A Scalable Privacy-Preserving Multi-Agent Deep Reinforcement Learning Approach for Large-Scale Peer-to-Peer Transactive Energy Trading

Yujian Ye, *Member, IEEE*, Yi Tang, *Senior Member, IEEE*, Huiyu Wang, *Student Member, IEEE*,
Xiao-Ping Zhang, *Fellow, IEEE*, and Goran Strbac, *Member, IEEE*

***Abstract*—Peer-to-peer (P2P) transactive energy trading has emerged as a promising paradigm towards maximizing the flexibility value of prosumers' distributed energy resources (DERs). Despite reinforcement learning constitutes a well-suited model-free and data-driven methodological framework to optimize prosumers' energy management decisions, its application to the large-scale coordinated management and P2P trading among multiple prosumers within an energy community is still challenging, due to the scalability, non-stationarity and privacy limitations of state-of-the-art multi-agent deep reinforcement learning (MADRL) approaches. This paper proposes a novel P2P transactive trading scheme based on the multi-actor-attention-critic (MAAC) algorithm, which addresses the above challenges individually. This method is complemented by a P2P trading platform that incentivizes prosumers to engage in local energy trading while also penalizes each prosumer's addition to rebound peaks. Case studies involving a real-world, large-scale scenario with 300 residential prosumers demonstrate that the proposed method significantly outperforms the state-of-the-art MADRL methods in reducing the community's cost and peak demand.**

***Index Terms*—Distributed energy resources, energy management, local energy community, multi-agent deep reinforcement learning, peer-to-peer transactive energy trading.**

NOMENCLATURE

$t, T, \Delta t$ Temporal index, horizon and resolution of coordination problem

i, N Index and number of prosumers

λ_t^b, λ_t^s Retail import and export prices at step t (pence/kWh)

$\lambda_t^{L,b}, \lambda_t^{L,s}$ Local market buy and sell prices at step t (pence/kWh)

$P_{i,t}^{pv}$ Photovoltaic (PV) production of prosumer i at step t (kW)

$P_{i,t}^{nd}$ Non-shiftable load of prosumer i at step t (kW)

$V_{i,t}^{ev}$ Binary variable indicating whether the electric vehicle (EV) of prosumer i charges ($V_{i,t}^{ev} = 1$) or discharges ($V_{i,t}^{ev} = 0$) at step t

$C_{i,t}^{ev}, D_{i,t}^{ev}$ Charging and discharging power of EV of prosumer i at step t (kW)

$\overline{P}_i^{ev}$ Power capacity of EV battery of prosumer i (kW)

$E_{i,t}^{ev}$ Energy in EV of prosumer i at step t (kWh)

$\underline{E}_i^{ev}, \overline{E}_i^{ev}$ Minimum and maximum energy limits of EV battery of prosumer i (kWh)

$E_{i,t}^{tr}$ Energy requirement of EV of prosumer i for commuting purposes at step t (kWh)

$\eta_i^{evc}, \eta_i^{evd}$ Charging and discharging efficiencies of EV of prosumer i

t_i^{dep}, t_i^{arr} Departure and arrival times of EV of prosumer i

$A_{i,t}^{ev}$ Binary parameter indicating whether the EV of prosumer i is available to be scheduled ($A_{i,t}^{ev} = 1$) or not ($A_{i,t}^{ev} = 0$) at step t.

$V_{i,t}^{es}$ Binary variable indicating whether the energy storage (ES) of prosumer i charges ($V_{i,t}^{es} = 1$) or discharges ($V_{i,t}^{es} = 0$) at step t

$C_{i,t}^{es}, D_{i,t}^{es}$ Charging and discharging power of ES of prosumer i at step t (kW)

$\overline{P}_i^{es}$ Power capacity of ES of prosumer i (kW)

$E_{i,t}^{es}$ Energy in ES of prosumer i at step t (kWh)

$\underline{E}_i^{es}, \overline{E}_i^{es}$ Minimum and maximum energy limits of ES of prosumer i (kWh)

$\eta_i^{esc}, \eta_i^{esd}$ Charging and discharging efficiencies of ES of prosumer i

τ Index of phases of the smart appliance (SA) cycle

$P_{i,\tau}^{cyc}$ Power demand at phase τ of the SA cycle of prosumer i (kW)

$P_{i,t}^{sa}$ Power demand of SA of prosumer i at step t (kW)

T_i^{dur} Duration of SA cycle of prosumer i

t_i^{in}, t_i^{ter} Earliest initiation and latest termination times of SA cycle of prosumer i

$A_{i,t}^{sa}$ Binary parameter indicating whether the SA of prosumer i is available to be scheduled ($A_{i,t}^{sa} = 1$) or not ($A_{i,t}^{sa} = 0$) at step t

Manuscript received December 27, 2020; revised May 28, 2021 and August 5, 2021; accepted August 7, 2021. Date of publication August 10, 2021; date of current version October 21, 2021. This work was supported by the National Natural Science Foundation of China under Grant 51877037. Paper no. TSG-01920-2020. *(Corresponding author: Yujian Ye.)*

Yujian Ye, Yi Tang, and Huiyu Wang are with the School of Electrical Engineering, Southeast University, Nanjing 210096, China (e-mail: yeyujian@seu.edu.cn; tangyi@seu.edu.cn; wanghuiyu@seu.edu.cn).

Xiao-Ping Zhang is with the School of Electronic, Electrical and Computer Engineering, University of Birmingham, Birmingham B15 2TT, U.K. (e-mail: x.p.zhang@bham.ac.uk).

Goran Strbac is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: g.strbac@imperial.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2021.3103917.

Digital Object Identifier 10.1109/TSG.2021.3103917

$z^{sa}_{i,t}$ Binary variable indicating whether the SA cycle of prosumer i is initiated ($z^{sa}_{i,t} = 1$) or not ($z^{sa}_{i,t} = 0$) at step t

$H^{in}_{i,t}$ Indoor temperature of prosumer i at step t (^{o}C)

H^{out}_{t} Outdoor temperature at step t (^{o}C)

$\overline{H}, \underline{H}$ Maximum and minimum comfort levels of the indoor temperature (^{o}C)

$P^{ac}_{i,t}$ Power demand of heating, ventilation, and air conditioning (HVAC) system of prosumer i at step t (kW)

$\overline{P}^{ac}_{i}$ Maximum power rate of HVAC system of prosumer i (kW)

η^{ac}_{i} Energy efficiency of HVAC system of prosumer i

C^{ac}_{i} Thermal capacity of HVAC system of prosumer i (kWh/^{o}F)

R^{ac}_{i} Thermal resistance of HVAC system of prosumer i (^{o}F/kW)

$l_{i,t}$ Net demand (positive)/generation (negative) of prosumer i at step t (kW)

P^{nc}_{t} Net demand of the community at step t (kW)

P^{ng}_{t} Net generation of the community at step t (kW).

I. Introduction

A. Background and Motivation

GROWING proliferation of distributed energy resources (DERs) at power distribution and utilization systems, complemented by advancements in smart metering, energy management, information and communication technologies has converted the traditionally passive electricity consumers to *prosumers* who are able to actively monitor and control their consumption, generation and storage of energy close to real-time with for higher cost savings. However, the bi-directional power flow brought by DERs and the intrinsic variability and limited controllability of distributed renewable generation sources introduce critical techno-economic challenges to electricity systems' operation [1].

In this context, conventional retail market is no longer fit-for-purpose, and alternative market schemes need to be devised to more effectively coordinate and provide incentives for prosumers to harvest the flexibility resided in their DERs and subsequently deal with the challenges brought by themselves. In this context, peer-to-peer (P2P) transactive energy trading has recently arisen as a new market paradigm, promising a more coordinated and comprehensive exploitation of the available renewable production and DER flexibility of the prosumers. Specifically, P2P market structures allow direct energy trading between prosumers within a local energy collective, reducing their energy reliance on incumbent suppliers, who tend to offer higher tariffs for selling energy to the prosumers and lower tariffs for purchasing excess energy from them [2]. In addition, electricity demand and generation are balanced locally, leading to deferral or even avoidance of distribution network reinforcements and improved power system reliability. Furthermore, residential prosumers' involvement level are also enhanced by shaping a local and socially cooperative energy identity [3], [4].

B. Literature Review

The literature converges so far on two main categories of approaches for coordinated management of energy collectives and P2P transcative trading. The first category focuses on *system-centric coordination* approaches which employs a central supervisory entity to collect the economic and technical parameters of DER and match the generating and consuming peers in order to maximize the overall welfare of the community [4]– [11]. While such centralized approaches can theoretically provide the optimal coordination solution from the system perspective [12], they exhibit multifarious practical bottlenecks: 1) high communication requirements and costs associated with the necessity of transmitting the diverse and complex operating parameters and constraints of a large number of prosumers to the central coordinator and prone to a single-point-of-failure; 2) poor computational scalability owing to the need of the central coordinator to solve an optimization problem involving a massive number of decision variables and constraints; and 3) provoking prosumers' privacy concerns since generally they are not willing to reveal private information (e.g., their energy usage behaviors and DER portfolio) and be directly managed by external bodies.

The second category focuses on *prosumer-centric coordination* approaches, enabling individual prosumers to retain complete control over their DERs. Such distributed energy management significantly lightens the communication and computational requirements and partially addresses the privacy concerns. Iterative auction-based market mechanisms are employed in [13]–[15] to encourage prosumers to trade energy with each other. However, how to integrate the scheduling model of the prosumers' DERs with such an information-based trading mechanism poses a major challenge for optimizing the trading strategies of the prosumers, who generally have a relatively low level of professionalism and their bidding/offering patterns may be volatile. Authors in [16], [17] consider a P2P energy sharing game for a small community of prosumers. Each prosumer first optimizes its individual cost without energy sharing. An iterative approach is then adopted to determine the prosumers' energy sharing profiles and the corresponding payments in the sharing game. However, the scalability of this approach is questionable in practical applications with a large number of prosumers. A P2P energy sharing model is proposed in [18]. In this work, the local trading prices are determined iteratively based on the supply and demand ratio of the prosumer community. However, the employed iterative pricing approach may not converge and its performance is impacted by the level of demand side participation. Authors in [19] explore the application of the mid-market rate (MMR) and bill sharing pricing mechanisms to encourage P2P energy sharing between prosumers.

In both the above literature categories, model-based energy management approaches are adopted to optimize the operation of prosumers' DERs. Such approaches rely on complete knowledge on the DER operating models and accurate forecasts of exogenous parameters (such as PV production). As a result, the quality of the resulting coordination decisions may be deteriorated by the inaccuracy of the employed DER models and the forecasting method. Although scenario-based stochastic programming approaches [4], [5], [11] can be employed to account for the inherent uncertainties associated with these exogenous parameters, they are often criticized by inaccuracy and high computational demand, driven by the challenges of determining a representative number of scenarios that is computationally manageable and encompass all significant variations of the uncertainties [20].

In view of these drawbacks, a growing interest in applying the model-free reinforcement learning (RL) approach, to optimize prosumers' energy management decisions, has been witnessed in recent years. RL constitutes the most relevant class of machine learning approaches (other classes include supervised learning and unsupervised learning) for decision-making problems, driven by the employment of reward or penalty functions guiding the actions of the involved agents [21]. Within this RL framework, an energy management system (EMS) installed in the prosumers' premises gradually learns its optimal management policies based on experiences from the recurrent interactions with the environment, without full identification and no *a priori* knowledge of the latter. Additionally, RL can harness the increasing volume of data collected from various sensors and perform successive interpretation of data and learn optimal energy management strategies which cope with the system uncertainties. Furthermore, combining RL with deep neural networks (DNN) enables effective learning of more sophisticated control policies than those discovered by traditional RL methods founded on look-up tables or shallow regression models based on hand-crafted features [22].

