energies

MDPI

Article

Designing an Incentive Contract Menu for Sustaining the Electricity Market

Ying Yu [1,*], Tongdan Jin [2] and Chunjie Zhong [1]

Received: 14 October 2015; Accepted: 7 December 2015; Published: 16 December 2015
Academic Editor: Ying-Yi Hong

[1] School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China; stephzcj@sina.cn
[2] Ingram School of Engineering, Texas State University, San Marcos, TX 78666, USA; tj17@txstate.edu
* Correspondence: squarey@shu.edu.cn; Tel.: +86-21-5633-1568; Fax: +86-21-6613-4021

Abstract: This paper designs an incentive contract menu to achieve long-term stability for electricity prices in a day-ahead electricity market. A bi-level Stackelberg game model is proposed to search for the optimal incentive mechanism under a one-leader and multi-followers gaming framework. A multi-agent simulation platform was developed to investigate the effectiveness of the incentive mechanism using an independent system operator (ISO) and multiple power generating companies (GenCos). Further, a Q-learning approach was implemented to analyze and assess the response of GenCos to the incentive menu. Numerical examples are provided to demonstrate the effectiveness of the incentive contract.

Keywords: stackelberg game; Q-learning; multi-agent simulation; electricity market; incentive mechanism

1. Introduction

As the vertically integrated power industry evolves into a competitive market, electricity now can be treated as a commodity governed by demand and generation interactions. Competition among generating companies (GenCos) is highly encouraged in order to lower the energy price and benefit end consumers. However, when GenCos are given more flexibility to choose their bidding strategies, larger uncertainties are also brought into the electricity markets. Many factors that affect bidding strategies include the risk appetite of Gen Cos, price volatility of fuels, weather conditions, network congestion and overloading. These factors and their interactions may lead to larger price volatility in the deregulated power market.

Many efforts have been made to design and optimize the bidding strategies in the presence of uncertainty. Zhang *et al.* [1] proposed an efficient decision system based on a Lagrangian relaxation method to find the optimal bidding strategies of GenCos. Kian and Cruz [2] modeled the oligopolistic electricity market as a non-linear dynamical system and used dynamic game theory to develop bidding strategies for market participants. Swider and Weber [3] proposed a methodology that enables a strategically behaving bidder to maximize the revenue under price uncertainty. Centeno *et al.* [4] used a scenario tree to represent uncertain variables that may affect price formation, which include hydro inflows, power demand, and fuel prices. They also presented a model to analyze GenCos' medium-term strategic analysis. In [5], a dynamic bid model was used to simulate the bidding behaviors of the players and study the inter-relational effects of the players' behaviors and the market conditions on the bidding strategies of players over time. Li and Shi [6] proposed an agent-based model to study the strategic bidding in a day-ahead electricity market, and found that applying learning algorithms could help increase the net earnings of the market participants. Nojavan and Zare [7] proposed an information gap decision theory model to solve the optimal

bidding strategy problem by incorporating the uncertainty of market price. Their case study further shows that risk-averse or risk-taking decisions could affect the expected profit and the bidding curve in the day-ahead electricity market. Qiu *et al.* [8] discussed the impacts of model deviations on the design of a GenCo's bidding strategies using the conjectural variation (CV) based methods, and further proposed a CV-based dynamic learning algorithm with data filtering to alleviate the influence of demand uncertainty. Kardakos *et al.* [9] point out that when making a bidding decision, a GenCo would take into accounts the behavior of its competitors as well as specific features and enacted rules of the electricity market. They further developed an optimal bidding strategy for a strategic generator in a transmission-constrained day-ahead electricity market. Other studies [10–12] emphasized that transmission constraints, volatile loads, market power exertions, and collusions may induce GenCos to bid higher prices than their true marginal costs, thereby aggravating the price volatility issue.

As concern for the sustainability of the power market increases, efforts also have been made to reduce the risk of price variation. Most studies focus on the employment of price-based demand response (DR) programs for the electricity users to control and reduce the peak-to-average load ratio [13–22]. For instance, Oh and Hildreth [13] proposed a novel decision model to determine whether or not to participate in the DR program, and further assessed the impact of the DR program on the market stability. Faria *et al.* [14] suggested that adequate tolls could motivate the potential market players to adopt the DR programs. Ghazvini *et al.* [15] showed that multi-objective decision-making is more realistic for retailers to optimize the resource schedule in a liberalized retail market. In [16], a two-stage stochastic programming model was formulated to hedge the financial losses in the retail electricity market. Zhong *et al.* [17] proposed a new type of DR program to improve social welfare by offering coupon incentives. The researchers in [18,19] handled the energy scheduling issue by optimizing the DR program in a smart grid environment. Yousefi *et al.* [20] proposed a dynamic model to simulate a time-based DR program, and used Q-learning methods to optimize the decisions for the market stakeholders. In [21–23], much more detailed reviews were provided for benefit analyses and applications of DR in a smart grid environment.

However, in electricity markets where the demand side is regulated, how to design and optimize GenCos' bidding strategies is treated as one of the most efficient ways to sustain the market price in the presence of uncertainty. Some studies have been made by proposing an incentive mechanism or contract for the GenCos to mitigate the risk of price variation caused by their subjective preferences during the bidding process [24–27]. Silva *et al.* [24] introduce an incentive compatibility mechanism, which is individually rational and feasible, to resolve the asymmetric information problem. Liu *et al.* [25] proposed an incentive electricity generation mechanism to control GenCos' market power and reduce the pollutant emissions using the signal transduction of game theory. Cai *et al.* [26] proposed a sealed dynamic price cap to prevent GenCos from exercising market power. Heine [27] performed a series of studies on the effectiveness of regulatory schemes in energy markets, and pointed out that potential improvements exist in contemporary systems when incentive-based regulations are appropriately implemented.

Although there is a large body of literature in bidding and incentive policy, most of the studies neglect assessment of the long-term effects of the incentive programs on the GenCos' learning behaviors. Besides, less attention is paid to the dynamic response of the GenCos to the volatile loads and incentive schemes. To maintain the market stability, it is highly desirable to understand the interplays between the incentive mechanism and the GenCos' adaptive responses to the variable market. This paper aims to fill this gap by proposing an incentive mechanism in a day-ahead power market to reduce price variance, and further assessing the subsequent long-term impacts of the incentive mechanism. To that end, a two-level Stackelberg gaming model was developed to analyze the bidding strategies of the market participants including one independent system operator (ISO) and several GenCos. An optimal menu of incentive contracts was derived under a one-leader and multi-followers game theoretic framework. Finally, a Stackelberg-based Q-learning approach was employed to assess the GenCos' response to the incentive-based generation mechanism.

The remainder of the paper is organized as follows: in Section 2, we introduce the menu of incentive contracts, and describe the workflow of the multi-agent game framework; in Section 3, we give a mathematical description of the problem, and present the details of the Stackelberg model; in Section 4, we use a Q-learning methodology to investigate the long-term effectiveness of the incentive contracts; in Section 5, numerical examples are provided to demonstrate the application and performance of the method; Section 6 concludes the paper.

2. Problem Statement

2.1. Description of the Menu of Incentive Contracts

A commercial agreement between the ISO and the GenCo is proposed, which defines a reward scheme in exchange for a consistent bidding behavior: the GenCo agreed to bid a reasonable power generation with a constant bidding curve during the contract period. Note that the reasonable power output should be larger than the regulated threshold of power output.

In general it is not efficient to design a uniform incentive contract with a constant threshold due to the fact that GenCos usually possess diverse bidding behaviors. It is also rather complex to design customized incentive contracts for all GenCos. One viable approach is to design a pertinent incentive menu comprised of key characteristic incentive contracts for certain target GenCos. These GenCos are selected as representatives from the entire group of power generators. Though the ISO cannot precisely predict the target GenCos' bidding information in a future bidding round, the customized incentive contracts still can be devised by incorporating the target GenCos' interest through inference of historical bidding data. For the target GenCo, its expected profit could be amplified if it complies with the incentive contract which takes into account its individual rational constraints and incentive compatibility constraints. Hence it is reasonable to assume that target GenCos would not reject the customized contract.

Though the incentive contracts in the menu are designed based on the individual rationality and incentive compatibility of the target GenCos, they also benefit non-target GenCos that are willing to accept the incentive contracts. For a non-target GenCo, the incentive contract would be appealing if the expected profit is higher by abiding with the agreement. To better illustrate how the menu of the incentive contracts works, some notations are given as follows.

2.1.1. Target GenCo and Contracted Generating Companies

The GenCos are termed the target GenCos if their individual rationality constraints and incentive compatibility constraints are considered so that they could be motivated to accept the incentive contract.

We define A^0 as set of all possible combinations of target GenCos. For each $a^0 \in A^0$, $a^0 = (a_1^0, a_2^0, ..., a_m^0)$, where a_i^0 represents whether GenCo i is chosen as a target GenCo or not. Further, we define $I = \{k | a_k^0 = 1, k \in M\}$ as a set of target GenCos.

Note that a is a combination of strategies of GenCos, $a = (a_1, a_2, ..., a_m)$, where a_i represents whether GenCo i decides to accept the menu of the incentive contracts or not. If $a_i = 0$, it is "not", and $a_i \neq 0$ is "yes". If $a_i \neq 0$, $a_i = k$, $k \in I$, meaning GenCo k accepts the incentive contract and becomes the target GenCo. The GenCos with $a_i \neq 0$ are termed as contracted GenCos.

2.1.2. Bidding Curve and Market Clearing Price

For electricity transactions, The study in [28] shows that GenCos submit power output in MW (Megawatts) along with associated prices for one bid in both discrete form and continuous form. A bid in discrete form with three different blocks is shown in Table 1. If the power output level is below 30 MW, the price is 10 $/MWh; If the power output level is between 30 MW and 60 MW, the price is 15 $/MWh; If the power output level is between 60 MW and 90 MW, the price is 20 $/MWh. Generally, this bid could also take a continuous form as shown in Figure 1. Without loss of generality,

a continuous bid curve model is adopted in this paper. For GenCo i, its bidding curve at time t is in the form of $p_{it} = \alpha_{it} + \beta_{it} q_{it}$, where α_{it} and β_{it} are the bidding coefficients of GenCo i at time t. Here p_{it} and q_{it} respectively, represent the bidding price and the bidding power output of GenCo i at time t.

