Approximating Nash Equilibrium in Day-ahead Electricity Market Bidding with Multi-agent Deep Reinforcement Learning

Yan Du, *Student Member*, *IEEE*, Fangxing Li, *Fellow*, *IEEE*, Helia Zandi, *Member*, *IEEE*, and Yaosuo Xue, *Senior Member*, *IEEE*

***Abstract*—In this paper, a day-ahead electricity market bidding problem with multiple strategic generation company (GENCO) bidders is studied. The problem is formulated as a Markov game model, where GENCO bidders interact with each other to develop their optimal day-ahead bidding strategies. Considering unobservable information in the problem, a model-free and data-driven approach, known as multi-agent deep deterministic policy gradient (MADDPG), is applied for approximating the Nash equilibrium (NE) in the above Markov game. The MADDPG algorithm has the advantage of generalization due to the automatic feature extraction ability of the deep neural networks. The algorithm is tested on an IEEE 30-bus system with three competitive GENCO bidders in both an uncongested case and a congested case. Comparisons with a truthful bidding strategy and state-of-the-art deep reinforcement learning methods including deep *Q* network and deep deterministic policy gradient (DDPG) demonstrate that the applied MADDPG algorithm can find a superior bidding strategy for all the market participants with increased profit gains. In addition, the comparison with a conventional-model-based method shows that the MADDPG algorithm has higher computational efficiency, which is feasible for real-world applications.**

***Index Terms*—Bidding strategy, day-ahead electricity market, deep reinforcement learning, Markov game, multi-agent deterministic policy gradient (MADDPG), Nash equilibrium (NE).**

I. Introduction

THE recent success of artificial intelligence (AI) computer program AlphaGo has brought the deep reinforcement learning (RL) and deep RL into the spotlight [1], [2]. Currently, it has been applied to a wide range of complex real-world multi-stage decision optimization scenarios including video game play and self-driving. In power systems, the strategic electricity market bidding problem of generation companies (GENCOs) belongs to the category [3]. The topic remains interesting ever since the deregulation of electricity markets. The problem is investigable because the deregulation introduces the competition among GENCOs and makes them become strategic price-makers that directly exert impacts on market clearing results.

The main goal of investigating the strategic electricity market bidding problem is to find the Nash equilibrium (NE), where no player in the market will benefit from a unilateral deviation from its current bidding strategy with the strategies of other players unchanged. In such a case, it is assumed that each player has offered its best responses to a fixed set of its rivals' strategies. In literature, NE in a strategic environment is the research focus. The most widely applied method for achieving NE is to formulate a bi-level mathematical problem with equilibrium constraints (MPEC), where the upper-level problem maximizes the profit of each strategic GENCO, and the lower-level problem solves the market clearing. The simultaneous optimization of multiple MPECs formulates an equilibrium problem with equilibrium constraints (EPEC), and it can be solved via the diagonalization algorithm [4]-[6]. Considering the uncertainty related to hydropower, [7] discusses two types of NE when solving the EPEC model. They are the Bayesian NE which focuses on maximizing the expected payoff in different scenarios and the robust NE which focuses on maximizing the worst-case payoff. In [8], the unit commitment is included in the optimization model of GENCO bidding strategies, which introduces additional binary variables and fails the conventional primal-dual approaches. A selective branch-and-cutting algorithm is designed instead of the optimization of the generator scheduling. In [9], the market equilibria are studied in a coupled electricity and natural gas market, where several objective functions are designed for the EPEC model to achieve NE solutions with different characteristics, e. g., maximizing the profits of the producer or maximizing the welfare of the consumer.

All the above work can be categorized as game-theoretic modelling methods, where the strategic bidding behaviors of the players are explicitly expressed by mathematical equations. While the model-based methods have exhibited soundness and success in achieving NE. It is worth noting that the

Manuscript received: July 21, 2020; accepted: October 27, 2020. Date of CrossCheck: October 27, 2020. Date of online publication: April 19, 2021.

This work was supported in part by the US Department of Energy (DOE), Office of Electricity and Office of Energy Efficiency and Renewable Energy under contract DE-AC05-00OR22725, in part by CURENT, an Engineering Research Center funded by US National Science Foundation (NSF) and DOE under NSF award EEC-1041877, and in part by NSF award ECCS-1809458.

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Y. Du and F. Li (corresponding author) are with University of Tennessee, Knoxville, USA (e-mail: ydu15@vols.utk.edu; fli6@utk.edu).

H. Zandi and Y. Xue are with Oak Ridge National Laboratory, Oak Ridge, USA (e-mail: zandih@ornl.gov; xuey@ornl.gov).

DOI: 10.35833/MPCE.2020.000502

model-based methods mainly rely on the assumption that each GENCO bidder has the full knowledge of its rivals' bidding strategies as well as the market clearing algorithm, which renders a limiting assumption in reality. In addition, the game-theoretic model generally demonstrates nonlinearity and nonconvexity due to the existence of a large number of complementarity conditions, the scale of which can grow rapidly with increasing players and multi-period constraints. As a result, a model that is mostly close to the physical world can be highly complex and computationally intensive.

The data-driven RL method stands out as an efficient alternative to the conventional game-theoretic modelling method. The RL method requires no knowledge of the exact mathematical model, but it gradually learns a decision-making strategy through continuous interaction with the actual environment, which circumvents any modelling or prediction error. In addition, since the RL method relies on no complementarity constraints, it is free from the afore-mentioned computational challenge. Finally, the RL method holds the generalization characteristic after learning and can adapt to unseen instances where the game-theoretic model needs to be recalculated with even the slightest changes of the model.

Existing researches have witnessed the application of *Q*-learning, which is a type of value-based RL method for solving the GENCO optimal bidding problem [10] - [13]. The main idea of *Q*-learning is to construct a look-up table that stores the action-values of all the state-action pairs. Thus, the actions with highest action-values will constitute a desired bidding strategy. The action-values are obtained through iterative interactions with the market. While the *Q*-learning does not suffer from modelling complexity, it does require the discretization of both the state and action spaces, which brings in the curse of dimensionality, especially with multi-dimensional continuous state or action. In such cases, the exponential growth of the look-up table will leave the problem intractable. Furthermore, the discretization of continuous variables also restrains the search space and may lead to sub-optimal solutions.

Popularized by the AI computer program AlphaGo, an advanced variant of RL, the deep RL, has recently become an eye-catching technique for solving time-sequential decision-making problem with partial or hidden information. Deep RL is a combination of a deep neural network (DNN) and RL. If compared with the conventional *Q*-learning, the highlight of the deep RL method is that it builds a generalized mapping between the state inputs and the action values with the universal function approximation property of DNN, instead of storing a concrete look-up table. Therefore, the deep RL method can be applied to continuous environment settings. The success of deep RL has been spotted in the fields of computer games, robotics, industrial automation, etc. In power systems, the potentials of implementing deep RL for demand-side energy management and electric vehicle charging/discharging scheduling are shown in [14], [15]. The deep deterministic policy gradient (DDPG) algorithm is applied to solve the bidding problem of a load serving entity and GENCOs in [16], [17].

When there exist multiple decision-makers with interdependent interactions, the multi-agent deep RL (MADRL) can play an important role in optimizing the strategy of each individual agent. In [18] and [19], MADRL is leveraged for coordinating load frequency control and voltage regulation for multi-area power systems, respectively. References [20] and [21] explore the market equilibrium with multiple strategic GENCO bidders under both constrained and non-constrained networks using policy-based deep RL methods.

