A Self-Governed Online Energy Management and Trading For Smart Micro/Nano-Grids

Milad Latifi, Amir Rastegarnia, Azam Khalili, Wael M. Bazzi, and Saeid Sanei, *Senior Member, IEEE*

***Abstract*—Joint energy consumption and trading management is still a major challenge in smart (micro-) grids. The main goal of solving such problems is to flatten the aggregate power consumption-generation curve and increase the local direct power trading among the participants as much as possible. Here, an inclusive formulation for energy management and trading of a Micro/Nano-grid (M/NG) is proposed. Subsequently, a holistic solution to jointly optimizing the internal energy consumption management and external local energy trading for a smart grid including several M/NGs is provided. As the problem is computationally intractable, the proposed approach involves three hierarchical stages. Firstly, a game-theoretic online stochastic energy management model is provided with a reinforcement learning solution by which the M/NGs can schedule their power consumptions. Secondly, an effective incentive-compatible double-auction is formulated by which the M/NGs can directly trade with each other. Thirdly, the central controller develops an optimal power allocation program to reduce the power transmission loss and the destructive effects of local energy trading. The simulation results validate the efficiency of the proposed framework.**

***Index Terms*—Double-auction, energy consumption management, energy trading, micro/nano-grids, reinforcement learning.**

I. Introduction

IT is well known that traditional power systems are unable to efficiently respond to the growing demand for energy [1]. This motivates the study and adoption of smart Micro/Nano-grids (M/NGs), which are autonomous small-scale power supply networks with distributed generation (including both conventional and renewable energy generators and storage units such as plug-in electric vehicles) and consumption (the local electrical loads from residential, commercial, and industrial consumers). The M/NGs are capable of meeting the growing energy demand by sustaining the renewable energy resources (RERs) in a reliable, efficient, and economical manner [2].

As the electric power cannot be stored in a large scale, a significant part of the efforts in the smart grids are to match the power supply trend to the load demand trend and fully utilize the available capacity of the energy sources [3]. This is challenging because the load demand is highly time-varying and hard to predict. Besides, emerging the RERs under the outstretch of the use of distributed energy generation policies, results in volatility and unpredictability of the supply side. This is the point of weakness and prevents wide investment on the renewable power generation and increases the number of M/NGs. As a result, the power system has a low load factor and is underutilized most of the time, while it is necessary to increase the generation capacity to supply the high demand at a short time peak demand [4].

To sustain the RERs and benefit from these green resources, it is necessary to provide an effective joint supply and demand sides energy management strategy between the prosumers (**pro**ducer-con**sumer**) capable of local direct energy sharing. There are various research directions for the energy consumption scheduling and trading approaches [5]–[18].

A. Related Work

A dynamic pricing and energy consumption scheduling program in the micro-grid was investigated in [5], where the service provider acts as a broker between the utility company and customers by purchasing electric energy from the utility company and selling it to the customers. In order to overcome the challenges of implementing such program under various sources of uncertainties, reinforcement learning algorithms was developed. Wang and Huang studied the interactions among interconnected autonomous micro-grids in [6], and developed a joint energy trading and scheduling strategy. Another energy trading framework based on the repeated game was proposed in [7], that enables each micro-grid to individually and randomly choose a strategy with probability to trade the energy in an independent market so as to maximize his/her average revenue. In [8]it is shown that the on-site wind power generation of high-rise buildings can potentially support all the electric vehicles in the city. Considering that the charging demand of EVs usually does not align with the uncertain wind power, the coordination of electric vehicle charging with the locally generated wind power in a micro-grid of buildings using a Markov decision process was investigated.

In [9], energy trading between smart grid prosumers and a grid power company was studied. The problem was formulated as a single-leader, multiple-follower Stackelberg game between the power company and multiple prosumers. A decentralized energy trading algorithm that can be executed by the entities in a real-time fashion was presented in [10]. To deal with uncertainty issues, a probabilistic load model and

Manuscript received Month, 2018; revised Month, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was AB.

Milad Latifi, A. Rastegarnia and A. Khalili are with the Department of Electrical Engineering, Malayer University, Malayer, Iran (e-mail: miladlatifi98@gmail.com, rastegarnia@malayeru.ac.ir, khalili@malayeru.ac.ir).

W. M. Bazzi is with the Department of Electrical Engineering, American University in Dubai, Dubai, United Arab Emirates, (email: wbazzi@aud.edu)

S. Sanei is with School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK (email: saeid.sanei@ntu.ac.uk).

Digital Object Identifier XXXXX/XXXXXXX

a robust framework for renewable generation were proposed in this work. Designing adaptive learning algorithms to seek the Nash equilibrium (NE) of the constrained energy trading game among individually strategic players with incomplete information was discussed in [11]. In this game, each player used the learning automaton scheme to generate the action probability distribution based on his/her private information for maximizing his own averaged utility. In [12], a decentralized energy trading framework was implemented enabling the independent system operators to incentivize the entities toward an operating point that jointly optimize the cost of load aggregators and profit of the generators, as well as the risk of shortage in the renewable energy generation. To address the uncertainties in the renewable resources, they applied a risk measure called conditional value-at-risk (CVaR) with the goal of limiting the likelihood of high renewable energy generation shortage with a certain confidence level.

Bahrami *et al.* studied the users' long-term load scheduling problem in [13] developing an online load scheduling learning algorithm based on the actor-critic method to determine the users' Markov perfect equilibrium (MPE) policy. The authors of [14] investigated auction mechanisms for energy trading in a smart multi-energy district, in which the district manager sells electricity, natural gas, and heating energy to users as well as trading with outer energy networks. According to feed-in-tariff of photovoltaic (PV) energy, a system model of energy sharing management (ESM) was introduced in [15], which included the profit model of micro-grid operator (MGO) and the utility model of PV prosumers. In [16] an energy-sharing model with price-based demand response was analyzed for micro-grids of peer-to-peer PV prosumers. A dynamical internal pricing model was formulated for the operation of energy-sharing zone, which was defined based on the supply and demand ratio (SDR) of shared PV energy.

An energy storage (ES)-equipped ESM was developed in [17] to facilitate the energy sharing of multiple PV prosumers. In this work, the autonomous PV prosumers were formed as an energy-sharing network, and the energy-sharing activities were categorized as direct and buffered sharing. A day-ahead scheduling model of the ESM was built using stochastic programming to increase the operation profit and improve the net power profile of the energy-sharing network considering various types of uncertainties. The authors of the paper [18] formulated a micro-grid energy trading game, in which each micro-grid trades energy according to the predicted renewable energy generation and local energy demand, the current battery level, and the energy trading history. They presented a reinforcement learning based energy trading scheme that applied the deep Q-network (DQN) to improve the utility of the micro-grid for the case with a large number of the connected micro-grids. In [19], the consensus alternating direction method of multipliers (ADMM) algorithm is used to apply a novel cost allocation policy in peer-to-peer electricity markets. In this work the market participants have knowledge about the ISO charges prior to the negotiation process enabling them to anticipate on the network trade cost.

B. Our Contributions

However, there is still a lack of jointly designing energy management and trading (EMT) mechanism to effectively balance the uncertain supply fluctuations of the uncertain load demands. To the best of our knowledge, this paper is the first one in providing an EMT mechanism for the M/NG concerning with maximizing the usage of local RERs, autonomous direct energy trading with high efficiency, and minimizing the power transmission losses, while characterizing the M/NGs' equipment in detail. The main contributions are:

Establishing a joint energy consumption and trading management: To promote sustainable development, i.e., using the available generating capacity more efficiently, we enhance the demand side management (DSM) technique, a tool for load shaping that can redistribute (shifting some amount of) the energy demand over a certain period, to match the renewable power generation pattern. A novel supply-bidding price function mechanism is designed which couples the prosumers' actions, encouraging them to cooperatively take optimal decisions in the online EMT. A post-decision state (PDS) reinforcement learning mechanism is developed as the best response to the formulated distributed game-theoretic DSM, to tackle the uncertainty in the resources for the system operation.

Designing a novel hybrid iterative double-auction: In different time, the prosumers may behave as sellers or buyers depending on the electricity trading price and their net power profiles. Sellers make profit by selling their surplus of energy stored in storage devices such as electric vehicle batteries. Buyers can save on their energy bill by buying energy from their neighbors, instead of the grid, at a lower price, which also decreases the load on the grid. As in the proposed autonomous EMT framework the electricity trading price have no predetermined[1] standard value and is affected by many circumstances at a specific time (e.g., amount of supply and demand and the prosumers' preferences), an auction model has been used to clear the electricity market. A self-interested simple, flexible, and scalable market auction has been designed to guarantee the individual benefit and the global system efficiency simultaneously. The proposed double-auction mechanism is practical as it has all the necessary features, i.e., incentive compatibility (IC), individual rationality (IR), and budget balance (BB) [21].

Formulating a distributed optimal power allocation: During the local energy trading, substantial reverse power flow from the prosumers to the substation can cause the voltage magnitude of some of the households to exceed the upper limit of the allowed voltage variation. This is referred to as the voltage-frequency rise problem. The probability of facing this problem increases when more users decide to inject their excess generation via the main feeder into the grid. This increases the loss and reduces the power quality [22]. Therefore, the ability of users to route their excess power directly to their neighbors reduces the probability of voltage-frequency rises. At the final stage of the proposed

[1]Determining a predetermined fixed sell/buy price can reduce the interest in the power trading among the M/NGs [20].

EMT mechanism, an optimal power allocation is formulated by which the prosumers deliver/receive energy to/from the nearest neighbor in order to minimize the energy transmission cost.

Notation: Throughout the paper, $|\cdot|$ denotes the cardinality operator, $[a]^+ = \max\{a, 0\}$, $\mathbb{E}[\cdot]$ denotes the expectation operator, $I(\cdot)$ is the indicator function equal to one if is the case and equal to zero otherwise, and Cartesian product of the set is denoted by $\times$.

II. System Model

Consider a smart grid with a set $\mathcal{K}$ of $|\mathcal{K}| = K$ M/NG prosumers. Here, a prosumer can be a single residential/commercial/industrial entity (nano-grid), or a group of components that act as a single demand response entity (micro-grid) operating at both grid-connected and islanded modes. Each prosumer $k \in \mathcal{K}$ potentially comprises of set $\mathcal{A}_k$ of $|\mathcal{A}_k| = A_k$ power consumers (appliances), set $\mathcal{J}_k$ of $|\mathcal{J}_k| = J_k$ distributed storage (DS) units (e.g., backup-battery-banks (BBBs) and plug-in electric vehicles (PEVs)), and set $\mathcal{M}_k$ of $|\mathcal{M}_k| = M_k$ conventional (fossil fuel) generators and RER facilities (e.g., wind turbines, PV panels, and tidal energies) as well as an appropriate two-way power transmission and communication lines connected with the other prosumers and the main grid. Further, each prosumer has an energy management unit (EMU) responsible for managing, controlling, and monitoring the operation of all the prosumer's assets and sharing the energy resources with the other prosumers or the main grid, if needed. Also, there is an independent system operator (ISO) responsible for coordination, monitoring, and supervision of the prosumers' interactions together and with the main grid.

The overall energy consumption management and power trading horizon (scheduling window) is denoted by $\mathcal{H} \triangleq \{1, 2, \cdots, H\}$, where $H = |\mathcal{H}|$ is the number of time-slots with equal lengths[2]. We assume that the sequence of time-slots h, $\forall \text{days} = \{0, 1, 2, \cdots\}$ is predetermined in a deterministic manner and repeated every day, i.e., $h = \text{mod}(\text{day}, H), \forall \text{days} \geq 0$. To avoid ambiguity, hereafter, we use index $h \in \mathcal{H}$ for a time-slot in general and use index τ specifically for the current time slot at which the operations are made. The main challenges of the EMU to take optimal decisions are the uncertainty about the load demand, electricity market prices, and renewable generation in the upcoming time slots $h > \tau$. So, to improve the performance, at the beginning of the current time slot τ, each EMU updates its belief on the state of the load demand and generated power of its own assets (represented in Section III) and the price behavior (through the learning mechanism formulated in Section IV) with the gradual revealed demand/generation/price information over the period $\mathcal{H}^\tau = \{\tau, \cdots, H\} \subseteq \mathcal{H}$. At first, the EMU of each prosumer $k \in \mathcal{K}$ should characterize its assets as follows:

[2]The duration of a period can be 5, 15, or 60 mins, based on the time resolution at which the energy dispatch or the demand response decisions are made.

