Deep Reinforcement Learning for Joint Bidding and Pricing of Load Serving Entity

Xu, Hanchen; Sun, Hongbo; Nikovski, Daniel N.; Kitamura, Shoichi; Mori, Kazuyuki; Hashimoto, Hiroyuki

TR2020-003 January 08, 2020

Abstract

In this paper, we address the problem of jointly determining the energy bid submitted to the wholesale electricity market (WEM) and the energy price charged in the retailed electricity market (REM) for a load serving entity (LSE). The joint bidding and pricing problem is formulated as a Markov decision process (MDP) with continuous state and action spaces, in which the energy bid and the energy price are two actions that share a common objective. We apply the deep deterministic policy gradient (DDPG) algorithm to solve this MDP for the optimal bidding and pricing policies. Yet, the DDPG algorithm typically requires a significant number of state transition samples, which is costly in this application. To this end, we apply neural networks to learn dynamical bid and price response functions from historical data to model the WEM and the collective behavior of the EUCs, respectively. These response functions explicitly capture the inter-temporal correlations of the WEM clearing results and the EUC responses, and can be utilized to generate state transition samples without any cost. More importantly, the response functions also inform the choice of states in the MDP formulation. Numerical simulations illustrated the effectiveness of the proposed methodology.

IEEE Transactions on smart grids

Deep Reinforcement Learning for Joint Bidding and Pricing of Load Serving Entity

Hanchen Xu, *Student Member, IEEE*, Hongbo Sun, *Senior Member, IEEE*, Daniel Nikovski, *Member, IEEE*, Shoichi Kitamura, Kazuyuki Mori, *Member, IEEE*, Hiroyuki Hashimoto

***Abstract*—In this paper, we address the problem of jointly determining the energy bid submitted to the wholesale electricity market (WEM) and the energy price charged in the retailed electricity market (REM) for a load serving entity (LSE). The joint bidding and pricing problem is formulated as a Markov decision process (MDP) with continuous state and action spaces, in which the energy bid and the energy price are two actions that share a common objective. We apply the deep deterministic policy gradient (DDPG) algorithm to solve this MDP for the optimal bidding and pricing policies. Yet, the DDPG algorithm typically requires a significant number of state transition samples, which is costly in this application. To this end, we apply neural networks to learn dynamical bid and price response functions from historical data to model the WEM and the collective behavior of the EUCs, respectively. These response functions explicitly capture the inter-temporal correlations of the WEM clearing results and the EUC responses, and can be utilized to generate state transition samples without any cost. More importantly, the response functions also inform the choice of states in the MDP formulation. Numerical simulations illustrated the effectiveness of the proposed methodology.**

***Index Terms*—electricity market, bidding, pricing, load serving entity, demand response, deep reinforcement learning.**

I. Introduction

IN a restructured power system industry, a load serving entity (LSE) needs to submit bids for electricity/energy in a wholesale electricity market (WEM), which is operated by an independent system operator (ISO), so as to meet the demand from its end use customers (EUCs). Conventionally, the LSE charges the EUCs for electricity/energy a fixed tariff that is regulated by the government. Therefore, the decision making process of the LSE involves only the bidding problem—the determination of the energy bids, which will typically rely on the forecast of the relatively inflexible EUC demand. However, due to the rapid development of smart grid technologies, demand-side management becomes feasible through demand response programs such as real-time pricing [1], [2]. An LSE may determine a real-time energy price in the retail electricity market (REM) it operates to incentivize the EUCs changing their energy consumption behaviors in a way that benefits the LSE. In this context, in addition to the bidding problem, the LSE is also faced with the pricing problem—the determination of the energy price that is charged to the EUCs [3].

Existing works are mostly concerned with either the bidding/offering problem [4]–[7] or the pricing problem [8]–[12]. In regards of the bidding problem, optimal bidding functions are developed for price-sensitive and price-insensitive demands in [5]. A genetic algorithm based method is proposed for finding the optimal bidding strategy in a two-settlement energy market in [6]. The strategic bidding problem of LSEs has also been formulated as a bi-level programing problem in [7], where the upper level problem is to maximize the LSE's net revenue and the lower level problem is the ISO's economic dispatch. For the pricing problem, optimization algorithms (see, e.g., [8]–[10]) have been applied to dynamically price the energy in the REM. Similar to optimization based bidding algorithms, optimization based pricing algorithms typically resort to bi-level programing techniques and thus require specific models of the demand-side resources. For example, constraints in the lower level problem need to be linear in order to transform the bi-level programing problem into a solvable mixed integer linear programing (MILP) problem. Moreover, all model parameters need to be known in order to solve the optimization problem, which may be impractical in reality. The Q-learning algorithm—a mode-free reinforcement learning (RL) algorithm—has also been applied to solving the pricing problem [11]. While it does not require model parameters and can handle nonlinear constraints in the EUCs, discretization of the state and action spaces are required, which may lead to a problem referred to as the "curse of dimensionality" [13] and thus limit its applicability.

The bidding problem and the pricing problem are inherently coupled, since the energy purchased in the WEM and that sold in the REM must balance, and the profit earned by the LSE is dependent on the results in both markets. Therefore, it is indeed more desirable to solve the bidding problem and the pricing problem jointly. Yet, methodologies that solve the joint bidding and pricing problem have been rarely studied. The few existing works on this topic such as [3] model the joint bidding and pricing problem as a bi-level programing problem, which is solved using MILP techniques. However, existing solutions typically assume all market participants are myopic, the parameters of all models, including all market participants in the WEM and all EUCs in the REM, are completely known to the LSE, and more importantly, all the models are linear. These assumptions, however, are very constraining

Hanchen Xu is with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Email: hxu45@illinois.edu.

Hongbo Sun, Daniel Nikovski, and Hiroyuki Hashimoto are with Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA. Email: {hongbosun, nikovski, hashimoto}@merl.com.

Shoichi Kitamura and Kazuyuki Mori are with Advanced Technology R&D Center, Mitsubishi Electric Corporation, Hyogo 661-8661, Japan. Email: Kitamura.Shoichi@dw.MitsubishiElectric.co.jp, Mori.Kazuyuki@ab.MitsubishiElectric.co.jp.

This work was done when Hanchen Xu was a Research Intern in MERL.

and impractical. Therefore, effective methodologies that can address the joint bidding and pricing problem in a more general setting, specifically, in a setting in which parameters of the potentially nonlinear model are not known, the states and actions are continuous, and the LSE may be far-sighted, are still to be investigated.

To address the aforementioned issues, we formulate the joint bidding and pricing problem as a Markov decision process (MDP), in which the energy bid and the energy price are two actions that share a common objective. To solve this MDP without the necessity to knowing the WEM and EUC models, the deep deterministic policy gradient (DDPG) algorithm, a policy-based deep RL (DRL) algorithm, is applied to learn the bidding and pricing policies, which determine the optimal action from the state. We note that RL/DRL has proven to be successful in many tasks such as playing games [14], control [15], robotic manipulation [16], as well as many others [17]. RL/DRL algorithms have also been widely applied in power systems, such as generation offer construction [18], demand response [19], [20], and voltage control [21], [22]. We refer interested readers to [13] for a comprehensive review on existing and potential applications of RL/DRL in power systems.

DRL algorithms typically require a large number of state transition samples. Yet, it is costly to obtain such samples from the actual environment. Moreover, it is also infeasible to generate samples from models since the models of other market participants in the WEM and all EUCs in the REM are not known in advance. To this end, neural networks are applied to learn a bid response function and a price response function from historical data to model the WEM and the collective behavior of the EUCs, respectively, from the perspective of the LSE. These response functions can explicitly capture the inter-temporal correlations of the WEM clearing results and the EUC responses, and can be utilized to generate state transition samples without any cost. More importantly, they also inform the choice of the states in the MDP formulation.