In [20] and [21], while each strategic GENCO bidder has its own deep RL agent, each RL agent only learns from its local observation. However, given that a Markov game is studied in a multi-agent environment, where each decision-maker constantly changes the bidding strategies, the local observation cannot capture the entire dynamics of the environment and the policies learnt from local observations can get stuck in local optimum. In [22], a centralized training and decentralized execution mechanism is designed for MADRL, where each RL agent can receive the state and action information of other agents during the training. The centralized training allows the RL agents to become more adaptive in a dynamic environment where every agent constantly changes its bidding strategy as the learning evolves. It is believed that a centralized training can lead to an improved learning performance by enveloping global information. This framework is acceptable because the global information is only needed during the training, e.g., from historical data. During the test, it is assumed that the RL agents have mastered the knowledge of its rivals, and the global information is no longer needed.

Motivated by the above observation, we also focus on finding NE in the day-ahead electricity market with multiple strategic GENCOs by using MADRL. The main contributions of the paper are summarized as follows.

1) A Markov game is formulated to describe the strategic day-ahead electricity market bidding process of multiple GENCOs as price makers. Each GENCO acts intelligently to maximize its own benefits with the consideration of bidding policies from other rivals.

2) A multi-agent deep deterministic policy gradient (MADDPG) algorithm is applied to solve the above Markov game. A centralized training and decentralized execution mechanism proposed in [22] is implemented. MADDPG can deal with non-stationary environments where the rational players constantly change their strategies as well as high-dimensional continuous state and action spaces.

3) An in-depth analysis of the RL policies from MADDPG algorithm is presented to provide insights into the logic and rationality behind the "black-box" learning. It is discovered that the learnt bidding strategies are explainable and sensible in reaching the NE status with different market scenarios, which supports their feasibility in the real-world applications.

4) A comprehensive simulation analysis is presented to prove that the well-trained MADDPG algorithm can approximate NE in unseen market environments with high computational efficiency. The superiority of the applied method is further verified by comparing with other state-of-the-art RL methods and the model-based method.

We aim to build a solid theoretical foundation for the prac-

tical implementation of the applied MADDPG algorithm in real-world electricity market bidding problems. The rest of the paper is organized as follows. Section II provides the formulation of the multi-agent day-ahead electricity market bidding problem in the context of Markov game. Section III introduces the MADDPG algorithm for multi-agent decision-making process within non-stationary environment. The simulation results and analysis are shown in Section IV. Finally, Section V concludes the paper.

II. Formulation of Multi-agent Market Bidding Problem

A. Brief on Electricity Market and Bidding Strategies

A deregulated electricity market is usually composed of two stages, a day-ahead market and a real-time market. In the day-ahead market, GENCOs submit the amount of energy they are willing to sell for the next 24 hours and the associated offer price. And consumers submit the amount of energy they are willing to buy and the associated bid price. The market operator clears the market by running an optimal power flow (OPF) calculation and releases the market clearing prices and quantities to the supply side and the demand side.

GENCOs with large capacity can execute market power through the following two ways: ① physical withholding, which means that they submit generation quantities that are less than their capacity; ② economic withholding, which means that they offer the prices that are higher than their marginal cost [23]. However, executing market power can also be risky since GENCO bidders have incomplete information of their rivals.

There exist two market settlement methods: ① the marginal price principle, where all suppliers receive the same market clearing price, which is the cost of the marginal bidding block; ② the pay-as-bid principle, where each winning supplier receives a price based on their respective bid prices, and it can be different from one to another. The marginal price principle is applied in almost all the organized wholesale markets in the United States [24], while the pay-as-bid principle is mostly adopted in European countries like France and Britain [25].

In term of auction theory, the pay-as-bid principle is a variation of the sealed first-price auction and marginal price is a variation of the sealed second-price auction [26]. It has been proven that in the sealed second-price auction, truthful bidding, which means no economic withholding, is a dominant strategy. While in the sealed first-price auction, truthful bidding is a necessary but not sufficient condition to reach NE [27].

We assume that the day-ahead electricity market is cleared based on the marginal price principle, and GENCOs can execute economic withholding to maximize their profits. Under such condition, an NE bidding strategy of the GENCO g should satisfy the following two conditions [27]: ① $\max_{g' \neq g} c_{g'} \geq b_g$ and ② $\max_{g' \neq g} b_{g'} \geq c_g$.

Under the above two conditions, index g' refers to all the other GENCO bidders except for GENCO g; c_g is the marginal generation cost of GENCO g; and b_g is the bid price of GENCO g. The profit of GENCO g will be $b_g - c_g$. Condition ① means that GENCO g should bid at a sufficiently low price to win the bid (lower than the marginal cost of all the other GENCO bidders). Otherwise, if there is one GENCO bidder g' with $c_{g'} < b_g$, then g' could bid between the open interval $(c_{g'}, b_g)$, causing GENCO g to lose the bid. Condition ② means that the marginal cost of GENCO g should be sufficiently low (lower than the bid price of all the other GENCO bidders) to make a profit. Otherwise, if there is one GENCO bidder g' with $b_{g'} < c_g$, since the marginal price principle is applied, the market clearing price received by GENCO g will also be $b_{g'}$, and its profit will be $b_{g'} - c_g$, which is negative. In Section IV, we will demonstrate that the applied deep RL method is able to achieve an NE strategy that corresponds with the above two conditions through truthful bidding.

B. Mathematical Formulation of Day-ahead Electricity Market Clearing

The day-ahead electricity market clearing model is shown as the following direct current (DC) OPF:

$$\min \sum_{t=1}^{N_T} \sum_{g=1}^{N_g} \sum_{b=1}^{N_b} \lambda_{g,b}^{bid}(t) P_{g,b}^{cleared}(t) \tag{1}$$

s.t.

$$\sum_{g=1}^{N_g} \sum_{b=1}^{N_b} P_{g,b}^{cleared}(t) = \sum_{d=1}^{N_d} P_d^{load}(t) \tag{2}$$

$$-limit_l \leq \sum_{i=1}^{n} GSF_{l-i} \left(\sum_{g \in i} \sum_{b=1}^{N_b} P_{g,b}^{cleared}(t) - \sum_{d \in i} P_d^{load}(t) \right) \leq limit_l \tag{3}$$

$$0 \leq P_{g,b}^{cleared}(t) \leq P_{g,b}^{bid}(t) \quad \forall g, \forall b \tag{4}$$

where N_T is the number of time intervals; N_g is the number of GENCOs; N_b is the number of bidding blocks submitted by the g^{th} GENCO; N_d is the number of loads; $\lambda_{g,b}^{bid}(t)$ and $P_{g,b}^{cleared}(t)$ are the bid price and the cleared quantity of the b^{th} bidding block, respectively; $limit_l$ is the line capacity of the l^{th} transmission line; n is the set of buses; GSF_{l-i} is the called generation shift factor, which represents the power flow change on the l^{th} transmission line if one unit power injection takes place at bus i; $P_d^{load}(t)$ is the quantity of the d^{th} load; and $P_{g,b}^{bid}$ is the bidding quantity of the b^{th} bidding block of the g^{th} GENCO. Equation (2) is the power balance constraint, where the total amount of cleared generation should equal the total amount of load; (3) is the transmission line capacity constraint; and (4) ensures that the cleared quantity does not exceed the bidding quantity submitted by the GENCO bidders.

