



# Deep Reinforcement Learning-Aided Bidding Strategies for Transactive Energy Market

Amirheekmat Taghizadeh , Mina Montazeri, and Hamed Kebriaei , *Senior Member, IEEE*

**Abstract**—The concept of transactive energy market (TEM) has been introduced to efficiently balance supply and demand across the electrical networks in a distributed manner. TEM allows consumers to sell/purchase a portion of their excess/lack of energy to/from other local consumers. In this article, a transactive framework is proposed for a distribution network in which transactive agents participate in a local market and submit their bids to local TEM for day-ahead planning. However, due to complexity of the market and effects of a multitude of factors in the market outcome, deriving the optimal bid is not a trivial task. In order to learn the optimal bidding strategy in such a complex system with incomplete information, soft actor-critic method is utilized. Two scenarios are analyzed and compared. The first scenario assumes only one agent uses the learning method, while in the second scenario a number of residents form a coalition with a representative learning agent. Then, allocation of power assigned to the whole coalition between its residents is discussed. Comprehensive simulations are conducted for the 37-bus IEEE distribution system in two different scenarios to compare nonlearner agents with learner agents in both scenarios.

**Index Terms**—Bidding strategies, deep reinforcement learning (DRL), distribution network, transactive energy market (TEM).

## I. INTRODUCTION

### A. Motivation

**E**VOLVING equipment in the power grid, such as distributed energy resources [1], distributed energy storage [2], plug-in hybrid electric vehicle [3], and smart electrical tools and appliances [4], is creating new challenges for grid operation. Moreover, due to the transformation of users from a single producer or consumer to a “prosumer,” which is a consumer equipped with photovoltaic (PV) and other renewable energy sources, the traditional power system is confronting new challenges of how to manage these new distributed agents. Coordinating these distributed prosumers on the one side, and uncertainty in power generation due to uncertainties in weather

forecasts, on the other side, impose much complexities to the power distribution system [5]. The concept of transactive energy has been emerged as a solution to the aforementioned challenges [6]. The term “transactive energy” was first well defined by the GridWise Architecture Council as a system of economic and control mechanisms that allows the dynamic balance of supply and demand across the entire electrical infrastructure using value as a key operational parameter [7].

### B. Proposed Approach

In this article, a framework for prosumers’ bidding and local transactive market clearing is proposed. Each transactive energy market (TEM) is placed in a distribution network bus. The distribution system operator (DSO), which is responsible for the safe and efficient operation of the distribution network, provides a linear supply curve based on the local marginal price (LMP) and historical data from TEM and typical loads at each bus. The TEM operator gathers transactive agents’ bids, which are linear demand functions of price, and clears TEM by balancing the aggregated local demand with the supply curve provided by DSO. Two types of agents are considered in our bidding framework: a single residential house and a coalition of residential houses, such as a residential complex with multiple units. In the coalition agent case, the goal is to maximize the social welfare of all coalition members. In both cases, an agent submits a bid to TEM and receives a cleared power at a cleared price. In the case of multiple residents, a closed-form solution is derived to optimally allocate total cleared power of coalition to residents.

Bidding optimally is not an easy task to achieve, due to the complexity of the market clearing process and numerous affecting factors, many of which are not accessible by the prosumers directly. These factors include uncertainties of environmental parameters and incomplete information about other agents’ bidding strategies. Thus, the lack of information and uncertainties call for reinforcement learning (RL) approaches to be used. RL is a model-free method that learns how to act based on the information available and the feedback it receives from the environment. Continuous information and bidding parameters require a continuous version of RL. Moreover, too many influencing factors and complexity of optimal bidding makes it hard for traditional RL methods to cope with this task. To this end, soft actor-critic (SAC) method [8], which is a deep reinforcement learning (DRL) method, is proposed.

Manuscript received March 2, 2021; revised September 5, 2021; accepted January 17, 2022. This work was supported in part by the Institute for Research in Fundamental Sciences (IPM) under Grant CS 1400-4-451. (Amirheekmat Taghizadeh and Mina Montazeri are co-first authors.) (Corresponding author: Hamed Kebriaei.)

Amirheekmat Taghizadeh and Mina Montazeri are with the School of ECE, College of Engineering, University of Tehran, Tehran 14395-515, Iran (e-mail: taghizadeh.amirheekmat@gmail.com; montazeri70@gmail.com).

Hamed Kebriaei is with the School of ECE, College of Engineering, University of Tehran, Tehran 14395-515, Iran and also with the School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran 19395-5746, Iran (e-mail: kebriaei@ut.ac.ir).

Digital Object Identifier 10.1109/JSYST.2022.3145102

Comprehensive simulations for both agent types are conducted to demonstrate efficiency of DRL method for bidding.

### C. Related Works

1) *TEM Literature Review*: There are two main categories for approaching TEM: peer-to-peer and community-based markets. In peer-to-peer markets, on the one hand, each prosumer engages in negotiations with other prosumers and ultimately trades with other prosumers directly [9]. On the other hand, a community market is a collection of prosumers who collaborate with each other and act as a single agent from the grid's point of view. Although peer-to-peer schemes provide substantial flexibility for prosumers to act according to their preferences, they have several challenges in scalability, slow convergence, and satisfaction of operation constraints [10]. To mitigate these problems, the community-based markets are considered in the literature. Gathering all agents' bids and clearing the market is more efficient than peer-to-peer trading, both economically and practically [10]. Cornélusse *et al.* [11] implemented an internal aggregator within the community, which maximizes the social welfare based on marginal pricing. The aggregator collects the community members' preferences and solves the resource allocation problem centrally. The main problem with this approach is the lack of autonomy and privacy for prosumers. Muttaqi *et al.* [12] also proposed a two-stage market. At the first stage, the day-ahead market is cleared by running optimal power flow (OPF). In the second stage, consumers decide on their bids in the real-time market. However, the problem of optimal bidding strategy is not studied in this article. Li *et al.* [13] proposed a TEM framework in which the agents announce their cost functions to the market operator and then the distribution network optimization is solved cooperatively in a decentralized way. Renani *et al.* [14] proposed a distributed system operator framework for a transactive market that can reduce the supply cost of prosumers in a local distributed area and increase generation companies' payoffs.