To the best of our knowledge, this is the first paper that applies DRL to solve the joint bidding and pricing problem of an LSE. The major contributions of this paper lie specifically in the following three aspects:

1) the formulation of the joint bidding and pricing problem as an MDP, which allows the consideration of an accumulative profit of the LSE in the long-term;
2) the development of dynamical bid and price response functions that model the WEM and the REM using historical data, which captures the inter-temporal correlations and informs the choice of states in the MDP formulation;
3) the application of a state-of-the-art DRL algorithm—the DDPG algorithm—to solve the MDP while taking into account its structural characteristics.

The remainder of this paper is organized as follows. Section II introduces the hierarchical market model. Section III proposes the bid and price response functions, defines the bidding and pricing policies, and formulates the joint bidding and pricing problem as an MDP. Learning algorithms for response functions and the bidding and pricing policies are presented in Section IV, and the application of the proposed methodology is illustrated in Section V. Section VI concludes the paper.

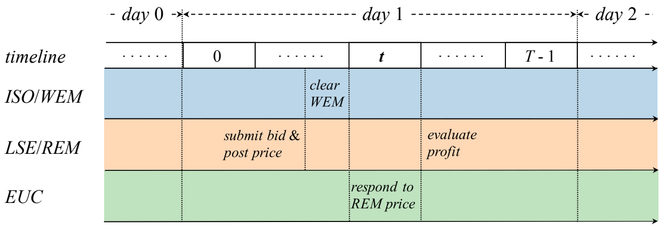

Fig. 1. Timeline of actions in the real-time market for interval t.

II. Hierarchical Market Model

In this section, we present a hierarchical market model that consists of a WEM operated by an ISO, a REM managed by a LSE, and a set of EUCs. Throughout this paper, all vectors and matrices are in bold and italics. A subscript t indicates the value of a variable at time interval t.

A. Overview

Assume one day is decomposed into T time intervals indexed by the elements in the set $\mathcal{T} = \{0, \cdots, T-1\}$. Let t index the time intervals; then $t \bmod T \in \mathcal{T}$, where mod denotes the modulo operation. The problem setting considered in this paper is described as follows. Priori to time interval t, each market participants, including the sellers and buyers, need to submit energy offers/bids for time interval t. Then, a WEM is cleared to yield a wholesale energy price, as well as the energy sales and purchases that are successfully cleared for each seller and buyer, respectively. In the meantime, the LSE, which is a buyer in the WEM, also determines a retail energy price (simply referred to as the price) for time interval t, at which it resells the energy to its customers, i.e., the EUCs, in the REM. During time interval t, the EUCs respond to the price signal by adjusting their energy consumptions. The LSE needs to make payments to the ISO for the energy consumed by the EUCs; meanwhile, it also collect payments from the EUCs. The total amount of profit resulted from energy trading in these two markets can be evaluated after time interval t. The sequence of actions taken by different parties in the real-time market is illustrated in Fig. 1. This process is repeated for all the time intervals.

B. Wholesale Market Model

Let $\mathcal{G} = \{g_1, \cdots, g_G\}$ denote the set of the sellers, and $\mathcal{B} = \{b_1, \cdots, b_B\}$ the set of buyers. Each seller $g \in \mathcal{G}$ submits an offer, denoted by $f_t^g(\cdot)$, which specifies the minimum price at which it is willing to sell energy during time interval t. Specifically, $f_t^g(q_t^g)$ is the minimum price at which seller g is willing to sell energy during time interval t with a quantity of q_t^g. Similarly, each buyer $b \in \mathcal{B}$ submits a bid, denoted by $f_t^b(\cdot)$, that specifies the maximum price at which it is willing to buy energy during time interval t. Specifically, $f_t^b(q_t^b)$ is the maximum price at which a buyer is willing to buy energy during time interval t with a quantity of q_t^b.

Then, assuming the bulk power system is lossless and congestion-free, the ISO clears the WEM by solving the following social welfare maximization problem:

$$\underset{\substack{q_t^{g_1},\cdots,q_t^{g_G}\\ q_t^{b_1},\cdots,q_t^{b_B}}}{\operatorname{maximize}} \sum_{b\in\mathcal{B}}\int_0^{q_t^b} f_t^b(q)\mathrm{d}q - \sum_{g\in\mathcal{G}}\int_0^{q_t^g} f_t^g(q)\mathrm{d}q, \tag{1a}$$

subject to

$$\sum_{b\in\mathcal{B}} q_t^b - \sum_{g\in\mathcal{G}} q_t^g = 0 \leftrightarrow \lambda_t, \tag{1b}$$

$$(q_t^{g_1},\cdots,q_t^{g_G},q_t^{b_1},\cdots,q_t^{b_B}) \in \mathcal{Q}_t, \tag{1c}$$

where (1b) is the power balance equation, λ_t is the dual variable associated with constraint (1b), $\mathcal{Q}_t$ is the feasible set of the decision variables, which may depend on the market clearing results in the previous time interval. For convenience, denote the total cleared energy sales/purchases by q_t, i.e., $q_t = \sum_{b\in\mathcal{B}} q_t^b = \sum_{g\in\mathcal{G}} q_t^g$.

The solution to (1) gives cleared energy sales and purchases, as well as the wholesale energy price for each market participant. In a uniform pricing market, all market participants receive a uniform price that equals to λ_t. When the WEM is competitive, a single market participant typically does not have the capability to influence the clearing price and the chances that it is the marginal unit are low. In such a setting, given λ_t, the cleared energy purchase for the buyer b when it is non-marginal, can be computed as follows:

$$q_t^b = \arg\max_{q^b} q^b|_{f_t^b(q^b)\geq\lambda_t}. \tag{2}$$

C. Retail Market Model

In the WEM, the LSE participates as a buyer that purchases energy through bidding. Without loss of generality, assume the LSE under consideration is buyer b in the WEM. The LSE resells the purchased energy to a set of EUCs in the REM and charges them at a typically regulated price that it needs to determine. Let ν_t denote the price at time interval t, and q_t^b the energy purchased from the WEM.

Let $\mathcal{C} = \{c_1,\cdots,c_C\}$ denote the set of the EUCs in the REM served by this LSE. For EUC $c\in\mathcal{C}$, it will respond to the price ν_t by adjusting its energy consumption, denoted by d_t^c. Denote the aggregate energy consumption of all EUCs measured at the substation during time interval t by d_t, i.e., $d_t = \sum_{c\in\mathcal{C}} d_\tau^c$. Then, the objective of the LSE considered here is to maximize its profit earned from time interval t and onwards, subject to the energy balance constraint, which can be mathematically express as follows:

$$\underset{\nu_t,\nu_{t+1},\cdots,\in[\underline{\nu},\overline{\nu}]}{\operatorname{maximize}} \ \mathbb{E}\left[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}((\nu_\tau-\lambda_\tau)d_\tau - \phi_\tau(d_\tau,q_\tau^b)\right]), \tag{3}$$

where $\mathbb{E}$ denotes expectation operation, $\gamma\in[0,1)$ is a discount factor that discounts the future profit, $\phi_\tau(\cdot,\cdot)$ is a non-negative scalar function, $\underline{\nu}$ and $\overline{\nu}$ are the minimum and maximum prices, respectively. Note that λ_τ and q_τ^b are determined by the WEM through(1), while d_τ is determined by the EUCs through (4) that is to be detailed in Section II-D. The objective consists of two components, of which the first is the profit earned from energy trading by the LSE, and the second is the cost incurred when the aggregate energy consumption deviates from the energy purchase. Note that the actual aggregate energy consumption d_τ is used when computing both the payment made to the WEM and that collected from the EUCs.

We emphasize that this objective function is general enough to reflect payment structure in various market schemes since there is no constraint on the form of the function ϕ. In addition, while the objective of the LSE under consideration is maximizing its profit, the methodology proposed in this paper also extends directly to LSEs with other objectives, for example, LSEs that aim to maximize the overall benefit of all EUCs.