In the DCOPF model (1)-(4), $\lambda_{g,b}^{bid}(t)$ and $P_{g,b}^{bid}(t)$ are known values submitted by GENCO bidders. The GENCO bidders decide the bid price $\lambda_{g,b}^{bid}(t)$ by solving the following profit-maximizing problem:

$$\sum_{t=1}^{N_T} \left(\sum_{b=1}^{N_b} P_{g,b}^{cleared}(t) \lambda_g^{cleared}(t) - C_g \sum_{b=1}^{N_b} P_{g,b}^{cleared}(t) \right) \tag{5}$$

s.t.

$$C_g \sum_{b=1}^{N_b} P_{g,b}^{cleared}(t) = \sum_{b=1}^{N_b} \lambda_{g,b}^{cost} P_{g,b}^{cleared}(t) \tag{6}$$

$$\lambda_{g,b}^{bid}(t) = \varepsilon_g(t) \lambda_{g,b}^{cost}(t) \tag{7}$$

$$1 \le \varepsilon_g(t) \le \varepsilon_{g,\max} \tag{8}$$

$$\sum_{b=1}^{N_b} P_{g,b}^{bid}(t) = P_g^{\max} \tag{9}$$

where $\lambda_g^{cleared}(t)$ is the locational marginal price (LMP) at the bus with the g^{th} generator connected; $\lambda_{g,b}^{cost}(t)$ is the marginal cost of the b^{th} bidding block; $\varepsilon_{g,\max}$ is the upper bound of the bidding factor; and $P_g^{\max}$ is the upper limit of the generation. In (5), the first item is the income of selling power at the day-ahead electricity market, and the second item is the generation cost, which is calculated by (6). Equation (7) indicates the economic withholding of the GENCO bidder. (8) shows the range of $\varepsilon_g(t)$, which indicates that the GENCO bidder can deliberately submit a higher marginal cost to increase its profit. Equation (9) is the capacity limit of the bidding block.

The above two mathematical models, i.e., model (5)-(9) and model (1)-(4), formulate a bi-level optimization problem. For the lower-level market clearing problem, the decision variable is the cleared quantity $P_{g,b}^{cleared}(t)$. At the upper-level profit-maximizing problem, the decision variable is the bidding factor $\varepsilon_g(t)$. For simplicity, we assume that only the bid price $\lambda_{g,b}^{bid}(t)$ is variable, and the capacity of the bidding block $P_{g,b}^{bid}(t)$ is constant. In literature, one of the most commonly used methods to solve the above bi-level optimization problem is to transform the lower-level market clearing model into its equivalent Karush-Kuhn-Tucker (KKT) conditions and add them as constraints to the upper-level profit-maximizing model, which formulates an MPEC. With multiple GENCO bidders, multiple MPECs will be formulated, and to obtain an equilibrium point among them is not a trivial task. In addition, solving MPECs requires the full knowledge of the market clearing process as well as the bidding information of the rival GENCO bidders, which remains as hidden information in real-world situations. Hence, the model-based method fails due to this unobservability. In the following sections, the multi-agent day-ahead electricity market bidding problem will be transformed to a Markov game, and a model-free deep RL method will be introduced as a solution method.

C. Markov Game Model of Day-ahead Electricity Market Bidding

Before building the Markov game model, we propose the following assumptions regarding the day-ahead electricity market bidding problem [28]:

1) GENCOs submit hourly bidding blocks for the next 24 hours in the day-ahead market. The bidding quantities are their true generation capacities, and only the bid price can be changed.

2) The bid price for the same bidding block can vary from hour to hour. However, the ratio of the highest bid price to the lowest bid price for the same bidding block should not exceed a threshold th_1.

3) For any two consecutive hours, the ratio of the bid prices for the same bidding block should not exceed a threshold th_2.

The reason for making assumption 3 is to avoid high fluctuation of marginal prices during peak hours and to prevent the GENCO bidders from speculation. From assumption 3, it can be discovered that the bid price for the current hour is related to the bid price in the previous hour, which leads to a finite Markov decision process (MDP) with discrete time steps. An MDP is composed of four essential elements (s, a, p, r), where s is the finite number of states; a is the finite number of actions; p is the state transition probability that falls within [0,1]; and r is the reward function.

In an MDP, at each time step, the agent firstly observes the environment state and takes an action. Then, the agent receives an immediate reward from the environment, and the environment transfers to the next state based on the transition probability. The process repeats until termination.

When multiple agents are considered in the day-ahead electricity market bidding, the above MDP is extended to a partially observable Markov game. A Markov game for N agents consists of a set of states s, a set of observations made by each agent at the current state, $o_1, o_2, \ldots, o_N$, and a set of actions $a_1, a_2, \ldots, a_N$ taken by each agent based on their respective observations. After the execution of the actions, the environment will transfer to the next state following a transition probability p: $s \times a_1 \times a_2 \times \ldots \times a_N \times s \to [0, 1]$. Each agent will receive a reward r_i: $s \times a_i \to R$ $(i = 1, 2, \ldots, N)$ and a private observation for the next state o_i: $s \to o_i$. The objective of each agent is to maximize the total discounted reward for the finite time steps: $R_i = \sum_{t=1}^{N_T} \gamma^{t-1} r_{i,t}$, where γ is a discount factor to convert future rewards to the present value.

In the day-ahead electricity market bidding problem, with the context of a Markov game, the agent is an independent GENCO bidder. The private observation for each GENCO is the demand quantity for the current hour and its bid price at the previous hour; the state is simply defined as the summation of the observations of all GENCOs; the action is the bid price for the current hour; and the reward is the hourly profit. The day-ahead electricity market bidding process is a sequential decision-making problem with multiple decision makers involved, which requires that each GENCO bidder is far-sighted enough to consider potential future outcomes in order to maximize the total profit.

Note that in the general day-ahead electricity market bidding, the GENCO bidders are required to submit their bidding blocks for the next 24 hours in one shot. While in the above Markov game, the bidding decision process is decomposed to discrete time steps and the bid price for each time step is decided sequentially. This decomposition is acceptable because at each time step, the private observation only includes the current hourly load and the bid price at the previous hour without involving any market clearing results. Hence, after the applied deep RL method is well-trained for achieving NE in the above Markov game model, it will only need the load data for the next 24 hours as the input and can generate the bid prices for the next day in one shot (with an initial bid

price) during the test process. Therefore, the algorithm can be physically implemented without violating market rules.

III. MADDPG Algorithm for Day-ahead Electricity Market Bidding

A. Overview of RL Method

The RL method aims to solve the MDP process and maximize the total discounted reward $R_i = \sum_{t=1}^{N_T} \gamma^{t-1} r_{i,t}$. An action-value function $Q_\pi(s_t,a_t)$ is further defined in RL as an estimation of the total discounted reward:

$$Q_\pi(s_t,a_t) = E_\pi\left(\sum_{k=0}^{N_T} \gamma^k r_{t+k+1} | s_t, a_t\right) \tag{10}$$

where E_π is the expectation. $Q_\pi(s_t,a_t)$ is equal to the expected return starting from state s_t, taking action a_t, and thereafter following policy π for a horizon with length N_T. The goal of RL is to find the optimal policy π^* that maximizes the action-value function:

$$Q^*(s_t,a_t) = \max_\pi Q_\pi(s_t,a_t) \tag{11}$$

One typical way for solving (11) is to update the action value based on the temporal difference (TD) error [29]:

$$Q_\pi^{(k+1)}(s_t,a_t) = r_t + \gamma \max_{a_{t+1}} Q_\pi^{(k)}(s_{t+1},a_{t+1}) \tag{12}$$

where k is the iteration index. In conventional RL methods such as Q-learning, a look-up table is established to store the action values of all the possible state-action pairs and it is updated iteratively according to (12) until convergence. However, the method encounters the curse of dimensionality when the state or action space becomes continuous. The deep RL method is developed to overcome the drawbacks of the tabular-based RL method. In deep RL, a neural network is designed to estimate the action-value function and it can form a continuous mapping between the state-action pair and the action value. In this way, more complex control or optimization problem with high dimensionality can be solved through tweaking the neural network model.

B. DDPG Algorithm for Continuous Control

In this subsection, we will briefly introduce a deep RL method and a DDPG to solve continuous control problems [30].

