A Multi-Agent Deep Reinforcement Learning Approach for a Distributed Energy Marketplace in Smart Grids

Arman Ghasemi, Amin Shojaeighadikolaei, Kailani Jones,
Morteza Hashemi, Alexandru G. Bardas, Reza Ahmadi
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
E-mails: {arman.ghasemi, amin.shojaei, kailanij, mhashemi, alexbardas, ahmadi}@ku.edu

***Abstract*—This paper presents a Reinforcement Learning (RL) based energy market for a prosumer dominated microgrid. The proposed market model facilitates a real-time and demand-dependent dynamic pricing environment, which reduces grid costs and improves the economic benefits for prosumers. Furthermore, this market model enables the grid operator to leverage prosumers' storage capacity as a dispatchable asset for grid support applications. Simulation results based on the Deep Q-Network (DQN) framework demonstrate significant improvements of the 24-hour accumulative profit for both prosumers and the grid operator, as well as major reductions in grid reserve power utilization.**

I. Introduction

Small-scale power generation and storage technologies, also known as Distributed Energy Resources (DERs), are changing the operational landscape of the power grid in a substantial way. Many traditional power consumers adopting a DER technology are starting to produce energy, thus morphing from a consumer to a *prosumer* (produces and consumes energy) [1]. The most common prosumer installations are the residential solar photovoltaic (PV) systems [2]. Although DER integration has the potential to provide multiple benefits to prosumers as well as grid operators [3], current grid operating strategies fail to leverage DER capabilities at a large scale, mostly due to the lack of modern and intelligent grid control strategies.

The residential PV systems likely have excess power generation during peak sun hours which usually do not coincide with peak demand hours [4]. In other words, current residential PV systems are likely to generate excess power during off-peak demand hours when electricity is not a valuable grid commodity, and this excess generation can even contribute to grid instability. Integration of energy storage into prosumer setups can potentially rectify this situation by allowing the prosumers to store their excess energy during the peak sun hours and inject it into the grid during the peak demand hours. Furthermore, proper coordination and aggregation of this dispatchable prosumers' generation capacity can be leveraged for various grid support services/applications [5], [6] .

Nevertheless, current popular net-metering compensation schemes do not properly incentivize the prosumers to engage in grid support applications [7]. The electricity meter in a net-metered household runs backwards when the prosumer injects power into the grid [8]. At the end of a billing cycle, the customer is billed for the "net" energy use, i.e., the difference between the overall consumed and produced energy, regardless of the actual schedule of injecting energy into the grid. Moreover, prosumers are compensated for the generated electricity at the same fixed retail price irrespective of the time of the day or any grid contingency at hand. Therefore, there is little incentive for prosumers to engage in any sort of grid support service.

In this paper, we propose a distributed energy marketplace framework that realizes a real-time, demand-dependent, dynamic pricing environment for prosumers and the grid operator. The proposed marketplace framework offers a plethora of vital properties to incentivize prosumers' engagement in grid support applications while providing improved economic benefits to prosumers as well as the grid operator, resulting in a "win-win" scenario. The contributions of the framework proposed in this paper can be summarized as follows,

- The proposed marketplace framework enables the grid operator to leverage prosumers' storage capacity as a dispatchable asset, while reducing grid cost through offsetting reserve power with prosumer generation.
- It incentivizes the prosumers to engage in grid support applications by providing higher economic benefits when supporting grid activities.
- Founded on a reinforcement learning (RL)-based decision-making, our framework handles the high dimensional, non-stationary, and stochastic nature of the problem without the need for abstract explicit modeling and deterministic rules used in traditional approaches.
- It models prosumers with generation, storage capacity, and bidirectional grid injection capability. This yields in a high degree of freedom for cost versus profit optimization and leads to improved overall benefits for all parties.

To enable all these properties, the proposed energy market leverages a multiagent RL framework with a single grid operator agent, and a network of distributed prosumer agents. The grid agent's goal is to maximize its economic benefit. To this end, the agent makes decisions on the optimal share of power purchased from a fleet of conventional generation facilities versus a cohort of prosumers with dispatchable generation capability, by considering the incremental cost of generation facilities versus the retail price of purchasing electricity from prosumers. In order to dispatch the prosumers' generation, the grid agent dynamically sets the retail electricity price to incentivize prosumers to adjust their generation level. On the other hand, the prosumer agents aim to maximize their own economic benefit by deciding on the level of grid support participation according to various factors such as electricity retail