D. End Use Customer Model

At the beginning of each time interval t, EUC c, $c\in\mathcal{C}$, receives a price ν_t from the LSE, it will then optimize its energy consumption so as to maximize its overall benefit. We next present a generic EUC model that is agnostic to the underlying components. Let e_t^c denote the energy need of EUC c at time interval t. Similar concepts are adopted in works such as [9], [11]. A myopic EUC finds its optimal action via solving the following utility maximization problem:

$$\underset{d_t^c\in\mathcal{D}_t^c}{\operatorname{maximize}} \ \beta^c(e_t^c,d_t^c) - \nu_t d_t^c, \tag{4a}$$

subject to

$$e_{t+1}^c = e_t^c + \eta_t^c(e_t^c - d_t^c) + \xi_t^c, \tag{4b}$$

where $\beta^c(\cdot)$ is the benefit function, which gives the benefit of the EUC at certain energy need and energy consumption, $\eta_t^c\in[0,1]$ is the backlog rate that represents the percentage of unmet energy need that is carried over to the next time interval, ξ_t^c is a random variable that models that newly generated incremental energy need, $\mathcal{D}_t^c$ is the feasible set of the energy consumption.

III. Problem Formulation

In this section, we first introduce dynamical bid and price response functions, followed by the bidding and pricing policies. Then, we formulate the joint bidding and pricing problem faced by the LSE as an MDP. Figure 2 illustrates the interaction between the LSE, the ISO, and the EUCs.

A. Bid and Price Response Functions

From the perspective of the LSE, it has to determine a bid f_t^b—the bidding problem, as well as a price ν_t—the pricing problem, for time interval t. Assume f_t^b is characterized by a parameter vector $\boldsymbol{\omega}_t$. Let $\{\lambda_\tau,q_\tau\}_{t-n_1}^{t-1}$ denote the set of WEM clearing results from time interval $t-n_1$ to $t-1$. Then, pursuing the a similar idea as in our earlier work in [23], we model the interaction between the LSE and the WEM defined

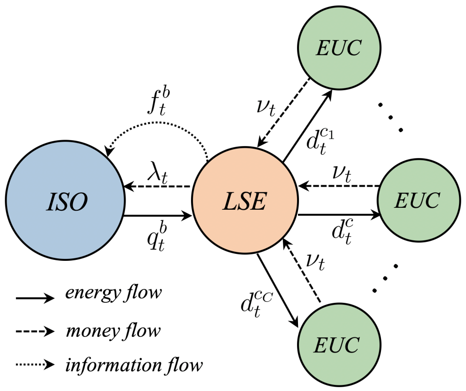

Fig. 2. Interaction between the ISO, the LSE, and the EUCs.

through (1) using a n_1-order bid response function, denoted by $\boldsymbol{\psi}(\cdot)$, as follows:

$$(\lambda_t, q_t) = \boldsymbol{\psi}(\{\lambda_\tau, q_\tau\}_{t-n_1}^{t-1}, \boldsymbol{\omega}_t, t \bmod T), \tag{5}$$

where $(t \bmod T)$ is included to model the time dependence. The cleared energy purchase can be computed using (2). In the special case when $n_1 = 0$, the WEM clearing results only depend on the bid from the LSE at the current time interval. For a perfectly competitive WEM, $\boldsymbol{\omega}_t$ has negligible impacts on the clearing results, and (5) essentially models the dynamics of the clearing results. From the perspective of the LSE, the WEM clearing results will evolve to λ_t, q_t, q_t^b from previous WEM clearing results, given its bid $\boldsymbol{\omega}_t$. The impacts from other market participants' actions are included in this bid response function. Therefore, when n_1 is large enough, the n_1-order bid response function can well capture the dynamics in the WEM.

In the meantime, the LSE may only have information on the aggregate energy consumption d_t in real time, rather than complete parameters in (4). Therefore, instead of adopting the complete EUC model in (4), we use a n_2-order price response function, denoted by $\varphi(\cdot)$, to characterize the collective behavior of all EUCs defined through the set of problems in (4), as follows:

$$d_t = \varphi(\{d_\tau, \nu_\tau\}_{t-n_2}^{t-1}, \nu_t, t \bmod T). \tag{6}$$

In the special case when $n_2 = 0$, the aggregate EUC demand only depends on the price at the current time interval. The core idea behind the price response function is similar to that of the bid response function. Compared to the complete WEM model and EUC models, the response functions are easier to learn from the data that are available to the LSE.

B. Bidding and Pricing Policies

The objective of the joint bidding and pricing problem to be solved by the LSE is to determine the bid and the price based on available information. As discussed earlier, prior to time interval t, the information related to the WEM that is available to the LSE includes $\boldsymbol{\omega}_\tau$, λ_τ, q_τ, $\forall \tau \leq t-1$. In the meantime, the information related to the REM that is available to the LSE includes ν_τ, d_τ, $\forall \tau \leq t-1$. Let $\mathcal{I}_{t-1} = \{\boldsymbol{\omega}_\tau, \lambda_\tau, q_\tau, \nu_\tau, d_\tau, \forall \tau \leq t-1\}$ denote the set of information available to the LSE before the WEM for time interval t is cleared. Then, the joint bidding and pricing problem for time interval t is to determine the bid $\boldsymbol{\omega}_t$ and the price ν_t from $\mathcal{I}_{t-1}$. The bidding problem and the pricing problem are inherently coupled, and thus need to be considered jointly. While it is feasible to define a joint bidding and pricing policy that maps $\mathcal{I}_{t-1}$ to $\boldsymbol{\omega}_\tau$ and ν_τ, yet, we will show next that in a competitive uniform pricing market, it is more desirable to define a bidding policy and a pricing policy separately since this allows more efficient utilization of the information for decision-making.

In a uniform pricing market, the LSE's bid will get cleared as long as its bid price is no smaller than λ_t. Meanwhile, to minimize the cost incurred due to the mismatch of the energy purchase and aggregate energy consumption, it is indeed desirable to bid for the amount of energy that equals to the aggregate energy consumption. In fact, when λ_t is not affected by $\boldsymbol{\omega}_t$, which is the case in a competitive uniform pricing market, for any ν_t, the optimal bid $\boldsymbol{\omega}_t$ that maximizes the profit defined in (3) is the one that gives $q_\tau^b = d_\tau$. Essentially, we only need to find the optimal price ν_t for the REM, and then construct the bid from ν_t.

Define a deterministic pricing policy, denoted by $\pi(\cdot)$, as the following function that maps $\mathcal{I}_{t-1}$ to the price ν_t:

$$\nu_t = \pi(\mathcal{I}_{t-1}). \tag{7}$$

Also, define a deterministic bidding policy, denoted by $\boldsymbol{\mu}(\cdot)$, as the following function that maps $\mathcal{I}_{t-1}$ and ν_t to a bid $\boldsymbol{\omega}_t$:

$$\boldsymbol{\omega}_t = \boldsymbol{\mu}(\mathcal{I}_{t-1}, \nu_t). \tag{8}$$

Assume the bid $\boldsymbol{\omega}_t$ consists of two components, a bid price ω_t^p in \$/MWh and a bid quantity ω_t^q in MWh. Then, the optimal bidding policy $\boldsymbol{\mu}^*$ is such that ω_t^p is set to ν_t and ω_t^q is set to the estimated aggregate energy consumption obtained using the price response function φ. Therefore, there is no additional parameter in $\boldsymbol{\mu}$ that needs to be learned beyond those in φ.

C. Markov Decision Process Formulation

We next formulate the joint bidding and pricing problem as an MDP. An MDP consists of a state space, an action space, a reward function, and a transition probability function that satisfies the Markov property [24], i.e., given the current state and action, the next state is independent of all states and actions in the past.