In DDPG, there are two types of neural networks: the critic network and the actor network [31]. The function of the critic network is to estimate the action value. The input to the critic network is the current state and the taken action, and the output is the associated action value. The mean square error (MSE) is used as the loss function for updating the parameters of the critic network, as shown as:

$$Q^{target(j)}(t) = r^{(j)}(t) + \gamma \max_{a^j(t+1)} Q(s^{(j)}(t+1), a^{(j)}(t+1); \theta^{Q'}) \tag{13}$$

$$L(\theta^Q) = \frac{1}{N_s}\sum_{j=1}^{N_s} (Q^{target(j)}(t) - Q(s^{(j)}(t), a^{(j)}(t); \theta^Q))^2 \tag{14}$$

Two critic networks are involved for calculating the MSE in (14). In the target critic network, the weights are noted as $\theta^{Q'}$. In the behavior critic network, the weights are noted as θ^Q. In (13), the target action value at time step t is the sum of the current reward $r(t)$ and the discounted value of the maximum action value at the next time step $t+1$, generated by the target critic network. The superscript j is the index of state-action pair samples. Then, the target action value is sent to (14) for calculating the loss. The output from the target critic network is served as the "labelled" data for the behavior network to learn. During the training, the target critic network is updated at a slower speed than the behavior critic network, which helps stabilize the learning process.

The actor network is designed to utilize the estimated action value to obtain the optimal policy, i. e., $\pi(s^{(j)}(t)) = \arg\max_{a(t)} Q(s^{(j)}(t), a^j(t))$ for all time step t. The input to the actor network is the current state $s^{(j)}(t)$, and the output is the action $a^{(j)}(t)$ that results in the maximum $Q(s^{(j)}(t), a^j(t))$. To achieve this goal, the loss function for the actor network is designed as:

$$\max J(\theta^\mu) = \frac{1}{N_s}\sum_{j=1}^{N_s} Q(s^{(j)}(t), a^j(t), \theta^Q)\Big|_{a^j(t)=\mu(s^{(j)}(t),\theta^\mu)} \tag{15}$$

where $\mu(s^{(j)}(t), \theta^\mu)$ is the current policy generated by the actor network and θ^μ is the network weights. θ^μ is updated in the direction of maximizing the Q value using the gradient:

$$\nabla J(\theta^\mu) = \frac{1}{N_s}\sum_{j=1}^{N_s} \nabla Q_\mu(s^{(j)}(t), a^j(t); \theta^Q) \nabla_{\theta^\mu} \mu(s^{(j)}(t); \theta^\mu) \tag{16}$$

$$\theta^\mu = \theta^\mu - \eta_\mu \nabla J(\theta^\mu) \tag{17}$$

where η_μ is the learning rate.

In (16), the chain rule is applied to calculate the gradient of the action value to the weights of the actor network.

The above introduction covers the basic idea behind the DDPG algorithm. Note that in the above actor and critic networks, only the action and the Q value at the current state are generated, and there is no need to store all the possible state-action pairs and their action values. The relationship between the state-action pair and the action value is encoded in the weights of the neural network. Therefore, DDPG can be applied to optimize continuous control strategies without suffering from the dimensionality explosion.

C. MADDPG for Solving Markov Game in Day-ahead Electricity Market Bidding

DDPG algorithm can be applied to optimize the single agent decision-making process. However, in the case of day-ahead electricity electricity market bidding, where multiple strategic GENCO bidders are involved, directly applying the above DDPG algorithm since each GENCO cannot achieve the ideal results. This is because when multiple agents are optimizing their decisions simultaneously, the environment becomes dynamic. And the reward received at the same state with the same action can constantly change due to the changing policies of other agents, which invalidates the experience learnt by the target critic network and can result in incorrect settings of target value and algorithm divergence.

Driven by the above concerns, there are some recent research works in AI that include multi-agent (MA) learning

as an extension of the DDPG algorithm to form MADDPG [22]. The main idea of MADDPG is to implement a centralized training, where the input to the critic network includes not only the observation and action of the current agent, but also the observations and actions of other agents. This assumption is acceptable because the critic network is only required during the training process. Once the algorithm is well-trained, only the actor network is needed to be tested in new environments, and the information of other agents are no longer required.

In this paper, the general-purposed MADDPG algorithm is applied and customized to solve the Markov game in day-ahead electricity market bidding. The proposed customized MADPPG algorithm flow is shown in Algorithm 1.

Algorithm 1: MADDPG algorithm for day-ahead electricity market bidding with N GENCO bidders

1: Initialize the parameters of the critic network $Q(s,a;\theta^Q)$ and actor network $\mu(o;\theta^\mu)$ for each GENCO bidder
2: Initialize the target networks with θ^Q and θ^μ
3: **for** episode is 1 to M **do**
4: Initialize the electricity market bidding from a random day
5: **for** $t=1$ to N_T **do**
6: Observe the current state $s(t)=[P_{load}(t),\lambda_1^{bid}(t-1),\lambda_2^{bid}(t-1),\ldots,\lambda_g^{bid}(t-1),\ldots,\lambda_N^{bid}(t-1)]$
7: For each GENCO bidder g, select the bid price $\lambda_g^{bid}(t)=\mu_g(o_g(t);\theta^\mu)$, where $o_g(t)=[P_{load}(t),\lambda_g^{bid}(t-1)]$
8: Run DCOPF (1)-(5) to complete the market clearing, obtain the cleared quantity $P_g^{cleared}(t)$, cleared price $\lambda_g^{cleared}(t)$, and the reward $r_g(t)$ for each GENCO, and observe the next state $[P_{load}(t+1),\lambda_g^{bid}(t)]$
9: Store the transition $(s(t),\lambda_g^{bid}(t),\lambda_g^{bid-}(t),r_g(t),s(t+1))$ for each GENCO
10: **for** GENCO $g=1$ to N **do**
11: Randomly sample a minibatch of S samples $(s^{(j)}(t),\lambda_g^{bid(j)}(t),\lambda_g^{bid-(j)}(t),r_g^{(j)}(t),s^{(j)}(t+1))$ from the stored transitions
12: Set $Q_g^{target(j)}(t)=r_g^{(j)}(t)+\gamma Q_g(s^{(j)}(t+1),\lambda_g^{bid(j)}(t+1),\lambda_g^{bid-(j)}(t+1);\theta^{Q'})$, for $\lambda_g^{bid(j)}(t+1)=\mu_g(o_g^{(j)}(t+1);\theta^{\mu'})$
13: Update the critic network by minimizing MSE:
14: $L_g(\theta^Q)=1/N_s\sum_j(Q_g^{target(j)}(t)-Q_g(s^{(j)}(t),\lambda_g^{bid}(t),\lambda_g^{bid-}(t);\theta^Q)$
15: $\theta^Q=\theta^Q-\eta_Q\nabla\theta^Q L_g(\theta^Q)$
16: Update the actor network by maximizing the expected Q value:
17: $\nabla J_g(\theta^\mu)=1/N_s\sum_j\nabla\mu_g Q_g(s^{(j)}(t),\lambda_g^{bid}(t),\lambda_g^{bid-}(t);\theta^Q)\nabla\theta^\mu\mu_g(o_g^{(j)}(t);\theta^\mu)$
18: $\theta^\mu=\theta^\mu-\eta^\mu\nabla J_g(\theta^\mu)$
19: **end for**
20: Update the target network parameters for each GENCO:
21: $\theta^{Q'}=(1-\tau)\theta^Q+\tau\theta^{Q'}$
22: $\theta^{\mu'}=(1-\tau)\theta^\mu+\tau\theta^{\mu'}$
23: **end for**
24: **end for**

In Algorithm 1, the state is defined as the hourly load and the bid prices of all the agents in the previous hour. The private observation of each agent is defined as the hourly load, and its bid price at the previous hour is shown by lines 6 and 7. For simplicity, we use $\lambda_g^{bid}(t)$ to represent the bid price and ignore the subscript b. The reward $r_g(t)$ is the hourly power selling profit:

$$r_g(t)=\sum_{b=1}^{N_b}P_{g,b}^{cleared}(t)\lambda_g^{cleared}(t)-C_g\sum_{b=1}^{N_b}P_{g,b}^{cleared}(t) \tag{18}$$

The state is sent to the critic network to calculate the target action value, as shown by line 12. $\lambda_g^{bid-(j)}(t+1)$ represents the bid prices of all GENCOs except for the g^{th} GENCO. Note that the bid price for the next time step $t+1$ is generated by $\mu_g(\theta^{\mu'})$ instead of $\mu_g(\theta^\mu)$ in line 12. Similar to the target critic network, $\mu_g(\theta^{\mu'})$ represents the target actor network, which also aims to stabilize the training process.