Table 1. Block bid.

Blocks	Price ($/MWh)	Power output level (MW)
Block 0	10	30
Block 1	15	60
Block 2	20	90

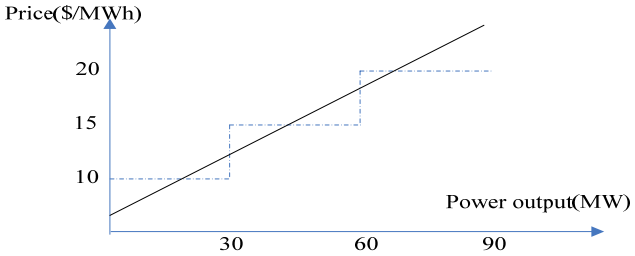

Figure 1. Block bid and continuous bid curves.

The market clearing price (MCP) is a uniform price shared by all GenCos, and the actual MCP depends on all GenCos' bidding behaviors. Assume the bidding curve of GenCo i is $p_{it} = \alpha_{it} + \beta_{it} q_{it}$, and the electricity demand at time t is D_t. The MCP at time t, which is denoted as p_t, can be obtained by solving the following power balance equation:

$$D_t = \sum_{i=1}^{m} q_{it} \tag{1}$$

$$p_t = \alpha_{it} + \beta_{it} q_{it}, \text{ for } i = 1, 2 \ldots, m \tag{2}$$

2.1.3. Menu of the Incentive Contracts

The menu of the incentive contracts is composed of multiple contracting terms in the form of (α_i, β_i, π_i), where α_i and β_i represent the thresholds of bidding coefficients respectively, and π_i is the relevant reward for meeting the incentive contract. Though each incentive contract is originally tailored to the rational and incentive-compatibility constraints of a certain target GenCo, it is also expected that these contracts are designed appropriately to motivate non-target GenCos to participate in the incentive program.

Assuming an incentive contract is customized for a target GenCo with bidding coefficients α_{i0} and β_{i0}, and the amount of the reward is π_{i0} by calculating the target GenCo's individual rationality and incentive compatibility conditions. This contract could be expressed by the triplet (α_{i0}, β_{i0}, π_{i0}) which specifies the reward and the obligation associated with the contracted GenCo: if the dispatched power output of the contracted GenCo during the contract period is always greater than the required level, which is prescribed as $\underline{q} = (p_t - \alpha_{i0})/\beta_{i0}$ with p_t being the MCP at time t, a reward of π_{i0} would be received at the end of the contract period.

All sets of (α_i, β_i, π_i) are further incorporated into (AL, B, π), where AL, B, π are the vectors of α_i, β_i, π_i, respectively. Hence the menu of the incentive contracts could be concisely specified in the form of (AL, B, π).

2.2. Scenario-Based Approach

At certain time, unexpected events like hot weather, network congestion, and demand spikes may occur randomly, which causes load soaring and demand forecast errors. Scenario based approach [16,29,30] is often employed to address these types of uncontrollable events. These uncertain events are characterized by scenarios with corresponding probabilities. The scenarios considered in this paper include both normal scenario and bad scenario. In the latter, the load is 20% higher than the average demand. Probability of each scenario could be inferred from historical data and experiences. The Monte Carlo method is adopted to simulate both normal and bad scenarios.

2.3. The Workflow of Multi-Agent System

In this paper, a multi-agent system, adapted to a simulated context with multiple GenCos and one ISO, is proposed to study a day-ahead electricity market based on the proposed incentive menu of contracts. The multi-agent system was developed in a Java platform that was partly inherited from the Repast platform [31]. Some actions of GenCos were coded by Matlab, and then packaged and implanted into the multi-agent Java platform.

Figure 2 shows the flowchart of multi-agent system (MAS) scheduling procedure.

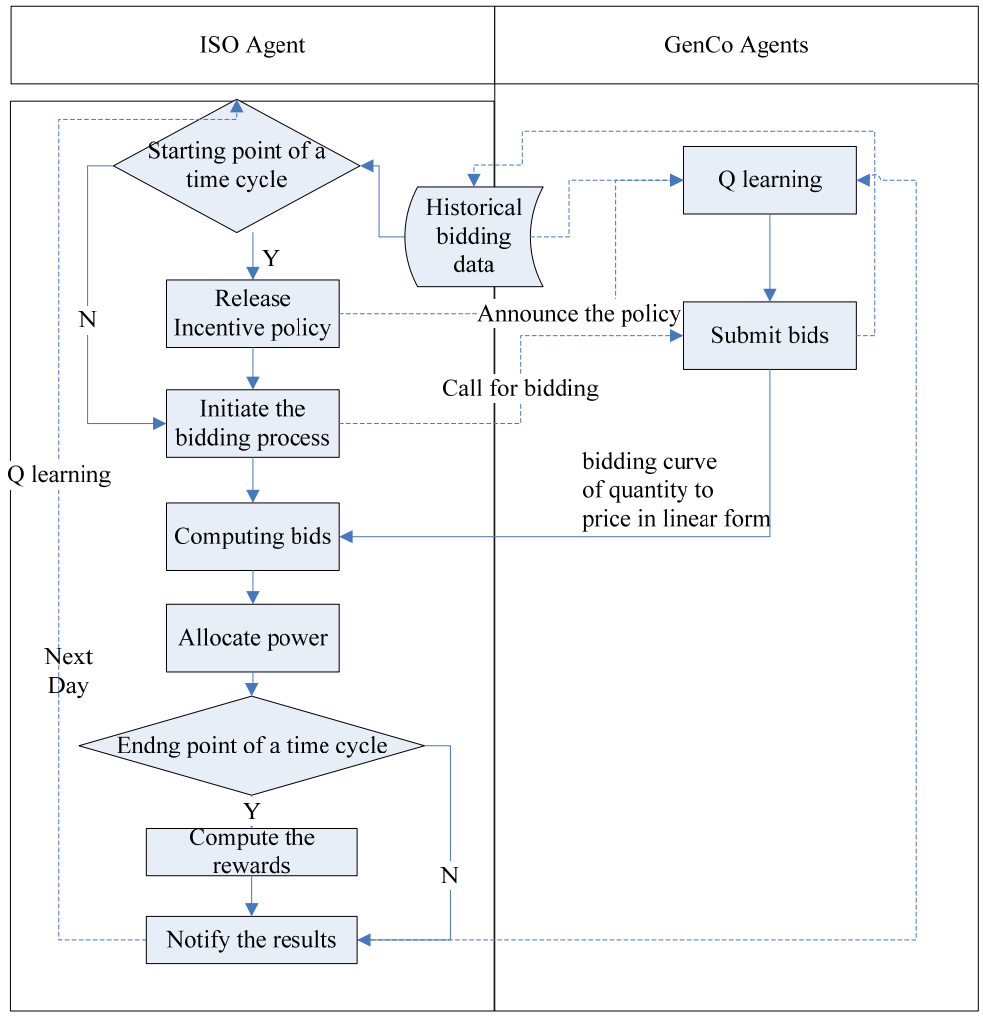

Figure 2. Flow chart of multi-agent system (MAS) scheduling mechanism (note: Y = yes and N = no). ISO: independent system operator.

At the beginning of a specified period $\widetilde{t}$ that consists of T days (*i.e.*, a period may include several days, or several months), ISO announces the menu of the incentive contracts. GenCos, which act for their own interest, decide whether they accept the incentive contract or not by using the periodic Q-learning method. This decision-making system resembles the one-leader and multi-followers Stackelberg game where the ISO is the leader and the GenCos are the followers. An algorithm using the idea of the Stackelberg game, which is further illustrated in Section 4.3, is presented for the ISO to find the initial optimal menu of incentive contracts in the first period. In addition, a periodic Stackelberg-based Q-learning method, which is illustrated in Section 4.2, is proposed for the ISO to find the subsequent optimal menu of incentive contracts over the following periods.

At the beginning of a period, GenCo i decides whether or not to accept the incentive program, by using its periodic Q-learning method. In each day of the period, GenCo i chooses to place a high bid or a normal bid by using its daily Q-learning method, and submits its bid, taking the form of $p_{it} = \alpha_{it} + \beta_{it} q_{it}$. After the ISO receives all the bids from the participants, the relevant information is aggregated and stored in a central repository. Based on the estimated hourly electricity demand of the next day, the ISO decides the unified hourly MCP of the next day, and announces the hourly power output schedule of individual GenCos for the next day.

At the end of period $\widetilde{t}$, the ISO computes the relevant rewards based on the bidding data retrieved from the central repository. If any GenCo's bidding data in the given period are constant, and always in alignment with certain contract in the menu of the incentive contracts, the GenCo would receive the relevant reward. During the repetitive bidding periods, both ISO and the GenCos improve their pricing policy and bidding strategies using the Q-learning algorithm.

3. Multi-Agent Stackelberg Game Model

3.1. Model Assumption and Description

Model assumptions are given as follows:

(1) To prevent GenCos reaping extra profits by modifying their bidding data to satisfy the incentive contract, it is stipulated that any GenCo using new bidding coefficients is not eligible to join the incentive program until after several rounds.

(2) At some time, due to the uncertainties in weather condition, network reliability and consumer behavior, unexpected demand spikes may occur, and the load may vary with a large degree of uncertainty. Probabilistic scenarios trees are adopted to accommodate the uncertain characteristics of the load profile. For instance, the electricity demand at time t is estimated to be 100 MW with probability of 0.8 for the normal demand scenario, and 150 MW with probability 0.2 for the high demand scenario, or bad scenario. Enumeration methods can be used to capture all possible scenarios if the problem size is not too large. Let Λ denote a set of uncertain scenarios, and λ_t denotes a realized scenario in Λ at time t. In addition, Λ^B is used to represent a set of bad scenarios.