Specifically, in the joint bidding and pricing problem, define the state at time interval t to be $\boldsymbol{s}_t = (\{\lambda_\tau, q_\tau\}_{t-n_1}^{t-1}, \{d_\tau, \nu_\tau\}_{t-n_2}^{t-1}, t \bmod T)$. Define the action for time interval t to be $a_t = \nu_t$. As discussed in the previous section, $\boldsymbol{\omega}_t$ can be constructed from ν_t through a set of deterministic procedures. Both the state and action spaces are continuous. Given $\boldsymbol{s}_t$ and a_t, $\boldsymbol{s}_{t+1}$ is determined through (5) and (6). Therefore, the Markov property is satisfied. However, the transition probability function is determined by all the market participants in the WEM as well all EUCs in the REM, and is unknown to the LSE.

Then, the pricing policy can equivalently become

$$\nu_t = \pi(\boldsymbol{s}_t), \tag{9}$$

and the bidding policy can be equivalently written as:

$$\boldsymbol{\omega}_t = \boldsymbol{\mu}(\boldsymbol{s}_t, \nu_t). \tag{10}$$

The objective of the joint bidding and pricing problem is to maximize the profit of the LSE; therefore, we define the reward for time interval t to be the profit earned by the LSE as follows:

$$r_t = (\nu_t - \lambda_t)d_t - \phi_t(d_t, q_t^b). \tag{11}$$

The cumulative discounted reward from time interval t and onwards, denoted by R_t and referred to as the return, is $R_t = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau$, where $\gamma \in [0, 1)$ is a discount factor. The action value function, also referred to as the Q function, under pricing policy π and bidding policy $\boldsymbol{\mu}$, at action a and state $\boldsymbol{s}$, denoted by $Q^{\pi,\boldsymbol{\mu}}(\boldsymbol{s}, a)$, is the expected return defined as

$$Q^{\pi,\boldsymbol{\mu}}(\boldsymbol{s}_t, a_t) = \mathbb{E}\left[R_t | \boldsymbol{s}_t, a_t; \pi, \boldsymbol{\mu}\right]. \tag{12}$$

The Q function under optimal pricing policy π^* and optimal bidding policy $\boldsymbol{\mu}^*$, denoted by $Q^*(\cdot,\cdot)$, satisfies the Bellman optimality equation:

$$Q^*(\boldsymbol{s}_t, a_t) = \mathbb{E}\left[r_t\right] + \gamma \int_{\mathcal{S}} \mathbb{P}\left\{\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, a_t\right\} \max_a Q^*(\boldsymbol{s}_{t+1}, a), \tag{13}$$

where $\mathbb{P}\left\{\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, a_t\right\}$ is the probability that the state transit into $\boldsymbol{s}_{t+1}$ conditioning on $\boldsymbol{s}_t$, a_t.

Since $\boldsymbol{\mu}^*$ does not need to be learned once we have φ, the joint bidding and pricing problem essentially becomes finding π that maximize the following performance function [25]:

$$J(\pi) = \mathbb{E}\left[R_1; \pi, \boldsymbol{\mu}^*\right], \tag{14}$$

which gives the expected return under given bidding and pricing policies. For ease of notation, we write $Q^{\pi,\boldsymbol{\mu}^*}(\cdot,\cdot)$ simply as $Q(\cdot,\cdot)$. The MDP problem can be solved leveraging a RL algorithm be detailed in Section IV.

We emphasize that although learning a joint bidding and pricing policy is reduced to learning a pricing policy, this problem is different from a pricing problem alone in the following aspects. First of all, the bidding problem and the pricing problem are addressed simultaneously, and their mutual impacts are implicitly captured in the proposed formulation. In addition, the bidding policy is constructed from the pricing policy π and the price response function φ, which are important components in the proposed methodology.

IV. Learning Algorithms

In this section, we first present learning methods for the response functions. Particularly, the optimal bidding policy is derived from the bid response function. Then, we apply the DDPG algorithm to finding the optimal pricing policy.

A. Response Function Learning

In RL algorithms, transitions $(\boldsymbol{s}_\tau, a_\tau, r_\tau, \boldsymbol{s}_{\tau+1})$ are critical for learning a good policy. Typically, a large number of transition samples are needed in order to learn a good policy. One approach to obtain the transitions is to sample from the actual environment online, i.e., to get samples from directly interacting with the ISO and the EUCs, till adequate samples are acquired. This approach, however, does not utilize the samples in an efficient manner. In addition, this may incur significant cost for the LSE during action exploration.

Alternatively, we can learn the bid response function $\boldsymbol{\psi}$ and the price response function φ from historical data and use the learned response functions as a substitute to the actual environment. The learned response functions can generalize the transition samples to new transitions, and if accurate enough, would allow the learning of good bidding and pricing policies without incurring any cost.

The response function learning problems can be cast as supervised learning problems. When learning the bid response function, the inputs are $(\{\lambda_\tau, q_\tau\}_{t-n_1}^{t-1}, \boldsymbol{\omega}_t, t \bmod T)$ and the outputs are (λ_t, q_t). When learning the price response function, the inputs are $(\{d_\tau, \nu_\tau\}_{t-n_2}^{t-1}, \nu_t, t \bmod T)$ and the output is d_t. The objective of the learning algorithm is to minimize the mean squared error between the predicted values and the actual values of the outputs. Then, the response functions can be represented using neural networks, the parameters of which can be learned using the backpropagation algorithm [26].

B. Policy Learning

We next introduce the learning algorithms for pricing policy π. Note that no additional parameter in the bidding policy $\boldsymbol{\mu}$ needs to be learned beyond those in φ. Assume π is parameterized by a vector $\boldsymbol{\theta}^\pi$. Then, finding the optimal pricing policy is essentially finding the optimal value for $\boldsymbol{\theta}^\pi$. One type of RL algorithms that can find (sub-optimal) values for $\boldsymbol{\theta}^\pi$ is the policy gradient methods, which update the parameter vector in the direction that maximizes $J(\pi)$. The gradient of J can be computed according to the Deterministic Policy Gradient Theorem [25]. Specifically, the gradient of J with respect to $\boldsymbol{\theta}^\pi$, referred to as the action gradient, is as follows:

$$\nabla_{\boldsymbol{\theta}^\pi} J = \mathbb{E}\left[\nabla_a Q(\boldsymbol{s}, a) \nabla_{\boldsymbol{\theta}^\pi} \pi(\boldsymbol{s})\right]. \tag{15}$$

Note that the gradient of the performance function J depends on the action value function Q, which is not known and needs to be estimated.

In the DDPG algorithm proposed by authors in [15], the actor-critic architecture is adopted, in which a critic is used to estimate the Q function, and an actor is used to estimate the policy. Neural networks as adopted to approximate these functions. Specific to the joint bidding and pricing problem, we represent the Q function by a neural network—referred to as the critic network—with a parameter vector $\boldsymbol{\theta}^Q$. The parameters of the critic network can be estimated using methods such as temporal-difference learning. Meanwhile, the pricing policy is represented by a neural network—referred to as the pricing policy network—with a parameter vector $\boldsymbol{\theta}^\pi$. The bidding policy is also represented by a neural network, which consists of the learned bid response function. The bidding and pricing policy networks are collectively referred to as the actor networks. The parameters of the pricing policy network can be estimated using the policy gradient method.

