Contents lists available at ScienceDirect

Energy Policy

journal homepage: www.elsevier.com

An Agent-based Simulation of Power Generation Company Behavior in Electricity Markets under Different Market-Clearing Mechanisms

Danial Esmaeili Aliabadi[a], Murat Kaya[a], Güvenç Şahin[a,*]

[a]*Sabanci University, Faculty of Engineering and Natural Sciences, Istanbul, Turkey.*

Abstract

Deregulated electricity markets are expected to provide affordable electricity for consumers through promoting competition. Yet, the results do not always fulfill the expectations. The regulator's market-clearing mechanism is a strategic choice that may affect the level of competition in the market. We conceive of the market-clearing mechanism as composed of two components: pricing rules and rationing policies. We investigate the strategic behavior of power generation companies under different market-clearing mechanisms using an agent-based simulation model which integrates a game-theoretical understanding of the auction mechanism in the electricity market and generation companies' learning mechanism. Results of our simulation experiments are presented using various case studies representing different market settings. The market in simulations is observed to converge to a Nash equilibrium of the stage game or to a similar state under most parameter combinations. Compared to pay-as-bid pricing, bid prices are closer to marginal costs on average under uniform pricing while GenCos' total profit is also higher. The random rationing policy of the ISO turns out to be more successful in achieving lower bid prices and lower GenCo profits. In minimizing GenCos' total profit, a combination of pay-as-bid pricing rule and random rationing policy is observed to be the most promising.

Keywords: Agent-based simulation, Reinforcement learning, Uniform pricing, Pay-as-bid pricing, DC-OPF, Game-theory

1. Introduction

Deregulated electricity markets procure most of the electricity through several trading floors some of which are designed as auction-based markets. In most regional/national markets, these trading floors are controlled and governed by an Independent System Operator (ISO).

We focus on the day-ahead market in which Power Generation Companies (GenCos) compete for the next day supply of an inelastic load demand. In the day-ahead market's auction, for each hour of the following day, each GenCo bids the minimum acceptable unit price of electricity for itself. Based on the predetermined market-clearing mechanism, the ISO determines the market-clearing price and each GenCo's assigned power. We address two questions regarding the ISO's market-clearance mechanism that are at the heart of policy discussions on deregulated electricity markets: Which mechanisms lead to (1) more competitive price bidding by GenCos, (2) lower GenCo profits, meaning higher customer benefit.

The market clearing mechanism that we consider includes a pricing rule and a rationing policy. We compare the two most common pricing rules in the literature: Uniform and pay-as-bid (or, discriminatory) pricing (see Cramton [6]). With uniform pricing, all GenCos with winning bids are paid the market-clearing price, whereas with pay-as-bid pricing, each GenCo is paid at its own bid price. In addition to uniform and pay-as-bid pricing, we also provide results under a DC-OPF rule under which, each region in the transmission grid may have a different electricity

*Corresponding Author

Email addresses: Danialesm@sabanciuniv.edu (Danial Esmaeili Aliabadi), Mkaya@sabanciuniv.edu (Murat Kaya), Guvencs@sabanciuniv.edu (Güvenç Şahin)

price due to physical constraints of the transmission lines.

By "rationing policy", we refer to the way remaining demand at the market cleaning price is auctioned when multiple GenCos' bids coincide at that price. The rationing decision is part of real electricity market exchange mechanics (See, for example Madlener and Kaufmann [17]), however it has not been addressed in electricity markets literature before.

Learning is an important aspect of electricity markets as GenCos engage in auctions repeatedly for every hour, and thus obtain experience that can change their bidding behavior. To capture this dynamic, we develop an agent-based simulation model where GenCos can learn from their own experience based on a variant of the well-known Q-learning algorithm. Using this model, we simulate the repetitive auction process under different market clearing mechanisms in a number of case studies. We compare the results of our simulations with the Nash equilibrium predictions of static game-theoretic models.

This work contributes to both managerial and academic literature in a number of ways. Our results can guide ISOs and GenCo managers regarding the merits of different market clearance mechanisms. Comparisons between uniform and pay-as-bid pricing is definitely not new in the literature. However, we extend this comparison with the rationing policy dimension, and we provide results that incorporate the interplay between learning, dynamic competition, and ISO's clearance mechanism. In particular, we show that GenCo learning can take the market to a different direction than predicted by standard game-theoretical models. Finally, unlike most papers in literature, we present results for a wide range of learning model parameters.

The remainder of this article is organized as follows. In Section 2, we review relevant literature from three different perspectives. In Section 3, we present the market-clearing mechanism of the electricity market, explaining the pricing rules and rationing policies. The learning procedure and the simulation model are discussed in Section 4. Our game-theoretical understanding of the auction mechanism in electricity markets and the significance of Nash equilibrium are addressed in Section 5. Section 6 presents the results of simulation experiments and our findings.

2. Related Work

We present the related literature in three parts: learning and game-theory, applications of agent-based simulation, and analysis of pricing rules.

2.1. Learning and Game-theory

Due to repetitive nature of auctions in the electricity markets, GenCos are expected to learn by gathering new information in each repetition of the auction and improve their performance over time. In this respect, analyzing GenCos' behavior without a learning mechanism would lead to inaccurate results. Even in the early years of game-theory, researchers have been interested in learning models. Studying convergence to Nash equilibrium in the presence of learning has attracted a lot of attention from game-theory modelers as well as energy-economics community.

Aumann [2] claims that Nash equilibrium concept is one of the most applied concepts in economics; yet, it is not crystal clear under what condition players might be expected to play a Nash equilibrium. Mailath [18] discusses various justifications that have been advanced for equilibrium analysis and points out learning as the least problematic justification. Also, Mailath notes that convergence to Nash equilibria is a necessary condition in the evolutionary dynamics for any reasonable model of social learning when the number of players is large enough. Kalai and Lehrer [13] show that under some simplifying assumptions, rational learning leads to Nash equilibrium.

Hart and Mas-Colell [10] propose "reinforcement" models in which all players can be led to an equilibrium of the stage game. Their learning procedure, unlike the "regret-matching" procedure [9], does not need to observe all past payoffs, and players do not need to know their own payoff function. Wang and Sandholm [27] state that even agents with non-conflicting interests may not be able to learn an optimal coordination policy in the presence of multiple Nash equilibria. As a solution, these authors propose a new learning mechanism based on reinforcement learning that converges to an optimal Nash equilibrium with probability one in any team Markov game.

2.2. Agent-based Simulation of Electricity Markets

Although analytical models can be employed to study learning mechanisms, the expected outcomes of these models are not necessarily observed in practice due to strict simplifying assumptions [7]. A widely accepted alternative tool is Agent-based Modeling and Simulation; it can provide better understanding of real-life markets especially when analytical models show poor tractability in investigating complicated problems. Li and Shi [16] claim that agent-based modeling and simulation is a viable approach which provides realistic insights for the complex interactions among various market players.

Existence of multiple Nash equilibria can disrupt GenCos' learning process in such a way that the long-run equilibrium is not necessarily achieved. Krause et al. [15] study a day-ahead market where GenCos learn by reinforcement learning (Q-learning). These authors' simulation does not converge in the existence of multiple Nash equilibria. The GenCos' strategies pendulate between those Nash equilibria. The oscillation between different Nash equilibria in the reinforcement learning process can be overcome by making better use of collected information. To this end, Wang [26] used the SA-Q-learning algorithm with Metropolis criterion.

Naghibi-Sistani et al. [20] apply Q-learning for agents' bidding in a pool-based power market with uniform pric-

ing. They show that a participant with reinforcement learning capability could ultimately learn the optimal policy and could adapt himself to unknown parameters in the environment. The authors also find that under reinforcement learning, bids can converge and stay in the Nash equilibrium for a two-participant case. Nevertheless, these authors have not studied other pricing rules than uniform pricing and their impact on convergence.

We propose a modified version of the standard Q-learning algorithm. Different from the standard algorithm, ours is a state-independent one where Q-values are expressed as functions of actions only (Similar to Krause et al. [15] and Krause and Andersson [14]). In addition, it is similar to the Simulated Annealing (SA) Q-learning method (Guo et al. [8]) in that both methods employ a time-decaying exploration parameter. We use a linear decay, whereas the SA Q-Learning method uses a geometric function. The time decaying exploration parameter reflects the increasing experience of agents in the decision making process, and helps the algorithm achieve convergence. Table 1 presents main features of popular learning algorithms in order to facilitate a comparison with our method.

2.3. Pricing Rules and Rationing Policy

Selecting a pricing rule is a vital decision for the ISO as it is likely to affect GenCos' strategic bidding behavior. Researchers have been investigating the characteristics of pricing rules to improve the functionality of underlying markets.

Kahn et al. [12] argue that the proposed shift from uniform to pay-as-bid pricing in California Power Exchange was a mistake and contrary to expectations, it will not reduce electricity prices. Under uniform pricing, GenCos have an incentive to bid their true marginal generation cost [21] which will contribute to efficiency in power dispatch. Under pay-as-bid pricing, on the other hand, GenCos will bid at their expectation of the market clearing price. For that reason, bid prices are expected to be higher under pay-as-bid. However, this does not necessarily result in a higher market price for electricity under pay-as-bid pricing. This is because under uniform pricing, all GenCos are paid at the market clearing price, whereas under pay-as-bid, they are paid at their own bids which are generally lower than the market clearing price. Variation in bid prices and consequently the short-run volatility in market prices is expected to be lower under pay-as-bid than under uniform pricing (see, for example, Tierney et al. [24] and Mount [19]). That is, pay-as-bid pricing will result in a flatter supply function. Power dispatch efficiency may be adversely affected under pay-as-bid because low-cost GenCos that overestimate the clearing price will not be dispatched. In addition, GenCos will need to spend resources on forecasting activities, which would provide larger GenCos with an advantage over the smaller ones [28]. Overall, the discussion about the pros and cons of these two pricing policies has not yet reached a conclusion [4].