A. Power Consumers (Appliances)

In general, the power consumers of prosumers k are classified into four categories $\mathcal{A}_k = \mathcal{A}_k^{nf} \cup \mathcal{A}_k^{lf} \cup \mathcal{A}_k^{mf} \cup \mathcal{A}_k^{hf}$, namely, non-flexible $\mathcal{A}_k^{nf}$, low-flexible $\mathcal{A}_k^{lf}$, mid-flexible $\mathcal{A}_k^{mf}$, and high-flexible $\mathcal{A}_k^{hf}$ appliances. This classification is based on the ability/flexibility to set the time and power consumption rate of the electrical appliances and the total energy demand to finish the obligated task. The non-flexible appliances (such as refrigerator and television) are not schedulable as they need to work with their nominal schedule with a predetermined power/time of consumption. So, there is no authority to manage their operations and they must consume the power immediately. Low-flexible appliances have less strict operation schedule in the sense that one can only manage the start point of their operation time. On the other hand, for the mid-flexible appliances, both the power consumption rate and the operation time can be altered and interrupted. Unlike previous appliances which need essential fixed energy demand for their task, in the high-flexible class of appliances the tasks can be performed with less energy or the whole task can be postponed to another scheduling window.

For each appliance $a \in \mathcal{A}_k^{lf} \cup \mathcal{A}_k^{mf} \cup \mathcal{A}_k^{hf}$ of prosumer k, we denote its power consumption at slot h by $x_{k,a}(h)$ and its consumption profile through one scheduling window $\mathcal{H}$ by $\mathbf{x}_{k,a} \triangleq [x_{k,a}(1), \cdots, x_{k,a}(H)]$. Further, the operation state at slot τ is denoted by $s_{k,a}^\tau \triangleq (r_{k,a}(\tau), d_{k,a}(\tau))$, where $r_{k,a}(\tau)$ is the number of remaining time slots to complete the current task and $d_{k,a}(\tau) = (\beta_{k,a} - r_{k,a}(\tau)) - \tau$ is the number of time slots for which the current task can be delayed, both updated at slot $\tau - 1$[3]. Each appliance $a \in \mathcal{A}_k$ should accomplish its work within its own allowed scheduling window $\mathcal{H}_{k,a} \triangleq \{\alpha_{k,a}, \cdots, \beta_{k,a}\} \subseteq \mathcal{H}$, where $\alpha_{k,a}$ is the declared time for operation of the appliance and $\beta_{k,a}$ is the deadline by that the task of the appliance a must be finished. The total load demand of the non-flexible appliances (called the base-load) of prosumer k at slot h is denoted by $l_{k,b}^{nf}(h)$. However, depending on the preferences of the prosumer, an appliance can be put into the non-flexible category in one day and as other categories in another day.

Low-flexible appliances: Washing and drying machines are examples of low-flexible appliances. Delaying their operations incurs significant dissatisfaction level, which is modeled by an incommodity cost with a non-decreasing and convex function $f^{lf}(\cdot)$. These appliances consume a fixed amount of energy at each slot. Further, once the operation of these appliances are started, they must continuously work until their tasks is finished. The incommodity obtained from scheduling the operation of these appliances depends on the operation time and defined as $C_{k,a}^{lf}(\mathbf{x}_{k,a}) = \sum_{h=\alpha_{k,a}}^{\beta_{k,a}} \delta_{k,a}^{lf}(h) f^{lf}(x_{k,a}(h) - x_{k,a}^{des}(h))$, with the desired power consumption $x_{k,a}^{des}(h)$ declared for slot h before the scheduling program and time dependent non-negative non-decreasing coefficients $\delta_{k,a}^{lf}(h)$.

Mid-flexible appliances: These kinds of appliances (such as water pump) are more flexible in the sense that they may consume a fixed or regulated power and their opera-

[3]See [13], [23] for the detailed description of updating the state of different appliances over time.

tion can be interrupted. For these appliances, both delaying the operation time and the number of interruptions impose some inconveniences. So, we propose the incommodity cost $C^{mf}_{k,a}(\mathbf{x}_{k,a}) = \sum_{h=\alpha_{k,a}}^{\beta_{k,a}} \delta^{mf}_{k,a}(h) f^{mf}(x_{k,a}(h) - x^{des}_{k,a}(h)) + \acute{\delta}^{mf}_{k,a} g^{mf}(\sum_{h=\alpha_{k,a}}^{\beta_{k,a}} \gamma_{k,a}(h))$, where $f^{mf}(x)$ and $g^{mf}(x)$ are some non-decreasing convex functions. The auxiliary variable $\gamma_{k,a}(h)$ captures the number of interruptions (i.e., $\gamma_{k,a}(h) = 1$ if appliance a of prosumer k is turned on at slot h.), $\delta^{mf}_{k,a}(h)$ is a time dependent non-negative non-decreasing coefficient, and $\acute{\delta}^{mf}_{k,a}$ is a fixed coefficient.

High-flexible appliances: These kinds of appliances (such as pool pump) have regulated power consumption and non-vital energy need. The incommodity $C^{hf}_{k,a}(\mathbf{x}_{k,a}) = \delta^{hf}_{k,a} f^{hf}(E^{max}_{k,a} - \sum_{h=\alpha_{k,a}}^{\beta_{k,a}} x_{k,a}(h))$ obtained via the operation of these appliances depends only on the total power consumption at the end of the deadline $\beta_{k,a}$, where $\delta^{hf}_{k,a}$ is a fixed coefficient and $E^{max}_{k,a}$ is the maximum desired energy the prosumer needs to be consumed by appliance a.

To schedule the appliances of prosumer k, his EMU is faced with the following constraints:

$$\begin{aligned}
x_{k,a}(h) = 0,\ \forall\ h \notin \mathcal{H}_{k,a},\ a \in \mathcal{A}^{lf}_k \cup \mathcal{A}^{mf}_k \cup \mathcal{A}^{hf}_k,\\
x^{min}_{k,a} \le x_{k,a}(h) \le x^{max}_{k,a},\ \forall\ h \in \mathcal{H}_{k,a},\ a \in \mathcal{A}^{mf}_k \cup \mathcal{A}^{hf}_k,\\
x_{k,a}(h) = x^{rat}_{k,a} \cdot \gamma_{k,a}(h),\ \forall\ h \in \mathcal{H}_{k,a},\ a \in \mathcal{A}^{lf}_k \cup \mathcal{A}^{mf}_k,\\
\sum_{h=\alpha_{k,a}}^{\beta_{k,a}} \gamma_{k,a}(h) = 1,\ \forall\ \ a \in \mathcal{A}^{lf}_k,\\
\sum_{h=\alpha_{k,a}}^{\beta_{k,a}} \gamma_{k,a}(h) \ge 1,\ \forall\ \ a \in \mathcal{A}^{mf}_k,\\
\sum_{h=\tau+1}^{\beta_{k,a}} x_{k,a}(h) = E^{des}_{k,a} - E^{\tau}_{k,a},\ \forall\ a \in \mathcal{A}^{lf}_k \cup \mathcal{A}^{mf}_k,\\
E^{min}_{k,a} - E^{\tau}_{k,a} \le \sum_{h=\tau+1}^{\beta_{k,a}} x_{k,a}(h) \le E^{max}_{k,a} - E^{\tau}_{k,a}, \forall\ a \in A^{hf}_k
\end{aligned} \tag{1}$$

where $E^{\tau}_{k,a} = \sum_{h=\alpha_{k,a}}^{\tau} x_{k,a}(h)$, $x^{min}_{k,a}$ and $x^{max}_{k,a}$ are respectively minimum and maximum power rate of high-flexible (and possibly mild-flexible) appliances, $x^{rat}_{k,a}$ is the rated power consumption of low-flexible (and possibly mid-flexible) appliances, $E^{des}_{k,a}$ is the desired fixed amount of energy the appliance $a \in \mathcal{A}^{lf}_k \cup \mathcal{A}^{mf}_k$ must consume before the deadline $\beta_{k,a}$, which for appliances $a \in \mathcal{A}^{hf}_k$ it is in the tolerable range $[E^{min}_{k,a} - E^{max}_{k,a}]$.

The first line of (1) implies that none of the appliances can consume power out of its scheduling windows $\mathcal{H}_{k,a}$, the fourth line ensures that the low-flexible appliances have a continuous working period, and the fifth line allows the mid-flexible appliances to have discrete power consumption pattern. At slot τ, the possibility of applying the DSM policies on the appliances is determined based on the state $s^{\tau}_{k,a}(r_{k,a}(\tau), d_{k,a}(\tau))$ of each appliance a updated through the two last lines of (1). To evaluate the potential possibility of DSM, the EMU divides all the appliances into two groups denoted by the sets $\mathcal{A}^{act}_k$ and $\mathcal{A}^{pas}_k$ according to their states. For example, all the non-flexible appliances are always in group $\mathcal{A}^{pas}_k$, while low/mid/high-flexible appliance are in group $\mathcal{A}^{pas}_k$ when $\beta_{k,a} - \tau < r_{k,a}(\tau) + 1$ (i.e., $d_{k,a}(h) < 1$) or when a mid-flexible appliance start working, it moves to this list. By updating the state $s^{\tau}_{k,a}(r_{k,a}(\tau), d_{k,a}(\tau))$ of each appliance a, the EMU rearranges the appliances which can be scheduled in group $\mathcal{A}^{act}_k$ and the load must be supplied into group $\mathcal{A}^{pas}_k$. Further, those challenging appliances which are not sent their state signal are considered to be in the set of off appliances $\mathcal{A}^{off}_k$ for capturing the uncertainty of the load demand.

B. DS Units

Devices with storing capability have an important role in the EMT program. The DS units do not only help to balance the operation of networks with high RER penetration, but also they contribute to an overall improvement of the system efficiency and smoothing of the frequency and voltage fluctuations [24]. To use the potential of each DS unit $j \in \mathcal{J}_k$ in the EMT program, the EMU is subject to the following constraints:

$$\begin{aligned}
&\sum_{h=\tau+1}^{\beta_{k,j}} \big(\eta^{c}_{k,j} x^{c}_{k,j}(h) + \frac{x^{d}_{k,j}(h)}{\eta^{d}_{k,j}}\big) = E^{des}_{k,j} - (E_{k,j}(\tau) + E^{0}_{k,j}),\\
&E_{k,j}(\tau) = E_{k,j}(\tau - 1) + \eta^{c}_{k,j} x^{c}_{k,j}(\tau) - \frac{x^{d}_{k,j}(\tau)}{\eta^{d}_{k,j}},\\
&E^{min}_{k,j} \le E_{k,j}(h) \le E^{max}_{k,j},\ \forall\ h \in \mathcal{H}_{k,j},\ E^{lb}_{k,j} \le E^{des}_{k,j} \le E^{ub}_{k,j},\\
&x^{c}_{k,j}(h) \cdot x^{d}_{k,j}(h) = 0,\ \forall\ h \in \mathcal{H}_{k,j},\\
&x^{c}_{k,j}(h) + x^{d}_{k,j}(h) = 0,\ \forall\ h \notin \mathcal{H}_{k,j}
\end{aligned} \tag{2}$$

where $E_{k,j}(\tau)$, $E^{0}_{k,j}$, $E^{min}_{k,j}$, and $E^{max}_{k,j}$ are the energy level at the end of slot τ, initial energy level, minimum acceptable energy level, and the storage capacity, respectively. $\mathcal{H}_{k,j} \triangleq [\alpha_{k,j}, \cdots, \beta_{k,j}]$, with $\alpha_{k,j}$ and $\beta_{k,j}$ denoting the first and last slots the DS unit j is available[4] to the EMU of prosumer k, and coefficients $\eta^{c}_{k,j}, \eta^{d}_{k,j} \in (0, 1]$ denote charging (with rate $x^{c}_{k,j}(h) \ge 0$) and discharging (with rate $x^{d}_{k,j}(h) \ge 0$) efficiencies, respectively. Usually, there is a desirable energy level $E^{des}_{k,j}$ each unit needs to consume before becoming unavailable to the EMU. For example, a PEV owner needs his PEV to have some level of energy for his trip before the departure at slot $\beta_{k,j}$, or sometimes it is necessary for the BBB to have some initial backup energy level before starting the next scheduling horizon. Such requirements are satisfied through the first dynamic equation in (2). The energy level evolution of the DS unit follows the second line of (2) and the energy level bounds and the tolerable deviations are provided in the third line. Further, the fourth and fifth lines imply that the DS unit cannot be charged and discharged at the same time and cannot be charged/discharged when is unavailable to the EMU, respectively. In particular, if we are going to use a DS unit only as a power resource (like the RERs), we can just let $E^{des}_{k,j} = 0$ and we have $x^{c}_{k,j}(h) = 0$ and $x^{d}_{k,j}(h) > 0$. In this case, the second line of (2) denotes that the part $\frac{x^{d}_{k,j}(\tau)}{\eta^{d}_{k,j}}$ is subtracted from the energy level $E_{k,j}(h)$ of the DS unit at each slot h. we can consume all the energy stored in the DS unit according to the first dynamic equation in (2) by letting $E^{min}_{k,j} = 0$.

