Received: 15 April 2024 | Revised: 20 July 2024 | Accepted: 1 August 2024

IET Generation, Transmission & Distribution

DOI: 10.1049/gtd2.13244

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Multi-agent reinforcement learning in a new transactive energy mechanism

Hossein Mohsenzadeh-Yazdi[1] | **Hamed Kebriaei**[1,2] | **Farrokh Aminifar**[1]

[1]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

[2]School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Correspondence
Hamed Kebriaei, School of Electrical and Computer Engineering, College of Engineering, University of Tehran; School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.
Email: kebriaei@ut.ac.ir

Funding information
Institute for Research in Fundamental Sciences (IPM), Grant/Award Number: CS 1402-4-208

Abstract

Thanks to reinforcement learning (RL), decision-making is more convenient and more economical in different situations with high uncertainty. In line with the same fact, it is proposed that prosumers can apply RL to earn more profit in the transactive energy market (TEM). In this article, an environment that represents a novel framework of TEM is designed, where all participants send their bids to this framework and receive their profit from it. Also, new state-action spaces are designed for sellers and buyers so that they can apply the Soft Actor-Critic (SAC) algorithm to converge to the best policy. A brief of this algorithm, which is for continuous state-action space, is described. First, this algorithm is implemented for a single agent (a seller and a buyer). Then we consider all players including sellers and buyers who can apply this algorithm as Multi-Agent. In this situation, there is a comprehensive game between participants that is investigated, and it is analyzed whether the players converge to the Nash equilibrium (NE) in this game. Finally, numerical results for the IEEE 33-bus distribution power system illustrate the effectiveness of the new framework for TEM, increasing sellers' and buyers' profits by applying SAC with the new state-action spaces. SAC is implemented as a Multi-Agent, demonstrating that players converge to a singular or one of the multiple NEs in this game. The results demonstrate that buyers converge to their optimal policies within 80 days, while sellers achieve optimality after 150 days in the games created between all participants.

1 | INTRODUCTION

Distribution systems are evolving into networks with proliferated penetration of renewable energy resources (RERs) to leverage their environmental and economic advantages [1], and it is expected that this increase will be over 60% of total generation by 2050 [2]. As another important development, consumers are using energy storage systems (ESS) to earn more profit by participating in the local market [3]. Not only do consumers get more rewards from using this equipment, but they also help the power grid to decrease demand at the peak of power consumption and improve environmental issues such as pollution [4]. Reducing the cost of buying the required power for buyers or increasing income through the sale of excess production power is the most critical factor that has led to the use of RERs. Important benefits, such as diminishing air pollution and raising the flexibility and reliability of the power systems, have made the operator willing to encourage consumers to use these resources. To increase the number of renewable energy sources and their participation in electricity markets, [5] examines various methods. The study discusses different policies such as investment support, production support, quantity targets, and limiting carbon emissions.

Despite the above-mentioned advantages of RERs, there are challenges that can limit their use. One of these challenges is the imbalance between production and consumption power in real-time for the owners of these resources. The transactive energy market (TEM) solves this challenge so that participants can compensate for their power imbalance by buying and selling power in this market. Small producers and consumers achieve more profit or less cost compared to participating in the traditional electricity market if they exchange power with each other in this market directly. In the last decades, transactive energy trading has been taken into contemplation in a considerable

number of countries. The European continent, especially Germany, UK, Spain, and the Netherlands are the leading countries in transactive energy trading from both the R & D projects and real-world trials points of view [6]. For example, Vandebron is a platform that is used in the Netherlands, and in this type of application, the consumer directly buys and supplies the electricity demand from independent generators. For example, farmers who buy and supply the electricity they need from the wind turbines in the fields. In fact, this model also acts as a supplier and balances the market by linking consumers and producers.

There are two main types of frameworks for TEMs that have been proposed in the literature: peer-to-peer (P2P) TEM and platform-based TEM. In the P2P TEM framework, participants exchange power directly with each other in a local market, while in the platform-based TEM framework, a centralized platform is used to facilitate transactions between participants. Several studies have proposed different approaches to designing TEMs. In [7], a two-level network-constrained P2P TEM in which the proposed scheme helps guarantee network security and incentivize microgrids (MGs) to trade energy. Reference [8] proposes a novel P2P transactive trading scheme based on the multi-actor-attention-critic algorithm. In [9], an integrated bilateral framework is introduced so that the distribution system operator (DSO) handles the interaction between consumers and producers. According to [10], the application of the sharing energy in local energy markets can be platform-based local energy sharing, and many studies have focused on this kind of energy sharing model. Reference [11] presents a new decentralized framework that buyers submit their offers in several steps to sellers and then the sellers announce the result of the market. Reference [12] uses a double-auction market to evaluate and compare two different TEMs. Reference [13] proposes a novel transactive energy control mechanism so that microgrids update their energy price according to their locations. In [14], a transactive energy management system for residential buildings is introduced with DERs as two phases that include day-ahead and real-time market. In [15], a TEM is proposed where the distributed generators act as independent suppliers. Reference [16] proposes a day-ahead local electricity market (DALEM) for transactive energy trading to facilitate the participation of various small and medium-sized DERs in the energy management program of the DSO at the distribution level.

The proposed framework aligns with the essential features outlined in [16] for an effective TEM: a competitive environment, technical constraints of the distribution system, and the integration of decentralized DERs. Many existing frameworks lack at least one of these features. For instance, the framework in [17], which employs a distributed optimization approach to clear a grid-connected DALEM, lacks the integration of decentralized DERs. Similarly, the market-based framework in [18] that facilitates energy trading between DSOs and multiple microgrids, and the model in [19] designed to promote DER deployment at the residential level, lack both the competitive environment and the integration of decentralized DERs. Our TEM ensures a competitive market where participants actively adjust their bids based on conditions, considers technical constraints of the distribution system to maintain grid stability and reliability, and fully integrates decentralized DERs for a more resilient and sustainable energy market.

Several studies have proposed TEM frameworks that resemble traditional electricity markets, where operators facilitate a bidding process to establish the market's clearing price. However, this approach is less effective for local electricity markets due to the negligible generation costs of DERs compared to traditional Generation Companies (GENCOs) [20]. For instance, a recent study in [21] applies a framework to local markets akin to those used in traditional electricity markets. Our proposed framework adopts a novel methodology for determining the clearing price, distinct from traditional electricity market frameworks. In the proposed framework, when demand exceeds supply, the clearing price converges towards the weighted average price of sellers, allowing buyers who offer higher prices to purchase more power. Conversely, when supply exceeds demand, the clearing price aligns more closely with the weighted average price of buyers, enabling sellers who offer lower prices to sell more power. The clearing price is bounded between the rate at which the grid purchases power from prosumers and the rate at which it sells power to consumers, accurately reflecting local market dynamics. This approach ensures a more accurate representation of local market conditions and enhances competition between all players in the TEM.

Despite the formation of TEMs, participating in these markets and sending bids by small producers and consumers is considered a challenge. In these markets, participants' bids have a significant impact on their income or costs, and finding the optimal bidding strategy in the formed games in this market is difficult. Furthermore, given the small scale of production and consumption, it is not cost-effective for participants to use expert human resources to submit bids to TEMs. Several studies address the challenges in developing effective bidding strategies. Reference [22] proposes an optimal bidding strategy for a Wind Energy Portfolio Manager (WEPM) that incorporates Electric Vehicle Parking Lots (EVPLs). Reference [23] introduces a bilevel and multistage framework for a Renewable Energy Portfolio Manager (REPM). This framework integrates RESs with an energy storage system (ESS) to reduce imbalances and participate in both the day-ahead market and the balancing market, including bilateral contracts.

Another method to address these kinds of decision-making challenges is using Reinforcement Learning (RL), which offers significant advancements over traditional mathematical optimization approaches [24, 25]. DRL's ability to continually adapt to the volatile nature of energy markets, characterized by frequent shifts in supply, demand, and pricing, renders it a more efficacious tool for real-time decision-making in TEM. DRL excellently manages the intricate web of interactions between multiple agents—encompassing buyers, sellers, and storage systems—amidst the complexities of pricing, demand fluctuations, and grid stability. DRL's capacity to bypass the limitations of predefined models and assumptions, instead drawing upon both historical and real-time market data, enables the derivation of optimal bidding and trading strategies that resonate more closely with actual market dynamics. Furthermore, DRL's robustness in the face of uncertainties such as variable

renewable energy outputs and consumer behavior variations significantly bolsters the dependability of market operations. Notably, as the TEM ecosystem expands with an increasing number of participants, DRL demonstrates remarkable scalability, maintaining robust performance despite escalating market complexity. Thus, DRL can be used to address the unique challenges in TEM to find the best policy.

