Comparisons of Auction Designs Through Multiagent Learning in Peer-to-Peer Energy Trading

Zibo Zhao, Chen Feng, and Andrew L. Liu

***Abstract*—Distributed energy resources (DERs), such as solar panels, are growing rapidly and reshaping power systems. To promote DERs, utility companies usually adopt feed-in-tariff (FIT) to pay DER owners (aka prosumers) fixed rates for supplying energy to the grid. As an alternative to FIT, consumers and prosumers can trade energy in a peer-to-peer (P2P) fashion. In this paper, we focus on a P2P market using double auctions, in which the payoffs of energy consumers/prosumers are determined by their bids and auction mechanisms. Special features of a P2P energy auction, however, including zero marginal cost and publicly-known reserve prices, may invalidate many theories on auction design and hinder market development. We discuss the impacts of such features on four specific clearing mechanisms: *k*-double, Vickrey, McAfee and maximum volume matching (MVM). Furthermore, we propose an automated bidding framework based on multi-agent, multi-armed bandit learning, in which each agent only needs to utilize their own bidding history to determine how to bid in the next round through certain regret-minimizing algorithms. Numerical results show that the *k*-double and McAfee auction appear to perform better in terms of bidders' surplus. However, if the auctioneer also requires compensation, MVM can yield the most profit for the auctioneer.**

***Index Terms*—Peer-to-peer market, double auction, multi-agent systems, bandit learning.**

I. Introduction

DISTRIBUTED energy resources (DERs) are a vital part of a smart grid, as such resources can improve system reliability and resilience with their proximity to load, and promote sustainability, with the majority of DERs being renewable energy. To incentivize investments in DERs, two general approaches exist: non-market-based versus market-based. A widely used policy in a non-market-based approach is feed-in-tariff (FIT) (including net-metering). While effective in promoting DERs, it may create equity issues as consumers without DERs would face increased electricity rates to pay for distribution systems' maintenance costs. In a market-based approach, DERs can choose to participate wholesale electricity markets, as specified in the recent FERC Order 2222. To do so, however, an aggregator is needed to pool DER resources and bid into a wholesale market on behalf of DER owners, as energy output from an individual owner (such as a household) is too small; nor do the owners have the required expertise. In addition, sending electricity from widely dispersed locations over long distance to a bulk transmission system will incur significant energy losses. An alternative market-based approach is to have a local marketplace for consumers and DER owners, also referred to as prosumers, to directly trade energy, hence the so-called peer-to-peer (P2P) market. The actual rates that market participants pay or receive will fluctuate over time, reflecting the dynamic supply and demand conditions.

There are two prevailing mechanisms to match supply and demand in a P2P market: bilateral matching and double-auction. In a bilateral matching market, as described in [1], buyers and sellers can continuously post their demand bids and supply offers at a marketplace, and a clearing mechanism similar to a stock exchange can be used to match the supply and demand. A specific implementation of such an approach within a distribution network can be found in [2]. In a double-auction approach, buyers and sellers submit their price and/or quantity bids and offers to an auctioneer within a bidding time window for a certain operation period. In addition to the two matching-based approaches, there have been numerous works that use distributed optimization algorithms to integrate DERs, such as using the well-known alternating direction method of multipliers (ADMM). All the three approaches have been extensively surveyed in several literature reviews, including [3], [4]. In this work, we focus on the double-auction approach. Note that this is not an endorsement of the double-auction over other approaches, as each of the three approaches described above has their merits and disadvantages. Our goal here is to highlight several unique features of a P2P energy auction market, if the auction approach is chosen. We identify the impacts of such features to the market outcomes under different auction designs and propose a simulation framework, based on multi-agent, multi-armed bandit games, that can be used to compare the auction designs with repeated interactions among agents with bounded rationality.

Manuscript received 11 November 2021; revised 12 April 2022 and 23 May 2022; accepted 27 June 2022. Date of publication 13 July 2022; date of current version 22 December 2022. This work was supported in part by U.S. National Science Foundation under Grant CMMI-1832688 and Grant ECCS-2129631, and in part by U.S. Department of Energy under Grant DE-OE0000921. Paper no. TSG-01806-2021. *(Corresponding author: Andrew L. Liu.)*

Zibo Zhao was with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA. He is now with Google Ads, Mountain View, CA 94043 USA (e-mail: zibozhao.purdue@gmail.com).

Chen Feng and Andrew L. Liu are with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: feng219@purdue.edu; andrewliu@purdue.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2022.3190814.

Digital Object Identifier 10.1109/TSG.2022.3190814

Game theory has been a predominant tool in analyzing P2P trading, both for bilateral matching and double auctions. The theoretical works of matching can be dated back to [5], and a more recent account is in [6]. As stated in [6], all the theoretical works for matching-based P2P trading requires the acyclicity assumption; that is, no agent can be both a seller and a buyer. This is easily violated in a P2P energy market with prosumers. For double-auction-based P2P trading, games of incomplete information are the prevailing approach, as stated in [7] and surveyed in [8]. Since an auction-based energy market inherently involves repeated auctions and exogenous uncertainties (e.g., wind/solar availability), the most fitting equilibrium concept is the Perfect Bayesian Nash equilibrium (PBNE) [9] for dynamic games of incomplete information.

A PBNE consists of the collection of each player's strategy profile, which is a function that maps the entire history of the game to each player's feasible set of actions. The requirements for PBNE are too strong to be practical: each player needs to choose a strategy profile that yields the best expected payoff (given that other players choose their corresponding PBNE strategies) over all possible histories of the game; then all players need to update their beliefs' of other players' (unknown) payoff functions by the Bayes' rule through their own observations in each time period. Not only such strategy profiles are not computable (as it would require to find the best mapping over the functional space of all possible mappings, leading to an infinite-dimension optimization problem), nor are electricity consumers/prosumers in reality such sophisticated, which leads to the prevalent bounded rationality of market participants, such as no information of other players in the game. In fact, they may not even know their own valuation of energy production and consumption. This is especially true for renewable-based DERs, such as wind and solar, since they have significant investment costs but zero marginal costs. It is not clear how such resources should bid in a double auction. Furthermore, the participants' valuations are likely dependent, such as in a hot summer day, all buyers would value high of energy consumption for air conditioning. This feature would nullify the assumptions of many results in auction theory, which require independent valuation among bidders [10]. Last but certainly not the least, when buyers or sellers' bids are not cleared in an auction, the buyers have to buy energy from a utility at a utility rate; while the sellers have to sell to a utility at a fixed rate (such as the FIT). Such rates essentially become the reserve prices for buyers and sellers in the P2P auction, respectively, and such reserve prices are public information. Such information may lead the sellers to aggressively offer high prices, close to the utility rate, and the buyers to aggressively bid low, close to the sell-back rate. This may cause frequent unclearing of bids and offers and extreme clearing prices, either close to the utility rate or to the sell-back rate, depending on the supply and demand ratio.

There have been an increasing amount of works to study double auctions in a P2P energy market, as surveyed in [3], [4]. Here we focus on the works that are directly related to ours. In [11], a double auction is proposed for residential users, with the focus on biding with HVAC. It uses a predetermined demand curve (based on desirable temperature) to determine price and quantity bids for buyers; for sellers, they all just bid a flat curve at some prevailing market price. This model is further extended in [12], which propose a mechanism to implement a social choice function. However, these works focus on demand bids only, and the theoretical results require the knowledge of consumers' utility functions. Reference [13] considers explicitly a double-auction, but only uses an averaging mechanism to determine clearing prices; that is, add all buyers and sellers bid/ask prices and then divide by the total number of bidders. Such a mechanism would make every bidder possible to manipulate clearing prices. [14] implements a variant of the McAfee auction, which will also be discussed in this paper, and proposes an approximate dynamic programming (ADP) approach to help bidders bid. The reward function in the ADP is set to be the economic cost of prosumers, which can be ambiguous for zero-marginal cost resources. This is the view shared in [15], which is the closest to our works. Reference [15] recognizes several issues that are also emphasized in this paper, including the zero-marginal-cost issue, and numerically compares three auction mechanisms: uniform-pricing, Vickrey and pay-as-bid, all of which will be analyzed in this paper as well. However, [15] only lets sellers bid levelized investment costs of the resources they own. This neither is a strategy nor does it have theoretical justifications, as fixed costs should not be factored into operation/bidding decisions.

Despite the existing works, significant knowledge gaps remain before actual implementing a P2P market. Most importantly, few studies have focused on the repeated nature of local energy trading, and the potential of market participants to learn to overcome their bounded rationality. A notable exception is [16]. In [16], the authors consider repeated double auctions in a P2P energy market, very similar to the settings in this paper, through multi-agent reinforcement learning (MARL). The paper uses a centralized learning, decentralized implementation approach (named the multi-agent deep deterministic policy gradient (MADDPG) algorithm, as developed in [17]) to let the central auctioneer run the reinforcement learning (RL) algorithm on behalf of all the bidders and train a policy jointly for all the bidders, and then pass the agent-specific value function information to each bidder to implement their own strategies. In this sense, the MADDPG algorithm is still a centralized algorithm,[1] which is likely to encounter practical hurdles, such as the lack of a such sophisticated central entity and the lack of transparency of the central algorithm from the bidders' perspective. There have been recent works on consensus-based decentralized MARL algorithms, such as [19], [20]. However, whether such algorithms can be applied to repeated double auctions (or to any P2P energy trading setting) needs further exploration. Regardless, all the MARL-based approaches mentioned above suffer the drawback of scalability, as the maximum number of agents presented in numerical examples in the papers of [16], [17], [19], [20] is 20, which already takes a significant number of episodes to train a policy. In addition, it is unclear if any of the MARL

[1]It is well-known that a naive decentralized MARL algorithm by simply letting each agent run their own RL does not work (see [18]) in the sense that the multi-agent system will not exhibit any stable outcomes in general.

algorithm can handle a dynamic agent pool; that is, the number of agents may change over time.