(3) It is assumed that each GenCo within the MAS framework have two bidding options: either place a high bid (*i.e.*, $b_{i,t} = 1$) or place a normal bid (*i.e.*, $b_{i,t} = 0$), where $b_{i,t}$ is the bidding strategy of GenCo i at time t. The coefficients for different bidding options are defined as follows:

$$\alpha_{i,t} = \begin{cases} {\alpha_i}^c & b_{i,t} = 0 \\ {\alpha_i}^h & b_{i,t} = 1 \end{cases} \tag{3}$$

$$\beta_{i,t} = \begin{cases} {\beta_i}^c & b_{i,t} = 0 \\ {\beta_i}^h & b_{i,t} = 1 \end{cases} \tag{4}$$

where ${\alpha_i}^{\mathrm{c}}$, ${\beta_i}^{\mathrm{c}}$ are parameters of the normal bidding curve for GenCo i, and ${\alpha_i}^{\mathrm{h}}$, ${\beta_i}^{\mathrm{h}}$ are parameters of the corresponding high bidding curve. Obviously, if a GenCo has accepted an incentive contract, we have $b_{it} = 0$, $\alpha_{i,t} = {\alpha_i}^{\mathrm{c}}$, $\beta_{i,t} = {\beta_i}^{\mathrm{c}}$ for $t \in \widetilde{t}$.

3.2. Single-Period Decision-Making Model of GenCo

Based on the given menu of incentive contracts, in each time period, a GenCo tries to maximize its profit by choosing the best bidding strategy as follows:

$$\max \quad \prod_i (a_i) \tag{5}$$

For a GenCo who does not accept any incentive contract, its profit is given as:

$$\prod_i (a_i = 0) = \sum_{t \in \widetilde{t}} \sum_{\lambda_t \in \Lambda} (Y \times \rho(\lambda_t)) \tag{6}$$

where $Y = \prod_i(\lambda_t, a_i = 0)$, and $\rho(\lambda_t)$ represents the probability of λ_t, and π_k is the reward specified in the incentive contract for target GenCo k.

It is usually difficult for a GenCo to know the actual bidding behavior of others, but it is reasonable to assume that the probability of its competitors' decision can be inferred from historical data. Hence $Y = \prod_i(\lambda_t, a_i = 0)$ can be obtained as:

$$Y = \sum_{t \in \widetilde{t}, \lambda_t \in \Lambda} X \tag{7}$$

where:

$$\begin{aligned} X = {} & \mathrm{pos}_i({b_t}^{i,C}) \prod_{j \neq i} \mathrm{pos}_j(b_{j,t}) \left(p\left(\lambda_t, {b_t}^{i,C}\right) K_{i,t} - c_{i1} K_{i,t} - 0.5 c_{i2} {K_{i,t}}^2 \right) \\ & + \mathrm{pos}_i({b_t}^{i,H}) \prod_{j \neq i} \mathrm{pos}_j(b_{j,t}) \left(p\left(\lambda_t, {b_t}^{i,H}\right) K_{i,t} - c_{i1} \mathrm{K}_{i,t} - 0.5 c_{i2} {K_{i,t}}^2 \right) \end{aligned} \tag{8}$$

where $K_{i,t} = q_{i,t}(\lambda_t, b_{i,t})$ is the power output of GenCo i when its bidding action is $b_{i,t}$ in a scenario λ_t. c_{i1} and c_{i2} are cost coefficients of GenCo i. ${b_t}^{i,c} = \{b_{1,t}, b_{2,t}, b_{i-1,t}, 0, b_{i+1,t}, \ldots, b_{m,t}\}$ represent a bidding combination when GenCo i places a normal bid. Note that $pos_j(b_{j,t})$ is the probability for GenCo j to take action $b_{j,t}$. Here $p(\lambda_t, b_t(a))$ is the expected electricity price when the bidding combination of GenCos is $b_t(a)$ in scenario λ_t. Note that $p(\lambda_t, b_t(a))$ is $p(\lambda_t, {b_t}^{i,C})$ when GenCos' bidding action is ${b_t}^{i,C}$ in scenario λ_t, and is $p(\lambda_t, {b_t}^{i,\mathrm{H}})$ when GenCos' bidding action is ${b_t}^{i,H}$ in scenario λ_t.

For a contracted GenCo who agrees on the acceptance of an incentive contract which is tailored to the target GenCo k, its profit could be calculated as follows:

$$\Pi_i(a_i = k) = \sum_{t \in \widetilde{t}} \sum_{\lambda_t \in \Lambda_t} \rho(\lambda_t) (\Pi_i(\lambda_t, a_i = k)) + \pi_k \tag{9}$$

Assuming the incentive contract is prescribed as $(\alpha_k, \beta_k, \pi_k)$ triplet, we have:

$$K_{i,t} \geqslant \frac{p_t - \alpha_k}{\beta_k} \tag{10}$$

$$\mathrm{pos}_i({b_t}^{i,C}) = 1 \tag{11}$$

$$\mathrm{pos}_i({b_t}^{i,H}) = 0 \tag{12}$$

So $\Pi_i(\lambda_t, a_i = k)$ could be calculated as:

$$\Pi_i(\lambda_t, a_i = k) = \sum_{t \in \widetilde{t}} \left(\prod_{j \neq i} \mathrm{pos}_j(b_{j,t}) \left(p\left(\lambda_t, {b_t}^{i,\mathrm{C}}\right) \mathrm{K}_{i,t} - c_{i1} \mathrm{K}_{i,t} - 0.5 c_{i2} {\mathrm{K}_{i,t}}^2 \right) \right), k \in I \tag{13}$$

subject to:

$$K_{i,t} \geqslant \frac{p_t - \alpha_k}{\beta_k} \tag{14}$$

$$\mathrm{pos}_i(b_t{}^{i,\mathrm{C}}) = 1 \tag{15}$$

$$\mathrm{pos}_i(b_t{}^{i,\mathrm{H}}) = 0 \tag{16}$$

The optimization problem faced by a GenCo is how to choose an optimal bidding strategy such that its expected profit is maximized. Hence the incentive-compatibility constraint could be formulated as:

$$\mathbf{IC} : a_i = \mathrm{argmax}\left\{\Pi_i(a_i = 0), \Pi_i(a_i = k)\right\}, k \in I \tag{17}$$

If a GenCo accepts an incentive contract, its expected profit should be higher than the alternative. Thus the personal rationality constraint could be re-formulated as:

$$\mathrm{PC} : \Pi_i(a_i = 0) < \Pi_i(a_i = k) + \pi_k, \qquad k \in I \tag{18}$$

3.3. Optimization Problem of Independent System Operator

From the ISO's point of view, its goal is to design an optimal menu of incentive contracts such that the average MCP during the period remains at a relatively stable level, or the volatility of price in the worst scenarios could be mitigated, while the total electricity payment is minimized. To that end, it is necessary for the ISO to identify the optimal set of target GenCos (*i.e.*, a^0) as well as designing the incentive menu for attracting contracted GenCos, so that its objectives could be optimized. Since an incentive contract, which is specified in the triplet form of (α_i, β_i, π_i), is dependent upon a^0, how to target suitable GenCos is the key to designing an optimal incentive menu of contracts. Hence the ISO's initial decision is to choose optimal a^0, so as to minimize the total cost with certain price stability.

As the leader of the Stackelberg game, the ISO can analyze the response of the followers (*i.e.*, GenCos) so as to find the optimal decision variable a^0. A two-level programming model is proposed to facilitate ISO's decision-making. The sub-problem at the first level enables the ISO to minimize the total cost with price stability by finding an optimal value of a^0. The sub-problem at the second level can be treated as GenCos' reaction model upon the release of the menu of incentive contracts from the first level decision:

$$\min_{a^0 \in A^0} \left(C(a^0)\right) \tag{19}$$

$$\min_{a^0 \in A^0} \left((1-\delta) \times \mathrm{EP}(a^0) + \delta \times \mathrm{BP}(a^0)\right) \tag{20}$$

subject to:

$$a = a(a^0) = (a_1(a^0), a_2(a^0), ..., a_m(a^0)) \triangleq (a_1, ..., a_i, ..., a_m) \tag{21}$$

$$C\left(a^0\right) = \sum_{t \in \widetilde{t}} \sum_{\lambda_t \in \Lambda_t} \left(\rho\left(\lambda_t\right) P(\lambda_t, a) K_{i,t}\right) + \sum_{i=1}^{m} \pi\left(a_i\right) \tag{22}$$

$$\pi\left(a_i\right) = \begin{cases} \pi_k & a_i = k, k \in I \\ 0 & a_i = 0 \end{cases} \tag{23}$$

$$\mathrm{EP}\left(a^0\right) = \sum_{t \in \widetilde{t}} \sum_{\lambda_t \in \Lambda_t} \rho\left(\lambda_t\right) \times P\left(\lambda_t, a\right) \tag{24}$$

$$\mathrm{BP}(a^0) = \sum_{t \in \widetilde{t}} \sum_{\lambda_t \in \Lambda_t^B} \rho\left(\lambda_t\right) \times \left[P\left(\lambda_t, a\right) - \mathrm{EP}^*\right]^2 \tag{25}$$

$$\sum_{i=1}^{m} \mathrm{K}_{i,t} = D\left(\lambda_t\right) \tag{26}$$

$$P\left(\lambda_t, a\right) = \prod_{j \in M} \mathrm{pos}_j(b_{j,t}) p\left(\lambda_t, b_t\left(a\right)\right) \tag{27}$$

$$p\left(\lambda_t, b_t\left(a\right)\right) = \alpha_{i,t} + \beta_{i,t} \mathrm{K}_{i,t}, i \in M \tag{28}$$

$$b_t\left(a\right) = \left(b_{1,t}\left(a\right), b_{2,t}\left(a\right), \cdots, b_{m,t}\left(a\right)\right) \tag{29}$$

$$b_{i,t}\left(a_i\right) = \begin{cases} 0 & a_i > 0 \\ 1 \quad or \quad 0 & a_i = 0 \end{cases} \tag{30}$$

where a_i is obtained by solving follows:

$$\mathrm{IC}: a_i = \mathrm{argmax}\left\{\Pi_i\left(a_i\left(a^0\right)\right)\right\} \tag{31}$$

$$\text{s.t. PC}: \Pi_i(a_i = 0) < \Pi_i(a_i = k) + \pi_k, \qquad k \in I \tag{32}$$

$$K_{i,t} \geqslant \frac{p_t - \alpha_k}{\beta_k} \text{ for } a_i > 0 \tag{33}$$

$$\mathrm{pos}_i(b_t{}^{i,\mathrm{C}}) = 1 \text{ for } a_i > 0 \tag{34}$$

$$\mathrm{pos}_i(b_t{}^{i,\mathrm{H}}) = 0 \text{ for } a_i > 0 \tag{35}$$

where $C(a^0)$ is the total power purchasing cost when the combination of the target GenCos is a^0, and δ is a balance parameter. EP(a^0) is the expected electricity price when the combination of the target GenCos is a^0, and EP* is the best expected price. BP(a^0) is the variance of mean price *versus* EP* when the combination of the target GenCos is a^0. Here $P(\lambda_t,a)$ is the expected MCP when the combination of contracted GenCos' is a in scenario λ_t. The MCP is $p(\lambda_t, b_t(a))$ when the bidding behavior of GenCos is $b_t(a)$ in scenario λ_t.