978-1-7281-6127-3/20/$31.00 ©2020 IEEE

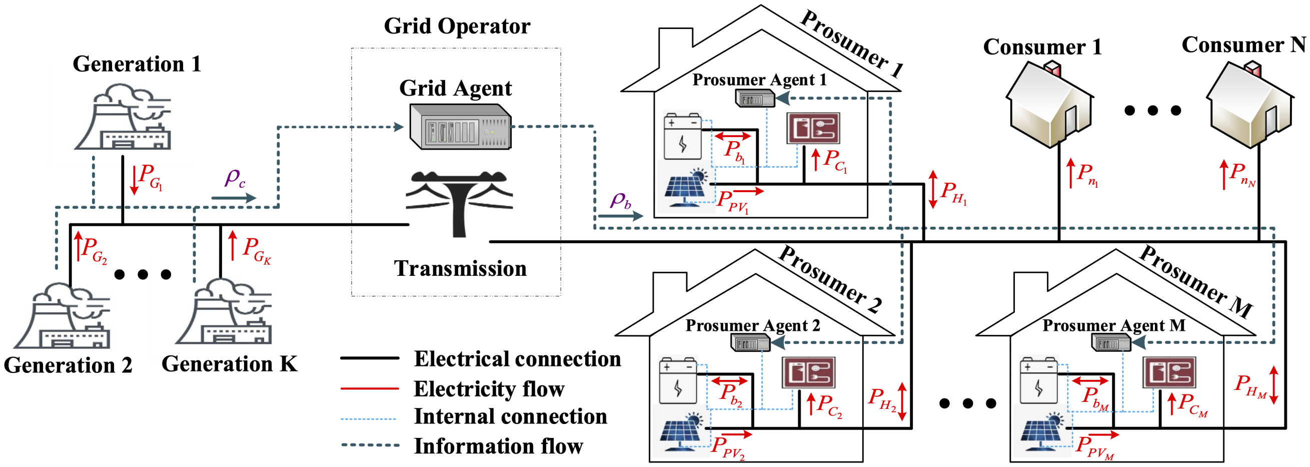


Fig. 1. Proposed electricity model market – The proposed energy marketplace includes several generation sources, household prosumers, and household consumers. By leveraging a reinforcement learning (RL) framework, our system enables a dynamic buy and sell pricing scheme handled by the grid as well as dynamic strategy for the prosumers to maximize benefits.

price, State of Charge (SoC) of storage device, PV generation level, household consumption level, etc. We demonstrate the efficiency of this marketplace through a simulation on a small scale microgrid as shown in Fig. 1. The microgrid [9] is under the management of a single grid operator entity and contains loads, distributed energy resources and/or storage devices that can be operated in a controlled and coordinated way.

This paper is structured as follows: Section II covers background and related works, while Section III provides the physical and learning system models for the proposed energy market place. Next, the simulation results for the small scale microgrid case study are presented in Section IV. Finally, Section V concludes this paper.

II. Background and Related Work

A brief survey of traditional energy marketplace models and dynamic pricing methods for smart grid applications is provided in [10]–[12]. On the other hand, research has explored RL-based energy market frameworks and dynamic pricing schemes that bring economic benefits to both costumer and grid operators. The authors in [13] proposed an RL algorithm that allows service providers and customers to learn pricing and energy consumption strategies without a priori of knowledge, leading to reduced system costs. Furthermore, [14] investigated an RL-based dynamic pricing scheme for achieving an optimal "price policy" in the presence of fast charging electric vehicles over the grid. In order to reduce the electricity bill of the residential customers, a mathematical model using RL for load scheduling was developed in [15], assuming that residential loads include schedulable loads, non-schedulable loads, and local PV generation.

More closely aligned to our paper are the works in [16] and [17]. [16] described an RL-based dynamic pricing, demand response algorithm using Q-learning approach for a hierarchical electricity market that considers both service providers and customers' profits as well as shows improvements in profiability and reduced costs. However, this work only examines regular customers without generation or storage capacity. The authors in [17] proposed an RL-based home energy management (HEM) framework which considers real-time electricity price and PV generation, and the framework achieve superior performance and cost-effective schedules for demand response in a HEM system. Nonetheless, the households in this work are modeled as traditional loads unable to sell back their excess power to the grid. Although the Electric Vehicle (EV) charging is modeled, the storage capacity of EVs is not leveraged for cost optimization, meaning the households do not have any energy storage capacity.

III. System Model

The proposed electricity market model is shown in Fig. 1. As pictured, this model encompasses a grid agent (GA) and several prosumer agents (PAs). The learning environment is a combination of governing equations of the grid and prosumer's physical systems, the operational limitations of the power grid and the prosumers, and external factors such as the time of day or PV generation level as explained in the physical model subsection below. Although consumers are depicted in Fig. 1, we do not consider them as an individual agent due to their constant consumption of energy.

Notations: We use the following notations throughout the paper. Bold letters are used for vectors, while non-bold letters are scalars. Sets are denoted by calligraphy fonts (e.g., $\mathcal{S}$). The grid and household variables are denoted by $(.)_{\mathrm{G}}$ and $(.)_{\mathrm{H}}$.