2) *Bidding Strategy Using RL*: There are a number of research works trying to utilize RL methods for bidding strategy in the wholesale market. The use of simulated annealing Q-learning for bidding in a market with the Vickrey pricing rule is discussed in [15]. In this work, the corresponding Q-table is filled in an iterative trading process, with revenue maximization as the generator's goal. Noncooperative Markov game is used in [16] to cope with the incomplete information of the electricity market. In this work, discrete multiagent Q-learning is used to aid optimal generators bidding. The technique of agent-based computational economics for bidding is used in [17]. Using an RL agent with adaptive actor-critic mapping augmented with particle swarm optimization method, this work optimizes the revenue of generator companies in the wholesale market. Ye *et al.* [18] utilized DRL in the wholesale market to aid generators bidding optimization only using limited information, which is practically available. They have employed a deep policy gradient with an innovative long short-term memory network to model continuous state and action present in the bidding market. Lu *et al.* [19] used convolutional neural network (CNN) to find optimal power trading between multiple microgrids with battery and a power plant.

The multiagent deep deterministic policy gradient (DDPG) method was used by Du *et al.* [20] to approximate the Nash equilibrium of bidding game between generation companies.

Although RL methods have been used for bidding in wholesale markets, to the best of our knowledge, there is no work applying these methods for TEM. Moreover, the methods, such as Q-Learning, CNN, and DDPG, have some drawbacks that make them not very suitable for the task of bidding in TEM. The high number of influencing factors that creates a complex environment and bidding for 24 h ahead that requires a high action space dimension makes the learning process complex; thus, these methods require high learning time and a large number of samples to learn the optimal actions. Moreover, tuning model parameters is often very hard for these methods because these methods have poor convergence if their parameters are not set correctly. Tuning these parameters often requires much time and it gets worse as the problem gets more complicated. The preferred method should have a high sample efficiency and performance while being robust in the face of model parameters. Also, the preferred method should be able to handle complex environment with high dimension continuous action space. In this article, we deploy SAC, which is a state-of-the-art DRL algorithm that uses an off-policy actor-critic method based on the maximum entropy framework, which has been shown to outperform many other methods robustly in a variety of tasks [8].

### D. Contributions and Organization of This Article

The main contributions of this article are as follows.

- 1) A framework for TEM is proposed in which transactive agents in the distribution network submit their daily bids to TEM and TEM is cleared among the local transactive agents, nontransactive agents, and DSO.
- 2) We propose a DRL module with the SAC algorithm to solve the bidding problem in TEM.
- 3) Single and coalition agents are considered, and in the latter case, a closed form optimal allocation solution is obtained that enables the coalition agent to allocate the cleared power to its members.
- 4) Comprehensive simulations are conducted on the 37-bus IEEE distribution network to show the efficacy of method.

The rest of this article is organized as follows. Section II describes the system model, which includes market overview, market clearing, and transactive agent modeling. In Section III, the proposed DRL technique is presented. In Section IV, two different simulation scenarios are simulated. Finally, Section V concludes this article.

## II. SYSTEM MODEL AND PROBLEM FORMULATION

### A. Market Overview

Fig. 1 depicts the distribution network under study, which consists of traditional loads and TEMs at different locations. DSO is responsible for running the power flow and ensuring that the operational constraints of the system are met. Each bus consists of numerous transactive and traditional users. Traditional users draw electrical power from the grid at a constant

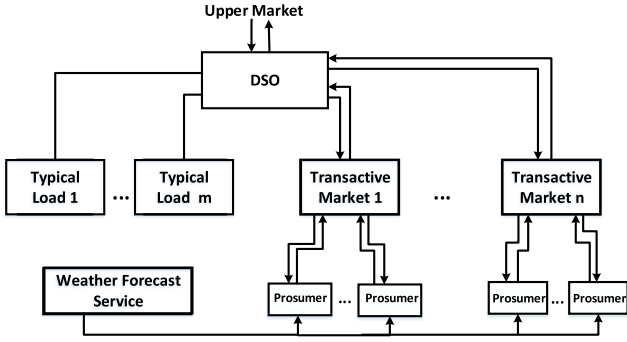


Fig. 1. Schematic of distribution network.

price, without any predefined consumption demand. On the other hand, a prosumer is a transactive participant that takes part in the day-ahead TEM by submitting a bidding profile to the market. Moreover, this type of agent is equipped with proper equipment that enables it to inject its excess energy production into the grid.

As illustrated in Fig. 1, prosumers are connected to a TEM, which is responsible for clearing the market and acts as an intermediary between prosumers inside a TEM and DSO. DSO offers electricity at a supply curve based on LMP to TEM. The TEM operator gathers the prosumers' bids and balances the aggregated demand of local TEM with DSO's supply curve, determining TEM price and energy exchange of each prosumer. After TEM is cleared, the aggregated demand of the transactive market is reported to DSO, which then is reported to the upper market. The day-ahead nature of the transactive market allows DSO for better planning and operation, in terms of performance and robustness.