Considering the importance of decision-making in situations with high uncertainty, such as decision-making in the electricity market, using RL and DRL methods have been considered in recent studies. In [21], it is proposed that owners of ESSs in distribution systems use the Q-learning algorithm, which is an RL algorithm for discrete space, to manage their batteries. Reference [26] proposes that participants who are big producers and consumers in the traditional power market employ an algorithm that is like Q-learning to increase their profit. Reference [27] proposes a bid selection model to tame the computational complexity of auction mechanisms in TESMs. The proposed design is carried out for single sided combinatorial auctions through a two-step procedure. In [28], DRL is employed for a single agent, acting as a prosumer, to maximize profit in TEM. Reference [28] highlights the Soft Actor-Critic (SAC) algorithm, a state-of-the-art approach, demonstrating its superior performance compared to other DRL algorithms. However, it acknowledges a potential limitation: the scenario assumes a single agent. This assumption overlooks the possibility of other prosumers also employing algorithms to optimize their strategies.

In recent years, several studies have explored multi-agent reinforcement learning in *traditional electricity markets*. Reference [29] introduces a Deep Policy Gradient (DPG) method with a novel Long-Short Term Memory (LSTM)-based representation network to optimize the offering strategies of multiple self-interested GENCOs. Similarly, [30] applies a Multi-Agent Deep Deterministic Policy Gradient (MAD-DPG) algorithm to address the Markov game between GENCOs. Meanwhile, reference [31] presents a new RL algorithm combined with Graph Convolutional Networks (GCN) for optimal bidding strategies for generation units. Reference [32] proposes a Multi-agent Simulation (MAS) theoretical framework for strategic bidding in electricity markets, using reinforcement learning for GENCOs. In [29]–[32], the focus is on GENCOs as the sole players selling power in traditional electricity markets. These studies have a relatively uncomplicated state space, as GENCOs submit daily bids for the next 24 h, and factors like energy storage are not considered. Reference [33] proposes a Multi-Agent Reinforcement Learning (MARL) method for numerous small-scale prosumers, aimed at autonomous control policies for P2P energy trading in the Continuous Double Auction (CDA) market. However, [33] employs a standard model of the traditional electricity market for calculating clearing prices and power allocation, which [20] suggests may not be reasonable due to the negligible generation costs of DERs.

This article diverges from these approaches by considering *a new platform-based TEM* as the environment for the problem. This framework allows all *sellers and buyers* to submit offers including the quality and price of power. Unlike most studies that consider GENCOs as the primary focus of multi-agent analyses in traditional electricity markets, this article's multi-agent problem considers *both sellers and buyers in TEM*. When buyers and sellers utilize DRL, they establish a game within TEM. In this article, we examine this game and investigate the probability of converging to the Nash equilibrium. We also incorporate ESSs for both sellers and buyers, requiring them to manage their ESSs throughout the day in the created game. The chosen algorithm for each player is SAC, as it is a state-of-the-art and best-performing algorithm in DRL for the single agent problem, according to [28].

To apply DRL algorithms to a problem, it is necessary to specify the agent and environment of the problem. In this problem, the agent is either a seller or a buyer who aims to maximize their profit by interacting with the TEM. The DRL algorithm helps the agent find the optimal bidding strategy, which involves determining the price and amount of power to trade in TEM within the shortest possible time. Besides the framework and the agent, state-action space is another essential parameter. We consider ESSs for prosumers and the dynamic state of these sources in the state-action state. Because of the designed continuous state-action, the agent should apply DRL algorithms to detect the optimal policy. In this article, SAC algorithm is used to find the optimal policy by the participants who have ESS. When participants apply this algorithm with the designed state-action, they can converge to the best policy in the shortest possible time.

Table 1 highlights the novelty of our paper. Our approach introduces a novel platform-based framework, utilizes the SAC algorithm, and implements extensive multi-agent SAC (MASAC). We provide a detailed convergence analysis to Nash Equilibrium (NE), integrate ESS for both buyers and sellers, and focus on small-scale prosumers and consumers, distinguishing our work from existing methodologies.

The summarized contributions of this article are:

I) A new TEM framework is designed so that the three main mentioned features of a standard framework are considered: (1) competitive environment, (2) technical constraints of the distribution system, and (3) integration of decentralized DERs.
II) Continuous state-action spaces are designed to use the SAC algorithm for buyers and sellers as a single agent with or without ESS. Charging of ESS at different times is the state space, and how its transformation is considered.
III) The problem is modeled as a competitive multi-agent reinforcement learning. The formed game in TEM is investigated, and the multi-agent SAC algorithm is implemented for the first time in *platform-based TEM*. In this situation, it is studied whether the sellers and buyers converge to the NE of this game.

The rest of the paper is organized as follows: Section 2 suggests the new TEM framework. Section 3 illustrates the games formed in the TEM and introduces a new state-action for participants in it to implement the SAC. Section 4 gives numerical results tested on the IEEE 33-bus distribution power system. Finally, conclusion is given in Section 5.

TABLE 1 Comparison of existing approaches and the novel contributions of this paper.

Feature/methodology	Existing approaches	Proposed method
Framework type	Peer-to-peer and platform-based frameworks: 1- Similar to traditional electricity markets frameworks 2- Does not aligns with the essential features outlined in [16]	Novel platform-based framework
Algorithm	Well-known RL algorithms and DDPG	SAC with a new state action space
Multi-agent implementation	Limited multi-agent focus	Extensive multi-agent SAC (MASAC)
Convergence analysis	Not provided	Convergence analysis to NE
Energy storage system (ESS) integration	Ignored	ESS considered for both buyers and sellers
Market participants	Mainly GENCOs	Both small-scale prosumers and consumers

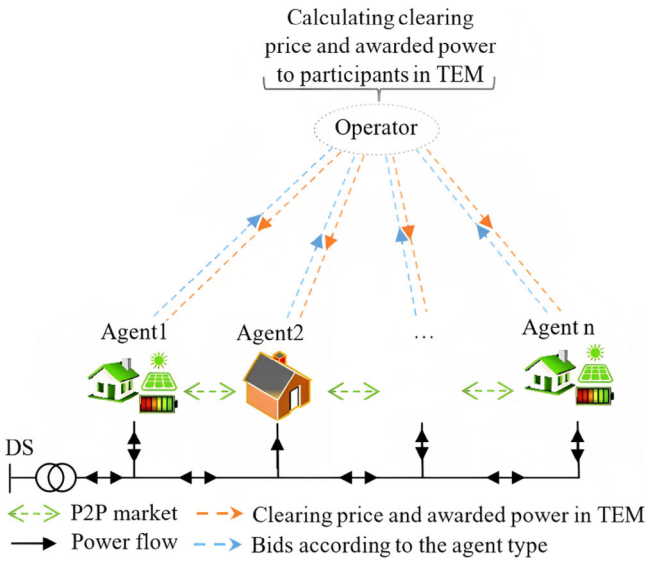

FIGURE 1 Bid and result exchange between agents and operator in TEM.

2 | NEW TEM FRAMEWORK

This section introduces a novel framework for TEM that enables participants to increase profits or reduce costs when selling or buying active power, as compared to the participating in the traditional electricity market. The framework calculates a clearing price that is between the selling and buying prices of power to/from the grid, leading to increased profits for both sellers and buyers. The clearing price adjusts according to the balance between demand and supply, reflecting the principles of a market economy. For example, when demand is high, the clearing price increases.

Renewable resources do not have a generation cost, making them less compatible with the traditional electricity market framework [20]. Therefore, the new framework is more appropriate for managing transactions related to prosumers' production and consumption within the TEM.

Figure 1 shows that small sellers and buyers submit their bids, which comprise price and amount of power, to the market operator. The operator then calculates the clearing price and assigns active power to participants based on their bids. This environment maintains competition between participants to trade more active power in this market, although all participants can trade a minimum active power based on differences between submitted prices. Thus, this framework has less risk for prosumers to participate in it. The proposed TEM framework is structured into two distinct phases, which are detailed below.

