iteration $T-1$ [19,25]. It is feasible to simulate the strategic bidding process of an existing double-side spot electricity markets with our proposed method by letting all participants know the historical *MCP* information when bidding in the first iteration of market competition ($T = 0$) etc. The main contents of this section are as follows:

Firstly, in order to demonstrate the superiority of our proposed model for double-side day-ahead electricity market over the classic table-based reinforcement learning model which was proposed in [19,24–26,29,33], three scenarios are established in sub-Section 4.2, among which Scenario 1 assumes that both the market state (*MCP*) and all participants' action (bidding strategy) sets are discrete, Scenario 2 assumes that GenCO1 is a GDCAC-based agent with continuous state and action sets while other participants are the same as that in Scenario 1, and Scenario 3 assumes that all participants in the market are our proposed GDCAC-based agents with continuous state and action sets.

Secondly, the comparison of profits of all participants in three scenarios after a given number of iterations is shown in Section 4.2, which demonstrates the superiority of our proposed model; Finally, the sensitivity analysis with respect to different numbers of training iterations are presented in sub-Section 4.3, which leads to two new topics to be studied by means of our proposed double side day-ahead electricity market model.

4.1. Data and Assumptions

The parameters of GenCOs' and DisCOs' bid functions are shown in Table 1 [20].

Table 1. Economical technological coefficients of GenCOs and DisCOs.

Participants	a_i (10^3 RMB yuan /MW^2)	b_i (10^3 RMB yuan /MW)	$k_{gi,min}$	$k_{gi,max}$	$P_{gi,min}$	$P_{gi,max}$
GenCO1	0.046	14	1.0	3.0	0	210
GenCO2	0.074	10	1.0	3.0	0	600
GenCO3	0.062	12	1.0	3.0	0	200
GenCO4	0.043	25	1.0	3.0	0	520
GenCO5	0.031	20	1.0	3.0	0	250
GenCO6	0.064	20	1.0	3.0	0	400
Participants	c_j (10^3 RMB yuan /MW^2)	d_j (10^3 RMB yuan /MW)	$k_{j,min}$	$k_{j,max}$	$P_{dj,min}$	$P_{dj,max}$
DisCO1	−0.052	25	0	1.0	0	250
DisCO2	−0.034	25	0	1.0	0	250
DisCO3	−0.031	20	0	1.0	0	300
DisCO4	−0.054	25	0	1.0	0	300
DisCO5	−0.013	20	0	1.0	0	300

It is assumed that $k_{g1}, k_{g2}, k_{g3}, k_{g4}, k_{g5}, k_{g6} \in [1,3]$, $k_{d1}, k_{d2}, k_{d3}, k_{d4}, k_{d5} \in [1,3]$, (actually, changing the value interval of any strategy (k_{gi} or k_{dj}) will not affect the final results of the Nash equilibrium). Table 2 presents the state and action sets of every participant while taking Scenarios 1, 2 and 3 into consideration, respectively. All participants are considered as the learning agents who bid strategically by using and adjusting their own strategic variables k_{gi} or k_{dj} (i = 1, 2, 3, 4, 5, 6 and j = 1, 2, 3, 4, 5). The related parameters of the GDCAC algorithm and classic table-based reinforcement learning algorithm which use the ε-greedy method to balance exploration and exploitation [19,24–26,29,33] are also listed in Table 2.

Table 2. Related information about the three scenarios.

