Competition, risk and learning in electricity markets: An agent-based simulation study

Danial Esmaeili Aliabadi *, Murat Kaya, Guvenc Sahin
Sabanci University, Faculty of Engineering and Natural Sciences, Istanbul, Turkey

HIGHLIGHTS
• The first ABMS study to consider both learning and risk aversion of GenCos.
• Presenting an agent-based simulation of GenCos’ behavior in electricity markets.
• Conducting a large-scale analysis with a wide range of learning model parameters.
• Analyzing effects of risk aversion on GenCos’ bid prices, profits, and learning.
• Showing that some level of risk aversion can improve GenCos’ profits.

ARTICLE INFO
Article history:
Received 21 December 2016
Received in revised form 24 March 2017
Accepted 25 March 2017

Keywords:
Electricity markets
Risk aversion
Q-learning
Agent-based simulation
Imperfect competition

ABSTRACT
This paper studies the effects of learning and risk aversion on generation company (GenCo) bidding behavior in an oligopolistic electricity market. To this end, a flexible agent-based simulation model is developed in which GenCo agents bid prices in each period. Taking transmission grid constraints into account, the ISO solves a DC-OPF problem to determine locational prices and dispatch quantities. Our simulations show how, due to competition and learning, the change in the risk aversion level of even one GenCo can have a significant impact on all GenCo bids and profits. In particular, some level of risk aversion is observed to be beneficial to GenCos, whereas excessive risk aversion degrades profits by causing intense price competition. Our comprehensive study on the effects of Q-learning parameters finds the level of exploration to have a large impact on the outcome. The results of this paper can help GenCos develop bidding strategies that consider their rivals’ as well as their own learning behavior and risk aversion levels. Likewise, the results can help regulators in designing market rules that take realistic GenCo behavior into account.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In this work, we present a wholesale electricity market simulation with learning agents. The agents are power generation companies (GenCos) that engage in repetitive hourly pool trading. We are concerned with how the learning behavior and risk aversion of competing GenCos will shape GenCo bid prices and profit levels.

Electricity markets are oligopolies, and electricity demand is often considered inelastic in the short term with respect to price. In addition, transmission line constraints, and the relative locations of electricity demand and supply can provide market power to individual GenCos. Due to all these reasons, GenCos can bid above their marginal costs and obtain positive profit. This possibility, and the importance of the power sector for the economy has triggered a

http://dx.doi.org/10.1016/j.apenergy.2017.03.121
© 2017 Elsevier Ltd. All rights reserved.
wave of research into the strategic bidding behavior of GenCos (see, for example, [1–4]).

In this study, building on [5], we develop a flexible agent-based simulation model (ABMS) to characterize the evolution of a dynamic electricity market under transmission grid constraints. ABMS models have gained popularity in electricity market research because they offer advantages over game-theoretical models such as the ability to model heterogeneous players and observe the dynamic evolution of the market. The distinction of our ABMS model is that it considers both the learning behavior and risk aversion of GenCos.

In practice, GenCos often make bid decisions without proper information on the characteristics (such as capacity, cost, and/or financial situation) and bid history of competing GenCos. There is, however, also the potential for learning due to the repetitive nature of trading as GenCos interact with each other every day and gain experience. Due to learning and adaptation, GenCos can be expected to exhibit time-variant bidding policies. While learning individually, through its bids, each GenCo also has an impact on all the prices and dispatch quantities in the market. In this study, we are interested in observing how the collective learning of the GenCos will change the market. For example, will a GenCo discover its strategic advantage, such as low cost or a favorable position in the network, and learn over time to take advantage of it? We model GenCo learning using the Q-learning approach. In particular, we extend the learning model of [5,6] by considering time-dependent learning model parameters, similar to [7].

The literature that addresses GenCo behavior generally assumes risk-neutral decision makers whose objective is to maximize only the expected profit. In reality, however, GenCos may act risk-averse because they are exposed to increased levels of risk due to fluctuations in hourly prices and dispatched power quantities. To study the effects of risk aversion, we adopt a model in which risk is captured through the variance of past realized profits. Second, we present a flexible simulation model that can characterize the evolution of a dynamic electricity market under transmission constraints and time-dependent learning parameters. Third, we show that a certain level of risk aversion can improve GenCos’ profits, whereas excessive risk aversion decreases profits due to intense price competition. Finally, we present a comprehensive study on the effects of Q-learning parameters, in which we find the level of exploration to have a large impact on results.

The remainder of the paper is organized as follows: In Section 2 we summarize the related literature. In Section 3, we present the model with risk-neutral GenCos, discussing the network and market structures as well as our learning model. In Section 4, we illustrate the learning model and simulation algorithm through two case studies. In Sections 5 and 6, we discuss the model with risk-averse GenCos and the related simulation study, respectively. Section 7 presents the simulation study regarding the effects of Q-learning parameters. Finally, we discuss the implications of our results in Section 8, and conclude in Section 9.

### 2. Literature survey

Ventosa et al. [8] provide a review of electricity market modeling approaches, classifying the literature into optimization, equilibrium and simulation models. Among these, game-theoretical models aim to characterize the equilibrium when players compete in quantity (Cournot competition [9]), in price (Bertrand competition [10]), or by submitting supply functions (supply function equilibrium [2,11]). There are also the more general conjectural variation type models [12–14]. Game-theoretical models have been extensively used because they offer insights into the strategic behavior of players and allow an easy derivation of equilibrium results. However, they are too stylized to reflect the realities of complex electricity markets [1]. Almost all game-theoretical models assume players to be rational, which often does not hold in practice, and implicitly assume GenCo behavior not to change over time. In addition, transmission grid constraints are ignored in most game-theoretic electricity market.