[4]For appliances such as PEVs which also have their own individual tasks, $\beta_{k,j}$ is the deadline similar to that in Section II-A.

To maximize the lifetime of DS unit, it is necessary to consider a cost function $C^{ds}_{k,j}(x^c_{k,j}(h), x^d_{k,j}(h), E_{k,j}(h)) = \psi^r_{k,j} f^{ds}\big(x^c_{k,j}(h)+x^d_{k,j}(h)\big)+\psi^f_{k,j} g^{ds}\big(E_{k,j}(h)-E_{k,j}(h-1)\big)+c^{fix}_{k,j}$ for its operation, with some non-decreasing convex functions $f^{ds}(\cdot)$ and $g^{ds}(\cdot)$, and some non-negative weights $\psi^r_{k,j}$ and $\psi^f_{k,j}$. The first term is the cost due to the charging/discharging rate, the second term is the cost due to the fluctuation of energy level, and $c^{fix}_{k,j}$ is the fixed investment and maintenance cost.

C. Power Resources

The future generation of RERs on the M/NG level will rely on the wind turbines and photovoltaic technologies [24]. So, we assume the prosumers have access to such energy sources somehow, and there is possibility for some prosumers to reserve diesel generators.

Diesel generator: The output of generator $m \in \mathcal{M}^{dg}_k \subseteq \mathcal{M}_k$ of prosumer k for slot h is denoted by $P^{dg}_{k,m}(h)$ with the following minimum $P^{min}_{k,m}$ and maximum $P^{max}_{k,m}$ generation limit:

$$P^{min}_{k,m} \leq P^{dg}_{k,m}(h) \leq P^{max}_{k,m} \tag{3}$$

The common cost function for the conventional generators is the quadratic cost function $C^{dg}_{k,m}(P^{dg}_{k,m}(h)) = a_m(P^{dg}_{k,m}(h))^2 + b_m P^{dg}_{k,m}(h) + c_{k,m}$, with non-negative coefficients a_m and b_m, and a fixed operational cost $c_{k,m}$ [25].

Photovoltaic panel: The PV power generators of the prosumer k are denoted by $m \in \mathcal{M}^{pv}_k \subseteq \mathcal{M}_k$. Assuming operation at maximum power point tracking (MPPT), the output power $P^{pv}_{k,m}(h)$ of the PV unit m of prosumer k is [26]:

$$P^{pv}_{k,m}(h) = \eta^{pv}_{k,m} \cdot A^{pv}_{k,m} \cdot R^{pv}_{k,m}(h) \tag{4}$$

where $\eta^{pv}_{k,m}$ is the PV panel efficiency, $A^{pv}_{k,m}$ is the panel area, and $R^{pv}_{k,m}(h) = I^{ci}(h) \cdot R^{max}_{si}$ is the solar irradiation at slot h, with the clearness index $I^{ci}(h)$ and the extraterrestrial solar radiation R^{max}_{esi}. The extraterrestrial solar radiation is approximated as $R^{max}_{esi} = I_{sc}(1 + 0.033\cos(360t/365))\sin\alpha(h)$, with $\sin\alpha(h) = \sin\phi\sin\gamma + \cos\phi\cos\gamma\cos\omega(h)$, where I_{sc} is a solar constant, t is the day of a year, $\alpha(h)$ is the altitude of the sun, ϕ is the latitude, γ is the declination of the sun, and $\omega(h)$ is the hour angle [27]. The clearness index $I^{ci}(h)$ at each slot h denotes an index that any extraterrestrial solar radiation tolerates by the natural factors such as cloud and temperature. Using Beta distribution, the probabilistic fluctuation of the clearness index is described as $p^{ci}(I^{ci}(h)) = \big(\Gamma(a+b)/\Gamma(a)\Gamma(b)\big)[I^{ci}(h)]^{a-1}[1-I^{ci}(h)]^{b-1}$, with $a = \big((\mu^{sr})^2(1-\mu^{sr})/(\sigma^{sr})^2\big) - \mu^{sr}$, $b = a(1-\mu^{sr})/\mu^{sr}$, and gamma function $\Gamma(\cdot)$, where μ^{sr} and σ^{sr} are the mean value and standard deviation of solar radiation supply computed according to the weather historical data, respectively [27].

Wind turbine: Let $\mathcal{M}^w_k \subseteq \mathcal{M}_k$ denote the set of all wind turbines belonging to prosumer k. The power output of each turbine $m \in \mathcal{M}^w_k$ is calculated based on the wind speed and the wind turbine power coefficient obtained from the basic expression $P^w_{k,m}(h) = \rho/2 \cdot \eta^w_{k,m} \cdot A^w_{k,m} \cdot [v_{k,m}(h)]^3$, where ρ is the air density, $\eta^w_{k,m}$ is the power coefficient, $A^w_{k,m}$ is the swept area of the wind rotor, and $v_{k,m}(h)$ is the wind speed (m/s) at the site of turbine m [26]. Since the wind generation output varies with wind speed, a probabilistic fluctuation analysis of wind speed can effectively handle the uncertainty of the wind generation. The wind speed fluctuations can be characterized using Weibull distribution $p^w(v_{k,m}(h)) = (a/b)(v_{k,m}(h)/b)^{a-1}\exp(-(v_{k,m}(h)/b)^a)$, with $a = (\sigma^w/\mu^w)^{-1.086}$ and $b = \mu^w/\Gamma(1+a^{-1})$, where, μ^w and σ^w are the mean value and standard deviation of wind speed based on the observed value, respectively [27]. Accordingly, the output power of wind turbine m corresponding to its rated power $P^{rat}_{k,m}$ (kW) is described as:

$$P^w_{k,m}(h) = \begin{cases} f^w(v_{k,m}(h)), & v^{in}_{k,m} \leq v_{k,m}(h) \leq v^{rat}_{k,m} \\ P^{rat}_{k,m}, & v^{rat}_{k,m} < v_{k,m}(h) \leq v^{out}_{k,m} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $f^w(v_{k,m}(h)) = P^{rat}_{k,m}[(v_{k,m}(h))^3 - (v^{in}_{k,m})^3]/[(v^{rat}_{k,m})^3 - (v^{in}_{k,m})^3]$ is the cubic function which yields the wind power [28], or we can let $f^w(v_{k,m}(h)) = P^{rat}_{k,m}(v_{k,m}(h)/v^{rat}_{k,m})^3$ according to [8], where v^{in}_k, v^{out}_k, and v^{rat}_k are the cut in, cut out, and rated speed (m/s) of wind turbine m of prosumer k, respectively.

III. Problem Formulation

In the online deterministic formulation of the EMT problem, there are three uncertainty sources; 1) the upcoming load demand of appliance $a \in \mathcal{A}^{off}_k$ of the prosumer and its scheduling capability (i.e., if they are in group $\mathcal{A}^{act}_k$ or group $\mathcal{A}^{pass}_k$), 2) the total generated power from the RERs at the upcoming slots, and 3) the electricity market selling/buying price which are challenging problems. In this section, we handle 1) and 2) and formulate the global social welfare maximization problem, and provide a game-theoretic reinforcement learning mechanism to tackle 3) in the next section.

At slot τ, the EMU of prosumer k has no information about the type and state of appliances $a \in \mathcal{A}^{off}_k$, since the actual load of these appliances and their operation time are not known in advance. So, it is impossible to deterministically group the appliances $a \in \mathcal{A}^{off}_k$ into non-schedulable group $\mathcal{A}^{pas}_k$ or schedulable group $\mathcal{A}^{act}_k$. To address the lack of information, one can collect the operation time/amount historical data record of each appliance for estimating the probability $p^{ap}_{k,a}(h)$ that each appliance $a \in \mathcal{A}_k$ sends the state signal at each time slot $h > \tau$. The conditional probability $p^{ap}_{k,a}(h|\tau)$ that the appliance $a \in \mathcal{A}^{off}_k$ goes on in an upcoming time slot $h > \tau$, given that it has been off before the current time slot, τ, is [10]:

$$p^{ap}_{k,a}(h|\tau) = \frac{p^{ap}_{k,a}(h)}{1 - \sum_{t=1}^{\tau} p^{ap}_{k,a}(t)} \tag{6}$$

The most conservative decision results from considering the worst-case scenario, in which the electric appliances that send a demand signal in the upcoming time slots $h > \tau$, are all located in group $\mathcal{A}^{pass}_k$, i.e., must consume power immediately. Accordingly, the total base-load of prosumer k at slot h is updated as $l_{k,b}(h) = l^{nf}_{k,b}(h) + l^{pas}_{k,b}(h) + l^{off}_{k,b}(h)$, where $l^{pas}_{k,b}(h)$ is the load caused by moving appliance a from group $\mathcal{A}^{act}_k$ to group $\mathcal{A}^{pas}_k$ (when $d_{k,a}(h) = 0$), and the expected worst-case

electric demand $l_{k,b}^{off}(h)$ is calculated as:

$$l_{k,b}^{off}(h) = \sum_{a \in \mathcal{A}_k^{off}} x_{k,a}^{rat} \Big[\sum_{t=\max\{\tau+1, h-\varpi_{k,a}+1\}}^{h} p_{k,a}^{ap}(t|\tau) \Big] \quad (7)$$

where parameter $\varpi_{k,a} = r_{k,a}(0)$ and $x_{k,a}^{rat}$ are the operation duration and the rated power consumption of the appliance $a \in \mathcal{A}_k^{off}$ that goes on in upcoming time slot $h > \tau$ and operate immediately with the probability $\sum_{t=\max\{\tau+1, h-\varpi_{k,a}+1\}}^{h} p_{k,a}^{ap}(t|\tau)$ [10].

To tackle the RER generation uncertainty, using PV model (4), wind generation model (5), and the historical data records, we can forecast the joint uncertainty power output bound $[P_{k,rer}^{min}(h), P_{k,rer}^{max}(h)]$ of all the RER units of prosumer k at each slot h. However, providing an EMT strategy by taking into account all possible scenarios $P_k^{rer}(h) = \sum_{m \in \mathcal{M}_k^{pv}} P_{k,m}^{pv}(h) + \sum_{m \in \mathcal{M}_k^{w}} P_{k,m}^{w}(h) \in [P_{k,rer}^{min}(h), P_{k,rer}^{max}(h)]$ is very conservative and possibly inefficient. One can consider an adaptive robust model by defining the uncertainty space for the joint generation profile $\boldsymbol{P}_k^{rer}(\tau) = [P_k^{rer}(\tau), \cdots, P_k^{rer}(H)]$ of the RER units of prosumer k at slot τ through the upcoming slots $\mathcal{H}^\tau$ as [29]:

$$\begin{aligned} \mathcal{P}_k^{rer}(\tau) = \{ \boldsymbol{P}_k^{rer}(\tau) | P_k^{rer}(h) \in [P_{k,rer}^{min}(h), P_{k,rer}^{max}(h)], \\ h \in \mathcal{H}^\tau, \sum_{h \in \mathcal{H}^\tau} \frac{P_{k,rer}^{max}(h) - P_k^{rer}(h)}{P_{k,rer}^{max}(h) - P_{k,rer}^{min}(h)} \leq \Delta_k(\tau) \} \end{aligned} \quad (8)$$

where $[0 \text{ (least-conservative)} \leq \Delta_k(\tau) \leq |\mathcal{H}^\tau| \text{ (most-conservative)}]$ is the confidence level parameter which should be optimized by prosumer k [10]. To encourage the prosumers to have good power generation evaluations and reasonable offers to the system by implementing the robust model (8), the ISO charges them by the unit price $\rho_{gs}(h)$ (\$/MW) for the generation shortage in time slot h. According to the cost imposed by prosumer k due to the power shortage becomes[5]:

$$C_k^{rer}(\boldsymbol{P}_k^{rer}(\tau)) = \sum_{h \in \mathcal{H}^\tau} \rho_{gs}(h) [P_k^{rer}(h) - \breve{P}_k^{rer}(h)]^+ \quad (9)$$

where $\breve{P}_k^{rer}(h)$ is the actual renewable generated power. It is clear from (9) that the least-conservative decision, $\Delta_k(\tau) = 0$ (i.e., offering $P_k^{rer}(h) = P_{k,rer}^{max}(h)$), can impose the most possible penalty to prosumer k.