2.1 | Energy allocation to participants

The newly proposed framework requires the involvement of an operator who receives bids from both producers and consumers and oversees the security of the grid. The proposed framework for TEM can be implemented for the next day or the next hour. First, all participants in TEM submit their bids (including active and reactive power and price for selling or buying energy) for the next hour to the TEM operator. Then the operator runs the Backward/Forward sweep load flow in order to check voltage constraints for all buses in the distribution grid. Based on the voltage magnitudes obtained from the load flow analysis, the operator may do one of the following two tasks:

1) If the voltage magnitude at any of the buses is found to be outside of the permissible range, the operator has to apply inductors and capacitors, according to the voltage condition, to keep the voltage magnitude in its permitted range in the distribution system. In case of voltage deviation remains after applying inductors or capacitors, which *rarely* happens, the operator is forced to curb the participants' bids. The operator changes sellers' and buyers' offers by linear optimization with the objective function represented in Equation (1). The equation indicates that sellers who offer lower prices to sell and buyers who offer higher prices to buy can trade more active power in the TEM. Therefore, participants aim to offer competitive prices in order to trade more active power. The mentioned objective function is subjected to several constraints, which are expressed by Equations (2)–(7). The second and third equations represent the balance of active and reactive power at each bus, respectively. Equations (4) and (5) are load flow equations in distribution systems [34]. The line and bus

voltage limitations are enforced by Equations (6) and (7) in this load flow.

$$\max_{ps_i^f[t],pl_i^f[t]} \sum_{i\in n} \left(\pi_g^s[t] - \pi_i^s[t]\right) \times ps_i^f[t] + \sum_{i\in n} \pi_i^b[t] \times pl_i^f[t] \quad \forall i,t, \tag{1}$$

$$ps_i^f[t] - pl_i^f[t] + \sum_{l,to=i} p_l^f[t] - \sum_{l,from=i} p_l^f[t] = 0 \quad \forall i,t,l, \tag{2}$$

$$qg_i^f[t] - ql_i^f[t] + \sum_{l,to=i} q_l^f[t] - \sum_{l,from=i} q_l^f[t] = 0 \quad \forall i,t,l, \tag{3}$$

$$v_k[t] - v_i[t] + \frac{r_l p_l^f[t] + x_l q_l^f[t]}{v_i[t]} \geq (\alpha_l[t] - 1)M, \quad \forall i,t,l, \tag{4}$$

$$v_k[t] - v_i[t] + \frac{r_l p_l^f[t] + x_l q_l^f[t]}{v_i[t]} \leq (1 - \alpha_l[t])M, \quad \forall i,t,l, \tag{5}$$

$$-\alpha_l[t]\, cap_l \leq p_l^f[t] \leq \alpha_l[t]\, cap_l, \quad \forall t,l, \tag{6}$$

$$V^{\min} \leq v_i[t] \leq V^{\max}, \quad \forall t,i. \tag{7}$$

In Equation (1), $\pi_g^s[t]$, $\pi_i^s[t]$, and $\pi_i^b[t]$ denote grid price, producer i offering price, and consumer i bid price at hour t. $ps_i^f[t]$ and $pl_i^f[t]$ represent the awarded active power to producer i and consumer i at hour t. In (2) and (3), $p_l^f[t]$ and $q_l^f[t]$ are variables that denote the active and reactive power passing through the line l at hour t. The reactive generation and consumption at bus i at hour t are shown as $qg_i^f[t]$ and $ql_i^f[t]$. In (4) and (5), r_l and x_l are constant and denote the impedance and inductance of the transmission line l. v_k and α_l are integer and binary variables that indicate voltage at bus k and whether the line l is connected at hour t. Another constant parameter is M, a big positive number, which relaxes these constraints if line l is open. In (6), cap_l is maximum transmission line capacity l.

In the TEM, participants typically consist of consumers and producers engaged in the exchange of relatively low-scale power. Consequently, instances where reducing the generation or demand of prosumers significantly affects voltage deviation are relatively *uncommon*. It is crucial to note that, in practice, the need for the TEM operator to curb the bids of participants, either in terms of demand or generation, is very limited occurrence. However, despite the *infrequency* of such situations, the importance of addressing all aspects for a comprehensive TEM framework should be considered.

2) If the voltage magnitudes at all buses are within the range defined by the DSO that is a *common* situation, the operator does not need to take any action. In this case, all bids submitted by the participants are accepted without any modifications.

After confirming or changing the participants' bids according to the voltage constraints, Equation (8) is used to determine the amount of active power each producer can trade if the total active power submitted by producers is greater than that submitted by consumers. Conversely, Equation (9) is used to calculate the amount of active power each consumer can purchase if the total active power submitted by consumers is greater than that submitted by producers. These two equations show each producer can sell more active power in TEM and earn more profit if they submit a lower price to the operator in competition with other sellers. On the other hand, consumers can buy more active power in TEM and pay lower cost if they submit a higher price to the operator.

$$ps_a^{f,al}[t] = \max\left(0, \left(\frac{\sum_{i\in s} \pi_i^s[t] - (n_s - 1)\pi_a^s[t]}{\sum_{i\in s} \pi_i^s[t]}\right)\right) \times \sum_{i\in b} pl_i^f[t], \tag{8}$$

$$pl_a^{f,al}[t] = \left(\frac{\pi_a^b[t]}{\sum_{i\in b} \pi_i^b[t]}\right) \times \sum_{i\in s} ps_i^f[t]. \tag{9}$$

In (8), $ps_a^{f,al}[t]$ is the active power that the seller a can sell in TEM at hour t. Thus, the seller has to sell $ps_a^f - ps_a^{f,al}$ to grid with a lower price. In (9), $pl_a^{f,al}[t]$ is the active power that the consumer a can buy in TEM at hour t, and has to buy $pl_a^f - pl_a^{f,al}$ from grid with a higher price. As we explained, not only are the participants' bids effective in the optimization and clearing price process, but they also affect the amount of active power sold or bought in TEM.

According to (8) and (9), it is possible that allocated energy to a seller or a buyer exceeds their request. In such cases, the surplus power is allocated to the seller who submitted the lowest bid or the buyer who offered the highest price.

2.2 | Clearing price in TEM

After calculating the energy allocation process, the operator must calculate some parameters to find the clearing price. First, the weighted average prices for sellers and buyers are calculated using Equations (10) and (11). These equations illustrate that the clearing price is influenced by all participants in proportion to the amount of active power awarded to them after the optimization.

$$wm^s[t] = \frac{\sum_{i\in S} \pi_i^s[t] \times ps_i^f[t]}{\sum_{i\in S} ps_i^f[t]}, \tag{10}$$

$$wm^b[t] = \frac{\sum_{i\in B} \pi_i^b[t] \times pl_i^f[t]}{\sum_{i\in B} pl_i^f[t]}. \tag{11}$$

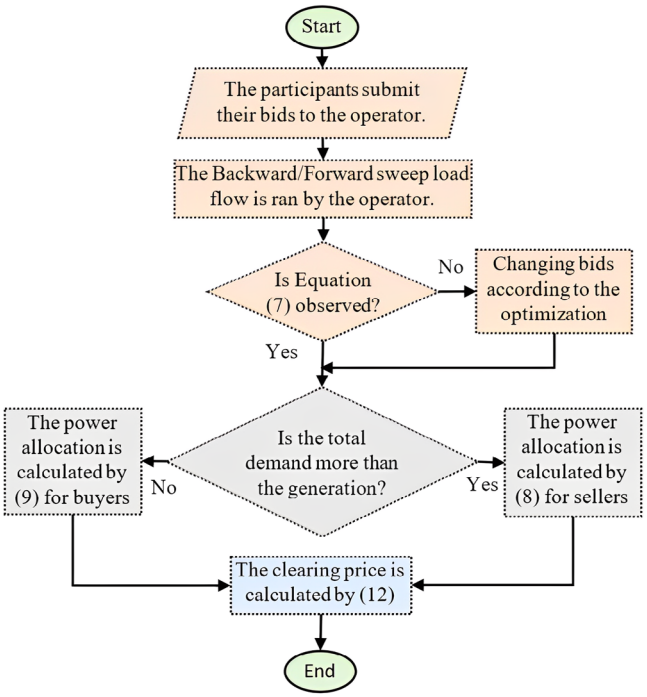

FIGURE 2 A brief of proposed framework.

In (10) and (11), $wm^s[t]$ and $wm^b[t]$ are the weighted average price of sellers and buyers at hour t.