Inspired by the work in [21], we propose an alternative framework to model multi-agent repeated interactions in double auctions. Instead of having each agent solving a reinforcement learning problem, we discretize the decisions and have each agent solve a multi-armed bandit problem. Specifically, each agent uses an algorithm to choose an action (referred to as an arm) from a finite number of available actions at each round of an auction. The collective actions of the agents determine the reward for each agent, who can compare it with the best-possible reward in hindsight, with the difference defined as regret. Each agent then seeks to find an algorithm to minimize their own cumulative regret. While regret-minimizing algorithms have been well-established for single-agent multi-armed bandit (MAB) problems, such as the famed Upper Confidence Bound (UCB) algorithm [22], multi-agent MAB games have only received attention recently, which is considered a special case of a MARL game, as surveyed in [23]. The multi-agent MAB game framework is a completely decentralized approach, as agents only need to use their own past bidding and reward information to make decisions, without needing to exchange information with any other agents; nor does it need any entity to run a centralized algorithm (other than, of course, an auctioneer to clear the market; though the clearing can be automated as well through a ledger-based system, such as blockchain). In addition, the MAB game framework is highly scalable, as it is built upon the mean-field concept; that is, each agent believes that the system is stabilized at a steady state such that their individual actions will not affect the steady state. It can also easily accommodate a dynamic agent pool.

In this paper, we demonstrate the applicability of the MAB-game approach in repeated games with incomplete information and bounded rationality, and use it to compare four specific double-auction mechanisms: k-double auction, Vikrey, McAfee, and maximum volume matching (MVM). Our contributions in this paper are threefold.

- On the modeling side, we present the detailed extension of a multi-agent MAB game from its general setting in [21] to the specific setting of a double-auction in an energy market.
- On the theoretical side (independent of the MAB game framework), we establish theoretical properties of the four auctions under the unique features of a P2P energy market. We show that the k-double auction and the MVM mechanism are not truth-revealing; while the specific setup of the Vikrey and McAfee auctions in our setting preserve their truthfulness property.
- On the simulation results and their market design implications, we show that while truthfulness is a desired theoretical result in general, when bidders have bounded rationality and do not know their own valuation, the best they could do is to be truthful with respect to the reserve prices, which would lead to bang-bang type of clearing prices. The clearing prices are also sensitive to the supply and demand ratio in a particular round of an auction. Such results may not be obtainable without the MAB-game framework, and we believe that these results are important for policy makers/market designers to know, as otherwise the outcomes of a double-auction-based P2P market may not bring the desired results to either consumers or prosumers.

The rest of the paper is structured as follows. Section II lays out the details of the MAB-game framework to study repeated double-auctions in a P2P energy market. Section III introduces four specific auction mechanisms. Their theoretical properties are discussed in Section IV. Numerical results are presented in Section V; while limitations of the current work and possible future research are discussed in Section VI.

II. Learning in Double Auctions

Consider an electricity distribution network. Without a marketplace, prosumers can only sell their generated energy to a utility or a DSO at some pre-defined FIT, denoted as P_{FIT}; similarly, consumers can only buy energy from a utility at the utility rate (UR), denoted as P_{UR}. Throughout this paper, it is assumed that $P_{FIT} < P_{UR}$.

A. Agents and Types

We consider three kinds of market participants: pure buyers, pure sellers, and prosumers. For the last group, their role is not fixed; namely, a prosumer can be either a buyer or a seller at any particular round of an auction, just not at the same time. More specifically, let $\mathcal{A}_b^h$ denote the set of buyers at round h of an auction (such as at a particular hour h in a particular day), and $\mathcal{A}_s^h$ be the set of sellers in the same round. Then $\mathcal{A}_b^h \cap \mathcal{A}_s^h = \emptyset$. Furthermore, let $\mathcal{A}^h := \mathcal{A}_b^h \cup \mathcal{A}_s^h$. Not only the sets $\mathcal{A}_b^h$ and $\mathcal{A}_s^h$ may change over h, so is the joint set $\mathcal{A}^h$. The changing agent sets are both to account for prosumers' altering positions and to reflect situations where some agents may leave the auctions (such as moving out of the local network) and new agents may join. We want to highlight this capability and flexibility of dealing with dynamic agents as a particular benefit of the learning-based approach.

In a double auction, both buyers and sellers need to decide their bid/ask prices and quantities of energy. As a starting point, we do not consider storage options in this work (and will discuss the challenges of considering storage and potential solutions in the conclusion section). Without storage, the energy quantities produced by DERs are likely not controllable, and hence we assume that any quantities minus self-consumption will be sold to the local market.[2] Consequently, agents only bid/ask energy prices. While the agents do not control quantities, their demand and output quantities from DERs are still stochastic, reflecting the generation variations from renewable resources. To account for agent heterogeneity, we use a generic variable y_i^h to denote the type of agent $i \in \mathcal{A}^h$ in a particular round h, which specifies what kind of market participant i is, as well as the distributions of their energy consumption and generation quantity at h. The set of all possible types for an agent i is denoted by Y_i.

[2]This is exactly the case of a grid-tied solar system, where the grid essentially serves as a battery.

B. Discrete Price Arms

Since the majority of DERs are solar and wind resources, we assume that the sellers' marginal costs are all 0.[3] Hence, any rate higher than P_{FIT} would be attractive to sellers; similarly, buyers would want rates lower than P_{UR}. In another words, any rate in the range of (P_{FIT}, P_{UR}) would be preferred by both the buyers and sellers. To implement a learning-based algorithm, we discretize the interval $[P_{FIT}, P_{UR}]$ into M elements, such as by a quarter or a dollar increment, and refer to each element $p(m) \in [P_{FIT}, P_{UR}]$, $m = 1, \ldots, M$, a price arm. At each round of an auction, each buyer/seller chooses a price arm to bid/ask. Since the zero marginal-cost of energy production is common-knowledge to all agents, as well as the de facto price ceiling (P_{UR}) and price floor (P_{FIT}), the agents (both buyers and sellers) try to learn how to choose the price arms in the repeated auction to maximize their rewards, with the rewards being explicitly defined in the next subsection.

C. Rewards

Conceptually, the (marginal) rewards for buyers in one-round of the auction is the difference between the prices they pay for one unit of energy and P_{UR}; similarly, the (marginal) rewards for sellers is the difference between the prices they are paid and P_{FIT}. To aid model development, it is more convenient to scale the agents' rewards between 0 and 1. To do so, we first define two benchmarking payoffs; that is, the lower and upper bound of an agent's payoff. (Note that for buyers, payoffs should be understood as payments to purchase energy. They are negative numbers in our modeling setup and hence, it is still that the bigger the payoff, the better for a buyer.)

Let q_i^h denote agent i's demand or generation at a round h of an auction. To ease the arguments, we drop the round (or time) superscript in the notation within this subsection, and it is understood that all discussions are within one round of the repeated auction. The quantity q_i is negative for a buyer and positive for a seller. For a buyer, the lower bound of the payoff is naturally to pay all of q_i at P_{UR}; for the upper bound, we define it to be $q_i \cdot P_{FIT}$, as no sellers would be willing to supply energy at a rate lower than P_{FIT}. (Note that q_i is negative for a buyer; hence $q_i P_{FIT} > q_i P_{UR}$ with the assumption that $P_{FIT} < P_{UR}$). For sellers, the lower and upper bounds are exactly reversed. To avoid discussing the buyers and sellers separately, we use the indicator function $\mathbb{1}_{\{\}}$ to define the lower and upper bound of payoff for any agent $i \in \mathcal{A}$ as follows:

$$\underline{\Lambda_i} = q_i \cdot \left[P_{UR} \cdot \mathbb{1}_{\{i \in \mathcal{A}_b\}} + P_{FIT} \cdot \mathbb{1}_{\{i \in \mathcal{A}_s\}}\right], \quad (1)$$

$$\overline{\Lambda_i} = q_i \cdot \left[P_{FIT} \cdot \mathbb{1}_{\{i \in \mathcal{A}_b\}} + P_{UR} \cdot \mathbb{1}_{\{i \in \mathcal{A}_s\}}\right], \quad (2)$$

where $\mathbb{1}_{\{i \in \mathcal{A}_b\}} = 1$ if agent i is a buyer and is zero otherwise. The definition for $\mathbb{1}_{\{i \in \mathcal{A}_s\}}$ is the same.

Now let Λ_i denote the actual payoff of agent i in a round of the auction. Throughout this paper, we assume that for the uncleared bid or ask quantities in an auction, they will be purchased by a utility company at P_{UR} or sold to a utility at P_{FIT}, respectively. Note that depending on the specific auction design, there can be partial clearing, meaning that for the same bidder, only part of their bid/ask quantity may be cleared. Hence, within any round, an agent's payoff consists of two components: the payoff from participating the auction (denoted as Λ_i^{au}), and the payoff from selling to or buying from the utility (denoted Λ_i^{ut}); that is,

$$\Lambda_i = \Lambda_i^{au} + \Lambda_i^{ut}. \quad (3)$$

More specifically, let q_i^{au} denote agent i's cleared quantity in an auction ($q_i^{au} \geq 0$ for $i \in \mathcal{A}_s$, and $q_i^{au} \leq 0$ for $i \in \mathcal{A}_b$), and p_i^{au} be the corresponding unit price determined by an auction for agent i to receive (if a seller) or to pay (if a buyer). Then $\Lambda_i^{au} = p_i^{au} \cdot q_i^{au}$. Similarly, we have that $\Lambda_i^{ut} = p_i^{ut} \cdot q_i^{ut}$, where $p_i^{ut} = P_{FIT}$ if $i \in \mathcal{A}_s$, and $p_i^{ut} = P_{UR}$ if $i \in \mathcal{A}_b$, and q_i^{ut} denotes the uncleared energy quantity for agent i.