Xiong et al. [29] compare uniform and pay-as-bid pricing rules using agent-based simulation and show that pay-as-bid results in lower market prices and price volatility. They also claim that demand side response has less effect on market prices with pay-as-bid rule. Bakirtzis and Tellidou [4] show that high price levels are due to exercised market power with both uniform and pay-as-bid policies. Azadeh et al. [3] study three different pricing rules (uniform, pay-as-bid, and Vickrey) by using Principal Component Analysis (PCA). They conclude pay-as-bid pricing rule with one permissible step to be the best pricing rule. Sugianto and Liao [23] use agent-based modeling approach to investigate the impact of different auction pricing rules on the market performance. They conclude that the pay-as-bid pricing rule can complicate the way bidders learn and react to each other's strategy. Also, their results suggest that Vickrey pricing provides a balance between managing the total cost and its stability in the presence of unequal GenCo market shares.

In addition to the pricing rule, we also discuss another aspect of the market-clearing mechanism, the "rationing policy". The rationing policy determines the allocation of the remaining demand at the market cleaning price when multiple GenCos' (the marginal GenCos) bids coincide at that price. This possibility arises due to the discrete nature of bid price and quantity. Rationing rule is especially important when the bid prices are likely to accumulate at certain values. Holmberg [11] and the references therein discuss rationing rules to break ties between multiple bids at the market clearing price in general multi-unit auctions. In auctions where all bids are cleared simultaneously, standard practice is pro-rata rationing where the same percentage of bid is accepted for each marginal bidder. In continuous trading, priority can be given to marginal bids that arrive early. Madlener and Kaufmann [17] describe the rationing rules employed in European power exchanges. For instance, in case of a supply surplus at the market clearing price, APX and OMEL exchanges distribute the demanded quantity in proportion to the bid quantities. Borzen, EEX and EXAA exchanges, on the other hand, prioritize according to the size of bid or time of submission. Different from these, we propose a rationing policy where the priority ordering of marginal GenCos is randomly determined (random rationing), and another policy where the remaining demand is equally distributed to marginal GenCos (equal rationing). We are not aware of any other work that models rationing policies in electricity markets.

In order to study the strategic behavior of GenCos under different market-clearing mechanisms, we employ an agent-based simulation model in which GenCos learn from their own previous actions by using a modified version of Q-learning algorithm. We then compare the results of our simulation with the Nash equilibria of the relevant stage games under different pricing rules and rationing policies. Finally, we investigate the effects of the market clearing mechanism on GenCos' competitive bidding behavior and their profits.

Table 1: Main features of learning algorithms in comparison with our method

Learning Algorithm	Exploration Parameter	Pros	Cons
Basic Q-Learning	Constant	• Simple to implement • Shorter CPU time in each iteration • Estimates the value of the new state	• May not converge
SA-Q-Learning	Changing with number of times an action is use	• Better convergence • Estimates the value of the new state	• Parameters should be tuned • Harder to implement
Erev and Roth	Action selection probabilities change based on fitness value	• Based on psychological findings about human learning	• Slow due to updating Q-values for each action at each iteration.
Our method	Changing with time	• Better convergence • Simple implementation • Considers experience • Fast due to updating only the selected action Q-value.	• Parameters should be tuned

3. Market-Clearing Mechanism

In the day-ahead market, the ISO clears the bids sequentially for each hour of the next day through an auction mechanism. For any hour, each GenCo (GenCo-i) participates in the auction with its maximum capacity (P_i^{max}, MW) and submits a bid price from a discrete set of available prices ($b_{ij} \in B_i$) to the ISO. The bid price alternatives b_{ij} (\$/MWh) should be above the marginal generation cost (C_i) and below market price cap (b_i^{cap}) which is determined by the ISO [5]. Based on the predetermined market clearing mechanism, the ISO determines the market price λ (\$/MWh) and the power to be dispatched by each GenCo ($P_i \leq P_i^{max}$) (MW). We now discuss different market-clearing mechanisms with respect to pricing rules and rationing policies.

The combination of price and generation quantity submitted to the ISO by a GenCo is referred to as the energy block of that GenCo. The ISO sorts received blocks in an increasing order of their prices ($b_{(i)} \leq b_{(i+1)}$), and accepts generation bid prices starting from the least expensive block ($b_{(1)} \times P_{(1)}^{max}$) until demand is completely satisfied [22]. This procedure is known as the merit order.

Under *uniform* pricing, the ISO determines the market clearing price λ as that of the last accepted energy block. Winning GenCos, whose bid prices were less than or equal to the market price are paid at λ. Since they are gaining no less than what they have asked in their bids, these GenCos accept the market price.

Figure 1 depicts energy blocks accepted under uniform pricing by the merit order procedure. The shaded energy blocks are those of the winning GenCos. The partially shaded block determines the market price while only part of the capacity of the corresponding GenCo is accepted by the ISO.

In this example, the market clearing price is determined as $b_{(3)}$ because $D \geq P_{(1)}^{max} + P_{(2)}^{max}$ and $D \leq P_{(1)}^{max} + P_{(2)}^{max} + P_{(3)}^{max}$. Thus, all winning GenCos are paid $\lambda = b_{(3)}$. As a result, payoff for GenCo-1 is $r_1 = P_1^{max}(\lambda - C_1)$, for GenCo-2 $r_2 = P_2^{max}(\lambda - C_2)$, and for GenCo-3 $r_3 = (D - P_1^{max} - P_2^{max})(\lambda - C_3)$.

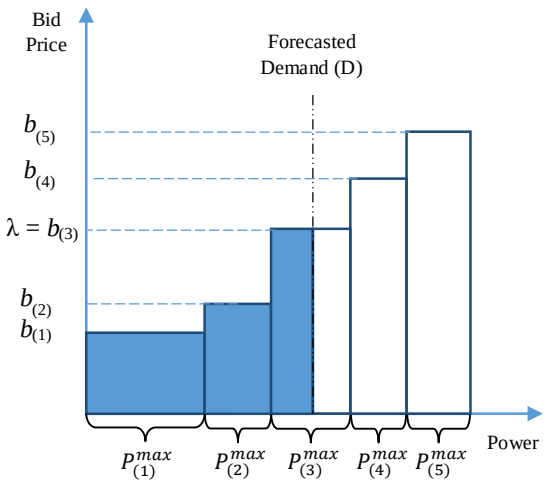

Figure 1: Supply curve from GenCos' bids

Different from *uniform* pricing, in *pay-as-bid* pricing rule, each GenCo is paid exactly at its own bid, but not more. According to the example in Figure 1, the GenCo with the least expensive block will receive a price of $b_{(1)j}$ instead of λ. The next block will receive $b_{(2)j}$ for $P_{(2)}^{max}$. Note that the last accepted block will receive λ for the accepted capacity under both uniform and pay-as-bid pricing rules.

Under both uniform and pay-as-bid pricing, an issue arises when the market clearing price λ coincides with multiple GenCo's bid prices: How to allocate the remaining demand? As Figure 2 illustrates with two such marginal GenCos (G_1 and G_2), two "rationing policies" (as we refer to it) can be used:

1. *Random rationing*: Marginal GenCos are put into a randomly determined priority order. If the GenCo

with the first priority cannot satisfy all remaining demand, the unmet demand is passed to the next marginal GenCo in priority order. This is illustrated in the left plot of Figure 2, where GenCo-2 (G_2) was chosen to be the first priority.

2. *Equal rationing*: Remaining demand is shared equally between the marginal GenCos. This is illustrated in the right plot of Figure 2. If a GenCo does not have sufficient capacity to meet his allocated demand, the unmet demand is distributed equally to the other marginal GenCo(s).

In addition to uniform and pay-as-bid pricing, we also study another pricing mechanism that takes the transmission network structure into consideration. Even though uniform pricing is the most common method to set prices in electricity markets, it may lead to infeasible solutions due to network constraints [25]. Therefore in a constrained network, a viable pricing method should also provide some economic signal to reflect the charge due to the physical constraints. This is what is done in the "Locational Marginal Pricing" (LMP) approach. In this approach, the ISO handles physical constraints by considering congestion cost in calculating price of electricity at different locations. An LMP at any node corresponds to the minimum cost of fulfilling the demand for one additional unit (MW) of power at that particular location. It includes marginal generation cost, transmission congestion cost, and cost of marginal losses. Transmission grid congestion is managed by the inclusion of congestion components in LMPs.

In order to implement locational marginal pricing, an ISO may employ two methods: AC-OPF and DC-OPF. In practice, AC-OPF problems are typically approximated by more tractable DC-OPF problems that focus exclusively on real power constraints in the linearized form. In this paper, we also use a DC-OPF approach as one of the market-clearing mechanism alternatives. We choose DC-OPF over AC-OPF because our long simulation study would simply be infeasible with the more complicated AC-OPF formulation, and because DC-OPF is the preferred alternative in the literature. We assume that the ISO solves a DC-OPF problem to clear the market for any hour, with the objective of maximizing social welfare (by minimizing the total cost of demanded electricity). By doing so, the ISO determines the dispatch (P_i) of each GenCo-i and the price of electricity, LMP_i, at each node i. The payoff of GenCo-i is then calculated as

$$r_i = P_i(LMP_i - C_i). \tag{1}$$

Note that the rationing policy is not relevant under a DC-OPF mechanism.

4. Simulation Process

In our agent-based simulation model, agents represent the GenCos that are expected to satisfy demand on the transmission grid. GenCos submit bids sequentially for each hour of the next 24 hours to the ISO. The bidding process is synchronic for all GenCos, and each iteration in the simulation corresponds to an hour in the day-ahead-market. The simulation runs for a finite number of iterations (max_t). At the end of each iteration/bid, each GenCo-i calculates its payoff r_i.