### Nomenclature

<table>
<thead>
<tr>
<th>Indices</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>i, k, l</td>
<td>GenCos or nodes</td>
</tr>
<tr>
<td>j</td>
<td>bids</td>
</tr>
<tr>
<td>t</td>
<td>simulation iterations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>set of transmission lines</td>
</tr>
<tr>
<td>Bi</td>
<td>set of available bids of GenCo-i</td>
</tr>
<tr>
<td>Hi</td>
<td>history of all realized profits for submitting bid price alternative b_i</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>number of nodes in the network</td>
</tr>
<tr>
<td>P_{max,i}</td>
<td>the maximum generation capacity of GenCo-i (MW)</td>
</tr>
<tr>
<td>C_i</td>
<td>production cost of GenCo-i ($/MW h)</td>
</tr>
<tr>
<td>D_k</td>
<td>power demand at node-k (MW)</td>
</tr>
<tr>
<td>P_{i,k}</td>
<td>thermal limit for real power flow on line k</td>
</tr>
<tr>
<td>y_{it}</td>
<td>negative of the susceptance value for line k</td>
</tr>
<tr>
<td>p_i</td>
<td>risk aversion level of GenCo-i</td>
</tr>
<tr>
<td>y_{it}</td>
<td>recency rate of GenCo-i at iteration t</td>
</tr>
<tr>
<td>y_{io}</td>
<td>initial recency rate of GenCo-i</td>
</tr>
<tr>
<td>e_{it}</td>
<td>exploration parameter of GenCo-i at iteration t</td>
</tr>
<tr>
<td>e_{io}</td>
<td>initial exploration parameter of GenCo-i</td>
</tr>
</tbody>
</table>

<p>| max_i     | number of iterations |</p>
<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\Theta_i</td>
<td>voltage angle at node-i (radians)</td>
</tr>
<tr>
<td>LMP_i</td>
<td>locational marginal price at node-i ($/MW h)</td>
</tr>
<tr>
<td>P_i</td>
<td>power injected by GenCo-i (MW)</td>
</tr>
<tr>
<td>b_{ij}</td>
<td>jth bid price alternative of GenCo-i ($/MW h)</td>
</tr>
<tr>
<td>\hat{b}_i</td>
<td>simplified notation for b_{ij} ($/MW h)</td>
</tr>
<tr>
<td>\hat{b}_i</td>
<td>the best identified bid price of GenCo-i ($/MW h)</td>
</tr>
<tr>
<td>P^*_i</td>
<td>average bid price submitted by GenCo-i over iterations ($/MW h)</td>
</tr>
<tr>
<td>r_j</td>
<td>realized profit of GenCo-i for submitting b_j ($)</td>
</tr>
<tr>
<td>\hat{r}_i</td>
<td>simplified notation for r_j ($)</td>
</tr>
<tr>
<td>\hat{r}_i</td>
<td>profit of the best identified bid price of GenCo-i ($)</td>
</tr>
<tr>
<td>\hat{r}^*_i</td>
<td>average realized profit of GenCo-i over iterations ($)</td>
</tr>
<tr>
<td>Q_{ij}</td>
<td>Q-value of GenCo-i for submitting b_{ij} ($)</td>
</tr>
<tr>
<td>\hat{Q}_{ij}</td>
<td>risk-modified Q-value ($)</td>
</tr>
<tr>
<td>\hat{Q}_{ij}</td>
<td>average Q-value of GenCo-i over different scenarios ($)</td>
</tr>
<tr>
<td>CP_i</td>
<td>cumulative profit of GenCo-i ($)</td>
</tr>
<tr>
<td>CP^*_i</td>
<td>average cumulative profit of GenCo-i over different scenarios ($)</td>
</tr>
</tbody>
</table>
studies [2,15,16], despite the fact that physical network character-
istics can lead to important differences in results [17].

Due to these shortcomings of game-theoretical models, Agent-
Based Modeling and Simulation has become a popular choice in
the modeling of electricity markets [18–21]. ABMS offers a flexible
computational approach to model GenCo agents that are heteroge-
eous in parameters such as generation cost and capacity, location
in the network, and possessed information. Moreover, ABMS
imposes minimal information requirements and avoids the multi-
ple equilibrium issues of the game-theoretical models. ABMS
allows the modeling of the two-way interaction between agents
and the market. The evolution of the resulting models can be
observed in detail, even for dynamic systems that are not in equi-
librium. Agent-based approach also assists in modeling learning
and adaptation in a dynamic environment.

In the ABMS electricity market literature, different types of rein-
forcement learning methods have been used to simulate the intel-
ligence of the players [21]. The most popular of these include Erev-
Roth reinforcement learning [22], Q-learning [23], and the tempo-
ral difference algorithm [24]. In [25,26], modified versions of the
Erev-Roth reinforcement approach has been applied to electricity
markets.

Q-learning is a model-free and state-dependent algorithm that
was originally designed to be used with a Markov Decision Process.
Q-learning algorithms have been extensively used in many appli-
cations such as industrial control, time sequence prediction and
robot soccer competition [27]. In [28], GenCos learn through a Q-
learning algorithm, yet when bidding for capacity, they also con-
sider their competitors’ actions through a conjectural-variation
based strategy. In [29], the effect of market power mitigation
strategies is analyzed through an agent-based study with Q-
learning. In the current paper, we apply the Q-learning model used
in [7], which extends the learning model of [5] by considering
time-decaying parameters. The use of time-decaying parameters
in Q-learning model is akin to the Metropolis criterion in simulated annealing (i.e., the SA-Q-learning algorithm in [30,28]). Similar to [6], we assume a state-independent version of the method in which Q-values are
expressed as functions of actions only.

Dahlgren et al. [31] provide an early review of risk assessment
methods in energy trading. In [32], the contract quantity determi-
nation problem is considered under uncertain generation and
imbalance prices. A number of researchers have formulated
stochastic programming models to develop bidding strategies
under supply and price risks [33–36]. Zheng et al. [37] provide a
recent review of the stochastic optimization literature that address
the unit commitment problem. GenCos’ self-scheduling problem is
also widely studied [38–44].