Driven by these desirable properties, previous works have employed various RL methods to optimize energy schedules of prosumers' DERs, as reviewed in [23], [24]. The majority of them, however, only considers energy management optimization for a single prosumer, employing *single-agent* RL (SARL) methods. On the other hand, prior research effort on the development and application of *multi-agent* RL (MARL) for the coordinated energy management of multiple prosumers and P2P transactive trading is still thin and emerging. The *concurrent* learning framework is adopted in [25]–[27] to train multiple EMS agents, each optimizes the corresponding prosumer's DER schedule. However, the selfish behavior of prosumers aiming at minimizing their individual costs may lead to demand/generation concentration effect, threatening the security of the local distribution network [28]. More importantly, since hundreds of prosumers are independently updating their management policies as learning progresses, the community environment appears to be *non-stationary* from the view of any individual prosumer [29].

The *centralized* learning framework [25], [27] provides an effective remedy for tackling the non-stationarity, which takes as input the observations and actions of all agents and learns actions jointly for all agents. However, this approach is not scalable since the input space increases exponentially with the number of agents. Secondly, both training and execution stages are centralized, resulting in high resource requirements for practical deployments. The *centralized training with decentralized execution* (CTDE) framework [30], [32]–[27] constitutes another remedy towards eliminating the environmental non-stationarity. CTDE allows the policies to use extra information to ease training, but this information is not used at test time. Similar to centralized learning, CTDE is not privacy-preserving and shares the same dimensionality drawback in training the central critic, which is problematic in large-scale multi-agent applications.

C. Paper Contributions

This paper addresses the challenging large-scale coordinated energy management and P2P transactive trading of multiple prosumers within an energy collective by developing a tailor-made, scalable and privacy-preserving MADRL method. Concretely, the contributions of this paper are fourfold.

- This paper considers the coordination of a large number of prosumers operating multiple and diverse DERs, including nonshiftable demand, a non-dipsatchable PV generator, an energy storage (ES) unit, and three types of shiftable demand, namely electric vehicles (EV) with flexible charging and Vehicle-to-Grid (V2G) / Vehicle-to-Home (V2H) capabilities, heating, ventilation, air conditioning (HVAC) with certain comfortable temperature margins, and smart appliances (SA) with deferrable cycles.

- Energy trading within the collective is administrated by a P2P trading platform which employs the MMR pricing mechanism to adequately remunerate prosumers for their participation in P2P trading, as well as a novel reward shaping mechanism which penalizes each prosumer' addition to rebound peaks, thereby preserving the security of the local distribution network.

- This paper proposes a novel P2P transactive trading scheme based on the *multi-actor-attention-critic* (MAAC) algorithm to overcome the drawbacks of state-of-the-art MARL approaches. This method inherits from the CTDE paradigm which keeps the critics centralized but the actors decentralized. Moreover, it employs the attention mechanism to selectively incorporate agents' representative information for estimating the critics; furthermore it enables learning of all agents' critics jointly by sharing a set of learnable parameters among agents. On this account, MAAC significantly improves the scalability, eliminates the environmental non-stationarity as well as preserves prosumers' privacy.

- Case studies involving a real-world, large-scale scenario with 300 residential prosumers demonstrate that the proposed method significantly outperforms the state-of-the-art methods in reducing the energy collective' cost and peak demand.

D. Paper Organization

The rest of this paper is organized as follows. Section II provides the detailed mathematical operating models of the examined DERs and the formulation of the model-based, system-centric coordination optimization. Section III details the P2P transactive trading mechanism in the local energy market. Section IV introduce the Markov Game formulation of the multi-agent coordination problem. Section V outlines the background of RL, compares the state-of-the-art MARL methods and orchestrates the MAAC method. Section VI presents the case studies demonstrating the effectiveness of the MAAC method. Finally, Section VII discusses conclusions and future extension of this work.

II. Formulation of Model-Based System-Centric Coordination Approach

The coordination and P2P trading problem of an energy community composed of a large group of residential electricity prosumers is investigated. The prosumers are residing to the same low-voltage substation and operate diverse portfolios of DERs, which consist of solar PV panels, ES units, non-shiftable demand (e.g., lighting loads), three different types of shiftable demand (including EV, HVAC system, and SA). DER operating parameters driven by prosumers' preferences and requirements are diversified, accounting for the natural variability of different prosumers (Section VI-A). These DERs are manged by EMS across a daily horizon of 48 half-hourly time steps.

A. Electric Vehicle and Energy Storage

The charging/discharging power of EV can be continuously adjusted between 0 and a maximum rate, and they need to obtain the energy required for the desired journeys within the interval they are connected to the grid [33]. Without loss of generality, each EV is assumed to depart from its grid connection point only once during the temporal horizon of the coordination problem (at time period t_i^{dep}) and subsequently arrive back to its grid connection point only once during the same horizon (at time period t_i^{arr}). The operating model of the EV of prosumer i includes the following constraints:

$$E_{i,t}^{ev} = E_{i,t-1}^{ev} + \Delta t C_{i,t}^{ev} \eta_i^{evc} + \Delta t D_{i,t}^{ev} / \eta_i^{evd} - E_{i,t}^{tr} \quad \forall t \tag{1}$$

$$\underline{E}_i^{ev} \leq E_{i,t}^{ev} \leq \overline{E}_i^{ev} \quad \forall t \tag{2}$$

$$0 \leq C_{i,t}^{ev} \leq V_{i,t}^{ev} A_{i,t}^{ev} \overline{P}_i^{ev} \quad \forall t \tag{3}$$

$$-\left(1 - V_{i,t}^{ev}\right) A_{i,t}^{ev} \overline{P}_i^{ev} \leq D_{i,t}^{ev} \leq 0 \quad \forall t \tag{4}$$

$$E_{i,t_i^{dep}}^{ev} \geq \sum_t E_{i,t}^{tr} \tag{5}$$

Constraint (1) corresponds to the EV battery's energy balance, taking into account the energy needed for commuting purposes as well as the losses caused by charging and discharging efficiencies. Constraint (2) expresses the lower and upper bounds of the battery's energy content. Constraint (3),(4) represent the limits of the battery's charging / discharging power, which depend on its power capacity $\overline{P}_i^{ev}$ and on whether the EV is available to be scheduled ($A_{i,t}^{ev} = 1$) or not ($A_{i,t}^{ev} = 0$), while the binary variable $V_{i,t}^{ev}$ is employed to avoid simultaneous charging and discharging. Constraint (5) ensures that the EV is sufficiently charged upon departure to satisfy the commuting requirements of its users.

The operating constraints of the ES of prosumer i are similar to those of EV apart from the fact that the commuting energy requirement $E_{i,t}^{tr}$ and the scheduling availability $A_{i,t}^{ev}$ are irrelevant and thus are removed [34].

We define the power rate of EV and ES as $P_{i,t}^{ev} = C_{i,t}^{ev} - D_{i,t}^{ev}$ and $P_{i,t}^{es} = C_{i,t}^{es} - D_{i,t}^{es}$, which leads to the design of actions $a_{i,t}^{ev}$ and $a_{i,t}^{es}$ in the RL formulation (Section IV-B).

B. HVAC System

The operation of HVAC system involves transition from electricity power to thermal comfort. The flexibility of the HVAC system lies in the allowance of an indoor temperature range by the users, within which their thermal comfort is preserved:

$$\underline{H} \leq H_{i,t}^{in} \leq \overline{H}, \quad \forall t \tag{6}$$

Analogous to [35], the employed HVAC dynamic model of indoor temperature is mainly influenced by the outdoor temperature H_t^{out}, the indoor temperature $H_{i,t}^{in}$ and the power demand $P_{i,t}^{ac}$ of prosumer i at step t:

$$H_{i,t+1}^{in} = H_{i,t}^{in} - \frac{\left(H_{i,t}^{in} - H_t^{out} + \eta^{ac} R_i^{ac} P_{i,t}^{ac}\right)\Delta t}{C_i^{ac} R_i^{ac}} \quad \forall t \tag{7}$$

In order to keep the indoor temperature within the thermal comfort range in (6), the power demand $P_{i,t}^{ac}$ of HVAC system can be continuously adjusted, also conduces to the design of action $a_{i,t}^{ac}$ in the RL formulation (Section IV-B)

$$0 \leq P_{i,t}^{ac} \leq \overline{P}_i^{ac}, \quad \forall t. \tag{8}$$

C. Smart Appliance

The operation of SA involves the execution of user-prescribed cycles that encompass a series of phases occurring in a fixed sequence with generally fixed duration and fixed electrical power requirements, which are immutable and uninterruptible; their flexibility is driven by the functionality to defer this cycle up to a maximum delay limit defined by their users [36]. Each SA is assumed to be activated only once by its users within the horizon of the coordination problem. The operating constraints of the SA of prosumer i includes the following constraints:

$$z_{i,t}^{sa} \in \{0, 1\}, \quad \forall t \in \left[t_i^{in}, t_i^{ter} - T_i^{dur} + 1\right] \tag{9}$$

$$z_{i,t}^{sa} = 0, \quad \forall t < t_i^{in} \text{ and } \forall t > t_i^{ter} - T_i^{dur} + 1 \tag{10}$$

$$\sum_{t=t_i^{in}}^{t_i^{ter} - T_i^{dur} + 1} z_{i,t}^{sa} = 1 \tag{11}$$

$$P_{i,t}^{sa} = \sum_{\tau=1}^{T_i^{dur}} z_{i,t+1-\tau}^{sa} A_{i,t}^{sa} P_{i,\tau}^{cyc}, \quad \forall t \tag{12}$$

Constraints (9),(10) enforce that the demand activity of SA can only be executed over the time window determined by t_i^{in} and t_i^{ter}. $z_{i,t}^{sa}$ is the binary variable expressing whether the SA cycle is initiated at step t, which also conduces to the action design of SA in RL formulation (Section IV-B). Constraint (11) enforces that the demand activity of SA can be carried out (and therefore initiated) at most once during the coordination horizon. Constraint (12) expresses that the SA demand at each time step is dependent on the initiation time $z_{i,t}^{sa}$, $A_{i,t}^{sa}$, T_i^{dur}, and $P_{i,\tau}^{cyc}, \forall \tau \in [1, T_i^{dur}]$.

D. Model-Based System Centric Coordination Optimization

The net demand/generation $l_{i,t}$ of prosumer i can then be expressed as:

$$l_{i,t} = P^{nd}_{i,t} - P^{pv}_{i,t} + C^{ev}_{i,t} - D^{ev}_{i,t} + C^{es}_{i,t} - D^{es}_{i,t} + P^{ac}_{i,t} + P^{sa}_{i,t} \quad (13)$$

A theoretical, idealized case, which involves a model-based representation of the community's DER (assuming perfect knowledge of their operating models and techno-economic parameters as well as accurate predictions of their uncertain parameters) and a system-centric coordination approach (assuming that the P2P trading platform collects all techno-economic parameters of the DER and determines their optimal energy management decisions by centrally solving a global optimization problem) is employed as a theoretical optimality benchmark to the proposed MADRL approach.