With the above notations, we can now define the normalized reward as $\pi_i := (\Lambda_i - \underline{\Lambda_i})/(\overline{\Lambda_i} - \underline{\Lambda_i})$. It is straightforward to see that if agent i's auction clearing price p_i^{au} is in $[P_{FIT}, P_{UR}]$, we have $\Lambda_i \in [\underline{\Lambda_i}, \overline{\Lambda_i}]$ and hence $\pi_i \in [0, 1]$. If agent i is not cleared in an auction at all, then $\Lambda_i = \underline{\Lambda_i}$ and $\pi_i = 0$.

Note that in our simulation, we do allow sellers to ask above P_{UR}, and buyers to bid below P_{FIT}. This is to represent the case that either the auction agents have bounded rationality (in the sense that they do not recognize the practical price ceilings/floors) or the agents are greedy, as if they want to try their luck to earn a higher payoff in one round of the auction. Our numerical results later show that indeed the auction clearing price can go beyond P_{UR} or below P_{FIT} in certain rounds. (But such clearing prices cannot be sustained in the repeated auctions as the counterpart agents can quickly learn to reverse the course.) To prevent the normalized reward to go outside the range of [0, 1], we expand the definition of π_i for all $i \in \mathcal{A}$ as follows:

$$\pi_i = \begin{cases} 1 \cdot \mathbb{1}_{\{i \in \mathcal{A}_b\}} + 0 \cdot \mathbb{1}_{\{i \in \mathcal{A}_s\}}, & \text{for } p_i^{au} < P_{FIT} \\ \left(\Lambda_i - \underline{\Lambda_i}\right)/\left(\overline{\Lambda_i} - \underline{\Lambda_i}\right), & \text{for } P_{FIT} \leq p_i^{au} \leq P_{UR} \\ 0 \cdot \mathbb{1}_{\{i \in \mathcal{A}_b\}} + 1 \cdot \mathbb{1}_{\{i \in \mathcal{A}_s\}}, & \text{for } p_i^{au} > P_{UR}. \end{cases} \quad (4)$$

D. Policies and Regret Minimization

Once the reward of each agent is defined, each agent learns from the history of the game to decide what to do in the next round. The history of the game for agent i is recorded in the state variables. For agent i at the h-th round of the repeated auction, the state variable, denoted by z_i^h, is a vector of $2M$ elements, with M defined earlier as the number of price arms. The first M elements record the number of times that each arm $m \in \{1, \ldots, M\}$ has been chosen by agent i; while the second M elements denote the average rewards (from the first round to the current round h) associated with each arm m. Let $\mathcal{Z}_i^h$ denote the set of all possible states for agent i at round h. A policy, also referred to as a strategy or simply an algorithm, is a mapping from $\mathcal{Z}_i^h$ to a probability distribution over the arms.

Note that an agent's reward at each round of the auction does not only depend on which arm they choose, but also depends on the collective actions of all agents. The outcomes of all agents' actions in a specific round h can be represented

[3]Note that we actually do not need any assumptions on marginal costs. Assuming zero marginal costs just makes it easier to gain insights from the numerical results.

by a quantity referred to as the population profile, which is the histogram of the arm choices by all the agents at round h. Let $\mathbf{f}^{|\mathcal{A}^h|}(m)$ denote the population profile for a specific arm m, where $|\mathcal{A}^h|$ denotes the number of total agents at round h. Then

$$\mathbf{f}^{|\mathcal{A}^h|}(m) = \frac{1}{|\mathcal{A}^h|} \sum_{i=1}^{|\mathcal{A}^h|} \mathbb{1}\left\{\sigma_i\left(z_i^h\right) = m\right\}, \tag{5}$$

where the function $\mathbb{1}\{\sigma_i(z_i^h) = m\} = 1$ if agent i chooses the arm m at round h and is zero otherwise.

A key technical difficulty with a multi-agent MAB game is that the population profile is both random and does not follow a stationary distribution. The essence of a mean-field approach is to assume that each agent believes that the population profile is in a steady state, denoted as $\boldsymbol{f} = \{f(m)\}_{m=1}^{M}$. Under a stationary population profile, we can define the best possible reward in one round for agent i as $\pi_i^*(\boldsymbol{f}) = \max_{m=1,\ldots,M} \mathbb{E}[\pi_i(\boldsymbol{f}, m)]$, where $\pi_i(\boldsymbol{f}, m)$ denotes the reward of agent i for choosing price arm m under the population profile $\boldsymbol{f}$. Let σ_i be an arbitrary policy for agent i to choose an arm at each round. Over D rounds, we use $\Gamma_{\sigma_i}(D, m)$ to denote the number of times that price arm m has been chosen by the policy σ_i. Then we can define agent i's cumulative regret under the policy σ_i as below:

$$\Delta_{\sigma_i} = \pi_i^*(\boldsymbol{f}) \cdot D - \sum_{m=1,\ldots,M} \mathbb{E}\left[\pi_i(\boldsymbol{f}, m) \cdot \Gamma_{\sigma_i}(D, m)\right]. \tag{6}$$

The regret Δ_{σ_i} in Eq. (6) is the expected cumulative loss due to the fact that the policy σ_i does not necessarily always pick up the optimal price arm under the stationary population profile $\mathbf{f}$, which is unknown to the agent. A policy σ_i is called a *no-regret* policy if the regret in Eq. (6) satisfies: $\Delta_\sigma / D < R(D, M)$, where the function R is $o(1)$ with respect to D, and M is the total number of arms. Regret-minimizing policies for a single-agent MAB problem have been well-studied (under the assumption that the distribution of the exogenous uncertainty is stationary). One popular policy is the UCB algorithm [22], whose idea is simple: at the D-th round of the game, an agent chooses the arm $\hat{m}$, with

$$\hat{m} \in \text{argmax}_{m \in \{1,\ldots,M\}} \left\{ \overline{\Pi}_i(m) + \sqrt{\frac{2\ln(D)}{\Gamma_i(D, m)}} \right\}, \tag{7}$$

where $\overline{\Pi}_i(m)$ represents the average reward for agent i when the arm m is chosen up to round D, and $\Gamma_i(D, m)$, as defined earlier, is the number of times that the arm m has been chosen up to round D. The above formulation reflects the trade-off between exploitation (choosing an arm with the highest payoff) and exploration (trying as many arms as possible).

E. Specific Implementation

In this section, we put things together and describe the detailed implementation of the MAB-game framework from a single agent's perspective, with an agent being broadly defined as a single household, a smart building or even a microgrid. Each round of a double-auction is assumed to be organized for some time ahead; that is, the auction is to determine the price and quantity of traded energy for a specific time in the future. For example, this can be day-ahead, which will be very similar to the wholesale energy market operation, hour-ahead, or even 15-minute-ahead. At each round, an agent is to choose a specific price from the discretized range of $[P_{FIT}, P_{UR}]$ to bid or ask, where the discretized price range has M elements. Each agent has a $2M$-dimension vector that is assumed to be saved on their local device (such as the home energy management system (HEMS)): $[\Gamma_i(D, 1), \ldots, \Gamma_i(D, M), \overline{\Pi}_i(1), \ldots, \overline{\Pi}_i(M)] \in \Re^{2M}$. As discussed in the previous subsection, the first M elements indicate how many times each price arm has been chosen; while the second M elements represent the (average) cumulative regret corresponding to each price arm m. For prosumers, however, since they can be either sellers or buyers in a particular round of an auction, depending on their net energy positions, they have to maintain two sets of the $2M$ vectors, one set to be used when the prosumer will be a buyer in the coming round of the auction, and the other set for the case that they will be sellers. With the $2M$ vector, each agent can choose a bandit learning algorithm, such as using the formula in (7). More bandit learning algorithms can be found in [24]. As mentioned in Section II-A, we assume that agents do not choose a quantity to bid; they simply bid for all their energy demand or offer all their net energy production. Once all the bid/ask prices are collected (by either an auctioneer or an open-ledger-based system) with the corresponding quantities, a clearing mechanism (to be discussed in Section III) will match the supply and demand bids to determine how much each agent will pay or get paid at what quantity (note that there can be partial clearing of an agent's bid quantities). Consumers with uncleared demand will then buy from their utility company, and sellers with uncleared supply offer will sell to the utility. The auction then moves to the next round.

We want to emphasize here some key benefits of the learning-based approach in a repeated game, including minimal information required for each agent, policy heterogeneity, and policy automation. As seen in (7) (and in other policies as well), only the vector of state variables is needed to implement a policy, and all such information is private (that is, with the agent i index). No other agents' information is needed; nor does the private information need to be shared with others. Implementing a specific policy usually requires very simple calculations, and no sophistic optimization is required. Bandit learning policies can also be programmed into electronic devices, such as HEMS, and can be automated. In addition, the overall framework does not require each agent to use the same policy. In all of our numerical results, agents are randomly assigned to use different polices and market outcomes, based on numerical experiments, are very robust with respect to policy heterogeneity.