### B. Transactive Market Clearing

At each day, transactive agents submit their 24 bids for the next day to TEM. Agent  $j$ 's bid is considered to be in the linear form  $p_j^r = a_j \lambda + b_j$ , where  $a_j$  and  $b_j$  are bidding parameters of transactive agent  $j$ ,  $\lambda$  is the unit electricity price, and  $p_j^r$  is the power consumed by transactive agent  $j$ . Note that if  $p_j^r > 0$ , the transactive agent is consuming power, and if  $p_j^r < 0$ , the transactive agent is injecting power to the grid. The cleared price of the market determines the power exchange of a transactive agent. If the price is low, the agent buys more electric power to consume or save in its battery. However, if the price is high, the agent forgoes the power consumption utility in favor of revenue achieved by selling power to grid. The clearing is performed by the TEM operator and consists of balancing the supply and demand of the agents. Considering all agents bids and  $p_{DSO} = a_{DSO} \lambda + b_{DSO}$ , which is DSO's supply curve, TEM is cleared by the following equation:

$$p_{DSO} + \sum_{j \in M} p_j^r = \left( a_{DSO} + \sum_{j \in M} a_j \right) \lambda + \left( b_{DSO} + \sum_{j \in M} b_j \right). \quad (1)$$

Balancing consumption and injection of energy into the network, we obtain

$$\lambda_{\text{cleared}} = - \frac{b_{DSO} + \sum_{j \in M} b_j}{a_{DSO} + \sum_{j \in M} a_j}. \quad (2)$$

After the price is cleared, the power exchange of every transactive agent and DSO is specified. We denote the cleared power by  $p_{\text{cleared}}^j$ , which is easily calculated by knowing the cleared price  $\lambda_{\text{cleared}}$  and the bid line.

### C. Transactive Agent Modeling

Transactive agents are residential consumers, possibly equipped with solar panels, batteries, household appliances, and suitable equipment that allows bidirectional power exchange with the grid. A transactive agent in TEM could be a single residential consumer or representative of a building complex with a number of residential consumers.

1) *Single Resident Agent*: The single resident agent's day cost function is represented as follows:

$$\begin{aligned} J^j[k] &= \sum_{t=1}^{24} J^j[k, t] \\ &= \sum_{t=1}^{24} \{ v_1^j (p_{\text{des}}^j[k, t] + p_{\text{bat}}^j[k, t] - p_{\text{cleared}}^j[k, t] - p_{\text{solar}}^j[k, t])^2 \\ &\quad + (\lambda_{\text{cleared}}[k, t] p_{\text{cleared}}^j[k, t]) + v_2^j f(p_{\text{bat}}^j[k, t]) \} \end{aligned} \quad (3)$$

where  $p_{\text{des}}^j[k, t]$  is the desired power consumption,  $p_{\text{bat}}^j[k, t]$  is the battery power, and  $p_{\text{solar}}^j[k, t]$  is the solar power generation which is predicted by knowing the weather condition predictions acquired from forecasting services, all for agent  $j$  at day  $k$  and hour  $t$ . The first term represents the loss in utility due to deviation from desired power consumption.  $v_1^j$  is a user-defined parameter that converts dissatisfaction of not consuming desired amount of electricity into monetary value. The second term represents utility loss (or gain) due to the money exchange, and the third term, denoted by  $f$ , models the degradation of battery due to charging and discharging.  $f$  can be modeled as  $f(p) = Dp^2 + Ep + F$ , where  $D$ ,  $E$ , and  $F$  are positive constants depending on the nominal voltage and the parameters of capacity loss function [21].  $v_2^j$  is a parameter that converts degradation of battery into monetary value. Battery's state of the charge (SoC) determines the amount of energy stored in battery and follows (4)

$$\text{SoC}^j[k, t + 1] = \text{SoC}^j[k, t] + p_{\text{bat}}^j[k, t] \quad (4)$$

$$\underline{\text{SoC}}^j \leq \text{SoC}^j[k, t] \leq \overline{\text{SoC}}^j \quad \forall k, t, j \quad (5)$$

where  $\underline{\text{SoC}}^j$  and  $\overline{\text{SoC}}^j$  are the min and max allowed SOC.

2) *Multiple Residents Agent*: For the case of multiple residents, also denoted by the coalition, the representative agent of the coalition tries to maximize the social welfare of its residential consumers. The learning scheme in this case is similar to a single resident agent. However, a problem that arises in this case is how to allocate the energy acquired by the coalition agent between its residential consumers after TEM is cleared each hour. To address

this issue, the coalition agent solves a local optimization problem among its residents to determine the share of each resident and the amount of money they have to pay.

Therefore, we consider the following coalitional objective function that includes the social welfare of the residents according to their desired electricity consumption, the billing cost, and the degradation cost of the central storage unit in the building:

$$\begin{aligned} J_C^l[k] &= \sum_{t=1}^{24} J_C^l[k, t] \\ &= \sum_{t=1}^{24} \left\{ \sum_{j \in N_C^l} \{v_1^j (p_{\text{alloc}}^j[k, t] + p_{\text{solar}}^j[k, t] - p_{\text{desire}}^j[k, t])^2 \right. \\ &\quad \left. + \lambda_{\text{cleared}}[k, t] p_{\text{alloc}}^j[k, t] \} + v_2^l f(p_{\text{bat}}^l[k, t]) \right\} \end{aligned} \quad (6)$$

where the first term represents the cost of deviating from consuming the desired amount of electricity, the second term represents the cost of electricity consumed, and the third term represents the battery degradation cost.  $N_C^l$  and  $p_{\text{bat}}^l$  represent the set of residential consumers and the battery power of coalition  $l$ , respectively.  $p_{\text{desire}}^j$  and  $p_{\text{alloc}}^j$  represent the desired amount of electricity and the amount of power allocated to residential consumer  $j$ , respectively.  $\lambda_{\text{cleared}}$  and  $p_{\text{cleared}}$  are cleared price and power of the coalition. Abovementioned parameters are at day  $k$  and hour  $t$ .  $v_1^j$  is residential consumers' parameters that converts dissatisfaction of not consuming desired amount of electricity into monetary value.  $v_2^l$  is a coalition parameter that converts degradation of battery into monetary value.