This optimization problem can be formulated as a mixed-integer linear program as:

$$\min \sum_{t=1}^{T} \left(\lambda^b_t [P^{re}_t]^+ + \lambda^s_t [P^{re}_t]^- \right) \quad (14)$$

where:

$$P^{re}_t = \sum_{i=1}^{N} l_{i,t} \quad (15)$$

subject to:

$$(1) - (13) \quad (16)$$

$$-P^{thr} \leq P^{re}_t \leq P^{thr} \quad (17)$$

where operators $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$ indicate taking the maximum/minimum value between $\cdot$ and 0. The objective function (14) lies in minimizing the sum of prosumers' energy costs when the P2P trading platform trades with the supplier in the event of energy imbalance within the local community, with the first/second term representing the cost/revenue of buying/selling energy from/to the supplier. Furthermore, to avoid the overloading of the substation that supplies power to the community, constraint (17) is added to the above optimization problem, explicitly restricting the net demand/generation of the community P^{re}_t to be lower (in absolute terms) or equal to a specified threshold P^{thr} (reflecting the capacity of the substation to which the prosumers are connected), the same constraint is employed in [8], [10], [25].

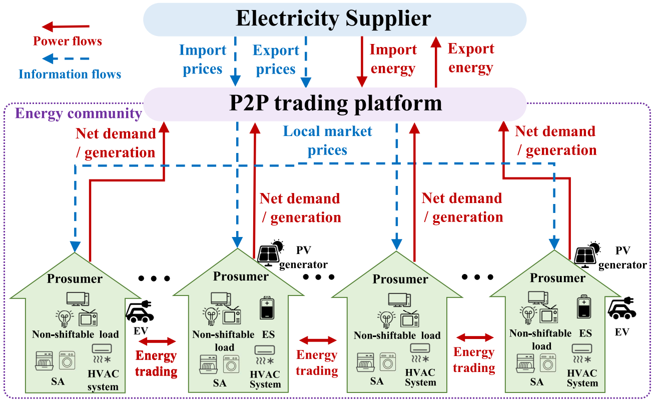

Fig. 1. Illustration of examined multi-agent coordination problem.

III. Peer-to-Peer Energy Trading Mechanism

We examine a local energy market with a large number of participating prosumers, each equipping a EMS agent to optimize its energy management decisions based on the available information, including the techno-economic characteristics of its DERs and local market prices. As shown in Fig. 1, each prosumer first trades energy directly with other prosumers within the community at the local market prices, in order to weaken their exposure to the supplier price differential (higher import prices and lower export prices, as discussed in Section I-A) and thus reduce their costs. Then, the whole community trades the energy deficit (excess demand) or surplus (excess generation) at each time step with the supplier at its offered import/export prices. A P2P trading platform is employed to manage the local energy trading and set the local market prices, as well as trading with the supplier any energy surplus or deficit that are not purchased or supplied in the community.

A. Mid-Market Rate Pricing

In the presence of the differentiated import/export prices offered by the supplier, instead of independently managing their DER, prosumers are encouraged to exchange energy directly with each other within the community in order to reduce their costs. Promoting such local energy trading and bypassing the incumbent supplier, however, is a challenging task that relies on a suitable P2P trading mechanism, providing prosumers with suitable monetary benefits to participate.

To this end, the considered P2P trading platform adopts the MMR pricing mechanism [9], [19] to suitably incentivize prosumers to participate in P2P trading, as demonstrated in [9]. Consider a set of prosumers $\mathcal{N} = \{1, \ldots, N\}$ and define the net demand P^{nc}_t and net generation P^{ng}_t of the community as functions of individual prosumers' net load $l_{i,t}$.

$$P^{nc}_t = \sum_{i \in \mathcal{N}^c} l_{i,t}, \, P^{ng}_t = \sum_{i \in \mathcal{N}^g} \mathcal{N}^g l_{i,t} \quad (18)$$

where $\mathcal{N}^c = \{i \in \mathcal{N} : l_{i,t} > 0\}$ and $\mathcal{N}^g = \{i \in \mathcal{N} : l_{i,t} \leq 0\}$ are sets of consumers and producers at step t, respectively.

The MMR method sets the local buy $\lambda^{L,b}$ and sell $\lambda^{L,s}$ prices as the average of the supplier import and export prices $\bar{\lambda}_t$ with some adjustments based on the difference between the net demand and generation of the community. Concretely, at time step t.

i) If the community's total demand matches generation (i.e., $P^{re}_t = 0$), then the local buy and sell prices are set as:

$$\lambda^{L,b} = \lambda^{L,s} = \bar{\lambda}_t = \left(\lambda^b_t + \lambda^s_t \right) / 2 \quad (19)$$

ii) If the community's total demand exceeds the total generation (i.e., $P^{re}_t > 0$), the energy deficit is bought from the supplier at its import price λ^b_t. Since $\lambda^b_t > \bar{\lambda}_t$, an extra payment is made and proportionally shared among consumers according to their net demand $l_{i,t}$. In this case, the producers are paid at $\bar{\lambda}_t$ whereas the consumers are charged at a higher local buy price as expressed by:

$$\lambda_t^{L,b} = \left(\bar{\lambda}_t |P_t^{ng}| + \lambda_t^b P_t^{re}\right) / P_t^{nc}. \tag{20}$$

iii) If the community's total generation exceeds the total demand (i.e., $P_t^{re} \leq 0$), the energy surplus is sold to the supplier at its export price λ_t^s. Since $\lambda_t^s < \lambda_t^b$, a revenue shortfall emerges and is proportionally shared among producers according to their net generation $|l_{i,t}|$. In this case, the consumers are charged at $\bar{\lambda}_t$ whereas the producers are paid at a lower local sell price as expressed by:

$$\lambda_t^{L,s} = \left(\bar{\lambda}_t P_t^{nc} + \lambda_t^s |P_t^{re}|\right) / |P_t^{ng}| \tag{21}$$

Under MMR pricing, the cost of prosumer i at step t can be calculated on the basis of the local buy/sell prices as:

$$Cost_{i,t} = -\left(\lambda_t^{L,b}\left[l_{i,t}\right]^+ + \lambda_t^{L,s}\left[l_{i,t}\right]^-\right). \tag{22}$$

B. Mitigating Rebound Demand Peaks

Although MMR pricing promotes efficient coordination of energy sharing activities among prosumers, the selfish behavior of the latter in minimizing their individual costs, may result in synchronization of flexible demand/generation schedules at the lowest-/highest-priced periods and consequently to significant new (rebound) demand/generation peaks, jeopardizing the security of the local distribution network (Section I-B).

To address this challenge, the P2P trading platform penalizes the contribution of prosumers to rebound peaks with the aim to reschedule the flexible DERs of prosumers such that the network capacity threshold P^{thr} is not violated at each time step, i.e., constraint (16) is satisfied.

To this end, we define $l_{i,t}^{flex}$ as the net demand (positive) / generation (negative) of the flexible DER (EV, ES, HVAC and SA) of prosumer i at step t:

$$l_{i,t}^{flex} = C_{i,t}^{ev} - D_{i,t}^{ev} + C_{i,t}^{es} - D_{i,t}^{es} + P_{i,t}^{ac} + P_{i,t}^{sa} \tag{23}$$

If the net demand of the community is higher than the specified threshold, we introduce a penalty term $Pen_{i,t}$ to each prosumer that penalizes its contribution to the community's flexible demand. Analogously, if the net generation of the community is higher than the threshold, we penalize each prosumer's contribution to the community's flexible generation:

$$Pen_{i,t} = \begin{cases} -\omega_1\left(l_{i,t}^{flex} / \sum_{i \in \mathcal{N}^{fc}} l_{i,t}^{flex}\right), \quad \forall i \in \mathcal{N}^{fc} & \text{if } P_t^{re} > P^{thr} \\ -\omega_1\left(l_{i,t}^{flex} / \sum_{i \in \mathcal{N}^{fg}} l_{i,t}^{flex}\right), \quad \forall i \in \mathcal{N}^{fg} & \text{if } P_t^{re} < -P^{thr} \\ 0, \quad \forall i \in \mathcal{N} & \text{otherwise} \end{cases} \tag{24}$$

where $\mathcal{N}^{fc} = \{i \in \mathcal{N} : l_{i,t}^{flex} > 0\}$ and $\mathcal{N}^{fg} = \{i \in \mathcal{N} : l_{i,t}^{flex} \leq 0\}$ and ω_1 is the weighting factor.

IV. Markov Game Formulation of the Multi-Agent Coordination Problem

It can be observed that the coordinated energy management and P2P trading problem of the energy community (Fig. 1) resembles a decision-making process that involves multiple agents. In this multi-agent scenario, agents need to take into account and interact with not only the environment (i.e., the local energy community) but also other learning agents. This type of multi-agent coordination problem is usually modeled through a *Markov Game* [37] or *Stochastic Game* [38], a similar practice is also adopted in [25], [31], [32].

Specifically, a Market Game can be described by a state space, $\mathcal{S}$, a collection of action spaces for the N agents, $\{\mathcal{A}_{1:N}\}$, a state transition function $\mathcal{T}$, which defines the probability distribution over possible next states, given the current state and actions for each agent, and a reward function for each agent $\mathcal{R}_i$. Note that the *private observation* o_i of each agent i contains only partial information from the global state, $s \in \mathcal{S}$, and therefore the game is referred to as a *Partially Observable Markov Game* (POMG). The objective of each agent lies in learning a policy $\pi_i : \mathcal{O}_i \rightarrow P(\mathcal{A}_i)$ (which maps each agent's observation to a distribution over its set of actions) that maximizes its expected discounted return (or the performance function), $J_i(\pi_i) = \mathbb{E}_{a_i \sim \pi_i, \forall i, s \sim \mathcal{T}}[\sum_{t=0}^{T} \gamma^t r_{i,t}(s_t, a_{1:N,t})]$, where $\gamma \in [0, 1)$ is the discount factor and T is the time horizon.

The main elements associated with the POMG formulation of the examined coordinated energy management and P2P trading problem are outlined as follows.

A. Observation

The observation $o_{i,t}$ for each agent i at time step t is composed of common features including the time step identifier t, the supplier import λ_t^b and export λ_t^s prices and features that are dependent to agent i's portfolio of DERs, these may include PV generation $P_{i,t}^{pv}$ and non-shiftable demand $P_{i,t}^{nd}$. Note that we propose to incorporate the past 24-hour profiles (with 48 half-hourly time steps) for the prices, non-shiftable load and PV in the agent's observation in order to infer their future trends, and thereby fostering more cost-effective energy management decisions to be devised. To this effect, the *long short term memory* network [39] is employed which takes as input these past 24-hour profiles and extracts discriminative features containing information concerning their future trends.

According to the mathematical models of the controllable DERs (Section II), for proumsers with EV and ES, their observation also include the energy content of EV $E_{i,t}^{es}$ and ES $E_{i,t}^{ev}$ and the EV scheduling availability indicator $A_{i,t}^{ev}$ for time step t. For proumsers with HVAC system, their observation also include the outdoor temperature H_t^{out} and their indoor temperature $H_{i,t}^{in}$. For proumsers with SA, their observation also include the SA scheduling availability indicator $A_{i,t}^{sa}$ for step t. For prosumers with all examined DERs (PV, non-shiftable load, EV, ES, SA and HVAC), their observation $o_{i,t}$ can be expressed as:

$$o_{i,t} = \Big[t, \lambda_{t-47:t}^b, \lambda_{t-47:t}^s, P_{i,t-47:t}^{pv}, P_{i,t-47:t}^{nd}, E_{i,t}^{ev}, E_{i,t}^{es}, A_{i,t}^{ev}, H_t^{out}, H_{i,t}^{in}, A_{i,t}^{sa}\Big] \tag{25}$$

Taking local observations of all agents at step t into account, we have $o_t = (o_{1,t}, o_{2,t}, \ldots, o_{N,t})$, for simplicity, the global state s_t is set to be o_t.

B. Action

The action $a_{i,t}$ of each agent i at step t consists of its employed control decisions for managing its controllable DERs including EV, ES, HVAC and SA. Depending on the managed DERs, agents' actions can be distinguished as continuous and discrete.