A. Physical System Model

Grid Operation: We assume a power system with K generators each with a power output level of P_{G_i} such that $i \in \{1, \ldots, K\}$, and M prosumers each with power injection level of P_{H_j} where $j \in \{1, \ldots, M\}$. In the context of an energy marketplace, the goal of the grid is to maximize its profit over a time horizon of T, which is denoted by $\psi_{\mathrm{G}}(T)$. The accumulative grid profit is then equal to the total grid revenue minus the total cost of operation, i.e.,

$$\psi_{\mathrm{G}}(T) = \Upsilon_{\mathrm{G}}(T) - \left\{ \sum_{i=1}^{K} \Omega_{G_i}(T) + \sum_{j=1}^{M} \Omega_{H_j}(T) \right\}. \quad (1)$$

In this case, $\Upsilon_{\mathrm{G}}(\cdot)$ denotes the accumulative grid revenue as a result of selling $P_D(t)$ of electricity to the loads at the selling price of $\rho_s(t)$ \$/kWh. Therefore, the accumulative revenue over a time horizon of T is defined as:

$$\Upsilon_{\mathrm{G}}(T) = \int_0^T P_D(t)\,\rho_s(t)dt. \quad (2)$$

Moreover, $\Omega_{G_i}(T)$ denotes the accumulative cost of buying electricity from the i^{th} generation facility. The $\Omega_{G_i}(T)$ is typically estimated using the incremental cost curves of the generation facilities. In addition to the cost of buying electricity from generation facilities, the grid is able to buy electricity from prosumers. Thus, the accumulative cost of buying electricity from the j^{th} prosumer is equal to:

$$\Omega_{H_j}(T) = \int_0^T P_{H_j}(t)\,\rho_b(t)dt \qquad \text{for} \quad P_{H_j}(t) > 0\ , \quad (3)$$

where $\rho_b(t)$ (in the unit of \$/kWh) is the price of purchasing electricity from prosumers, referred to as *buy price* hereinafter.

The GA's goal is to maximize (1) subject to the fundamental grid power balance equation,

$$P_D(t) - \sum_{i=1}^{K} P_{G_i}(t) - \sum_{j=1}^{M} P_{H_j}(t) = 0\ , \quad \forall t. \quad (4)$$

It should be noted that due to heterogeneous generation facilities, we assume that the output of the i^{th} facility is constrained by practical limitations such as:

$$P_{G_i}^{\min} \leq \ P_{G_i}(t) \leq \ P_{G_i}^{\max}, \quad \text{for } i = 1, ..., K. \quad (5)$$

Prosumer's Operation: A typical prosumer setup with a PV deployment and energy storage is shown in Fig.1. According to this figure, the goal of the j^{th} prosumer's agent is to maximize its own accumulative profit $\psi_{H_j}(T)$ defined as:

$$\psi_{H_j}(T) = \Upsilon_{H_j}(T) - \Omega_{H_j}(T), \quad (6)$$

where $\Upsilon_{H_j}(T)$ is the accumulative revenue of the j^{th} prosumer for selling electricity to the grid, and $\Omega_{H_j}(T)$ is the accumulative cost of buying electricity from the grid defined by:

$$\Upsilon_{H_j}(T) = \int_0^T P_{H_j}(t)\rho_b(t)dt \quad \text{for} \quad P_{H_j}(t) > 0, \quad (7)$$

$$\Omega_{H_j}(T) = \int_0^T P_{H_j}(t)\rho_s(t)dt \quad \text{for} \quad P_{H_j}(t) \leq 0. \quad (8)$$

Assuming that for the j^{th} prosumer, $P_{PV_j}(t)$ is the PV generation, $P_{bj}(t)$ is battery charge/discharge power, and $P_{Cj}(t)$ is the consumption power, the internal power balancing is then described as follows:

$$P_{H_j}(t) = P_{PV_j}(t) - P_{b_j}(t) - P_{C_j}(t)\,. \quad (9)$$

In order to model realistic scenarios, we also pose the following constraints on each of these parameters:

(i) If $P_{H_j}^{\max}$ is the maximum allowable power injection, then we have: $\left|P_{H_j}(t)\right| \leq \ P_{H_j}^{\max}$.

(ii) $P_{PV_j}^{\max}$ denotes the peak PV generation such that $0 \leq P_{PV_j}(t) \leq \ P_{PV_j}^{\max}$.

(iii) Given that $P_{b_j}^{\max}$ is the maximum allowable battery charge/discharge power, then $\left|P_{b_j}(t)\right| \leq \ P_{b_j}^{\max}$.