The coalition agent solves a local optimization problem at the start of each hour. (Thus, the day and hour are fixed for all terms in the optimization, and in the rest of this section, the indices  $k$  and  $t$  are dropped.) At the start of the hour, we assume that actual weather condition is known, and since the market is cleared, cleared amount of energy, cleared price, and battery power of coalition are also known. The coalition agent aims to optimally allocate available power between its residents to maximize their social welfare (sum of the reward). Note that the total bill of the coalition agent at each hour is equal to  $\sum \lambda_{\text{cleared}} p_{\text{alloc}}^j = \lambda_{\text{cleared}} (p_{\text{cleared}} - p_{\text{bat}})$ , which is a fixed value in the optimization. The total bill is proportionally distributed between residents, i.e., each resident pays  $\lambda_{\text{cleared}} p_{\text{alloc}}^j$ . Thus, the second term of (6) is considered as a constant in this step. Also, battery power is determined by a DRL module and is considered as a known constant term in the objective function [the third term of (6)]. In addition, the output power of the solar panels are assumed to be predicted beforehand by a weather forecasting service. The discrepancy between weather forecast and actual weather condition is assumed to be negligible in hour-ahead prediction. Therefore, the optimization problem for coalition  $l$  is as follows:

$$\min_{p_{\text{alloc}}^j} \sum_{j \in N_C^l} v_1^j (p_{\text{alloc}}^j + p_{\text{solar}}^j - p_{\text{desire}}^j)^2 \quad (7)$$

$$\text{s.t.} \sum_{j \in N_C^l} p_{\text{alloc}}^j + p_{\text{bat}} = p_{\text{cleared}}. \quad (8)$$

Note that the constraint (8) ensures that the allocated power to residents matches the power available from TEM considering the battery power.

*Theorem 1:* The closed-form solution of optimization (7) is

$$p_{\text{alloc}}^z = d_z + \frac{x^* - \sum_{j \in N_C^l} d_j}{v_1^k \sum_{j \in N_C^l} \frac{1}{v_1^j}} \quad \forall z \in N_C^l \quad (9)$$

where  $d_j = p_{\text{desire}}^j - p_{\text{solar}}^j$  and  $x^* = p_{\text{cleared}} - p_{\text{bat}}$ .

*Proof:* Using Lagrangian of (7) and applying first-order optimality condition, the theorem is proven. ■

Using the previous closed-form solution, the coalition agent is able to allocate the cleared power among the residents and collect the corresponding payments from them.

### III. PROPOSED REINFORCEMENT LEARNING METHOD

#### A. Environment, State, Action, and Reward

The environment in the DRL framework includes anything that can affect RL agent. The environment should be sufficiently captured in the state representation. In this article, the environment includes three main parts: other agents, market-clearing process, and weather condition. The RL is composed of three components: state space, action space, and reward function. The state vector is comprised of the next day information and historical data, given as follows:

$$\begin{aligned} s[k] &= [d[k], \vec{w}_f[k], \text{SoC}[k] \\ &\quad \vec{w}_f[k-1], \vec{w}_r[k-1], \vec{\lambda}_{\text{cleared}}[k-1], \dots, \\ &\quad \vec{w}_f[k-L], \vec{w}_r[k-L], \vec{\lambda}_{\text{cleared}}[k-L]] \end{aligned} \quad (10)$$

where  $d[j]$ ,  $\vec{w}_f[j]$ ,  $\vec{w}_r[j]$ ,  $\text{SoC}[j]$ , and  $\vec{\lambda}_{\text{cleared}}[j]$  denote the day of the week, the weather forecast vector, actual weather condition vector, SoC, and cleared price vector of TEM, all at step  $j$ , respectively.  $d$ ,  $\vec{w}_f$ ,  $\vec{w}_r$ , and  $\vec{\lambda}_{\text{cleared}}$  are repeated for  $L$  historical steps. Note that  $\vec{\lambda}_{\text{cleared}}$  is determined after the market is run and cleared, so its value at step  $k$  is not known when the agent is submitting its bid. In our problem, the action of the agent is choosing its linear bid and desired battery power for the next day. The linear bid is determined by a slope and y-intersect. To make it more precise, the action of the agent is as follows:

$$A = [\vec{a}, \vec{b}, \vec{p}_{\text{bat}}] \quad (11)$$

where  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{p}_{\text{bat}}$  denote the slope vector, y-intersect vector, and desired battery power vector of the learner agent that consist of 24 values for hours of the next day.

The reward function for each single agent is the negation of  $J_C^j$  in (3), and for multiple residents, agent is the negation of  $J_C^l$  in (6). For an agent in TEM, the state could include any information that affects the agent's decision. Possible affecting information for bidding could include time, agent's desired demand, weather forecast for next day (that has error and, thus, is a source of uncertainty), battery charge, cleared price, consumed power, and weather condition of previous days.

Note that in reward functions [see (3) and (6)], the term  $p_{\text{solar}}^j[k, t]$  represents the actual solar power generated by solar

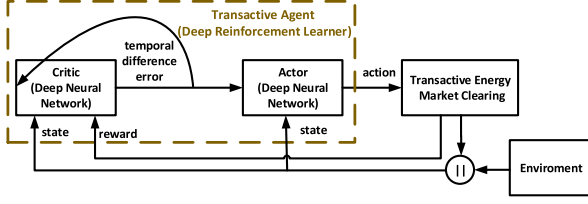


Fig. 2. Overall system overview, consisting a DRL agent as a transactive agent, TEM, environment, and their connections.

panels at the real time market. However, the next day weather condition in state space is the prediction since the market is cleared in day-ahead fashion and the actual weather condition is not known. The prediction and actual weather conditions for the previous days are placed in the state space as historical data. Although the weather prediction is a source of uncertainty and can cause unexpected costs to the agent, it learns to maximize its expected reward so that average reward over time with different situations of weather prediction error is maximized.

### B. Deep Reinforcement Learning

The overall system overview is depicted in Fig. 2. The learner agent submits a bid to TEM. After the bids are gathered and DSO's supply curve is acquired, the TEM operator clears the TEM market, determining the cleared price and power of agents, and consequently, the reward is calculated. The DRL concatenates TEM clearing information and environmental signals and forms the state vector.