After the weights of the behavior critic network and the behavior actor network are updated as shown by line 15 and 18, the weights of the target critic network and the target actor network are updated accordingly at a slower speed as shown by lines 21 and 22, where τ has a value close to 1. The reason for this slow update is also to increase the stability of the learning.

To obtain an easy and clear understanding of the MADDPG algorithm for an MA day-ahead electricity market bidding problem, an illustration of the algorithm is shown in Fig. 1.

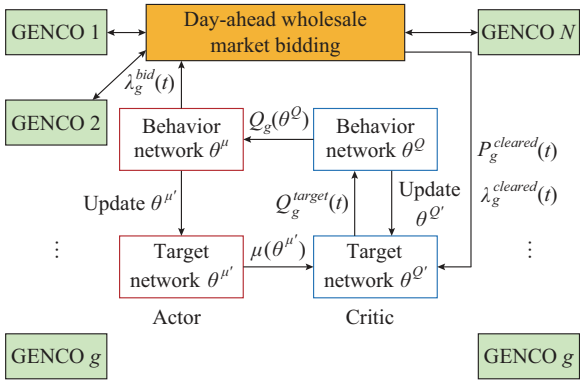

Fig. 1. MADDPG algorithm for solving Markov game in day-ahead electricity electricity market bidding.

IV. Simulation Analysis

A. Test System Description

IEEE 30-bus system with 9 generators is applied as the transmission-level electricity market. The topology of the system is shown in Fig. 2.

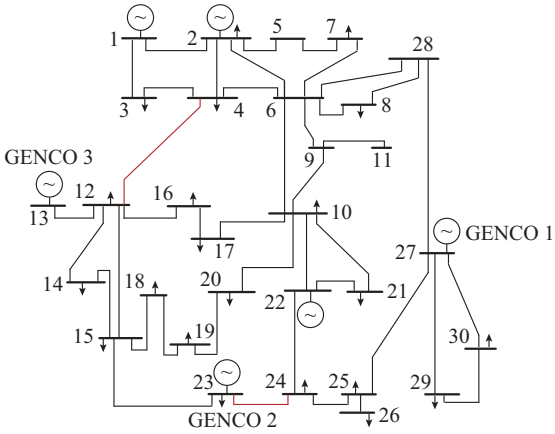

Fig. 2. Topology of IEEE 30-bus system.

The generators at bus 27, bus 23, and bus 13 are considered as strategic bidders that will conduct economic withholding to maximize their profits. All other generators will submit their true marginal cost. In addition, the transmission lines 4-12 and 23-24 have a capacity limit of 10 MW. Therefore, nearby GENCOs, GENCO 2 and GENCO 3, will be given the market power to manipulate the clearing price, which will be shown in the simulation results.

The generation cost function of GENCOs is assumed to be a piecewise linear function, which includes three segments. The parameters of the cost function are shown in Table I.

TABLE I
GENCO GENERATION COST FUNCTION

Segment	Marginal price ($/MWh)	Generation range (MW)
1	20	0-12
2	40	12-36
3	50	36-60

It is assumed that at each hour, only one bidding block is submitted by each GENCO. The bidding quantity is 60 MW, which is also their capacity. For GENCOs 1-3, the bid price is $\lambda_g^{bid}(t)=\varepsilon_g(t)\lambda_g^{cost}(t)$. For other GENCOs, the bid price is $\lambda_g^{cost}(t)$. In this case, $\lambda_g^{cost}(t)$ is 50 $/MWh.

Following the assumptions presented at Section II-C, the values of the bid price thresholds, th_1 and th_2, are set to be 1.5 and 1.1, respectively, which means $\lambda_g^{bid}(t)$ should comply with the following condition:

$$\begin{cases}\dfrac{\max \lambda_g^{bid}(t)}{\min \lambda_g^{bid}(t)} \le 1.5 \\ 0.9 \le \dfrac{\lambda_g^{bid}(t)}{\lambda_g^{bid}(t-1)} \le 1.1\end{cases} \quad \forall t, g=1,2,3 \tag{19}$$

B. Design of Neural Network and Simulation Platform

The detailed structures of the actor network and critic network in the proposed MADDPG as well as the structure of the deep Q network (DQN) are shown in Table II.

TABLE II
DESIGN OF DNNS IN DEEP RL

Neural network	Actor	Critic	DQN
Input	$[P_{load}(t), \lambda_g^{bid}(t-1)]$	$[P_{load}(t), \lambda_g^{bid}(t-1), \lambda_g^{bid-}(t-1), \lambda_g^{bid}(t), \lambda_g^{bid-}(t)]$	$[P_{load}(t), \lambda_g^{bid}(t-1)]$
No. of hidden layers	2	2	2
No. of neurons	[2, 64], [64, 64]	[7, 64], [64, 64]	[2, 64]
Output	$\delta \in [0, 1]$	$Q_g(s(t), \lambda_g^{bid}(t), \lambda_g^{bid-}(t)$	$Q_g(s(t), \lambda_g^{bid}(t))$
Activation function	ReLU (hidden layer); sigmoid (output layer)	ReLU (hidden layer)	ReLU (hidden layer)
Learning η	0.001	0.002	0.001
τ	0.99	0.99	0.99
Optimizer	Adam	Adam	Adam

The output from the actor network is a value δ between 0 and 1. The bid price $\lambda_g^{bid}(t)$ is calculated as:

$$\lambda_g^{bid}(t)=(0.9+0.2\delta)\lambda_g^{bid}(t-1) \tag{20}$$

The value of $\lambda_g^{bid}(t)$ will be further adjusted to be within the range of $1\lambda_g^{cost}(t)$ to $1.5\lambda_g^{cost}(t)$:

$$\begin{cases}\lambda_g^{bid}(t)=1.5\lambda_g^{cost}(t) & \lambda_g^{bid}(t)>1.5\lambda_g^{cost}(t) \\ \lambda_g^{bid}(t)=\lambda_g^{cost}(t) & \lambda_g^{bid}(t)<\lambda_g^{cost}(t) \\ \delta=(\lambda_g^{bid}(t)/\lambda_g^{bid}(t-1)-0.9)/0.2\end{cases} \tag{21}$$

The neural network model is built and trained by the open-source deep learning platform TensorFlow. The day-ahead market clearing process is completed by the smart market module in MATPOWER toolbox [32]. The hardware environment is a laptop with Intel®Core™ i7-7600U 2.8 GHz CPU, and 16 GB RAM.