(iv) Assuming that ϕ_j is the State of Charge (SoC) of the battery, and $\phi_j^{\min}$ and $\phi_j^{\max}$ are the minimum and maximum allowable state of charge of battery, we have $\phi_j^{\min} \leq \phi_j \leq \phi_j^{\max}$. The state of charge of battery for the j^{th} prosumer is calculated from,

$$\phi_j(t) = \phi_j(0) + \frac{1}{C_{B_j}} \int_0^t P_{b_j}(\tau)d\tau, \quad (10)$$

where C_{B_j} is the battery capacity and $\phi_j(0)$ represents the initial SoC of the battery.

Next we describe a deep reinforcement learning framework to enable the grid and prosumers to dynamically take optimal actions at each time slot.

B. Reinforcement Learning Model

In this work, the dynamic pricing problem is formulated as a Markov Decision Process (MDP) such that given a state s^t at time t, the goal is choosing the *optimal action* for transitioning to a new state s^{t+1} at time $t+1$, where $s^t, s^{t+1} \in \mathcal{S}$ such that $\mathcal{S}$ is the set of all possible environment states. This problem can be viewed as an instance of Reinforcement Learning (RL) that is concerned with studying how an agent or a group of agents learn(s) the environment by collecting **observations**, choosing **actions**, and receiving **rewards**. Assuming that $\mathcal{A}$ is the set of feasible actions available to each agent, as a result of taking an action $a^t \in \mathcal{A}$, the agent receives an immediate reward r^t, and the environment transitions from the state s^t to s^{t+1}.

In the proposed energy marketplace, we have a set of agents denoted by $\mathcal{N} = \{\mathrm{GA}, \mathrm{PA}_1, \ldots, \mathrm{PA}_M\}$ in which GA is the grid agent and PA_j is the agent for prosumer j. Next, we provide details on the observations, actions, and rewards for each agent type (i.e., grid agent or prosumer agent). In this framework, all the continuous variables are discretized using a zero-order hold to find the values at each time slot t.

Grid Agent: The GA observes the following state variables:

(i) cost of buying electricity from K generation facilities at time t, which is denoted by $\boldsymbol{\omega}_G^t = [\omega_1^t, \ldots, \omega_K^t]$,

(ii) cost of grid operator for buying electricity from M prosumers, which is denoted by $\boldsymbol{\omega}_H^t = [\omega_{H_1}^t, \ldots, \omega_{H_M}^t]$,

(iii) the total grid demand P_D^t,

We use the notation s_{GA}^t to represent all observations of the grid agent at time t. Thus, based on the observations of the grid at time t, the grid agent **action** is to determine the electricity buy price. As described in the physical model, the buy price is denoted by $\rho_b^t \in \mathcal{A}_{GA}$, where $\mathcal{A}_{GA}$ is the finite set of available actions to GA (i.e., all possible buy prices).

The **reward function for the grid** at time t is defined as the grid profit, i.e.,

$$r^t_{\mathrm{GA}} = \upsilon^t_G - \left\{ \sum_{i=1}^{K} \omega^t_{G_i} + \sum_{j=1}^{M} \omega^t_{H_j} \right\} , \tag{11}$$

where υ^t_G denotes the grid revenue at time slot t as a result of selling P^t_D electricity, which is obtained by $\upsilon^t_G = P^t_D \times \rho^t_s$. In addition, $\omega^t_{G_i}$ is the grid cost to buy $P^t_{G_i}$ from the i^{th} generation facility at time slot t. The value of $P^t_{G_i}$ is obtained using incremental cost curve of the i^{th} generation facility. Finally, the grid cost to buy $P^t_{H_j}$ from prosumer j at time slot t is denoted by $\omega^t_{H_j}$ that can be calculated as,

$$\omega^t_{H_j} = P^t_{H_j} \times \rho^t_b \qquad \text{for} \quad P^t_{H_j} > 0. \tag{12}$$

Given the definition for immediate reward r^t_{GA}, the ultimate goal is to maximize the agent cumulative reward over an infinite time horizon that is also known as expected return:

$$\Gamma^t_{\mathrm{GA}} = \sum_{k=0}^{\infty} \gamma^k r^{t+k+1}_{\mathrm{GA}} , \tag{13}$$

where $0 \le \gamma \le 1$ is the discount rate for the grid agent.

Prosumer Agent: The prosumer agent j observes the following state variables:

(i) state of charge of battery that is denoted by ϕ^t_j,
(ii) PV generation denoted by $P^t_{PV_j}$,
(iii) buy price ρ^t_b determined by the grid agent,
(iv) local power consumption denoted by $P^t_{C_j}$.