So, the aggregate cost imposed on prosumer k at slot h is defined as:

$$\begin{aligned} C_k(h) = & \sum_{a \in \mathcal{A}_k^{lf}} C_{k,a}^{lf}(\mathbf{x}_{k,a}) + \sum_{a \in \mathcal{A}_k^{mf}} C_{k,a}^{mf}(\mathbf{x}_{k,a}) \\ & + \sum_{a \in \mathcal{A}_k^{hf}} C_{k,a}^{hf}(\mathbf{x}_{k,a}) + \sum_{j \in \mathcal{J}_k} C_{k,j}^{ds}(x_{k,j}^c(h), x_{k,j}^d(h), E_{k,j}(h)) \\ & + C_{k,m}^{dg}(P_{k,m}^{dg}(h)) + C_k^{rer}(\boldsymbol{P}_k^{rer}(h)) \end{aligned} \quad (10)$$

Further, prosumer k is subject to the following power balance constraints to maintain the stability of the power system at each slot:

$$x_k^{sl}(h) + x_k^{by}(h) = l_k^{ap}(h) + l_k^{ds}(h) - P_k(h) \quad (11)$$

where $x_k^{sl}(h) < 0$ and $x_k^{by}(h) \geq 0$ are the total power to be sold (due to the power surplus) and to be bought (due to the power deficit), respectively. The aggregate power consumption by all appliances $a \in \mathcal{A}_k$ of prosumer k is denoted by $l_k^{ap}(h)$, the net charge (≥ 0) or discharge (< 0) profile of all the DS units $j \in \mathcal{J}_k$ by $l_k^{ds}(h)$, and the aggregate generated power from all the available sources $m \in \mathcal{M}_k$ by $P_k(h)$. Accordingly, the total revenue prosumer k acquires from the energy trading is:

$$\begin{aligned} R_k(h) = & \rho_{sl}(h)[P_k(h) - l_k^{ds}(h) - l_k^{ap}(h)]^+ \\ & - \rho_{by}(h)[l_k^{ap}(h) + l_k^{ds}(h) - P_k(h)]^+ - C_k^{tr}(h) \end{aligned} \quad (12)$$

where $\rho_{sl}(h)$, $\rho_{by}(h)$, and $C_k^{tr}(h)$ are the selling price, the buying price, and the transmission cost, respectively. Accordingly, the global social welfare maximization problem can be defined as:

$$\begin{aligned} & \max \sum_{h \in \mathcal{H}} W^{iso}(h) = \sum_{h \in \mathcal{H}} \sum_{k \in \mathcal{K}} \big(R_k(h) - C_k(h) \big), \\ & \text{subject to: } (1) - (3), (11), \text{ and } \boldsymbol{P}_k^{rer}(h) \in \mathcal{P}_k^{rer}(\tau), \ \forall \ k \in \mathcal{K}, \\ & \text{variables: } \{ x_{k,a}(h), \gamma_{k,a}(h), x_{k,j}^c(h), x_{k,j}^d(h), x_k^{sl}(h), x_k^{by}(h), \\ & P_{k,m}^{dg}(h), P_k^{rer}(h), \rho_{gs}(h), \rho_{by}(h), \rho_{sl}(h) \}, \ \forall \ k \in \mathcal{K} \end{aligned} \quad (13)$$

The EMT problem in (13) is a mixed-integer nonlinear non-convex NP-hard optimization problem. Centrally solving this problem by the ISO bears major challenges such as; 1) imposing a huge communication and computational burden to the ISO, 2) putting into danger the prosumers' privacy as all the prosumers' information must be available to the ISO, 3) reducing the reliability of the system due to creating the critical single point of failure[6], 4) reducing the incentives for the prosumer to participate in the EMT program as the ISO decides on the traded power and selling/buying prices with the possibility of cheating and reducing the prosumers' revenue.

IV. Proposed Distributed-Hierarchical Framework

To tackle the existing challenges, we decompose problem (13) into three hierarchical sub-problems, i.e., the DSM problem (as the internal problem) and the energy trading and power allocation problems (as external problems).

A. Optimal DSM

To separate the internal DSM part of problem (13), it is necessary for the ISO to provide the prosumers with some supply-function bidding, representing the lower and upper bound of the transactions selling/buying prices. To encourage the prosumer for maximum possible local energy trading, it is reasonable to consider the highest/lowest buying/selling price for procuring/selling power from/to the main grid. Accordingly, it is expected that the lowest transaction cost is achieved when all the prosumers' load demands are satisfied locally. The supply function as a strategic variable allows to adapt better to changing the market conditions (such as uncertain and stochastic load demand and renewable power generation) than committing to keeping a price or quantity fixed. That is because no matter what the value of the supply

[5]Without loss of generality, we have not considered the investment, operation, and maintenance costs of the RERs.

[6]A single point of failure (SPOF) is a part of system that, if it fails, will stop the entire system from working. So, in the centralized solution, if the ISO (which is the SPOF) fails, the whole system stops working.

deficit/surplus is, the ISO can apply the supply function bid to the prosumers to clear the deficit/surplus [30]. It also respects practical informational constraints in the power network, as a properly-chosen parameterized supply function controls the information revelation [31]. Our model also prevents prosumers from selling more power at slots with very low demand by reducing the selling price and vice versa. So, each prosumer can approximate his individual effectual price $\rho_k^{eff}(h)$ as:

$$\rho_k^{eff}(h) = \rho_k(h) + f_{wsp}(\sum_{k\in\mathcal{K}} x_k^{sl}(h) + \sum_{k\in\mathcal{K}} x_k^{by}(h)),$$
$$\rho_k(h) = [\upsilon_k(h)/x_k^{net}(h)]\rho_{mcp}(h) + [(1-\upsilon_k(h))/x_k^{net}(h)]$$
$$[\rho_{iso}^{by}(h)I(x_k^{by}(h)) + \rho_{iso}^{sl}(h)I(x_k^{sl}(h))] \quad (14)$$

where $f_{wsp}(\cdot)$ is a supply-function bidding, $\upsilon_k(h)$ is the portion of energy directly traded with the other prosumers, $\rho_{mcp}(h)$ is the local electricity trading price (called the market clearing price (MCP)), and $\rho_{iso}^{by}(h)$ and $\rho_{iso}^{sl}(h)$ are the prices of buying and selling power from/to the ISO, respectively. We know that if $\sum_{k\in\mathcal{K}} x_k^{sl}(h) = \sum_{k\in\mathcal{K}} x_k^{by}(h)$, then, there is no need for the ISO to trade power with the wholesale market, which subsequently affects the prosumers' payments. The actual value of $\rho_k^{eff}(h)$ is not known to prosumer k in advance and depends on the scheduling and trading decisions of all the prosumers and the realizations of random events (e.g., the power generation from RERs and the load demand). For example, when the RER power generation is high and most of the prosumers have surplus energy, two factors reduces the chance of selling power with high price for all the prosumers as $x_k^{sl}(h) < 0$ increases. 1) The supply-function bidding $f_{wsp}(\cdot)$ reduces. 2) The chance of trading power with high MPC price $\rho_{mcp}(h)$ reduces and the chance of trading power with low prices $\rho_{iso}^{by}(h)$ and $\rho_{iso}^{sl}(h)$ increases as $\upsilon_k(h)$ reduces. These factors encourage the prosumers to store the surplus energy in their DS units as much as possible and sell it in another time. This behavior helps smoothing the power consumption/trading curves and alleviating the voltage and frequency fluctuation problems. So, the natural choice to model these interactions is the game theory. As in the game-theoretic model the best response is the dominant strategy (or strategies) which produces the most favorable outcome for the prosumers [32], we assume that each prosumer k has access to information $\boldsymbol{x}_{-k}^{net}(h) = \sum_{n\in\mathcal{K}/k}\big(x_n^{sl}(h) + x_n^{by}(h)\big)$ and decides to best respond to it, resulting to an equilibrium [33]. Accordingly, we can define the DSM objective function of each prosumer k at slot h as:

$$C_k^{dsm}(\mathscr{S}_k(h)) = \rho_k^{eff}(h)[x_k^{sl}(h) + x_k^{by}(h) + \boldsymbol{x}_{-k}^{net}(h)] + C_k(h) \quad (15)$$

where $\mathscr{S}_k(h) \in \mathbb{S}_k^h$ is the state of prosumer k at slot h determined according to $[s_{k,a}^{\tau}(r_{k,a}(\tau), d_{k,a}(\tau))]_{a\in\mathcal{A}_k}$ and $\boldsymbol{x}_{-k}^{net}(h)$, the state of the charge of the DS units, the power produced from the power resources, and the prosumer attitude (e.g., belief on the worthiness of price $\rho_{mcp}(h)$). The feasible state set $\mathbb{S}_k^h = \mathbb{A}_k \times \mathbb{J}_k \times \mathbb{M}_k \times \mathbb{P}_k$ is constructed of $\mathbb{A}_k$ (feasible state of appliances according to constraints in (1)), $\mathbb{J}_k$ (feasible state of the DS units according to constraints in (2)), $\mathbb{M}_k$ (feasible state of the power resources according to constraint (3), (11), and $\boldsymbol{P}_k^{rer}(h) \in \mathcal{P}_k^{rer}(\tau)$), and $\mathbb{P}_k$ (feasible state of the effectual price (14) according to the personal historical data of the customer k's payment). To take the best response, the constraints in (1) and (2) temporarily couple the prosumer decision through the scheduling horizon H. So, the challenge is to understand how a current action/state will affect the future profits, meaning that, for scheduling the equipment, the prosumer must infer (trade-off) that consuming/buying/selling power in the current slot is more profitable or the next slots. On the other, the decision making only depends on the belief of the effectual price $\rho_k^{eff}(h)$ and the state $\mathscr{S}_k(h)$, while it is independent of the time slot index. We propose using post-decision state reinforcement learning for each prosumer to foresee the change of the effectual price, learn his action, and determine the best response over the time [5]. A good way to model this task is with Markov decision processes (MDP), which is the dominant approach in the reinforcement learning theory [34].

As a decision maker, each prosumer chooses an energy consumption function $\mathscr{E}_k^h$ at each time-slot h among the set of energy consumption functions $\mathcal{E}_k^h = \{\mathscr{E}_{k,1}, \mathscr{E}_{k,2}, \cdots, \mathscr{E}_{k,E_k}\}$. Then, the actual energy consumption profile of prosumer k which is constructed of concatenating all the decision variables $\{x_{k,a}(h), \gamma_{k,a}(h), x_{k,j}^c(h), x_{k,j}^d(h), P_{k,m}^{dg}(h)\}$, i.e., $\boldsymbol{X}_k(h)$, is calculated based on the energy consumption function $\mathscr{E}_k^h$ and the prosumer's state $\mathscr{S}_k(h)$, i.e., $\boldsymbol{X}_k(h) = \mathscr{E}_k^h(\mathscr{S}_k^h)$.

The prosumer k decides its energy consumption function $\mathscr{E}_k^h$ based on the observation of its state $\mathscr{S}_k^h \in \mathbb{S}_k^h$. We denote prosumer k's stationary policy that maps its state sets $\mathbb{S}_k^h$ to the action sets $\mathcal{E}_k^h$ by $\pi_k : \mathbb{S}_k^h \to \mathcal{E}_k^h$, i.e., $\mathscr{E}_k^h = \pi_k(\mathscr{S}_k^h)$. Accordingly, the prosumer aims to solve the following MDP learning problem which aims to minimize the expected long-term (discounted) cost in the upcoming time slots [35]:

$$\min_{\pi_k:\mathbb{S}_k^h\to\mathcal{E}_k^h} \mathbb{E}\Big[\sum_{h=\tau}^{\infty} \gamma_k^h C_k^{dsm}(\mathscr{S}_k^h)\Big] \quad (16)$$

where the discount factor $0 \le \gamma_k^h \le 1$ can be used to characterize a wide range of the prosumers' behavior and the expectation is with respect to the demand and generation uncertainties in the upcoming time slots. When γ_k^h is close to zero, the prosumers are myopic, i.e., they aim to minimize their short-term cost without considering the consequences of their short-term policy on their future cost and vice versa. In the previous section, we provided a mechanism for the prosumers to estimate (and update) their load demand and generated power for the next slots. In order to exploit this available information for improving the learning accuracy and speed, we develop PDS learning algorithm to exploit the available information about the system which is revealed slot-by-slot [5]. We define prosumer k's PDS as the state where the known information is reflected based on prosumer k's decision on $\mathscr{E}_k^h$, but the unknown information is not reflected. Accordingly, we denote prosumer k's PDS at time-slot h by $\bar{\mathscr{S}}_k^h(\boldsymbol{X}_k(h+1), h+1, \mathscr{E}_k^h) \in \mathbb{S}_k^h$. To solve the MDP problem (16), the state transition probability from $\mathscr{S}_k^h$ to $\mathscr{S}_k^{h+1}$ is:

$$p(\mathscr{S}_k^{h+1}|\mathscr{S}_k^h, \mathscr{E}_k^h) = \sum_{\bar{\mathscr{S}}_k^h\in\mathbb{S}_k^h} p_{kn}(\bar{\mathscr{S}}_k^h|\mathscr{S}_k^h, \mathscr{E}_k^h) p_{un}(\mathscr{S}_k^{h+1}|\bar{\mathscr{S}}_k^h) \quad (17)$$

where $p_{kn}(\cdot)$ and $p_{un}(\cdot)$ denote the known and unknown probabilities, respectively [5]. The optimal PDS policy $\bar{\pi_k}^*(\bar{\mathscr{S}}_k^h)$ can be well defined by using the optimal action-value function $Q^* : \mathbb{S}_k^h \times \mathcal{E}_k^h \rightarrow \mathbb{R}$, which satisfies the Bellman optimality equation[7]:

$$Q^*(\mathscr{S}_k^h, \mathscr{E}_k^h) = r(\mathscr{S}_k^h, \mathscr{E}_k^h) + \gamma \sum_{\acute{\mathscr{S}}_k^h \in \mathbb{S}_k^h} p(\acute{\mathscr{S}}_k^h | \mathscr{S}_k^h, \mathscr{E}_k^h) V^*(\acute{\mathscr{S}}_k^h) \tag{18}$$

where $V^*(\acute{\mathscr{S}}_k^h) = \min_{\mathscr{E}_k^h \in \mathcal{E}_k^h} Q^*(\acute{\mathscr{S}}_k^h, \mathscr{E}_k^h),\ \forall \mathscr{S}_k^h \in \mathbb{S}_k^h$ is the optimal state-value function and $r(\mathscr{S}_k^h, \mathscr{E}_k^h)$ is the reward observed for the current state $\mathscr{S}_k^h$ [36]. As the action-state space in our stochastic MDP problem is potentially huge (continuous), to guarantee the convergence, one can apply a recursive equation with a suitable learning rate (step-size) $\alpha_k(h)$ to approximate the value of Q function as follows [37]:

$$\begin{aligned} Q(\mathscr{S}_k^h, \mathscr{E}_k^h) \leftarrow (1 - \alpha_k(h)) Q(\mathscr{S}_k^h, \mathscr{E}_k^h) + \\ \alpha_k(h) \left[r(\mathscr{S}_k^h, \mathscr{E}_k^h) + \gamma_k^h \max_{\acute{\mathscr{E}}_k^h \in \mathcal{E}_k^h} Q(\acute{\mathscr{S}}_k^h, \acute{\mathscr{E}}_k^h) \right] \end{aligned} \tag{19}$$

where $\max_{\acute{\mathscr{E}}_k^h \in \mathcal{E}_k^h} Q(\acute{\mathscr{S}}_k^h, \acute{\mathscr{E}}_k^h)$ is an estimate of the optimal future value. A detailed analysis for choosing optimal (step-size) $\alpha_k(h)$ for different MDPs are drawn in [38]. Using the Bellman optimality equation and (19), the state value functions of prosumer k's state, the PDS, and the optimal PDS policy becomes:

$$\begin{aligned} \bar{V}^*(\bar{\mathscr{S}}_k^h) &= \gamma_k^h \sum_{\acute{\mathscr{S}}_k^h \in \mathbb{S}_k^h} p_{un}(\acute{\mathscr{S}}_k^h | \bar{\mathscr{S}}_k^h, \mathscr{E}_k^h) V^*(\acute{\mathscr{S}}_k^h) \\ V^*(\mathscr{S}_k^h) &= \min_{\mathscr{E}_k^h \in \mathcal{E}_k^h} \Big[C_k^{dsm}(\mathscr{S}_k^h) + \\ &\sum_{\bar{\mathscr{S}}_k^h \in \mathbb{S}_k^h} p_{kn}(\bar{\mathscr{S}}_k^h | \mathscr{S}_k^h, \mathscr{E}_k^h) \bar{V}^*(\bar{\mathscr{S}}_k^h) \Big] \\ \bar{\pi_k}^*(\bar{\mathscr{S}}_k^h) &= \arg\min_{\mathscr{E}_k^h \in \mathcal{E}_k} \Big[C_k^{dsm}(\mathscr{S}_k^h) + \\ &\sum_{\bar{\mathscr{S}}_k^h \in \mathbb{S}_k^h} p_{kn}(\bar{\mathscr{S}}_k^h | \mathscr{S}_k^h, \mathscr{E}_k^h) \bar{V}^*(\bar{\mathscr{S}}_k^h) \Big] \end{aligned} \tag{20}$$

where, as we show in Section V, by exploitation of the known parts of probability $p_{kn}(\cdot)$ and updating the information about $\mathscr{S}_k^h$ the learning accuracy and speed are improved compared to the conventional Q-learning algorithm [37]. In the adopted learning method, the exploration parameter ϵ of ϵ-greedy is adaptively chosen corresponding to the temporal-difference error observed from value-function backups, which is considered as a measure of the prosumer's uncertainty about the environment. Balancing the ratio between exploration and exploitation, i.e., choosing appropriate ϵ, is one of the most challenging tasks in the reinforcement learning with great impact on the prosumer's learning performance. In one hand, too long exploration prevents the prosumer from maximizing short-term reward because the selected exploration actions may yield negative reward from the environment. On the other hand, exploiting uncertain environment knowledge prevents maximization of long-term reward since the selected actions may remain suboptimal. This problem is well known as the dilemma of exploration and exploitation [34]. The desired behavior is to have the prosumers more explorative in situations where the knowledge about the environment is uncertain, i.e. at the beginning of the learning process, which is recognized by large changes in the value function. Then, the exploration rate should be reduced as the prosumer's knowledge becomes certain about the environment, which can be recognized as very small or no changes in the value function. As an alternative we can adopt an adaptive value-difference-based ϵ-greedy exploration, according to a Softmax-Boltzmann distribution of the value-function estimates similar to that in [39].

[7]It is a necessary condition for optimality associated with the dynamic programming methods. It writes the value of a decision problem at a certain point in time in terms of the payoff from some initial choices and the value of the remaining decision problem that results from those initial choices. This breaks a dynamic optimization problem into simpler sub-problems [36].

B. Optimal Energy Trading

After scheduling the internal energy consumptions (i.e., the DSM part), the prosumers with $x_k^{net}(h) = x_k^{sl}(h) + x_k^{by}(h) < 0$ and $x_k^{net}(h) > 0$ participate in the energy trading stage as the sellers and buyers, respectively. The ISO as an auctioneer classifies them into set $\mathcal{S}$ of $S = |\mathcal{S}|$ sellers and set $\mathcal{B}$ of $B = |\mathcal{B}|$ buyers. At this stage, each potential seller $s \in \mathcal{S}$ sends the quantity of energy $x_s^{sl}(h)$ that it intends to supply and its reservation bid ρ_s^{sl} to the auctioneer (ISO). The reservation bid sent by the potential sellers corresponds to the minimum price at which the seller is willing to sell its offered amount of energy. On the other side, each potential buyer $b \in \mathcal{B}$ proposes a bid ρ_b^{by} and the quantity it requests, denoted by $x_b^{by}(h)$ to the auctioneer. As the MPC price $\rho_{mcp}(h)$ and its lower and upper bounds $[\rho_{iso}^{by}(h) - \rho_{iso}^{sl}(h)]$ depend on the prosumers' actions, for optimally determining the parameters ρ_s^{sl} and ρ_b^{by}, one can develop a game-theoretic competition among the prosumers similar to [23]. However, the proposed trading structure must have the following necessary economic properties to be a legitimate double auction mechanism:

1) Truthfulness or incentive compatibility (IC): The bidders cannot benefit from bidding different from their true valuation, i.e., cheating always harms. 2) Individual rationality (IR): Bidders get non-negative utilities, i.e., the sellers are paid no less than what they ask for and buyers do not pay more than their bids. 3) Budget-balance (BB): The total amount paid to the sellers is no more than the total amount received from the buyers. This prevents the auctioneer, who runs the auction, from losing money. 4) Efficiency or social welfare: The aggregate profit acquired by the prosumers from participating in the energy trading market. To achieve the maximum efficiency, the power should be sold to the buyers at the lowest price and bought from the sellers at the highest price, while minimizing the number of losers. In [21], it is shown that it is impossible for an auction mechanism to maximize the social welfare whilst being IR, IC, and BB at the same time. So, for the auctioneer to maintain the BB property in an IR and IC mechanism, it is necessary to compromise on the optimality of the social welfare as the IR and IC properties are essential. Thus, in this paper, we aim to retain features IR, IC, and BB, while achieving high (but not maximum) efficiency.

To satisfy the mentioned properties, we develop our hybrid double-auction mechanism as follows:

1) The sellers are ordered in an increasing order of their reservation price $\rho_1^{sl} < \rho_2^{sl} < \cdots < \rho_S^{sl}$; 2) the buyers are ordered in a decreasing order of their reservation bids $\rho_1^{by} > \rho_2^{by} > \cdots > \rho_B^{by}$; 3) if two sellers (respectively, buyers) have equal reservation prices (bids), they are aggregated into one single virtual seller (or buyer); 4) we consider the ISO as a virtual buyer/seller with the lowest buy $\rho_{iso}^{by}(h)$ and the highest sell $\rho_{iso}^{sl}(h)$ prices, make the balance in the energy trading marker (e.g., trade with the losers and supply the aggregate excess demand if $\sum_{b\in\mathcal{B}} x_b^{by}(h) + \sum_{s\in\mathcal{S}} x_s^{sl}(h) > 0$ or buy the aggregate surplus of energy if $\sum_{b\in\mathcal{B}} x_b^{by}(h) + \sum_{s\in\mathcal{S}} x_s^{sl}(h) < 0$) by procuring/selling some power from/to the wholesale market at the deficit/surplus of energy.

Following this sorting process, the supply curve (sellers' bids $[\rho_s^{sl}]_{s\in\mathcal{S}}$ as a function of their energy amounts $[x_s^{sl}(h)]_{s\in\mathcal{S}}$) and the demand curve (buyers' bids $[\rho_b^{by}]_{b\in\mathcal{B}}$ as a function of their load demands $[x_b^{by}(h)]_{b\in\mathcal{B}}$) can be generated. These two curves intersect at a point that corresponds to a given seller L and buyer M with bids $\rho_M^{by} \geq \rho_L^{sl}$. This intersection point is easily computed using routine numerical and graphical techniques [40]. Once we determine the seller L and a buyer M at the supply and demand intersection point, double auction theory implies that $L-1$ and $M-1$ buyers will practically participate in the market and directly trade with each other (as they are the winners). Here, as shown in [41], we must exclude seller L and buyer M from the market so as to guarantee that the total supply and demand will match while ensuring a strategy proof and truthful auction mechanism. To determine the trading price (i.e., the MCP), once the intersection is identified, one can select any suitable point (payment rule) within the interval $[\rho_L^{sl}, \rho_M^{by}]$ [41]. In our case, one can simply set $\rho_{mcp}(h) = (\rho_L^{sl} + \rho_M^{by})/2$ [20]. Once the trading price and the winners are found, different approaches can be applied to find the quantity of energy traded between each of the $L-1$ participating sellers and $M-1$ participating buyers [40]. We propose to apply the technique of [41], in which the entire traded volume is divided in a way to maintain the truthfulness of the auction. Using this approach, the total amount $q_s(x_s^{sl}(h))$ sold by any seller s, for a given strategy vector is:

$$q_s(x_s^{sl}(h)) := \begin{cases} x_s^{sl}(h), & \text{if } \sum_{s=1}^{L-1} |x_s^{sl}(h)| \leq \sum_{b=1}^{M-1} x_b^{by}(h) \\ [x_s^{sl}(h) - \sigma_s]^+, & \text{if } \sum_{s=1}^{L-1} |x_s^{sl}(h)| > \sum_{b=1}^{M-1} x_b^{by}(h) \end{cases} \quad (21)$$

where $\sigma_s = [\sum_{s=1}^{L-1} |x_s^{sl}(h)| - \sum_{b=1}^{M-1} x_b^{by}(h)]^+/(L-1)$ represents the fraction of the oversupply that is allotted to seller s. The mechanism in (21) implies that whenever the total demand at the auction's outcome exceeds the supply, then every seller would sell all of the energy that it introduced into the market. However, when the total supply exceeds the total demand, all the sellers receive an equal share of the oversupplied amount. Nonetheless, if for a seller $s1$, we have $\sigma_{s1} > |x_s 1^{sl}(h)|$, the seller does not sell any energy as per the second case in (21). So, for other sellers we have $\sigma_s = [\sum_{s=1}^{L-1} |x_s^{sl}(h)| - \sum_{b=1}^{M-1} x_b^{by}(h) - |x_s^{s1}(h)|]^+/(L-2)$. This scheme will be repeated as long as each seller sells a non-negative quantity [42]. Further, the fraction σ_s of each winner seller and all the power $x_s^{sl}(h)$ of the loser sellers (e.g., sellers with index $s \geq L$) is sold to the virtual buyer (i.e., the ISO) at the price $\rho_{iso}^{by}(h) < \rho_{mcp}(h)$. For the buyers, we have the same procedure, except the fraction σ_b and the power of all the losers (e.g., buyers with index $b \geq M$) is bought from the ISO at the price $\rho_{iso}^{sl}(h) > \rho_{mcp}(h)$.