C. NE Strategy from MADDPG: Uncongested Case

In this subsection, we first investigate the bidding strategies of the three GENCOs with marginal pricing mechanism in uncongested case, where the capacity limits on lines 4-12 and 23-24 are removed. The goal is to show that MADDPG algorithm can achieve the NE strategies that satisfy the two conditions as introduced in Section II-A, when no GENCO bidder has access to market power.

The load profile in June, 2019 from PJM wholesale market [33] is used to train the deep RL. The load data in the 31 days in July, 2019 from PJM market is used to test the deep RL after training. The load profiles of the training days and the test days are shown in Fig. 3.

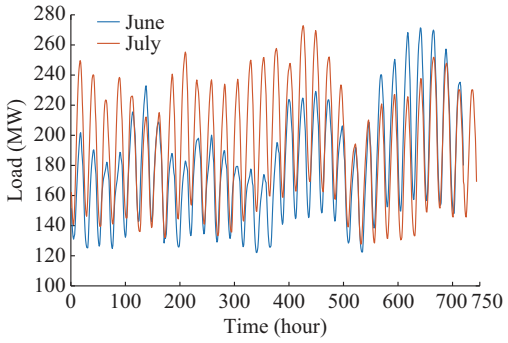

Fig. 3. Load profiles of training days and test days in June and July of year 2019.

There exist differences between the load levels in the two months. However, since the deep RL is a generalized model, it can adapt to the changes in the environment and produces optimal strategies, which will be shown in the following test results.

Figure 4 presents the convergence of MADDPG in uncongested case over 500 training episodes. The average rewards for the three GENCOs gradually stabilize as the training proceeds, which indicates the convergence of the MADDPG algorithm.

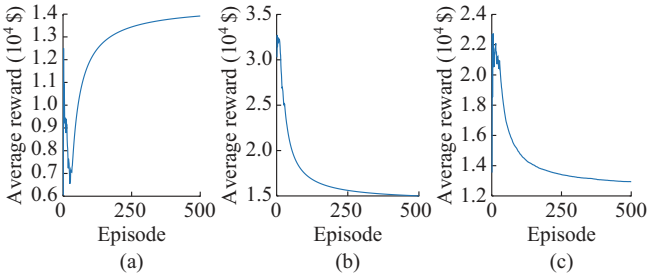

Fig. 4. Convergence of MADDPG in uncongested case. (a) GENCO 1. (b) GENCO 2. (c) GENCO 3.

The test results of MADDPG with the data of July are shown in Fig. 5 and are also compared with the truthful bidding baseline case in Table III, where the three GENCOs always bid at their true marginal cost.

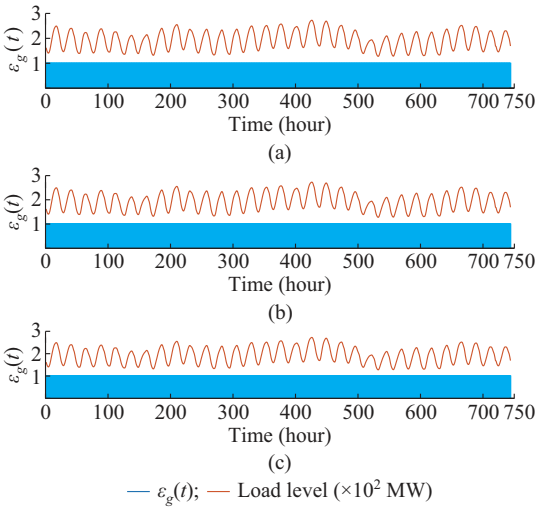

Fig. 5. Bidding strategy of three GENCOs with MADDPG in uncongested case. (a) GENCO 1. (b) GENCO 2. (c) GENCO 3.

TABLE III
COMPARISON OF MADDPG WITH BASELINE IN UNCONGESTED CASE

GENCO	Total profit (10^4 $)	
	MADDPG	Truthful bidding
1	44.64	44.64
2	44.64	44.64
3	40.26	40.26

In Fig. 5, y axis is the bidding parameter $\varepsilon_g(t)$ in (7). One thing should be pointed out is that since the state of GENCO bidder requires the bid price at the previous hour, we assume that the bid price at hour zero is always the true marginal cost. It can be observed that in the uncongested case, all GENCOs bid at their true marginal cost, regardless of the system load level. This is a logical and explainable behavior because when the capacity limit is removed, GENCOs 2 and 3 cannot manipulate the market clearing price. Since all other GENCOs are bidding at their true marginal cost, the optimal bidding strategy for GENCOs 1-3 is also truthful. According to the NE conditions in Section II-A, c_g and b_g are $\lambda_g^{cost}(t)$ and $\lambda_g^{bid}(t)$, respectively. In the truthful bidding, $\lambda_g^{bid}(t)=\lambda_g^{cost}(t)$ satisfies the equality constraint in conditions 1 and 2. In Table III, the total profit from MADDPG is the same as the baseline because they both bid truthfully. Therefore, it can be safely concluded that the well-trained MADDPG algorithm can find the optimal bidding strategy of the three GENCO bidders in a constraint-free market environment.

D. Solving Markov Game with MADDPG: Congested Case

In this subsection, the MADDPG algorithm is applied to solve the Markov Game in day-ahead electricity electricity market bidding with congestions. The same training data are used and the training results are shown in Fig. 6.

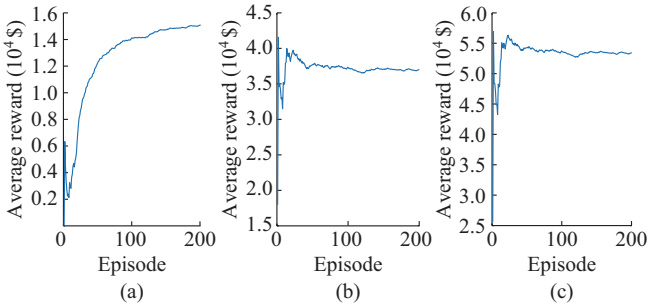

Fig. 6. Convergence of MADDPG in congested case. (a) GENCO 1. (b) GENCO 2. (c) GENCO 3.

The algorithm is trained for 200 episodes. The average reward converges for all three GENCO bidders. The well-trained RL agents are then tested with the data of July, and are also compared with the truthful bidding case, as shown in Figs. 7 and 8, and Table IV.

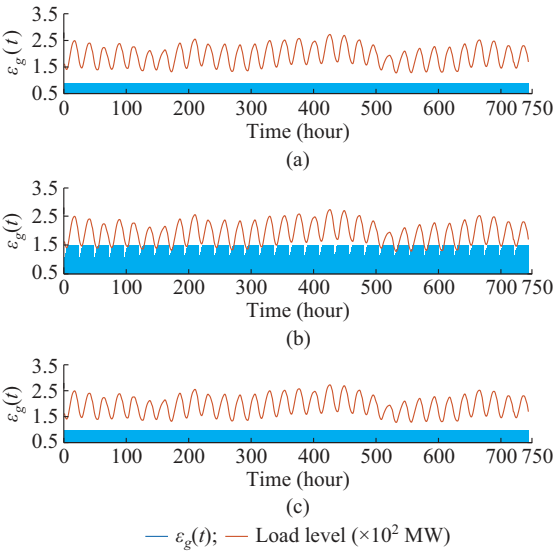

Fig. 7. Bidding strategy of three GENCOs with MADDPG in congested case. (a) GENCO 1. (b) GENCO 2. (c) GENCO 3.