In addition to using the neural networks, there are two more important ideas in the DDPG algorithm. First, target

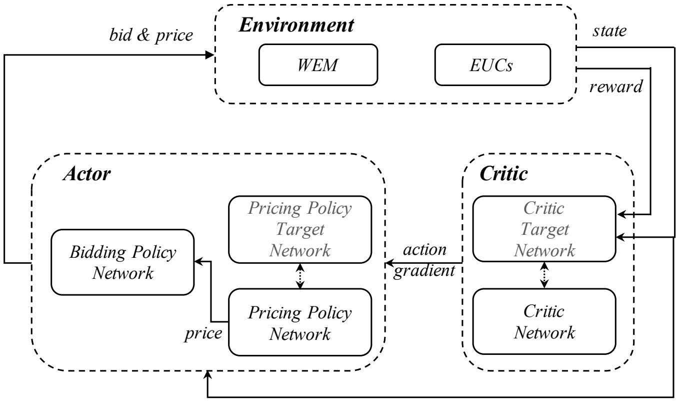

Fig. 3. Interaction of components in the joint bidding and pricing algorithm.

networks, the parameters of which slowly track those of the actor network and the critic network, are used to stabilize the algorithm. The parameter vector of the target network for the critic is denoted by $\boldsymbol{\theta}^{Q'}$, and that of the pricing network network is denoted by $\boldsymbol{\theta}^{\pi'}$. Second, a replay buffer $\mathcal{R}$ is used to store the transitions of the MDP and at each time instant, a mini-batch of size m is sampled from the $\mathcal{R}$ and used to estimate the gradients. The interaction between each component in the algorithm is illustrated in Fig. 3. Note that in the training stage, response functions are used to substitute the WEM and EUCs.

The detailed DDPG based RL algorithm for solving the joint bidding and pricing problem is presented in Algorithm 1. At each step, $\boldsymbol{\theta}^Q$ is updated in the direction that minimizes the following loss function:

$$\ell = \frac{1}{m}\sum_i (r_i + \gamma Q'(\boldsymbol{s}_{i+1}, \pi'(\boldsymbol{s}_{i+1})) - Q(\boldsymbol{s}_i, a_i))^2. \quad (16)$$

The intuition behind this is to actually find a critic network that satisfy the Bellman optimality equation in (13). Note that the target networks are used to compute the action value as well as the next action, i.e., $\boldsymbol{\pi}'(\boldsymbol{s}_{i+1})$. Meanwhile, $\boldsymbol{\theta}^\pi$ is updated in the direction that maximizes J, specifically, the direction of action gradient that is approximated using samples as follows:

$$\nabla_{\boldsymbol{\theta}^\pi} J \approx \frac{1}{m}\sum_i \nabla_a Q(\boldsymbol{s}_i, \pi(\boldsymbol{s}_i))\nabla_{\boldsymbol{\theta}^\pi}\pi(\boldsymbol{s}_i). \quad (17)$$

Algorithm 1: DDPG-based Policy Learning [15]

```
Input:
    $\boldsymbol{\psi}$: bid response function
    $\varphi$: price response function
    $\alpha^a$: actor learning rate
    $\alpha^c$: critic learn rate
    $M$: number of episodes
    $m$: mini-batch size
Output:
    $\pi$: policy
Randomly initialize critic network $Q(\boldsymbol{s}, a)$ and actor
  network $\pi(\boldsymbol{s})$, with weights $\boldsymbol{\theta}^Q$ and $\boldsymbol{\theta}^\pi$, respectively
Initialize target networks $Q'$, and $\pi'$, with weights
  $\boldsymbol{\theta}^{Q'} \leftarrow \boldsymbol{\theta}^Q$ and $\boldsymbol{\theta}^{\pi'} \leftarrow \boldsymbol{\theta}^\pi$, respectively
Initialize replay buffer $\mathcal{R}$
for $episode = 1, \cdots, M$ do
    Initialize a random process $\zeta$ for price exploration
    Receive initial state $\boldsymbol{s}_0$
    for $\tau = 0, \cdots, T-1$ do
        Select a price according to
            $\nu_\tau = \pi(\boldsymbol{s}_\tau) + \zeta_\tau$
          and a bid according to
            $\boldsymbol{\omega}_\tau = \boldsymbol{\mu}^*(\boldsymbol{s}_\tau, \nu_\tau)$
        Obtain $\lambda_\tau$, $q_\tau$ from (5) and $d_\tau$ from (6)
        Compute reward $r_\tau$ according to (11)
        Store transition $(\boldsymbol{s}_\tau, a_\tau, r_\tau, \boldsymbol{s}_{\tau+1})$ into $\mathcal{R}$
        Sample from $\mathcal{R}$ a mini-batch of $m$ transitions
          $(\boldsymbol{s}_i, a_i, r_i, \boldsymbol{s}_{i+1})$ if $|\mathcal{R}| > m$ else continue
        Update critic network by minimizing $\ell$ in (16):
            $\boldsymbol{\theta}^Q = \boldsymbol{\theta}^Q - \alpha^c \nabla_{\boldsymbol{\theta}^Q} \ell$
        Update actor network by maximizing $J$ in (14)
          using sampled gradients in (17):
            $\boldsymbol{\theta}^\pi = \boldsymbol{\theta}^\pi + \alpha^a \nabla_{\boldsymbol{\theta}^\pi} J$
        Update target networks:
            $\boldsymbol{\theta}^{Q'} \leftarrow \rho\boldsymbol{\theta}^Q + (1-\rho)\boldsymbol{\theta}^{Q'}$,
            $\boldsymbol{\theta}^{\pi'} \leftarrow \rho\boldsymbol{\theta}^\pi + (1-\rho)\boldsymbol{\theta}^{\pi'}$
    end
end
```

V. Numerical Simulation

In this section, we illustrate the application of the joint bidding and pricing algorithm through numerical simulations.

A. Simulation Setup

The WEM model is constructed based on the IEEE 300-bus test system, which has 69 generators, each corresponding to one seller, and 195 loads, each corresponding to one buyer. For an illustrative purpose, assume each offer/bid is a pair of offer/bid price (in \$/MWh) and quantity (in MW). Then, $\boldsymbol{\omega}_t$ is a two-dimensional vector that consists of a bid price and a bid quantity. The offer quantities of the sellers are taken from the generator capacities in test system, and the offer prices are sampled uniformly from $[10, 30]$ \$/MWh. The bid quantities of the buyers are taken from historical loads in PJM in 2017 [27], with their peak values scaled to the nominal loads in the test case, and the bid prices are sampled uniformly from $[20, 40]$ \$/MWh. In addition, an inelastic load, the peak value of which equals to 50% of the total generator capacity, is also added. System losses and line congestions are ignored in the WEM clearing problem, and only generation capacity limits are considered.

Assume the LSE under study serves 100 EUCs. The backlog rate η_t^c is sampled uniformly from $[0, 0.5]$. The newly generated incremental energy need ξ_t^c is simulated using historical incremental loads in PJM scaled by a value that is sampled uniformly from $[0.1, 2]$ MW, and added with a zero-mean Gaussian noise that has a scaled standard deviation (SD) of 0.1. The benefit function takes the following quadratic form:

$$\beta^c(e_t^c, d_t^c) = \kappa_t^c(e_t^c - d_t^c)^2 + \varsigma_t^c d_t^c,$$

where κ_t^c (in \$/MWh^2) is sampled from a Gaussian distribution with a mean of 10 and a SD of 1, and ς_t^c is sampled uniformly from $[20, 30]$ \$/MWh. The feasible set of the energy consumption is $\mathcal{D}_t^c = \{d_t^c \geq 0\}$.

Other parameters are set as follows: $T = 24$, i.e., one day is decomposed into 24 segments, and $\phi_t(x_1, x_2) = 5|x_1 - x_2|$, i.e., the LSE will loss \$5 if the aggregate energy consumption in the REM deviates from purchase energy quantity in the WEM by 1 MW. We create two scenarios, a winter scenario in which historical load data from PJM during January to March in 2017 are used, and a summer scenario in which historical load data from PJM during June to August in 2017 are used. In both scenarios, data from the first two months are used for training, while data for the last month are used for testing.

The neural networks are implemented using TensorFlow [28]. All hyperparameters, such as the number of layers in neural networks and the number of neurons in each layer, are chosen based on common practice recommended by the deep learning community [26], and are tuned using the training data.