The aforementioned papers assume price-taking GenCos oper-
ating under perfect competition. Ventosa et al. [8]’s survey cites
[45] as the only work that addresses the risk management problem
of GenCos under imperfect competition, in which case GenCos
become price-makers. In [45], a Monte Carlo simulation model
is developed to capture hydro production and demand risks in elec-
tricity markets under Cournot competition. The authors use risk
measures such as value-at-risk (VaR) and profit-at-risk (PaR). In
[46], similar to our study, a competitive market with network con-
straints is considered. The authors solve a bi-level optimization
problem in which competing GenCos submit linear supply func-
tions at the first stage, and the ISO determines the dispatch at the
second stage. In [47], both pool and bilateral-contract structures
are analyzed. In [48], the integrated risk management problem
of a hydrothermal GenCo in an oligopolistic market is considered.
The risk exposure due to fuel price, water inflow, electrical
price and power demand uncertainties are represented by the
conditional value-at-risk (CVaR) approach. In [49], bidding
strategies for a single price-taker hydro GenCo are studied. Uncer-
tainty about competitor GenCo offers are represented through the
residual demand curve. In [50], the effect of risk aversion and for-
ward markets on capacity expansion and forward hedging deci-
sions of GenCos are analyzed. Similar to our work, these authors
observe that due to competition, GenCos in a Cournot duopoly
may obtain higher expected profits as they become more risk
averse. A similar observation is made in [51] regarding the invest-
ment decisions of natural gas suppliers in an oligopolistic gas
market.

Another stream of risk-related papers are those that address
the generation portfolio selection problem of a single GenCo. For
instance, [52] presents a stochastic programming model for the
integrated portfolio selection and scheduling problem of a risk-
averse hydro producer. In [53], a Monte-Carlo simulation tool is
developed to optimize a power portfolio composed of physical
and financial assets. In [54], a detailed production model with nat-
ural gas, wind and cascaded hydro units is considered. Using agent-
based simulation and Monte Carlo approaches, [55] discusses the
effect of risk aversion on power plant investment decisions. In
[56], a Monte Carlo simulation is developed for assessment of
low-carbon power plant proposals.

As discussed in the paragraphs above, the learning behavior and
risk attitude of GenCos have been separately studied in the litera-
ture. Yet, their joint effect has not so far been investigated, which
presents a gap in the literature. We are aware of only two pieces
of work that address both learning behavior and risk attitudes;
however, their model structures as well as definitions of learning
and risk are different from our work. Liu and Wu [57] present a
stochastic optimal control mechanism in an oligopolistic market
in which GenCos engage in Cournot competition. The adaptation
mechanism in this model is somewhat similar to the learning in
our model. The [57] model, however, ignores the bidding and price
formation details in the market, and the network physical struc-
ture. Rahimyan and Mashhadi [58] consider both Q-learning (a fuzzy version) and risk attitude of GenCos. However, different from
our model and the literature in general, GenCo risk attitude is char-
acterized as a combination of certain Q-learning parameter values,
without using a separate risk component in the model.

3. Model with risk-neutral GenCos

In this section, we assume risk-neutral GenCos that aim to max-
imize expected profit. We first discuss the network representation
and market structure in the framework of our study. Then, we pro-
vide the details of the GenCo learning model.

3.1. The network and market structure

We consider only the day-ahead market, ignoring the futures
markets and real-time markets. The market is an oligopoly with
a relatively small number of GenCos each having a single produc-
tion unit. All parameters related to GenCos, including demand,
capacity and costs, are steady. Line or generation outages are
ignored.

The physical transmission grid is represented using a network
in which nodes correspond to GenCos and arcs correspond to
transmission lines between GenCos. The GenCo that connects to
the system at node-i is referred to as GenCo-i. GenCo-i has gen-
eration capacity \( P_{i}^{\text{max}} > 0 \) and marginal production cost \( C_i \). Power
demand \( (\text{load}) \) in node-i, \( D_i \), is assumed to be constant and price-
elastic. The transmission line between nodes \( k \) and \( l \) has capacity
\( P_{kl}^\text{max} \) and susceptance \( y_{kl} \).

For every period, corresponding to an hour in the day-ahead
market, each GenCo submits a bid composed of a power quantity and price:
• Bid quantity: The GenCo is assumed to bid all its production capacity, without capacity withholding. This assumption implies that systems with storage capacity are ignored.

• Bid price: The GenCo chooses a bid price $b_i$ ($$/MW$$ h) among the exogenously-given bid price alternatives ($b_i \in B_i$). The bid price alternatives range from the GenCo’s marginal production cost $C_i$ to a given price cap in the market.

The information set of each GenCo-$i$ consists of the history $H_i$ of its own realized profit values for each bid price alternative $b_i$. The GenCo has no information about the generation capacity, marginal cost, bid prices or profits of other GenCos, or the total number of GenCos in the system.

The market is run by the Independent System Operator (ISO), which collects the bids and clears the market. For each period, the ISO solves the following DC-OPF problem as an approximation to the underlying AC-OPF problem. Note that AC-OPF problems are typically approximated by the more tractable DC-OPF problems that consider linearized power constraints [59].

\[
\min \sum_{i=1}^{n} b_i P_i \quad \text{(1)}
\]

subject to

\[
P_k - D_k = \sum_{(k,i) \in BR} y_{kj}(\theta_k - \theta_i), \quad \forall k \in \{1, \ldots, n\}, \quad \text{(2)}
\]

\[
P_i \leq P_{ij}^{\text{max}}, \quad \forall i \in \{1, \ldots, n\}, \quad \text{(3)}
\]

\[
[y_{kj}(\theta_k - \theta_i)] \leq P_{ij}^{\text{max}}, \quad \forall (k, l) \in BR. \quad \text{(4)}
\]

The objective (1) is to minimize the system-wide cost of power supply. Constraint (2) allows the surplus power in each node to flow via the transmission lines to the connected nodes. Constraint (3) is the generation capacity constraint of each GenCo. Constraint (4) is the power flow constraint on each transmission line. The problem (1)-(4) is a linear optimization problem as we assume a DC representation of the transmission network.