Based on this set of observations, the charge/discharge command to the energy storage in prosumer j is the **action** determined by PA_j, which is shown by $\sigma^t_j \in \mathcal{A}_{PA_j}$. In this case, $\mathcal{A}_{PA_j}$ is the finite set of available actions to PA_j. The **reward** function for PA_j is defined as,

$$r^t_{PA_j} = \upsilon^t_{H_j} - \omega^t_{H_j}, \tag{14}$$

where $\upsilon^t_{H_j} = P^t_{H_j} \times \rho^t_b$ for $P^t_{H_j} > 0$ is the j^{th} prosumer's revenue from selling $P^t_{H_j}$ to the grid at time slot t and, $\omega^t_j = P^t_{H_j} \times \rho^t_s$ for $P^t_{H_j} \le 0$ is the j^{th} prosumer's cost from buying $P^t_{H_j}$ from the grid at time slot t. Similar to the grid agent, the j^{th} prosumer tries to maximize its infinite-horizon accumulative reward defined as:

$$\Gamma^t_{PA_j} = \sum_{k=0}^{\infty} \tilde{\gamma}^k_j r^{t+k+1}_{PA_j} , \tag{15}$$

where $0 \le \tilde{\gamma}_j \le 1$ is the discount rate for PA_j.

C. Q-Learning Framework

In this work, the agents use Deep Q-Network (DQN) to solve their respective MDPs and maximize their accumulative rewards in (13) and (15). The DQN algorithm uses deep learning for each agent using the bellman iterative equation. In particular, for the grid agent we have,

$$\begin{aligned} Q(s^t_{GA}, \rho^t_b) &\leftarrow Q(s^t_{GA}, \rho^t_b) + \\ &\alpha[r^{t+1}_{GA} + \gamma \max_{\rho^{t+1}} Q(s^{t+1}_{GA}, \rho^{t+1}_b) - Q(s^t_{GA}, \rho^t_b)] , \end{aligned} \tag{16}$$

and similarly, for the prosumer agent we have,

$$\begin{aligned} Q(s^t_{PA_j}, \sigma^t_j) &\leftarrow Q(s^t_{PA_j}, \sigma^t_j) + \\ &\tilde{\alpha}_j[r^{t+1}_{PA_j} + \tilde{\gamma}_j \max_{\sigma^{t+1}_j} Q(s^{t+1}_{PA_j}, \sigma^{t+1}_j) - Q(s^t_{PA_j}, \sigma^t_j)] , \end{aligned} \tag{17}$$

where α and $\tilde{\alpha}_j$ are the learning rates for GA and PA_j, respectively. The estimated Q-values are used to find the optimal policy that maximizes the accumulative rewards. The DQN framework for the grid and prosumer agents is illustrated in Algorithms 1 and 2, respectively.

Algorithm 1 Q-learning Algorithm for the Grid Agent

```
Initialize $Q(s^t_{GA}, \rho^t_{GA})$ to zero
for each Episode do
    for each Iteration do
        $t := t + 1$
        Set buy price $\rho^t_b$ according to policy $\pi_{GA}$
        Observe reward $r^{t+1}_{GA}$ at new state $s^{t+1}_{GA}$
        Update $Q(s^t_{GA}, \rho^t_b)$ using (16)
        $s^t_{GA} := s^{t+1}_{GA}$
    end for
end for
```

Algorithm 2 Q-learning Algorithm for the j^{th} Prosumer Agent

```
Initialize $Q(s^t_{PA_j}, \sigma^t_j)$ to zero
for each Episode do
    for each Iteration do
        $t := t + 1$
        Set charge/discharge $\sigma^t_j$ according to policy $\pi_{PA_j}$
        Observe reward $r^{t+1}_{PA_j}$ at new state $s^{t+1}_{PA_j}$
        Update $Q(s^t_{PA_j}, \sigma^t_j)$ using (17)
        $s^t_{PA_j} := s^{t+1}_{PA_j}$
    end for
end for
```

In this framework, to balance exploration versus exploitation, the epsilon greedy strategy π is used for GA and PA as follow [18],

$$\pi = \begin{cases} \arg\max\limits_{a^t} E\left[Q\left(s^t, a^t\right)\right] & \text{with probability } 1 - \varepsilon, \\ \text{random action} & \text{with probability } \ \varepsilon. \end{cases}$$

The probability of random actions ε starts at 1 for the first 300 episodes, and then decays to 0.01 over the training episodes.