Scenarios	Participants	State Set (RMB yuan/MWh)	Action Set	ε	γ	α	β	σ	m
Scenario 1	Gen1	$\{X_1, X_2, \ldots, X_{20}\}$	$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen2		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen3		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen4		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen5		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen6		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Dis1		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis2		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis3		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis4		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis5		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
Scenario 2	Gen1	[10 34]	[1 3]	-	0.5	0.1	0.1	4	1
	Gen2	$\{X_1, X_2, \ldots, X_{20}\}$	$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen3		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen4		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen5		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Gen6		$\{U_{g1}, U_{g2}, \ldots, U_{g20}\}$	0.1	0.5	-	-	-	-
	Dis1		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis2		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis3		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis4		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
	Dis4		$\{U_{d1}, U_{d2}, \ldots, U_{d20}\}$	0.1	0.5	-	-	-	-
Scenario 3	Gen1	[10 34]	[1 3]	-	0.5	0.1	0.1	4	1
	Gen2		[1 3]	-	0.5	0.1	0.1	4	1
	Gen3		[1 3]	-	0.5	0.1	0.1	4	1
	Gen4		[1 3]	-	0.5	0.1	0.1	4	1
	Gen5		[1 3]	-	0.5	0.1	0.1	4	1
	Gen6		[1 3]	-	0.5	0.1	0.1	4	1
	Dis1		[0.3 1]	-	0.5	0.1	0.1	4	1
	Dis2		[0.3 1]	-	0.5	0.1	0.1	4	1
	Dis3		[0.3 1]	-	0.5	0.1	0.1	4	1
	Dis4		[0.3 1]	-	0.5	0.1	0.1	4	1
	Dis5		[0.3 1]	-	0.5	0.1	0.1	4	1

Note: X_1 represents the interval [10 11.2) RMB yuan/MWh, X_2 represents the interval [11.2 12.4) RMB yuan/MWh, ... , X_{19} represents the interval [31.6 32.8) RMB yuan/MWh, and X_{20} represents interval [32.8 34] yuan/MWh; U_{g1} represents the interval [1 1.1), U_{g2} represents the interval [1.1 1.2), ... , U_{g19} represents the interval [2.8 2.9), and U_{g20} represents the interval [2.9 3]; U_{d1} represents the interval (0.3 0.335), U_{d2} represents interval [0.335 0.37), ... , U_{d19} represents interval [0. 93 0.965), and U_{d3} represents interval [0. 965 1].

Set the central point parameters in the Gauss radial basis function to form the following set:

$$\boldsymbol{C} = \{10, 14, 18, 22, 26, 30, 34\}$$

4.2. Simulation Result and Comparative Analysis

For the simulation on the three scenarios, every participant of the market will go through a process of training with 3000 iterations in which all participants' actions selecting policy consider the balance of exploration and exploitation. After the training process, decision making process with 500 iterations will be implemented by all participants, in which only greedy policy will be adopted by every participant when selecting actions in face of a given state of the market. Using Matlab R2014a software to program and run the models of three scenarios, the profits of all participants in three scenarios when the market reaches the dynamic stability (namely Nash equilibrium [11–13]) can be obtained, which are

listed in Table 3. At this time, the profit and bidding strategy of every participant and *MCP* in the market are not changing over time (iterations). Figure 2 shows the dynamic adjusting processes of *MCP* in Scenario 3. The dynamic adjusting process of all participants' profits in Scenario 3 from horizontal and vertical perspectives are depicted in Figures 3 and 4, respectively. From Figure 3, the variations of eleven participants' profits with 3500 iterations can be respectively seen. From Figure 4, the profits of eleven participants can be compared with 3500 iterations.

Table 3. The profits of all participants when the market reaches the dynamic stability in three scenarios. (Unit: 10^3 RMB yuan). Scen: Scenario.

Participants		Gen1	Gen2	Gen3	Gen4	Gen5	Gen6	Dis1	Dis2	Dis3	Dis4	Dis5	Sum
Profits	Scen 1	1.3353	1.4581	1.3870	1.1777	1.8035	0.6384	1.2555	1.9202	1.4146	1.0530	1.4774	14.9207
	Scen 2	1.4030	1.5142	1.3622	1.1502	1.7634	0.6217	1.2828	1.9620	1.4529	1.0759	1.5414	15.1297
	Scen 3	1.4952	1.6809	1.5363	1.3604	2.0445	0.7442	1.1860	1.7686	1.3464	0.9963	1.4221	15.5809