Continuous actions are employed for managing DERs featuring continuously adjustable power demand/generation, namely EV, ES, and HVAC. Accordingly, actions $a_{i,t}^{ev}$ and $a_{i,t}^{es} \in [-1, 1]$ are used to represent the size of the charging (positive) and discharging (negative) power of EV and ES as a ratio of $A_{i,t}^{ev}\overline{P}_i^{ev}$ and $\overline{P}_i^{es}$; action $a_{i,t}^{ac} \in [0, 1]$ is used to represent the size of the power demand of HVAC as a percentage of $\overline{P}_i^{ac}$. For continuous actions, the policy $\pi_i(a_{i,t}|o_{i,t})$ is usually approximated by a Gaussian distribution (referring as a Gaussian policy) [40] $\mathcal{N}(\mu(o_{i,t}), \sigma^2)$, where $\mu(o_{i,t})$ and σ^2 are the mean and standard deviation for the actions of EV, ES and HVAC, respectively.

Discrete (binary) actions are employed for managing DERs featuring deferrable but non-interruptible power demand, namely SA. Accordingly, action $a_{i,t}^{sa} = z_{i,t}^{sa} \in \{0, 1\}$ represents whether the SA cycle is initiated or not at step t, it also satisfies $a_{i,t}^{sa} = 0$, if $\sum_{t'=t_i^{in}}^{t} a_{i,t'}^{sa} = 1, \forall t$, which enforces that the demand activity of SA can be initiated at most once during the daily coordination horizon (constraint (11)). For binary actions, the policy $\pi_i(a_{i,t}|o_{i,t})$ naturally can be approximated by a Bernoulli distribution $\mathcal{B}(p(o_{i,t}))$ (which is utilized to represent the probability mass function of a binary random variable), where $p(o_{i,t})$ represents the probability of initiating the operating cycle of SA.

For prosumers with all examined controllable DERs, their action $a_{i,t}$ can be expressed as:

$$a_{i,t} = \left[a_{i,t}^{ev}, a_{i,t}^{es}, a_{i,t}^{ac}, a_{i,t}^{sa}\right]. \tag{26}$$

C. State Transition

After the execution of actions $a_{i:N,t}$, the community environment maps the actions to the energy management decisions of the DERs and the net load/generation of each prosumer, and subsequently determines the next state and reward.

The transition is not only affected by all agents' actions but also by the randomness that characterizes some state features. In the examined problem, the transitions for the *exogenous features* (i.e., the ones characterized by inherent uncertainty and variability and are not affected by agents' actions) included in $[\lambda_t^b, \lambda_t^s, P_{i,t}^{nd}, P_{i,t}^{pv}, A_{i,t}^{ev}, H_t^{out}, A_{i,t}^{ev}]$ are subject to the variability and uncertainty of the supplier's pricing strategies, end-user's electricity usage behavior and weather conditions. In this regard, it presents significant challenges to identify suitable probabilistic models which can fully capture such randomness. RL remedies this problem in a model-free fashion which does not rely on accurate mathematical modeling of the underlying uncertainties. Alternatively, MARL resorts to machine learning techniques to learn the state transitions of these exogenous features from real system data. In this setting, the values of exogenous features $[\lambda_{t+1}^b, \lambda_{t+1}^s, P_{i,t+1}^{nd}, P_{i,t+1}^{pv}, A_{i,t+1}^{ev}, H_{t+1}^{out}, A_{i,t+1}^{ev}]$ will be directly taken from the data-set indexed at $t + 1$. On the other hand, the state transitions for endogenous features (i.e., the ones that are directly affected by agent's actions) $E_{i,t}^{ev}$, $E_{i,t}^{es}$ and $H_{i,t}^{in}$ will be determined deterministically by the actions taken at step t. Mutually exclusive quantities $C_{i,t}^{ev}$ and $D_{i,t}^{ev}$ (as the charging and discharging activity of EV cannot occur simultaneously at a given step) are managed by action $a_{i,t}^{ev}$, and are also limited by the energy content $E_{i,t}^{ev}$ and the EV's operating parameters: $A_{i,t}^{ev}$, $\underline{E}_i^{ev}$, $\overline{E}_i^{ev}$, η_i^{evc}, and η_i^{evd}.

$$C_{i,t}^{ev} = \min\left(a_{i,t}^{ev}A_{i,t}^{ev}\overline{P}_i^{ev}, (\overline{E}_i^{ev} - E_{i,t}^{ev})/(\eta_i^{evc}\Delta t)\right) \tag{27}$$

$$D_{i,t}^{ev} = \min\left(-a_{i,t}^{ev}A_{i,t}^{ev}\overline{P}_i^{ev}, (E_{i,t}^{ev} - \underline{E}_i^{ev})\eta_i^{evd}/\Delta t\right) \tag{28}$$

Equations (27),(28) ensure that the variables $C_{i,t}^{ev}$, $D_{i,t}^{ev}$ and $E_{i,t}^{ev}$ satisfy the operational constraints of EV (i.e., energy balance and maximum energy and power limits given in (1)-(4)). Similar measure to ensure feasibility or safety of actions are commonly adopted in RL-based energy management literature [24], [41]. Based on $C_{i,t}^{ev}$ and $D_{i,t}^{ev}$, and according to the energy balance constraint (1) of EV, the transition of $E_{i,t}^{ev}$ can be expressed as:

$$E_{i,t+1}^{ev} = E_{i,t}^{ev} + C_{i,t}^{ev}\eta_i^{evc}\Delta t + D_{i,t}^{ev}\Delta t/\eta_i^{evd} - E_{i,t}^{tr} \tag{29}$$

The derivation of $C_{i,t}^{es}, D_{i,t}^{es}$, and $E_{i,t+1}^{es}$ follows the analogous logic in (27)-(29) except that the scheduling availability and the traveling energy requirement are not considered. Furthermore, based on the HVAC model in Section II-B, the state transition of $H_{i,t+1}^{in}$ of the HVAC based on $P_{i,t}^{hvac}$ can be expressed as (30).

$$H_{i,t+1}^{in} = H_{i,t}^{in} - \frac{\left(H_{i,t}^{in} - H_t^{out} + \eta_i^{ac}R_i^{ac}a_{i,t}^{ac}\overline{P}_i^{ac}\right)\Delta t}{C_i^{ac}R_i^{ac}} \tag{30}$$

The power demand of the SA $P_{i,t}^{sa}$ can be expressed through (12) after substituting $z_{i,t}^{sa}$ with $a_{i,t}^{sa}$ as:

$$P_{i,t}^{sa} = \sum_{\tau=1}^{T_i^{dur}} a_{i,t+1-\tau}^{sa}A_{i,t}^{sa}P_{i,\tau}^{cyc} \tag{31}$$

Finally, the net load/generation of prosumer i can be calculated using (13).

D. Reward

The basic reward $r_{i,t}^{cost}, r_{i,t}^{comf}$ of agent i at time step t is set as to minimize the energy cost $Cost_{i,t}$ in (22) while ensuring the indoor temperature vary within a comfortable range:

$$r_{i,t}^{cost} = -Cost_{i,t} = -\left(\lambda_t^{L,b}\left[l_{i,t}\right]^+ + \lambda_t^{L,s}\left[l_{i,t}\right]^-\right) \tag{32}$$

$$r_{i,t}^{comf} = -\omega_2\left(\left[H_{i,t}^{in} - H^{max}\right]^+ + \left[H^{min} - H_{i,t}^{in}\right]^+\right) \tag{33}$$

where ω_2 is a penalty factor that decides on the relative importance of maintaining the thermal comfort with respect to the energy cost.

Furthermore, note that in (27),(28), the charging/discharging power of EV only respects the minimum/maximum power/energy limits of the EV, but does not ensure that the EV is sufficiently charged upon departure, i.e., constraint (5) may not be satisfied. To properly account for the time-coupling constraint of EV, a penalty term $r_{i,t}^{evcons}$ is superimposed on the basic reward function, which penalizes the extent of constraint violation with a penalty factor ω_3. Similar measures to ensure such constraint satisfaction of EV are adopted in [27], [33].

$$r_{i,t}^{evcons} = -\omega_3 \left[E_{i,t^{dep}}^{ev} - \sum_{t} E_{i,t}^{tr} \right]^{+} \text{ if } t = t_i^{dep} \tag{34}$$

Finally, as discussed in Section II-D, model-based system-centric coordination approaches safeguard the security of local distribution network by imposing relevant global constraints [14] (this is constraint (17) in the context of our paper). However, in the examined model-free MADRL framework, constraints coupling actions dimensions cannot be explicitly imposed [42]. In order to tackle this challenge, we propose a novel reward shaping mechanism which implicitly satisfies the above constraint by penalizing each prosumer's contribution to rebound peaks. To this effect, we introduce a penalty term $r_{i,t}^{thr} = Pen_{i,t}$ in each agent's reward function. To the best of the authors' knowledge, our work is the first one considering and addressing (through this reward shaping mechanism) distribution network congestion and rebound peak effects in a model-free RL framework, as none of the relevant MADRL papers [25]–[27], [30]–[32] does so.

The final reward function $r_{i,t}$ of each agent i at step t can be expressed as:

$$r_{i,t} = r_{i,t}^{cost} + r_{i,t}^{thr} + r_{i,t}^{comf} + r_{i}^{evcons} \tag{35}$$

As can be observed in equations (27)-(35), the community environment implements the mapping between the agents' energy management actions $a_{i,t}^{ev}$, $a_{i,t}^{es}$, $a_{i,t}^{ac}$, $a_{i,t}^{sa}$ to the power demand/generation of the respective DERs $C_{i,t}^{ev}$, $D_{i,t}^{ev}$, $C_{i,t}^{es}$, $D_{i,t}^{es}$, $P_{i,t}^{ac}$, $P_{i,t}^{sa}$, and subsequently determines the state transitions of $E_{i,t}^{ev}$, $E_{i,t}^{es}$ and $H_{i,t}^{in}$, and the rewards for all the agents.

V. Multi-Agent Reinforcement Learning Framework

Before introducing the proposed methodology, preliminaries of single-agent RL (SARL) and MARL are first presented in this section, followed by a comprehensive review on the state-of-the-art MARL frameworks.

A. Preliminaries of SARL and MARL

In the single-agent setting, SARL involves an agent acts in an environment by sequentially taking actions over a sequence of time steps, in order to maximize a cumulative reward. In general, SARL can be described as a *Markov Decision Process* (MDP) which includes: 1) a state space $\mathcal{S}$; 2) an action space $\mathcal{A}$; 3) a transition probability function $p(s_{t+1}|s_t, a_t)$; and 4) a reward function r: $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The decision as to which action a_t is chosen in a certain state s_t is governed by a policy $\pi(a_t|s_t)$. The agent employs its policy to interact with the MDP. The return $R_t = \sum_{l=t}^{T} \gamma^{(l-t)} r_l$ is the discounted reward where $\gamma \in [0, 1]$ is the discount factor. The agents' goal through SARL is to construct a policy that maximizes the cumulative discounted reward from the start state s_1, denoted by the *performance function* $J(\pi) = \mathbb{E}[R_1|\pi] = \mathbb{E}_{s \sim \rho^{\pi}, a \sim \pi}[r]$, where ρ^{π} denotes the discounted state distribution.

TABLE I
Summary of MARL Frameworks

Approach Feature	**Concurrent**	**Centralized**	**MADDPG**	**MAAC**
Training	Decentralized	Centralized	Centralized	Centralized
Execution	Decentralized	Centralized	Decentralized	Decentralized
Non-stationarity	Yes	No	No	No
Privacy preserving	Yes	No	No	Yes
Scalability	No	No	No	Yes

In the multi-agent setting, similar to the single-agent setting, each agent is still trying to solve the sequential decision-making problem from its repeated interaction with the environment. The difference is that the evolution of the environmental state and the reward function is now influenced by all agents' joint actions. As a result, agents need to take into account and interact with not only the environment but also other learning agents. Such a decision-making process involving the coordination of multiple agents is usually modeled through a POMG, as introduced in Section IV.