In the simulation model, we assume that

- The demand is inelastic and constant, i.e., it does not change from one hour to the next.
- GenCos participate only in the day-ahead market (no futures or real-time markets).
- No line or generation outage is experienced.
- GenCos do not change their technology (no change in P_i^{max} or C_i).
- Capacity withholding is not allowed. That is, each GenCo bids its maximum generation capacity.
- GenCos do not share information with each other. They are not aware of others' generation costs, available bid prices and submitted bids.

In essence, the bidding process of GenCos is a decision-making problem with incomplete information as each GenCo is only aware of its own cost and bids. Each GenCo determines what price to bid through a Q-learning mechanism (to be explained) based on historical payoff information from its own bids in previous iterations. Thus, the profit at an iteration affects the GenCo's subsequent bid decisions.

We model GenCos' learning mechanism by reinforcement learning. In particular, we improve the standard Q-learning mechanism by making the two following parameters time-dependent:

- Recency rate ($\alpha_{it} \in [0, 1]$) determines the weight given by GenCo-i to the most recent observed outcome (profit).
- Exploration parameter ($\epsilon_{it} \in [0, 1]$) measures the tendency of GenCo-i at iteration t to explore, i.e., to use a randomly selected bid rather than using its best identified bid.

Recall that GenCo-i has a set of bid prices ($b_{ij} \in B_i$) to choose from. For each bid price, the Q-value in the learning algorithm denoted by Q_{ij} corresponds to the average realized profit of GenCo-i when b_{ij} was used in the previous iterations. Initially, all Q-values are zero. At the end of each iteration t, based on the observed payoff r_i, the Q-value of the submitted bid price is updated as follows

$$Q_{ij} = (1 - \alpha_{it})Q_{ij} + \alpha_{it}(r_i). \tag{2}$$

A high α value represents a GenCo that is primarily concerned about the most recent outcomes it experienced,

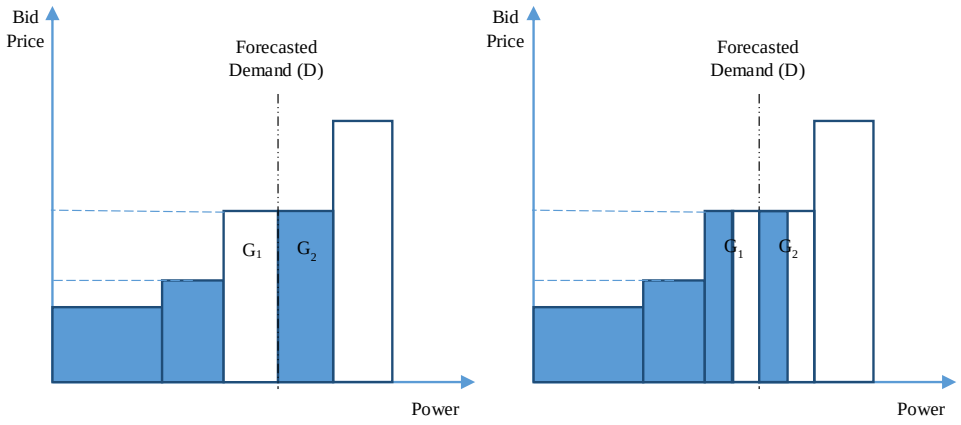

Figure 2: Random rationing (left) versus Equal rationing (right)

and is less affected by the distant ones. In our modified Q-learning algorithm, α_{it} starts from a high value (α_{i0}) at the beginning and diminishes linearly over iterations to a lower value of ($\frac{\alpha_{i0}}{10}$). To this end, we use a linear decreasing function of time for α_{it} as

$$\alpha_{it} = \left(1 - \frac{t}{max_t}\right)(\alpha_{i0}) + \left(\frac{t}{max_t}\right)\left(\frac{\alpha_{i0}}{10}\right). \quad (3)$$

We use a descending recency rate because the GenCo is assumed to become less sensitive to individual recent observations over time due to gained experience.

We refer to the bid price b_i^* that maximizes the Q_{ij} value as GenCo-i's "best identified bid price". In the proposed Q-learning algorithm, at iteration t, GenCo-i either selects its best identified bid price (with prob $1 - \epsilon_{it}$), or it explores by choosing a random bid price from B_i (with with probability ϵ_{it}). Therefore, with lower ϵ, the GenCo explores less and sticks to its best identified bid price more often. At the beginning of the simulation, b_i^* is determined randomly from B_i since Q-values of all bid prices are zero. We assume that GenCos explore more in initial iterations using random bids, but they are more likely to use their best identified bid price in latter iterations. To represent such behavior, the exploration parameter (ϵ_{it}) decreases linearly from the base value of ϵ_{i0} (when $t = 0$) to almost zero at iteration number $\lceil \frac{max_t \epsilon_{i0}}{8(1-\epsilon_{i0})} \rceil$. Exploration is minimum and exploitation is maximum afterwards.

Our Q-learning algorithm is presented in Algorithm 1. In line 2, Q-values are set to zero for initialization. In lines (7 - 11), the GenCo determines its bid price to the ISO. This decision is governed by the Q-learning parameters. Following the market clearance by the ISO in line 12, GenCo-i will update the Q-value of the selected bid price in line 13.

The increasing exploitation of the best identified bid price cancels out the effect of GenCos' random choices at early iterations. Therefore, the Q-value of the best identified bid price b_i^* for each GenCo-i converges to the profit (r_i) of GenCo-i in the equilibrium state in the long-run.

Algorithm 1 The simulation model with the proposed Q-learning algorithm for each GenCo-i.

1: $t \leftarrow 1$
2: $Q_{ij} \leftarrow 0 \quad \forall j$
3: **repeat**
4: R $\leftarrow$ Random Number $\in [0, 1]$
5: $\epsilon_{it} \leftarrow max\{0.001, 1 - (1 - \epsilon_{i0})\left(1 + \frac{8t}{max_t}\right)\}$
6: $\alpha_{it} = (1 - \frac{t}{max_t})(\alpha_{i0}) + (\alpha_{i0}/10)(\frac{t}{max_t})$
7: **if** $R > \epsilon_{it}$ **then**
8: $b_{ij} \leftarrow$ Select Best bid price (b_i^*)
9: **else**
10: $b_{ij} \leftarrow$ Select a bid price randomly $(b_{ij} \in B_i)$
11: **end if**
12: $r_i \leftarrow$ ClearMarket$(\{b_{ij} : \forall i\})$
13: $Q_{ij} \leftarrow (1 - \alpha_{it})Q_{ij} + \alpha_{it}(r_i)$
14: $t \leftarrow t + 1$
15: **until** $(t < max_t)$

5. Nash Equilibrium

We assume that GenCos start bidding with no information about the potential profit of each bid price in their set of bid prices, and the possible bid prices of other GenCos. Throughout the iterations, each GenCo experiences the outcome of its own bids, but not those of competitors. In our simulation analysis, we would like to understand if the market reaches a Nash equilibrium of the stage game. In this respect, each iteration of the simulation corresponds to a stage of the multi-stage interaction between the competing GenCos.

The market clearance process for an hour in the day-ahead market can be modeled as a non-cooperative single-stage game G with finite number of players, $\mathcal{F} = \{$GenCo-1, ..., GenCo-$n\}$, an action space of $\mathfrak{B} = (B_1 \times \ldots \times B_n)$ and a vector of payoffs $r = (r_1, \ldots, r_n)$. Thus, the normal-form representation of G is denoted by the triplet $(\mathcal{F}, \mathfrak{B}, r)$. In each iteration, collection of submitted bids $(b_1 \in B_1, \ldots, b_n \in B_n)$ defines the "state" of the game.

When the random rationing policy is used (under both uniform and pay-as-bid pricing rules), the same set of bid prices from GenCos can lead to different power dispatches, resulting in different profit vectors. This is because of the ISO's random prioritization among the GenCos that submit the same bid at the market-clearing price. Therefore, in the random rationing policy, we calculate the average payoff of each GenCo with respect to the probability of each realized profit for a given state. The vector of average payoffs will be used to identify Nash equilibria.

In our context, a bidding strategy $N = (b_1^N, \ldots, b_n^N)$ is called a Nash equilibrium if any GenCo-i cannot make a better payoff than the payoff of the Nash equilibrium (r_i^N) by choosing another bid price $(b_{ij} \in B_i)$ as long as the other GenCos are not changing their bid prices, i.e.

$$(r_1^N, \ldots, r_i^N, \ldots, r_n^N) \geq (r_1^N, \ldots, r_i, \ldots, r_n^N), \quad i \in \{1, \ldots, n\}. \tag{4}$$

We also make the following new definition: A state $S = (b_1^S, \ldots, b_n^S)$ is a *semi-Nash* state if

$$(r_1^S, \ldots, r_n^S) = (r_1^N, \ldots, r_n^N) \quad \exists j \text{ such that } b_j^S \neq b_j^N.$$

A semi-Nash state is defined with respect to a particular Nash equilibrium. If one of the GenCos (say, GenCo-i) in a Nash equilibrium can change its bid price without affecting the payoff of any GenCo including itself, we refer to the resulting state as a semi-Nash (as long as the resulting state is not a Nash equilibrium itself). Because a semi-Nash state is not a Nash equilibrium, at least one of the GenCos (other than GenCo-i) can increase its payoff by deviating from this state. During our experimental simulations, however, such a GenCo may or may not realize this profitable deviation opportunity. Hence, the simulation may end up converging to a semi-Nash state just like it may converge into a Nash equilibrium. Indeed, this is what we observed in our experiments as we will present in the following section.