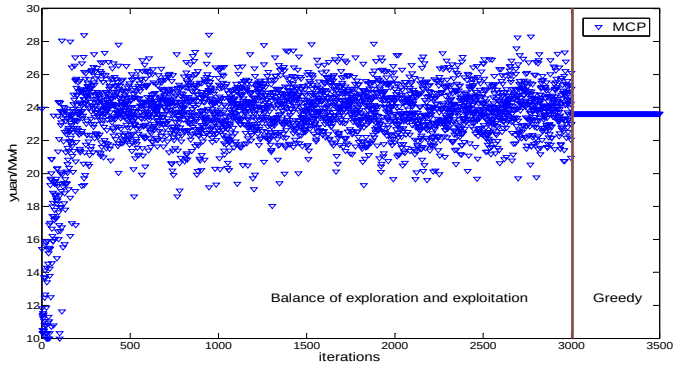

Figure 2. The dynamic adjusting processes of *MCP* in Scenario 3.

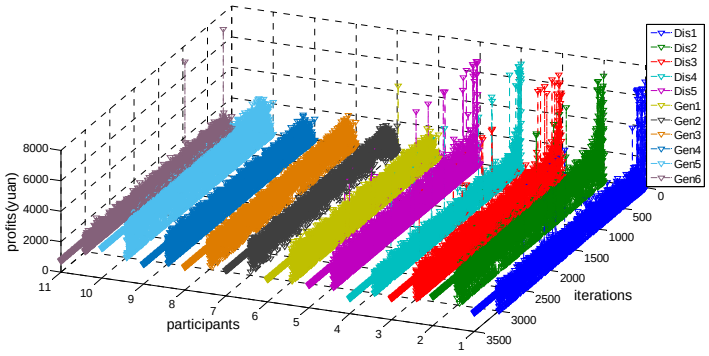

Figure 3. The dynamic adjusting processes of every participant's profit (from a horizontal perspective).

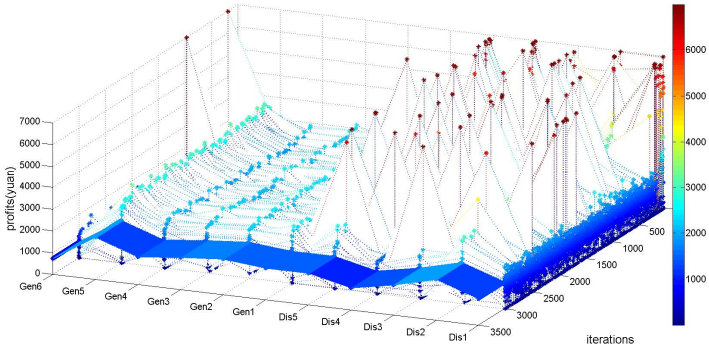

Figure 4. The dynamic adjusting processes of every participant's profit (from a vertical perspective).

From Table 3, it can be seen that:

(1) After the same number of iterations (including 3000 iterations of training and 500 iterations of decision making), GenCO 1's profit in Scenario 2 is 1.4030×10^3 yuan which is higher than GenCO 1's profit in Scenario 1 (e.g., 1.3353×10^3 yuan). This indicates one can get more profit by using our proposed GDCAC reinforcement learning model to bid in the market than using the traditional Q-learning model with the same conditions (namely the same parameters values, number of iterations, and adaptive learning mechanism of other participants);

(2) If we ignore the externality, the total social welfare of the electricity market is equal to the summation of all participants' profits. Therefore, after the same number of iterations, the social welfare in Scenario 3 is higher than that in Scenario 2, and the social welfare in Scenario 2 is higher than that in Scenario 1. This indicates with the increase in the number of participants by using our proposed GDCAC reinforcement learning model to bid in electricity market, the total social welfare can be higher and higher;

Regarding to the profit of a specific participant and the total social welfare of the electricity market, our simulation of this case study shows the superiority of our proposed GDCAC model over the table-based Q-learning one. The main reasons of this result are: (1) because the traditional table-based reinforcement learning algorithm can hardly store the value function information about continuous data sets, which will cause the curse of dimensionality; and (2) no matter how many sub-intervals are divided from the original continuous state and action sets, the state and action sets in traditional table-based Q-learning electricity model are still discrete, and the globally optimal action solution can hardly be found to cope with the issues with continuous state and action sets such as double-side day-ahead electricity market simulation.