B. Response Functions

The response functions are critical since they replace the actual environment during the learning process of the bidding and pricing policy, and also are used to determine the state in the MDP formulation. To illustrate the application of the response functions, we first generate a set of historical data of the WEM, i.e., $\{\boldsymbol{\omega}_\tau, \lambda_\tau, q_\tau\}$, using the WEM model in (1), and a set of historical data of the REM, i.e., $\{\nu_\tau, d_\tau\}$ using the EUC model in (4). When generating the data of the WEM, the bid quantities from the LSE under study are sampled uniformly from $[0, 80]$ MW, and the bid prices are sampled uniformly from $[20, 40]$ \$/MWh.

A neural network with 2 hidden layers, each consisting of 128 neurons are used as the bid response function. An L_2 regularizer with a scale of 0.01 is used. Rectified linear unit (ReLU) is used as the activation function for the two hidden layers and the output layer. Adam optimizer with a learning rate of 0.001 is adopted to train the neural network for 10000 steps. The performance of the response functions are measured by the mean and SD of the absolute error between the actual and predicted responses. Table I shows the mean and SD of the absolute error in the wholesale energy price under different orders of the bid response function. The mean wholesale energy prices of the training data in the winter scenario and the summer scenario are 22.98 \$/MWh and 23.72 \$/MWh, respectively, and those of the testing data in the winter scenario and the summer scenario are 22.63 \$/MWh and 23.40 \$/MWh, respectively. Note that a zero-order bid response function takes only information on the time as well as the bid when predicting the WEM clearing results. Both the mean and SD of the absolute error decrease as the order of the response function increases. Yet, the decrease is not significant when the order is great than 1 in both scenarios. Therefore, an appropriate order of the bid response function for this case would be $n_1 = 1$.

TABLE I
ABSOLUTE ERROR IN WHOLESALE ENERGY PRICE (IN \$/MWH).

		winter scenario			summer scenario		
	order	0	1	2	0	1	2
Train	mean	0.82	0.27	0.26	1.02	0.26	0.26
	SD	0.77	0.23	0.21	0.73	0.18	0.18
Test	mean	0.82	0.26	0.23	1.03	0.25	0.25
	SD	0.64	0.22	0.20	0.69	0.18	0.18

TABLE II
ABSOLUTE ERROR IN AGGREGATE ENERGY CONSUMPTION (IN MW).

		winter scenario				summer scenario			
	order	0	1	2	3	0	1	2	3
Train	mean	6.24	3.15	1.40	1.31	8.40	3.66	1.45	1.41
	SD	5.21	2.41	1.40	1.32	6.22	2.62	1.20	1.16
Test	mean	6.27	3.21	1.43	1.43	8.07	3.61	1.51	1.48
	SD	4.78	2.33	1.23	1.22	6.03	2.57	1.26	1.20

The neural network adopted for the price response function is similar to that for the bid response function except that the number of neurons in each hidden layer is 256 and the scale of the L_2 regularizer is 0.001. The neural network is trained with a learning rate of 0.0002 for 20000 steps. Table II shows the mean and SD of the absolute error in the aggregate energy consumption under different orders of the price response function. The mean aggregate energy consumptions in the training data in the winter and summer scenarios are 40.75 MW and 50.45 MW, respectively, and those of the testing data in the winter and summer scenarios are 38.16 MW and 47.08 MW, respectively. A zero-order price response function takes only information on the time and the price when predicting the aggregate energy consumption. Similar to the argument made for the bid response function, an appropriate order for the price response function would be $n_2 = 2$.

We emphasize that the appropriate order of response functions may vary from case to case, and need to be determined from the historical data following the procedures presented here. Based on learned response function, the state is

$$\boldsymbol{s}_t = (\lambda_{t-1}, q_{t-1}, d_{t-2}, \nu_{t-2}, d_{t-1}, \nu_{t-1}, t \bmod T).$$

C. Bidding and Pricing Policies

The pricing policy network and the critic network each has 2 hidden layers each with 128 neurons. ReLU is used as the activation function for all hidden layers. The output layer of the pricing policy network adopts the $\tanh$ function as the activation function, while that of the critic network does not use any activation function. An L_2 regularizer with a scale of 0.01 is used for the critic network. The learning rates for the pricing policy network and the critic network are $\alpha^a = 0.0001$ and $\alpha^c = 0.001$, respectively. Note that the bidding policy

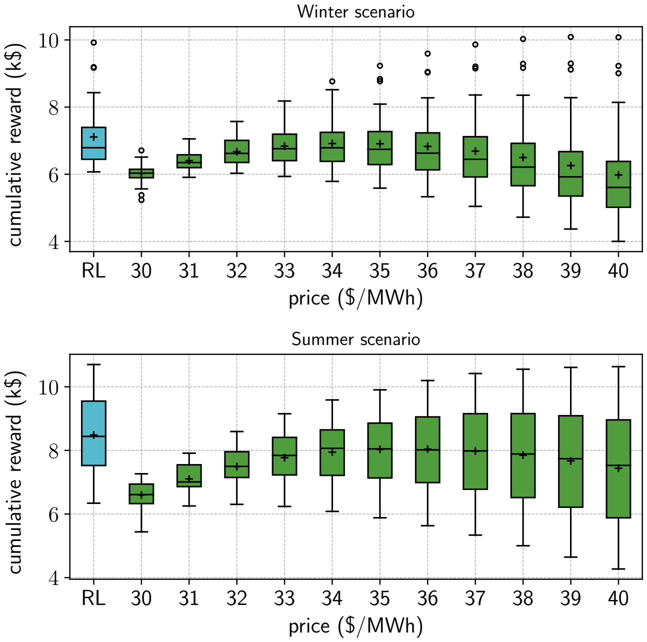

Fig. 4. Cumulative rewards under baseline policies and RL policy.

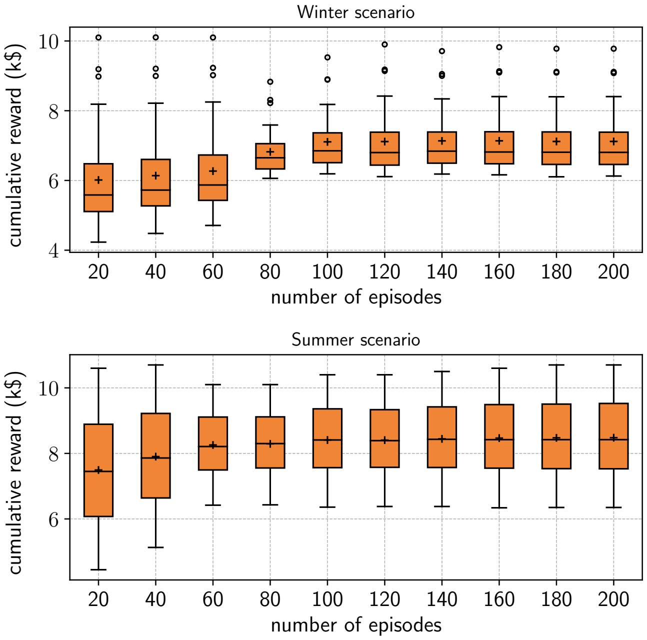

Fig. 5. Cumulative rewards under RL policy obtained under various training episodes.

network essentially the bid price to ν_t and the bid quantity to the estimated aggregate energy consumption obtained using the price response function φ. Therefore, there is no parameter for the bidding part needs to be trained. The minimum price is $\underline{\nu} = 20$ \$/MWh and the maximum price is $\overline{\nu} = 40$ \$/MWh. The update rate for the target networks is $\rho = 0.001$. The size of a mini-batch is chosen to be $m = 64$. The discount rate is $\gamma = 0.9$. The policy is trained over $M = 200$ episodes.