F. Network Constraints

As pointed out in [3], P2P energy trading includes two layers: virtual layer and physical layer. Our proposed work clearly focuses on the virtual layer, which is to provide a platform for energy buyers and sellers to have equal access to create and execute orders. The physical layer, on the other hand,

refers to the physical network (such as the location distribution network) to facilitate the actual delivery of electricity from the cleared sellers to buyers. The physical layer should no doubt be an integral part of any P2P energy market, as transactions at the virtual layer may lead to network infeasibility and hence are invalid. While we do not consider the physical layer in this work, we want to point out that it is not necessarily the limitation of the MAB-game framework. For example, after collecting the bids/asks, the auctioneer can run a power flow model to ensure physical feasibility. If the bids would lead to infeasibility or reliability concerns, the auctioneer (likely a utility company or a distribution system operator (DSO)) can just reject all bids/asks and re-solicit them. The agents can modify their algorithms a bit to assign a big penalty to the arms they chose in this situation so that jointly the infeasible situation is much less likely to occur in the future (although this would work better when agents can bid both price and quantity than just bidding price only). This is the idea we used in another line of our work [25], in which we used the MAB-game approach to model consumers' response to real-time prices of wholesale energy markets. There we explicitly embedded a system operator's optimal power flow problem into the MAB-game and the results showed that agents can even learn to alleviate congestion at different time and locations to reduce their electric bills. These being said, incorporating distribution network constraints in a double auction is not a trivial matter. We refer to this as a future research direction.

G. Discussion of Convergence

In [21], a mean-field approach is proposed to address the theoretical issue of an MAB game, which contains a large number of agents, and each agent believes that the system is stabilized at a mean field steady state (MFSS) such that their individual actions will not affect the steady state. Under this belief and the assumption that each agent's reward is continuous with respect to the population profile, [21] shows the existence of the MFSS. In addition, they show that if the reward function is also Lipschitz continuous with respect to the population profile, then the MFSS is unique; as the number of agents grows to infinity, the dynamics of the multi-agent system will converge to the unique MFSS. These results are really desirable, as it shows the robustness of the approach (with a unique MFSS to converge to) and its scalability, which overcomes the computational issues of other MARL algorithms as mentioned in the introduction section that may not scale well with more agents. Unfortunately, the key continuity assumption does not hold in the auction situation, as each agent's reward depends on if they win or lose the auction, which can change abruptly with a small change in population profile. Despite the lack of theoretical results, our numerical experiments do show the emergence of a steady-state population profile in all the simulations. We also want to point out that the reward function is usually required to have a finite support in a single-agent bandit learning algorithm, such as the UCB algorithms [22]. Without the finite support, the regret-minimization property of the established bandit learning algorithm may not hold. Consequently, the numerical outcomes of the multi-agent MAB-game in our specific setting may not be as good as what we obtained (in terms of convergence) without the specific approach we designed to convert agents' energy bills or payments into rewards defined on [0, 1], as in (4). Hence, we consider our framework a non-trivial extension of the multi-agent MAB game into repeated double-auctions, with such a framework's theoretical properties certainly call for further research.

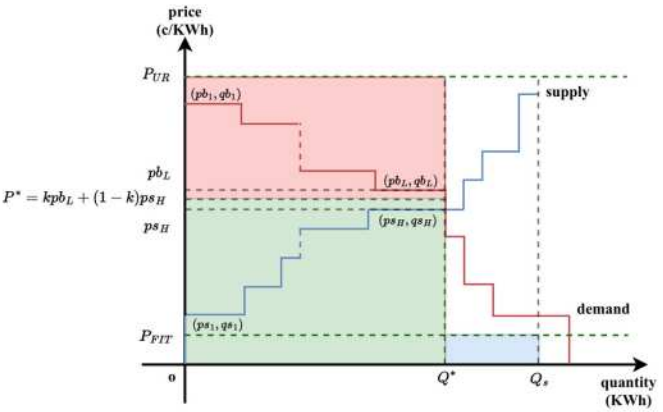

Fig. 1. k-double auction.

III. Double Auction Designs

In this section, we lay out four specific auction designs to realize a multi-unit, double-side auction. While some of the auction designs are well-known, such as the k-double auction and the Vickrey auction, we do want to provide the complete details of the clearing mechanisms (such in the case of over-supply or over-demand, or there are ties in the bids) in the specific context of a P2P energy market so that there is no ambiguity in terms of market clearing.

A. k-Double Auction

To start with, all bids/asks are sorted by their bidding/asking prices, which result in the stair-wise demand/supply curves as shown in Fig. 1. At the quantity of the intersection point Q^*, where the aggregate demand and supply meet, the last step of the cleared bids from buyers are denoted as (pb_L, qb_L), and highest cleared asks from sellers are denoted as (ps_H, qs_H). The k-double auction represents a whole class of auctions with similar designs, with the market clearing price $P^* := kpb_L + (1-k)ps_H$, where $k \in [0, 1]$. For the clearing mechanism to work under any supply/demand conditions, we define two rules. **Rule 1 –** If $\sum_{l=1}^{L} qb_l \geq \sum_{h=1}^{H} qs_h$ (referred to as over-demand), all the asks with $h \leq H$ are cleared, and sell all their quantities at price P^*; all the bids with $l \leq L$ are cleared, and the clearing price for all the buyers is also P^*. However, the quantity for each cleared buyer l is not the quantity they bid; it is scaled back by the amount of over-demand. Specifically, for each buyer $l \leq L$, their cleared quantity is $qb_l - (\sum_{l=1}^{L} qb_l - \sum_{h=1}^{H} qs_h)/L$. As mentioned earlier, uncleared supply is assumed to sell to a utility at P_{FIT}; while uncleared demand is to buy at from the utility

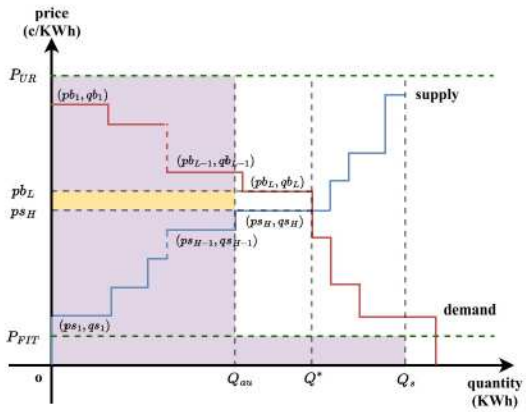

Fig. 2. A Vickrey-variant double auction market.

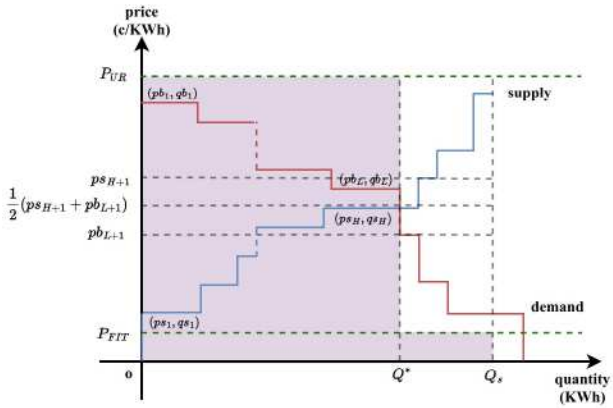

Fig. 3. McAfee double auction market (Case A).

at P_{UR}. This is the same across all the rules and all the auctions below, and hence will not be stated again. **Rule 2 –** If $\sum_{l=1}^{L} qb_l \leq \sum_{h=1}^{H} qs_h$ (referred to as over-supply), all the buyers' bids with $l \leq L$ are cleared and buy all their demand bids qb_l at the price P^*; all the supply asks with $h \leq H$ are cleared and sell at the price P^*, but their cleared quantities are scaled back to $qs_h - (\sum_{h=1}^{H} qs_h - \sum_{l=1}^{L} qb_l)/H$. Note that we add the scaling-back rule here to ensure that no sellers or buyers are arbitrarily favored by the clearing mechanism. This is inspired by the rule in [26] and may not be present in the common uniform-pricing-based auction design in the literature.

To aid the comparison among different auction designs, we define the concept of auction surplus as the total surpluses of buyers and sellers, and denote it as $\widehat{S}$. We also denote the auctioneer's surplus by S_{au}. In a k-double action, since the market clearing price P^* is the same for both buyers and sellers, the auctioneer always has zero surplus. For buyers, we define their surplus as $\sum_{i \in \mathcal{A}_b} (\Lambda_i - \underline{\Lambda}_i)$, with Λ_i and $\underline{\Lambda}_i$ being defined in (1) and (3), respectively. The buyers surplus in a k-double auction is the upper rectangle (of dark pink color) in Fig. 1. For sellers, since we assume that all sellers' marginal costs are zero, their surpluses equal price times quantity, with the price being set by an auction if the supply ask is cleared, or being P_{FIT} if the ask is not cleared. In Fig. 1, suppliers surplus is the green area plus the blue area. With the above definition of buyers and sellers' surplus, we can write their surplus in a simple way for the k-double auction as $\widehat{S}^{k-double} = P_{UR} \cdot Q^* + P_{FIT} \cdot (Q_s - Q^*)$.