Figures 5 and 6 show the dynamic adjusting processes of every participant's bidding strategy from the horizontal perspective and vertical perspective respectively. From Figures 2–6, it can be seen that in Scenario 3 (actually the same as Scenarios 1 and 2), when we assume every participant employs our proposed GDCAC reinforcement learning method to bid in the double-side day-ahead electricity market, all market-related factors including *MCP*, profit and bidding strategy of every participant will reach a dynamically stable state respectively and simultaneously. Even when the number of training iterations and learning algorithm are set to be different among all participants, all market factors will also reach dynamically stable states after enough iterations, which may have different values from the former one. This dynamically stable state can be considered as a Nash Equilibrium (NE) [11–13]. Moreover, it only takes about 2.23 s on a 2.5 GHz laptop computer for the double-side day-ahead electricity market including eleven participants in Scenario 3 to find the equilibrium through 3500 iterations, which is attributed to the low time complexity of our proposed method.

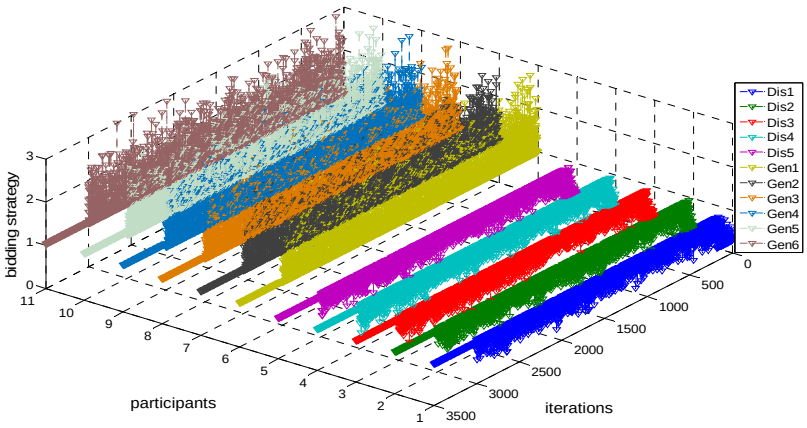

Figure 5. The dynamic adjusting processes of every participant's bidding strategy (from a horizontal perspective).