By solving the problem, the ISO determines the power $P_i$ to be dispatched by each GenCo-$i$, the voltage angle $\theta_i$ at each node-$i$, and the Locational Marginal Price $\text{LMP}_i$ at each node-$i$, which is the shadow price for Constraint (2). LMP at a node corresponds to the minimum cost of fulfilling the demand for an additional unit ($$MW$$) of power at that particular node. Based on the solution, each GenCo is paid the LMP at its location node multiplied by its power dispatch. Thus, the profit $r_i$ of GenCo-$i$ from bidding price $b_i$ at a particular period becomes

\[
r_i = P_i(\text{LMP}_i - C_i). \quad \text{(5)}
\]

Note that because $P_i$ and LMP$_i$ values are determined as a function of all bids submitted to the ISO, each GenCo’s profit is affected by the bid price choices of all GenCos. We refer to GenCo-$i$’s profit simply by $r_i$ when the particular bid price $b_i$ that resulted in the profit is not relevant for the discussion.

3.2. The learning model

To model the learning behavior of GenCos, we use the modified Q-learning algorithm of [7] in which each GenCo (agent) experiments with bid price alternatives and learns through experience. GenCo-$i$ keeps a set $H_i$ of past realized profits from bidding the price alternative $b_i$. This includes the zero profit realizations due to rejected bids. For each price alternative $b_i$, the GenCo calculates the Q-value $Q_{ij}$ which denotes the weighted average of past realized profits from bidding $b_i$. $Q_{ij}$ captures the expected profit GenCo-$i$ believes to obtain by bidding this price in the subsequent period.

When GenCo-$i$ bids price $b_i$ and obtains the profit $r_i$ in a particular period, the relevant Q-value is updated as

\[
Q_{ij} = (1 - \alpha_i)Q_{ij} + \alpha_i r_i. \quad \text{(6)}
\]

The history set is also updated as $H_i \leftarrow H_i \cup r_i$. The other bid price alternatives’ Q-values are unchanged. In this equation, the recency rate $\alpha_i$ determines the weight given to the most recent profit observation. If $\alpha_i = 1$, the last obtained profit ($r_i$) becomes the Q-value. In this case, the agent uses only the last period information for that price alternative, ignoring history. At the other extreme, if $\alpha_i = 0$, the Q-value will not be updated. To facilitate convergence, we assume $\alpha_i$ to start at $\alpha_0$, and decay linearly over periods to $\alpha_0/10$ according to the equation

\[
x_i = (1 - t/\text{max}_t)(\alpha_0) + (\alpha_0/10)(t/\text{max}_t), \quad \text{where max}_t \text{ is the number of periods. We have also studied the case of exponentially decaying $\alpha_i$ and found no significant difference in results.}
\]

The bid price alternative that maximizes the expected profit in a particular period is labeled as the GenCo’s best identified bid price. Note that the best identified bid price values are by definition time-dependent. We ignore their time index because the meaning will be clear from the context.

\[
b_i' = \max_{b_i} Q_{ij}. \quad \text{(7)}
\]

In choosing its bid price, the GenCo uses an $\epsilon$-greedy action selection rule [60], which is characterized by the exploration parameter $\epsilon$: In each period, GenCo-$i$ submits its best identified bid price $b_i'$ with probability $1 - \epsilon$. With probability $\epsilon$, the GenCo submits a randomly chosen price $b_{ij}$ from its set of bid price alternatives. Each alternative has an equal probability of being chosen. This randomization helps the GenCo in assessing the performance of different bid price alternatives. The approach aims to strike a balance between exploitation of the best identified bid price and exploration of possibly better bid prices. A high exploration parameter $\epsilon$ would cause the GenCo to search for better bid prices most of the time, slowing learning. A low $\epsilon$, on the other hand, may lead to local optimum solutions by causing the GenCo to stick prematurely to a particular $b_i'$. We use a time-decaying exploration parameter. This approach, which is similar to [30], is different from most works in the literature, in which $\epsilon$ is fixed (e.g., [61]). Starting from a relatively high initial value of $\epsilon_0$, the parameter decreases over time towards zero according to the equation $\epsilon_t = \max(0, \epsilon_0 + 8t(\epsilon_0 - 1)/\text{max}_t)$. Hence, exploration is favored in the initial periods. Over the course of the simulation, GenCo’s experience about the profitability of different bid price alternatives builds up, decreasing the need for exploration. Thus, the GenCo is more likely to exploit its experience by submitting its $b_i'$. Because $\epsilon$ decreases over time to zero, the GenCo’s bid price choice will converge to an alternative that hopefully maximizes its Q function.

The aforementioned learning model captures the dynamics of a GenCo’s bidding behavior over time. We refer to a GenCo that bids to maximize its utility and is subject to learning through this model as a learning GenCo. Note that in the model, the GenCo does not take any strategic action, that is, the GenCo does not consider the actions of other GenCos explicitly in its decision process. In fact, it does not have information on other GenCos. The GenCo is modeled as a simple agent that learns only from its own experience. GenCos’ collective behavior, however, may lead to strategic outcomes.

4. Simulation study with risk-neutral GenCos

A simulation run in our study consists of max$_i$ iterations. Each iteration corresponds to the settlement of an hourly auction (one period) in the day-ahead market. In each iteration, GenCos simultaneously bid prices $b_i$ to the ISO. The ISO then solves the DC-OPF
problem given in (1)–(4) to determine the power $P_i$ to be dispatched by each GenCo-$i$, and the nodal price $LMP$, at each node.

At the beginning of the simulation, the following variables are initialized as $t = 1$, $Q_{ij} = 0$, $H_{ij} = \emptyset$, whereas $\alpha_0$ and $\epsilon_0$ are set to their initial values. One simulation run with 2000 iterations takes less than two seconds on an Intel Core i7 @ 3.2 GHz computer with 24 GB RAM. To obtain robust results, we report the average result over a number of simulation runs, each having a different random number seed. This is similar to the literature [15]. The random numbers are used for simulating the $\epsilon$-greedy action selection rule of each GenCo in each iteration: First, in determining whether the GenCo submits its best identified bid price, and if this price is not to be submitted, in determining the bid price to submit among the alternatives.