B. Vickrey Variant Double Auction

Vickrey auction [27] is one of the most famed auction designs, mainly due to its truth-revealing property (which we will discuss in the next section). The original Vickrey auction has been extended to a double-side version for multiple-units goods in [26]. Here we refer to it as the Vickerey-variant auction. Similar to the k-double auction, all bids/asks are sorted by price as shown in Fig. 2,

The detailed setup of the Vickrey-like auction is documented in [26] and is omitted here. However, we do want to point out that the key difference between the Vikrey and the k-double auction is that the Vikrey auction clears only up to $(H-1)$-th sellers and $(L-1)$-th buyers, and their clearing prices can be different (such as pb_L for cleared buyers and ps_H for cleared sellers in Fig. 2.)

Based on the clearing rules, the total cleared quantity in the Vickrey-like auction (denoted by Q^V) is $Q^V := min(\sum_{l=1}^{L-1} qb_l, \sum_{h=1}^{H-1} qs_h)$. Regarding to auction surplus, the auctioneer now has a positive surplus due to the difference between the cleared buying and selling price (as illustrated in the yellow shaded area in Fig. 2). Mathematically, $S_{au}^V = (pb_L - ps_H) \cdot Q^V$. The total agents' surplus in a Vickrey-like auction is: $\widehat{S}^V = [(P_{UR} - pb_L) + ps_H] \cdot Q^V + P_{FIT} \cdot (Q_s - Q^V)$, as represented by the light purple area in Fig. 2.

C. McAfee's Double Auction

This mechanism is a variant of the Vickrey-like auction, suggested by McAfee [28]. Consider $P_0 := \frac{1}{2}(pb_{L+1} + ps_{H+1})$, as shown in Fig. 3. There are two cases – Cases A: if $P_0 \in [ps_H, pb_L]$, then the auction is the same as the k-double auction, with a uniform clearing price of $P^* = P_0$. The first L buyers and H sellers are cleared. Case B: if $P_0 \notin [ps_H, pb_L]$, the mechanism becomes the Vickrey-like auction, with up to the $(L-1)$-th buyer and $(H-1)$-th seller being cleared.

D. Maximum Volume Matching (MVM) Double Auction

While the above three mechanisms differ in the details of how market clearing prices and quantities are determined, they all build upon the same idea of stacking supply and demand to find the intersection point. With a drastically different idea, [29] proposes a pay-as-bid auction whose sole purpose is to maximize the cleared volume of the traded goods. Such an auction is referred to as the MVM auction. The idea of maximizing traded volume is appealing in our context, as it could help promote the penetration of renewable energy. The mechanism of the market clearing process is as follows. We start with the regular stacked supply asks/demand bids curves as in Fig. 1. Then the supply stack is flipped horizontally around the vertical axis, as illustrated in Fig. 4.

The flipped supply stack is then shifted rightward towards the demand curve until that any part of the two curves touch the first time. The distance from the origin to the right end of the flipped supply curve after shifting, denoted by Q^{MVM}

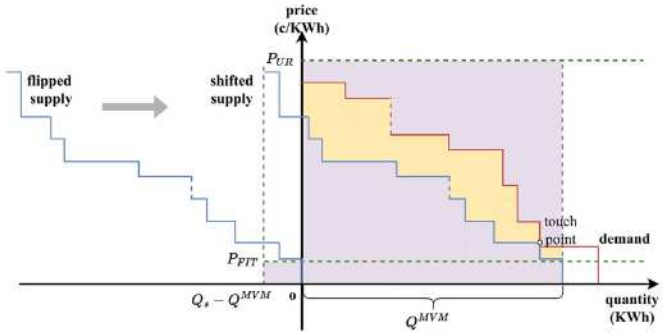

Fig. 4. A maximum volume matching auction market.

and illustrated in Fig. 4, is exactly the maximum trading volume that the auction can achieve. Consequently, Q^{MVM} amount of demand quantities of the highest bids are cleared; similarly, the same amount of the lowest ask quantities from suppliers are cleared. The cleared bids/asks are then matched in an ascending order with respect to bid/ask prices. Let $\mathcal{C}_b$ and $\mathcal{C}_a$ denote the set of cleared bids and asks, respectively. The total surplus of all agents can be written as $\widehat{S}^{MVM} = \sum_{l\in\mathcal{C}_b}(P_{UR} - pb_l)\cdot qb_l + \sum_{h\in\mathcal{C}_s} ps_h \cdot qs_h + P_{FIT}(Q_s - Q^{MVM})$, as represented by the light purple area in Fig. 4. The auctioneer's surplus (represented by the yellow area in Figure 4) can be written as $S_{au}^{MVM} = \sum_{l\in\mathcal{C}_b}(pb_l \cdot qb_l) - \sum_{h\in\mathcal{C}_a}(ps_h \cdot qs_h)$.

IV. Analysis of the Auction Mechanisms

In this section we will analyze the theoretical properties of the four auction designs, with focuses on strategy-proofness and budget balance. While the main focus of this paper is to develop a multi-agent learning framework to implement a double-side auction, which does not rely on any of the theoretical properties, the purpose of the discussions here is to highlight the special settings of a P2P energy auction and their consequences. One thing to note is that the theoretical properties discussed below are only for a single-round auction, which may not hold in a repeated setting [30]. The task of establishing theoretical results for repeated double auction is very daunting, which again makes the multi-agent learning-based framework a valuable approach.

To make this paper stand-alone, we first present the definitions of the various well-established concepts in auction theory and mechanism design. To ease the presentation, we let n denote the total number of buyers and sellers, and again drop the round superscript h.

Let A denote a set of alternatives. In our setting, $A = \{(p_i; q_i){:}i = 1, \ldots, n\}$, where agent i can choose the price p_i to bid or ask while the quantity q_i is assumed to be fixed. The preference of each agent i is modeled by a valuation function $v_i{:}A \to \mathcal{R}$, where $v_i \in V_i$, with V_i being the set of possible valuation functions for player i. We first give the formal definition of a mechanism.

Definition 1 [31, Definition 9.14]: A ***mechanism*** is a social choice function $f : V_1 \times \cdots \times V_n \to A$ and a vector of payment functions $p_1, \ldots, p_n$, where $p_i : V_1 \times \cdots \times V_n \to \mathcal{R}$ is the amount that agent i pays or receives.

Definition 2 [31, Definition 9.15]: For each $i = 1, \ldots, n$. Let $v_{-i} \coloneqq (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$ be the vector with the i-th component removed, and $V_{-i} \coloneqq \Pi_{j\neq i}^n V_j$ be the Cartesian product of the sets $V_1, \ldots, V_n$ without V_i. A mechanism $(f, p_1, \ldots, p_n)$ is called ***strategy-proof*** (also known as incentive compatible or truthful) if for every agent i, every $v_i' \in V_i$ and every $v_{-i}' \in V_{-i}$, $v_i(a) - p_i(v_i, v_{-i}') \geq v_i(a') - p_i(v_i', v_{-i}')$, where $a \coloneqq f(v_i, v_{-i}')$ and $a' \coloneqq f(v_i', v_{-i}')$.

Note that the price $p_i(v_i, v_{-i}')$ is positive for buyers and negative for sellers. The above definition can be understood as follows: while v_i represents agent i's true valuation, the agent may claim other (non-truthful) valuations v_i'. If a mechanism is strategy-proof, then the social choice f, together with the corresponding payment functions, will ensure that an agent will not be better off in terms of the net utility $v_i(a') - p_i(v_i', v_{-i}')$, if they do not reveal their true valuation.

A big issue with the above definition in our specific context is that consumers (and prosumers) do not really know their valuations of electricity consumption (and generation with zero-marginal costs). Such valuations also change over time. For example, it is more valuable to consume energy in a hot summer day than in a calm spring night. In addition, all consumers' valuation will be higher than P_{UR}, as otherwise they would choose not to use electricity at all. However, if they all truthfully reveal their valuation in an auction, the auction clearing price will be higher than P_{UR}, which would make consumers worse-off and the auction pointless.

While many existing works assume that consumers maximize their utility functions (valuations minus costs) to make decisions, we do not believe that such is a reasonable assumption.[4] The learning-based framework, as described in the previous section, does not use agents' valuation at all.

Without using agents' true valuation, the next best goal for an auction design is strategy-proof with respect to reservation price, as proposed in [26], which means that each agent will truthfully report their reservation price. However, while reservation prices are private information in a typical auction (such as bidding on eBay), they are clearly public information in a local P2P energy market, with P_{UR} and P_{FIT} being the buyers and sellers reservation prices, respectively. We will show below that in certain auction designs, such public information will result in the bang-bang type of market clearing prices; that is, the clearing prices will be either P_{UR} or P_{FIT}, depending on the total supply and demand quantities.

To further compare the outcomes of the four auctions, we introduce an additional concept of budget balance [26], which states that an auction is *strongly budget balanced* if the payments from cleared buyers and sellers always sum to zero exactly. If the sum is always nonnegative, it is *weakly budget balanced*.

In the following, we will analyze the above-defined concepts for the four auction mechanisms.

[4]We understand that the valuation also includes consumers' preference of comfortable level, such as the temperature level of an air conditioner. We will address this issue when we discuss the limitations and extensions of the current work.

A. *k*-Double Auction

We first show that the mechanism of k-double auction is not strategy-proof with respect to reserve prices.

Assumption 1: All buyers and sellers truthfully submit their quantities to buy or sell.

Assumption 2: The reservation prices of buyers and sellers P_{FIT} and P_{UR}, respectively, are both public information.

