Multi-Agent Reinforcement Learning for Automated Peer-to-Peer Energy Trading in Double-Side Auction Market

Dawei Qiu[*] , **Jianhong Wang** , **Junkai Wang** and **Goran Strbac**
Imperial College London
{d.qiu15, jianhong.wang16, junkai.wang17, g.strbac}@imperial.ac.uk

Abstract

With increasing prosumers employed with distributed energy resources (DER), advanced energy management has become increasingly important. To this end, integrating demand-side DER into electricity market is a trend for future smart grids. The double-side auction (DA) market is viewed as a promising peer-to-peer (P2P) energy trading mechanism that enables interactions among prosumers in a distributed manner. To achieve the maximum profit in a dynamic electricity market, prosumers act as price makers to simultaneously optimize their operations and trading strategies. However, the traditional DA market is difficult to be explicitly modelled due to its complex clearing algorithm and the stochastic bidding behaviors of the participants. For this reason, in this paper we model this task as a multi-agent reinforcement learning (MARL) problem and propose an algorithm called DA-MADDPG that is modified based on MADDPG by abstracting the other agents' observations and actions through the DA market public information for each agent's critic. The experiments show that 1) prosumers obtain more economic benefits in P2P energy trading w.r.t. the conventional electricity market independently trading with the utility company; and 2) DA-MADDPG performs better than the traditional Zero Intelligence (ZI) strategy and the other MARL algorithms, e.g., IQL, IDDPG, IPPO and MADDPG.

1 Introduction

Power systems are undergoing a significant transition from the fossil fuel resources to the decarbonlization of renewable energy resources (RES), providing the environmental energy challenges [Haller *et al.*, 2012]. However, the less controllable and predictable output of RES introduce the new challenges to the power system planning and operation. In this respect, there has been a significant increase in developing small-scale *distributed energy resources* (DER) that are connected to the end consumers and provide flexibility toward to a more reliable system. The key categories of DER include distributed generation, energy storage (ES) and demand response [Jiayi *et al.*, 2008]. In this context, traditional consumers evolve to *prosumers*, who can manage their energy consumption, production and storage of electricity through a *home energy management system* (HEMS) targeted to minimize their cost of energy bills and provide the system flexibility [Parag and Sovacool, 2016].

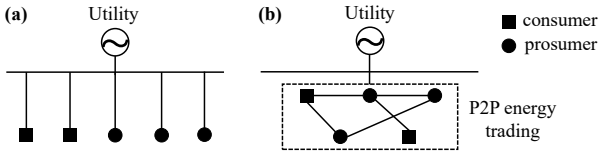


Figure 1: Paradigms of (a) traditional electricity market and (b) P2P energy trading.

With the change of system structure, the electricity market will have to accommodate a large number of small-scale renewable energy prosumers. Feed-in Tariff (FiT) scheme has been introduced as the most common production subsidies across the world in which prosumers receive payments for the energy exported to the grid at FiT offered by the upstream utility company [Qiu *et al.*, 2020]. This scheme allows prosumers to make use of the flexibility of their self-generated PV electricity and feed the surplus into the grid. Nevertheless, this subsidy may not recover the installation and operation cost of flexible DER. On the other hand, consumers who want to import energy from the grid may complain about the expensive Time-of-Use (ToU) price offered by the upstream utility company [Qiu *et al.*, 2020]. Furthermore, with the traditional electricity market in Figure 1 (a), prosumers and consumers under such an independent and uncoordinated fashion only benefit from their own DER flexibility to ensure the self-temporal balancing of generation and consumption, but do not benefit from their sufficient energy flexibility interacted with each other to satisfy the overall system energy balance [Morstyn *et al.*, 2018].