B. State-of-the-Art MARL Frameworks

As discussed in Section I-B, for the coordinated energy management and P2P trading problem of the energy community, prior works have identified three majors MARL frameworks, including Centralized, Concurrent, and CTDE approaches [43], the main features of which are summarized in Table I.

Concurrent learning: At one end of the spectrum, Concurrent learning trains each agent independently to maximize their individual reward, while regarding other agents as part of the environment it interacts. However, as a large number of agents are learning and adapting their policies individually, the frequent change in these policies renders the environment dynamics non-stationary, contributing to instability or even divergence. Furthermore, since each agent needs to train its own policy, significant computational and memory burdens arise when the policies are represented by complex models like DNNs. Finally, because the agents do not share experiences, this approach is often criticized by its low sampling and learning efficiency.

Centralized learning: At the other end of the spectrum, centralized learning involves modeling all agents collectively as a single agent whose action space is the concatenation of all agents' action spaces. This approach exhibits poor scalability since the dimension of the action space increases exponentially with the number of agents, which quickly becomes intractable in real-world applications. Additionally, it exhibits significant communication requirement during execution, as the computation of the central policy necessitates input of all observations from and distribute actions to the individual agents. Furthermore, the implementation of this approach violates prosumers' privacy since they are generally unwilling to expose their private information and directly exchange such information with others.

Multi-agent Deep Deterministic Policy Gradient (MADDPG): MADDPG constitutes the most commonly employed algorithm within the CTDE framework [44], which attempts to combine the strengths of the concurrent and centralized MARL approaches. Concretely, MADDPG involves learning a number of critics centrally with all agents' private information as input. The actors, however, receive private observations only from their corresponding agents. Namely, during testing, an agent's policy execution is conducted without the knowledge of other agents' information. This framework effectively circumvents the challenge of environmental non-stationary during learning. However, as in centralized approach, MADDPG is not privacy preserving and suffers from an analogous curse of dimensionality in training the central critics, which is problematic in practical large-scale multi-agent.

C. Multi-Actor-Attention-Critic (MAAC) Method

As discussed above, state-of-the-art MARL approaches are still far from being scalable to large-scale multi-agent systems and largely neglect privacy concerns of the prosumers. However, both properties are deemed imperative in successful establishment of a P2P transactive trading mechanism. To this end, a tailor-made approach, namely the *multi-actor-attention-critic* (MAAC) method [45], is adopted to tackle these challenges. MAAC also belongs to the CTDE MADRL paradigm, but as opposite to MADDPG which indifferently incorporates the physical private information (observations and actions) of all agents to train their critics, MAAC enables learning of all agents' critics jointly by sharing a set of learnable parameters among agents. Furthermore, the employment of the *attention mechanism* [46] enables selectively paying attention to the relevant information (an abstract embedding of agents' private information) of other agents at each step during training. As such, MAAC offers significantly greater scalability, lower computational complexity, and is able to protect the privacy of prosumers compared to the state-of-the-art MARL approaches.

1) SAC Method: Since the MAAC method constitutes a multi-agent extension of the *soft actor-critic* (SAC) method, the latter is briefly introduced in this subsection. Interested readers are referred to [47] for a more detailed introduction.

SAC falls within the category of *actor-critic* SARL algorithms [42], it employs a parameterized critic network Q^{θ}, taking as input a state s and action a and outputting a scalar estimate of the Q-value function $Q^{\theta}(s,a)$. The parameterized actor network π^{ϕ} takes as input a state s and outputs the action selection probabilities $\pi^{\phi}(a|s)$ at state s.

To maximize the performance function, i.e., $J(\pi^{\phi})$, the policy parameters ϕ are updated by taking steps in the direction of the *performance gradient* $\nabla_{\phi}J(\pi^{\phi})$ which takes the following form [40]:

$$\nabla_{\phi}J\left(\pi^{\phi}\right)=\mathbb{E}_{s\sim\mathcal{B},a\sim\pi}\left[\nabla_{\phi}\log\left(\pi^{\phi}(a|s)\right)Q^{\theta}(s,a)\right]. \quad (36)$$

where $\mathcal{B}$ denotes the experience replay buffer [22] which stores the past experiences $(s,a,r,\tilde{s})$.

The critic network implements policy evaluation, criticizing the policy by producing a Q-value function estimate.

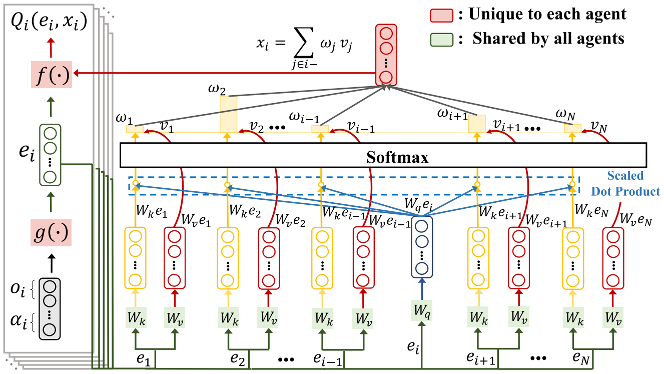

Fig. 2. Q-value function evaluation under MAAC.

This is achieved using *temporal difference* (TD) learning by minimizing the following mean-squared TD error:

$$L(\theta)=\mathbb{E}_{(s,a,r,\tilde{s})\sim\mathcal{B}}\left[\left(Q^{\theta}(s,a)-y\right)^{2}\right] \quad (37)$$

where $y = r(s,a) + \gamma\mathbb{E}_{\tilde{a}\sim\pi(\tilde{s})}[Q^{\bar{\theta}}(\tilde{s},\tilde{a})]$ and $Q^{\bar{\theta}}$ is the parameterized target Q-value function.

To foster exploration and also evade premature converging to non-optimal deterministic policies (adopted in DDPG for example), the performance gradient (36) is modified to incorporate an *entropy* term:

$$\nabla_{\phi}J\left(\pi^{\phi}\right)=\mathbb{E}_{s\sim\mathcal{B},a\sim\pi}\left[\nabla_{\phi}\log\left(\pi^{\phi}(a|s)\right)\xi(s,a)\right]. \quad (38)$$

where $\xi(s,a)=-\beta\log(\pi^{\phi}(a|s))+Q^{\theta}(s,a)-z(s)$, β is the temperature parameter which governs the balance between maximizing entropy and reward, and $z(s)$ is a state-dependent baseline for $Q^{\theta}(s,a)$. Subsequently, the target value y employed in the loss function (37) can be restated as:

$$y=r(s,a)+\gamma\mathbb{E}_{\tilde{a}\sim\pi(\tilde{s})}\left[Q^{\bar{\theta}}(\tilde{s},\tilde{a})-\beta\log\left(\pi^{\bar{\phi}}(\tilde{a}|\tilde{s})\right)\right] \quad (39)$$

where $\pi^{\bar{\phi}}$ is the target policy function parameterized by $\bar{\phi}$.

2) Working Principal of MAAC: Extending SAC in a multi-agent setting, the MAAC method inherits from the CTDE paradigm which trains critics centrally and executes the learned policies distributedly. Additionally, the attention mechanism is employed which allows each agent to query other agents regarding information about their observations and actions, and then selectively incorporates relevant information of other agents in the said agent's Q-value estimate.

In MADDPG, the critic takes as input the observations, $o=(o_1,\ldots,o_N)$, and actions, $a=(a_1,\ldots,a_N)$ of all agents to calculate the Q-value of agent i, $Q_i(o,a)$. It can be observed that the critic's input dimension increases exponentially with the action/observation dimension of each agent, as well as the number of agents, soon rendering the problem intractable. Furthermore, o and a comprise physical quantities pertaining to the operation of all agents' DERs; however, such data can expose prosumers' daily habits and their types of DERs, leading to violation of their privacy. To address this, P2P trading platform acts as a trusted third party which provides prosumers with information that reflect the collective behavior of other prosumers in the community during centralized training.

Fig. 2 illustrates the estimation of $Q_i^{\theta}(o,a)$ under MAAC. It can be observed that the Q-value estimation only takes input agent i's local observation o_i and action a_i, and other agents' *contributions*, x_i, which is an abstract, learned representation of other agents' local information. Considering x_i in the Q-value function estimate, each agent can make informed decisions based on the influence of the collective behavior of other agents in the community environment, but without knowing their specific local information, which therefore achieves privacy preservation. Furthermore, it can be observed that the input dimension of the critic is significantly compressed from $N \cdot (|\mathcal{O}_i| + |\mathcal{A}_i|)$ (if MADDPG is employed and assuming all agents have the same dimension for o_i and a_i) to the dimension of the concatenation of embedding e_i and x_i, avoiding the action/observation space explosion, and substantially improving scalability for large-scale multi-agent applications.

$$Q_i^{\theta}(o, a) = f_i(g_i(o_i, a_i), x_i) \tag{40}$$

where f_i is a two-layer multi-layer perceptron (MLP) and $g_i(o_i, a_i) = e_i$ is a one-layer MLP embedding function. Note that f_i and g_i are functions with learnable parameters that are private to each agent and inaccessible to others. The total contribution from other agents, x_i, is evaluated as a weighted sum of each agent's *value* v_i as:

$$x_i = \sum_{j \in i-} \omega_j v_j = \sum_{j \in i-} \omega_j h\big(W_v e_j\big) \tag{41}$$

where the set of all agents other than i is denoted as $i-$ and it is indexed with j, W_v is a shared matrix of all agents which transforms an agent's embedding e_j into a *value*, h is a nonlinear activation function, and ω_j denotes the *attention weight* assigned to each agent j's value (i.e., the extent of attention that agent i paid to the value of agent j). It is obtained by comparing the similarity between agent i's embedding e_i with other agents' embeddings e_j, and then passes the similarity value to a *softmax* operator:

$$\omega_j = \exp^{(W_k e_j)^T W_q e_i} \Big/ \sum_{j=1}^{N} \exp^{(W_k e_j)^T W_q e_i} \tag{42}$$

where W_k and W_q are shared matrices of all agents that transform e_j/e_i into a *key*/*query*, respectively.

The sharing of the three trainable parameter matrices W_v, W_k, and W_q allows the critics to be updated jointly by minimizing a single loss function as expressed by:

$$L(\theta) = \sum_{i=1}^{N} \mathbb{E}_{(o,a,r,\tilde{o}) \sim \mathcal{B}} \Big[\big(Q_i^{\theta}(o, a) - y_i\big)^2 \Big] \tag{43}$$

where $y_i = r_i(o, a) + \gamma \mathbb{E}_{\tilde{a} \sim \pi^{\bar{\phi}}(\tilde{o})}[Q_i^{\bar{\theta}}(\tilde{o}, \tilde{a}) - \beta \log(\pi^{\bar{\phi}_i}(\tilde{a}_i|\tilde{o}_i))]$. On the other hand, each agent i updates the weights π^{ϕ_i} of its own actor network using the performance gradient:

$$\nabla_{\phi_i} J\big(\pi^{\phi_i}\big) = \mathbb{E}_{o \sim \mathcal{B}, a \sim \pi} \big[\nabla_{\phi_i} \log\big(\pi^{\phi_i}(a_i|o_i)\big) \xi_i(o_i, a_i) \big] \tag{44}$$

where $\xi_i(o_i, a_i) = -\beta \log(\pi^{\phi_i}(a_i|o_i)) + Q_i^{\theta}(o, a) - z(o, a_{i-})$, $z(o, a_{i-}) = \mathbb{E}_{a_i \sim \pi^{\phi_i}(o_i)}[Q_i^{\theta}(o, (a_i, a_{i-}))]$, and $Q_i^{\theta}(o, a) - z(o, a_{i-})$ signifies the multi-agent *advantage function* which compares the value of a specific action a_i to the mean value averaged over all actions of agent i (treating all other agents' actions fixed). It therefore indicates whether the current action a_i will cause an increase in expected return.