In the next step, the clearing price in TEM is calculated by the operator according to Equation (12). This equation defines the basic market principle that states an increase in active power for selling in the market leads to a decrease in active power price and vice versa.

$$\pi_c[t] = \frac{(wm^s[t] \times \sum_{i \in B} pl_i^f[t]) + (wm^b[t] \times \sum_{i \in S} ps_i^f[t])}{\sum_{i \in B} pl_i^f[t] + \sum_{i \in S} ps_i^f[t]}. \tag{12}$$

In (12), $\pi_c[t]$ is the clearing price in TEM at hour t.

Figure 2 is a flowchart that provides an overview of the new TEM framework incorporating the three key principles discussed. The first principle is the submission of active power by producers and consumers to the operator. The second principle is the allocation of active power to producers and consumers based on their bid prices. The final principle involves the calculation of the clearing price by the operator, which is based on the total demand and the total amount of power submitted by sellers.

3 | GAME IN TEM AND THE ROLE OF DRL

In this section, we delve into the role of RL within TEM, providing a comprehensive explanation. Following this, we offer a detailed introduction to the SAC algorithm, outlining its functionalities and significance. Subsequently, we examine the dynamics of the game within TEMs and proceed to design and present a novel state-action space, tailored to this specific context.

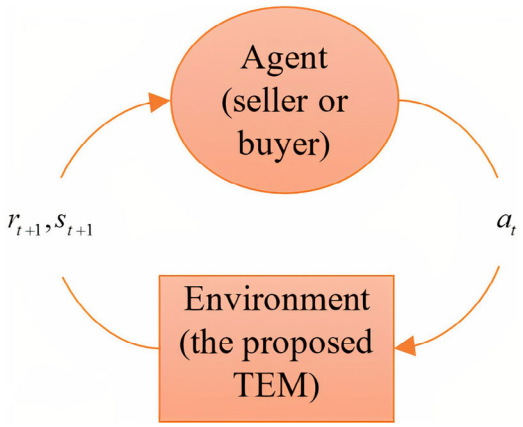

FIGURE 3 Basic concept of RL.

3.1 | Reinforcement learning in TEM

Participating in the TEM can be financially beneficial for both sellers and buyers as they have the opportunity to sell or purchase active power at better prices than trading with the grid. However, determining the optimal bid to submit can be challenging since it impacts the optimization, clearing price, and energy allocation processes. One potential solution is to use RL algorithms to find the best bidding strategy efficiently. By doing so, agents can increase their profits when participating in the TEM.

Agents who are participants in the TEM can use the RL algorithms to earn more payoff during learning with the highest convergence speed. The concept of RL is shown in Figure 3, where the agent selects an action, and the environment, which is TEM, returns a reward and a next state. After that, the agent updates the policy according to the action, reward, and next state.

Agents in the TEM select a continuous action including a price and amount of active power, and the states can be continuous if the agents own ESS. Thus, they should use the DRL algorithms to find the optimal bidding strategy.

SAC algorithm [35], which is a state-of-the-art DRL algorithm, is an off-policy algorithm renowned for its stability and efficiency in continuous action spaces. Its effectiveness is particularly notable in complex, multi-agent environments such as TEM, primarily due to its entropy-based exploration strategy that significantly enhances performance. In contrast to other RL algorithms like Nash Q-learning [36], designed for non-cooperative games with a focus on equilibrium strategies in discrete action spaces, which makes it less suitable for the continuous state-action context presented in this problem. SAC algorithm does not need any information from other players and the inherent design of this algorithm is adept at managing continuous state spaces, which is crucial in the TEM context,

where variables such as the varying capacities of batteries need accurate modeling and response to dynamically reflect energy storage and consumption patterns.

SAC algorithm includes two parts: actor and critic network implemented by deep neural networks. One of the most important factors for SAC is entropy regularization, which causes a close connection between exploration and exploitation during learning. By increasing entropy, the algorithm encourages exploration. The agent gets a reward in each state and wants to increase the reward by using an optimal policy described as (13).

$$\pi^* = \arg\max_{\pi} \underset{\tau\sim\pi}{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(R(s_t,a_t,s_{t+1}) + \alpha H(\pi(\cdot|s_t))\right)\right]. \tag{13}$$

In (13), α and γ are the trade-off coefficient and discount factor. The entropy H is computed from distribution D according to the (14).

$$H(D) = \underset{x\sim D}{E}[-\log D(x)]. \tag{14}$$

Three main functions named state value function, self Q-function, and policy are estimated by deep neural networks. The state value function's neural network is trained to minimize squared error with the estimation of parameter ψ:

$$J_V(\psi) = \mathrm{E}_{s_t\sim D}\left[\frac{1}{2}\left(V_\psi(s_t) - \mathrm{E}_{a_t\sim\pi_\phi}[Q_\theta(s_t,a_t) - \log\pi_\phi(a_t|s_t)]\right)^2\right]. \tag{15}$$

In (15), D is a replay buffer that collects states, actions, and next states. For updating ψ to reduce error, we can use gradient by (16):

$$\widehat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(s_t)\times(V_\psi(s_t) - Q_\theta(s_t,a_t) + \log\pi_\phi(a_t|s_t)). \tag{16}$$

Parameter θ for the Q-function's neural network is trained to reduce the error (17) by (18) similar to ψ.

$$J_Q(\theta) = \mathrm{E}_{(s_t,a_t)\sim D}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t) - \widehat{Q}(s_t,a_t)\right)^2\right], \quad \text{where} \quad \widehat{Q}(s_t,a_t) = r(s_t,a_t) + \gamma\mathrm{E}_{s_{t+1}}[V_{\bar{\psi}}(s_{t+1})] \tag{17}$$

$$\widehat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(s_t,a_t) \times (Q_\theta(s_t,a_t) - r(s_t,a_t) - \gamma V_{\bar{\psi}}(s_{t+1})). \tag{18}$$

Parameter ϕ in policy's neural network is updated for better estimation by (19) that is gradient Equation (20).

$$J_\pi(\phi) = \mathrm{E}_{s_t\sim D,\varepsilon_t\sim N}[\log\pi_\phi(f_\phi(\varepsilon_t;s_t)|s_t) - Q_\theta(s_t, f_\phi(\varepsilon_t;s_t))], \tag{19}$$

$$\widehat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi\log\pi_\phi(a_t|s_t) + (\nabla_{a_t}\log\pi_\phi(a_t|s_t) - \nabla_{a_t}Q(s_t,a_t)) \times \nabla_\phi f_\phi(\varepsilon_t;s_t). \tag{20}$$

In (19), $f_\phi(\varepsilon_t,s_t)$ is a result of the reparameterization trick with noise ε from normal distribution [35].

3.2 | Game in TEM

In TEM, the fundamental objectives of the participants are distinct: sellers consistently aim to maximize their rewards in the ensuing games, while buyers are focused on minimizing their costs. A critical aspect of TEM is the incorporation of power flow equations into the outcomes, an inclusion vital for accurately representing the dynamics and constraints inherent within TEM. Given these considerations, the model can be more aptly described as a Generalized Nash Equilibrium Problem (GNEP). GNEP is a mathematical model used to analyze strategic interactions among multiple decision-makers in a decentralized manner. In a classical Nash equilibrium, each player's strategy is optimal given the strategies chosen by the other players. In a Generalized Nash equilibrium, the players may have different decision variables, and their objectives and constraints may also be interdependent. The GNEP in TEM can be formulated as follows:

$$\begin{aligned}&\text{Seller } i\text{: } \min_{\mathcal{A}_i^s} \ u_i^s\left(\mathcal{A}_i^s, \mathcal{Z}\right),\\ &\text{Buyer } j\text{: } \min_{\mathcal{A}_j^b} \ u_j^b\left(\mathcal{A}_j^b, \mathcal{Z}\right),\\ &\text{subject to: (1)–(12),}\end{aligned} \tag{21}$$

where $\mathcal{Z}$ represents the vector of decision variables from all players except player i or j.