To illustrate our learning model and simulation algorithm, we use two case studies that are based on [6]. In both case studies, we consider the five-node transmission grid presented in Fig. 1. This network topology, which follows from [5], is inspired by the real Pennsylvania-NewJersey-Maryland (PJM) five node power system. The (negative) susceptance and the maximum flow values of the transmission lines are summarized in Table 1. Node-3 is the reference bus in this system.

4.1. Case 1: Two learning GenCos and a unique Nash equilibrium

In this case, GenCo-1 and GenCo-5 are the learning GenCos, both having $\epsilon_0 = 0.9$ and $\alpha_0 = 0.1$. Other GenCo parameters are summarized in Table 2. We report results from a single sample simulation run that has 300 iterations.

Table 3 shows the profits resulting from each possible bid price profile $(b_i, b_5)$ for the learning GenCos. The highlighted profile $(20, 50)$ is the only pure strategy Nash equilibrium of the stage game. This profile also happens to provide the maximum total profit for GenCos. Fig. 2 illustrates the evolution of $Q$-values for each bid price alternative during the simulation. We observe that the two learning GenCos eventually reach the Nash equilibrium with bids $(b_1 = 20$ and $b_5 = 50)$. The graph for GenCo-5 shows that it takes some iterations to learn to bid $\$50$/MW h.

4.2. Case 2: Three learning GenCos and multiple Nash equilibria

In this case, GenCo-2 in Case 1 also becomes a learning GenCo, and its bid price alternatives are extended from $\{20\}$ to $\{20, 30, 40, 50\}$. With this new setting, as seen in the profit values of Table 4, the stage game has two Nash equilibria as $\{20, 40, 50\}$ and $\{30, 50, 50\}$. The initial $Q$-learning parameters are $\epsilon_0 = 0.85$ and $\alpha_0 = 0.15$ for all GenCos.

We aim to observe where the simulations will converge. To this end, 10,000 simulation runs are conducted, each having 2000 iterations. Overall, we observe most simulation runs to converge to either one of the two Nash equilibria, or a state that provides a similar profit profile to a Nash equilibria. In fact, 65.6% of the runs converge to the Nash equilibrium $(20, 40, 50)$ and 34.4% converge to the Nash equilibrium $(30, 50, 50)$. Compared to the latter, the former equilibrium provides higher social welfare, i.e., lower cost of power as measured by the DC-OPF objective function value, and a more equitable profit distribution among GenCos.

The simulation runs also converge to state $(30, 40, 50)$ with probability 25% and to $(40, 50, 50)$ with probability 34%. These Non-Nash states provide an identical profit profile to one of the Nash equilibria. Such states are referred to as semi-Nash by [7]. Apparently, in search of better profits, the $Q$-learning behavior of GenCos can make the market converge to even a non-Nash state if this state provides reasonable profits.

5. Model with risk-averse GenCos

Here, we present a model with risk averse GenCos, which encompasses the risk-neutral model of Section 3 as a special case. In this risk-averse model, the utility of bid price alternative $b_i$ to GenCo-$i$ is increasing in its $Q$-value $Q_{ij}$, and decreasing in the standard deviation of past realized profits from this alternative, which are recorded in the set $H_i$. Accordingly, GenCo-$i$’s best identified bid price is determined using the risk-modified $Q$-values as $b_i = \max_{\bar{b}_i} Q_{ij} \beta$.

where $Q_{ij}^\beta = (1 - \beta)|Q_{ij} - \beta| \sum_{k=1}^n \frac{H_{ij} - Q_{ij}^\beta}{|H_{ij} - 1|}$.

Parameter $\beta \in (0, 1)$ denotes the risk aversion level of the GenCo where $\beta = 0$ corresponds to the risk-neutral case and higher $\beta$ values correspond to more risk-aversion. $Q$-values are updated similar to the risk-neutral model, based on realized profits as given in Eq. (6).

6. Simulation study with risk-averse GenCos

To address risk aversion, two modifications are made in the simulation algorithm. First, $b_i^\prime$ is now determined based on the
risk-modified $Q'$-values as shown in Eq. (8). Second, the variance component of Eq. (8) is ignored in the initial half of the iterations because there may be too few observations to support the required variance calculations.

We study the effects of risk aversion on GenCos’ bid prices and profits using a new case study, Case 3. The grid structure and GenCos’ parameters for this case are presented in Fig. 3 and Table 5, respectively. This structure provides GenCo-3 advantage due to zero generation cost, whereas GenCo-4 is at an unfavorable position with a relatively high generation cost. All reported results are averages over 30 random runs each having 2000 iterations. The initial Q-learning parameters are $c_0 = 0.85$ and $x_0 = 0.15$ for all GenCos.

### 6.1. Identical risk aversion level

Here, we assume all three GenCos to have the same $\beta$. We initially focus on the picture at the end of the simulation. Fig. 4(a) provides the $b'_i$ value at iteration 2000 for each GenCo, averaged

![Graph showing Q-value evolutions for GenCo-1 and GenCo-5 in Case 1.](image-url)