Algorithm 1 Training Phase of MAAC

```
Initialize the community environment with $N$ agents
for episode = 1 : $E$ do
    Reset environment to get initial $o_i$ for each agent $i$
    for time step = 1 : $T$ do
        Choose an action $a_i \sim \pi^{\phi_i}(\cdot|o_i)$ for each agent $i$
        Execute actions $a_{1:N}$ in the environment and get $\tilde{o}_i$ and $r_i$ for all agents
        Deposit experiences $(o_{1:N}, a_{1:N}, \tilde{o}_{1:N}, r_{1:N})$ in $\mathcal{B}$
        if $N_B \geq N_L$ and $\mathrm{mod}(t, T^u) = 0$ then
            Sample a minibatch of experiences $\{(o^l_{1:N}, a^l_{1:N}, r^l_{1:N}, \bar{o}^l_{1:N})\}_{l=1}^{N_L}$ from $\mathcal{B}$
            For each experience $l$, evaluate $Q_i^{\theta}(o^l_{1:N}, a^l_{1:N})$, $Q_i^{\bar{\theta}}(\tilde{o}^l_{1:N}, \tilde{a}^l_{1:N})$, $a^l_i \sim \pi^{\phi_i}(\tilde{o}^l_i)$, and $\tilde{a}^l_i \sim \pi^{\bar{\phi}_i}(\tilde{o}^l_i)$ for all $i$ in parallel
            Update critic using $\nabla_{\theta} L(\theta)$ in (43)
            Update actor using $\nabla_{\phi_i} J(\pi^{\phi_i})$ in (44)
            Update target network weights as:
            $\bar{\theta} \leftarrow \tau\bar{\theta} + (1 - \tau)\theta$ and $\bar{\phi}_i \leftarrow \tau\bar{\phi}_i + (1 - \tau)\phi_i, \forall i$
        end if
    end for
end for
```

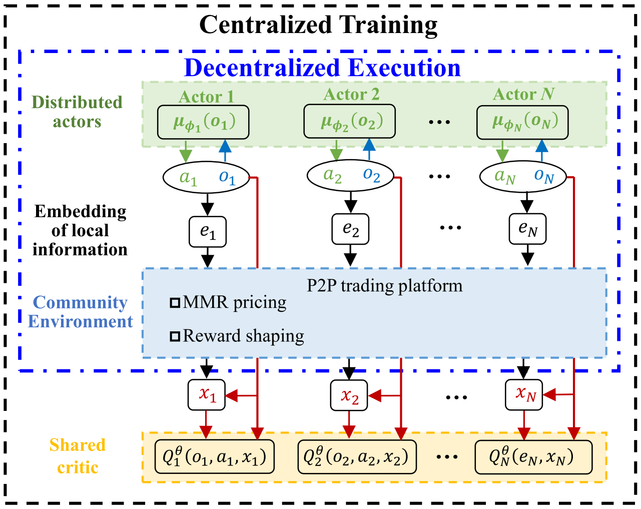

Fig. 3. Training and execution architecture of MAAC.

3) Training and Execution of MAAC: The training and execution architecture of the MAAC method is depicted in Fig. 3. Algorithm 1 outlines the training procedure of agents' actor and critic networks employed by MAAC. As illustrated in Fig. 3, during training, each agent i communicates its embedded information e_i to the P2P trading platform, the latter then calculates and informs agent i the contributions of all other agents, x_i, in their estimations of Q-values (equation (36)). It can be observed that the agents communicate with each other implicitly through the P2P trading platform and without directly exchanging private information among them. Since the centralized training process is performed in an offline fashion, communication failure is not a relevant concern.

After the training process, during the execution phase, the critic network is no longer required (Fig. 3), and the weights of the actor networks are fixed. During this phase, the trained

Algorithm 2 Execution Phase of MAAC

```
Load the neural network parameters $\phi_i^*$ of the online actor networks $\pi^{\phi_i^*}$ trained by Algorithm 1.
for episode = 1 : $E$ do
    Reset environment to get initial $o_i$ for each agent $i$
    for time step = 1 : $T$ do
        Select action $a_i \sim \pi^{\phi_i^*}(o_i)$ for each agent $i$
        Execute actions $a_{1:N}$ in the environment and get $\tilde{o}_i$ and $r_i$ for all agents
    end for
end for
```

TABLE II
COMPUTATIONAL COMPLEXITY OF INVESTIGATED COORDINATION METHODS

Method	Number of DNNs	Actor input / output dimensions	Critic input / output dimensions
U-SAC	$2N$	$\lvert o_i\rvert$ / $\lvert a_i\rvert$	$\lvert o_i + a_i\rvert$ / 1
Cen-SAC	2	$N\lvert o_i\rvert$ / $N\lvert a_i\rvert$	$N\lvert o_i + a_i\rvert$ / 1
Con-SAC	$2N$	$\lvert o_i\rvert$ / $\lvert a_i\rvert$	$\lvert o_i + a_i\rvert$ / 1
MADDPG	$2N$	$\lvert o_i\rvert$ / $\lvert a_i\rvert$	$N\lvert o_i + a_i\rvert$ / 1
MAAC	$N+1$	$\lvert o_i\rvert$ / $\lvert a_i\rvert$	$\lvert e_i + x_i\rvert$ / 1

[1] $|\cdot|$ denotes the cardinality.

actors are deployed for coordinated energy management of the community, as outlined by Algorithm 2. For a specific test day and each step t, each agent executes its policy using the trained actor network $\pi^{\phi_i^*}(o_{i,t})$ with the learned parameters ϕ_i^*, based only on its local observation $o_{i,t}$. In other words, during execution, the agents do not communicate with each other, and the coordination of their energy management and P2P trading decisions is performed in a fully decentralized fashion through the deployed actor networks.

It is constructive to clarify that the P2P trading platform functions within the community environment, which is treated as a blackbox to all the prosumer agents who interacts with it. Furthermore, The platform is assumed to act as a trustworthy third party which requires secure communication channel between the agents. The communicated information therefore needs to be encrypted in order to enhance security against attacks. Within the proposed framework, as explained previously, participated prosumers' privacy is preserved through information embedding and excluding the need of direct and repeated exchange of private information between agents.

VI. CASE STUDIES

A. Simulation Setup and Implementation

The MAAC method is trained and validated through case studies on a real-world dataset recorded by the Australian electricity distribution company Ausgrid [48]. This dataset provides residential load and PV generation data from 01/07/2011 to 30/06/2012 with a half-hourly resolution for 300 residential prosumers. Furthermore, the relevant outdoor temperature data is collected from the Australian government's open database [49]. The examined dataset enables us to evaluate the performance of the proposed method in a real-world, large-scale scenario for P2P transactive energy trading.

Considering the need to explore the generalizability of the examined RL coordination methods, and following the routine also employed in [24]–[27],[30]–[32], we use the data of the first 11 months as the training set and the data of the last month as the test set, thus evaluating the performance of the examined methods based on a set of conditions that has not been previously encountered by the agents during training. The supplier import and export prices are provided in [50] and [51], respectively. The assumed operating parameters of EV, ES, HVAC and SA are derived from [28], [35]. In addition, to capture the intrinsic variability of prosumers' requirements, we diversify each prosumer's EV departure and arrival times, initial energy level in the EV battery, energy requirements for traveling, as well as the earliest initiation and latest termination times of the SA using the truncated normal distributions. The employed simulation data is provided in a supplementary datasheet uploaded to IEEE DataPort [52].

For the purposes of our analysis, we compare 5 alternative coordination methods (except for MADDPG, the SAC learning algorithm (Section V-C) is employed by each prosumer).

U-SAC: Every prosumer trades independently with the supplier and P2P trading is not considered among the prosumers.

Cen-SAC: Prosumers coordinate by participating in P2P energy trading and employing the Centralized learning approach (Section V-B).

Con-SAC: Prosumers coordinate by participating in P2P trading and employing the Concurrent learning approach (Section V-B).

MADDPG: Prosumers coordinate by engaging in P2P trading and employing the MADDPG approach (Section V-B).

MAAC: Prosumers coordinate by participating in P2P trading and employing the MAAC method (Section V-C).

Table II summarizes the computational complexity of the examined coordination methods, by presenting the number of DNNs (with learnable parameters) that need to be trained as well as the input and output dimensions of the involved actor and critic networks. The highest complexity is observed in Cen-SAC (where both actor and critic dimensions are proportional to the number of agents) and MADDPG (where both number of DNNs and critic dimensions are proportional to the number of agents), impeding effective training of the DNNs and imposing impractical computational requirements. As a result, these two methods could not be successfully executed in the examined large-scale scenario (with 300 agents and 48 half-hourly steps), and thus the following sections present only results of the U-SAC, Con-SAC and MAAC.

On the other hand, MAAC exhibits the lowest computational complexity. This beneficial property is driven by the fact that i) the input dimension of the critic is substantially reduced due to the employment of the attention mechanism and ii) MAAC enables learning of all agents' critics jointly by sharing a set of learnable parameters among all agents.

The RMSProp optimizer [53] is employed to update the parameters of the online actor and critic, respectively. A soft update rate of $\tau = 10^{-3}$ is used. A discount factor of $\gamma = 0.99$ is used for the critic. The LSTM network extracts a

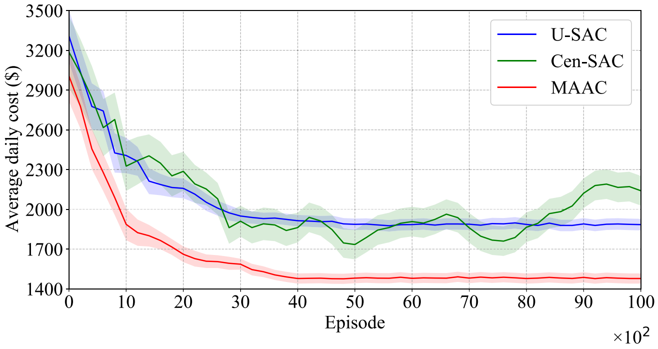

Fig. 4. Average daily cost of the community under different methods.

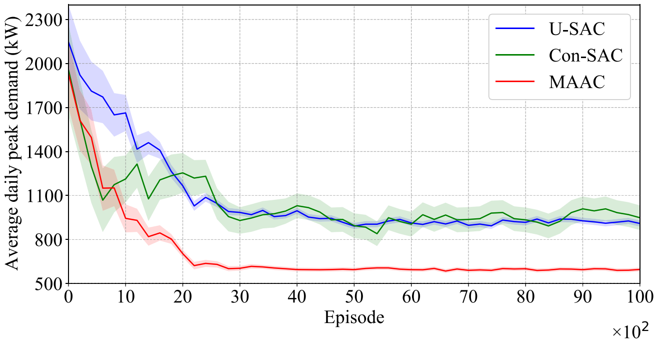

Fig. 5. Average daily peak demand of the community under different methods.

feature vector comprising future trends in retail import prices, inflexible demand, and PV production. This vector is then concatenated with the other relevant features and fed into the input layer of the actor and critic networks. The minibatch size and the replay buffer size are set as 128 and 10^5, respectively. Concerning the three penalty factors ω_1, ω_2 and ω_3, they need to attain sufficiently large values in order to ensure satisfaction of the respective physical constraints; the values employed in the case studies are $\omega_1 = 100$ and $\omega_2 = \omega_3 = 5$. The investigated methods have been implemented in Python with PyTorch [54], an open source platform for machine learning. The training process of the examined learning algorithms have been carried out on a computer with a 4-core 2.80 GHz Intel Core i7-7700HQ CPU and 16 GB of RAM.