As shown in Equation (19), one of the ISO's objectives is to minimize the electricity payment. Equation (20) is another objective of the ISO, that contains dual goals: Firstly, minimizing the common price in the contract period; and secondly minimizing the volatility of price. Both goals are combined by a balance factor. Equation (21) indicates that a is also decided by a^0. Equations (22) and (23) are the mathematical descriptions of the cost and the reward, respectively. Equation (24) calculates the average price in one period under multiple scenarios. Equation (25) calculates the variation of mean price *versus* EP* in one period in multiple scenarios. Equation (26) ensures that the electricity demand is always satisfied. Equation (27) computes the average price by multiplying the price for certain bid combination in a specified scenario with its occurrence possibility. Equations (28)–(30) provide the mathematical descriptions for $p\left(\lambda_t, b_t\left(a\right)\right)$, $b_{t(a)}$, $b_{i,t}\left(a_i\right)$, respectively. Equation (31) represents the GenCos' objective which is also their incentive-compatibility constraint with a_i being the decision variable for GenCo i. Equation (32) gives the personal rational constraint of the GenCos who is willing to accept an incentive contract. Finally, Equations (33)–(35) defines the constraints of contracted GenCos including power output capacity of contracted GenCos, and the possibilities of contracted GenCos to place high bids or normal bids.

A multi-objective optimization can be solved by turning it into a single objective model through appropriately assigning weight to each objective function. Using a weight w to combine the two objectives in Equations (19) and (20), the ISO's decision model could be further expressed as:

$$\max_{a^0 \in A^0} J = w\left(\frac{C_{\max} - C\left(a^0\right)}{C_{\max} - C_{\min}}\right) + (1 - w)\left(\frac{\mathrm{EPM}_{\max} - \mathrm{EPM}\left(a^0\right)}{\mathrm{EPM}_{\max} - \mathrm{EPM}_{\min}}\right) \tag{36}$$

subject to:

$$\mathrm{EPM}(a^0) = \left((1 - \delta) \times \mathrm{EP}\left(a^0\right) + \delta \times \mathrm{BP}(a^0)\right) \tag{37}$$

$$a = a(a^0) = (a_1(a^0), a_2(a^0), ..., a_m(a^0)) \triangleq (a_1, ..., a_i, ..., a_m) \tag{38}$$

$$C\left(a^0\right) = \sum_{t\in\widetilde{t}} \sum_{\lambda_t\in\Lambda_t} \left(\rho\left(\lambda_t\right) P(\lambda_t, a) K_{i,t}\right) + \sum_{i=1}^{m} \pi\left(a_i\right) \tag{39}$$

$$\pi\left(a_i\right) = \begin{cases} \pi_k & a_i = k, k \in I \\ 0 & a_i = 0 \end{cases} \tag{40}$$

$$\mathrm{EP}\left(a^0\right) = \sum_{t\in\widetilde{t}} \sum_{\lambda_t\in\Lambda_t} \rho\left(\lambda_t\right) \times P\left(\lambda_t, a\right) \tag{41}$$

$$\mathrm{BP}(a^0) = \sum_{t\in\widetilde{t}} \sum_{\lambda_t\in\Lambda_t^B} \rho\left(\lambda_t\right) \times \left[P\left(\lambda_t, a\right) - \mathrm{EP}^*\right]^2 \tag{42}$$

$$\sum_{i=1}^{m} \mathrm{K}_{i,t} = D\left(\lambda_t\right) \tag{43}$$

$$P\left(\lambda_t, a\right) = \prod_{j\in M} \mathrm{pos}_j(b_{j,t}) p\left(\lambda_t, b_t\left(a\right)\right) \tag{44}$$

$$p\left(\lambda_t, b_t\left(a\right)\right) = \alpha_{i,t} + \beta_{i,t}\mathrm{K}_{i,t}, i \in M \tag{45}$$

$$b_t\left(a\right) = \left(b_{1,t}\left(a\right), b_{2,t}\left(a\right), \cdots, b_{m,t}\left(a\right)\right) \tag{46}$$

$$b_{i,t}\left(a_i\right) = \begin{cases} 0 & a_i > 0 \\ 1 \quad or \quad 0 & a_i = 0 \end{cases} \tag{47}$$

where a_i is obtained by solving follows:

$$\mathrm{IC}: a_i = \mathrm{argmax}\left\{\Pi_i\left(a_i\left(a^0\right)\right)\right\} \tag{48}$$

$$\text{s.t. PC}: \Pi_i(a_i = 0) < \Pi_i(a_i = k) + \pi_k, \qquad k \in I \tag{49}$$

$$K_{i,t} \geqslant \frac{p_t - \alpha_k}{\beta_k} \text{ for } a_i > 0 \tag{50}$$

$$\mathrm{pos}_i(b_t{}^{i,\mathrm{C}}) = 1 \text{ for } a_i > 0 \tag{51}$$

$$\mathrm{pos}_i(b_t{}^{i,\mathrm{H}}) = 0 \text{ for } a_i > 0 \tag{52}$$

where $C_{\max}$ is the maximum available $C(a^0)$; $C_{\min}$ is the minimum available $C(a^0)$; $\mathrm{EPM}(a^0)$ is a balance between price minimization and price variation minimization when the decision variable is a^0 . $\mathrm{EPM}_{\max}$ is the maximum available $\mathrm{EPM}(a^0)$; and $\mathrm{EPM}_{\min}$ is the minimum available $\mathrm{EPM}(a^0)$. Equations (38)–(52) are the same with Equations (21)–(35).

4. Q-Learning for Agents' Optimal Decision Making

Each agent interacts in the volatile market environment due to the uncertain load and lack of precise knowledge of its competitors. It is imperative for the agents to evolve their actions through the learning of repeated bidding processes. Q-learning is one of the reinforcement learning methods, and could guide the agents to improve the performance of their decision making over time. In each period, an agent perceives the state of the market environment, and takes certain actions based on its perception and past experience, which result in a new state. This sequential learning process would reinforce its subsequent actions. Quite a few studies have been done on Q-learning, and its application in the electricity market has been reported. For instance, Rahimiyan and Mashhadi [32] propose a fuzzy Q-learning method to model the GenCos' strategic bidding behavior in a competitive market condition, and find that GenCos could accumulate more profit by using fuzzy Q-learning. Naghibi-Sistani *et al.* [33] developed a modified reinforcement learning based

on temperature variation, and applied it to the electricity market to determine the GenCos' optimal strategies. Attempts also have been made to combine the Q-learning with Nash-Stackelberg games for reaching a long-run equilibrium. Haddad *et al.* [34] incorporate a Nash-Stackelberg fuzzy Q-learning into a hierarchical and distributed learning framework for decision-making, with which mobile users are guided to enter the equilibrium state that optimizes the utilities of all the network participants.

In this paper, Q-learning methods are adopted by ISO and GenCos for the making decisions. Different learning algorithms are designed for ISO and GenCos because they have different goals. For GenCos, both a periodic Q-learning method and a daily Q-learning method are applied to the bidding decision process. For ISO, a Stackelberg-based Q-learning is adopted to design the menu of incentive contracts in each period.

4.1. Periodic and Daily Q-Learning Methods for Generating Companies

At the starting point of a period, a GenCo decides whether the incentive contract should be accepted or declined. In each day of the period, the GenCo should choose to place a high bid or place a normal bid. Especially, if a contracted GenCo decides to place a high bid, the reward at the end of the period, would be cancelled. To calculate the potential reward, a multi-step Q-learning method is adopted by the GenCo to decide its bidding strategy in daily basis. Two Q-learning methods are proposed for the GenCo's periodic and daily decision making. The state, actions, reward and Q-value function are defined as follows:

4.1.1. State Identification

State $s_{\widetilde{t}}$ is defined for GenCo's Q-learning method for a period, and it is composed of values of all possible average electricity prices over one period.

State s_t denote the states for GenCo's Q-learning method in each day, and it is composed of values of all possible average electricity prices over one day.

4.1.2. Action Selection

Let a discrete set of actions $a_{i,\widetilde{t}} = \{0, k\}, k \in I$, denote the action selection of GenCo i at the starting point of a period for GenCo's Q-learning method for a period. When $a_{i,\widetilde{t}} = 0$, GenCo i chooses not to accept the incentive contract over period $\widetilde{t}$. When $a_{i,\widetilde{t}} = k, k \in I$, GenCo i accepts the incentive contract which is tailored to the target GenCo k over period $\widetilde{t}$.

$a_{i,t}$ denotes the action selection of GenCo i in each day for GenCo's Q-learning method for a day, and its value is 0 or 1. When $a_{i,t} = 1$, GenCo i adopts a normal bidding strategy; When $a_{i,t} = 0$, GenCo i adopts a high bidding strategy.

When $a_{i,\widetilde{t}} \neq 0$, in each day of the period $\widetilde{t}$, GenCo i places a normal bidding strategy. So for a contracted GenCo, $a_{i,t} = 1$, with a high probability.