**Fig. 2.** Q-value evolutions for GenCo-1 and GenCo-5 in Case 1.

<table>
<thead>
<tr>
<th>$b_1$ = 30</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(428.57, 0, 0)</td>
<td>(3000, 785.71, 0)</td>
<td>(3000, 0, 0)</td>
<td>(3000, 0, 0)</td>
</tr>
<tr>
<td>30</td>
<td>(0, 3000, 0)</td>
<td>(0, 3000, 0)</td>
<td>(2500, 0, 0)</td>
<td>(2500, 0, 0)</td>
</tr>
<tr>
<td>40</td>
<td>(0, 3000, 0)</td>
<td>(0, 3000, 0)</td>
<td>(5000, 0, 2500)</td>
<td>(5000, 0, 2500)</td>
</tr>
<tr>
<td>50</td>
<td>(0, 3000, 0)</td>
<td>(0, 3000, 0)</td>
<td>(0, 5000, 2500)</td>
<td>(0, 7500, 5000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$b_1$ = 40</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(857.14, 0, 1214.29)</td>
<td>(3428.57, 785.71, 1214.29)</td>
<td>(6000, 1571.43, 1214.29)</td>
<td>(6000, 0, 2000)</td>
</tr>
<tr>
<td>30</td>
<td>(416.67, 2000, 3166.67)</td>
<td>(3428.57, 785.71, 1214.29)</td>
<td>(6000, 1571.43, 1214.29)</td>
<td>(6000, 0, 2000)</td>
</tr>
<tr>
<td>40</td>
<td>(0, 6000, 2000)</td>
<td>(0, 6000, 2000)</td>
<td>(6428.57, 1571.43, 2428.57)</td>
<td>(9000, 2357.14, 2428.57)</td>
</tr>
<tr>
<td>50</td>
<td>(0, 6000, 2000)</td>
<td>(0, 6000, 2000)</td>
<td>(0, 6000, 2000)</td>
<td>(0, 7500, 5000)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$b_1$ = 50</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(1285.71, 0, 2428.57)</td>
<td>(3857.14, 785.71, 2428.57)</td>
<td>(6428.57, 1571.43, 2428.57)</td>
<td>(9000, 4000)</td>
</tr>
<tr>
<td>30</td>
<td>(416.67, 2000, 3166.67)</td>
<td>(3857.14, 785.71, 2428.57)</td>
<td>(6428.57, 1571.43, 2428.57)</td>
<td>(9000, 2357.14, 2428.57)</td>
</tr>
<tr>
<td>40</td>
<td>(833.33, 5500, 3166.67)</td>
<td>(833.33, 5500, 3166.67)</td>
<td>(6428.57, 1571.43, 2428.57)</td>
<td>(9000, 2357.14, 2428.57)</td>
</tr>
<tr>
<td>50</td>
<td>(0, 9000, 4000)</td>
<td>(0, 9000, 4000)</td>
<td>(0, 9000, 4000)</td>
<td>(0, 9000, 4000)</td>
</tr>
</tbody>
</table>

*Table 4* 

Profits $(r_1, r_2, r_3)$ of bid profiles $(b_1, b_2, b_3)$ in Case 2 where Rows: $B_1$, Columns: $B_2$ and separated tables: $B_3$. 

![Graph showing Q-value evolutions for GenCo-1 and GenCo-5 in Case 1.](image-url)
observed in Fig. 4(a) are driven by the changes in 

to different levels of risk aversion. In fact, the profit results 

profit all GenCos.

Table 5
GenCo parameters in Case 3.

<table>
<thead>
<tr>
<th>ID</th>
<th>p\textsuperscript{max} (MW)</th>
<th>C\textsubscript{i} ($/MW h)</th>
<th>B\textsubscript{i} ($/MW h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1200</td>
<td>10</td>
<td>{10, 20, 30, 40}</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>0</td>
<td>{9, 18, 20}</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>15</td>
<td>{15, 25, 35, 45}</td>
</tr>
</tbody>
</table>

over 30 runs, as a function of the identical 

becomes excessively concerned about the variability in profits and bids its marginal cost 10 more frequently. For even higher \( \beta \) values, GenCo-2 only bids 10, resulting in zero profits. Such low bids by GenCo-2 causes a significant reduction in the profit of competitor GenCo-3 as well. For sufficiently high \( \beta \) values, both GenCo-2 and GenCo-3 submit their marginal generation costs to minimize the variability in their profits. GenCo-4, meanwhile, is observed to obtain zero profit at the end of the simulation independent of \( \beta \).

We have discussed the end-of-simulation results when each GenCo-i bids its \( b_i^0 \) as of iteration 2000. While these converged results are of interest, they do not necessarily represent what has happened throughout the simulation, especially during the initial iterations in which most of the learning takes place. Figs. 4(a) and (b) provide the average results over all 2000 iterations, again averaged over 30 runs. A comparison between Figs. 4 and 5 illustrate the effects of GenCo learning and competition over time.

The similarities in shapes indicate strong convergence in bid prices. The differences in bid prices point to changes in bidding behavior over time. In particular, the effect of risk aversion on \( b_i \) becomes sharper over iterations. GenCo-3 bids higher prices than 9, and GenCo-4 bids lower prices than 45 throughout the iterations. Accordingly, GenCos’ profits converge to more extreme levels at the end of the simulation. For \( \beta < 0.74 \), the competing GenCos, GenCo-2 and GenCo-3, achieve higher profits at the end of the simulation than they do in the initial iterations. For higher \( \beta \) values, however, extreme risk aversion of GenCo-2 causes a reduction in both GenCos’ profits. Meanwhile, as expected, GenCo-4’s profits converge to zero over iterations. All these observations underscore the importance of risk aversion on GenCos’ bidding behavior and resulting profit levels in an environment shaped by dynamic learning and competition.

Fig. 6 presents the corresponding DC-OPF optimal value, \( \sum b_i P_i \) and the total payment to GenCos, \( \sum (LMP_i \times P_i) \). Comparing the end-of-simulation and simulation-average results, we make the following two observations:

- **DC-OPF optimal value:** For almost all \( i \), the end-of-simulation DC-OPF optimal value is lower than the simulation-average value. Hence, the ISO’s auction mechanism seems to be successful in driving GenCos’ bid prices down throughout the simulation.

- **Total payment to GenCos:** For \( \beta < 0.45 \), the total payment to GenCos, hence, their total profit is higher at the end of the simulation than the simulation-average payment. When the GenCos are not very risk-averse, they collectively learn to obtain better profits over time. For \( \beta > 0.45 \), however, the observation is reversed; risk-averse behavior of GenCo-2 causes a reduction in total GenCo profits. This reduction is especially acute for \( \beta > 0.74 \).

![Fig. 3. Transmission grid in Case 3.](image)

![Fig. 4. End-of-simulation results. (a) The best identified bid prices. (b) Profits.](image)
This analysis sheds light onto the effect of risk aversion level on GenCo bids and profits. Overall, while some level of risk aversion can be beneficial to GenCos’ total profits, high levels of risk aversion is observed to degrade profits due to extreme price competition.