6. Computational Experiments

We conduct simulation experiments on four case studies. In each case study, GenCos are subject to challenges due to a variety of environmental settings. The underlying characteristics of the case studies can be summarized as follows:

- Case 1: A public GenCo and a private GenCo compete in a limited competition market where the public GenCo always bids its generation cost.
- Case 2: Two private GenCos and a public GenCo participate in a competitive market where only the private GenCos are learning agents.
- Case 3: The public GenCo in Case 2 is replaced with a learning GenCo.
- Case 4: Three learning GenCos compete to satisfy demand of a single node. As the demand can be satisfied to a great extent by any one of three GenCos, competition between GenCos is tight.

The details of the case studies are presented in Appendix A. Case 2 and Case 3 are adopted from [15, 14] with slight modifications. The network structure of these cases are simplified versions of the real Pennsylvania-New-Jersey-Maryland (PJM) five node power system. Table 2 reports the number of Nash equilibria and semi-Nash states under each pricing rule.

Table 2: Number of Nash equilibria and semi-Nash states in case studies

Case Study	Learning GenCos	Uniform Pricing		Pay-as-bid Pricing		DC-OPF Pricing	
		Nash	semi-Nash	Nash	semi-Nash	Nash	semi-Nash
Case 1	1	1	0	1	0	1	0
Case 2	2	3	1	3	1	1	0
Case 3	3	6	4	3	2	2	3
Case 4	3	3	10	3	7	3	15

In what follows, we first use Case 1 to illustrate how our Q-learning algorithm operates. Then, with other cases, we study whether our simulations converge to theoretical Nash equilibria and/or semi-Nash states. Then, we study the competitive bidding behavior and realized profits of GenCos under different pricing rule and rationing policy combinations.

6.1. Illustration of the Learning Algorithm

To describe the behavior of the only learning agent (the private GenCo) in Case 1, we limit the public GenCo to bid one price; therefore, the described Q-learning algorithm does not apply to this GenCo.

Figure 3 shows the evolution of expected profits (Q-values) for the private GenCo for each bid price option throughout 300 iterations. The results clearly indicate one outcome: The private GenCo gradually learns to bid higher prices to the ISO as it discovers along iterations that the public company cannot fulfill the demand. Eventually, the private GenCo reaches the maximum Q-value of 2700 by bidding 40. In fact, this bid, along with the fixed bid of the public GenCo correspond to the unique Nash equilibrium of the stage game. The bid also happens to be the optimal one for the private GenCo. We observed this result to hold true in Case 1 independent of the pricing rule and rationing policy.

We examine the results with respect to three aspects: convergence to Nash equilibria, GenCos' competitive bidding behavior and GenCos' profits. All results are based on the data of the simulation study. For each case study, we run simulations for $51 \times 51 = 2601$ different combinations of the initial values of the two learning model parameters (α_{i0} and ϵ_{i0}). For each parameter, we use all values between 0 and 1.00 with an increment of 0.02. All learning GenCos are subject to the same learning algorithm and assumed to possess the same learning parameters.

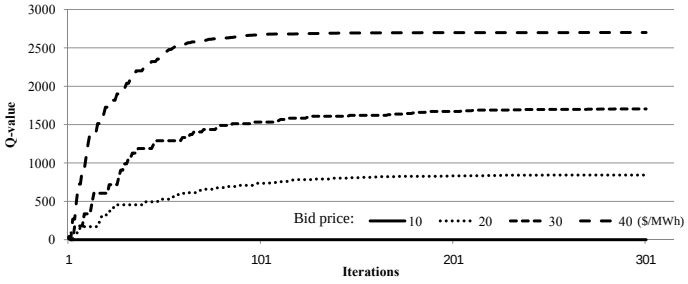

Figure 3: Evolution of Q-values of the private GenCo over iterations

To minimize the effects of random factors, we conduct 30 simulation replications for each parameter setting and report the average over these 30 replications. Each simulation replication consists of ($max_t = 2000$) iterations. That corresponds to solving the ISO's market clearing optimization problem and running the Q-learning algorithm $2601 \times 30 \times 2000 = 156,060,000$ times for each case study under each market-clearing mechanism. Due to such detailed analysis, the reported simulation study took longer than 700 hours on a powerful computer (Intel Core i7 @ 3.2GHz with 24GB RAM).

6.2. Convergence to Nash Equilibria and Semi-Nash States

We investigate whether the simulation converges to a Nash equilibrium, a semi-Nash state or neither at its termination. For a particular parameter setting (α_{i0} and ϵ_{i0}), we define "convergence frequency to Nash equilibrium" as the proportion of replications (over 30 total replications) in which the simulation converged to a Nash equilibrium. That is, if the simulation converged to a Nash equilibrium in 18 out of 30 replications, this frequency becomes $18/30 = 0.6$. Table 3 displays convergence frequency to Nash equilibria under different pricing rules and rationing policies for each case. The lighter a point on charts, the higher the convergence frequency of the simulation under the corresponding parameter setting to Nash equilibria.

We observe that the convergence frequency to Nash equilibria decreases as the complexity of the case increases (from Case 2 to Case 4) because the number of states is substantially higher in the more complex cases. In all case studies, GenCos bids converge to Nash equilibria more frequently when $\epsilon_{i0} \in [0.7, 0.9]$, and they fail to converge when $\epsilon_{i0} < 0.1$. That is, convergence frequencies tend to decline as one moves left in any figure. In fact, in Appendix B, we show very low exploration to disrupt learning. Furthermore, we observe a high-convergence zone around $\epsilon_{i0} \approx 0.9$ for any α_{i0}; this is due to the shape of the linear decay function of ϵ_{it}.

An important difference between the equal and random rationing policies is seen in the figures: Under equal rationing, the regions of high and low convergence are clearly separated from each other, whereas under random rationing, we do not observe such separation. Recall that the random rationing policy of the ISO introduces another uncertain event to the decision-making process of GenCos when the ISO prioritizes among marginal GenCos with the same bid price. Imposing more randomness to the system disrupts the learning process of GenCos by blurring the link between successful bids and high profits. Thus, the ISO can use a random rationing policy in cases where it needs to interfere with GenCos' learning. This can be the case, for instance, when collusion opportunities are present for GenCos.

The figures also allow comparing the effectiveness of changing the pricing rule and changing the rationing policy in convergence to Nash equilibria. We observe the result of the comparison to be case-dependent: While the rationing policy change is more influential in Case 2, pricing rule is more important in Case 4. In Case 3, on the other hand, both factors seem to be influential.

Under DC-OPF pricing, settings with a high convergence frequency create a distinctive wedge shape extending from $\alpha_{i0} \approx 0.04$ and $\epsilon_{i0} \approx 0.9$ to $\alpha_{i0} \approx 1$ and $\epsilon_{i0} \approx 0.9$ while it is curved around $\alpha_{i0} \approx 0.1$ and $\epsilon_{i0} \approx 0.4$. This observation suggests that GenCos with low tendency to explore while giving sufficient importance to the last observed outcome are more likely to converge to Nash equilibria especially in Case 2 and Case 3.

Table 4 summarizes the observations we make from Table 3 by presenting the average (over different parameter settings) observed convergence frequencies to Nash equilibrium under different pricing rules and rationing policies. As stated before, convergence frequency is high for the simple case (Case 1) and relatively low for the more complex ones (say, Case 4). Uniform pricing causes higher convergence frequency to Nash equilibrium than pay-as-bid for most cases (except Case 3, under equal rationing). For Case 1, we observe no impact of either the pricing rule or the rationing policy on the behavior of the private GenCo.

Comparisons become clearer when one also includes convergence to semi-Nash states. Table 5 presents the convergence frequency to semi-Nash states and to either Nash or semi-Nash states (in parentheses). That latter frequency is observed to be higher under uniform than under pay-as-bid pricing rule; and higher under random than under equal rationing policy. Overall, we can claim that our simulations with learning agents do converge to Nash equilibria, or a state that has identical payoffs with a particular Nash equilibrium (a semi-Nash state) for the majority of parameter settings.

6.3. Competitive Bidding Analysis

In Section 6.2, we studied the convergence of GenCos' bid prices to Nash equilibria or semi-Nash states. In this section, we are interested in the competitive bidding behavior of the GenCos. To this end, we analyze the difference between the bid prices and the marginal cost of GenCos. In a highly competitive market, GenCos are likely to reduce their bid prices, leading to a smaller difference

Table 3: Convergence frequency to Nash equilibria

Uniform / Equal	Case 2	Case 3	Case 4
Uniform / Random	Case 2	Case 3	Case 4
Pay-as-bid / Equal	Case 2	Case 3	Case 4
Pay-as-bid / Random	Case 2	Case 3	Case 4
DC-OPF pricing	Case 2	Case 3	Case 4