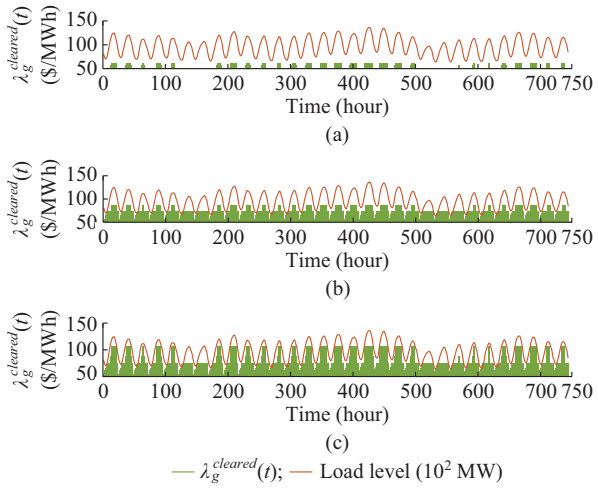

Fig. 8. Hourly market clearing price with MADDPG bidding strategy. (a) GENCO 1. (b) GENCO 2. (c) GENCO 3.

TABLE IV
COMPARISON OF MADDPG WITH BASELINES: CONGESTED CASE

GENCO	Total profit (10^4 $)		Profit increase (%)
	MADDPG	Truthful bidding	
1	57.24	50.01	14.45711
2	131.36	88.96	47.66187
3	184.37	114.34	61.24716

Figure 7 shows that GENCO 2 always bids at the highest price, which is $1.5\lambda_g^{cost}(t)$ at the peak hours when congestions are mostly likely to occur. Since GENCO 1 does not have market power, it always bids at the true marginal cost, where $\varepsilon_g(t)$ is 1. GENCO 3 also bids at the true marginal cost. However, because the principle of marginal price is applied to clear the market, GENCOs 1 and 3 still benefit from the high bid price offered by GENCO 2. As can be observed in Fig. 8, the cleared prices for all three GENCOs are higher than their marginal generation cost 50 $/MWh. In Table IV, all three GENCOs obtain higher total profits than the truthful bidding case. On average, the increase of the profit is 41%. This phenomenon is called "free riding" in game theory, where GENCOs 1 and 3 can bid at a lower price to get more of their quantity cleared at a higher marginal price.

The above bidding strategies form one NE and the reason is as follows: according to the definition of NE, no player can benefit by changing its strategy while the strategies of other players remain unchanged. Firstly, since GENCO 1 has no market power, increasing its bid price will only reduce its cleared quantity and the profit. GENCO 1 will not bid higher than the true marginal cost. Secondly, if GENCO 2 decreases the bid price to the marginal cost, all three GENCOs will bid truthfully like the baseline case, and all of them will receive a lower profit. Lastly, if GENCO 3 also adopts a similar strategy like GENCO 2, which is to bid high price at peak hours, the amount of its cleared power will be greatly reduced, which results in a lower profit. And this has been tested through the simulation. Therefore, no GENCO is willing to change its bidding strategy alone with the other two unchanged, which indicates an NE status.

E. Comparison with Model-based Method

To further verify that biding strategies obtained from MADDPG are approximated NE strategies, we compare the bidding results from MADDPG with those from solving the original MPEC models (5)-(9) and (1)-(4). Because there is one MPEC for each GENCO bidder, we apply the diagonalization algorithm (DIAG), which solves each MPEC iteratively. In each iteration, MPEC of each GENCO bidder is solved by setting the bid prices of all other GENCO bidders to the values from the last iteration. The iteration continues until there is no change in the bid price or the maximum number if iterations is reached. Table V compares the total profit from MADDPG and MPEC for one test day in the congested case.

TABLE V
COMPARISON OF MADDPG WITH MPEC

GENCO	Total profit (10^4 $)		Relative error (%)
	MADDPG	MPEC	
1	2.01	2.07	2.81
2	4.51	4.94	8.80
3	6.38	6.07	−5.23

Table V shows that the total profits from the two methods are close to each other, which verifies that the well-trained MADDPG algorithm can obtain an approximated NE for unseen cases. With regard to computation efficiency, the computation time for solving MPEC via DIAG is around 212 s for one test day. However, the total computation time for the well-trained MADDPG algorithm to generate bidding strategies for three GENCOs in the 31 test days is around 160 s. The acceleration ratio of the latter over the former is over 40 times, which proves the high computational efficiency of the deep RL method. This is because the well-trained algorithm does not require any iterative calculation process, but it can directly map the current observation to the bidding strategy with the function approximation property of DNN, which greatly saves computation efforts.

F. Comparison with State-of-the-art RL Methods

The major merit of MADRL over the existing deep RL method lies in its centralized training mechanism. During the training, the observations and actions of the rival agents are collected and analyzed by the critic network of each RL agent to develop a good grasp of the environment dynamics. It is believed that the extra information provided by the centralized training can facilitate a more effective and intelligent policy learning. The mechanism is applicable since during the testing phase, only the actors of the RL agents will be executed, who only have access to local observations without global information.

To demonstrate the advantages of MADDPG over the other deep RL methods, we compare the learning performance of MADDPG with two representative deep RL methods with the congested case. For each algorithm, we randomly generate 10 seeds, and for each seed, the algorithm is trained for

200 episodes, with each episode containing 24 time steps. In Fig. 9, the solid line stands for the average episodic reward over the five random seeds from the three RL methods, and the shaded area is the one-tenth of the standard deviation over the five runs. As can be observed in the figure, the average reward gradually becomes stable in all the three algorithms, which indicates the convergence. However, the final average reward obtained from MADDPG is higher than that from DQN and DDPG for all the three generators. The average reward increases of the three GENCO bidders over DQN and DDPG are 38.5% and 18.2%, respectively. This increase in the learning performance can be attributed to the deployment of the centralized training mechanism in MADDPG, where global information is adopted to obtain a comprehensive approximation of the multi-agent models. While in both DQN and DDPG algorithms, only local observations are utilized to train the neural networks within each RL agent, which can be biased from the actual market dynamics. Note that the average reward obtained by DQN method is lower than that of MADDPG and DDPG, owing to that DQN applies discretization to the continuous action domain, which can limit the search space and results in sub-optimal policy.

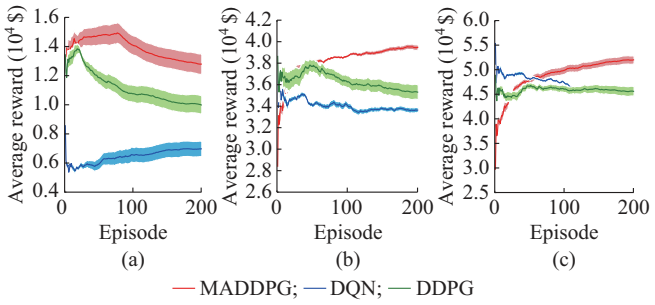

Fig. 9. Comparison of learning performance of different RL methods. (a) GENCO 1. (b) GENCO 2. (c) GENCO 3.

V. Conclusion

We present an MADDPG algorithm to approximate NE of the Markov game in the day-ahead electricity market. The MADDPG algorithm can learn a profitable bidding strategy for multiple GENCO bidders through centralized training and decentralized execution. In the simulation studies, the MADDPG-based bidding strategy is compared with a naive truthful bidding strategy, and the former achieves an average profit increase of 41% over the latter. The MADDPG algorithm is also compared with a model-based method to demonstrate its computational efficiency. The acceleration ratio of the former over the latter is over 40 times. Finally, MADDPG is compared with other state-of-the-art RL methods including DQN and DDPG, and achieves an average reward increase of 38.5% and 18.2% over the two methods. The utilization of complete information during the training allows the individual RL agent to formulate a more accurate approximation of the system dynamics and gain an improved learning performance.

One limitation of the proposed work is that only the day-ahead electricity market is considered in the paper, while in the real-world, the electricity market is composed of multiple market stages including day-ahead, intra-day, and real-time market. In such cases, MADDPG can still be applied to solve the associated Markov game through continuous interaction with the environment. Another limitation is that in the proposed work, only GENCOs act as strategic bidders, while in the real-world market, large consumers can also submit their bids to change the market clearing results. Modeling the multi-stage electricity market bidding and two-sided bidding will be the directions for our future researches.