1) Players:

$$N^s = \{1,2,\ldots,n_s\}, \quad N^b = \{1,2,\ldots,n_b\}. \tag{22}$$

2) Actions for each player:

$$\mathcal{A}_i^s = \left\{\pi_i^s \in \left[\pi_g^b, \pi_g^s\right], ps_i\right\} \ i\in N^s, \quad \mathcal{A}_j^b = \left\{\pi_j^b \in \left[\pi_g^b, \pi_g^s\right], pl_j\right\} \ j\in N^b. \tag{23}$$

3) Payoffs:

$$u_i^s\left(\mathcal{A}_i^s,\mathcal{Z}\right) = \left(\pi_c[t]\times ps_i^{f,al}[t] + \pi_g^b[t]\times\left(ps_i^f[t] - ps_i^{f,al}[t]\right)\right) - ps_i[t]\times\pi_g^b[t] \quad \forall i\in N^s, \tag{24}$$

$$u_j^b\left(\mathcal{A}_j^b, z\right) = \left(\pi_c[t] \times pl_j^{f,al}[t] + \pi_g^s[t] \times \left(pl_j^f[t] - pl_j^{f,al}[t]\right)\right) - pl_j[t] \times \pi_g^s[t] \quad \forall j \in N^b. \tag{25}$$

In Equation (24), the first part delineates the rewards accrued by seller i through their participation in TEM. The latter part details the rewards that seller i would earn by selling all their power directly to the grid, rather than through TEM. Similarly, Equation (25) initially presents the costs incurred by buyer j when participating in TEM. Subsequently, it illustrates the costs that would be incurred by this buyer if they opted to purchase all their power directly from the grid, bypassing TEM.

The game formulated within TEM possesses two distinctive features: first, the strategy set for this game is finite and nonempty; second, the associated payoff function demonstrates continuity. Given these attributes, the game is classified as a continuous game. The implication of such a classification, as per Glicksberg's Theorem [37], is the guaranteed existence of at least one NE for continuous games. However, it is possible for the created game to have more than one NE.

Overall, when all players in the game employ the SAC algorithm, convergence towards an NE may occur. The primary advantage of SAC, distinguishing it from other DRL algorithms, is its entropy feature. This characteristic encourages a thorough exploration of the action space, substantially enhancing the probability of uncovering an optimal policy. Consequently, this increases the likelihood of achieving an NE, potentially optimizing the outcome of the game in TEM.

The theoretical guarantee for the convergence of multi-agent RL to an NE in a GNEP is indeed a challenging and open problem [38]. In this article, the payoffs for sellers and buyers are influenced by numerous factors and variables. The policies of all players affect the payoffs for each seller and buyer, and the results of power flow and distribution constraints also impact these payoffs. Therefore, mathematically proving convergence to NE in a GNEP remains an open problem.

To prove the existence of an NE in the formed game, the definition of NE can be used. In the context of TEM, an NE is a state where no participant can improve their payoff by unilaterally changing their strategy, assuming other participants' strategies remain unchanged. Achieving NE ensures that all market participants operate optimally given the actions of others, leading to a stable and efficient market.

Verification of convergence to Nash equilibrium in TEM:

1. Individual policy convergence: All players, including sellers and buyers, individually use the SAC algorithm to find the best policy in the game. It is assumed that they converge to their respective best policies.
2. Single player re-optimization: Once all players have converged to their best policies, we fix the policies of all players except one. The selected player then applies a new SAC algorithm to re-optimize and find the best policy again.
3. Policy convergence check: After applying the new SAC algorithm, if the re-optimized best policy for the selected player converges to the previous point, it suggests that this point could be an NE in the created game.
4. Repeated verification: To confirm that the converged point is an NE, we repeat steps 2 and 3 for all players. If each player converges to the same policy after using the new SAC algorithm while other players' strategies remain fixed, it proves that the converged policy is indeed an NE of the game.

By following this strategy, it can be effectively verified the convergence to NE within the created game, ensuring stable and efficient market operation.

3.3 | New state-action

In the TEM, agents can either be sellers or buyers, and they may or may not own an ESS. First, we assume the agent is a seller who desires to increase their payoff. If the seller owns an ESS, their action will be to determine the price and the amount of active power they intend to sell. On the other hand, if the seller does not own an ESS, they will only need to submit a price. Thus, the actions of sellers with and without ESS are represented by Equation (26).

$$\mathcal{A}^s[t] = \begin{cases} [\pi^s[t], ps[t]] & \text{with ESS} \\ [\pi^s[t]] & \text{without ESS} \end{cases}. \tag{26}$$

On the other hand, the agent as a buyer with or without ESS submits their action as (27). Buyers with ESS can select a price and amount of active power as an action to trade in TEM, while buyers without ESS should choose a price and have to send their demand.

$$\mathcal{A}^b[t] = \begin{cases} [\pi^b[t], pl[t]] & \text{with ESS} \\ [\pi^b[t]] & \text{without ESS} \end{cases}. \tag{27}$$

For agents (both sellers and buyers) who own an ESS, the state space in the TEM is defined by the state of charge (SoC) of their ESS at each hour. However, for those who do not own an ESS, the state space only includes time. Thus, the state is SoC of SEE at each hour for agents with SEE presented as (28) with battery charge limitation as (29) and (30). Equation (29) shows the relation between state, action, and next state, and Equation (30) demonstrates minimum and maximum battery capacity.

$$s = [SoC[0], SoC[1], \ldots, Soc[23]], \tag{28}$$

$$\begin{cases} SoC[t+1] = SoC[t] + pg[t] - ps[t] & \text{for sellers} \\ SoC[t+1] = SoC[t] + pl[t] - pd[t] & \text{for buyers} \end{cases}, \tag{29}$$

$$0 \le SoC[t] \le SoC^{\max}. \tag{30}$$

In (29), $pg[t]$ and $pd[t]$ are the generation of seller and the demand of buyer at hour T.

4 | SIMULATIONS

The simulation is done on the 33-bus IEEE distribution system with assumption of TEM with eight participants on different buses. The Backward/Forward sweep load flow is run by PANDAPOWER in ANACONDA (python). In case the voltage magnitudes are found to be outside the permitted ranges, the optimization problem was solved using PYOMO. The SAC algorithm is implemented by the TORCH library in ANACONDA with a laptop machine own Intel Core i7 CPU with 12 GB of RAM.

In our SAC implementations, the actor network, comprising two fully connected layers with 256 neurons each, estimates the mean and standard deviation for the action distribution. Similarly, the critic network, structured with two 256-neuron layers, estimates the Q-value based on state-action pairs. The value network, also consisting of two fully connected layers with 256 neurons each, estimates the state value. All networks utilize ReLU activation functions, while the actor network's output is normalized using the tanh function. Key hyperparameters include learning rates set at 0.003, a discount factor (γ) of 0.99, a batch size of 64, a replay buffer size of 1,000,000, and a soft update parameter (τ) of 0.005. The training process involves initializing networks and optimizers, collecting experiences, sampling batches, updating the critic and value networks to minimize their respective losses, updating the actor network to maximize the expected return while considering entropy, and softly updating the target networks. This comprehensive setup ensures that the SAC algorithm effectively manages the trading process within our proposed framework.

In this section, we implement the new environment of TEM, analyze its performance under varying conditions of voltage magnitude, demand, and supply. Then we assume a single agent as a buyer or a seller who tends to increase their payoff by using the SAC algorithm. Subsequently, we extend our analysis to consider multiple agents, where four buyers or sellers are involved in a game and use the SAC algorithm to maximize their respective payoffs. The existence of NE in this game is then examined.

4.1 | Implementation of the proposed TEM

The new TEM is implemented for the modified 33-bus IEEE distribution system. We consider four sellers and four buyers in the TEM. The voltage conditions are categorized into three states: (1) normal voltage magnitude, (2) overvoltage at one or more buses, and (3) voltage magnitude lower than normal.

4.1.1 | Normal condition for voltage magnitude

Table 2 illustrates participants' offers and outcomes in TEM when offers create a normal condition for voltage magnitude. In this table, the prices and active powers for selling or buying that submit to the operator and the results of the TEM after three main processes are reported. Since the total demand in the TEM is more than the total active power for selling, the operator divides the active power among the buyers, and sellers are able to sell all their surplus active power in the TEM. As you can see, each buyer who submits a higher price for buying energy can buy more active power in the TEM according to the Equation (12) and they have to buy lack of active power from the grid with a higher price. Also, the clearing price in the TEM is nearer to the weighted average price of sellers because demand is more than production.

TABLE 2 Offers and results of participants in the TEM with normal condition of voltage magnitude.

Seller or buyer	Surplus energy (kwh)	Price for selling or buying (cents/kwh)	Awarded power in TEM (kwh)	The energy clearing price (cents/kwh)
bus1	4	12	4	11.42
bus2	8	13	8	
bus3	7	11	7	
bus4	3	12.5	3	
bus5	−12	9	**−5.42**	
bus6	−13	8	**−4.82**	
bus7	−14	9.5	**−5.73**	
bus8	−11	10	**−6.03**	

TABLE 3 Offers and results of participants in the TEM with over voltage condition.