IV. Case Study and Numerical Results

The proposed energy market place model is implemented on a small-scale microgrid system, illustrated in Fig 1, to demonstrate the operation of the agents and their effectiveness for improving the economic benefit of the grid operator and the prosumers. As pictured, the system under the study is comprised of two generation facilities ($K = 2$), three prosumers ($M = 3$) that host the PA_1 to PA_3 agents, the grid operator that hosts the grid agent (GA), and one nongenerational household (a.k.a., consumer, $N = 1$). The parameters of the system are tabulated in Table I. The employed PV generation and

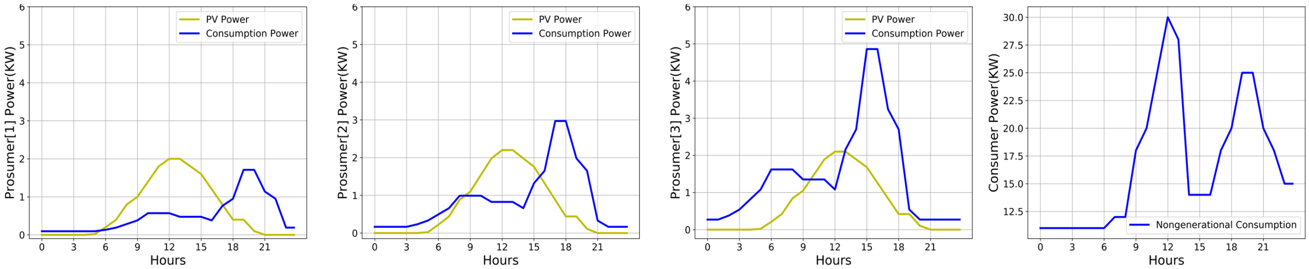


Fig. 2. Generation and consumption waveform sample for prosumers and consumer

Parameter	Description	Value
$P_{pv_j}^{\max}$	Max. PV Generation	[2-2.5] kW
$P_{b_j}^{\max}$	Max. allowable charge/discharge	2/-2 kW
$P_{H_j}^{\max}$	Max. allowable power injection	10 kW
$\phi_j^{\max}$	Max. state of charge	$0.9 \times C_{b_j}$
$\phi_j^{\min}$	Min. state of charge	$0.1 \times C_{b_j}$
C_{b_j}	Energy storage capacity	[8-10] kWh
$\phi_j(0)$	Initial state of charge	[3-4] kWh
ρ_s	Sell price [before 11am, after 11am]	[0.05, 0.095] \$/kWh
ρ_b^t	Buy price for agent-based scenario	$\{0.05, 0.06, 0.07,$ $0.08, 0.09, 0.1\}$\$/kWh
ρ_b^t	Buy price for conventional scenario	0.05 \$/kWh
$\left[P_{G_1}^{\min}, P_{G_1}^{\max}\right]$	Limitation of base generation	[5, 20] kW
$\left[P_{G_2}^{\min}, P_{G_2}^{\max}\right]$	Limitation of reserve generation	[0, 50] kW
$[\beta_1, \beta_2]$	Incremental cost of two generators	[0.03, 0.3] \$/kWh

TABLE I. Simulation parameters used for the proposed energy market place model on a small-scale microgrid

local consumption profiles for the last episode of the three prosumers are illustrated in Fig 2. These waveforms are constructed to be representative of real-world data available from California ISO website [4]. The peak value of generation and consumption for each prosumer is listed in Table I. The demand profile for last episode for the nongenerational household is also shown in Fig 2, and its peak value is listed in Table I. Each prosumer is equipped with an energy storage system (ESS) which includes a constant charge/discharge rate and a capacity provided in Table I.

In order to establish a baseline for the economic benefit of the grid operator and the households, a conventional system with a fixed buy price and no intelligent prosumer agents is simulated. In this scenario, the prosumers only sell electricity to the grid when their generation is more than their local consumption and their ESS is fully charged, which is likely to happen during the peak sun hours [19]. The described microgrid model for trading electricity between grid and residential loads is shown in Fig. 1. This scenario is referred to as the *conventional scenario*.

In the next scenario, we leverage the grid and prosumer agents to help implement the proposed market model, and these results are compared with the conventional scenario to demonstrate the economic improvements. This scenario is referred to as the *agent-based scenario*. In this work, we use PyTorch framework (v. 1.5.0 with Python3) to implement the DQN agents [20]. For training and testing the neural network, we leverage an Intel Xeon processor running at 3 GHz with 16 GB of RAM. The DQN algorithm hyperparameters used for simulations are provided in Table II. The simulations for both the conventional and agent-based scenarios are carried out via episodic iterations for 10,000 episodes. Each episode represents a 24 hour cycle and consists of 96 iterations, meaning that the simulation timeslots are 15 minutes.