The DRL deploys the SAC algorithm to learn its optimal action using received state and reward from the environment [8]. This process is done through two main submodules named critic and actor, which are implemented using deep neural networks. There are three main function estimators: state value function  $V(s_t|\theta_V)$ , soft Q-function  $Q(s_t, a_t|\theta_Q)$ , and policy  $\pi(a_t|s_t, \theta_\pi)$ .  $s_t \in S$ , where  $S$  is the state space,  $a_t \in A$ , where  $A$  is the action space, and  $\theta_V$ ,  $\theta_Q$ , and  $\theta_\pi$  are the deep neural networks weights. The weights of state value function's neural network are trained to minimize the squared residual error

$$J_V(\theta_V) = E_{s_t \sim D} \left[ \frac{1}{2} (V(s_t|\theta_V) - E_{a_t \sim \pi} [Q(s_t, a_t|\theta_Q) - \log \pi(a_t|s_t, \theta_\pi)])^2 \right] \quad (12)$$

where  $D$  is a replay buffer that provides the distribution of previously sampled states and actions. To learn the weights the gradient is needed which can be estimated using the following unbiased estimator:

$$\nabla_{\theta_V} J_V(\theta_V) = \nabla_{\theta_V} V(s_t|\theta_V) (V(s_t|\theta_V) - Q(s_t, a_t|\theta_Q) + \log \pi(a_t|s_t, \theta_\pi)). \quad (13)$$

To make the learning more stable, other networks use target value network  $\hat{V}(s_t|\hat{\theta}_V)$ , whose weights are exponentially moving average of the weights of the original value network. The weights of soft Q-function neural network are found to minimize soft

Bellman residual function

$$J_Q(\theta_Q) = E_{(s_t, a_t) \sim D} \left[ \frac{1}{2} (Q(s_t, a_t|\theta_Q) - \hat{Q}(s_t, a_t))^2 \right] \quad (14)$$

$$\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma E_{s_{t+1} \sim p} [\hat{V}(s_{t+1}|\hat{\theta}_V)] \quad (15)$$

where  $r(s_t, a_t)$  is the reward of the agent,  $\gamma$  is the discount factor, and  $p$  is the unknown state transition probability density that determines the probability of getting to a state in the next step given current state and action. Since this probability density function is not available, the learning of the parameters is done using stochastic gradients

$$\nabla_{\theta_Q} J_Q(\theta_Q) = \nabla_{\theta_Q} Q(s_t, a_t|\theta_Q) (Q(s_t, a_t|\theta_Q) - r(s_t, a_t) - \gamma \hat{V}(s_{t+1}|\hat{\theta}_V)). \quad (16)$$

Finally, the policy parameters, i.e.,  $\theta_\pi$  are chosen in such a way to minimize the expected Kullback–Leibler divergence using reparameterization trick technique which allows the use of backpropagation through stochastic node [22]

$$J_\pi(\theta_\pi) = E_{s_t \sim D, \epsilon_t \sim \mathcal{N}} [\log \pi(f(\epsilon_t, s_t|\theta_\pi)|s_t, \theta_\pi) - Q(s_t, f(\epsilon_t, s_t|\theta_\pi)|\theta_Q)] \quad (17)$$

where  $f(\epsilon_t, s_t|\theta_\pi)$  is the reparameterization of the policy using a neural network transformation that adds  $\epsilon_t$  as an input noise vector to introduce randomness and entropy maximization into the algorithm.  $\epsilon_t$  is sampled from a Gaussian distribution. The method extends the DDPG policy gradient to stochastic policy with use of the following unbiased gradient estimator:

$$\nabla_{\theta_\pi} J_\pi(\theta_\pi) = \nabla_{\theta_\pi} \log \pi(a_t|s_t, \theta_\pi) + (\nabla_{a_t} \log \pi(a_t|s_t, \theta_\pi) - \nabla_{a_t} Q(s_t, a_t|\theta_Q)) \nabla_{\theta_\pi} f(\epsilon_t, s_t|\theta_\pi) \quad (18)$$

where  $a_t$  is evaluated at  $f(\epsilon_t, s_t|\theta_\pi)$ . The SAC method also make use of two independent soft Q-functions to help mitigate the positive bias problem and performance degradation due to it in value-based methods [23].

## IV. SIMULATIONS

To evaluate the proposed method, we performed simulations on the standard 37-bus IEEE distribution system [24], consisting of ten TEMs. OPF is solved using MATPOWER [25], and RL method is implemented using MATLAB. The simulations were done on a laptop machine with Intel Core i7 CPU with 12 GB of RAM. Each TEM consists of 50 to 150 residential transactive agents. Fig. 3 depicts the distribution network. Demand profiles of agents follow a double-peaked figure, one around the morning and the other in the evening [26].

Nontransactive agents have a fixed demand at each hour and they draw power from the grid accordingly. Nonlearning transactive agents offer bids to TEM, which are linear functions of the available solar energy and their demand. The learner agent utilizes the algorithm, state, action, and rewards, as described in Section III. To investigate the efficiency of the proposed DRL method, two scenarios are compared. These scenarios correspond to types of agents as in Section II-C.

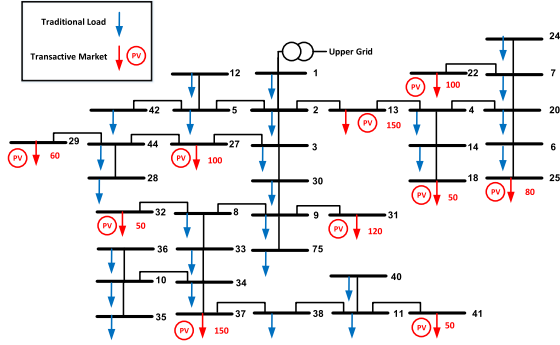


Fig. 3. 37-bus IEEE distribution network with ten TEMs. The number beside each TEM is the number of residential agents present in that TEM.

- 1) *Single resident agent*: The first scenario considers the case where agents play noncooperatively, and only one agent in TEM is considered to be using DRL for bidding, and its reward and performance are analyzed.
- 2) *Multiple residents agent*: The second scenario considers residents in the same building complex cooperatively solving the bidding problem to maximize their aggregated reward. In this case, the reward and performance of the coalition agent are analyzed. The allocation of cleared power follows Theorem 1.

Since the learning agent behaves stochastically, the plotted figures are averaged over five different runs.

#### A. Scenario I: Single Resident Agent

In this scenario, we assume that only one agent uses the DRL method.