Remark 1. *It is worth mentioning that, when the surplus/deficit energy of some prosumers are not significant (especially the nano-grids), one can easily develop a coalition among them, similar to work in [43]. Further, when the number of the M/NGs in the market is very large, one can split up the main market into several smaller sub-markets with still the same efficiency and features (IC, IR, BB, and high efficiency) according to the analysis in [44].*

Proposition 1. *The proposed auction mechanism has all the properties IC, IR, BB, and high efficiency. The proof is removed due to the space limitation, while it is routine and the same as the works in [20], [40], [45].*

C. Optimal Power Allocation

Once the trading amounts (from Section IV-A) and their corresponding prices (from Section IV-B) are determined, each seller will be indifferent to who buys his energy because his profit only depends on the energy quantity sold and the settled price. So, we can let the ISO determine who sells energy to whom in an efficient manner to minimize the transmission cost (e.g., power transmission loss). So, let $q_{sb}(h)$ denote the amount of energy provided by seller $s \in \mathcal{S}$ to buyer $b \in \mathcal{B}$ at slot h. Accordingly, we define the number of transmitter units (the inter-connector hops[8]) between seller s to buyer b by ℓ^{sb} and we denote the multi-hop transmission cost per each hop by c_{hp}[9]. Let p_{tc} denote the fixed transmission cost per kWh of energy. Then, the global transmission cost minimization problem at each slot h becomes:

$$\begin{aligned} &\min_{q_s^{sb}, s\in\mathcal{S}, b\in\mathcal{B}} C_{trans}(h) = \sum_{s\in\mathcal{S}} \sum_{b\in\mathcal{B}} p_{tc} q_{sb}(h) \ell^{sb} c_{hp} \\ &s.t.\ 0 \leq q_{sb}(h) \leq x_s^{sl}(h),\ \sum_{s\in\mathcal{S}} q_{sb}(h) = x_b^{by}(h), \\ &\sum_{b\in\mathcal{B}} q_{sb}(h) = x_s^{sl}(h) \end{aligned} \quad (22)$$

where, the ISO is located in both sets $\mathcal{S}$ and $\mathcal{B}$, as it is the virtual buyer/seller. The first constraint ensures that the power allocated to sell by seller s to each buyer b cannot exceed the total surplus power $x_s^{sl}(h)$, the second constraint ensures that the buyer b receives all his needed energy, and the third constraint ensures that the seller s sells all his surplus energy. The linear optimization problem (22) can be easily solved by some well-known optimization techniques [47]. However, if the number of prosumers is very large, one can develop a game theoretic mechanism between the sellers, by which, each seller

[8]Each hop reflects a transition of energy value and its associated information from one node to another [46].

[9]In our framework we can consider c_{hp} as the cost of sell power to a buyer through other intermediate prosumers' infrastructures (with ℓ^{sb} be a function of number of intermediate prosumers.) or the ISO's infrastructure (with ℓ^{sb} be a function of distance between the seller and the corresponding buyer).

Algorithm 1 The EMT Mechanism: Repeat for $h = 1, \cdots, H$.

I. **Load Demand and RER Generation Estimation: Executed by each prosumer** k
  Estimate the state of appliances $a \in \mathcal{A}_k^{off}$ and put them into non-schedulable A_k^{pas} or schedulable A_k^{act} groups according to (7).
  Define the uncertainty space for the joint RER generation profile as (8) using (4), (5), and the historical data records.
II. **The DSM Stage: Executed by each prosumer** k
  Receive $\boldsymbol{x}_{-k}^{net}(h)$ from the ISO.
  Use (14) to approximate effectual price $\rho_k^{eff}(h)$.
  Set discount factor γ_k^h.
  Choose the exploration parameter ϵ (e.g., similar to that in [39]).
  Choose optimal step-size $\alpha_k(h)$ (e.g., similar to that in [38]).
  Determine the state $\mathscr{S}_k(h) \in \mathbb{S}_k^h$ according to appliances' states and constraints, load demand, power resources, and the effectual price.
  Update the PDS $\bar{\mathscr{S}}_k^h(\boldsymbol{X}_k(h+1), h+1, \mathscr{E}_k^h) \in \mathbb{S}_k^h$ using new data about the state $\mathscr{S}_k(h)$ revealed at each slot h.
  Approximate the value of Q function using (19).
  Use (18), the value of Q function and (20) to determine the optimal stationary policy $\pi_k : \mathbb{S}_k^h \rightarrow \mathcal{E}_k^h$.
  Determine the energy consumption function $\mathscr{E}_k^h \in \mathcal{E}_k^h$ based on $\mathscr{S}_k(h)$ and $\mathscr{E}_k^h = \pi_k(\mathscr{S}_k^h)$.
  Determine consumption profile $\boldsymbol{X}_k(h) = \mathscr{E}_k^h(\mathscr{S}_k^h)$ accordingly.
  Calculate $x_k^{net}(h)$ and declare it to the ISO.
  If $x_k^{net}(h) \neq 0$ go to the energy trading stage,
  Else Stop the algorithm.
III. **The Energy Trading Stage:**
  The ISO as an auctioneer classifies prosumer k into set $\mathcal{S}$ of sellers if $x_k^{net}(h) < 0$ or set $\mathcal{B}$ buyers if $x_k^{net}(h) > 0$.
  All the sellers $s \in \mathcal{S}$ send their quantity of energy $x_s^{sl}(h)$ that they intend to supply and their reservation bid ρ_s^{sl} to the ISO.
  All the buyers $b \in \mathcal{B}$ send their quantity of energy $x_b^{by}(h)$ that they intend to buy and their proposed bid ρ_b^{by} to the ISO.
  Develop the hybrid double-auction mechanism in Section IV-B to determine the trading price (i.e., the MCP) $\rho_{mcp}(h) = (\rho_L^{sl} + \rho_M^{by})/2$.
  Determine the total amount of sold power $q_s(x_s^{sl}(h))$ by any seller s using (21).
  Determine the total amount of purchased power $q_b(x_b^{by}(h))$ by any buyer b similar to (21).
  The fraction σ_s (σ_b) of each winner seller (buyer) and all the power $x_s^{sl}(h)$ ($x_b^{by}(h)$) of the loser sellers (buyers) is sold to the ISO at the price $\rho_{iso}^{by}(h) < \rho_{mcp}(h)$ ($\rho_{iso}^{sl}(h) > \rho_{mcp}(h)$).
IV. **The Power Allocation Stage: Executed by the ISO**
  The ISO solves (22) and determines $q_{sb}(h)$.
  The rest of energy transactions are done directly with the ISO.

s can determine $q_{sb}(h),\ \forall\ b \in \mathcal{B}$ in a decentralized manner, similar to the work in [48].

Eventually, the whole energy management and trading process are performed as; 1) Each prosumer k predicts and schedules the energy consumption/production time/amount of its equipments through the optimal DSM mechanism Section IV-A and determines $x_k^{net}(h)$. 2) If $x_k^{net}(h) > 0 (< 0)$, the prosumer enters to the optimal energy trading mechanism Section IV-B as a seller(buyer) and infers the effectual price $\rho_k^{eff}(h)$ at which he will sell(buy) energy. 3) By the optimal power allocation mechanism Section IV-C, it is independently determined which seller s sell how much energy to which buyer b, i.e., $q_{sb}(h)$. This process is elaborated in Algorithm 1, and the whole smart micro/nano-grid model is depicted in Fig. 1.

V. NUMERICAL RESULTS

For evaluating the learning capability of the prosumers, we have considered 10 M/NGs each having 100 appliances randomly chosen between low/mid/high-flexible appliances and market clearing price of Pennsylvania-New Jersey-Maryland Interconnection (PJM) electricity market similar to the data in [49]. Each low/mid-flexible appliance has two possible actions (on and off) and the power consumption of each high-flexible appliance is quantized into 10 consumption (action) levels. The states of the appliances and the DS units are also assumed to have 100 different conditions which are updated at each slot (15 min) according to constraints (1) and (2). For the cost function of the equipment and ISO, simple quadratic functions are adopted, e.g., $f^{lf}(x) = f^{mf}(x) = g^{mf}(x) = f^{hf}(x) = f^{ds}(x) = g^{ds}(x) = f^{wsp}(x) = x^2$. As a benchmark, we assume that when there is no EMT mechanism in the system, the prosumers consume power once needed, sell/buy power only to/from the main grid, and cannot effectively manage the charge/discharge schedule of the DS units. Each seller is assumed to has a surplus between 50 kWh and 150 kWh that can be sold. The reservation prices of the sellers are chosen randomly from a range of [20, 60] cents per kWh, while the reservation bids of the buyers are chosen randomly from a range of [10, 70] cents per kWh. The demand of each buyer is chosen randomly within a range of [45, 200] kWh. The cost per energy sold is set to p_{tc} = 5 cents, ℓ^{sb} is chosen randomly from 1 to 10, and c_{hp} is chosen randomly from 5 to 10 cents for all $s \in \mathcal{S}$. All statistical results are averaged over all possible random values for the different parameters (prices, bids, demand, etc.) using a large number of independent simulation runs.

For analyzing the DSM part of the proposed EMT program, the behavior of a randomly chosen prosumer is depicted in Fig. 2. For the learning mechanism, we have considered the discount factor $\gamma = 0.9$, the step-size $\alpha = 0.2$, and the exploitation-exploration rate $\epsilon = 0.1$ for iteration (19). As we can see, the prosumer tries to consume low power at slots with high effectual prices (14) and self-generated powers, and consume more power at slots with low effectual prices and high self-generated powers. Further, the prosumer discharges the DS unite to sell power at the peak load demand at which the selling price is high, and charge them at the low-demand slots at which the buy price is low. In another simulation with 20 sellers and 20 buyers, the consequences of this consumption behavior are shown in Figs. 3(a) and (b). In these results, the aggregate cost and the utility level of each prosumer for one time-slot after the convergence of the learning algorithm are normalized to one. Fig. 3(a) presents the negative costs (profits) and utility levels of 20 sellers. As is clear, participating in the proposed EMT framework increases the profit of all the sellers compared to those when there is no EMT program. However, as denoted, the utility levels of all the sellers are reduced. This is due to shifting the consumption time/amount of some appliances to other time-slots, which imposes some discomfort to the end users. The same rule is always applied to the buyers' behavior. According to Fig. 3(b), the buyers should make a trade-off between reducing their payment and their utility level.