To this end, we introduce a *peer-to-peer* (P2P) energy trading paradigm that enables prosumers and consumers to trade locally independent of the upstream utility company [Morstyn *et al.*, 2018], as illustrated in Figure 1 (b). In order to coordinate this energy trading activity, *double-side auction* (DA) is set up as an incentive-driven market mechanism to

[*]Corresponding author.

attract prosumers and consumers to cooperatively participate in local trading with the application of *information and communication technologies* (ICTs), and is validated as a highly efficient market mechanism [Friedman, 2018]. In DA market, buyers (consumers) and sellers (prosumers) strategically bid and offer their willing price and corresponding energy quantity, respectively. The market dynamics are expected to maximize the market clearing efficiency by shifting demand to the periods of PV generation and/or low prices as well as enabling sellers with PV and/or ES to sell their self-generated energy to buyers. We use *multi-agent reinforcement learning* (MARL) to learn individual trading strategies and energy schedules of flexible DER to minimize the energy bills. The experiments on a real-world scenario demonstrate that 1) the above effects in P2P energy trading can potentially result in economic benefits for all the market participants as well as higher balance of local demand and generation; and 2) the proposed DA-MADDPG obtains the superior performance over the Zero Intelligence (ZI) strategy [Friedman, 2018] and the other MARL algorithms (e.g., IQL [Watkins and Dayan, 1992], IDDPG [Lillicrap *et al.*, 2016], IPPO [Schulman *et al.*, 2017] and MADDPG [Lowe *et al.*, 2017]).

2 Related Work

An appropriate market mechanism is required to facilitate the P2P energy trading among prosumers and consumers. [Alam *et al.*, 2019] proposed a central coordinator that directly manages the energy schedules of all participants by maximizing the overall economic benefits. In this setting, the central coordinator has to require all participants' economic and technical parameters, violating their security and privacy issues of energy preferences and usage behaviors [Aitzhan and Svetinovic, 2018]. In this contrast, the introduced DA market under a decentralized manner has shown its market efficiency [Guerrero *et al.*, 2019], and only a small amount of information is delivered to participants at relatively low computational costs.

In DA market, buyers and sellers are faced with a complex quotation decision process. Thus, an appropriate trading strategy is challenging to select in such a complicated market environment. Zero Intelligence (ZI) is a fundamental and popular trading strategy adopted by traders in DA market [Friedman, 2018]. ZI traders set their order price as a random surplus offset from its valuation, based on a uniform distribution from a specified range (e.g., FiT and ToU) without considering market transactions. In particular, the extant ZI strategy has been typically developed assuming that the market is static and all participants' energy availability is pre-optimized by solving a day-ahead energy planning with perfect information, which means there is no change in demand and generation at the beginning of each trading day [Guerrero *et al.*, 2019]. However, the real market is typically very dynamic considering that all participants' strategies are adjusted in real time as well as the information of price signals, PV generation and energy usage behaviors are stochastic.

To address this challenge, *reinforcement learning* (RL) is a framework to study sequential decision-making problems of agents (prosumers and consumers) gradually learning the optimal trading strategies by utilizing experiences acquired from its repeated interactions with the environment (P2P energy trading in DA market) [Sutton and Barto, 2018], which has been applied in many smart grid applications [Zhang *et al.*, 2018]. Specifically, for DA market, [Nicolaisen *et al.*, 2001] applied a modified Roth-Erev RL algorithm to help energy traders determine their price-quantity strategies in each auction round. Authors in [Sun *et al.*, 2015] presented a general RL bidding strategy for controlling and coordinating HVAC (heat, ventilating, and air conditioning) systems in a DA market. Authors in [Pedasingu *et al.*, 2020] applied DQN to adjust the price-quantity strategies in a day-ahead DA market. However, the majority of them only consider discrete state and/or action spaces by simply discretizing the original continuous spaces. Thus, the naïve discretization throws away valuable information regarding the structure of the state and action domain, which may be essential to achieve optimal trading strategies. To this end, this paper has employed the deep MARL algorithm based on DDPG [Lillicrap *et al.*, 2016] method to obtain multi-dimensional and continuous state and action spaces.

3 Energy Trading in DA Market

In this section, we define the energy trading rules in DA market and then formulate them into Markov Games.