Proposition 1: Consider a single round k-double auction with $k \in [0, 1]$, where buyers and sellers only bid prices. Assume that Assumption 1 and 2 hold. Furthermore, let $\widehat{Q}_s$ and $\widehat{Q}_b$ denote the total supply and demand quantities (not just cleared bids) in the auction, respectively. If the relationship between $\widehat{Q}_s$ and $\widehat{Q}_b$ is also public information (that is, $\widehat{Q}_s \geq \widehat{Q}_b$ or $\widehat{Q}_s \leq \widehat{Q}_b$), then the k-double auction is not strategy-proof with respect to reservation prices.

Proof: If all agent bids truthfully with respect to the reservation price, then the market clearing price in the k-double auction will be $kP_{UR} + (1-k)P_{FIT}$. For the case where $\widehat{Q}_s \leq \widehat{Q}_b$, consider a seller g: if all other sellers ask P_{FIT}, and all buyers bid P_{UR}, then seller g can ask P_g. All the agent's ask quantities will still be cleared, since $\widehat{Q}_s \leq \widehat{Q}_b$. Then the market clearing price would be $kP_{UR} + (1-k)P_g$, strictly greater than $kP_{UR} + (1-k)P_{FIT}$, and hence, seller g's utility is strictly higher. Similar arguments can be made for buyers when $\widehat{Q}_s \geq \widehat{Q}_b$. ■

When $k = 0$ or 1, Assumption 2 actually can be dropped, meaning that even without knowing the actual supply and demand situation, the agents still do not have incentives to bid the reserve price. The proof is a simple extension of the above argument and is omitted here.

The non-strategy-proofness of the k-double auction is not surprising, as in such a mechanism, the marginal bidders (namely, the lowest bids and the highest asks that are cleared in an auction) set the market clearing prices, which gives agents incentives to manipulate the clearing prices in their favor.

While Proposition 1 is about the property of the auction mechanism, in the following we study from the perspective of the agents, and discuss what bidding strategies can lead to an ex-post Nash equilibrium, as defined below.

Definition 3 [31, Definition 9.22]: A profile of strategies $s_1, \ldots, s_n$ of n agents is an ex-post Nash equilibrium if for all $i = 1, \ldots, n$, all types $y_i \in Y_i$, and all feasible actions a_i' of i, we have that $\pi_i(y_i, s_i(y_i), s_{-i}(y_{-i})) \geq \pi_i(y_i, a_i', s_{-i}(y_{-i}))$, where π_i is i's utility function.

In essence, an ex-post Nash equilibrium requires that $s_i(y_i)$ is the best response to $s_{-i}(y_{-i})$ for all possible y_{-i}, which is the collection of other agents' types.

Proposition 2: Under the same context and assumptions in Proposition 1, with a given $k \in [0, 1]$, we have the following.

(1) If $\widehat{Q}_s > \widehat{Q}_b$, all agents bidding/asking P_{FIT} is an ex-post Nash equilibrium.
(2) If $\widehat{Q}_s < \widehat{Q}_b$, all agents bidding/asking P_{UR} is an ex-post Nash equilibrium.
(3) If $\widehat{Q}_s = \widehat{Q}_b$, and let $\widetilde{P}$ be a given price in $[P_{FIT}, P_{UR}]$. Then all agents bidding/asking $\widetilde{P}$ is an ex-post Nash equilibrium.

Proof: The proofs for all the three situations are similar; hence, we only show the proof of case (3) here. In (3), the strategy for any agent (buyer or seller) is that $s_i(y_i) = \widetilde{P}$, for all $i \in \mathcal{A}$, and all $y_i \in Y_i$. With this strategy, all buyers' and sellers' quantities are cleared (since $\widehat{Q}_s = \widehat{Q}_b$). Now assume that a particular buyer j chooses $pb_j \neq \widetilde{P}$. If $pb_j > \widetilde{P}$, buyer j's bid will be cleared, but the market clearing price is still $\widetilde{P}$, and buyer j's utility is the same as bidding $\widetilde{P}$. If $pb_j < \widetilde{P}$, j's bid will not be cleared, since all sellers ask $\widetilde{P}$. Then buyer j's utility is zero, strictly less than bidding $\widetilde{P}$. The arguments for the seller side are the same. Hence, the strategy $s_i(y_i) = \widetilde{P}$, for all $i \in \mathcal{A}$, and all $y_i \in Y_i$ is an ex-post Nash equilibrium. ■

Note that under situation (1) and (2) in Proposition 2, the market outcomes are susceptible to the bang-bang type results; that is, either the buyers or the sellers will reap all the benefits, depending on the total supply versus total demand, leaving the other parties of zero benefits. Under situation (3), the result in Proposition 2 does not appear to be useful, as $\widetilde{P}$ can be anything between P_{FIT} and P_{UR}, and there is no way for the agents to agree upon a common point a priori. However, the public information of the two reserve prices for buyers and sellers, respectively, provide a focal point in a k-double auction; that is, $\widetilde{P} = (P_{FIT} + P_{UR})/2$. This is indeed what we observe in our simulations, along with the outcomes as predicted by Proposition 2 under situation (1) and (2).

Regarding the other property, the k-double auction is strongly budget-balanced because the cleared selling and buying quantities are equal, and the clearing price is the same for buyers and sellers. This means that the auctioneer's surplus is always zero in a k-double auction, which can be a desirable outcome in some cases, but can be a downside here. As the role of an auctioneer in a P2P energy market is likely played by a utility or a DSO, they may require payments to provide such service. While a double auction may run on a distributed ledger system (aka blockchain) without a central auctioneer, physical delivery of the cleared energy still needs access to distribution networks, which entail maintenance and repair costs. If such costs need to be recouped from the auction process, then a weakly budget-balanced mechanism, such as the Vickrey variant auction, can be a choice; or some non-market-based approach must be implemented to cover the additional costs.

B. Vickrey and McAfee Double Auction

The key difference between a k-double auction and a Vickrey-like auction is that the marginal winners in the later do not affect clearing prices. Hence, the strategy-proofness of a Vickrey auction still holds in a multi-unit, double auction setting, as proved in [26], and it is weakly budget-balanced. While the truthfulness of an auction is usually a desired property, it is actually the opposite in a P2P energy market. This so because the goal of a local energy trading market is different than a traditional auction. In a traditional auction, the objective is to allocate resources to people who value them the most. In a local energy market, everyone needs electricity, and the goal is not about efficient allocation. Instead, the P2P

market is to help buyers and sellers achieve more favorable rates than P_{UR} and P_{FIT}, respectively.

While the theorem predicts that in our specific setting, the outcomes of the Vickrey-variant double auction will be always P_{UR} for buyers and P_{FIT} for sellers, we will see that this is not always the case in our numerical results. This highlights the point that all the theoretical results in this section are only established for a one-shot auction, and may not be true in a repeated game.

For the McAfee auction, its strategy-proofness for a single-unit good is provided in [28], and the idea is the same as the Vickrey auction; that is, the marginal winners do not affect clearing prices. Note that in Case A of the McAfee auction as described in Section III-C, even though the clearing mechanism is similar to the k-double auction, the clearing price is determined by the $(L+1)$-th buyer and $(H+1)$-th seller; while only up to the L-th buyers and the H-th sellers are cleared, hence achieving the strategy-proofness. In terms of budget balance, under Case A, the McAfee auction will result in the same clearing price for buyers and sellers. Hence, it is strongly budget-balanced; while under Case B, it is the same as the Vickrey-variant, which is weakly budget-balanced. Hence, overall the McAfee auction is weakly budget-balanced.

C. Maximum Volume Matching Double Auction

The MVM double auction is not strategy-proof (with respect to reservation price). By Definition 2, it suffices to find one particular instance under which an agent has an incentive to be not truthful. One such instance is that when all other buyers bid P_{UR} except buyer i, and all sellers ask P_{FIT}. Suppose that the total supply is greater than total demand. If buyer i bids any price strictly between P_{UR} and P_{FIT}, the quantity allocated to the agent remains the same and the price buyer i pays is lower than P_{UR}, since the MVM is pay-as-bid. As a result, buyer i will get strictly better utility of not bidding P_{UR}. Hence, the MVM double auction is not strategy-proof with respect to reservation price.

On the other hand, since the buying price is always no lower than the selling price for each matching (as shown in Fig. 4), the auctioneer's payoff is always non-negative, which means that the MVM double auction is weakly budget-balanced.

V. Numerical Simulations

A. Input Data

1) Decision Epochs and Temporal Resolution: We consider daily auctions with an hourly temporal resolution. More specifically, one round of an auction happens in a specific hour, such as 9 - 10 AM, which is repeated daily. Other hours are considered as different auctions, and agents will learn different strategies. In our simulations, we perform seven auctions in a day, representing the seven hours between 9 AM and 4 PM, and we run for 365 days.

2) P_{UR}, P_{FIT} and the Decision Space: P_{UR} and feed-in tariff P_{FIT} are fixed throughout the simulation, and are set at $P_{UR} = 11$ ¢/KWh and $P_{FIT} = 5$ ¢/KWh. For price bids/asks, as mentioned in Section II-C, we expand the range of bidding prices to [0, 14 ¢/KWh], and discretize the range with 1 ¢/KWh increment. This is to further account for agents' bounded rationality or aggressive bidding strategies.