The effects of changing the parameters of the proposed learning framework on the convergence speed and level are analyzed in Fig. 4(a) for tuning the exploitation-exploration rate ϵ of a buyer $b \in \mathcal{B}$, and Fig. 4(b) for tuning the discount

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2945280, IEEE Transactions on Industrial Electronics

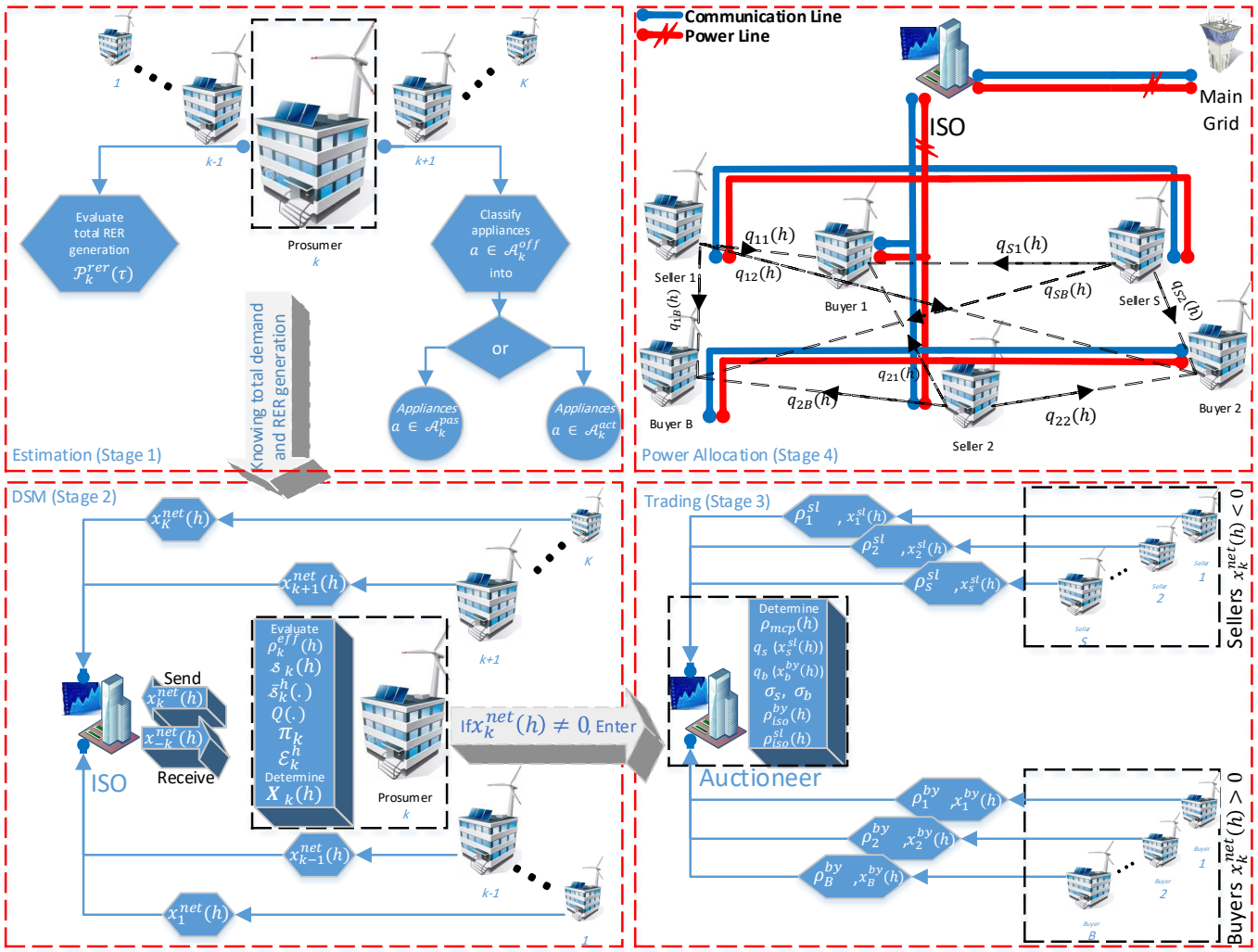


Fig. 1. Block diagram model of the proposed energy management and trading framework.

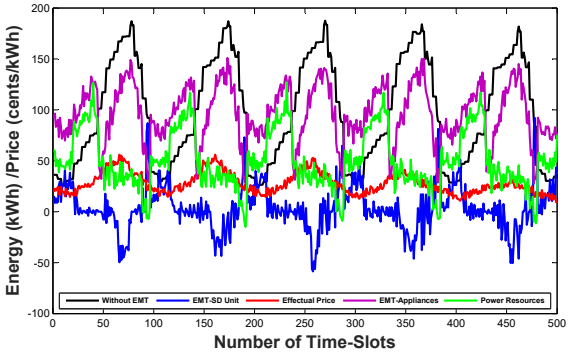


Fig. 2. The prosumer power consumption pattern with/without the EMT and the effectual price parameter.

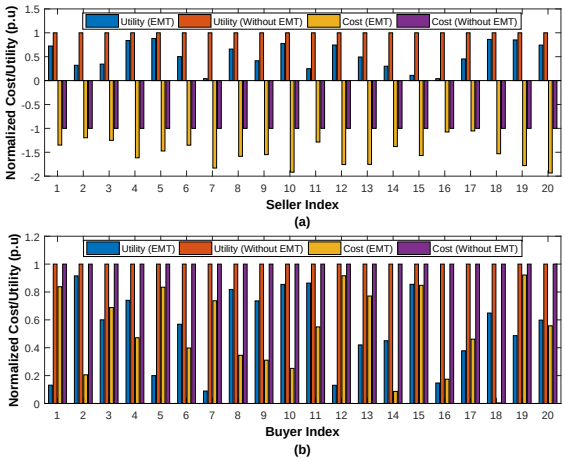


Fig. 3. Comparison of the aggregate costs and utility levels of the prosumers with/without applying the proposed EMT mechanism; a) Sellers evaluation, b) Buyers evaluation.

factor γ of a seller $s \in \mathcal{S}$. In Fig. 4(a), we can see that as the buyer sets smaller ϵ at each slot, the PDS algorithm adopts random actions with smaller probability and achieves a lower average payment by more searching the Q-function (19). That means, adopting smaller values for ϵ lets the PDS algorithm use the available and updated information more efficiently. The discount factor γ determines the importance of the future rewards. A factor of 0 makes the prosumer myopic (or short-sighted) by only considering the current reward, while a factor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2945280, IEEE Transactions on Industrial Electronics

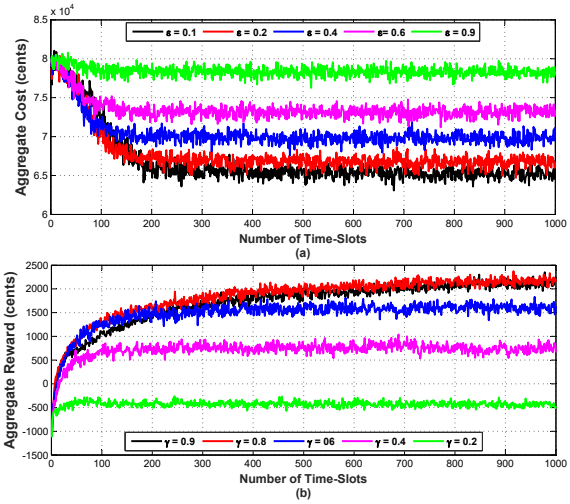


Fig. 4. Effects of parameter tuning on the learning behavior of the proposed reinforcement learning; a) Average cost of buyer b for the bought power, b) Average revenue of seller s for the sold power.

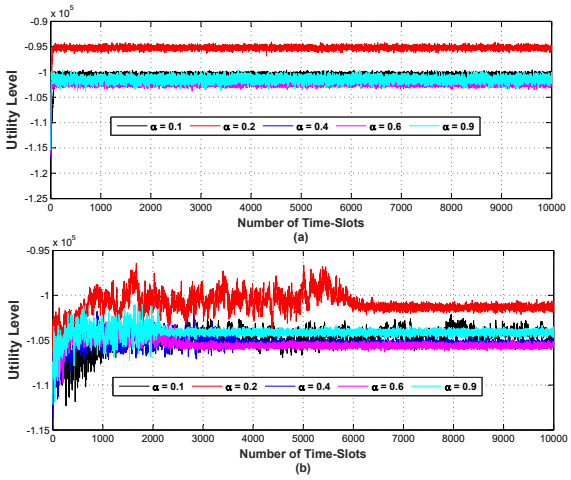


Fig. 5. Comparison between the performance of; a) The proposed PDS reinforcement learning and, b) The conventional Q-learning.

approaching 1 makes it strive for a long-term high reward. In Fig. 4(b), we can see that when the seller is myopic (sets lower γ_k), he achieves a lower average reward. This is because the seller does not consider the impact of his current actions on the future states of the equipment and the acquired rewards.

To justify using the PDS learning instead of the conventional Q-learning techniques, Figs. 5(a) and (b) are depicted. As illustrated by these figures, the PDS learning mechanism converges much faster and achieves a higher average utility level compared with the conventional Q-learning methods. So, although the PDS learning method imposes a more computational burden on the system, it is worth using that, as the proposed framework is online and fully distributed.

Given the outcome of the DSM part, Figs. 6(a) and (b) show the competition between 8 sellers and 6 buyers. In Fig. 6(a), the intersection point demonstrates that seller 4 and buyer 3 determine the trading price (MCP). The total amount sold by the participating sellers (seller 1 to 3) must be equal to the one bought by the participating buyers (buyer 1 and 2)

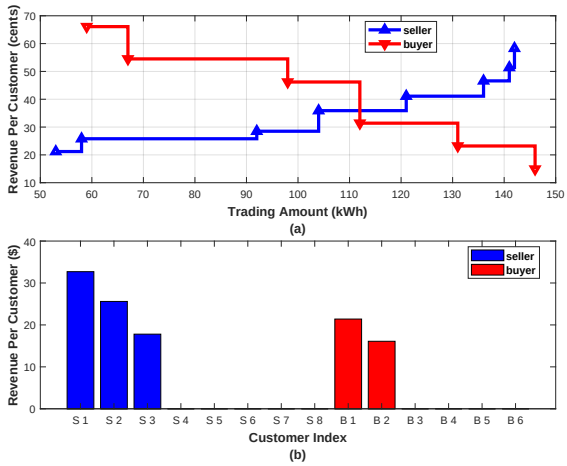


Fig. 6. Competition between 8 sellers and 6 buyers; a) The winner determination rule, b) The resulted revenue/frugality.

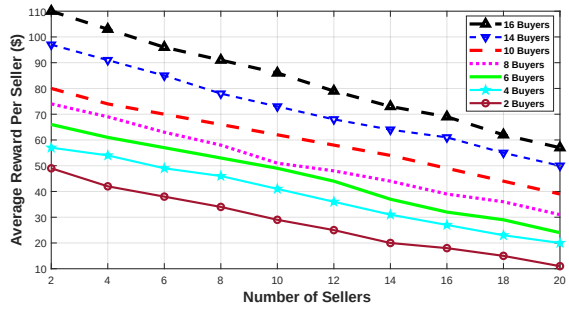


Fig. 7. Comparison of the average utility per seller resulted from the proposed trading mechanism for different numbers of buyers and sellers.

at the MCP. As the total procured power by the winning sellers is 203 kWh and the total demand of the winning buyers is 126 kWh, each seller must sell his over-supply fraction $(203 - 126)/3 = 25.6667$ kWh to the ISO with a lower price than the MCP $\rho_{mpc}(h)$. The revenue of the sellers for selling power with a higher MCP price and the frugality of the buyers for buying power with a lower MPC price compared with the ISO's high selling price $\rho^{sl}_{iso}(h)$ and low buy price $\rho^{by}_{iso}(h)$ are demonstrated in Fig. 6(b). Clearly, other losing prosumers acquire no revenue/frugality, since they trade all their energy with the ISO. Therefore, we can conclude that the acquired benefit from local energy trading is proportional to the number of market participants and the traded amount, as denoted in Fig. 7. From the results of this figure, we can see that the most profitable scenario for a seller is the condition with the least sellers and the most buyers. The reason for this is that the competition extremity between the sellers for supplying power is reduced.

The cost of transmitting power from some sellers to their buyers are illustrated in Fig. 8 for both with/without the EMT-optimal power allocation scenarios. The results reveal that by letting the ISO decide which seller sells power to which buyers, the power loss and transmitting costs are reduced significantly. This is because the ISO dispatches the surplus power of each seller to his nearest buyer neighbors to also reduce the destructive effects such as voltage-frequency rise

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2019.2945280, IEEE Transactions on Industrial Electronics

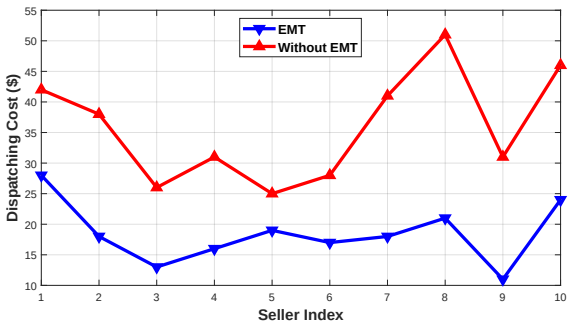


Fig. 8. Comparison between the two scenarios (without EMT and with EMT-optimal power allocation) in terms of the average dispatching cost.

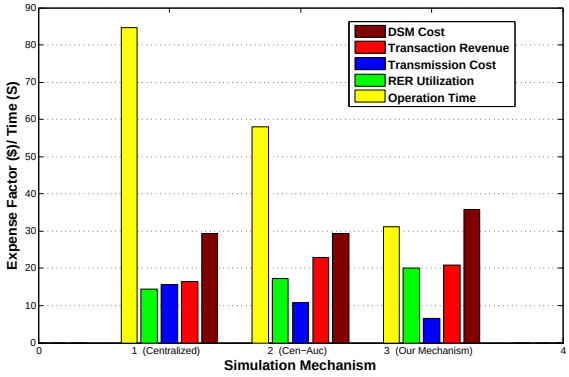


Fig. 9. Performance comparison between the centralized solution of (13), the centralized solution of (13) with auction ability (i.e., Cen-Auc case), and the proposed EMT framework.

problems.