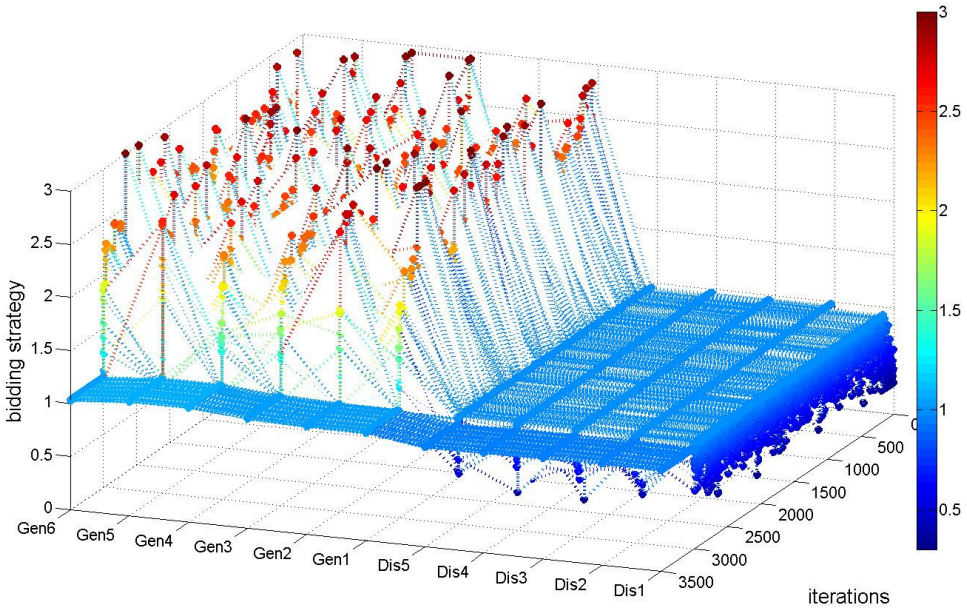

Figure 6. The dynamic adjusting processes of every participant's bidding strategy (from a vertical perspective).

4.3. Sensitivity Analysis

In order to examine the influence of different numbers of training iterations on NE of double-side day-ahead electricity market in Scenario 3, we set five cases related to the number of training iterations, and the results are listed in Tables 4 and 5, respectively.

Table 4. The obtained profits with different cases in Scenario 3 (unit: 10^3 RMB yuan).

Cases		Gen1	Gen2	Gen3	Gen4	Gen5	Gen6	Dis1	Dis2	Dis3	Dis4	Dis5	Sum
Numbers of training iterations	1000	1.5441	1.7012	1.5932	1.4133	2.0992	0.7905	1.1380	1.7204	1.2790	0.9585	1.3209	15.5583
	2000	1.4744	1.6543	1.4986	1.2826	1.9602	0.7249	1.2252	1.8838	1.3884	1.0241	1.4600	15.5765
	3000	1.4952	1.6809	1.5363	1.3604	2.0445	0.7442	1.1860	1.7686	1.3464	0.9963	1.4221	15.5809
	4000	1.5686	1.6696	1.4343	1.3713	2.1683	0.7986	1.1329	1.6471	1.2650	0.9255	1.2714	15.2526
	5000	1.4315	1.6855	1.5455	1.3334	2.0525	0.7525	1.1844	1.8216	1.2907	1.0006	1.3901	15.4883

Table 5. The obtained strategies with different cases in Scenario 3.

Cases		Gen1	Gen2	Gen3	Gen4	Gen5	Gen6	Dis1	Dis2	Dis3	Dis4	Dis5	*MSE* *
Numbers of training iterations	1000	1.1117	1.0986	1.1723	1.0576	1.0976	1.0308	0.9555	0.9398	0.9828	0.9477	0.9225	0.0854
	2000	1.0719	1.1116	1.1238	1.0693	1.0494	1.0877	0.9661	0.9243	0.9518	0.8928	0.9645	0.0797
	3000	1.0185	1.0186	1.0343	1.0791	1.0771	1.0291	0.9571	0.9640	0.9552	0.9471	0.9392	0.0491
	4000	1.1068	1.1478	1.1146	1.0257	1.0332	1.0508	0.9145	0.9592	0.9623	0.9813	0.9294	0.0968
	5000	1.0706	1.3096	1.1995	1.0471	1.0427	1.1625	0.9285	0.9928	0.9721	0.9899	0.9447	0.0881

Note: *MSE* represents the mean square error between the strategies of all participants and 1. For example, 0.0854 = $\sqrt{\frac{1}{11}\left[(1.1117-1)^2+(1.0986-1)^2+(1.1723-1)^2+(1.0576-1)^2+(1.0976-1)^2+(1.0308-1)^2+(0.9555-1)^2+(0.9398-1)^2+(0.9828-1)^2+(0.9477-1)^2+(0.9225-1)^2\right]}$.

From Tables 4 and 5, it can be seen that:

(1) There is no monotonic relationship between the social welfare and the number of training iterations, which may be caused by the system noises during training process. Therefore, in the market simulation with our proposed GDCAC reinforcement learning model, how to find the globally optimal number of training iterations that can bring the highest social welfare may be a new topic to be studied.