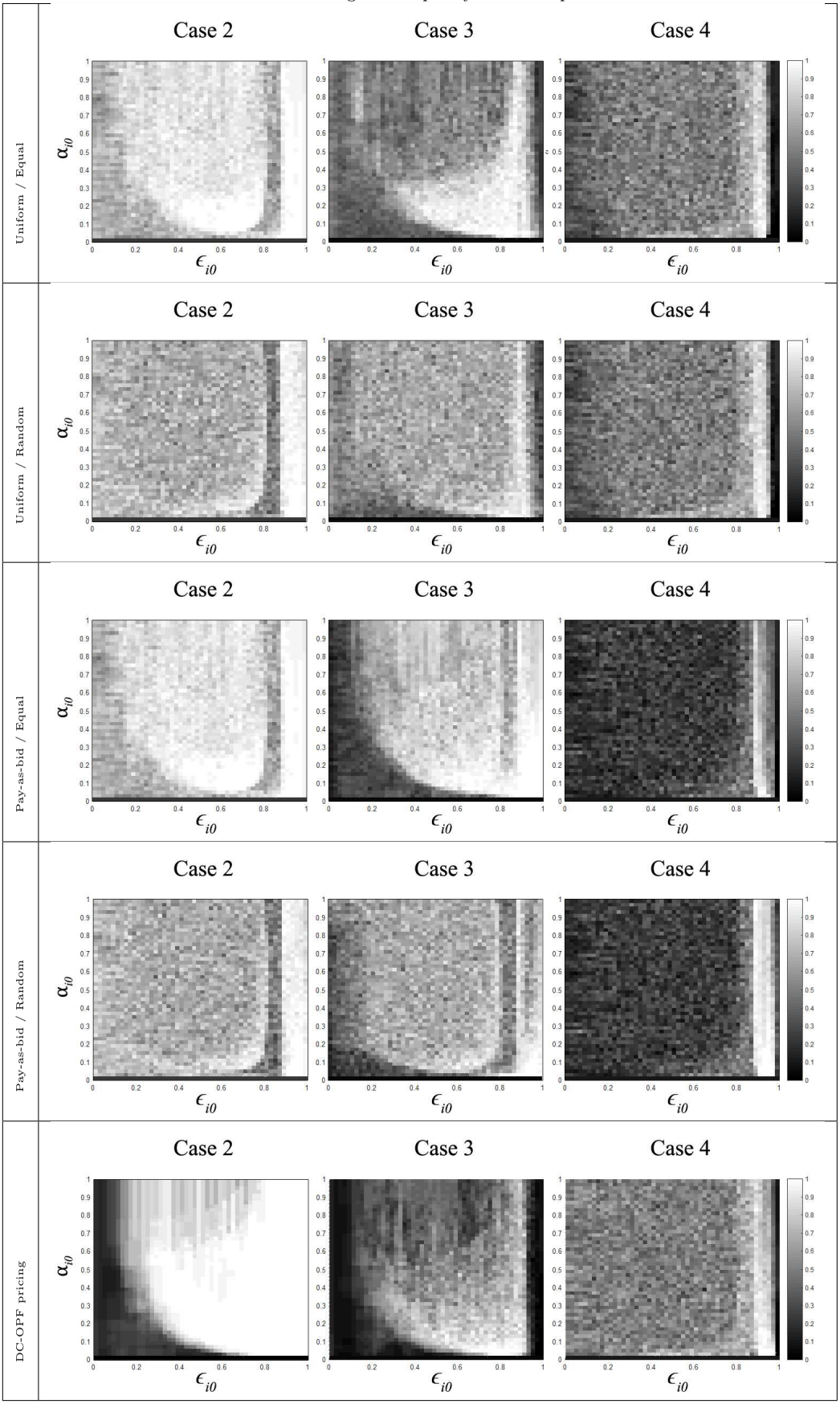

Table 4: Convergence frequency to Nash equilibria

Case Study	Uniform Equal	Uniform Random	Pay-as-bid Equal	Pay-as-bid Random	DC-OPF
Case1	0.8242	0.8242	0.8242	0.8242	0.8242
Case2	0.8396	0.7096	0.8396	0.7049	0.7865
Case3	0.5776	0.6230	0.6673	0.6133	0.4443
Case4	0.4848	0.4891	0.2906	0.3269	0.5246

between the bid price and the generation cost. In general, uniform pricing provides more incentives for bidding a closer price to marginal cost than pay-as-bid pricing does.

To investigate whether this expectation holds in our simulation results, we define $\Delta_i^k = b_i^{(k)*} - C_i$ for GenCo-i in replication k. Here, $b_i^{(k)*}$ denotes the best identified bid price at the end of replication k. For a given parameter setting (ϵ_{i0} and α_{i0}), we calculate the average Δ_i^k over all N GenCos and 30 replications as $\Delta = \frac{\sum_k \sum_i \Delta_i^k}{30 \times N}$.

Figure 4 presents the differences in Δ between the two pricing rules and the two rationing policies for Case 3 (as an example). For instance, Figure 4(A) shows the difference in Δ between equal and random rationing, under uniform pricing. All indicated difference values are found to be positive with the only exceptions marked with dark coloring. These exceptions, where the Δ difference is negative, are found in the rightmost side of Figure 4(A), and in a few islands in the middle of Figure 4(C). Thus, uniform pricing is found to be more successful in making GenCos submit closer bid prices to their marginal costs for almost all parameter settings. Likewise, equal rationing is observed to be more successful than random rationing. These observations are summarized in Table 6 which provides the average Δ values ($\overline{\Delta}$) over all parameter settings. A combination of uniform pricing and random rationing lead to the lowest $\overline{\Delta}$ values. Hence an ISO can use this combination to stimulate GenCos to submit lower bid prices in the market. DC-OPF rule, on the other hand, is again not observed to perform better than the other mechanisms.

6.4. GenCo Profit Analysis

In the previous section, we observed uniform pricing to achieve lower bid prices than pay-as-bid. However, this difference does not necessarily lead to lower GenCo profits under uniform pricing because of the difference in payment mechanisms. In this section, we compare GenCos' total profits under different market clearing mechanisms. This analysis is important because higher GenCos profits indicates that the market fails to provide affordable electricity for consumers.

For a given parameter setting $(\alpha_{i0}, \epsilon_{i0})$, we first calculate the average profit of each GenCo-i over 2000 iterations and 30 replications as $\overline{r_i} = \frac{\sum_{k=1}^{30} \sum_{t=1}^{max_t} r_{it}^k}{30 \times max_t}$. Next, we calculate GenCo-i's average profit over all parameter settings. These are reported for all GenCos over Case 2, Case 3, and Case 4 in Table 7, Table 8, and Table 9, respectively.

We observe clearly that switching from uniform to pay-as-bid pricing rule decreases GenCos' total profits, regardless of the rationing policy. For Case 2, the effect is more tangible because GenCo-2 has only one bid price which is equal to its generation cost.

We shall also investigate the profit difference between the market-clearance mechanisms statistically using 2601 parameter settings $(\alpha_{i0}, \epsilon_{i0})$ as the samples. For each case study and each rationing policy, we test whether the difference of the median profits between uniform pricing and pay-as-bid pricing is zero or positive while the difference is calculated by subtracting the profit under pay-as-bid pricing from that under uniform pricing. Figure 5 shows the histograms for the three case studies. We observe almost all differences to be positive, indicating higher GenCo profits under uniform pricing for almost all parameter settings. In fact, the median profit difference is found to be statistically higher than zero (*Nonparametric Sign Test* with $p-$values around 0.0000). In addition, in Case 2 and Case 3, the profit difference between the pricing rules under equal rationing is observed to be more pronounced from the one under random rationing. Finally we observe the profit distribution among GenCos to be quite different under the DC-OPF rule compared to those under the other clearance mechanisms. This result highlights the role of locational marginal pricing in DC-OPF, which takes the transmission grid structure and constraints into accounts in determining local electricity prices.

7. Conclusion

We study the effect of the ISO's market-clearing mechanism on the bidding behavior of GenCos in an electricity market in the presence of GenCo learning. We compare the results under two well-known pricing rules along with two different rationing policies. We also consider the DC-OPF formulation as an alternative mechanism that takes transmission line constraints into account.

We develop an agent-based simulation model that captures the learning dynamics of competing GenCos in the repetitive market. Our learning algorithm is an extension of the standard Q-learning algorithm with time-decaying parameters. The simulation model is implemented on four case studies representing different levels of market complexity. Results are reported under all possible combinations of the two key learning-model parameters.

Simulation results indicate that under most parameter settings, the market does converge to either a Nash equilibrium or a state that has identical payoffs with a particular Nash Equilibrium (a semi-Nash state, as we define it). Thus, GenCos individual learning moves the market towards the game-theoretical prediction in a good number of, if not in all cases. This is an interesting result as individual GenCos are not aware of the parameters of other GenCos. The convergence frequency to Nash Equilibria is found to be lower for more complex cases. Hence, learn-

Table 5: Convergence frequency to semi-Nash* states

Case Study	Uniform Equal rat.	Uniform Random rat.	Pay-as-bid Equal rat.	Pay-as-bid Random rat.	DC-OPF
Case 2	0.1081 (0.9477)	0.2683 (0.9779)	0.1081 (0.9477)	0.2722 (0.9771)	N/A (0.7865)
Case 3	0.1864 (0.7640)	0.2748 (0.8978)	0.0845 (0.7518)	0.2517 (0.8650)	0.2253 (0.6696)
Case 4	0.4369 (0.9217)	0.4367 (0.9258)	0.3314 (0.6220)	0.3430 (0.6699)	0.4418 (0.9664)

* Values inside parentheses indicate the summation of convergence frequencies to Nash equilibria and semi-Nash states.