6.2. Differing risk aversion levels

Here, we analyze the effects of changes in the $\beta$ values of individual GenCos, focusing initially on GenCo-2. Fig. 7 presents the average profits and bid prices of each GenCo in a separate column, as a function of $\beta_2$ (in the y axis) and $\beta_3 = \beta_4$ (in the x axis) over the whole simulation. If $\beta_2$ increases, while keeping $\beta_3 = \beta_4$ constant, GenCo-2’s profit decreases. This is expected as this GenCo now bids lower prices. Interestingly, GenCo-3’s profit also decreases due to increased competition. GenCo-4’s average profit, too, is generally reduced. The only exception with GenCo-4 occurs for very high $\beta_2$ values, in which GenCo-2 bids its minimum price most frequently. In this case, GenCo-4 has a chance to make some profit only if $\beta_4$ is relatively low. Fig. 8 presents the end-of-simulation version of the same analysis.

Next, we investigate the effects of a simultaneous increase in $\beta_3 = \beta_4$, while keeping $\beta_2$ fixed, for example, at zero. When GenCos 3 and 4 become more risk-averse, they might be expected to reduce their bid prices, leading to a decrease in GenCo-2’s profit. Our simulation results, however, suggest the opposite. As $\beta_3$ and $\beta_4$ increase, we observe GenCo-2 to increase its bid price, leading to an increase in its profit. The key to understanding this counterintuitive result is GenCo-4’s behavior, who simply bids its highest price alternative 45. For this price, GenCo-4 is assigned no dispatch. If GenCo-4 bids one of the lower prices, there is a slight chance that it will be assigned some dispatch and earn some profit. When this happens, however, the variability in profit also increases which is not desirable from a risk-aversion point of view.

Fig. 9 illustrates the average profit and bid price of each GenCo-i (in a separate column) over the simulation as a function of its own $\beta_i$ (in the y axis) and the other GenCos’ $\beta$ values (in the x axis). The leftmost column is the same as that of Fig. 7. From the middle column, for example, we observe that rather than its own $\beta_3$, GenCo-3’s profit depends mostly on the risk aversion levels of the other GenCos, particularly that of GenCo-2. GenCo-3 sticks to the advantageous bid price of 9 unless $\beta_3$ is very low. Given this $\beta_3$, GenCo-3’s profit becomes a function of $\beta_2$, which decreases if $\beta_2$ increases. Note the emergence of $\beta = 0.38$ as a critical value again in this graph. GenCo-3 profits, in particular, are maximized when $\beta_2$ and $\beta_4$ are around 0.38. GenCo-4 makes a much lower profit compared to GenCos 2 and 3. GenCo-4’s profit is maximized when $\beta_4$ is at intermediate values, while the other two GenCos’ risk aversion levels are low, and consequently they do not engage in intense price competition.

Fig. 10 presents the end-of-simulation versions of these graphs. Compared to the simulation-average values, we observe GenCo-2 and GenCo-3’s profits to be higher. GenCo-4’s end-of-simulation profit, meanwhile, converges to zero independent of $\beta_4$.

6.3. Learning dynamics

Here, we drill further into the detailed workings of the learning model during the simulation. Fig. 11 presents how the three Gen-
Cos’ $Q$-values, hence the best identified bid prices, change over iterations for a given risk profile $(\beta_2, \beta_3, \beta_4)$. The graphs on left present the case of the risk profile $(0, 0, 0)$, corresponding to the bottom left corner of the relevant graph in Fig. 8. The graphs on right present the case of the risk profile $(0, 1, 1)$. Recall that our model ignores the effect of risk during the first half of the iterations; the risk model kicks in after iteration 1000.

When all GenCos are risk-neutral (Fig. 11(a)), we observe $\hat{b}_2$ and $\hat{b}_3$ to take some time to converge, due possibly to the tight competition between GenCos 2 and 3. Once the equilibrium between
these two GenCos with $b_1 = 20$ and $b_1 = 9$ is reached, GenCo-4’s bids become irrelevant as this GenCo is driven out of the market.

When GenCo-2 is risk-neutral but GenCos 3 and 4 are extremely risk-averse (Fig. 11(b)), $b_2$ changes from 20 to 30 once risk aversion kicks in at iteration 1000. For GenCo-3, the price 9 arises as $b_3$. Bidding 9 brings in a decent profit to GenCo-3 while not having the profit variability disadvantage of the higher bid prices.

This discussion illustrates how the learning, risk aversion and competition components of our model interact with each other.

7. Effects of the Q-learning parameters

Here, we study the effects of the initial values, $\epsilon_0$ and $\alpha_0$, of the time-decaying Q-learning parameters, $\epsilon_k$ and $\alpha_k$, on GenCo profits. A comprehensive simulation study is conducted using the network structure of Case 3. In all simulation runs 2000 iterations are conducted.

We report the results from the perspective of one GenCo at a time (the GenCo-$i$), which is assumed to be risk neutral. For this GenCo, we consider $21 \times 21 = 441$ parameter combinations of $(\epsilon_0, \alpha_0)$, in which each of the two parameters range between 0 and 1 with an increment size of 0.05.