(2) Social welfare increases with the decrease of *MSE* between all participants' strategy values and 1. It is known that every participant will respectively bid at its marginal cost or revenue when all

participants' strategy values equal to 1, which also means the perfect competition and the highest welfare. Therefore, how to design the double-side electricity market mechanism especially for China to pursue higher efficiency of resource allocation by means of our proposed GDCAC reinforcement learning market model may be another new topic to be studied.

5. Conclusions

China is experiencing a new round of electricity market reforms, and the double-side day-ahead electricity market will become more and more important in China's energy trading area in the future. On one hand, the participant who expects to pursue more profit and less business risk, needs to employ a suitable and feasible technology to simulate the dynamic market environment and return it the optimal bidding strategy under any market environment state. On the other hand, the government hopes to effectively design the double-side day-ahead electricity market mechanism and formulate the relevant policies, and also needs to employ a suitable and feasible technology to simulate the market dynamic process and equilibrium consequence.

This paper a new double-side day-ahead electricity market modeling and simulating method based on GDCAC algorithm is proposed. Some conclusions can be drawn as follows:

(1) Our proposed GDCAC reinforcement learning market model needs no common knowledge of every participant's cost or revenue, strategy probability distribution function of every participant, *MCP* probability distribution function of the market, and scheduling result of every participant, which need be more or less assumed to be known by every participant in most game-based models.

(2) Our proposed GDCAC reinforcement learning market model can cope with the issues with continuous state and action sets without causing trouble of 'curse of dimensionality', which cannot be overcome by using traditional table-based reinforcement learning algorithms. Therefore, our proposed model is more suitable and feasible for simulating the practical double-side day-ahead electricity market in which both the state (*MCP*) and action (every participant's bidding strategy) sets are continuous.

(3) Because the time complexity of GDCAC reinforcement learning algorithm is only O(n), our proposed model can be used in large-scale electricity market system simulation with a lot of participants competing with each other simultaneously, which can hardly be achieved by using game-based models or table-based reinforcement models.

(4) The simulation results show that by using our proposed model, a participant can get more profit than that without using it. Meanwhile, if every participant in the market adopts our proposed model simultaneously, the Nash equilibrium result of electricity market will bring higher social welfare, which is very close to the situation of every participant using marginal cost or revenue based bidding strategy.

Our proposed GDCAC reinforcement learning market model, which can simulate the dynamic bidding process and market equilibrium in the double-side day-ahead electricity market is not only of importance to some developed countries but also to China. For the participants (GenCOs or DisCOs), it can provide a bidding decision-making tool to get more profits in the market competition. For the government, it can provide an economic analysis tool to help design proper market mechanism and policies.

Acknowledgments: This study is supported by the National Natural Science Foundation of China under Grant No. 71373076, the Fundamental Research Funds for the Central Universities under Grant No. 2015 XS28.