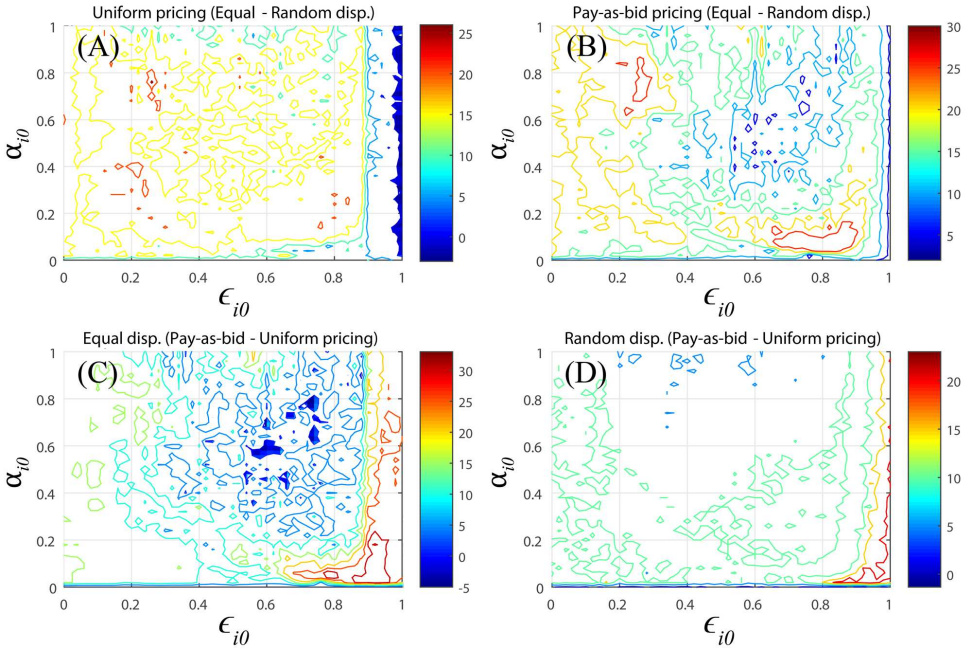

Figure 4: Difference in Δ between the two pricing rules and the two rationing policies for Case 3

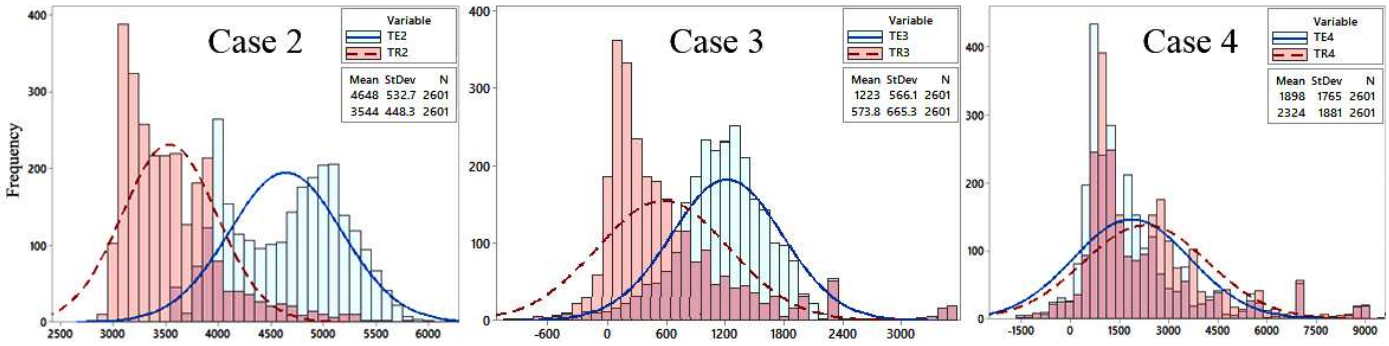

Figure 5: Histogram plot of profit differences between uniform and pay-as-bid pricing where TXY denotes the difference vector under rationing policy X (E represents Equal and R represents Random) in Case Y

Table 6: $\overline{\Delta}$ comparison

Case Study	Uniform Equal rat.	Pay-as-bid Equal rat.	Uniform Random rat.	Pay-as-bid Random rat.	DC-OPF
Case 1	**27.510**	**27.510**	**27.510**	**27.510**	**27.510**
Case 2	21.244	21.244	**10.896**	10.915	20.592
Case 3	26.247	37.176	**12.781**	22.107	40.680
Case 4	19.209	27.696	**19.110**	26.969	19.753

ing GenCos are not likely to behave as predicted by game theory in more sophisticated settings.

We use the simulation study to investigate two aspects of different market clearing mechanisms: Competitiveness of GenCo bid prices and GenCos' total profits. Uniform pricing is found to be more successful than pay-as-bid pricing in making GenCos submit closer bids to their marginal

Table 7: GenCos' profits under different market-clearing mechanisms

Case Study 2					
	Uniform Equal rat.	Uniform Random rat.	Pay-as-bid Equal rat.	Pay-as-bid Random rat.	DC-OPF
GenCo-1	1944.07	1966.18	1944.07	1965.28	1007.61
GenCo-2	4647.61	3542.70	0.00	0.00	1060.85
GenCo-5	571.73	172.90	571.73	172.04	2264.00
Total	7163.40	5681.78	2515.79	2137.32	4332.47

Table 8: GenCos' profits under different market-clearing mechanisms

Case Study 3					
	Uniform Equal rat.	Uniform Random rat.	Pay-as-bid Equal rat.	Pay-as-bid Random rat.	DC-OPF
GenCo-1	3393.48	2822.74	2709.80	2504.06	5085.20
GenCo-2	3434.35	2829.77	2756.82	2517.76	1844.68
GenCo-5	755.65	275.91	893.53	332.77	2233.11
Total	7583.48	5928.43	6360.15	5354.60	9162.99

costs. At the same time, however, GenCos' total profit is also higher under uniform pricing, indicating lower consumer benefit. Because all GenCos are paid at the highest accepted bid price, uniform pricing can lead to higher total profits than pay-as-bid although the submitted bid prices are lower. This result approximately holds true at GenCo level (except GenCo-5 of Case 3 and GenCo-2 of Case 4). Therefore, in accordance with Azadeh et al. [3], our results suggest pay-as-bid as the better pricing rule for the ISO.

In addition to the pricing rule, we find the rationing policy to have significant effects on GenCos' bids and their profits. Compared with equal rationing, random rationing policy is seen to be more effective in reducing total profits. This can partly be explained by the disruption in GenCos' learning process due to random rationing. This finding can be instrumental if the ISO needs to prevent GenCos' learning towards, for instance, a collusive equilibrium. Overall, a combination of pay-as-bid pricing and random rationing turns out to be the best combination to limit GenCos' total profits.

The agent-based simulation model in this study is a detailed and versatile one. We plan to extend this model to address further questions on strategic interactions in electricity markets. One future research direction is the study of GenCos' collusive behavior [1]. Considering capacity withholding would allow GenCos bid pairs of price and quantity, and it may facilitate to a more realistic model. Assuming that the GenCo submits its whole capacity, we chose to ignore the quantity choice. It helps to focus on our research questions and avoid additional computational burden due to curse of dimensionality. Thus, capacity withholding can also be investigated by making quantity a second dimension of the GenCos' bid. Yet another subject of study might be the effects of a second market.

Table 9: GenCos' profits under different market-clearing mechanisms

Case Study 4					
	Uniform Equal rat.	Uniform Random rat.	Pay-as-bid Equal rat.	Pay-as-bid Random rat.	DC-OPF
GenCo-2	5273.27	5318.39	5505.16	5422.45	4962.80
GenCo-3	14783.93	14796.97	12757.13	12454.59	15900.05
GenCo-4	439.73	421.12	336.70	335.89	375.96
Total	20496.93	20536.48	18598.98	18212.93	21238.81

Table A.10: Transmission line properties

Src (k)/ Dst (l)	y_{kl}	F_{kl}^{max} (MW)
1/2	4	20
2/3	4	No Limit

Table A.11: Parameters of the GenCos

ID	P_i^{max} (MW)	C_i (\$/MWh)	B_i (\$/MWh)
1	110	10	$\{10\}$
3	100	10	$\{10, 20, 30, 40\}$

Appendix A. Details of Case Studies

Appendix A.1. Case 1: Public GenCo versus Private GenCo

The simplified transmission grid is illustrated in Figure A.6. Node 1 represents the public GenCo company and Node 3 represents the private one while Node 2 represents a load/demand center. The properties of transmission lines are shown in Table A.10; the first column $Src(k)/Dst(l)$ shows the source and the destination nodes of transmission line; the second column depicts the value of the admittance parameter of the line and the last column shows the maximum flow on the line. The public GenCo benefits from subsidies to keep the price of energy as low as possible. As a result, it offers a bid price which is equal to its marginal cost (C_1). The list of possible bid prices (\$/MWh) for both companies are given in Table A.11 along with their generation capacities and marginal generation costs.

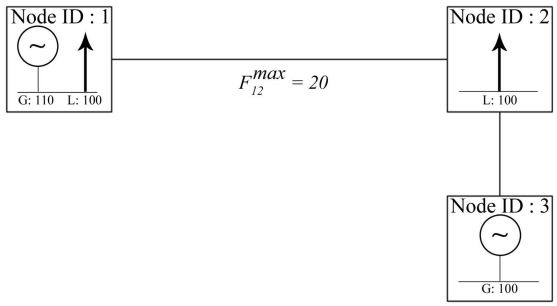

Figure A.6: The transmission grid for Case 1

Appendix A.2. Case 2: Two GenCos as Learning Agents

In the second case study, we have a five-node transmission grid governing the power market. The properties of transmission lines are given in Table A.12. The network structure along with generation capacities and demand load data are given in Fig.A.7 and Table A.13, respectively. Node 3 is the reference bus in this system (i.e. the voltage angle of reference bus is zero in DC-OPF). Node 1 and Node 5 are the GenCos that benefit from learning.

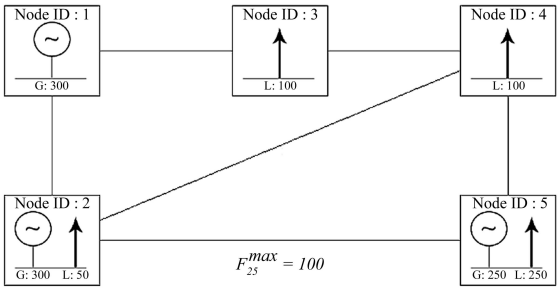

Figure A.7: The network for Case 2 and Case 3

Table A.12: Transmission line properties in Case 2

Src (k)/ Dst (l)	y_{kl}	F_{kl}^{max} (MW)
$\{1/2, 1/3, 2/4, 3/4, 4/5\}$	4	No Limit
2/5	4	100

Appendix A.3. Case 3: Three GenCos behaving as learning agents

If we change the second case study's bid alternatives for b_2 from $\{20\}$ to $\{20, 30, 40, 50\}$ the system would have multiple Nash equilibria. Table A.14 shows all possible outcomes under DC-OPF pricing: light-gray cells are the best responses of the GenCo-2 to the action of GenCo-1, bold-text cells are the best responses of the GenCo-1 to the action of the GenCo-2, and italic-text cells are the best responses of the GenCo-5 to the given action of GenCo-1 and GenCo-2. The intersection of all best responses are highlighted in dark-gray; they represent the Nash equilibria.