4.1.3. Reward Calculation

The periodic reward function for Q-learning method over period $\widetilde{t}$ is defined as:

$$r(s_{\widetilde{t}}, a_{i,\widetilde{t}}) = \sum_{t \in \widetilde{t}} \Pi_{i,t}(a_{i,t}) + R(a_{i,\widetilde{t}}) \tag{53}$$

Equation (53) represents the reward assigned to action $a_{i,\widetilde{t}}$ from the old state $s_{\widetilde{t}}$. If $a_{i,\widetilde{t}} = 0$, which means that the menu of incentive contracts is not accepted over period $\widetilde{t}$, $R(a_{i,\widetilde{t}}) = 0$. If $a_{i,\widetilde{t}} = k, k \in I$, which means that GenCo i accepts the incentive contract which is tailored to the target GenCo k over period $\widetilde{t}$, an amount of $R(a_{i,\widetilde{t}})$ is received as the reward for meeting the incentive contract. The reward would further influence the periodic Q-value which guides the GenCo to determines the next action as whether or not to accept the incentive menu.

Every day the GenCo agent evaluates the current state, and chooses the best action that optimizes its objectives. Then the current state evolves to the new state, with a transition probability, and the agent receives a reward.

The reward r for daily Q-learning is made up of two parts. One is the direct profit subject to all the GenCo agents' bidding behaviors, loads, and cost of the GenCo agent. The second is a portion of the expected reward if the GenCo accepts the incentive contract. If a GenCo agent accepts the incentive menu and fulfills the contractual obligations in the contract period, a reward would be obtained at the end of the period, so the reward is a delayed reward. A multi-step reward function, which captures the characteristic of the delayed reward, is defined to describe GenCo agent's daily Q-learning as follows:

$$r(s_t, a_{i,t}) = \Pi_{i,t}(a_{i,t}) + R(a_{i,t}) \tag{54}$$

subject to:

$$R(a_{i,t}) = \begin{cases} \Gamma & \text{for } \prod\limits_{t\in\widetilde{t}} a_{i,t} = 1 \\ 0 & \text{for } \prod\limits_{t\in\widetilde{t}} a_{i,t} = 0 \end{cases} \tag{55}$$

where,

$$\begin{aligned} \Gamma &= T_s \frac{1}{T}\pi\left(a_{i,t}\right) + (T - T_s)\left(\varphi\frac{1}{T}\pi\left(a_{i,t}\right) + \varphi^2\frac{1}{T}\pi\left(a_{i,t}\right) + \ldots + \varphi^{T-T_s}\frac{1}{T}\pi\left(a_{i,t}\right)\right) \\ &= \frac{1}{T}\pi(a_{i,t})\left[T_s + (T - T_s)\sum_{i=1}^{T-T_s}\varphi^i\right], (T_s = 1, 2, \ldots, T) \end{aligned} \tag{56}$$

where φ is a discount factor, and $\pi(a_{i,t})$ is the reward for GenCo i to meet the incentive contract terms at time t, and T is the total number of days in a contract period, and T_s is the number of the days elapsed in the period.

4.1.4. Q-Value Update

By Q-learning, using the Bellman optimality in Equations (57) and (58), each GenCo agent tries to find the optimal action to maximize the Q-value of each state in a long run.

The periodic Q-value function defined for GenCo i over period $\widetilde{t}$ is given as follows:

$$Q_{\widetilde{t}+1}(s_{\widetilde{t}}, a_{i,\widetilde{t}}) = Q_{\widetilde{t}}(s_{\widetilde{t}}, a_{i,\widetilde{t}}) + \ell_{\widetilde{t}}\left[r(s_{\widetilde{t}}, a_{i,\widetilde{t}}) + \gamma_{\widetilde{t}}\max_{a_{i,\widetilde{t}+1}} Q(s_{\widetilde{t}+1}, a_{i,\widetilde{t}+1}) - Q_{\widetilde{t}}(s_{\widetilde{t}}, a_{i,\widetilde{t}})\right] \tag{57}$$

where $\ell_{\widetilde{t}}$ is a positive learning rate at period $\widetilde{t}$, and $\gamma_{\widetilde{t}}$ is a discount parameter at period $\widetilde{t}$.

These action-state value functions $Q_{\widetilde{t}+1}(s_{\widetilde{t}}, a_{i,\widetilde{t}})$ (i = 1, ... m), which are greatly affected by the reward function as illustrated in Equation (53), determine the GenCo agents' most suitable actions for the next run. That is, if the Q-value for accepting the incentive menu is less than the Q-value for not accepting it, the GenCo agent would not take the action of accepting the incentive menu. Conversely, it would. The daily Q-value function defined for GenCo i at each day is given as follows:

$$Q_{t+1}(s_t, a_{i,t}) = Q_t(s_t, a_{i,t}) + \ell_t\left[r(s_t, a_{i,t}) + \gamma_t\max_{a_{i,t+1}} Q(s_{t+1}, a_{i,t+1}) - Q_t(s_t, a_{i,t})\right] \tag{58}$$

where ℓ_t is a positive learning rate in day t, and γ_t is a discount parameter in day t.

4.2. Q-Learning for the Leader of the Stackelberg Game (Independent System Operator)

4.2.1. State Identification

State $\overrightarrow{s}_{\widetilde{t}} = \left\{\left(s_{\widetilde{t}}, a^0\right)\right\}$ for ISO's Q-learning is composed of two state variables, one is the values of all possible average electricity prices during that period, and the other is the decision variable for menu of the incentive contracts.

4.2.2. Action Selection

The set of all possible combinations of target GenCos, or A^0 is defined as the set of action selection of ISO agent. The ISO takes the action at each step, or at the starting point of each period. $\left(a^0\right)_{\widetilde{t}}$ denotes the action selection of ISO in period $\widetilde{t}$.

4.2.3. Reward Calculation

Reward function $r\left(s_{\widetilde{t}}, \left(a^0\right)_{\widetilde{t}}\right)$ is given by:

$$r\left(s_{\widetilde{t}}, \left(a^0\right)_{\widetilde{t}}\right) = w\left(\frac{C_{\max} - C\left(\left(a^0\right)_{\widetilde{t}}\right)}{C_{\max} - C_{\min}}\right) + (1-w)\left(\frac{\mathrm{EPM}_{\max} - \mathrm{EPM}\left(\left(a^0\right)_{\widetilde{t}}\right)}{\mathrm{EPM}_{\max} - \mathrm{EPM}_{\min}}\right) \tag{59}$$

$$\text{s.t. } a = a\left[\left(a^0\right)_{\widetilde{t}}\right] = \left[a_{1,\widetilde{t}}\left(a^0\right)_{\widetilde{t}}, a_{2,\widetilde{t}}\left(a^0\right)_{\widetilde{t}}, \ldots, a_{m,\widetilde{t}}\left(a^0\right)_{\widetilde{t}}\right] \triangleq \left(a_{1,\widetilde{t}}, a_{2,\widetilde{t}}, \ldots, a_{m,\widetilde{t}}\right) \tag{60}$$

$$a_{i,\widetilde{t}} = \operatorname{argmax}(Q_{\widetilde{t}+1}(s_{\widetilde{t}}, a_{i,\widetilde{t}})) \tag{61}$$

Equations (59)–(61) illustrates that $a_{i,\widetilde{t}}$, which is the GenCo i's action in period t, depends on its Q-learning, and so the reward of the ISO is obtained by using a Stackelberg-based Q-learning method.

4.2.4. Q-Value Update

As the leader of the Stackelberg game, Q-learning algorithm for ISO is given as follows:

$$Q^0{}_{\widetilde{t}+1}\left(s_{\widetilde{t}}, \left(a^0\right)_{\widetilde{t}}\right) = Q^0{}_{\widetilde{t}}\left(s_{\widetilde{t}}, \left(a^0\right)_{\widetilde{t}}\right) + \ell_{\widetilde{t}}\left[r_{\widetilde{t}}\left(s_{\widetilde{t}}, \left(a^0\right)_{\widetilde{t}}\right) + \gamma \max_{\left(a^0\right)_{\widetilde{t}+1}} Q^0\left(s_{\widetilde{t}+1}, \left(a^0\right)_{\widetilde{t}+1}\right) - Q^0{}_{\widetilde{t}}\left(s_{\widetilde{t}}, \left(a^0\right)_{\widetilde{t}}\right)\right] \tag{62}$$

4.3. Solution Methodology for Independent System Operator's Initial Q Value

For problems with multiple decision variables, chaos search is more capable of hill-climbing and escaping from the local optima than the random search [35]. Hence a chaos optimization algorithm is proposed to solve the problem

The detailed procedure of the algorithm is given as follows:

Step 1: set initial parameters incorporating bidding coefficients of GenCos and its power capacity.

Step 2: set $\nu = 1$.

Step 3: generate a non-zero chaos variable $\eta_{\nu+1}$ using cube mapping method as shown below:

$$\eta_{\nu+1} = 4\eta_\nu^3 - 3\eta_\nu \tag{63}$$

Step 4: decoding the chaos variables into a binary variable which represents a value for the sets of target GenCos.

Step 5: calculate the tailoring values of $(\alpha_i, \beta_i, \pi_i)$ for all target GenCos using Equation (18).

Step 6: for the designed menu of the incentive contracts, check each GenCo's optimal reaction by solving Equation (17).

Step 7: calculate the corresponding objective value and the state $\overrightarrow{s}_{\widetilde{t}}$ for ISO's Q-learning. The latter includes the mean electricity price during the period, and the menu of the incentive

contracts. If the obtained objective value is larger than the existing one, substitute the existing one.

Step 8: substitute the chaos variables into Equation (64) to yield new chaos variables:

$$x_\nu = c_\nu - \left[d_\nu \eta_{\nu+1}\right] \tag{64}$$

where c_ν and d_ν are two constant vectors, and $\left[d_\nu \eta_{\nu+1}\right]$ is the integr part of $d_\nu \eta_{\nu+1}$.

Step 9: Set $\nu = \nu + 1$, $k = k + 1$.

Step 10: If $\nu > \nu_{\max}$, stop searching, else go to Step 4.

5. Simulations and Analysis

The simulation is performed in a day-ahead electricity market with the participation of five GenCos. The original probability for a GenCo to bid high or normal is 0.5. Electricity demand at each hour varies between 170 MW and 230 MW. The probability of high-demand scenarios, which are also termed as bad scenarios, is less than 0.2. The length of a contract could be several days, or a couple of months. For computational convenience, firstly it is assumed that each period consists of seven days or one week.