Author Contributions: Huiru Zhao guided the research; Yuwei Wang established the model, implemented the simulation and wrote this article; Sen Guo guided and revised the paper and refined the language; Mingrui Zhao and Cao Zhang collected references.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mou, D. Understanding China's electricity market reform from the perspective of the coal-fired power disparity. *J. Energy Policy* **2014**, *74*, 224–234. [CrossRef]
2. Ma, C.; Zhao, X. China's electricity market restructuring and technology mandates: Plant-level evidence for changing operational efficiency. *J. Energy Econ.* **2015**, *47*, 227–237. [CrossRef]
3. Sun, Q.; Xu, L.; Yin, H. Energy pricing reform and energy efficiency in China: Evidence from the automobile market. *J. Resour. Energy Econ.* **2016**, *44*, 39–51. [CrossRef]
4. Website of National Development and Reform Commission (NDRC) People's Republic of China. Available online: http://www.sdpc.gov.cn/zcfb/zcfbtz/201511/t20151130_760016.html (accessed on 26 November 2015).
5. Prabavathi, M.; Gnanadass, R. Energy bidding strategies for restructured electricity market. *J. Int. J. Electr. Power Energy Syst.* **2015**, *64*, 956–966. [CrossRef]
6. Ringler, P.; Keles, D.; Fichtner, W. Agent-based modeling and simulation of smart electricity grids and markets—A literature review. *J. Renew. Sustain. Energy Rev.* **2016**, *57*, 205–215. [CrossRef]
7. Kardakos, E.G.; Simoglou, C.K.; Bakirtzis, A.G. Optimal bidding strategy in transmission-constrained electricity markets. *J. Electr. Power Syst. Res.* **2014**, *109*, 141–149. [CrossRef]
8. Al-Agtash, S.Y. Supply curve bidding of electricity in constrained power networks. *J. Energy* **2010**, *35*, 2886–2892. [CrossRef]
9. Langary, D.; Sadati, N.; Ranjbar, A.M. Direct approach in computing robust Nash strategies for generating companies in electricity markets. *J. Int. J. Electr. Power Energy Syst.* **2014**, *54*, 442–453. [CrossRef]
10. Borghetti, A.; Massucco, S.; Silvestro, F. Influence of feasibility constrains on the bidding strategy selection in a day-ahead electricity market session. *J. Electr. Power Syst. Res.* **2009**, *79*, 1727–1737. [CrossRef]
11. Gao, F.; Sheble, G.B.; Hedman, K.W.; Yu, C.N. Optimal bidding strategy for GENCOs based on parametric linear programming considering incomplete information. *J. Int. J. Electr. Power Energy Syst.* **2015**, *66*, 272–279. [CrossRef]
12. Kumar, J.V.; Kumar, D.M.V. Generation bidding strategy in a pool based electricity market using Shuffled Frog Leaping Algorithm. *J. Appl. Soft Comput.* **2014**, *21*, 407–414. [CrossRef]
13. Wang, J.; Zhou, Z.; Botterud, A. An evolutionary game approach to analyzing bidding strategies in electricity markets with elastic demand. *J. Energy* **2011**, *36*, 3459–3467. [CrossRef]
14. Liu, Z.; Yan, J.; Shi, Y.; Zhu, K.; Pu, G. Multi-agent based experimental analysis on bidding mechanism in electricity auction markets. *J. Int. J. Electr. Power Energy Syst.* **2012**, *43*, 696–702. [CrossRef]
15. Nojavan, S.; Zare, K.; Feyzi, M.R. Optimal bidding strategy of generation station in power market using information gap decision theory (IGDT). *J. Electr. Power Syst.* **2013**, *96*, 56–63. [CrossRef]
16. Wen, F.; David, A.K. Optimal bidding strategies and modeling of imperfect information among competitive generators. *IEEE Trans. Power Syst.* **2001**, *16*, 15–21.
17. Kumar, J.V.; Kumar, D.M.V.; Edukondalu, K. Strategic bidding using fuzzy adaptive gravitational search algorithm in a pool based electricity market. *J. Appl. Soft Comput.* **2013**, *13*, 2445–2455. [CrossRef]
18. Azadeh, A.; Skandari, M.R.; Maleki-Shoja, B. An integrated ant colony optimization approach to compare strategies of clearing market in electricity markets: Agent-based simulation. *J. Energy Policy* **2010**, *38*, 6307–6319. [CrossRef]
19. Rahimiyan, M.; Mashhadi, H.R. Supplier's optimal bidding strategy in electricity pay-as-bid auction: Comparison of the Q-learning and a model-based approach. *J. Electric Power Syst. Res.* **2008**, *78*, 165–175. [CrossRef]
20. Shivaie, M.; Ameli, M.T. An environmental/techno-economic approach for bidding strategy in security-constrained electricity markets by a bi-level harmony search algorithm. *J. Renew. Energy* **2015**, *83*, 881–896. [CrossRef]
21. Menniti, D.; Pinnarelli, A.; Sorrentino, N. Simulation of producers' behaviour in the electricity market by evolutionary games. *J. Electr. Power Syst. Res.* **2008**, *78*, 475–483. [CrossRef]
22. Ladjici, A.A.; Boudour, M. Nash–Cournot equilibrium of a deregulated electricity market using competitive coevolutionary algorithms. *J. Electr. Power Syst. Res.* **2011**, *81*, 958–966. [CrossRef]
23. Ladjici, A.A.; Tiguercha, A.; Boudour, M. Nash Equilibrium in a two-settlement electricity market using competitive coevolutionary algorithms. *J. Int. J. Electr. Power Energy Syst.* **2014**, *57*, 148–155. [CrossRef]