For a given $(\epsilon_0, \alpha_0)$ combination, we speak of different scenarios characterizing the parameters of the other two GenCos (GenCo-$k$ where $k \neq i$) which are chosen from the following sets: $\epsilon_0 \in \{0.2, 0.4, 0.8\}$, $\alpha_0 \in \{0.2, 0.4, 0.8\}$, $\beta_0 \in \{0.0, 0.4, 0.8\}$. In each of the $3^3 = 27$ considered scenarios, the same stream of random numbers is used and the results are averaged over 10 runs. GenCo-$i$ is assumed to have no information about the parameters of the other GenCos; hence, it believes all scenarios to be equally likely. Consequently, for each $(\epsilon_0, \alpha_0)$ combination of GenCo-$i$, the average Q-value ($\overline{Q}$) and the average cumulative profit ($\overline{CP}$) over all 27 scenarios are reported. All in all, this comprehensive simulation study required the DC-OPF problem to be solved 19,289,340,000 times (3 GenCo-$i$ × 441 combinations × 2729 scenarios).
Fig. 12 presents the results (averaged over 729 scenarios) for each GenCo’s expected profit, that is, the Q-value of the best identified bid at the end of the simulation, whereas those in the second row illustrate the cumulative profit (CP) throughout the simulation. From the figure, we observe $x_0$ not to have a major impact on profit results unless its value is very low. Thus, the profits are robust to the initial value of the recency rate as long as some updating of Q-values occur. The exploration parameter $\epsilon_0$, on the other hand, is seen to have a significant impact on profits. The direction of this impact, however, is ambiguous. For GenCo-2, high $\epsilon_20$ lead to better profits. For GenCo-3, this is true for the end-of-simulation profit, yet the cumulative profit first increases then decreases with $\epsilon_0$. For GenCo-4, profit is uniformly increasing in $\epsilon_20$. Recall that GenCo-4 is at a disadvantageous position compared to the other GenCos. As suggested in Fig. 12, this GenCo can maximize its expected profit by acting as randomly as possible (corresponding to $\epsilon_0 = 1$), thereby disrupting the learning of the other two GenCos.

Next, we compare the end-of-simulation and cumulative profit values. For GenCos 2 and 4, $(\epsilon_0, x_0)$ combinations that yield the highest expected profit at the end of the simulation (the first row graphs) are also observed to provide the highest cumulative profit (the second row graphs). For GenCo-3, on the other hand, we observe significant differences. This GenCo can identify better profit opportunities at the end of the simulation by exploring excessively, however, this comes at the cost of achieving a lower cumulative profit. Overall, the only parameter that has a major profit impact at the end of the simulation turn out to be $\epsilon_20$. High values of this parameter is observed to increase the expected profit of GenCo-2 significantly.

8. Discussion

The academic literature on electricity markets is somewhat overly concerned with theoretical issues such as convergence to Nash equilibria [61]. Market participants, meanwhile, need studies that address the profits and risks associated with realistic GenCo bidding strategies. In fact, one of the research suggestions in Dahlgren et al. [31]’s survey on risk assessment in energy trading is “risk assessment would be more accurate if the bidding behaviors of market players can be modeled”. This is what we do in the present study by considering the effects of two behavioral factors on GenCo bidding and market evolution: learning from experience, and risk aversion. Our findings imply that one should be cautious in using static models to investigate dynamic markets such as the day-ahead electricity market. This is because these models fail to capture the dynamics of the interaction between competing GenCos that learn from experience.

We present the first ABMS study to analyze the joint effects of learning and risk aversion, leading to interesting observations. In particular, different from the literature, we obtain non-monotonous results concerning the effect of risk aversion on profits. As Liu and Wu [57] mention, most studies in GenCo risk management literature (e.g., [46,62,63]) neglect the market dynamics by considering known (or fixed) probabilistic distributions for the price of electricity, demand, or rivals’ behavior. Consequently, these models generally find a monotonous decrease in profit and the taken risk as GenCos become more risk averse. In practice, such results may not hold true since the interaction of learning GenCos can affect the distribution of the aforementioned factors through time [50,51].

Our results have important regulatory and managerial implications. Most importantly, simulation models such as our model can be used in the development of testbeds that are tailored to the learning behavior and risk aversion levels of GenCos in a particular electricity market. For instance, a large and established GenCo with a strong financial status would be modeled as less risk averse than a small GenCo. Risk aversion levels can be modeled as time-dependent, to capture the changes in risk attitude due to changing economic and financial conditions. Likewise, the learning behavior of GenCos would reflect the overall firm strategy, experience in the market and capabilities of the firm’s human resources. For example, a conservative GenCo would be modeled with a smaller $\epsilon$ parameter value compared to a GenCo that is more open to trying alternative bid prices.

ISOs may use the aforementioned testbeds to develop market rules that manipulate GenCo behavior in directions that provide higher social welfare. Such testbeds would allow an ISO to study the likely impacts of market rule changes prior to costly real market implementation. These studies would be particularly important in the prevention of tacit collusion among GenCos that can arise from GenCo learning [29,64,65]. GenCos, too, would benefit from testbeds, in formulating bidding policies that consider their own as well as their competitors’ bidding behavior.

9. Conclusions

This study analyzes how learning dynamics and risk aversion shape GenCo bidding behavior in a competitive electricity market, using an agent-based simulation model. GenCos are modeled as agents that bid prices repeatedly for each hour of the day-ahead market. Learning is modeled through a modified Q-learning algorithm, and risk aversion is captured as aversion to variability in profits. Given GenCos’ bids, to determine locational marginal prices and GenCo power dispatches, the ISO solves a DC-OPF problem that considers the physical network characteristics.

First, considering risk-neutral GenCos, the mechanics of our learning algorithm is illustrated on two simple case studies. The simulation runs achieve convergence thanks to time-decaying Q-learning parameters. In the case of a unique Nash equilibrium, the simulation easily converges to the equilibrium. In the presence of multiple equilibria, however, simulation runs converge to either one of the Nash equilibria, or a state that provides identical profits to a Nash equilibrium. Thus, the individual learning of GenCos is observed to drive the market into a reasonable outcome.

When the model is extended to consider risk-averse GenCos, the results show that some level of risk aversion may indeed be beneficial for GenCo total profits compared to the risk-neutral case. On the other hand, high levels of risk aversion are observed to intensify price competition and degrade profits. The findings illustrate how altering the risk aversion level of even one GenCo can trigger changes in the bidding behavior and profit levels of all GenCos due to learning and market interaction.

This study can be extended in numerous directions. First and foremost, one could use other measures of risk such as CVaR. A different learning model might be used, or the Q-learning algorithm employed in this study may be extended to achieve better performance. For instance, a GenCo may improve its profit by considering changes in the Q-values in addition to the Q-values themselves. Definitely, the results may depend on the specific risk aversion or learning method being used, which calls for follow-up studies.

References