Appendix A.4. Case 4: Three learning GenCos with a centralized demand node

In this case, we have created a small market with three GenCos; Fig.A.8 shows structure of undertaken market. The second GenCo benefits from wind power technology; this is why, generation cost is negligible in Table A.15. Thus, GenCo-2 can bid a lower price to the ISO (cost of not fulfilling promised demand is considered in the price of electricity).

Appendix B. Boundary Analysis of the Proposed Learning Algorithm

To keep tracking the evolution of Q-values over iterations, we modify the Q-value and payoff notations. $Q_{ij}^{(t)}$ stands for the Q-value of b_{ij} at iteration t. Also, the received payoff of GenCo-i at iteration t is $r_i^{(t)}$.

Table A.13: Parameters of GenCos in Case 2

ID	P_i^{max} (MW)	C_i (\$/MWh)	B_i (\$/MWh)
1	300	20	$\{20, 30, 40, 50\}$
2	300	20	$\{20\}$
5	250	30	$\{30, 40, 50\}$

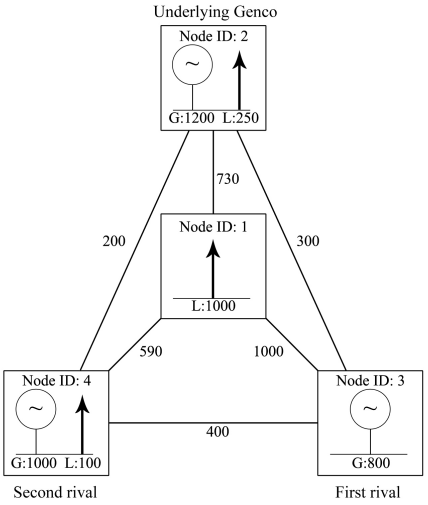

Figure A.8: Structure of market

Appendix B.1. Conventional Q-learning

by using mathematical induction we know,

$$\begin{aligned} &Q_{ij}^{(n)} = (1-\alpha_i)Q_{ij}^{(n-1)} + \alpha_i r_i^{(n)}; Q_{ij}^{(0)} = 0 \\ &\text{when } t=1 \Rightarrow Q_{ij}^{(1)} = (1-\alpha_i)Q_{ij}^{(0)} + \alpha_i r_i^{(1)} \\ &\text{when } t=2 \Rightarrow Q_{ij}^{(2)} = (1-\alpha_i)Q_{ij}^{(1)} + \alpha_i r_i^{(2)} \\ &= \alpha_i r_i^{(2)} - (\alpha_i^2 - \alpha_i) r_i^{(1)} \end{aligned} \tag{B.1}$$

The closed form is as follow.

$$Q_{ij}^{(n)} = (\alpha_i) \sum_{t=1}^{n} r_i^{(t)} (1-\alpha_i)^{(n-t)} \tag{B.2}$$

Boundary Condition 1.

When $\alpha_i = 1 \Rightarrow Q_{ij}^{(n)} = r_i^{(n)}$ from Eq.(B.2) by setting $\alpha_i = 1$ just for $t = n$ we get 1 from $(1-\alpha_i)^{n-t} = 0^0 = 1$.

Boundary Condition 2.

When $\epsilon_i = 0$. ϵ parameter determines how much exploration should be done. Therefore, if $\epsilon_i = 0$ then the GenCo only selects b_{ij} corresponds to $Q_{ij}^{(1)}$ (by assuming positive $r_i^{(1)}$), because all other $Q_{iJ}^{(1)} = Q_{iJ}^{(0)} = 0$ when $J \neq j$.

$$\begin{aligned} Q_{ij}^{(n)} &= r_i^{(1)}(1-\alpha_i)^n \left(\left(\frac{1}{1-\alpha_i} \right)^n - 1 \right) \\ &= r_i^{(1)}(1-(1-\alpha_i)^n) = r_i^{(1)} - (1-\alpha_i)^n r_i^{(1)} \end{aligned} \tag{B.3}$$

So, if we assume to conduct experiment up to infinity ($max_t = \infty$) then $\lim_{n\to\infty} Q_{ij}^{(n)} = r_i^{(1)}$. Therefore, GenCos cannot make more profit per iteration by increasing the number of iterations.

Boundary Condition 3.

Finally, if $\alpha_i = 0$ then $Q_{ij}^{(n)} = 0$ from Eq.(B.2).

Table A.14: Profit of each policy $\{r_1, r_2, r_5\}$ - Rows: B_1, Columns: B_2 and separated tables: B_5

$b_5 = 30$	20	30	40	50
20	**(428.57, 0, 0)**	**(3000, 785.71, 0)**	**(3000, 0, 0)**	(3000, 0, 0)
30	(0, 3000, 0)	(0, 3000, 0)	(2500, 0, 0)	(2500, 0, 0)
40	(0, 3000, 0)	(0, 3000, 0)	*(0, 5000, 2500)*	***(5000, 0, 2500)***
50	(0, 3000, 0)	(0, 3000, 0)	(0, 5000, 2500)	*(0, 7500, 5000)*

$b_5 = 40$	20	30	40	50
20	**(857.14, 0, 1214.29)**	**(3428.57, 785.71, 1214.29)**	**(6000, 1571.43, 1214.29)**	**(6000, 0, 2000)**
30	(416.67, 2500, 1583.33)	**(3428.57, 785.71, 1214.29)**	**(6000, 1571.43, 1214.29)**	**(6000, 0, 2000)**
40	(0, 6000, 2000)	(0, 6000, 2000)	(0, 6000, 2000)	*(5000, 0, 2500)*
50	(0, 6000, 2000)	(0, 6000, 2000)	(0, 6000, 2000)	*(0, 7500, 5000)*

$b_5 = 50$	20	30	40	50
20	***(1285.71, 0, 2428.57)***	***(3857.14, 785.71, 2428.57)***	***(6428.57, 1571.43, 2428.57)***	***(9000, 0, 4000)***
30	*(416.67, 2000, 3166.67)*	***(3857.14, 785.71, 2428.57)***	***(6428.57, 1571.43, 2428.57)***	***(9000, 2357.14, 2428.57)***
40	*(833.33, 5500, 3166.67)*	*(833.33, 5500, 3166.67)*	**(6428.57, 1571.43, 2428.57)**	**(9000, 2357.14, 2428.57)**
50	*(0, 9000, 4000)*	*(0, 9000, 4000)*	*(0, 9000, 4000)*	(0, 9000, 4000)

Table A.15: GenCos bidding sets and costs

ID	P_i^{max} (MW)	C_i (\$/MWh)	B_i (\$/MWh)
2	1200	10	$\{10, 20, 30, 40\}$
3	800	0	$\{9, 18, 20\}$
4	1000	15	$\{15, 25, 35, 45\}$

Appendix B.2. Q-learning with variable learning rate

The closed form of Q-values in presence of variable learning rate is as follows.

$$Q_{ij}^{(n)} = \sum_{t=0}^{n} \alpha_{it} r_i^{(t)} \Pi_{\hat{t}=t+1}^{n} (1 - \alpha_{i\hat{t}}) \tag{B.4}$$

Also, we have $\alpha_{it} = \alpha_{i0} - \frac{t}{n}(\alpha_{i0} - \frac{\alpha_{i0}}{10}) = \alpha_{i0} - \frac{9t\alpha_{i0}}{10n}$. Because in each iteration, $Q_{ij}^{(n)}$ is a convex combination of non-negative r_is, $Q_{ij}^{(n)}$ is non-negative.

Boundary Condition 4.

if $\epsilon = 0$ then from Eq.(B.4) we get following equation.

$$Q_{ij}^{(n)} = r_i^{(1)} \sum_{t=0}^{n} \alpha_{it} \Pi_{\hat{t}=t+1}^{n} (1 - \alpha_{i\hat{t}}) \tag{B.5}$$

because $r_i^{(n)} = r_i^{(1)}$. In this situation $Q_{ij}^{(n)}$ is not only affected by $r_i^{(1)}$ but is also a function of maximum number of iterations and the initial learning rate. We refer to $f(n) = \sum_{t=0}^{n} \alpha_{it} \Pi_{\hat{t}=t+1}^{n} (1 - \alpha_{i\hat{t}})$ as learning function when $\alpha_{it} = \alpha_{i0} - \frac{9t\alpha_{i0}}{10n}$. Fig.B.9 depicts the evolution of $f(n)$ over different n, considering different α_{i0}.