B. Performance Evaluation of Coordination Methods

To assess the average performance of the examined coordination methods as well as the associated variability, 10 different random seeds are generated, and each method is trained for 10^4 episodes for each seed, where each episode signifies a random day selected from the training set. During training, the performance of each method is assessed regularly (every 100 episodes) on the test set. Figs. 4 and 5 illustrate, respectively, the average daily cost and peak demand of the community (over the 31 test days) for the examined methods. The mean and the standard deviation of the average daily cost and peak demand over the 10 seeds are illustrated in Fig. 4 and 5 through the solid lines and the shaded areas, respectively, and their values at convergence (or at termination in the case of Con-SAC) are presented in Table III.

TABLE III
MEAN AND STANDARD DEVIATION OF AVERAGE DAILY COST AND PEAK DEMAND OF THE COMMUNITY UNDER DIFFERENT COORDINATION METHODS (AT CONVERGENCE/TERMINATION)

Method	**Cost ($)**	**Peak demand (kW)**
U-SAC	1,884.86 ± 43.36	909.42 ± 21.34
Con-SAC	2,140.53 ± 126.42	948.63 ± 83.77
MAAC	1,478.08 ± 36.16	595.93 ± 13.27
TOB	1,406.11	600

It can be observed that the Concurrent method exhibits the highest standard deviation and an unstable learning behavior, eventually failing to reach convergence after 10^4 episodes. As discussed in Section V-B, this is because a large number (300) of agents are learning and adapting their energy management strategies independently, rendering the environment dynamics non-stationary for any agent. Consequently, the average daily cost and peak demand at termination are even higher than the ones achieved by U-SAC (the average daily cost is even higher compared to U-SAC), despite the fact that the latter involves no coordination among the prosumers. By contrast, MAAC selectively incorporates relevant information of agents to aid centralized training which effectively eliminates the non-stationarity effects. On this account, it converges after approximately 3,000 episodes, and exhibits the lowest standard deviation at convergence among all methods. Furthermore, MAAC significantly outperforms the other methods in terms of both community cost and peak demand, with the latter beneficial impact being also attributed to the reward shaping mechanism presented in Section IV-D. In relative terms, MAAC achieves 21.58%/30.95% lower average daily cost and 34.47%/37.18% lower average peak demand over U-SAC/Con-SAC, respectively (Table III), and is the only method satisfying the network capacity threshold.

Finally, we further validate the performance of the MAAC method by comparing it against the model-based optimization approach (Section II-D). The average daily cost and peak demand of the community over the 31 test days, under this *theoretical optimality benchmark* (TOB), are included in Table III. These results demonstrate that the solution produced by the MAAC method is very close to the theoretical benchmark (around 4.9% higher in community cost), despite the fact that it drops the highly unrealistic assumptions of model-based representation and fully centralized coordination, which would practically entail massive monitoring, communication, and computational costs (Section I-B).

C. Computational Performance and Scalability Analysis

Table IV summarizes the computational evaluation of the investigated methods, with regard to a) the average training time per episode, b) the number of episodes, c) the total training time required to attain convergence, and d) the average execution (or test) time per prosumer (i.e., the computational time required to determine the energy management actions of a prosumer in a test day). As previously discussed, Con-SAC fails to reach convergence and thus the reported number

TABLE IV
COMPUTATIONAL PERFORMANCE OF INVESTIGATED COORDINATION METHODS

Method	**U-SAC**	**Con-SAC**	**MAAC**
Average training time per episode (s)	2.19	2.30	1.37
Number of episodes	5,000	10,000*	4,000
Total training time (h)	3.04	6.39*	1.52
Average test time per prosumer (ms)	5.45	5.41	5.37

* Failure to attain convergence.

TABLE V
SCALABILITY ANALYSIS OF THE MAAC METHOD

Number of prosumer agents	**100**	**200**	**300**
% increase in community cost	5.03	4.88	4.90
Total training time (h)	0.51	0.85	1.33
Average test time per prosumer (ms)	4.91	5.08	5.37

of episodes and total training time corresponds to the 10^4 episodes executed in our experiments.

As shown in Table IV, the average training time per episode is the highest in U-SAC and Con-SAC, since both methods involve training of $2N$ DNNs at each time step. On the other hand, the training time per episode is significantly lower in MAAC, as it involves training N actor networks and a shared critic network. The number of episodes and the total training time required to reach convergence are highest in Con-SAC (since convergence is not achieved at termination), lower in U-SAC (since its learning process does not benefit from other agents' experiences, which results in slow convergence), and the lowest in MAAC owning to the employment of attention mechanism. Furthermore, it can be observed that all investigated MADRL methods exhibit a similar average test time per prosumer is in the order of milliseconds, implying that they can be effectively deployed in real-time energy management coordination applications. On the other hand, the computational time for the TOB is around 3.5 seconds, as it necessitates the solution of a large-scale optimization problem. Overall, these results reveal that beyond accomplishing a lower average community cost and peak demand (Table III), MAAC also exhibits a higher computational efficiency.

Going further, the scalability of the MAAC method is evaluated with increasing number of prosumer agents in terms of the solution quality (reflected in the percentage increase of community cost over TOB) and the overall computational performance (reflected in the total training time and test time per prosumer), as illustrated in Table V. The solution quality of MAAC is high with respect to the theoretical model-based benchmark, regardless of the number of prosumers. Furthermore, the total training time almost linearly increase with the number of agents, while the test time remains at around 5ms to execute the energy management decision of each agent, validating excellent scalability of MAAC.

D. Value of P2P Transactive Energy Trading

Having demonstrated the superiority of the MAAC method with regard to the state-of-the-art Centralized, Concurrent and MADDPG methods, the aim of this section lies in exploring the value stream from local demand-supply balancing and peak reduction provided by the coordinated P2P trading paradigm (under MAAC) over the conventional, uncoordinated peer-to-grid trading paradigm (under U-SAC). For each of these two paradigms, Fig. 6 illustrates the net demand (positive)/generation (negative) of the community (including for comparison purposes the net load of non-shiftable load and PV), while Fig. 7 illustrates the aggregate schedules of the four types of flexible DER (SA, HVAC, ES and EV) of the community (negative values indicate ES and EV discharging), averaged over the 31 test days.

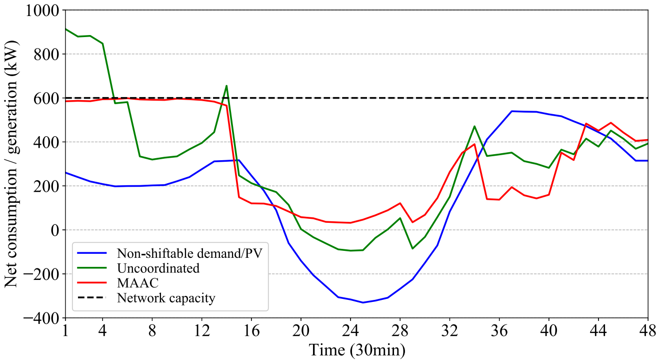

Fig. 6. Net demand/generation of the community under U-SAC and MAAC, averaged over the 31 test days.

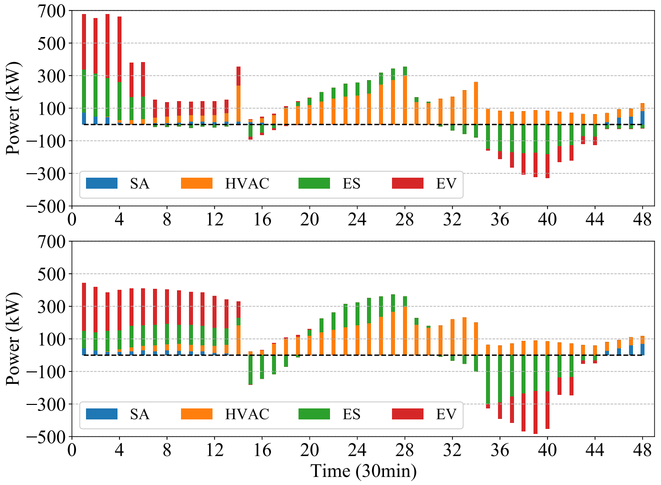

Fig. 7. Aggregate schedules of flexible DER of the community under (a) U-SAC and (b) MAAC, averaged over the 31 test days.

It can be observed that certain trends are prevailing under both the U-SAC and MAAC methods: i) ES and flexible EV charging, as well as the operation of flexible SA, are scheduled during the off-peak hours 22:00-7:00 to exploit the low, off-peak ToU prices; ii) the abundant PV generation during mid-day hours 8:00-18:00 is locally absorbed by charging the ES and operating the HVAC (the flexibility of EV and SA cannot be utilized for that purpose since their scheduling intervals do not encompass the mid-day periods; iii) ES and EV discharge during hours 16:00-21:00 which are characterized by high non-shiftable demand and no/low PV production; and iv) the HVAC systems are mainly operating during the mid-day hours 8:00-17:00 to ensure the indoor temperature within the comfort range when the outdoor temperature is high.

However, these two methods also exhibit evident differences. First of all, both PV generation absorption at hours 9:00-15:00 and the net demand reduction at hours 17:00-21:00 are more prominent under MAAC (Fig. 7), since P2P trading allows a more comprehensive harnessing of complementary DERs in the community. As a result, this allows the prosumers to weaken their exposure to the differentiated supplier prices (higher import/lower export prices) more effectively and thus to reduce their costs with regard to U-SAC (Table III). Furthermore, significant rebound demand peak is observed under U-SAC, resultant from the synchronization of the load activities of SA, ES and EV charging during the off-peak hours (when the offered supplier buy price is at its lowest value). On the contrary, such rebound peak is not presented under MAAC, thanks to the reward shaping scheme which effectively restricts prosumers' contribution to the demand peaks (Section IV-D). As a result, the peak demand of the community is reduced with regard to U-SAC and satisfies the network capacity threshold.

VII. Conclusion and Future Work

In this paper, we proposed a scalable, privacy-preserving MAAC-based method tailor-made to address the challenging coordinated energy management and P2P transactive trading of a large group of prosumers operating multiple and assorted DERs in an energy community. This method is founded on the CTDE MADRL paradigm and leverages its performance through the employment of the attention mechanism. The latter enables selecting relevant encoded information to estimate the critic, as opposite to indifferently incorporating the local information of all agents for such estimation. Furthermore, it enables learning of all agents' critics jointly by sharing a set of learnable parameters among agents. As such, the proposed method overcomes the non-stationarity, computational complexity, and privacy drawbacks of state-of-the-art MADRL approaches. The proposed method is complemented by a P2P trading platform which i) provides adequate financial incentives for prosumers' participation in local energy trading through an MMR pricing mechanism; ii) safeguards the security of the local distribution network by penalizing each prosumer's addition to rebound peaks through a new reward shaping mechanism, and iii) acts as a trusted third party providing prosumers with information regarding the collective trading behavior of other prosumers, thereby mitigating the non-stationarity while preserving prosumers' privacy.