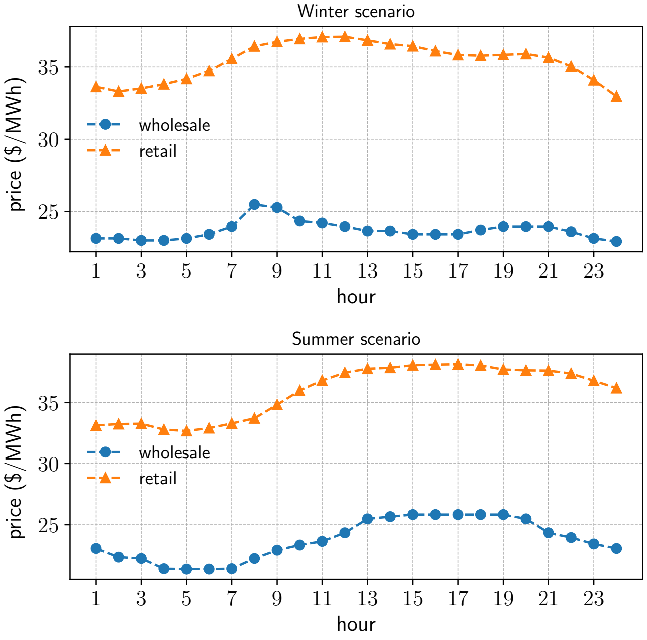

Fig. 6. Wholesale and retail energy prices under RL policy during a typical day.

We benchmark the proposed methodology with a baseline bidding and pricing policy which sets the price ν_t to a constant, and submits a bid price that equals to ν_t and a bid quantity that equals to the estimated aggregate energy consumption obtained using the price response function φ. Figure 4 presents a box-plot of the cumulative rewards, i.e., the sum of immediate rewards during a day, under the policy learned by the DDPG algorithm—referred to as the RL policy—and the baseline policies with various constant prices. The mean cumulative reward under the RL policy is higher than those under baseline policies in both two scenarios. Specifically, the mean cumulative reward under the RL policy in the winter scenario and the summer scenario are 7.111 k\$ and 8.485 k\$, respectively, while the highest mean cumulative reward under the baseline policies are 6.914 k\$ and 8.041 k\$, respectively. Figure 5 shows that impacts of training episodes on cumulative rewards. As can be seen from Fig. 5, the proposed methodology achieves good performance after 100 episodes, over which the performance improvement becomes relatively small.

The wholesale and retail energy prices under the RL policy during a typical day are shown in Fig. 6. It is obvious from Fig. 6 that the optimal retail energy price has a similar trend as the wholesale energy price, which makes sense since the cumulative reward depends on the difference of these two prices. The bid quantities and the aggregate energy consumptions under the RL policy and the baseline policy with a constant price of 35 \$/MWh during the same day is presented in Figs. 7 and 8, respectively. In addition, results for the case where the EUCs have direct access to the WEM, i.e., the LSE sets the retail price equal to the wholesale price (hence, the LSE is non-profit), are also plotted in Figs. 7 and 8. We make two observations for this particular simulation setup here. First,

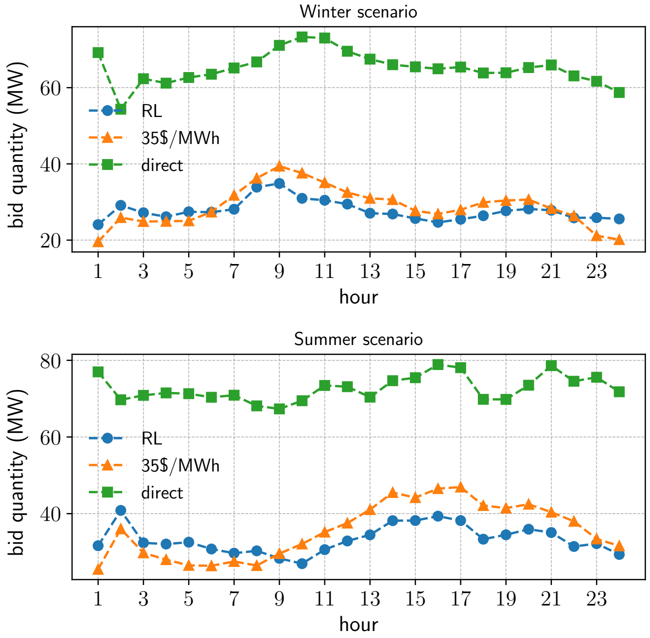

Fig. 7. Bid quantities during a typical day under the RL policy, the baseline policy with a constant price of 35 $/MWh, and the case where EUCs have direct access to the WEM.

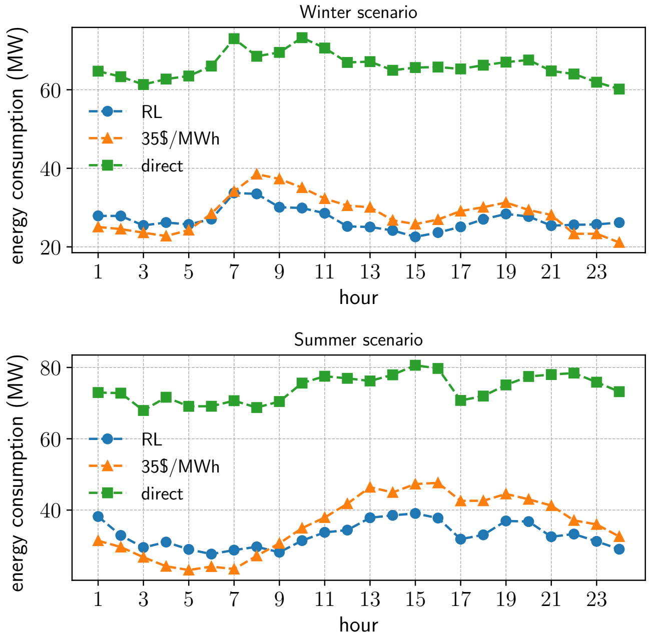

Fig. 8. Aggregate energy consumptions during a typical day under the RL policy, the baseline policy with a constant price of 35 $/MWh, and the case where EUCs have direct access to the WEM.

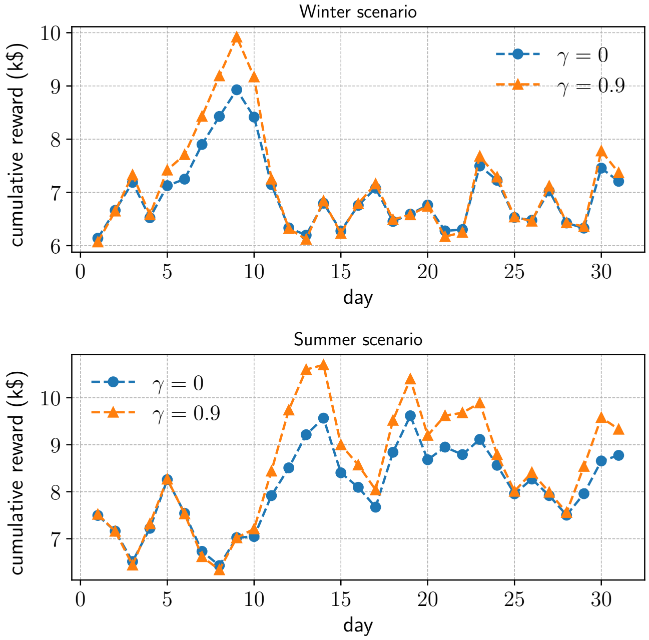

Fig. 9. Impacts of discount factor on cumulative rewards.

allowing the profit-seeking behavior of the LSE will lead to a situation in which the LSE would set a retail price that is (potentially much) higher than thed wholesale price such that its profit is maximized yet the total EUC energy consumption is greatly reduced, compared to the case where the EUCs have direct access to the WEM. Second, the aggregate energy consumptions under the RL policy has lower variance than under the baseline policy, which results into a smoother load curve. These phenomena are also observed in most of the days.