Hyperparameters	Value for GA	Value for PA_j
Batch size	64	64
Discount factor	γ=[0.95-0.99]	$\tilde{\gamma}_j$=[0.95-0.99]
Learning rate	α=1e-3	$\tilde{\alpha}_j$=1e-3
Soft update interpolation	1e-5	1e-5
Hidden Layer-nodes	1-[1000]	2-[1000,1000]
Activation	Sigmoid	Sigmoid
Optimizer	Adam	Adam

TABLE II. DQN hyperparametrs

The action space for all prosumer agents (i.e., set $\mathcal{A}_{PA}$) includes three options: charge, no charge or discharge, and discharge. As a result, these actions command the battery power to one of the following three levels at each time slot t:

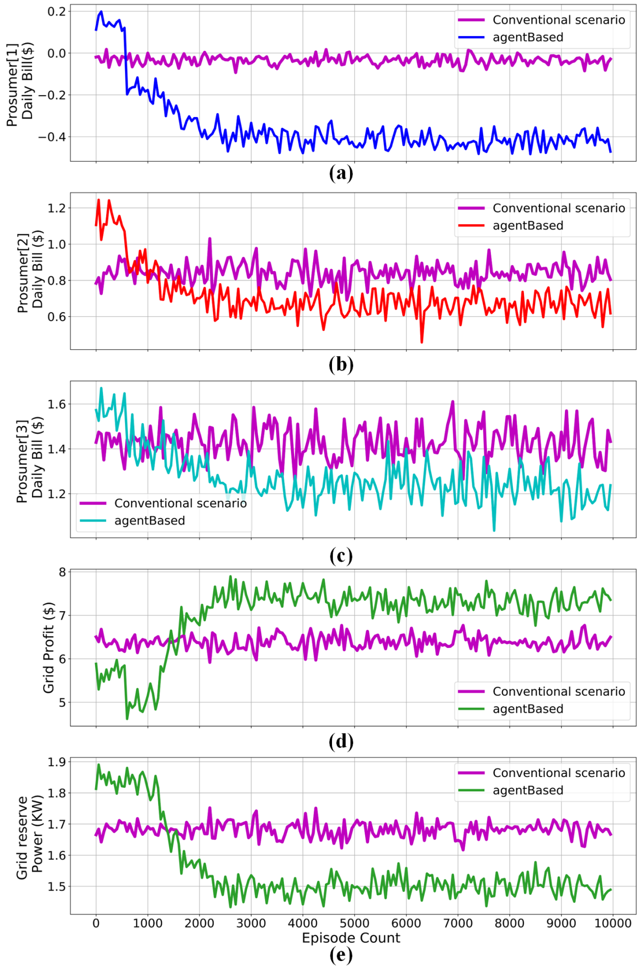
$$P_{b_j}^t = \begin{cases} P_{b_j}^{\max} & \text{Charge action,} \\ 0 & \text{No charge or discharge action,} \\ -P_{b_j}^{\max} & \text{Discharge action.} \end{cases} \qquad (18)$$

The action space for GA (i.e., buy price) is defined as $\mathcal{A}_{GA}$= $\{0.05,\ 0.06,\ 0.07,\ 0.08,\ 0.09,\ 0.1\}$ in which all numbers represent \$/kWh values. The sell price $\left(\rho_s^t\right)$ is defined at a constant rate in this work as provided in Table I. The incremental cost of the two generators in terms of \$/kWh are defined as,

$$\begin{cases} \omega_{G_1}^t = \beta_1 & \text{for} \quad P_{G_1}^{\min} \leq P_{G_1}^t \leq P_{G_1}^{\max} \\ \omega_{G_2}^t = \beta_2 & \text{for} \quad P_{G_2}^{\min} \leq P_{G_2}^t \leq P_{G_2}^{\max} \end{cases} . \qquad (19)$$

where $\beta_2 > \beta_1$ (see Table I). Consequently, the P_{G_1} provides baseline generation capacity at a lower incremental cost while P_{G_2} provides reserve capacity at a much higher cost.

The simulation results comparing the conventional and agent-based scenarios throughout 10,000 episodes are illustrated in Fig. 3 (a)-(c), where we compare the daily bill of the three prosumers over a 24-hour period. From the results, we note that while the daily bill resulting from a conventional scenario remains fairly constant throughout the episodes, the prosumer agents start converging to a lower bill as the agents explore the environment further and learn the optimal policy. As shown, the daily bill for households 1-3 are lowered by 1400%, 27%, and 13%, respectively. The unusually high daily bill reduction for household 1 is attributable to the conventional daily bill that is close to zero since the beginning (i.e., high PV generation), and the household's smaller peak consumption according to Fig. 2.



(a)

(b)

(c)

(d)

(e)

Fig. 3. Simulation results for conventional vs. agent-based scenarios over 10000 episodes:(a)-(c) 24-hour accumulative reward comparison for three prosumers, (d) grid 24-hour accumulative reward comparison, (e) grid reserve power utilization.