3) Energy Consumption, Supply and Types of Agents: We simulate three groups of agents: 1,000 pure consumers, 1,000 pure suppliers, and 500 prosumers. For the last group, the net position of their supply and demand in a particular hour determines their roles in the corresponding round of the auction; namely, in each round, a prosumer will be a seller if they have excess energy, and a buyer if their self-produced energy is not sufficient to meet their own non-flexible demand. For DER generation, we assume that it is all from rooftop photovoltaic (PV) panels. The energy demand and generation of each buyer and seller are all randomly generated based on the System Advisor Model (SAM) National Renewable Energy Laboratory (NREL).[5] The details of how such data are generated, along with the data we used in our simulation, are available on Github.[6] In addition, at each round, each agent has a 0.005 of being regenerated independently. The learning history of the agent will be cleared to all-zero vector after regeneration.

B. Individual Agent's Bandit Learning Algorithm

In our simulation, each agent has an equal probability of using one of the three algorithms to choose a specific arm in each round: UCB1, UCB2, and ϵ-greedy. Once an algorithm is chosen, the agent will use the same algorithm throughout the repeated auction till regeneration. A regenerated agent again will have equal probability to choose from one of the three algorithms. The detailed codes of the three algorithms for our simulation are available at the same Github location.

C. Numerical Results

This subsection reports the simulation results of the four different auction designs. Note that for the k-double auction, we only consider the case where $k = 0.5$ and refer it as $k05$ double auction in this subsection.

In the simulation, for each auction (such as the auction for 9 - 10 AM on each day), we run four times with different random seeds, with each time consisting of 365 rounds of the auction. The codes are written in Python and run on a PC with Windows 10 OS, Intel Core i7-6700K processor, and 16 GB RAM. The time of one particular run (i.e., one time of the 365 rounds) of the k-double, Vickrey-variant, McAfee and MVM auction are 281, 285, 293, and 276 seconds, respectively. The time for the other three runs are similar. We first show results corresponding to one particular hour, 9 - 10 AM. The observed trends of the outcomes are exactly the same in the other hours. The solid lines in Fig. 5, 6, and 7 represent the average of the four simulation runs; while the shades in the figures represent the values within one standard deviation of the mean.

Fig. 5 compares the cleared quantities from the four different auction mechanisms. It can be seen that the Vickrey auction results in lower cleared quantities than $k05$-double and McAfee. This is expected as the Vickrey mechanism design

[5]Details can be found at https://sam.nrel.gov.

[6]https://github.com/feng219/MAB_Algorithm

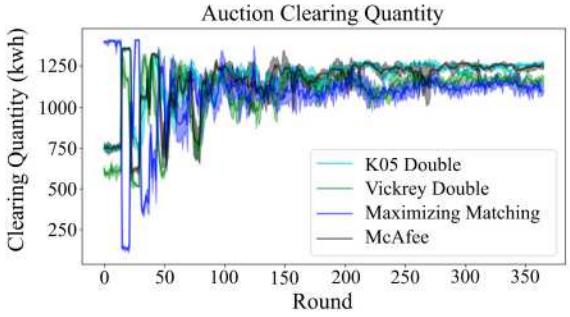

Fig. 5. Clearing quantity in the 9 - 10 AM auctions.

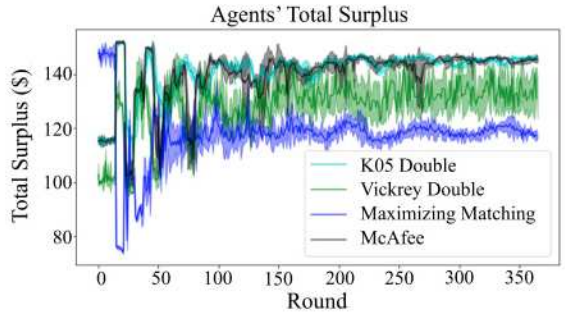

Fig. 6. Agents total surplus in the 9 - 10 AM auctions.

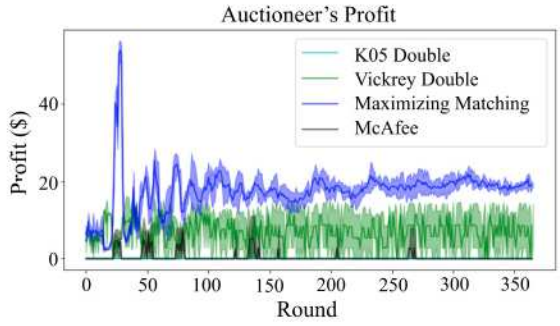

Fig. 7. Auctioneer's profit ($) in the 9-10 AM auctions.

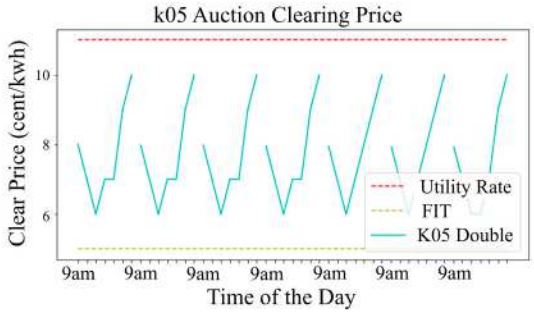

Fig. 8. k05 auction clearing price (consecutive 7 days).

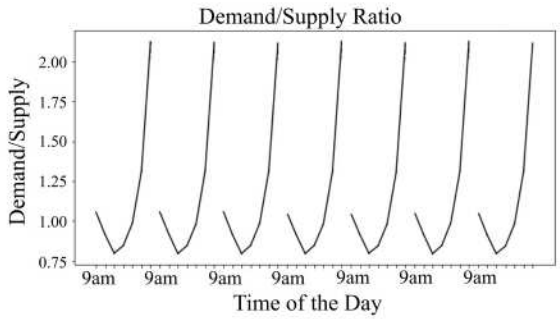

Fig. 9. Demand/supply ratio (consecutive 7 days).

sacrifices some traded volume (as it only clears up to the $(L-1)$-th buyer and $(H-1)$-th seller) to achieve strategy-proofness. What is surprising is that the cleared quantity of the MVM auction, which is designed (and proved in [29]) to maximize the traded volume, is also lower than k05-double and McAfee auction. The reason is as follows: in a single-round auction, with everything being equal, the MVM auction will be able to reach the theoretical upper bound of the maximum cleared quantity. However, in a dynamic setting with the learning-based approach, agents will learn different things in different auction designs. Hence, at the beginning of each round, the assumption that "everything being equal" no longer holds, which results in the observed outcomes that on average, the cleared quantity in MVM auction can be less than other auction designs. The cleared quantities between k05 double and McAfee auctions are very similar.

Next we compare the total surplus. Fig. 6. It can be seen in Fig. 6 that the total surplus of the Vickrey auction is notably less than the k05 double auction, exactly for the same reason as why the cleared quantity in the Vickrey auction is less. Since the McAfee auction is a hybrid of the k-double and the Vickrey auction, agents in the McAfee auction can likely learn the surplus differences between the k-double and the Vickrey auction through repeated interactions, and will make the McAfee auction outcome the same as the k-double auction most of the time. The total surplus is the lowest in the MVM auction, which likely is due to the surplus transfer from agents to the auctioneer, as shown in Figure 7.

Fig. 7 shows the auctioneer's surplus from the four auction mechanisms. Clearly the MVM auction results in the most auctioneer surplus. The Vickrey auction also results in positive auctioneer's payoff as buyers pay more than what sellers receive. The k-double auction, regardless k's value always results in zero auctioneer surplus. For McAfee auction, again, due to its hybrid nature, the auctioneer does have positive surplus over the course of the repeated auction, but the surplus is significantly less than the Vickrey auction.

Based on the numerical results, it appears that the k-double auction and the McAfee auction are the better performing mechanisms, so long as that an auctioneer's payoff is not a concern. To have a more complete picture, we present the clearing prices from the k05 auction in chronological order over seven days. All the results shown below are taken from the later rounds of the simulation when the market outcomes appear to stabilize. Note that the clearing prices are taken from just one of the four simulation runs, not the average of the four runs, as all the runs exhibit the same patterns.

The pattern of the auction clearing price closely resembles that of the demand/supply ratio in the market, as shown in Fig. 9.

Fig. 10. k05 auction clearing price (9 AM) over the rounds.

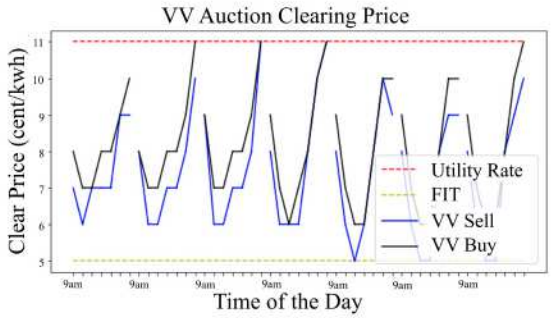

Fig. 11. Vickrey-variant auction clearing price over 7 days.

In our simulations, the demand/supply ratio is close to 1 at 9 AM, as shown in Fig. 9. When it is around noon time, however, the supply is more than the demand (due to high PV generation and low demand at residential home, as people are out at work), and the corresponding clearing price in the $k05$ auction gets close to P_{FIT}. In late afternoons, as the PV production winds down and residential demand picks up, the clearing price approaches to P_{UR}. These two situations are exactly as what predicted in Proposition 2. To have a closer look of the market outcomes when the demand/supply ratio is 1, we present the clearing prices for the 9 - 10 AM auction of the $k05$ auction over the entire simulated horizon in Fig. 10.