Seller or buyer	Surplus energy (kwh)	Price for selling or buying (cents/kwh)	Awarded power after optimization (kwh)
bus1	16	12	**13.22**
bus2	16	13	**12.8**
bus3	14	11	**11.2**
bus4	11	12.5	**8.8**
bus5	−11		
bus6	−10		
bus7	−12		
bus8	−13		

4.1.2 | The overvoltage condition

Table 3 shows offers and results in the TEM when the voltage magnitude in the slack bus is 1.05 per unit. This voltage magnitude happens in the middle of the day because of low demand and high generation. Thus, the PV injection can cause overvoltage conditions in the distribution system. In these situations, the TEM operator runs the optimization to modify the sellers' offers according to their bids. Besides the grid physical condition, the price that sellers submit to the operator affects the awarded power by the optimization. As a result, the sellers who submit a lower price for selling ackWhe power can sell more active power in TEM.

TABLE 4 Offers and results of participants in the TEM with under voltage condition.

Seller or buyer	Surplus energy (kWh)	Price for selling or buying (cents/kwh)	Awarded power after optimization (kWh)
bus1	11		
bus2	10		
bus3	12		
bus4	13		
bus5	−15	9	**−13.18**
bus6	−15	8	**−12**
bus7	−14	9	**−11.2**
bus8	−12	10	**−9.6**

4.1.3 | The low voltage condition

Table 4 presents offers and outcomes in the TEM, when the voltage magnitude in the slack bus is 0.95 per unit. Since the demand is more than the generation in the peak hours, the voltage magnitude is the lowest amount in the day. Therefore, the operator may reduce the amount of buyers' demand by optimization. As we mentioned, this optimization is affected by prices and physical condition, which the buyers with higher prices can buy more active power in the TEM.

4.2 | Implementation of the proposed TEM for a market with more participants

In this section, we further extend our analysis through an additional case study designed to evaluate the scalability and robustness of our platform-based framework under more complex conditions. For this purpose, we employ the IEEE 33 bus system as a model to simulate a TEM scenario involving 15 sellers and 15 buyers. In this simulated market, each seller attempts to sell 5 kW of active power, and each buyer is set to purchase 3 kW. The sellers' and buyers' weighted average prices are calculated to be 8.93 cents/kW and 8.07 cents/kW, respectively. Consequently, the TEM's clearing price is established at 8.39 cents/kW, closer to the buyers' average price, indicative of a supply exceeding demand situation. This simulation demonstrates the effectiveness of our framework in a diversified market environment. Notably, Seller 1, offering power at a competitive price of 7.5 cents/kW, successfully sells their entire 5 kW allocation in TEM, while Seller 15, quoting a higher price of 9 cents/kW, is only able to sell 2.69 kW in TEM and should sell the remaining power to the grid at a lower price.

4.3 | SAC for single-agent problem

The agents in the TEM want to raise their rewards. To improve the payoff, the agent must find the best policy quickly. DRL can help the agents discover the optimal bidding strategy during the

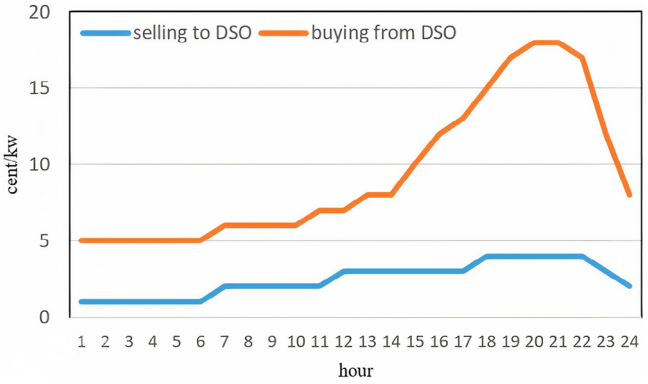

FIGURE 4 The trading prices with the grid.

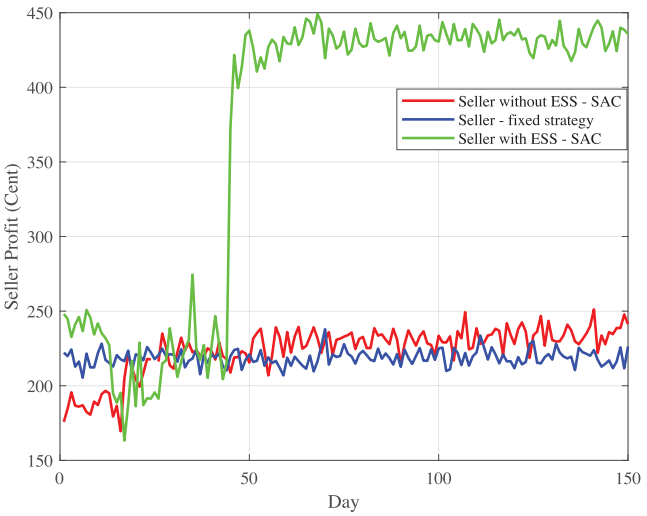

FIGURE 5 Agent's profit (as a seller) with and without ESS when using the SAC algorithm or heuristic-based method, with other participants submitting their offers following a normal distribution.

day. In this section, we consider an agent as a buyer and a seller who can use SAC algorithm while the remaining participants submit their offers based on a normal distribution with a reasonable mean and standard deviation, which are not revealed to other players.

In our analysis, we consider an agent who possesses an ESS and aims to enhance their payoff over a day. The states in the DRL algorithm are defined by the agent's SoC at each hour. To expedite convergence, we divide the day into four intervals with comparable generation and demand patterns, which results in a smaller state space dimension.

Participants in the TEM must submit a price between the price that grid buys and sells active power in distribution systems. We use Figure 4 as a reference that sellers and buyers can select their offers, which is similar to prices in [7].

The agent should submit the price and amount of active power to the TEM operator. The SAC algorithm is used by the agent to increase their reward in the day from selling the active power in the TEM. Figure 5 illustrates the various types of profits an agent can earn as a seller in three different

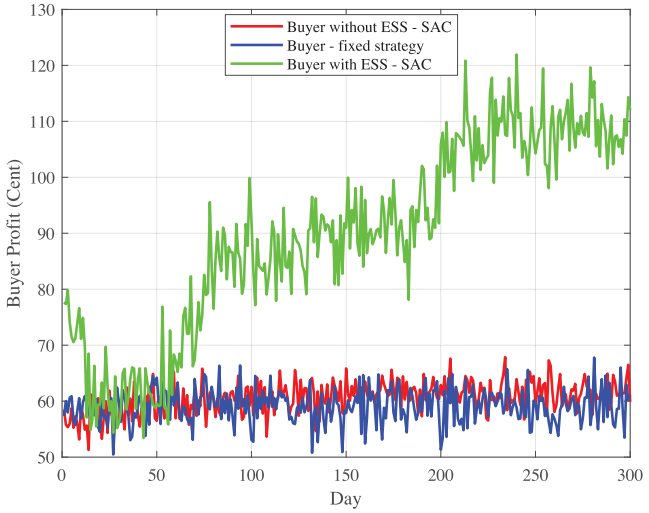

FIGURE 6 Agent's profit (as a buyer) with and without ESS when using the SAC algorithm or heuristic-based method, with other participants submitting their offers following a normal distribution.

scenarios: (1) selling active power in the TEM without ESS, (2) owning ESS and selling active power in the TEM, and (3) selling active power using a fixed strategy, which involves setting the price as the average of the minimum and maximum prices for the given hour and the power bid equal to the excess generation power. By participating in the TEM and selling active power, the agent can earn higher rewards, which can be further increased if the agent owns ESS. The application of the SAC algorithm shows significant improvement compared to the heuristic-based method. The SAC algorithm enables the agent to converge to the optimal bidding strategy within 30 days without ESS and 60 days with ESS. After convergence, fluctuations in the offers submitted by other participants to the TEM are limited, as these offers are assumed to follow a normal distribution.

To evaluate the optimality of the convergence point of the SAC algorithm, it can be compared with a strategy based on knowing complete information. If a seller knows all other players' bids, the seller can find the best strategy without using the SAC algorithm, meaning the seller has complete information. With this information, the seller charges the ESS in state 1, partially charges in states 2 and 3, and discharges the ESS in state 4. When using the ESS, selecting the amount of power for trading is more important than selecting the price for participating in the TEM, because selling power during peak hours increases the seller's profit. According to this strategy, which is based on complete information and considers other players' bids as constant, the maximum profit is 427.0266 cents per day. By comparing this amount with the data in Figure 5, it can be concluded that the SAC algorithm performs exceptionally well. After converging, the profit oscillates close to the profit calculated based on the complete information strategy, demonstrating the effectiveness of the SAC algorithm.