Parameters of GenCos' bidding curves are listed in Table 2, and the GenCos' cost parameters are listed in Table 3, and the weights of the two objectives are listed in Table 4. Three cases are investigated for the comparative analysis.

Case 1: no menu of incentive contracts or Q-learning.

Case 2: menu of incentive contracts without Q-learning in one period. Eight sub-cases are further analyzed, and the comparative results are listed in Tables 5 and 6.

Case 3: menu of incentive contracts with Q-learning in multiple periods. Note that load demand over the multiple periods varies between 170 MW and 230 MW.

Table 2. Parameters of GenCos' bidding curves. (Unit for $\alpha_{i,t}{}^{c}$, $\alpha_{i,t}{}^{h}$: \$/MW per hour, unit for $\beta_{i,t}{}^{c}$, $\beta_{i,t}{}^{h}$: \$/(MW)^2 per hour).

GenCo No.	Case 1, Cases 2.1–2.5, Case 3				Cases 2.6–2.8			
	$\alpha_{i,t}{}^{c}$	$\beta_{i,t}{}^{c}$	$\alpha_{i,t}{}^{h}$	$\beta_{i,t}{}^{h}$	$\alpha_{i,t}{}^{c}$	$\beta_{i,t}{}^{c}$	$\alpha_{i,t}{}^{h}$	$\beta_{i,t}{}^{h}$
1	10	0.5	10.5	0.525	10	0.5	10.5	0.525
2	11	0.8	12	0.84	11	0.8	12	0.84
3	8	0.6	8.4	0.63	8	0.9	8.4	0.945
4	15	0.5	15.75	0.525	15	0.5	15.75	0.525
5	20	0.9	21	0.945	20	0.6	21	0.63

Table 3. GenCos' cost parameters. (Unit for c_{i1}: \$/MW, unit for c_{i2}: \$/(MW)^2).

GenCo No.	Cases 2.1–2.4 and 2.6–2.8		Case 2.5	
	c_{i1}	c_{i2}	c_{i1}	c_{i2}
1	10	0.5	5	0.25
2	11	0.55	6	0.35
3	8	0.5	4	0.25
4	15	0.8	7	0.4
5	20	0.9	10	0.45

Table 4. Objectives of Cases 2.1–2.8.

Cases	Cost	EPM (0.5 × EP + 0.5 × BP)
2.1, 2.5, 2.6	1	0
2.2, 2.7	0	1
2.3, 2.8, 3	0.5	0.5
2.4	Without menu of incentive contracts	

Table 5. Comparative results for Cases 2.1–2.3 and 2.5 (unit: $,). (Note: Y = saying "Yes" to offer of the incentive contract menu and N = saying "No" to the offer).

Items	Case 2.1				Cases 2.2 and 2.3							Case 2.5			
Target GenCos	1	0	0	0	1	1	0	0	1	1	0	1	0	1	1
GenCos' response	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Expected reward	208,750				543,530							123,070			
Expected cost saving (compared with Case 2.4)	597,830				512,120							932,580			
Expected price (EP)	37.17				36.94							36.94			
BP (mean price variance in bad scenarios)	2.78				2.53							2.53			
EPM (0.5 × EP + 0.5 × BP)	19.97				19.74							19.74			

Table 6. Comparative results for Cases 2.6–2.8 (unit: $, N/A = Not Applicable).

Items		Case 2.6				Cases 2.7 and 2.8					
Target GenCos		0	0	1	0	1	0	0	1	1	1
GenCos' response		Y	Y	Y	N	Y	Y	Y	Y	Y	Y
Expected reward		215,980							581,130		
Expected cost saving (compared with Case 2.9)		607,780							497,270		
EP		38							36.94		
BP (meanprice variance in bad scenarios)		2.78							2.53		
EPM (0.5 × EP + 0.5 × BP)		20.39							20.15		
Threshold for GenCo's power output	1	6558							6477		
	2	7145							7091		
	3	7145							7091		
	4	N/A							9852		
	5	6558							6477		

Case 2.1 aims at minimizing cost. Cost in Case 2.1 is less than that in Cases 2.2 and 2.3, but EP and BP in Case 2.1 are larger compared with that in Cases 2.1 and 2.3.

In Case 2.5, though it also aims at minimizing cost, as GenCos have small cost coefficient, they could gain more profits compared with Cases 2.1–2.4, and so GenCos in Case 2.5 prefer to make normal bids since they could obtain more power output and hence more profit, and so both cost and EPM could be minimized.

Since GenCo 3 has higher bidding coefficients in Cases 2.6–2.8, it has more influence on MCP than in Cases 2.1–2.5. Hence GenCo 3 is more likely to be the target GenCo in Cases 2.6–2.8 than in Cases 2.1–2.5.

Figures 3 and 4 show the simulation results of variations in price and cost across 112 days in Cases 1 and 3.

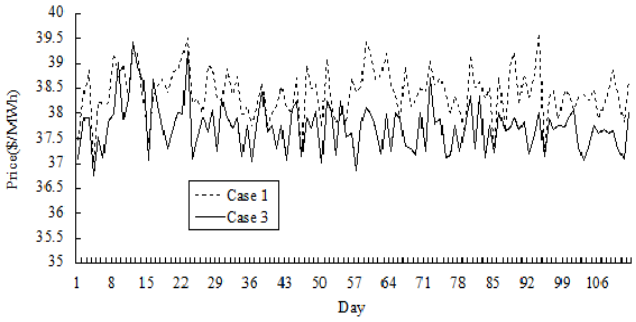

Figure 3. Comparative results of price variation for Cases 1 and 3 (for 112 days).

It can be seen that the variations of electricity price and cost could be reduced in the long term provided that the incentive contract is adopted. It could be seen that in the early phases, the effect of the incentive contract is not obvious as the daily price and daily electricity purchase cost do not significantly decrease in Case 3. However, as time evolves, GenCos can enhance their bidding experience through learning from past bidding processes and realize that accepting the incentive contract could help improve their profitability. They become more interested in participating in the incentive program. As a result, the electricity price is kept at a low and stable level.

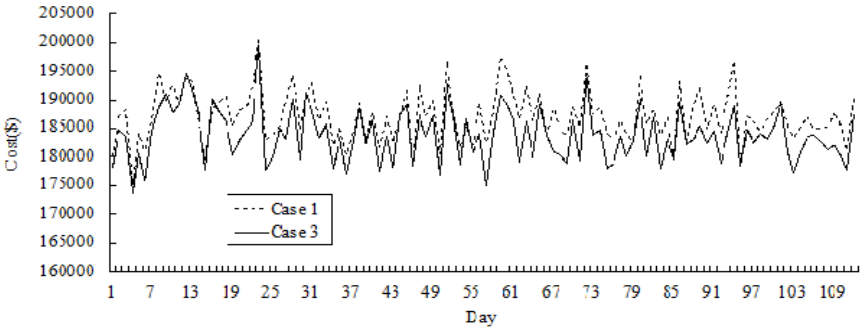

Figure 4. Comparative results of cost variation for Cases 1 and 3 (for 112 days).

Figures 5 and 6 show the simulation results of the variations of price and the cost across 112 days in Cases 2 and 3. It could be seen that in the early periods, the cost in Case 3 may be higher than that in Case 2 over certain number of days, and the price in Case 3 is higher than that in Case 2. As GenCos and ISO accumulate more bidding experiences by Q-learning, optimum decisions in Case 3 could be made by both players, and the cost and the price could be reduced compared with that in Case 2.

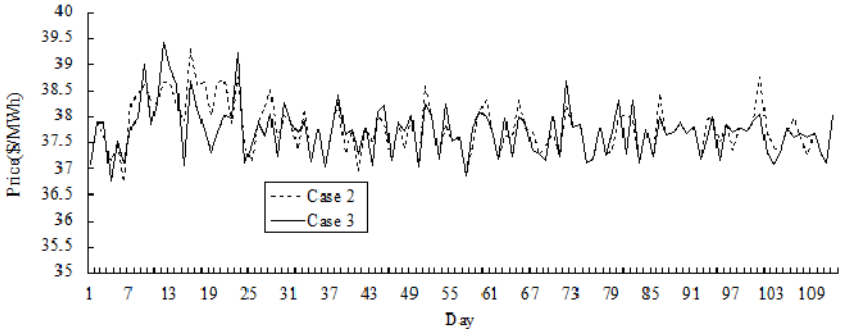

Figure 5. Comparative results of price variation for Cases 2 and 3 (for 112 days).

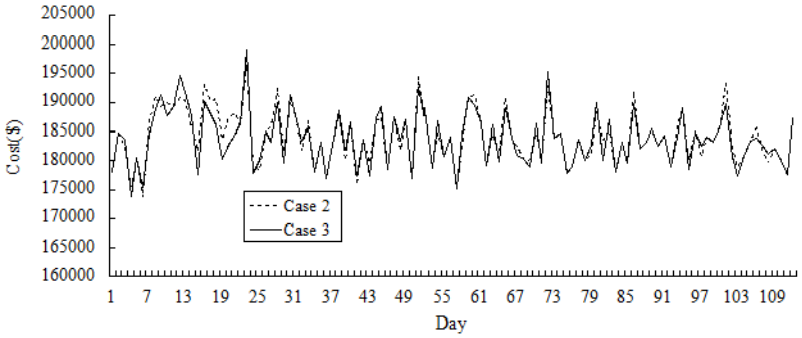

Figure 6. Comparative results of cost variation for Cases 2 and 3 (for 112 days).

Figures 7 and 8 show the comparative results of average price variation and average cost variation for three cases, respectively. Based on Case 3, it could be seen that both the cost and the price could be reduced and remain stable in a long run.

Extending the length of the contract period to 14 days, the comparative results for the duration of 224 days (*i.e.*, 16 periods) are shown in Figure 9, and it could be seen that price variation in the electricity market with incentive mechanism is less than the market without incentive mechanism. In fact, the price variance in the former market is 0.242 *versus* 0.270 in the latter, and the average price in the former market is 38.33 *versus* 38.50 in the latter (price unit is $/MWh).