3.1 DA Market Mechanism

The DA market matches multiple buyers (consumers) and sellers (prosumers) who are interested in (energy) trading, and is deemed as highly efficient mechanism. They are widely used in the trading of various types of commodities, such as stocks and electricity. A DA market lasts a fixed period of time, known as the *auction period* (e.g., hourly resolution in electricity market). It allows traders to submit their bids/offers at the beginning of an auction period, then the *auctioneer* (e.g., market operator) clears the market and publishes the public market outcomes (e.g., trading prices and quantities) at the end of each auction period [Friedman, 2018]. More specifically, a DA market comprises:

- A set of *buyers* $\mathcal{B}$, where each $i \in \mathcal{B}$ defines its trading price p_i^b and the amount of energy to buy q_i^b, which means the buyer i would like to buy the q_i^b amount of energy at the price p_i^b.
- A set of *sellers* $\mathcal{S}$, where each $j \in \mathcal{S}$ defines its trading price p_j^s and the amount of energy to sell q_j^s, which means the seller j would like to sell the q_j^s amount of energy at the price p_j^s.
- A public order book managed by an auctioneer, where the accepted bids and offers are listed, respectively. Buy orders queue in order book $k^b(i, p_i^b, q_i^b)$ and are sorted by decreasing submitted buy prices, while sell orders queue in order book $k^s(j, p_j^s, q_j^s)$ and are sorted by increasing submitted sell prices.

The pseudo-code of the matching process in DA market is given in Algorithm 1. Once an auction period begins, traders submit their order information with a trading price and a corresponding energy quantity to the market. All submitted or-

Algorithm 1 DA market clearing algorithm

1: Allocate order book k_t^b and k_t^s at auction period t
2: Initialize $i = j = 1$
3: **while** $p_{i,t}^b \geq p_{j,t}^s$ **do**
4: match the trading energy: $q_t^l = min(q_{i,t}^b, q_{j,t}^s)$
5: calculate the trading price: $p_t^l = (p_{i,t}^b + p_{j,t}^s)/2$
6: update buy order book $q_{i,t}^b \leftarrow q_{i,t}^b - q_t^l$
7: **if** $q_{i,t}^b = 0$ **then**
8: $i \leftarrow i + 1$
9: update sell order book $q_{j,t}^s \leftarrow q_{j,t}^s - q_t^l$
10: **if** $q_{j,t}^s = 0$ **then**
11: $j \leftarrow j + 1$
12: **break if**
13: $i >$ length of k_t^b or $j >$ length of k_t^s
14: **end while**
15: Balance the unmatched quantity at ToU λ_t^b and FiT λ_t^s

ders are allocated in the order book (step 1). The matching algorithm iterates down the order books and attempts to match each buy order with sell order (steps 4-11) until the buy price is less than the sell price or no unmatched sell/buy order exists anymore (steps 12-13). Specifically, when two orders get matched, the auctioneer calculates the market clearing price between the matched buy price and sell price, using the traditional mid-pricing method [Friedman, 2018] (step 5), while the transaction quantity is equal to the minimum quantity between the matched orders (step 4). Due to the sorting principle and clearing algorithm, the clearing results promise the social welfare maximization [Friedman, 2018]. Finally, at the end of the auction period, the remaining quantity of energy and the unmatched orders are balanced by the auctioneer with the utility company at the grid prices of ToU and FiT. It should be noted that the pricing strategies of all traders are bounded between FiT and ToU to ensure the economic benefits.

3.2 Numerical Example of DA Market Mechanism

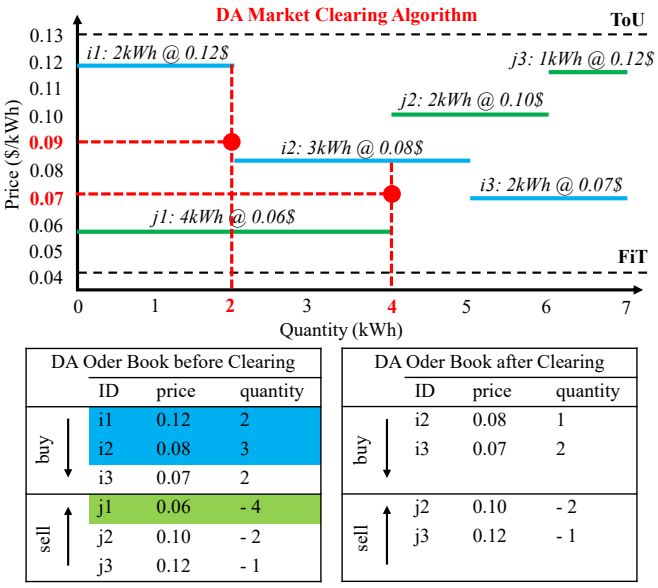
To better illustrate the DA market clearing algorithm, we take the scenario in Figure 2 as an example.