To optimize daily costs, the agent utilizes the SAC algorithm for submitting bids on both price and active power quantity to the TEM operator. Figure 6 depicts the three different profit scenarios achievable by a buyer within a day, contingent on

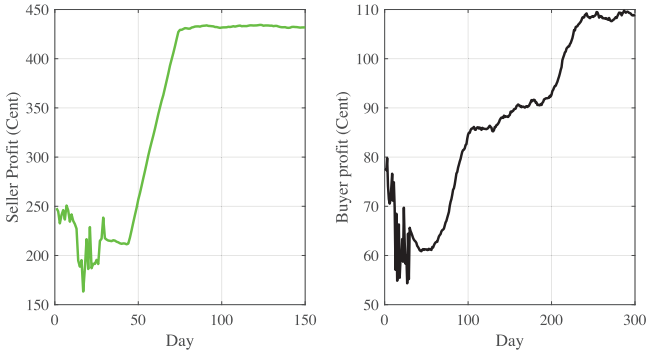

FIGURE 7 Average profit for a seller and buyer over 30 consecutive days.

the presence or absence of ESS. The application of the SAC algorithm shows improvement compared to the heuristic-based method. By applying SAC, a buyer without ESS attains the optimal policy within 50 days, whereas a buyer with ESS takes approximately 200 days due to the complexity of the state space. Consequently, the agent earns more rewards with the SAC algorithm when participating in the TEM and can further increase these rewards by utilizing ESS.

To evaluate the optimality of the SAC algorithm for a buyer, consider a scenario where the buyer knows all information about bids in the TEM, meaning the buyer has complete information. With this information, the buyer purchases power and charges the ESS in states 1 and 2, then discharges the ESS in state 4 to supply their demand during peak hours. According to this strategy, which is based on complete information and considers other players' bids as constant, the maximum profit is 108.2931 cents per day. By comparing this amount with the data in Figure 6, it can be concluded that the SAC algorithm performs exceptionally well for a buyer.

The stability of the SAC algorithm in applications involving both sellers and buyers can be observed by examining the average reward over the last 30 days, as shown in Figure 7. As shown in this Fig, after an initial exploration phase, the profit curve demonstrates increasing over time, ultimately converging towards the optimal policy.

4.4 | Implementation of SAC algorithm as multi-agent

The simulation, employing the SAC algorithm in a multi-agent setting with eight participants, covers a span of 350 days, each represented by continuous 24-h cycles, and is completed in just 16 min. Notably, a substantial part of this time is dedicated to running the SAC algorithm for each participant, a process that can be parallelized in real-world applications. The simulation is executed using Google Colab Pro with an Nvidia V100 GPU and 32 GB of RAM.

In the strategic game played within the TEM, agents are incentivized to adopt the SAC algorithm to optimize their rewards throughout the day. This approach is applicable for both sellers and buyers participating in the market. Consider a

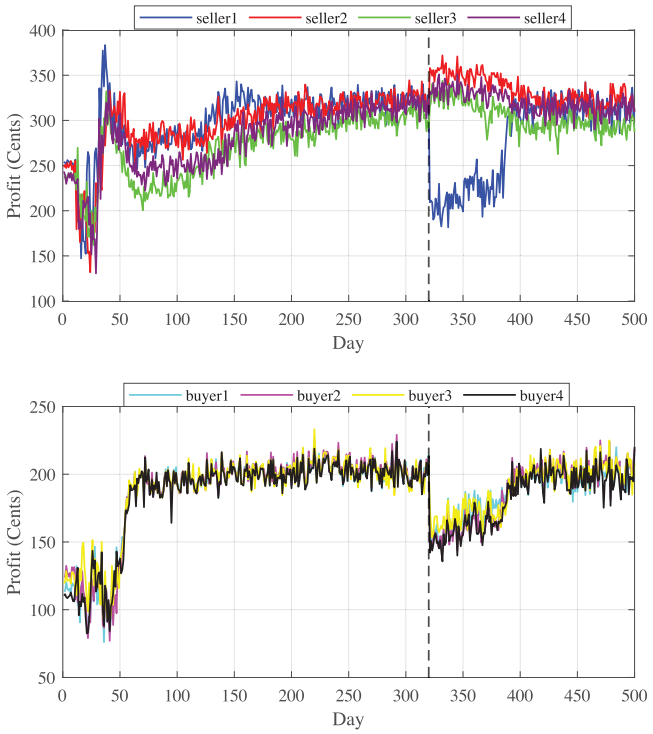

FIGURE 8 Game dynamics in TEM showing sellers and buyers using the SAC algorithm, with a focus on the probability convergence to NE.

scenario where four sellers and four buyers implement the SAC algorithm in TEM, each with varying ESS capacities (5 kWh, 5 kWh, 3 kWh, and 4 kWh, respectively, for sellers, mirrored by the buyers). Figure 8 illustrates the rewards of all eight participants, demonstrating the competitive dynamics as each one strives to maximize their profit against the backdrop of simultaneous actions by other market players.

In Figure 8, the convergence of all players to a specific point is evident before before seller 1 uses a new SAC algorithm, indicating that it may represent an NE in this game. To thoroughly examine this point, we conduct a controlled experiment by changing the actions of Seller 1 for a duration of 30 days. The outcomes reveal that Seller 1 consistently earns lower profits when the policies of other players remain unchanged. On day 350, Seller 1 starts using a new SAC algorithm while the strategies of other agents are fixed (Single agent learning). Remarkably, the strategy of this test agent converges to the same strategy observed in the multi-agent learning case, demonstrating that the converged policy for Seller 1 in multi-agent learning serves as the best response strategy to the strategies of other players. This comprehensive investigation provides support for the assertion that the identified point, where all players converge, indeed constitutes an NE of this game.

5 | CONCLUSION

Small prosumers in distribution systems can earn more profit or spend less cost when they participate in TEMs. In this market, they can balance their demand and supply by trading with other consumers or producers in the real-time. They should select a price and amount of active power to trade in this market that is challenging because their bids have a direct impact on their reward. In this article, DRL is proposed to prosumers to apply them for finding the best policy in the shortest time. To use RL and DRL, an environment, a state-action space, and an algorithm should be determined. First, an novel environment that shows the TEM framework is introduced so that participants can trade active power in it. Then the created game between all participants including buyers and sellers is investigated in this market, and the SAC algorithm is presented for players to improve their payoffs in this game. This algorithm is applied by an agent, who is a seller or a buyer, with a novel continuous state-action space. After using the SAC algorithm as a single agent, we consider all players can apply this algorithm as Multi-Agent in TEM. In the multi-agent problem, it is demonstrated, through the application of the NE definition, that buyers or sellers attain either the unique NE or one of multiple NEs when all participants employ SAC algorithm.

Future research could explore several promising directions to enhance the proposed framework and its applications in transactive energy markets. One potential area is the integration of more advanced machine learning techniques to improve the accuracy of demand and generation forecasting. Additionally, the inclusion of other types of DERs, such as electric vehicles and demand response programs, could provide a more comprehensive and robust TEM model. Investigating the impact of different market conditions and regulatory environments on the performance of the proposed algorithms would also be valuable. Furthermore, real-world pilot projects could be implemented to validate the theoretical models and algorithms in practical scenarios. These studies could provide insights into the scalability and adaptability of the proposed TEM framework in diverse geographic and economic contexts.

AUTHOR CONTRIBUTIONS

All authors played integral roles in this study. The specific contributions of each author are as follows: **Hossein Mohsenzadeh-Yazdi**: Methodology; software; writing—original draft. **Hamed Kebriaei**: Project administration; supervision; validation; writing—review and editing. **Farrokh Aminifar**: Project administration; supervision; validation; writing—review and editing. These contributions collectively contributed to the completion and success of the research.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The hourly price data used in this study is sourced from reference [7]. The IEEE 33-bus system was utilized in this research. No additional data was used in this paper.