Finally, the performance comparison results between the three different mechanisms for energy consumption management and trading framework is depicted in Fig. 9. As is clear from the operation time results, the centralized method in solving (13) is not practical in the real-time real world applications. This is because in a practical smart grid model we may have thousands of prosumers sending their private information to the central controller which can violate the privacy of the prosumers as well as increasing the operational time significantly. However, the aggregate cost imposed on the prosumers as the results of the DSM program is the lowest for the centralized method. This is because the centralized solutions (i.e., Centralized and Cen-Auc) converge to the global optimal point, while our mechanism converges to the vicinity of this point as it is autonomous and distributed. In terms of the transaction revenue, Cen-Auc is the most attractive scenario for the prosumers. In this case the DSM program is run at the central controller effectively and then the prosumers are allowed to trade directly with the other participants who have higher trading prices than the ISO (i.e., the centralized case). This can reduce the prosumers motivation for participating in the EMT program as they have to sell/buy energy to/from the ISO with a pre-determined values.

In terms of transmission cost, our proposed framework has the lowest expense factor as it tries to trade energy as locally as possible. However, in the centralized method all the energy are sold/bought directly from the ISO with increase in the transmission cost significantly. Cen-Auc has lower transmission cost than the centralized method as the prosumers are allowed to trade directly after the DSM program ends. But still the cost is higher than that of our method as the third stage (i.e., optimal power allocation) is not performed in this case. In terms of motivations for implementing the RER facilities, our method is the best one and acquires the highest revenue for the prosumers. This is because in the proposed method the prosumers can manage the RER power generation and autonomously decide whether they want to produce and sell the power with a preferred price or store it for a better (higher) price at the later slots. In the other cases they have no right to act autonomously. However, in the case of Cen-Auc, the RER utilization factor is slightly higher than the centralized case as the prosumers are allowed to declare their bides and have a chance to trade directly.

VI. Conclusions

A detailed characterization and formulation of the M/NGs' components have been provided in this paper. For the first time, a novel distributed hierarchical online method has been proposed by which the M/NGs can schedule their power consumption, efficiently manage their produced power, and trade the surplus/deficit energy with their neighbors and the ISO to make the profit. Further, an optimal power allocation has been introduced to reduce the power loss and the destructive effects (e.g., voltage and frequency rise problems) of the local energy trading as much as possible. Extensive numerical simulations have been carried out and verified that all the prosumers benefit from participating in it.

References

[1] H. Farhangi, "The path of the smart grid," *IEEE power and energy magazine*, vol. 8, no. 1, 2010.

[2] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart gridthe new and improved power grid: A survey," *IEEE communications surveys & tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[3] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE transactions on industrial informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[4] D. E. Olivares, A. Mehrizi-Sani, A. H. Etemadi, C. A. Cañizares, R. Iravani, M. Kazerani, A. H. Hajimiragha, O. Gomis-Bellmunt, M. Saeedifard, R. Palma-Behnke *et al.*, "Trends in microgrid control," *IEEE Transactions on smart grid*, vol. 5, no. 4, pp. 1905–1919, 2014.

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

[6] H. Wang and J. Huang, "Incentivizing energy trading for interconnected microgrids," *IEEE Transactions on Smart Grid*, 2016.

[7] H. Wang, T. Huang, X. Liao, H. Abu-Rub, and G. Chen, "Reinforcement learning in energy trading game among smart microgrids," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 8, pp. 5109–5119, 2016.

[8] Y. Yang, Q.-S. Jia, G. Deconinck, X. Guan, Z. Qiu, and Z. Hu, "Distributed coordination of ev charging with renewable energy in a microgrid of buildings," *IEEE Transactions on Smart Grid*, 2017.

[9] G. El Rahi, S. R. Etesami, W. Saad, N. Mandayam, and H. V. Poor, "Managing price uncertainty in prosumer-centric energy trading: A prospect-theoretic stackelberg game approach," *IEEE Transactions on Smart Grid*, 2017.

[10] S. Bahrami and M. H. Amini, "A decentralized framework for real-time energy trading in distribution networks with load and generation uncertainty," *arXiv preprint arXiv:1705.02575*, 2017.

[11] H. Wang, T. Huang, X. Liao, H. Abu-Rub, and G. Chen, "Reinforcement learning for constrained energy trading games with incomplete information," *IEEE transactions on cybernetics*, vol. 47, no. 10, pp. 3404–3416, 2017.

[12] S. Bahrami and M. H. Amini, "A decentralized trading algorithm for an electricity market with generation uncertainty," *arXiv preprint arXiv:1705.02577*, 2017.

[13] S. Bahrami, V. W. Wong, and J. Huang, "An online learning algorithm for demand response in smart grid," *IEEE Transactions on Smart Grid*, 2017.

[14] W. Zhong, K. Xie, Y. Liu, C. Yang, and S. Xie, "Auction mechanisms for energy trading in multi-energy systems," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1511–1521, April 2018.

[15] N. Liu, X. Yu, C. Wang, and J. Wang, "Energy sharing management for microgrids with pv prosumers: A stackelberg game approach," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1088–1098, 2017.

[16] N. Liu, X. Yu, C. Wang, C. Li, L. Ma, and J. Lei, "Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3569–3583, 2017.

[17] N. Liu, M. Cheng, X. Yu, J. Zhong, and J. Lei, "Energy-sharing provider for pv prosumer clusters: A hybrid approach using stochastic programming and stackelberg game," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6740–6750, Aug 2018.

[18] L. Xiao, X. Xiao, C. Dai, M. Pengy, L. Wang, and H. V. Poor, "Reinforcement learning-based energy trading for microgrids," *arXiv preprint arXiv:1801.06285*, 2018.

[19] T. Baroche, P. Pinson, R. L. G. Latimier, and H. B. Ahmed, "Exogenous cost allocation in peer-to-peer electricity markets," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2553–2564, July 2019.

[20] Y. Wang, W. Saad, Z. Han, H. V. Poor, and T. Başar, "A game-theoretic approach to energy trading in the smart grid," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1439–1450, 2014.

[21] R. B. Myerson and M. A. Satterthwaite, "Efficient mechanisms for bilateral trading," *Journal of economic theory*, vol. 29, no. 2, pp. 265–281, 1983.

[22] Q. Fu, L. F. Montoya, A. Solanki, A. Nasiri, V. Bhavaraju, T. Abdallah, and C. Y. David, "Microgrid generation capacity design with renewables and energy storage addressing power quality and surety," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 2019–2027, 2012.

[23] P. Samadi, V. W. Wong, and R. Schober, "Load scheduling and power trading in systems with high penetration of renewable energy resources," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1802–1812, 2016.

[24] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE power and energy magazine*, vol. 7, no. 2, pp. 52–62, 2009.

[25] T. Li and M. Shahidehpour, "Price-based unit commitment: A case of lagrangian relaxation versus mixed integer programming," *IEEE transactions on power systems*, vol. 20, no. 4, pp. 2015–2025, 2005.

[26] R. Palma-Behnke, C. Benavides, E. Aranda, J. Llanos, and D. Sáez, "Energy management system for a renewable based microgrid with a demand side management mechanism," in *Computational intelligence applications in smart grid (CIASG), 2011 IEEE symposium on*. IEEE, 2011, pp. 1–8.

[27] A. Sobu and G. Wu, "Optimal operation planning method for isolated micro grid considering uncertainties of renewable power generations and load demand," in *Innovative Smart Grid Technologies-Asia (ISGT Asia), 2012 IEEE*. IEEE, 2012, pp. 1–6.

[28] N. K. Paliwal, R. Mohanani, N. K. Singh, and A. K. Singh, "Demand side energy management in hybrid microgrid system using heuristic techniques," in *Industrial Technology (ICIT), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1910–1915.

[29] D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng, "Adaptive robust optimization for the security constrained unit commitment problem," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 52–63, 2013.

[30] P. D. Klemperer and M. A. Meyer, "Supply function equilibria in oligopoly under uncertainty," *Econometrica: Journal of the Econometric Society*, pp. 1243–1277, 1989.

[31] N. Li, L. Chen, and M. A. Dahleh, "Demand response using linear supply function bidding," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1827–1838, 2015.

[32] M. J. Osborne and A. Rubinstein, *A course in game theory*. MIT press, 1994.

[33] A.-H. Mohsenian-Rad, V. W. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE transactions on Smart Grid*, vol. 1, no. 3, pp. 320–331, 2010.

[34] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[35] Y. Shoham and K. Leyton-Brown, *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[36] R. Bellman, *Dynamic programming*. Courier Corporation, 2013.

[37] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[38] E. Even-Dar and Y. Mansour, "Learning rates for q-learning," *Journal of Machine Learning Research*, vol. 5, no. Dec, pp. 1–25, 2003.

[39] M. Tokic and G. Palm, "Value-difference based exploration: adaptive control between epsilon-greedy and softmax," in *Annual Conference on Artificial Intelligence*. Springer, 2011, pp. 335–346.

[40] D. Friedman and J. Rust, *The double auction market: institutions, theories, and evidence*. Westview Press, 1993, vol. 14.

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

[42] W. Saad, Z. Han, H. V. Poor, and T. Başar, "A noncooperative game for double auction-based energy trading between phevs and distribution grids," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE, 2011, pp. 267–272.

[43] M. Zeng, S. Leng, S. Maharjan, S. Gjessing, and J. He, "An incentivized auction-based group-selling approach for demand response management in v2g systems," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1554–1563, 2015.

[44] F. Kojima and T. Yamashita, "Double auction with interdependent values: Incentives and efficiency," *Theoretical Economics*, vol. 12, no. 3, pp. 1393–1438, 2017.

[45] O. Karaca, P. G. Sessa, N. Walton, and M. Kamgarpour, "Game theoretic analysis of auction mechanisms modeled by constrained optimization problems," *arXiv preprint arXiv:1711.06774*, 2017.

[46] U. M. Pešović, J. J. Mohorko, K. Benkič, and Ž. F. Čučej, "Single-hop vs. multi-hop–energy efficiency analysis in wireless sensor networks," in *18th Tlcommunictions Forum, TELFOR*, 2010.

[47] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997, vol. 6.

[48] N. Yaagoubi and H. T. Mouftah, "A distributed game theoretic approach to energy trading in the smart grid," in *Electrical Power and Energy Conference (EPEC), 2015 IEEE*. IEEE, 2015, pp. 203–208.

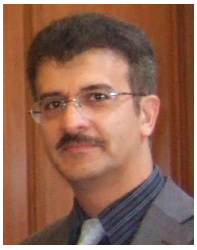
[49] M. Latifi, A. Khalili, A. Rastegarnia, and S. Sanei, "Fully distributed demand response using the adaptive diffusion–stackelberg algorithm," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2291–2301, 2017.

Milad Latifi received the M.Sc degree in communication engineering from Malayer University, Hamedan, Iran, in 2017. His research interests include advanced signal processing, adaptive filtering, cooperative learning, multi-agent networking, and adaptive optimization. Mr. Latifi is a student member of the IEEE.

Amir Rastegarnia completed his PhD degree in the electrical engineering at the University of Tabriz, Tabriz, Iran, in 2011. In 2011, he joined the Department of Electrical Engineering, Malayer University, as Assistant Professor. His current research interests are theory and methods for adaptive and statistical signal processing, distributed adaptive estimation, as well as signal processing for communications. He is a Member of IEEE.

Azam Khalili received the PhD degree in electrical engineering from the University of Tabriz, Tabriz, Iran, in 2011. In 2011, she joined the Department of Electrical Engineering, Malayer University, as Assistant Professor. Her current research interests are theory and methods for adaptive filtering, distributed adaptive estimation, as well as signal processing for communications. She is a Member of IEEE.

Wael M. Bazzi received the graduated degree from the American University of Beirut (AUB), Lebanon, in 1996, the ME degree from AUB, in 1999 and the PhD degree from the University of Waterloo, Canada, in 2001. He is currently an associate professor with the American University in Dubai. His research interests include wireless communication and networks, especially the optimization and modeling aspects of communication networks and systems.



Saeid Sanei (SM05) received his PhD in signal processing from Imperial College London, UK. He has been a member of academic staff in Iran, Singapore, and the UK. He has published three monograms, a number of book chapters, and over 320 papers in peer reviewed journals and conference proceedings. His research interest is in adaptive filtering, cooperative learning, multi-way, multimodal, and multichannel signal processing with applications to biomedical, audio, biometrics, and communication signals and images. He has served as an Associate Editor for the IEEE Signal Processing Letters, IEEE Signal Processing Magazine, and Journal of Computational Intelligence and Neuroscience.