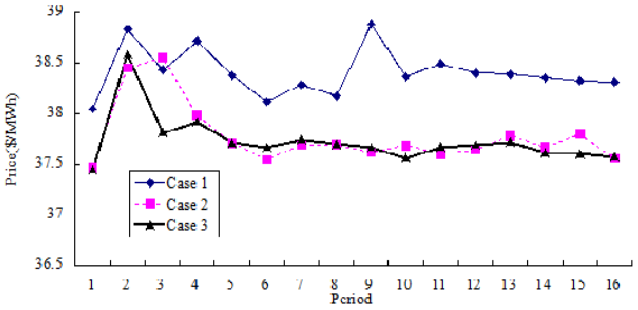

Figure 7. Comparative results of average price variation for 3 Cases (for 16 periods).

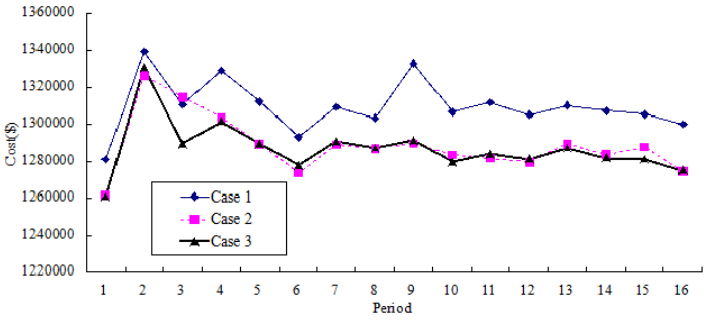

Figure 8. Comparative results of average cost variation (for 16 periods).

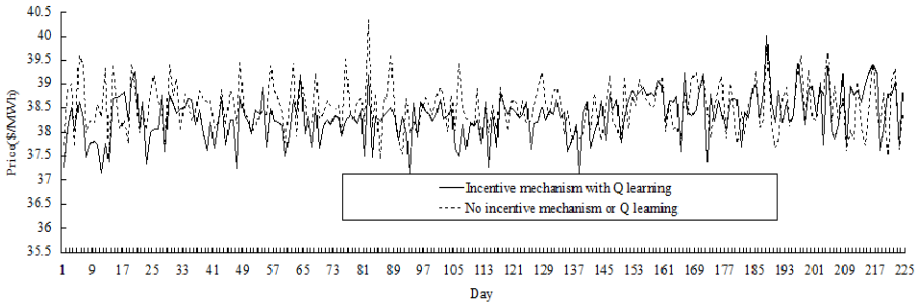

Figure 9. Comparative results of price variation (for 224 days).

Extending the single contract period to 28 days, the comparative results for 448 days (*i.e.*, 16 periods) are shown in Figure 10. It could be seen that price variation in the electricity market with incentive mechanism is less than the market without incentive mechanism. In fact, the price variance in the former market is 0.252 *versus* 0.310 in the latter, and the average price in the former market is 37.86 *versus* 38.43 in the latter (price unit is $/MWh).

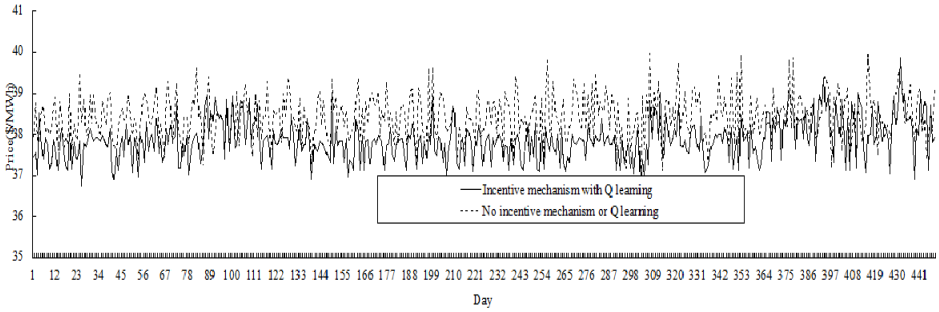

Figure 10. Comparative results of price variation (for 448 days).

6. Conclusions

In this paper a menu of incentive contracts is presented in a Stackelberg game model, aiming at seeking an incentive bidding mechanism with which the electricity price could be kept at a low and stable level. To ensure market equilibrium, a Stackelberg game based Q-learning is proposed for the ISO to analyze the responses of GenCos to the market as well as searching for the optimal menu of incentive contracts. For GenCos, a periodic Q-learning method is adopted to determine whether the incentivized menu should be accepted or not. In addition, a multi-step Q-learning method is adopted by the GenCo to decide its daily bidding policy. Based on the multi-agent platform, the long-term effectiveness of the incentive program is validated up to 14 months using simulation methods. Numerical results show that an incentivized menu, which is suitably designed by the ISO with the perspective of a central planner, could lead to desirable bidding behavior of GenCos, and hence guarantees the market sustainability. When multiple types of Q-learning methods are adopted by the ISO and the GenCos for decision makings, both the electricity price and purchasing cost could be reduced. Hence a desirable trade-off between the price variation and the purchasing cost could be reached at equilibrium. Future efforts could be directed to analyzing GenCos' reactions to the menu of incentive contracts under different risk preferences or generation uncertainties with wind and solar power integration.

Acknowledgments: This research is supported by National Natural Science Funds of China (Grant No. 71201097) and Action plan for scientific and technological innovation Program of Science and Technology Commission Foundation of Shanghai (Grant No. 15511109700). We would like to thank the anonymous reviewers and the editor for their valuable time and constructive comments for the improvement of the original manuscript.

Author Contributions: This paper designs an incentive contract menu to achieve long-term stability for electricity prices in a day-ahead electricity market. A bi-level Stackelberg game model is proposed to search for the optimal incentive mechanism under a one-leader and multi-followers gaming framework. A multi-agent simulation platform was developed to investigate the effectiveness of the incentive mechanism using an independent system operator (ISO) and multiple power generating companies (GenCos). Further, a Q-learning approach was implemented to analyze and assess the responses of GenCos to the incentive menu.

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

Notations

N1. Set Parameters

A^0	Set of all possible a^0, which denotes a combination of target GenCos.
M	Set of all serial numbers of GenCos.
I	Set of target GenCos, $I = \{k \mid a_k^0 = 1, k \in M\}$.
Λ	Set of uncertain scenarios during a bidding period.
Λ^B	Set of all bad scenarios.
(AL, B, π)	Data set for the menu of incentive contracts.

N2. Decision Variables

a_i^0	Whether a GenCo is chosen as the target GenCo. $a_i^0 = 1$ is "yes" and 0 is "not".
a^0	$a^0 = (a_1^0, a_2^0, ..., a_m^0)$, $a^0 \in A^0$.
a_i	*Whether GenCo* i *accepts the incentive menu.* $a_i = 0$ *means "not", and* $a_i \neq 0$ *means "yes". Moreover, if* $a_i \neq 0$, $a_i = k$, $k \in I$, *meaning GenCo* i *accepts the incentive contract with GenCo* k *as the target GenCo.*
a	$i \in M$.

N3. Model Parameters

$C(a^0)$	Total electricity purchasing cost when the combination of the target GenCos is a^0.
$\pi(a_i)$	Award received by GenCo i.
t	Time interval in one period.
$\widetilde{t}$	Length of contract period.
λ_t	A certain scenario in Λ at time t.
$\rho(\lambda_t)$	Probability of λ_t.
$EP(a^0)$	Expected electricity price when the combination of the target GenCos is a^0.
EP*	The best expected price.
δ	Balance parameter.
w	Weight of the objective functions.
$BP(a^0)$	Variance of mean price in bad scenarios *versus EP** when the set of the target GenCos is a^0.
EPM	A representative symbol of the model objective which combines the expected price (EP) and robustness of the price (BP) with a balance factor.
$P(\lambda_t, a)$	In scenario λ_t, the expected market price when the set of the GenCos' options for the accepted incentive contract is a.
m	Number of GenCos.
$K_{i,t}$	Power output of GenCo i when the bidding combination of GenCos is $b_t(a)$ in scenario λ_t.
α_{it}, β_{it}	Parameters of GenCo i's bidding curve at time t.

b_{it}	Bidding strategy of GenCo i at time t. $b_{it} = 1$ means a high bidding; and $b_{it} = 0$ implies a normal bidding.
$\text{pos}_i(b_{it})$	Probability for GenCo i to make a bid b_{it} at time t.
$\Pi_i(a_i)$	Expected profit of GenCo i when its contract decision is a_i.
$\Pi_i(\lambda_t, a_i)$	Expected profit for GenCo i in scenario λ_t. When $a_i = 0$, GenCo i accepts the menu of incentive contracts; when $a_i \neq 0$ or $a_i = k$, GenCo i does not accept the incentive contract which is tailored to the target GenCo k ($k = a_i$).
${b_t}^{i,c}$	${b_t}^{i,c} = \{b_{1t}, b_{2t}, b_{i-1,t}, 0, b_{i+1,t}, ..., b_{mt}\}$.
${b_t}^{i,H}$	${b_t}^{i,H} = \{b_{1t}, b_{2t}, b_{i-1,t}, 1, b_{i+1,t}, ..., b_{mt}\}$.
$b_t(a)$	Combination of GenCos' bidding strategy with $b_t(a) = (b_{1,t}(a_1), b_{2,t}(a_2), \cdots, b_{m,t}(a_m))$.
$p(\lambda_t, b_t(a))$	The expected electricity price when the bidding combination of GenCos is $b_t(a)$. The value of $b_t(a)$ could be ${b_t}^{i,c}$ or ${b_t}^{i,H}$.
$D(\lambda_t)$	Electricity demand in scenario λ_t.
c_{i1}, c_{i2}	Cost coefficients of GenCo i.
$(\alpha_i, \beta_i, \pi_i)$	Parameters of an incentive contract. Note α_i and β_i represent the bidding coefficients of a target GenCo, and π_i denotes per-period reward.
p_{it}	Bidding price of GenCo i at time t.
p_t	MCP at time t.
q_{it}	Bidding power output of GenCo i at time t.
D_t	Electricity demand at time t.
${\alpha_i}^{\text{c}}, {\beta_i}^{\text{c}}$	Parameters of the normal bidding curve for GenCo i.
${\alpha_i}^{\text{h}}, {\beta_i}^{\text{h}}$	Parameters of the high bidding curve for GenCo i.