Fig. 3 (d)-(e) compare the accumulative grid profit and use of costly reserve power (PG2) over a 24-hour period. The agent-based scenario starts with a lower profit than the conventional scenario but converges to a much higher profit level than the conventional scenario as the agent learns the optimal policy. In this case, the grid profit improved around 15%. According to Fig. 3(e), the grid profit improvement is mostly attributable to the lower usage of costly reserve power in the agent-based scenario. In fact, in this experiment, the grid agent learns to rely on the prosumers' generation for balancing the grid's power rather than using the reserve power which is more expensive. The use of reserve power is decreased by 10% in this experiment.

V. Conclusions

This paper proposes an RL-based distributed energy marketplace framework that enables a real-time, demand-dependent, dynamic pricing environment to incentivize prosumers' grid support engagement while improving the economic benefit of both, prosumers and the grid operator. Simulation results, when implementing the proposed market model, show major economic improvements for the prosumers and the grid (through a reduced reserve power utilization by the grid).

References

[1] US Energy Department. Consumer vs prosumer: What's the difference? Accessed 5/2020. [Online]. Available: https://www.energy.gov/eere/articles/consumer-vs-prosumer-whats-difference

[2] "Annual energy outlook 2019 with projections to 2050," US Energy Information Administration, Tech. Rep., 2019. [Online]. Available: https://www.eia.gov/outlooks/aeo/pdf/aeo2019.pdf

[3] G. El Rahi, W. Saad, A. Glass, N. B. Mandayam, and H. V. Poor, "Prospect theory for prosumer-centric energy trading in the smart grid," in *2016 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2016, pp. 1–5.

[4] California ISO. Current and forecasted demand. [Online]. Available: http://www.caiso.com/TodaysOutlook/Pages/default.aspx

[5] M. Ruiz-Cortés, E. González-Romera, R. Amaral-Lopes, E. Romero-Cadaval, J. Martins, M. I. Milanés-Montero, and F. Barrero-González, "Optimal charge/discharge scheduling of batteries in microgrids of prosumers," *IEEE Transactions on Energy Conversion*, vol. 34, no. 1, pp. 468–477, 2019.

[6] O. Ciftci, M. Mehrtash, F. Safdarian, and A. Kargarian, "Chance-constrained microgrid energy management with flexibility constraints provided by battery storage," in *2019 IEEE Texas Power and Energy Conference (TPEC)*, 2019, pp. 1–6.

[7] G. C. Christoforidis, I. P. Panapakidis, T. A. Papadopoulos, G. K. Papagiannis, I. Koumparou, M. Hadjipanayi, and G. E. Georghiou, "A model for the assessment of different net-metering policies," *Energies*, vol. 9, no. 4, 2016.

[8] A. Poullikkas, "A comparative assessment of net metering and feed in tariff schemes for residential pv systems," *Sustainable Energy Technologies and Assessments*, vol. 3, pp. 1 – 8, 2013.

[9] B. Nordman, "Local grid definitions," Smart Grid Interoperability Panel and Lawrence Berkeley National Laboratory, Berkeley,USA, Tech. Rep., 2016.

[10] M. Khoshjahan, M. Soleimani, and M. Kezunovic, "Optimal participation of pev charging stationsintegrated with smart buildings in the wholesale energy and reserve markets," in *IEEE Power Energy Society Innovative Smart Grid Technologies*, 2020, pp. 1–5.

[11] I. S. Bayram, M. Z. Shakir, M. Abdallah, and K. Qaraqe, "A survey on energy trading in smart grid," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 258–262.

[12] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, "Load forecasting, dynamic pricing and dsm in smart grid: A review," *Renewable and Sustainable Energy Reviews*, vol. 54, 2016.

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

[14] C. Fang, H. Lu, Y. Hong, S. Liu, and J. Chang, "Dynamic pricing for electric vehicle extreme fast charging," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2020.

[15] T. Remani, E. A. Jasmin, and T. P. I. Ahamed, "Residential load scheduling with renewable generation in the smart grid: A reinforcement learning approach," *IEEE Systems Journal*, vol. 13, no. 3, pp. 3283–3294, 2019.

[16] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Applied Energy*, vol. 220, pp. 220–230, 2018.

[17] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning based data-driven method for home energy management," *IEEE Transactions on Smart Grid*, pp. 1–1, 2020.

[18] F.-L. Vincent, H. Petr, R. Islam, G. Marc, and P. Loelle, "An introduction to deep reinforcement learning," *Foundations and Trends in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[19] Q. Sun, M. E. Cotterell, Z. Wu, and S. Grijalva, "An economic model for distributed energy prosumers," in *Proceedings of the 46th Annual Hawaii International Conference on System Sciences*, 2013.

[20] N. Naderializadeh and M. Hashemi, "Energy-aware multi-server mobile edge computing: A deep reinforcement learning approach," in *53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 383–387.