Order Books

There are six agents participating into DA market. They are divided into three buyers and three sellers corresponding to the sign of their submitted quantities (positive for buyer, negative for seller). Then, the auctioneer allocates the order books according to the principle of *price first* (i.e., the quotes of buyers are in high-to-low order, whereas the quotes of sellers are in low-to-high), as presented in the left table.

Market Transactions

Transaction 1: the first transaction occurs when the first bid price ($0.12/kWh) is higher than the first ask price ($0.06/kWh), the matched quantity is the minimum quantity of buyer $i1$ and seller $j1$ (i.e., 2kWh) and the transaction price is the average of $0.12/kWh and $0.06/kWh (i.e., $0.09/kWh). In this transaction, buyer $i1$ is completed matched and should be removed, and buyer $i2$ is updated on



DA Oder Book before Clearing

	ID	price	quantity
buy	i1	0.12	2
	i2	0.08	3
	i3	0.07	2
sell	j1	0.06	- 4
	j2	0.10	- 2
	j3	0.12	- 1

DA Oder Book after Clearing

	ID	price	quantity
buy	i2	0.08	1
	i3	0.07	2
sell	j2	0.10	- 2
	j3	0.12	- 1

Figure 2: Example case of DA market clearing algorithm.

the top of buy order book. On the other hand, the quotes of sellers do not change but the quantity of seller $j1$ is reduced to 2kWh.

Transaction 2: the second transaction occurs when the first bid price ($0.08/kWh) is still higher than the first ask price ($0.06/kWh), the matched quantity is the minimum quantity of buyer $i2$ and unmatched seller $j1$ (i.e., 2kWh) and the transaction price is the average of $0.08/kWh and $0.06/kWh (i.e., $0.07/kWh). In this transaction, seller $j1$ is completely matched and should be removed, while seller $j2$ is updated on the top of sell order book. The quantity of buyer $i2$ is reduced to 1kWh accordingly.

Transaction ends: this matching process stops when the first bid price ($0.08/kWh) is lower than the first ask price ($0.10/kWh). Finally, the unmatched quantities in the order book (right table) are balanced via ToU ($0.13/kWh) for buyer $i2$ at 1kWh and buyer $i3$ at 2kWh while FiT ($0.04/kWh) for seller $j2$ at 2kWh and seller $j3$ at 1kWh.

Market Outcomes

After all transactions occur in the DA market, the auctioneer publishes the market clearing outcomes that comprises: 1) the local trading price ($0.09/kWh for buyer $i1$ & seller $j1$, $0.07/kWh for buyer $i2$ & seller $j1$); 2) the cleared quantity in DA market for each agent; 3) the remaining/unmatched quantity traded with the utility company for each agent; and 4) the updated order books for all agents, as presented in the right table.

3.3 P2P Energy Trading as Markov Games

The DA market clearing algorithm outlined above can be formulated as a multi-agent coordination problem in the form of a finite *Partially Observable Markov Game* (POMG) [Shoham and Leyton-Brown, 2008] with discrete time steps. The POMG is defined by N agents with a set of state $\mathcal{S}$ describing global state, a collection of private observations

$\{\mathcal{O}_{1:N}\}$, a collection of action sets $\{\mathcal{A}_{1:N}\}$, a collection of reward functions $\{\mathcal{R}_{1:N}\}$ and a state transition function $\mathcal{T}$. The time interval between two consecutive steps is one auction period (Δt = 1 hour). At time step t, each agent n chooses an action $a_{n,t}$ according to its policy $\mu_n(o_{n,t})$ based on its private observation $o_{n,t}$. The environment then moves into the next state according to the state transition function conditioned on the actions of all agents. Each agent n obtains a reward $r_{n,t}$ and a private observation for next step $o_{n,t+1}$. Each agent n aims to maximize its cumulative discounted reward $R_n = \sum_{t=0}^{T} \gamma^t r_{n,t}$, where $\gamma \in [0, 1)$ is the discount factor and T is the time horizon of the coordination problem (24 hours). The components of the POMG are detailed as:

Observation. Agent n at time step t has its observation $o_{n,t} = [P^{inf}_{n,t}, E^{es}_{n,t}, \lambda^b_t, \lambda^s_t]$, which comprises: 1) the local information of its inflexible load $P^{inf}_{n,t}$[1], ES battery energy content $E^{es}_{n,t}$; and 2) the grid information of ToU λ^b_t and FiT λ^s_t.