Case studies on a real-world, large-scale scenario with 300 prosumers with diverse portfolio of DERs (including PV generators, ES units, HVAC systems with certain comfortable temperature range, EV with V2G/V2H flexibility and SA with deferrable cycles) have carried out numerous numerical simulations demonstrating that the MAAC method significantly outperforms the state-of-the-art MADRL methods. More specifically, the Cen-SAC and MADDPG approaches could not be successfully executed in the examined large-scale scenario due to computational intractability, while the Con-SAC approach has failed to reach convergence after 10,000 training episodes due to non-stationarity effects, and its achieved community cost and peak demand are higher to the ones achieved under no coordination among the prosumers (U-SAC). On the contrary, MAAC exhibits the fastest convergence (even faster than the case without coordination among the prosumers) and achieves 21.58%/30.95% lower average daily cost and 34.47%/37.18% lower average peak demand over U-SAC/Con-SAC methods, respectively.

Future work aims at applying the MAAC method to property account for the detailed representation of distribution network constraints (including nonlinear AC power flow equations capturing losses, voltage limits, and current thermal limits) of the distribution system in order to further enhance the potential of local energy trading and ensure that these constraints are not violated by energy trading activities.

References

[1] A. Shakoor, G. Davies, and G. Strbac, *Roadmap for Flexibility Services to 2030, A Report to the Committee on Climate Change*, Climate Change Committee, London, U.K., May 2017.

[2] D. Qiu, Y. Ye, and D. Papadaskalopoulos, "Exploring the effects of local energy markets on electricity retailers and customers," *Elect. Power Syst. Res.*, vol. 187, Dec. 2020, Art. no. 106761.

[3] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nat. Energy*, vol. 3, no. 2, pp. 94–101, Feb. 2018.

[4] F. Lezama, J. Soares, P. Hernandez-Leal, M. Kaisers, T. Pinto, and Z. Vale, "Local energy markets: Paving the path toward fully transactive energy systems," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 4081–4088, Sep. 2019.

[5] L. Ma, N. Liu, J. Zhang, and L. Wang, "Real-time rolling horizon energy management for the energy-hub-coordinated prosumer community from a cooperative perspective," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1227–1242, Mar. 2019.

[6] M. R. Alam, M. St-Hilaire, and T. Kunz, "Peer-to-peer energy trading among smart homes," *Appl. Energy*, vol. 238, pp. 1434–1443, Mar. 2019.

[7] M. S. H. Nizami, M. J. Hossain, and E. Fernandez, "Multiagent-based transactive energy management systems for residential buildings with distributed energy resources," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1836–1847, Mar. 2020.

[8] T. Morstyn and M. D. McCulloch, "Multiclass energy management for peer-to-peer energy trading driven by prosumer preferences," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 4005–4014, Sep. 2019.

[9] J. Li, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Computationally efficient pricing and benefit distribution mechanisms for incentivizing stable peer-to-peer energy trading," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 734–749, Jan. 2021.

[10] M. Khorasany, A. Najafi-Ghalelou, and R. Razzaghi, "A framework for joint scheduling and power trading of prosumers in transactive markets," *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 955–965, Apr. 2021.

[11] J. L. Crespo-Vazquez, T. AlSkaif, Á. M. González-Rueda, and M. Gibescu, "A community-based energy market design using decentralized decision-making under uncertainty," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1782–1793, Mar. 2021.

[12] W. Tushar *et al.*, "Peer-to-peer energy systems for connected communities: A review of recent advances and emerging challenges," *Appl. Energy*, vol. 282, Jan. 2021, Art. no. 116131.

[13] P. Shamsi, H. Xie, A. Longe, and J.-Y. Joo, "Economic dispatch for an agent-based community microgrid," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2317–2324, Oct. 2016.

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

[15] W. Liu, D. Qi, and F. Wen, "Intraday residential demand response scheme based on peer-to-peer energy trading," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1823–1835, Mar. 2020.

[16] S. Cui, Y.-W. Wang, Y. Shi, and J.-W. Xiao, "A new and fair peer-to-peer energy sharing framework for energy buildings," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 3817–3826, Sep. 2020.

[17] S. Cui and J.-W. Xiao, "Game-based peer-to-peer energy sharing management for a community of energy buildings," *Int. J. Elect. Power Energy Syst.*, vol. 123, Dec. 2020, Art. no. 106204.

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

[19] C. Long, C. Zhang, J. Wu, L. Thomas, M. Cheng, and N. Jenkins, "Peer-to-peer energy trading in a community microgrid," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Chicago, IL, USA, Jul. 2017, pp. 1–5.

[20] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, 2nd ed. New York, NY, USA: Springer, 2011.

[21] M. Farhoumandi, Q. Zhou, and M. Shahidehpour, "A review of machine learning applications in IoT-integrated modern power systems," *Elect. J.*, vol. 34, no. 1, Jan./Feb. 2021, Art. no. 106879.

[22] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[23] K. Mason and S. Grijalva, "A review of reinforcement learning for autonomous building energy management," *Comput. Elect. Eng.*, vol. 78, pp. 300–312, Sep. 2019.

[24] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3068–3082, Jul. 2020.

[25] C. Zhang, S. R. Kuppannagari, C. Xiong, R. Kannan, and V. K. Prasanna, "A cooperative multi-agent deep reinforcement learning framework for real-time residential load scheduling," in *Proc. Int. Conf. Internet Things Design Implement.*, Montreal, QC, Canada, Apr. 2019, pp. 59–69.

[26] J. R. Vázquez-Canteli, S. Ulyanin, J. Kämpf, and Z. Nagy, "Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities," *Sustain. Cities Soc.*, vol. 45, pp. 243–257, Feb. 2019.

[27] H.-M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2752–2763, Apr. 2021.

[28] D. Papadaskalopoulos and G. Strbac, "Nonlinear and randomized pricing for distributed management of flexible loads," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1137–1146, Mar. 2016.

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

[30] J. Vazquez-Canteli, T. Detjeen, G. Henze, J. Kämpf, and Z. Nagy, "Multi-agent reinforcement learning for adaptive demand response in smart cities," *J. Phys. Conf. Series*, vol. 1343, no. 1, Nov. 2019, Art. no. 012058.

[31] H. Kazmi, J. Suykens, A. Balint, and J. Driesen, "Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads," *Appl. Energy*, vol. 238, pp. 1022–1035, Mar. 2019.

[32] R. Lu, Y.-C. Li, Y. Li, J. Jiang, and Y. Ding, "Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management," *Appl. Energy*, vol. 276, Oct. 2020, Art. no. 115473.

[33] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[34] Y. Ye, D. Papadaskalopoulos, R. Moreira, and G. Strbac, "Investigating the impacts of price-taking and price-making energy storage in electricity markets through an equilibrium programming model," *IET Gener. Transm. Distrib.*, vol. 13, no. 2, pp. 305–315, 2019.

[35] Y. F. Du, L. Jiang, C. Duan, Y. Z. Li, and J. S. Smith, "Energy consumption scheduling of HVAC considering weather forecast error through the distributionally robust approach," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 846–857, Mar. 2018.

[36] Y. Ye, D. Papadaskalopoulos, and G. Strbac, "Factoring flexible demand non-convexities in electricity markets," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 2090–2099, Jul. 2015.

[37] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 157–163.

[38] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci.*, vol. 39, no. 10, pp. 1095–1100, 1953.

[39] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[40] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, Jun. 2014, pp. 1–9.

[41] L. Yu *et al.*, "Deep reinforcement learning for smart home energy management," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 2751–2762, Apr. 2020.

[42] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[43] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826–3839, Sep. 2020.

[44] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems (NIPS)*. Red Hook, NY, USA: Curran, 2017, pp. 6379–6390.

[45] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2961–2970.

[46] J. Jiang and Z. Lu, "Learning attentional communication for multi-agent cooperation," in *Proc. 32nd Int. Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 7254–7264.

[47] 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. Res.*, Stockholm, Sweden, 2018, pp. 1861–1870.

[48] E. L. Ratnam, S. R. Weller, C. M. Kellett, and A. T. Murray, "Residential load and rooftop PV generation: An australian distribution network dataset," *Int. J. Sustain. Energy*, vol. 36, no. 8, pp. 787–806, 2017.

[49] (2017). *Rainfall and Temperature Forecast and Observations—Verification 2016-05 to 2017-04*. [Online]. Available: https://data.gov.au/data/dataset/0bfba2bc-2042-4ae3-91a1-17e4414e4391

[50] Ausgrid. (2020). *Time of Use Pricing*. [Online]. Available: https://www.ausgrid.com.au/Your-energy-use/Meters/Time-of-use-pricing

[51] EnergyAustralia. (2020). *Solar Rebates and Feed-in Tariffs*. [Online]. Available: https://www.energyaustralia.com.au/home/electricity-and-gas/solar-power/feed-in-tariffs

[52] *IEEE Dataport Dataset*. Accessed: May 26, 2021. [Online]. Available: https://dx.doi.org/10.21227/80dj-hs08

[53] O. Wichrowska *et al.*, "Learned optimizers that scale and generalize," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 3751–3760.

[54] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst (NIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 8026–8037.

Yujian Ye (Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Northumbria University, Newcastle Upon Tyne, U.K., in 2011, and the M.Sc. degree in control systems and the Ph.D. degree from Imperial College London, London, U.K., in 2013 and 2017, respectively. He is currently an Associate Professor with Southeast University, Nanjing, China, and a Visiting Researcher with Imperial College London. His current research interests include development and application of artificial intelligence techniques in power and energy systems modeling, analysis control as well as modeling, and optimization of economics of power system operation and planning.

Yi Tang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2000, 2002, and 2006, respectively. He is currently a Professor with the School of Electrical Engineering, Southeast University, Nanjing, China. His research interest includes power system analysis, renewable energy, and the application of artificial intelligence techniques in power systems.

Huiyu Wang (Student Member, IEEE) received the B.S. degree in electrical engineering and automation from the Nanjing University of Science and Technology, Nanjing, China, in 2020. She is currently pursuing the M.Eng. degree in electrical engineering with Southeast University, Nanjing, China. Her research interests include the energy market modeling and renewable generation planning investment.

Xiao-Ping Zhang (Fellow, IEEE) is currently a Professor of Electrical Power Systems with the University of Birmingham, U.K., and he is also the Director of Smart Grid with Birmingham Energy Institute and the Co-Director of Birmingham Energy Storage Center. He has been appointed recently to the Expert Advisory Group of U.K. Government's Offshore Transmission Network Review. He has coauthored the first and second edition of the monograph *Flexible AC Transmission Systems: Modeling and Control* (Springer, 2006 and 2012). He has coauthored the book *Restructured Electric Power Systems: Analysis of Electricity Markets With Equilibrium Models* (IEEE Press/Wiley, 2010). His research interests include modeling and control of HVDC, FACTS and wind/wave generation, distributed energy systems and market operations, and power system planning. He has been the Advisor to IEEE PES U.K. and Ireland Chapter and chairing the IEEE PES WG on Test Systems for Economic Analysis. He has been made a Fellow of IEEE for contributions to modeling and control of high-voltage DC and AC transmission systems. He is an IEEE PES Distinguished Lecturer on HVDC, FACTS, and Wave Energy Generation. He is also a Fellow of IET.

Goran Strbac (Member, IEEE) is a Professor of Energy Systems with Imperial College London, London, U.K. He led the development of novel advanced analysis approaches and methodologies that have been extensively used to inform industry, governments, and regulatory bodies about the role and value of emerging new technologies and systems in supporting cost effective evolution to smart low carbon future. He is currently the Director of the joint Imperial-Tsinghua Research Centre on Intelligent Power and Energy Systems, a Leading Author in IPCC WG 3, a member of the European Technology and Innovation Platform for Smart Networks for the Energy Transition and the Joint EU Programme in Energy Systems Integration of the European Energy Research Alliance.