In the rest of this section, we first describe the distribution network and analyze the learning process of the agent. Then, the effect of  $L$  (the number of previous steps to consider in the state vector) and battery size on the agent's reward are analyzed. Moreover, the battery power, which is one of the actions learned by the DRL method, is depicted alongside agents' demand and cleared power. Also, the average reward for SAC method is compared with DDPG method. The agent is assumed to be residing at bus 13 of Fig. 3. Other transactive agents offer a linear function of the price of energy. A typical bid is depicted as follows:

$$p_i^{tr}[k] = c_i \lambda[k] + h_i[k] \quad (19)$$

where  $c_i$  and  $h_i$  are the agent's parameters, where the latter is a function of the agent's available solar energy and demand.

As depicted in Fig. 3, there are ten transactive markets in the distribution network, each with different number of transactive agents residing in them. However, in this scenario, only one transactive agent in bus 13 utilizes the DRL method. Initially, the agent acts randomly to explore the environment. The agent explores the environment in the first stage, and learns the environment model. After learning the environment, the agent moves toward exploiting its knowledge by choosing an optimal strategy. In all figures, each step represents a day. Since the demand and weather condition of each hour is different, different hours could

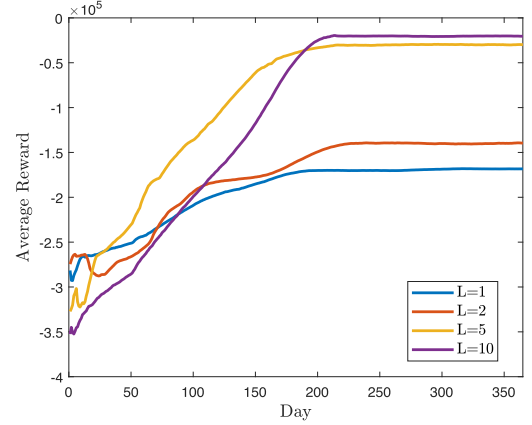


Fig. 4. Average reward of the agent for different values of  $L$ .

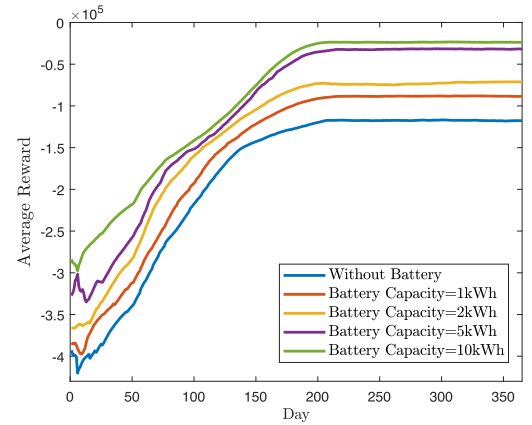


Fig. 5. Average reward of the agent for different battery capacities.

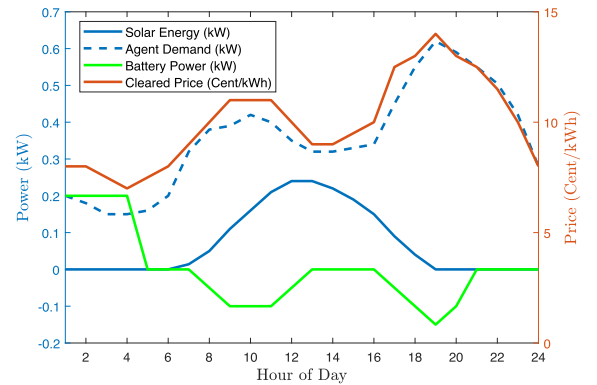


Fig. 6. Battery power, agent's demand, solar energy generation, and market clearing price for a day.

not be compared to each other fairly. To compensate this problem, the total reward of a day is considered and compared in figures. To reach better figures and better comparisons, the figures in this article are depicted starting after the exploration period, which according to Fig. 7(a) is assumed to be 30 days. During this period, the agent acts with high levels of randomness and shows great fluctuations in term of its reward. Fig. 4 compares the

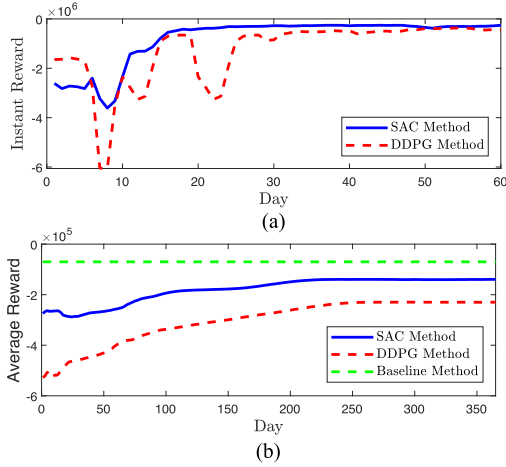


Fig. 7. Comparing the SAC method with deep deterministic policy gradient and baseline (with full information available), for  $L = 2$  and battery capacity of 5 kWh. (a) Instant reward of the agent. (b) Average reward of the agent.

average reward of the agent for different values of  $L$ , with battery capacity of 5 kWh. The higher value of  $L$  means more historical data available for the agent; hence, it might be suitable for better learning the environment. However, a higher dimension of state makes the learning process more computationally expensive. The simulations are run over 395 days, in which the first month is allocated for exploring the environment, and then, the rewards of the next 365 days are compared to simulate a full year. To observe the overall trend of the reward of the agent in different cases, the rewards of the last 50 days are averaged and plotted. As in Fig. 4, having more historical data increases the average reward. However, after a certain point ( $L = 5$ ), a higher  $L$  does not help the average reward substantially. This means that the information available in the extra historical data does not help the learning of the environment. Using more historical data means the environment is much more complex. Hence, as  $L$  increases, the learning becomes generally slower. Also, since the size of state space increases, each learning step becomes much more computationally intense.