ORCID

Hamed Kebriaei https://orcid.org/0000-0002-3781-2163
Farrokh Aminifar https://orcid.org/0000-0003-2331-2798

REFERENCES

1. Akrami, A., Doostizadeh, M., Aminifar, F.: Power system flexibility: An overview of emergence to evolution. J. Mod. Power Syst. Clean Energy 7(5), 987–1007 (2019)
2. Siano, P., De Marco, G., Rolán, A., Loia, V.: A survey and evaluation of the potentials of distributed ledger technology for peer-to-peer transactive energy exchanges in local energy markets. IEEE Syst. J. 13(3), 3454–3466 (2019)
3. Zhang, Z., Tang, H., Wang, P., Huang, Q., Lee, W.-J.: Two-stage bidding strategy for peer-to-peer energy trading of nanogrid. IEEE Trans. Ind. Appl. 56(2), 1000–1009 (2019)
4. Tushar, W., Saha, T.K., Yuen, C., Morstyn, T., McCulloch, M.D., Poor, H.V., Wood, K.L.: A motivational game-theoretic approach for peer-to-peer energy trading in the smart grid. Appl. Energy 243, 10–20 (2019)
5. Cicek, A., Güzel, S., Erdinc, O., Catalao, J.P.: Comprehensive survey on support policies and optimal market participation of renewable energy. Electr. Power Syst. Res. 201, 107522 (2021)
6. Vahidinasab, V., Mohammadi-Ivatloo, B.: Demand-Side Peer-to-Peer Energy Trading. Springer Nature, Cham (2023)
7. Yan, M., Shahidehpour, M., Paaso, A., Zhang, L., Alabdulwahab, A., Abusorrah, A.: Distribution network-constrained optimization of peer-to-peer transactive energy trading among multi-microgrids. IEEE Trans. Smart Grid 12(2), 1033–1047 (2020)
8. Ye, Y., Tang, Y., Wang, H., Zhang, X.-P., Strbac, G.: A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading. IEEE Trans. Smart Grid 12(6), 5185–5200 (2021)
9. Li, J., Zhang, C., Xu, Z., Wang, J., Zhao, J., Zhang, Y.-J.A.: Distributed transactive energy trading framework in distribution networks. IEEE Trans. Power Syst. 33(6), 7215–7227 (2018)
10. Wu, Z., Wang, J., Zhong, H., Gao, F., Pu, T., Tan, C.-W., Chen, X., Li, G., Zhao, H., Zhou, M., et al.: Sharing economy in local energy markets. J. Mod. Power Syst. Clean Energy 11(3), 714–726 (2022)
11. Bedoya, J.C., Ostadijafari, M., Liu, C.-C., Dubey, A.: Decentralized transactive energy for flexible resources in distribution systems. IEEE Trans. Sustainable Energy 12(2), 1009–1019 (2020)
12. Lian, J., Ren, H., Sun, Y., Hammerstrom, D.J.: Performance evaluation for transactive energy systems using double-auction market. IEEE Trans. Power Syst. 34(5), 4128–4137 (2018)
13. Liu, W., Zhan, J., Chung, C.: A novel transactive energy control mechanism for collaborative networked microgrids. IEEE Trans. Power Syst. 34(3), 2048–2060 (2018)
14. Nizami, M.S.H., Hossain, M.J., Fernandez, E.: Multiagent-based transactive energy management systems for residential buildings with distributed energy resources. IEEE Trans. Ind. Inf. 16(3), 1836–1847 (2019)
15. Gu, N., Cui, J., Wu, C.: Power-electronics-enabled transactive energy market design for distribution networks. IEEE Trans. Smart Grid 13(5), 3968–3983 (2021)
16. Haghifam, S., Laaksonen, H., Shafie-Khah, M.: Modeling a local electricity market for transactive energy trading of multi-aggregators. IEEE Access 10, 68 792–68 806 (2022)
17. Lilla, S., Orozco, C., Borghetti, A., Napolitano, F., Tossani, F.: Day-ahead scheduling of a local energy community: An alternating direction method of multipliers approach. IEEE Trans. Power Syst. 35(2), 1132–1142 (2019)
18. Xiong, L., Tang, Y., Mao, S., Liu, H., Meng, K., Dong, Z., Qian, F.: A two-level energy management strategy for multi-microgrid systems with interval prediction and reinforcement learning. IEEE Trans. Circuits Syst. I Regul. Pap. 69(4), 1788–1799 (2022)
19. Saif, A., Khadem, S.K., Conlon, M., Norton, B.: Hosting a community-based local electricity market in a residential network. IET Energy Syst. Integr. 4(4), 448–459 (2022)
20. Hirth, L.: The market value of variable renewables: The effect of solar wind power variability on their relative price. Energy Econ. 38, 218–236 (2013)
21. Nunna, H.K., Sesetti, A., Rathore, A.K., Doolla, S.: Multiagent-based energy trading platform for energy storage systems in distribution systems with interconnected microgrids. IEEE Trans. Ind. Appl. 56(3), 3207–3217 (2020)
22. Cicek, A., Erdinc, O.: Risk-averse optimal bidding strategy for a wind energy portfolio manager including ev parking lots for imbalance mitigation. Turk. J. Electr. Eng. Comp. Sci. 29(2), 481–498 (2021)
23. Çiçek, A., Erdinç, O.: Optimal bidding strategy considering bilevel approach and multistage process for a renewable energy portfolio manager managing RESs with ESS. IEEE Syst. J. 16(4), 6062–6073 (2021)
24. Almasan, P., Suárez-Varela, J., Wu, B., Xiao, S., Barlet-Ros, P., Cabellos-Aparicio, A.: Towards real-time routing optimization with deep reinforcement learning: Open challenges. In: 2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR), pp. 1–6. IEEE, Piscataway (2021)
25. Mosavi, A., Faghan, Y., Ghamisi, P., Duan, P., Ardabili, S.F., Salwana, E., Band, S.S.: Comprehensive review of deep reinforcement learning methods and applications in economics. Mathematics 8(10), 1640 (2020)
26. Tajeddini, M., Kebriaei, H., Imani, M.: Bidding strategy in pay as bid markets by multi-agent reinforcement learning. In: 28th International Power System Conference. IEEE, Piscataway (2013)
27. Sabir, S., Kelouwani, S., Hosseini, S.S., Henao, N., Fournier, M., Agbossou, K.: A bid selection model for computational cost reduction of transactive energy aggregator in smart grids. In: 2022 North American Power Symposium (NAPS), pp. 1–6. IEEE, Piscataway (2022)
28. Taghizadeh, A., Montazeri, M., Kebriaei, H.: Deep reinforcement learning-aided bidding strategies for transactive energy market. IEEE Syst. J. 16(3), 4445–4453 (2022)
29. Ye, Y., Qiu, D., Li, J., Strbac, G.: Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach. IEEE Access 7, 130515–130529 (2019)
30. Du, Y., Li, F., Zandi, H., Xue, Y.: Approximating nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning. J. Mod. Power Syst. Clean Energy 9(3), 534–544 (2021)
31. Rokhforoz, P., Montazeri, M., Fink, O.: Multi-agent reinforcement learning with graph convolutional neural networks for optimal bidding strategies of generation units in electricity markets. Expert Syst. Appl. 225, 120010 (2023)
32. Wang, J., Wu, J., Kong, X.: Multi-agent simulation for strategic bidding in electricity markets using reinforcement learning. CSEE J. Power Energy Syst. 9(3), 1051–1065 (2023)
33. Qiu, D., Wang, J., Dong, Z., Wang, Y., Strbac, G.: Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading. IEEE Trans. Power Syst. 38(5), 4853–4866 (2023)
34. Taheri, B., Safdarian, A., Moeini-Aghtaie, M., Lehtonen, M.: Distribution system resilience enhancement via mobile emergency generators. IEEE Trans. Power Delivery 36(4), 2308–2319 (2020)
35. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning, pp. 1861–1870. PMLR, New York (2018)
36. Hu, J., Wellman, M.P.: Nash q-learning for general-sum stochastic games. J. Mach. Learn. Res. 4, 1039–1069 (2003)
37. Ozdaglar, A.: Game theory with engineering applications. In: Lecture Notes. MIT, Cambridge, MA (2011)
38. Zhong, H., Yang, Z., Wang, Z., Jordan, M.I.: Can reinforcement learning find Stackelberg-Nash equilibria in general-sum Markov games with myopically rational followers? J. Mach. Learn. Res. 24(35), 1–52 (2023)

How to cite this article: Mohsenzadeh-Yazdi, H., Kebriaei, H., Aminifar, F.: Multi-agent reinforcement learning in a new transactive energy mechanism. IET Gener. Transm. Distrib. 18, 2943–2955 (2024). https://doi.org/10.1049/gtd2.13244