Proposition 1. (Bounded Learning Function) *$f(n)$ is an increasing bounded function that converges to a number between $[0, 1]$.*

Proof. we know that $f(n) = \sum_{t=0}^{n} \alpha_{i0} T_t \Pi_{\hat{t}=t+1}^{n} (1 - \alpha_{i0} T_{\hat{t}})$ when $T_t = 1 - \frac{9t}{10n} \in [1, 0.1]$ is a scale parameter and decreasing. Therefore, $f(n) \geq 0$ and $f(n) = 0$ when $\alpha_{i0} = 0$ which means no learning employed. However, we also need to find an upper bound for $f(n)$. The most dramatic increase in $f(n)$ happens when $\alpha_{i0} = 1$. Hence,

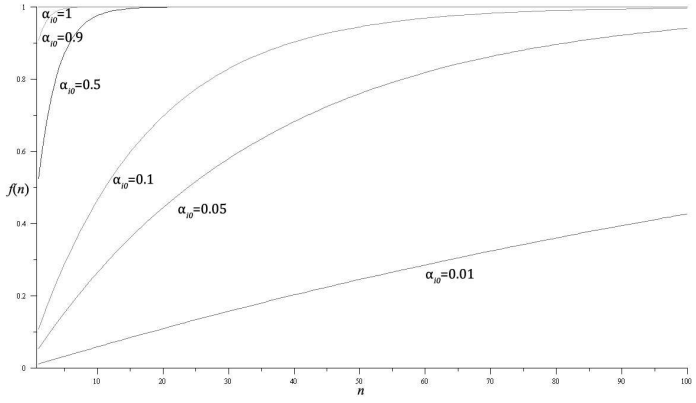

Figure B.9: effect of $f(n)$ for different α_{i0} over different n

$$\begin{aligned} f(n)|_{\alpha_{i0}=1} &= \sum_{t=0}^{n} T_t \Pi_{\hat{t}=t+1}^{n} (1 - T_{\hat{t}}) \\ &= \left(\frac{9}{10n}\right)^{n+1} \Gamma(n+1) \sum_{t=0}^{n} \frac{(1 - \frac{9t}{10n})}{\left(\frac{9}{10n}\right)^{t+1} \Gamma(t+1)} \\ &= -\frac{9(n+1)\left(1 - \frac{9(n+1)}{10n}\right)\left(\frac{9}{10n}\right)^{n+1} \Gamma(n+1)}{(-n+9)(\frac{9}{10n})^{n+2} \Gamma(n+2)} = 1 \end{aligned}$$

Thus, $f(n)$ is a bounded function $|f(n)| \leq 1$. Also, from definition of $f(n)$ we know, $f(n+1) - f(n) = \alpha_{in+1}(1 - f(n))$ and when $f(n) \leq 1$ therefore, $f(n+1) \geq f(n)$ hence $f(n)$ is an increasing sequence.

By using the Bolzano-Weistrass theorem, $f(n)$ converges to some point such as $f' \in [0, 1]$. □

By using Proposition.1, $Q_{ij}^{(n)} \leq r_i^{(1)}$. Also, $Q_{ij}^{(n)} = r_i^{(1)}$ when $\alpha_{i0} = 1$ for every $n > 0$ and $Q_{ij}^{(n)} = 0$ when $\alpha_{i0} = 0$.

Proposition 2. (Continuity of Q-value function) *for every* $\varepsilon > 0$ *there exists* $\delta_\alpha > 0$ *such that* $|\alpha_{i0} - 1| < \delta_\alpha$ *then* $\left|Q_{ij}^{(n)} - r_i^{(1)}\right| < \varepsilon$

Proof. By assuming n as real number, $f(n)$ is a continuous function, $Q_{ij}^{(n)} = r_i^{(1)} f(n)$ is also continuous and by definition of continuous function, we can find such an interval. Therefore, by increasing (n) a GenCo cannot make better profit than $r_i^{(1)}$. □

Thus, by checking our algorithm, one can comprehend by choosing $\epsilon_{i0} < \frac{8}{9}$ there exists $t \in [1, max_t]$ such that $\epsilon_{i\hat{t}} = 0$ when $\hat{t} \geq t$. Thus, conducting simulation for more iterations than t won't help GenCo-i to gain more profit per iteration than the best bid at time t (see Fig.B.10).

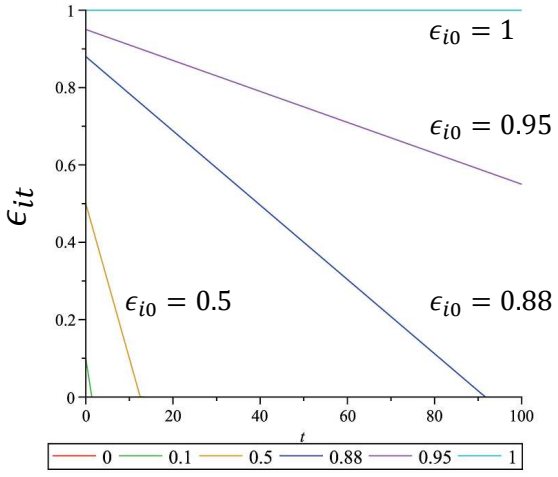

Figure B.10: Linear decaying function with different ϵ_{i0} over time

Boundary Condition 5.

Contrary to **Boundary Condition 1**, $Q_{ij}^{(n)} \neq r_i^{(n)}$ when $\alpha_{i0} = 1$ because α-value changes during iterations. Hence, the historical payoff information from GenCo-i's bids in previous iterations are considered when α_{it} is a monotone decreasing sequence ($\alpha_{it} > \alpha_{i\hat{t}}$ for $t < \hat{t}$).

References

[1] Aliabadi, D.E., Kaya, M., Şahin, G., 2016. Determining collusion opportunities in deregulated electricity markets. Electric Power Systems Research 141, 432–41.

[2] Aumann, R.J., 1987. Correlated equilibrium as an expression of Bayesian rationality. Econometrica: Journal of the Econometric Society , 1–18.

[3] Azadeh, A., Skandari, M., Maleki-Shoja, B., 2010. An integrated ant colony optimization approach to compare strategies of clearing market in electricity markets: agent-based simulation. Energy Policy 38, 6307–19.

[4] Bakirtzis, A.G., Tellidou, A.C., 2006. Agent-based simulation of power markets under uniform and pay-as-bid pricing rules using reinforcement learning, in: Power Systems Conference and Exposition, pp. 1168–73.

[5] Borenstein, S., Bushnell, J., 2000. Electricity restructuring: deregulation or reregulation? Regulation 23, 46–52.

[6] Cramton, P., 2004. Alternative pricing rules, in: Power Systems Conference and Exposition, IEEE. pp. 1621–3.

[7] David, A., Wen, F., 2000. Strategic bidding in competitive electricity markets: a literature survey, in: Power Engineering Society Summer Meeting, pp. 2168–73.

[8] Guo, M., Liu, Y., Malec, J., 2004. A new Q-learning algorithm based on the metropolis criterion. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34, 2140–3. doi:10.1109/TSMCB.2004.832154.

[9] Hart, S., Mas-Colell, A., 2001a. A general class of adaptive strategies. Journal of Economic Theory 98, 26–54.

[10] Hart, S., Mas-Colell, A., 2001b. A reinforcement procedure leading to correlated equilibrium. Springer.

[11] Holmberg, P., 2014. Pro-competitive rationing in multi-unit auctions IFN Working Paper No. 1037. Available at SSRN: http://ssrn.com/abstract=2485252.

[12] Kahn, A.E., Cramton, P.C., Porter, R.H., Tabors, R.D., 2001. Uniform pricing or pay-as-bid pricing: a dilemma for California and beyond. The Electricity Journal 14, 70–9.

[13] Kalai, E., Lehrer, E., 1993. Rational learning leads to nash equilibrium. Econometrica: Journal of the Econometric Society , 1019–45.

[14] Krause, T., Andersson, G., 2006. Evaluating congestion management schemes in liberalized electricity markets using an agent-based simulator, in: 2006 IEEE Power Engineering Society General Meeting, IEEE. pp. 8–15.

[15] Krause, T., Andersson, G., Ernst, D., Vdovina-Beck, E., Cherkaoui, R., Germond, A., 2004. Nash equilibria and reinforcement learning for active decision maker modelling in power markets, in: Proceedings of the 6th IAEE European Conference: Modelling in Energy Economics and Policy.

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

[17] Madlener, R., Kaufmann, M., 2002. Power exchange spot market trading in Europe: theoretical considerations and empirical evidence. Technical Report. Optimisation of Cogeneration Systems in a Competitive Market Environment. Available at http://www.oscogen.ethz.ch.

[18] Mailath, G.J., 1998. Do people play Nash equilibrium? lessons from evolutionary game theory. Journal of Economic Literature , 1347–74.

[19] Mount, T., 2001. Market power and price volatility in restructured markets for electricity. Decision Support Systems 30, 311–25.

[20] Naghibi-Sistani, M., Akbarzadeh-Tootoonchi, M., Javidi-Dashte Bayaz, M., Rajabi-Mashhadi, H., 2006. Application of Q-learning with temperature variation for bidding strategies in market based power systems. Energy conversion and management 47, 1529–38.

[21] Oren, S., 2004. When is pay-as-bid preferable to uniform price in electricity markets, in: Power Systems Conference and Exposition.

[22] Ott, A.L., 2003. Experience with PJM market operation, system design, and implementation. IEEE Transactions on Power Systems 18, 528–34.

[23] Sugianto, L.F., Liao, K.Z., 2014. Comparison of different auction pricing rules in the electricity market. Modern Applied Science 8, 147–63.

[24] Tierney, S.F., Schatzki, T., Mukerji, R., 2008. Uniform-pricing versus pay-as-bid in wholesale electricity markets: does it make a difference? Analysis Group AMD New York Independent System Operator , 1–24.

[25] Veit, D.J., Weidlich, A., Krafft, J.A., 2009. An agent-based analysis of the german electricity market with transmission capacity constraints. Energy Policy 37, 4132–44.

[26] Wang, J., 2009. Conjectural variation-based bidding strategies with Q-learning in electricity markets, in: 42nd Hawaii International Conference on System Sciences, IEEE. pp. 1–10.

[27] Wang, X., Sandholm, T., 2002. Reinforcement learning to play an optimal Nash equilibrium in team Markov games, in: Ad-

vances in neural information processing systems, pp. 1571–8.
[28] Wolfram, C.D., 1999. Electricity markets: Should the rest of the world adopt the united kingdom's reforms. Regulation 22, 48–53.
[29] Xiong, G., Okuma, S., Fujita, H., 2004. Multi-agent based experiments on uniform price and pay-as-bid electricity auction markets, in: Electric Utility Deregulation, Restructuring and Power Technologies, IEEE. pp. 72–6.