As discussed earlier, the consideration of the long-term behavior is beneficial, compared to the myopic decision making, in which no future rewards are taken into account. To illustrate this, we compare the cumulative rewards under the RL policy with $\gamma = 0.9$ and those under a myopic policy, or equivalently, the RL policy with $\gamma = 0$. Figure 9 shows that the RL policy with $\gamma = 0.9$ outperforms the myopic policy in both two scenarios. This indeed justifies the motivation of modeling the joint bidding and pricing problem as an MDP.

VI. Concluding Remarks

In this paper, we developed an MDP formulation for the joint bidding and pricing problem of the LSE, and applied a state-of-the-art DRL algorithm—the DDPG algorithm to solve it. Dynamical bid response and price response functions represented by neural networks are learned from historical data to model the WEM and the EUCs, respectively. These response functions explicitly capture the inter-temporal correlations of the WEM clearing results and the EUCs, and can be utilized to generate state transition samples required by the DDPG algorithm without any cost. Numerical simulation results show that the LSE can make more profit using the bidding and pricing policies learned via the proposed methodology, yet the aggregate energy consumptions may be significantly reduced, compared to the case where the EUCs have direct access to the WEM. An interesting phenomenon is that the more profitable bidding and pricing policies result in smoother aggregate energy consumptions.

There are several potential directions for future work. The first is to develop RL algorithms that incorporate risk management in the decision making process and construct bids and prices with profit guarantees, The second is to extend the proposed methodology in a multi-agent setting where all participants in the WEM, as well as participants in the REM, i.e., EUCs, also have learning capabilities and they may compete or cooperate with each other.

References

[1] F. Wang, H. Xu, T. Xu, K. Li, M. Shafie-Khah, and J. P. Catalão, "The values of market-based demand response on improving power system reliability under extreme circumstances," *Applied energy*, vol. 193, pp. 220–231, 2017.

[2] F. Rahimi and A. Ipakchi, "Demand response as a market resource under the smart grid paradigm," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 82–88, 2010.

[3] H. Xu, K. Zhang, and J. Zhang, "Optimal joint bidding and pricing of profit-seeking load serving entity," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5427–5436, Sept 2018.

[4] D. Fooladivanda, H. Xu, A. D. Domínguez-García, and S. Bose, "Offer strategies for wholesale energy and regulation markets," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7305–7308, 2018.

[5] H. Oh and R. J. Thomas, "Demand-side bidding agents: Modeling and simulation," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1050–1056, 2008.

[6] R. Herranz, A. M. San Roque, J. Villar, and F. A. Campos, "Optimal demand-side bidding strategies in electricity spot markets," *IEEE Trans. Power Syst.*, vol. 27, no. 3, pp. 1204–1213, 2012.

[7] X. Fang, Q. Hu, F. Li, B. Wang, and Y. Li, "Coupon-based demand response considering wind power uncertainty: a strategic bidding model for load serving entities," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1025–1037, 2016.

[8] A. J. Conejo, J. M. Morales, and L. Baringo, "Real-time demand response model," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 236–242, 2010.

[9] S. Yousefi, M. P. Moghaddam, and V. J. Majd, "Optimal real time pricing in an agent-based retail market using a comprehensive demand response model," *Energy*, vol. 36, no. 9, pp. 5716–5727, 2011.

[10] D. T. Nguyen, H. T. Nguyen, and L. B. Le, "Dynamic pricing design for demand response integration in power distribution networks," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3457–3472, 2016.

[11] B.-G. Kim, Y. Zhang, M. Van Der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2187–2198, 2016.

[12] A. Halder, X. Geng, P. Kumar, and L. Xie, "Architecture and algorithms for privacy preserving thermal inertial load management by a load serving entity," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3275–3286, 2017.

[13] M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement learning for electric power system decision and control: Past considerations and perspectives," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6918–6927, 2017.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[15] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[16] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. of IEEE International Conference on Robotics and Automation*, 2017, pp. 3389–3396.

[17] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau *et al.*, "An introduction to deep reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[18] G. Gajjar, S. Khaparde, P. Nagaraju, and S. Soman, "Application of actor-critic learning algorithm for optimal bidding problem of a genco," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 11–18, 2003.

[19] Z. Wen, D. O'Neill, and H. Maei, "Optimal demand response using device-based reinforcement learning," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2312–2324, 2015.

[20] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2149–2159, 2017.

[21] J. Zhang, C. Lu, J. Si, J. Song, and Y. Su, "Deep reinforcement leaming for short-term voltage control by dynamic load shedding in china southem power grid," in *Proc. of IEEE International Joint Conference on Neural Networks*, 2018, pp. 1–8.

[22] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *arXiv preprint arXiv:1807.10997*, 2018.

[23] H. Xu, H. Sun, D. Nikovski, K. Shoichi, and K. Mori, "Learning dynamical demand response model in real-time pricing program," in *Proc. of IEEE Power Energy Society Innovative Smart Grid Technologies Conference*, Feb. 2019, pp. 1–5.

[24] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[25] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *ICML*, 2014.

[26] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016.

[27] "PJM metered load data," http://www.pjm.com/markets-and-operations/ops-analysis/historical-load-data.aspx.

[28] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning," in *OSDI*, vol. 16, 2016, pp. 265–283.

Hanchen Xu (S'12) received the B.Eng. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2012 and 2014, respectively, and the M.S. degree in applied mathematics from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2017, where he is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering. His current research interests include control, optimization, reinforcement learning, with applications to power systems and electricity markets.

Hongbo Sun (SM'00) received his Ph.D. degree in electrical engineering from Chongqing University in Chongqing, China in 1991. He is currently a Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories in Cambridge, Massachusetts, USA. His research interests include power system operation and control, power system planning and analysis, and smart grid applications.

Daniel Nikovski (M'03) received his Ph.D. degree in robotics from Carnegie Mellon University in Pittsburgh, USA, in 2002, and is currently the group manager of the Data Analytics group at Mitsubishi Electric Research Laboratories in Cambridge, Massachusetts, USA. His research interests include artificial intelligence, robotics, machine learning, optimization and control, and numerical methods for analysis of complex industrial systems.

Shoichi Kitamura received the B.E. and M.E. degrees in biophysical engineering from Osaka University in 2000 and 2002, respectively, and the Dr. of Eng. degree in electrical engineering from Osaka University in 2013. He joined the Advanced Technology R&D Center, Mitsubishi Electric Corporation, Hyogo, Japan, in 2002, where he was engaged in research on the factory energy management system. At present, he is engaged in research on the smart grid and smart community-related technologies.

Kazuyuki Mori (M'98) received the B.E. and M.E. degrees in industrial administration from Tokyo University of Science in 1985 and 1987, respectively, and the Dr. of Eng. degree in electrical engineering from Osaka University in 1998. He joined Mitsubishi Electric Corporation, Amagasaki, Hyogo, Japan, in 1987 where he is currently engaged in research on discrete event systems, production scheduling, systems optimization, and energy solution. At present, he is a manager of the Advanced Technology R&D Center. He is also a Guest Professor at Osaka University and a Visiting Professor at Kyushu University. Dr. Mori is a fellow of the Institute of Electrical Engineers in Japan, and a member of the Society of Instrument and Control Engineers, and the Institute of Systems, Control and Information Engineers.

Hiroyuki Hashimoto received his B.E. and M.E. degrees in electrical engineering from Kyoto University in 1993 and 1995, respectively, and his Dr. Eng. degree from the University of Tokyo, Japan, in 2015. He joined Mitsubishi Electric Corporation in 1995 and has engaged in research on electric power grid analysis, stability control, power generation scheduling and smart grid. He is currently with Mitsubishi Electric Research Laboratories. His current research interests include optimization and its application in industrial systems.