Fig. 5 illustrates the effect of battery size on the average reward of the agent over 50 days, for  $L = 5$ . As seen in this figure, a higher battery size increases the average reward greatly. However, a much larger battery will not help the average reward, since the maximum battery requirement is saving energy when the price is low and using it when the price is high, if the difference is higher than the battery degradation cost. In Fig. 6, the learned battery power (for the case of battery capacity of 1 kWh) is shown alongside agent demand, solar power generation, and cleared price of the market for a full day at day 300. The agent uses its battery to hoard electricity between  $h = 1$  and  $h = 4$  when the price is lowest. It later uses its energy to meet part of its demand when price is peaking.

To investigate the effectiveness of the SAC method, the simulations were also conducted for DDPG. Also, full-information case is depicted as a baseline where the agent has full knowledge of clearing process and its parameters, weather condition, and other agents bids. In this case, due to having full information, the agent can use gradient descent method to reach its optimal

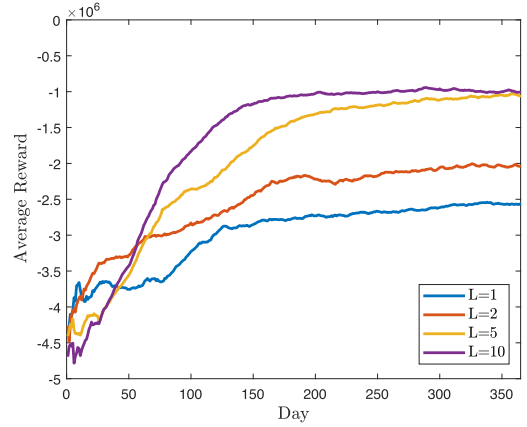


Fig. 8. Average reward of the coalition agent for different values of  $L$ .

solution. The comparison of two DRL methods' instant and average rewards are shown in Fig. 7 alongside the baseline. As seen in this figure, the SAC method learns faster and achieves the higher rewards ( $-1.4e5$  for SAC and  $-2.2e5$  for DDPG). As expected, baseline reward ( $-0.7e5$ ) is higher than DRL methods.

### B. Scenario II: Multiple Residents Agent

In this scenario, we assume that ten of the agents in TEM at bus 13 in Fig. 3 reside in a building and form a coalition to minimize their aggregated cost function. Instead of individual bids, the coalition agent must learn the aggregation bid of all its residential agents. The clearing process is the same as the previous scenario. Only instead of individual bids for each agent in the coalition, the aggregated bid of them is learned and submitted to the market. The state vector is considered as in (10), where  $\text{SoC}[k]$  is the state of the charge for the single centralized battery inside the building at step  $k$ .

Fig. 8 compares the average reward of the agent for different amounts of historical data, with the battery capacity of 10 kWh. Although more historical data help better learning of the environment, they make learning process longer and more computationally complex. As the last scenario, the plots are depicted for a year after the exploration period. This figure shows having more historical data increases the average reward. After a certain point ( $L = 5$ ), a higher  $L$  does not help the average reward substantially, since the available information is sufficient to learn the environment. Moreover, more historical data mean longer state vector that leads to a much more complex environment; thus, it requires more exploration to be learned and each step of the learning becomes more computationally complex.

Fig. 9 illustrates the effect of total battery capacity present in the residential complex of the coalition agent on the average aggregated reward for  $L = 10$ . As expected, adding battery capacitance improves the average reward; however, much more capacity does not improve the reward.

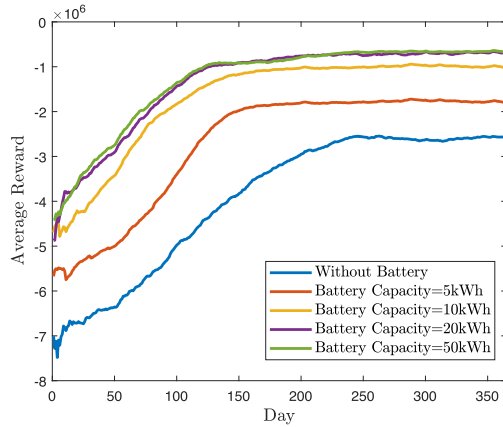


Fig. 9. Average reward of the coalition agent for different battery capacities.

TABLE I  
SUM OF REWARD OVER A DAY IN DIFFERENT SCENARIOS

Type of Agent	Reward Over a Day
Learner Transactive Agent in Scenario 1	-31905
Non-learner Transactive Agent in Scenario 1	-69312
Average coalition agent in Scenario 2	-57120

### C. Summary of Simulations

In this section, we have shown that the increasing  $L$  and battery capacity increases the agent's reward. This is expected. However, the simulations show that increasing these parameters indefinitely does not increase the reward; however, it increases the complexity of the model and computational intensity of each step and decreases the learning speed. For the battery capacity, it does not have a negative effect on the learning process since state and actions remain in the same dimensions.

Moreover, comparing two scenarios show that total battery capacitance needed to reach optimal performance is less in the coalition case. As the number of agents in the coalition agent is assumed to be ten, if all agents had the same battery, the total battery capacity of the coalition agent would be ten times the battery capacity of the previous case. However, as Fig. 9 shows, the best reward is not achieved at ten times battery capacitance of the best reward of Scenario 1. As seen in Fig. 5, having the battery larger than 5 kWh does not improve the reward. In Fig. 9, the best reward is reached at 20 kWh, which is much less than 50 kWh. This shows an important benefit of the cooperation of agents in a building: The battery needed is much less than when they work individually. This is due to the fact that the batteries now could be shared between agents. To compare the learner and nonlearner transactive agents in Scenario 1 and the average reward of the transactive agent of coalition agent learning in Scenario 2, Table I is drawn. For both the scenarios,  $L = 5$ . For Scenario 1, the battery capacity is 5 kWh, and for Scenario 2, the battery capacity is 20 kWh. In Table I, the summation of reward over a day is compared. In the first scenario, the learner agent outperforms the nonlearner agents, since it submits its bid smartly. When all agents in a building learn collectively as a coalition agent, the average reward improves. However, since the profits are distributed over all agents in the building,

agents' rewards become closer, and the single learner agent's reward is reduced compared with the previous scenario. In the first scenario, since other agents do not use learning methods, the learner agent can utilize its learning fully and improve its reward greatly.