24. Salehizadeh, M.R.; Soltaniyan, S. Application of fuzzy Q-learning for electricity market modeling by considering renewable power penetration. *Renew. Sustain. Energy Rev.* **2016**, *56*, 1172–1181. [CrossRef]
25. Thanhquy, B. Using Reinforcement Learning to Study the Features of the Participants' Behavior in Wholesale Power Market. Ph.D Thesis, Hunan University, Hunan, China, 2013.
26. Naghibi-Sistani, M.B.; Akbarzadeh-Tootoonchi, M.R.; Javidi-Dashte, B.M.H.; Rajabi-Mashhadi, H. Application of Q-learning with temperature variation for bidding strategies in market based power systems. *J. Energy Convers. Manag.* **2006**, *47*, 1529–1538. [CrossRef]
27. Li, H.; Tesfatsion, L. Co-learning patterns as emergent market phenomena: An electricity market illustration. *J. Econ. Behav. Organ.* **2012**, *82*, 395–419. [CrossRef]
28. Pinto, T.; Sousa, T.M.; Morais, H.; Praça, I.; Vale, Z. Metalearning to support competitive electricity market players' strategic bidding. *J. Electr. Power Syst. Res.* **2016**, *135*, 27–34. [CrossRef]
29. Lim, Y.; Kim, H.M. Strategic bidding using reinforcement learning for load shedding in microgrids. *J. Comput. Electr. Eng.* **2014**, *40*, 1439–1446. [CrossRef]
30. Sheikhi, A.; Rayati, M.; Ranjbar, A.M. Dynamic load management for a residential customer: Reinforcement Learning approach. *J. Sustain. Cities Soc.* **2016**, *24*, 42–51. [CrossRef]
31. Mahvi, M.; Ardehali, M.M. Optimal bidding strategy in a competitive electricity market based on agent-based approach and numerical sensitivity analysis. *J. Energy* **2011**, *36*, 6367–6374. [CrossRef]
32. Bublitz, A.; Genoese, M.; Fichtner, W. An agent-based model of the German electricity market with short-time uncertainty factors. In Proceedings of the 2014 11th International Conference on European Energy Market (EEM) IEEE, Cracow, Poland, 28–30 May 2014.
33. Raju, L.; Sibi, S.; Milton, R.S. Distributed optimization of solar micro-grid using multi agent reinforcement learning. *J. Procedia Comput. Sci.* **2015**, *46*, 231–239. [CrossRef]
34. Wang, Y.H.; Li, T.H.S.; Lin, C.J. Backward Q-learning: The combination of Sarsa algorithm and Q-learning. *J. Eng. Appl. Artif. Intell.* **2013**, *26*, 2184–2193. [CrossRef]
35. Xu, M.L.; Xu, W.B. Fuzzy Q-learning in continuous state and action space. *J. China Univ. Posts Telecommun.* **2010**, *17*, 100–109. [CrossRef]
36. Xu, X.; Zuo, L.; Huang, Z. Reinforcement learning algorithms with function approximation: Recent advances and applications. *J. Inf. Sci.* **2014**, *261*, 1–31. [CrossRef]
37. Chen, X. Study of Reinforcement Learning Algorithms Based on Value Function Approximation. Ph.D Thesis, Nanjing University, Jiangsu, China, 2013.
38. Chen, G. Research on Value Function Approximation Methods in Reinforcement Learning. Master's Thesis, Soochow University, Jiangsu, China, 2014.