As seen in Fig. 10, the undesirable bang-bang type of clearing prices (that is, either P_{UR} or P_{FIT}) did appear at the initial rounds of the auction. As the learning progresses, the market clearing prices tend to stabilize at around 8 ¢/kwh, which is exactly half way between P_{UR} (11 ¢/kwh) and P_{FIT} (5 ¢/kwh). This is consistent to what we speculated when discussing the results of Proposition 2. We consider this market outcome fair as it splits the total surplus equally between buyers and sellers.

The outcomes of the McAfee double auction are very similar to the $k05$ auction. The clearing prices of the Vickrey are shown in Fig. 11. The Vickrey auction outcomes exhibit similar trends following the demand/supply ratio in a day, but the clearing prices are more extreme than the $k05$ auction when demand and supply are imbalanced. This is so because the averaging mechanism in determining the clearing price in a $k05$ auction helps alleviate the extreme prices. The clearing prices of the MVM auction cannot be shown in a picture, as every cleared bids/asks pay/receive different prices. The quantity-averaged clearing prices in the MVM auction do exhibit similar patterns to the other three auctions.

VI. Conclusion and Future Research

In this work, we proposed a multi-agent MAB-game approach to help automate the bidding strategies for agents participating a P2P energy market. This approach also helps study complicated games, such as repeated games with incomplete information, where theoretical results are either scarce or of little practical use. The approach has shown to be very useful in numerically comparing market designs, which can be a handy tool for policy makers to test their market design before actual implementation. The framework is also flexible as it does not require agents to know their utility functions, can easily incorporate heterogeneous agents, and requires minimal storage and computing power on the agent side.

Independent of the MAB-game, we studied theoretical properties of four double-auction designs in the context of a P2P energy market, which presents distinct features of all zero-marginal-cost supplies and publicly known reserve prices for both buyers and sellers. We showed that such features can be undesirable as they may lead to bang-bang type clearing prices, depending on the demand/supply ratio and auction mechanism. These results also highlight the needs of sophisticated simulation framework that can capture the essence of repeated auctions with a large number of agents with bounded rationality, as theories need to be tested to see if they can indeed emerge from agents' repeated interactions.

The current work only means to be a starting point for the general topic of decentralized multi-agent games and their applications in better utilizing DERs in a decentralized fashion. It has several notable limitations and can certainly benefit from future research. First, the presented work does not consider physical network constraints, which should be essential for any P2P market design to be implementable in the real-world. We point out that the auction mechanism can be used to pre-commit resources some time ahead, such as an hour ahead, exactly the same way as the day-ahead market in a wholesale market. After each clearing, a utility or a DSO can run power flow equations to determine if the cleared bids are physically feasible. If not, the past round of the auction can be re-run, and each cleared agent will update their rewards of the last round to zero. Granted that such an approach still cannot address real-time feasibility issues, which are likely to be dealt with in a completely different framework. In addition, specific rules or mechanisms need to be designed to compensate power losses and distribution network maintenance costs.

Another notable limitation of the MAB-game approach is that it cannot easily handle time-linkage decisions, such as injection/withdraw decisions for energy storage. We are currently developing a multi-agent reinforcement learning framework to include energy storage. Along this direction, we can also consider thermal load (HVAC) and consumers' preference of comfort (such as reflected by temperature settings). Such works will be reported in follow-up papers.

Last, but certainly not the least, privacy and cybersecurity issues should also be at the center of any local energy market design. One particular strength of our approach is that only bids are needed to submit to an auctioneer, while no private information needs to be shared or communicated. However,

the central auctioneer can become a single-point failure if it is compromised by cyber-attacks. In addition, further research is needed to study the robustness of the MAB game approach and auction designs in the event of malicious bidders, such as those consumers/prosumers who are hacked to send out misleading bids.

REFERENCES

[1] Y. Liu, L. Wu, and J. Li, "Peer-to-peer (P2P) electricity trading in distribution systems of the future," *Electricity J.*, vol. 32, no. 4, pp. 2–6, 2019.

[2] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized P2P energy trading under network constraints in a low-voltage network," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5163–5173, Sep. 2019.

[3] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-peer trading in electricity networks: An overview," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3185–3200, Jul. 2020.

[4] J. Guerrero, D. Gebbran, S. Mhanna, A. C. Chapman, and G. Verbič, "Towards a transactive energy system for integration of distributed energy resources: Home energy management, distributed optimal power flow, and peer-to-peer energy trading," *Renew. Sustain. Energy Rev.*, vol. 132, Oct. 2020, Art. no. 110000.

[5] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[6] J. W. Hatfield and S. D. Kominers, "Matching in networks with bilateral contracts," *Amer. Econ. J. Microecon.*, vol. 4, no. 1 pp. 176–208, 2012.

[7] D. Friedman, "The double auction market institution: A survey," in *The Double Auction Market: Institutions, Theories, and Evidence*, D. Friedman and J. Rust, Eds. London, U.K.: Routledge, 1993, pp. 3–25.

[8] D. Friedman and J. Rust, *The Double Auction Market Institutions, Theories, and Evidence: Proceedings of the Workshop on Double Auction Markets Held June, 1991 in Santa Fe, New Mexico*, 1st ed. New York, NY, USA: Routledge, 1993.

[9] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[10] R. P. McAfee and J. McMillan, "Auctions and bidding," *J. Econ. Literature*, vol. 25, no. 2, pp. 699–738, 1987.

[11] J. C. Fuller, K. P. Schneider, and D. Chassin, "Analysis of residential demand response and double-auction markets," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2011, pp. 1–7.

[12] S. Li, W. Zhang, J. Lian, and K. Kalsi, "Market-based coordination of thermostatically controlled loads—Part I: A mechanism design formulation," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1170–1178, Mar. 2016.

[13] M. Khorasany, Y. Mishra, and G. Ledwich, "Auction based energy trading in transactive energy market with active participation of prosumers and consumers," in *Proc. Aust. Universities Power Eng. Conf. (AUPEC)*, 2017, pp. 1–6.

[14] D. Kiedanski, D. Kofman, J. Horta, and D. Menga, "Strategy-proof local energy market with sequential stochastic decision process for battery control," in *Proc. IEEE Power Energy Soc. Innovat. Smart Grid Technol. Conf. (ISGT)*, 2019, pp. 1–5.

[15] C. Gerwin, R. Mieth, and Y. Dvorkin, "Compensation mechanisms for double auctions in peer-to-peer local energy markets," *Current Sustain. Renew. Energy Rep.*, vol. 7, pp. 165–175, Oct. 2020.

[16] D. Qiu, J. Wang, J. Wang, and G. Strbac, "Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 2913–2920.

[17] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Assoc., 2017.

[18] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*. Cham, Switzerland: Springer, 2021, pp. 321–384.

[19] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5872–5881.

[20] C. Qu, S. Mannor, H. Xu, Y. Qi, L. Song, and J. Xiong, "Value propagation for decentralized networked deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY, USA: Curran Assoc., 2019.

[21] R. Gummadi, R. Johari, S. Schmit, and J. Y. Yu. "Mean Field Analysis of Multi-Armed Bandit Games." [Online]. Available: http://dx.doi.org/10.2139/ssrn.2045842 (Accessed: Aug. 11, 2016).

[22] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 235–256, 2002.

[23] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity," 2017, *arXiv:1707.09183*.

[24] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," 2014, *arXiv:1402.6028*.

[25] Z. Zhao, A. L. Liu, and Y. Chen, "Electricity demand response under real-time pricing: A multi-armed bandit game," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2018, pp. 748–756, doi: 10.23919/APSIPA.2018.8659687.

[26] P. Huang, A. Scheller-Wolf, and K. Sycara, "Design of a multi-unit double auction E-market," *Comput. Intell.*, vol. 18, no. 4, pp. 596–617, 2002.

[27] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *J. Finan.*, vol. 16, no. 1, pp. 8–37, 1961.

[28] R. P. McAfee, "A dominant strategy double auction," *J. Econ. Theory*, vol. 56, no. 2, pp. 434–450, 1992.

[29] J. Niu and S. Parsons, "Maximizing matching in double-sided auctions," Feb. 2013, *arXiv:1304.3135*.

[30] B. F. Hobbs, M. H. Rothkopf, L. C. Hyde, and R. P. O'Neill, "Evaluation of a truthful revelation auction in the context of energy markets with nonconcave benefits," *J. Regulatory Econ.*, vol. 18, no. 1, pp. 5–32, 2000.

[31] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

Zibo Zhao received the B.E. degree in automation of control from the Huazhong University of Science and Technology, Wuhan, China, and the M.S. degree in electrical engineering from the Columbia University, New York City, and the Ph.D. degree in industrial engineering from Purdue University, West Lafayette, IN. He is currently a Senior Software Engineer with Google, Mountain View, CA.

Chen Feng received the B.B.A. degree from the Department of Economics and Finance, City University of Hong Kong, Hong Kong, and the M.S. degree from the Department of Applied Mathematics and Statistics, Johns Hopkins University, Maryland, USA. He is currently pursuing the Ph.D. degree with the School of Industrial Engineering, Purdue University. His research interests include game theory, mechanism design, reinforcement learning, and their applications in the energy system.

Andrew L. Liu received the B.S. degree in applied mathematics from the Beijing Institute of Technology, Beijing, China, in 2000, and the Ph.D. degree in applied mathematics and statistics from the Johns Hopkins University, Baltimore, MD, in 2009. He is currently an Associate Professor with the School of Industrial Engineering, Purdue University, West Lafayette, IN. Before joining Purdue, he was a Senior Associate with the Wholesale Power Group, ICF International, Fairfax, VA. His research interests include optimization, game theory and their applications in power systems, smart grid, and environmental policy analysis.