N4. Q-Learning Parameters

$s_{\widetilde{t}}$	State identification for GenCo's Q-learning method in a period.
s_t	State identification for GenCo's Q-learning method in each day.
$a_{i,\widetilde{t}}$	Periodical action selection of GenCo i at the starting point of a period for GenCo's Q-learning method.
$a_{i,t}$	Daily action selection of GenCo i in each day for GenCo's Q-learning method.
$r(s_{i,\widetilde{t}}, a_{i,\widetilde{t}})$	Periodical reward function for Q-learning method.
$R(a_{i,\widetilde{t}})$	Reward obtained by GenCo i over period $\widetilde{t}$.
$r(s_{i,t}, a_{i,t})$	Daily reward function for Q-learning method.
$R(a_{i,t})$	Reward obtained by GenCo i over period at a day.
φ	Discount factor.
T	Number of days in the contract period.
T_s	Number of days elapsed over a contract period $\widetilde{t}$.
$Q_{\widetilde{t}+1}(s_{\widetilde{t}}, a_{i,\widetilde{t}})$	Periodical Q-value function defined for GenCo i.
$Q_{t+1}(s_t, a_{i,t})$	Daily Q-value function defined for GenCo i.
$\ell_{\widetilde{t}}$	Positive learning rate for periodical Q-learning function.
ℓ_t	Positive learning rate for daily Q-learning function.
$\gamma_{\widetilde{t}}$	Discount parameter for periodical Q-learning function.
γ_t	Discount parameter for daily Q-learning function.
$\overrightarrow{s}_{\widetilde{t}} = \{(s_{\widetilde{t}}, a^0)\}$	State identification for ISO's Q-learning function.
$(a^0)_{\widetilde{t}}$	Periodic action selection of ISO.
$r\left(s_{\widetilde{t}}, (a^0)_{\widetilde{t}}\right)$	Periodic reward calculation for ISO's Q-learning function.
${Q^0}_{\widetilde{t}+1}\left(s_{\widetilde{t}}, (a^0)_{\widetilde{t}}\right)$	Periodic Q-value function defined for ISO.

N5. Algorithms Parameters

ν	An iteration number.
η_ν	Non-zero chaos variable.
c_ν, d_ν	Constant vectors.
$\nu_{\max}$	Max iteration times.

References

1. Zhang, D.; Wang, Y.; Luh, P.B. Optimization based bidding strategies in the deregulated market. *IEEE Trans. Power Syst.* **2000**, *15*, 981–986. [CrossRef]
2. Kian, A.R.; Cruz, J.B. Bidding strategies in dynamic electricity markets. *Decis. Support Syst.* **2005**, *40*, 543–551. [CrossRef]
3. Swider, D.J.; Weber, C. Bidding under price uncertainty in multi-unit pay-as-bid procurement auctions for power systems reserve. *Eur. J. Oper. Res.* **2007**, *181*, 1297–1308. [CrossRef]
4. Centeno, E.; Renese, J.; Barquin, J. Strategic analysis of electricity markets under uncertainty: A conjectured-price-response approach. *IEEE Trans. Power Syst.* **2007**, *22*, 423–432. [CrossRef]
5. Sahraei-Ardakani, M.; Rahimi-Kian, A. A dynamic replicator model of the players' bid in an oligopolistic electricity market. *Electr. Power Syst. Res.* **2009**, *79*, 781–788. [CrossRef]
6. Li, G.; Shi, J. Agent-based modeling for trading wind power with uncertainty in the day-ahead wholesale electricity markets of single-sided auctions. *Appl. Energy* **2012**, *99*, 13–22. [CrossRef]
7. Nojavan, S.; Zare, K. Risk-based optimal bidding strategy of generation company in day-head electricity market using information gap decision theory. *Int. J. Electr. Power Energy Syst.* **2013**, *48*, 83–92. [CrossRef]
8. Qiu, Z.; Gui, N.; Deconick, G. Analysis of equilibrium-oriented bidding strategies with inaccurate electricity market models. *Int. J. Electr. Power Energy Syst.* **2013**, *46*, 306–314. [CrossRef]
9. Kardakos, E.G.; Simoglou, C.K.; Bakirtzis, A.G. Optimal bidding strategy in transmission-constrained electricity markets. *Electr. Power Syst. Res.* **2014**, *109*, 141–149. [CrossRef]
10. Anderson, E.J.; Cau, T.D.H. Implicit collusion and individual market power in electricity markets. *Eur. J. Oper. Res.* **2011**, *211*, 403–414. [CrossRef]
11. Nam, Y.W.; Yoon, Y.T.; Hur, D.; Park, J.; Kim, S. Effects of long-term contracts on firms exercising market power in transmission constrained electricity markets. *Electr. Power Syst. Res.* **2006**, *76*, 435–444. [CrossRef]
12. David, A.K.; Wem, F.S. Market power in electricity supply. *IEEE Trans. Energy Convers.* **2001**, *16*, 352–360. [CrossRef]
13. Oh, S.; Hildreth, A.J. Decisions on energy demand response option contracts in smart grids based on activity-based costing and stochastic programming. *Energies* **2013**, *6*, 425–443. [CrossRef]
14. Faria, P.; Vale, Z.; Baptista, J. Demand response programs design and use considering intensive penetration of distributed generation. *Energies* **2015**, *9*, 6230–6246. [CrossRef]
15. Ghazvini, M.A.F.; Soares, J.; Horta, N.; Neves, R.; Castro, R.; Vale, Z. A multi-objective model for scheduling of short-term incentive-based demand response programs offered by electricity retailers. *Appl. Energy* **2015**, *151*, 102–118. [CrossRef]
16. Ghazvini, M.A.F.; Faria, P.; Ramos, S.; Morais, H.; Vale, Z. Incentive-based demand response programs designed by asset-light electricity providers for the day-ahead market. *Energy* **2015**, *82*, 786–799. [CrossRef]
17. Zhong, H.; Xie, L.; Xia, Q. Coupon incentive-based demand response: Theory and case study. *IEEE Trans. Power Syst.* **2013**, *28*, 1266–1276. [CrossRef]
18. Fakhrazari, A.; Vakilzadian, H.; Choobineh, F.F. Optimal energy scheduling for a smart entity. *IEEE Trans. Smart Grid* **2014**, *5*, 2919–2928. [CrossRef]
19. Christopher, O.A.; Wang, L. Smart charging and appliance scheduling approaches to demand side management. *Int. J. Electr. Power Energy Syst.* **2014**, *57*, 232–240.
20. Yousefi, S.; Moghaddam, M.P.; Majd, V.J. Optimal real time pricing in an agent-based retail market using a comprehensive demand response model. *Energy* **2011**, *36*, 5716–5727. [CrossRef]
21. Shariatazadeh, F.; Mandal, P.; Srivastava, A.K. Demand response for sustainable energy systems: A review, application and implementation strategy. *Renew. Sustain. Energy Rev.* **2015**, *45*, 343–350. [CrossRef]

22. Gu, W.; Yu, H.; Liu, W.; Zhu, J.; Xu, X. Demand response and economic dispatch of power systems considering large-scale plug-in hybrid electric vehicles/electric vehicles (PHEVs/EVs): A review. *Energies* **2013**, *6*, 4394–4417. [CrossRef]
23. Bradley, P.; Leach, M.; Torriti, J. A review of the costs and benefits of demand response for electricity in the UK. *Energy Policy* **2013**, *52*, 312–327. [CrossRef]
24. Silva, C.; Wollenberg, B.F.; Zheng, C.Z. Application of mechanism design to electric power markets. *IEEE Trans. Power Syst.* **2001**, *16*, 1–8. [CrossRef]
25. Liu, Z.; Zhang, X.; Lieu, J.; Li, X.; He, J. Research on incentive bidding mechanism to coordinate the electric power and emission-reduction of the generator. *Int. J. Electr. Power Energy Syst.* **2010**, *32*, 946–955. [CrossRef]
26. Cai, X.; Li, C.; Lu, Y. Price cap mechanism for electricity market based on constraints of incentive compatibility and balance accounts. *Power Syst. Technol.* **2011**, *35*, 143–148.
27. Heine, K. Inside the black box: Incentive regulation and incentive channeling on energy markets. *J. Manag. Gov.* **2013**, *17*, 157–186. [CrossRef]
28. Weber, J.D.; Overbye, T.J. A two-level optimization problem for analysis of market bidding strategies. In Proceedings of the IEEE Power Engineering Society Summer Meeting, Edmonton, AB, Canada, 18–22 July 1999; Volume 2, pp. 682–687.
29. Lei, W.; Shahidehpour, M.; Zuyi, L. Comparison of scenario-based and interval optimization approaches to stochastic SCUC. *IEEE Trans. Power Syst.* **2012**, *27*, 913–921.
30. Wang, B.; Yang, X.; Li, Q. Bad-scenario set risk-resisting robust scheduling model. *Acta Autom. Sin.* **2012**, *38*, 270–278. [CrossRef]
31. North, M.J.; Collier, N.T.; Vos, J.R. Experiences creating three implementations of the repast agent modeling toolkit. *ACM Trans. Model. Comput. Simul.* **2006**, *16*, 1–25. [CrossRef]
32. Rahimiyan, M.; Mashhadi, H.R. An adaptive Q-learning algorithm developed for agent-based computational modeling of electricity market. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2010**, *40*, 547–556. [CrossRef]
33. Naghibi-Sistani, M.B.; Akbarzadeh-Tootoonchi, M.R.; Bayaz, M.H.J.D.; Rajabi-Mashhadi, H. Application of Q-learning with temperature variation for bidding strategies in market based power systems. *Energy Convers. Manag.* **2006**, *47*, 1529–1538. [CrossRef]
34. Haddad, M.; Altmann, Z.; Elayoubi, S.E.; Altaman, E. A Nash-Stackelberg fuzzy Q-learning decision approach in heterogeneous cognitive networks. In Proceedings of the IEEE Global Telecommunications Conference, Miami, FL, USA, 6–10 December 2010.
35. Zuo, X.Q.; Fan, Y.S. A chaos search immune algorithm with its application to neuro-fuzzy controller design. *Chaos Solitons Fractals* **2006**, *30*, 94–109. [CrossRef]