## V. CONCLUSION

The proposed TEM framework in this article allows the grid to utilize small distributed energy sources, such as solar panels installed on each residential user, by allowing the transactive agents to engage in a bidirectional trade with the grid. At each day, local TEM clears the market between the transactive agents residing in a bus using DSO's supply curve, which is based on demands of buses and physical constraints of the network. Agents are considered to have solar panels and battery units installed. Agents submit demand bids to TEM and TEM aggregates the agents' demand and clears the market by DSO's supply curve. To analyze the effect of the DRL method, two scenarios were considered. In the first scenario, only one agent applied this method to optimize its bidding profile. In the second scenario, it was assumed that ten residential members of a building form a coalition in TEM and bid as one agent. After that, the cleared power was optimally allocated to the coalition members. Comparing two scenarios shows that cooperative bidding of residents in a building reduces the need for battery capacity to reach the maximum reward. Also, when only one agent used the learning method, its reward was greater than other transactive agents who submitted a fixed bidding profile. However, when residents of a building cooperated, the average reward of the cooperating agents were increased. We have shown the battery power of the learner agent and how it tries to lower costs by buying cheap and use it when the price is high. Also, we have shown that the SAC method outperforms the DDPG method in learning speed and performance.

## REFERENCES

- [1] B. M. Eid, N. Abd Rahim, J. Selvaraj, and A. H. El Khateb, "Control methods and objectives for electronically coupled distributed energy resources in microgrids: A review," *IEEE Syst. J.*, vol. 10, no. 2, pp. 446–458, Jun. 2016.
- [2] J. Confrey, A. H. Etemadi, S. M. Stuban, and T. J. Eveleigh, "Energy storage systems architecture optimization for grid resilience with high penetration of distributed photovoltaic generation," *IEEE Syst. J.*, vol. 14, no. 1, pp. 1135–1146, Mar. 2020.
- [3] G. Du, Y. Zou, X. Zhang, T. Liu, J. Wu, and D. He, "Deep reinforcement learning based energy management for a hybrid electric vehicle," *Energy*, vol. 201, 2020, Art. no. 117591.
- [4] S. L. Arun and M. P. Selvan, "Intelligent residential energy management system for dynamic demand response in smart buildings," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1329–1340, Jun. 2018.
- [5] P. Fairley, "The unruly power grid," *IEEE Spectr.*, vol. 41, no. 8, pp. 22–27, Aug. 2004.
- [6] V. S. K. Murthy Balijepalli, V. Pradhan, S. A. Khaparde, and R. M. Shereef, "Review of demand response under smart grid paradigm," in *Proc. ISGT2011-India*, 2011, pp. 236–243.
- [7] T. Council, "Gridwise transactive energy framework version 1.0," The GridWise Architecture Council, Richland, WA, USA, Tech. Rep. PNNL-22946, 2015.
- [8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

- [9] H. Kebriaei and V. J. Majd, "A simultaneous multi-attribute soft-bargaining design for bilateral contracts," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4417–4422, 2009.
- [10] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin, "Peer-to-peer and community-based markets: A comprehensive review," *Renewable Sustain. Energy Rev.*, vol. 104, pp. 367–378, 2019.
- [11] B. Cornélusse, I. Savelli, S. Paoletti, A. Giannitrapani, and A. Vicino, "A community microgrid architecture with an internal local market," *Appl. Energy*, vol. 242, pp. 547–560, 2019.
- [12] K. M. Muttaqi *et al.*, "Transactive energy-based planning framework for VPPs in a co-optimised day-ahead and real-time energy market with ancillary services," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 11, pp. 2024–2035, 2019.
- [13] J. Li, C. Zhang, Z. Xu, J. Wang, J. Zhao, and Y.-J. A. Zhang, "Distributed transactive energy trading framework in distribution networks," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7215–7227, Nov. 2018.
- [14] Y. K. Renani, M. Ehsan, and M. Shahidehpour, "Optimal transactive market operations with distribution system operators," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6692–6701, Nov. 2018.
- [15] Z. Liao and L. Sugianto, "Using Q-learning to model bidding behavior in electricity market simulation," in *Proc. IEEE Symp. Comput. Intell. Multicriteria Decis.- Mak.*, 2011, pp. 1–7.
- [16] N. Rashedi, M. A. Tajeddini, and H. Kebriaei, "Markov game approach for multi-agent competitive bidding strategies in electricity market," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 15, pp. 3756–3763, 2016.
- [17] B. M. Radhakrishnan *et al.*, "A reinforcement learning algorithm for agent-based computational economics (ACE) model of electricity markets," in *Proc. IEEE Congr. Evol. Comput.*, 2015, pp. 297–303.
- [18] Y. Ye, D. Qiu, J. Li, and G. Strbac, "Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 130515–130529, 2019.
- [19] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, "Reinforcement learning-based microgrid energy trading with a reduced power plant schedule," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10728–10737, Dec. 2019.
- [20] Y. Du, F. Li, H. Zandi, and Y. Xue, "Approximating nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning," *J. Modern Power Syst. Clean Energy*, vol. 9, no. 3, pp. 534–544, May 2021.
- [21] M. A. Tajeddini, H. Kebriaei, and L. Glielmo, "Decentralized hierarchical planning of PEVs based on mean-field reverse Stackelberg game," *IEEE Trans. Automat. Sci. Eng.*, vol. 17, no. 4, pp. 2014–2024, Oct. 2020.
- [22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.
- [23] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
- [24] W. H. Kersting, "Radial distribution test feeders," *IEEE Trans. Power Syst.*, vol. 6, no. 3, pp. 975–985, Aug. 1991.
- [25] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.
- [26] J. Love *et al.*, "The addition of heat pump electricity load profiles to GB electricity demand: Evidence from a heat pump field trial," *Appl. Energy*, vol. 204, pp. 332–342, 2017.