Action. Agent n at time step t controls its action $a_{n,t} = [a^p_{n,t}, a^q_{n,t}]$, which comprises: 1) the price decision $a^p_{n,t} \in [0, 1]$ representing the magnitude of willing price submitted to DA market as a ratio of FiT and ToU price differentials ($p_{n,t} = \lambda^s_t + a^p_{n,t}(\lambda^b_t - \lambda^s_t)$); and 2) the energy decision $a^q_{n,t} \in [-1, 1]$ representing the magnitude of charging (positive) and discharging (negative) power of ES as a ratio of its power capacity $\overline{P}^{es}_n$.

State Transition. The state transition from time step t to $t + 1$ is governed by a function: $s_{t+1} = \mathcal{T}(s_t, a_{1:N,t}, \omega_t)$. It can be observed that the transition is influenced partly by all agents' actions $a_{1:N,t}$ and partly by the environment stochasticity ω_t. In the examined problem, this corresponds to the exogenous state features $[P^d, P^{pv}, \lambda^b, \lambda^s]$, which are decoupled from agent's actions and are characterized by inherent variability and uncertainty. In this context, it presents significant challenges to identify suitable probabilistic models which can fully capture such randomness since it is influenced by many exogenous factors, such as energy usage behaviors, solar radiation, and utility pricing mechanism. RL remedies this problem in a data-driven approach thta does not rely on accurate mathematical models of the underlying uncertainties.

By contrast, the state transition for the endogenous state features $E^{es}_{n,t}$ is determined by the action $a^q_{n,t}$ adopted at step t. Let $C^{es}_{n,t}$ and $D^{es}_{n,t}$ denote the charging and discharging power of ES, respectively. The mutually exclusive quantities $C^{es}_{n,t}$ and $D^{es}_{n,t}$ (as the charging and discharging activity of ES cannot occur simultaneously at a given step) are managed by action $a^q_{n,t}$, and are also limited by $E^{es}_{n,t}$ and the ES operating parameters: 1) minimum and maximum energy levels $\underline{E}^{es}_n, \overline{E}^{es}_n$; 2) charging and discharging efficiencies $\eta^{esc}_n, \eta^{esd}_n$.

$$C^{es}_{n,t} = \min\big(a_{n,t}\overline{P}^{es}_n, (\overline{E}^{es}_n - E^{es}_{n,t})/(\eta^{esc}_n \Delta t)\big) \quad (1)$$

$$D^{es}_{n,t} = \min\big(a_{n,t}\overline{P}^{es}_n, (\underline{E}^{es}_n - E^{es}_{n,t})\eta^{esd}_n/\Delta t\big) \quad (2)$$

[1]for consumers: inflexible load equals to inflexible demand $P^{inf}_{n,t} = P^d_{n,t}$; for prosumers: inflexible load equals to the difference between inflexible demand and PV generation $P^{inf}_{n,t} = P^d_{n,t} - P^{pv}_{n,t}$.

Based on $C^{es}_{n,t}$ and $D^{es}_{n,t}$, the state transition of $E^{es}_{n,t}$ can be expressed as:

$$E^{es}_{n,t+1} = E^{es}_{n,t} + C^{es}_{n,t}\Delta t\eta^{esc}_n + D^{es}_{n,t}\Delta t/\eta^{esd}_n \quad (3)$$

Consequently, the quantity submitted to DA market $q_{n,t}$ of agent n at step t can be expressed as the summation of inflexible load and ES charging / discharging power, where the positive value represents the net demand to buy while the negative value represents the net generation to sell in DA market:

$$q_{n,t} = P^{inf}_{n,t} + C^{es}_{n,t} + D^{es}_{n,t} \quad (4)$$

After collecting the price-quantity strategies $(p_{n,t}, q_{n,t})$ of all agents, the auctioneer allocates the order book k^b_t and k^s_t, clears the market (Algorithm 1) and publishes the market clearing outcomes $[\lambda^l_{n,t}, q^{da}_{n,t}, q^{grid}_{n,t}, k^b_t, k^s_t]$, which comprises: the local trading price $\lambda^l_{n,t}$, the cleared quantity in DA market $q^{da}_{n,t}$, the remaining/unmatched quantity traded with the utility company $q^{grid}_{n,t}$, and the updated public order books k^b_t, k^s_t.