References

[1] F. Li and Y. Du, "From AlphaGo to power system AI," *IEEE Power and Energy Magazine*, vol. 16, no. 2, pp. 76-84, Feb. 2018.
[2] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp.1066-1076, Mar. 2020.
[3] H. Huang and F. Li, "Bidding strategy for wind generation considering conventional generation and transmission constraints," *Journal of Modern Power Systems and Clean Energy*, vol. 3, no. 1, pp. 51-62, Mar. 2015.
[4] T. Dai and W. Qiao, "Finding equilibria in the pool-based electricity market with strategic wind power producers and network constraints," *IEEE Transactions on Power Systems*, vol. 32, no. 1, pp. 389-399, Jan. 2017.
[5] C. Wang, W. Wei, J. Wang *et al*., "Strategic offering and equilibrium in coupled gas and electricity markets," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 290-306, Jan. 2018.
[6] Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Investigating the ability of demand shifting to mitigate electricity producers market power," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 3800-3811, Jul. 2018.
[7] E. Moiseeva and M. R. Hesamzadeh, "Bayesian and robust nash equilibria in hydrodominated systems under uncertainty," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 818-830, Apr. 2018.
[8] M. Loschenbrand and M. Korpas, "Multiple Nash equilibria in electricity markets with price-making hydrothermal producers," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 422-431, Jan. 2019.
[9] S. Chen, A. J. Conejo, R. Sioshansi *et al*., "Equilibria in electricity and natural gas markets with strategic offers and bids," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1956-1966, May 2020.
[10] H. Kebriaei, A. Rahimi-Kian, and M. N. Ahmadabadi, "Model-based and learning-based decision making in incomplete information cournot games: a state estimation approach," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 4, pp. 713-718, Apr. 2015.
[11] M. R. Salehizadeh and S. Soltaniyan, "Application of fuzzy Q-learning for electricity market modeling by considering renewable power penetration," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 1172-1181, Apr. 2016.
[12] D. E. Aliabadi, M. Kaya, and G. Sahin, "Competition, risk and learning in electricity markets: an agent-based simulation study," *Applied energy*, vol. 195, pp. 1000-1011, Jun. 2017.
[13] N. Rashedi, M. A. Tajeddini, and H. Kebriaei, "Markov game approach for multi-agent competitive bidding strategies in electricity market," *IET Generation, Transmission & Distribution*, vol. 10, no. 15, pp. 3756-3763, Jan. 2016.
[14] E. Mocanu, D. C. Mocanu, P. H. Nguyen *et al*., "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698-3708, Jul. 2019.
[15] Z. Wan, H. Li, H. He *et al*., "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246-5257, Sept. 2019.
[16] H. Xu, H. Sun, D. Nikovski *et al*., "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6366-6375, Nov. 2019.
[17] Y. Ye, D. Qiu, M. Sun *et al*., "Deep reinforcement learning for strategic bidding in electricity markets," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343-1355, Mar. 2020.
[18] Z. Yan and Y. Xu, "A multi-agent deep reinforcement learning method for cooperative load frequency control of multi-area power systems," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4599-4608,

Nov. 2020.
[19] S. Wang, J. Duan, D. Shi *et al*., "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4644-4654, Nov. 2020.
[20] Y. Liang, C. Guo, Z. Ding *et al*., "Agent-based modeling in electricity market using deep deterministic policy gradient algorithm," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4180-4192, Nov. 2020.
[21] Y. Ye, D. Qiu, J. Li *et al*., "Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: a novel multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 130515-130529, Sept. 2019.
[22] R. Lowe, Y. Wu, A. Tamar *et al*., "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of 2017 Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, USA, Jan. 2017, pp. 6379-6390.
[23] D. R. Biggar and M. R. Hesamzadeh, *The Economics of Electricity Markets*. New York: John Wiley & Sons, 2014.
[24] S. F. Tierney, T. Schatzki, and R. Mukerji. (2018, Mar.). Uniform-pricing versus pay-as-bid in wholesale electricity markets: does it make a difference? [Online]. Available: http://citeseerx. ist. psu. edu/viewdoc/download?doi=10.1.1.365.2514&rep=rep1&type=pdf
[25] A. G. Vlachos and P. N. Biskas, "Demand response in a real-time balancing market clearing with pay-as-bid pricing," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 1966-1975, Dec. 2013.
[26] Y. Ren and F. D. Galiana, "Pay-as-bid versus marginal pricing – part I: strategic generator offers," *IEEE Transactions on Power Systems*, vol. 19, no. 4, pp. 1771-1776, Nov. 2004.
[27] K. R. Apt. (2020, Oct.) Strategic games: chapter 7 sealed-bid auctions. [Online]. Available: https://homepages.cwi.nl/~apt/stra/ch7.pdf
[28] Y. Huang, J. Shang, and C. Kang, "An operation mechanism and model of the day-ahead electricity market," *Automation of Electric Power Systems*, vol. 27, no. 3, pp. 23-27, Feb. 2003.
[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge: MIT press, 2018.
[30] T. P. Lillicrap, J. J. Hunt, A. Pritzel *et al*. (2015, Jan.). Continuous control with deep reinforcement learning. [Online]. Available: http:arXiv:1509.02971
[31] V. Mnih, K. Kavukcuoglu, D. Silver *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, Jun. 2015.
[32] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12-19, Feb. 2011.
[33] PJM. (2020, Jul.). PJM website. [Online]. Available: https://www.pjm.com/markets-and-operations.aspx

Yan Du received the B.S. degree from Tianjin University, Tianjin, China, in 2013, the M.S. degree from the Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing, China, in 2016, and the Ph.D. degree from University of Tennessee, Knoxville, USA, in 2020. Her research interests include power system control and optimization and deep learning.

Fangxing Li received the B.S.E.E. and M.S.E.E. degrees from Southeast University, Nanjing, China, in 1994 and 1997, respectively, and the Ph.D. degree from Virginia Tech, Blacksburg, USA, in 2001. He is currently a James McConnell Professor at the University of Tennessee, Knoxville, USA. He is a Fellow of IEEE (Class of 2017) and a recipient of the 2020 R&D 100 Award. Presently, he is the Editor-in-Chief of the IEEE Open Access Journal of Power and Energy and the Chair of IEEE Power System Operation, Planning, and Economics (PSOPE) Committee. His research interests include renewable energy integration, demand response, power markets, power system control, and power system artificial intelligence.

Helia Zandi received the B.S. degree in computer science from the University of Tehran, Tehran, Iran, in 2010, and the M.S. degree in computer engineering from the University of Florida, Gainesville, USA, in 2012. She is currently pursuing the Ph.D. degree in computer engineering with the University of Tennessee, Knoxville, USA. She is also a Modeling and Simulation Software Engineer with the Oak Ridge National Laboratory, Oak Ridge, USA. Her research interests include home energy management, statistical methods, machine learning and their applications in buildings, and smart grid related applications.

Yaosuo Xue received the B.Sc. degree from East China Jiaotong University, Nanchang, China, in 1991, and the M.Sc. degrees in electrical engineering from the University of New Brunswick, Fredericton, Canada, in 2004. He is currently a R&D Staff with the Oak Ridge National Laboratory, Oak Ridge, USA. He is currently serving as an Associate Editor for the IEEE Transactions on Power Electronics and an Editor for the IEEE Open Access Journal of Power and Energy. His research interests include multilevel converters and smart inverter controls for renewable energy and utility applications.