Reward Function. Agent n at step t obtains its reward $r_{n,t}$ as the negative cost of energy bills developing from the DA market clearing outcomes. Specifically, for these agents who are successfully cleared in DA market will receive the local price $\lambda^l_{n,t}$ and its cleared quantity $q^{da}_{n,t}$, then each agent n can calculate its corresponding cost in DA market and the remaining/unmatched quantity $q^{grid}_{n,t}$ will be bought or sold through the utility company at ToU λ^b_t or FiT λ^s_t. For these agents who are unsuccessfully cleared in DA market, their quantity $q^{grid}_{n,t} = q_{n,t}$ will be directly traded at ToU λ^b_t or FiT λ^s_t.

$$r_{n,t} = -(\lambda^l_{n,t} q^{da}_{n,t}\Delta t + \lambda^b_t [q^{grid}_{n,t}]^+ \Delta t + \lambda^s_t [q^{grid}_{n,t}]^- \Delta t) \quad (5)$$

where $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$.

4 MARL Method

Directly applying single-agent RL methods to the multi-agent setting by treating other agents as part of the environment is problematic as the environment appears non-stationary from the view of any one agent, violating Markov assumptions required for convergence [Shoham and Leyton-Brown, 2008]. Specifically, this non-stationary issue becomes more severe in the case of deep RL with neural networks as function approximators. To this end, we propose an extension of MADDPG [Lowe *et al.*, 2017], namely DA-MADDPG that learns a centralized Q-function for each agent (to alleviate the non-stationary problem and stabilize training) by abstracting the other agents' observations and actions through the DA market information (to promise the scalability and protect the private information). The general architecture of DA-MADDPG is illustrated in Figure 3.

Concretely, we consider a game of N agents with policies parameterized by $\boldsymbol{\theta^\mu} = \{\theta^\mu_{1:N}\}$, and let $\boldsymbol{\mu} = \{\mu_{1:N}\}$ be the set of all agents' policies. Then the gradient of the expected return for agent n with policy $\mu_n(o_n|\theta^\mu_n)$, $J_n = \mathbb{E}_{\mu_n}[R_n]$ is written as follows:

$$\begin{aligned}\nabla_{\theta^\mu_n} J_n = \mathbb{E}_{\mathrm{x},a\sim\mathcal{D}}\big[&\nabla_{\theta^\mu_n}\mu_n(o_n|\theta^\mu_n) \\ &\nabla_{a_n} Q_n(\mathrm{x}, a_{1:N}|\theta^Q_n)|_{a_n=\mu_n(o_n|\theta^\mu_n)}\big]\end{aligned} \quad (6)$$

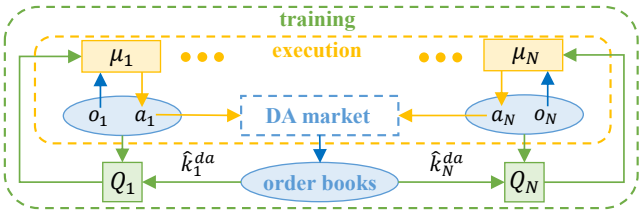


Figure 3: Architecture of DA-MADDPG.

here $Q_n(\mathrm{x}, a_{1:N}|\theta_n^Q)$ parameterized by θ_n^Q is a *centralized action-value function* that takes as input all agents' actions $a_{1:N}$, in addition to some state information x (e.g., all agents' observations $o_{1:N}$) and outputs the Q-value for agent n.

The conventional approach in MADDPG involves training all agents using a centralized critic estimator which takes as input the observations and actions of all agents. However, it is evident that the input dimension of such centralized critic grows exponentially with the number of agents, quickly rendering the problem intractable. Furthermore, driven by prosumers/consumers' privacy concerns, they are not willing to exchange their local observations (inflexible demand and/or PV generation) and actions (ES energy behaviors) with each other. Motivated by [Yang *et al.*, 2019], this paper assumes the DA auctioneer as a trusted third party supplying agents with the public and dynamic market information of order books that epitomize the collective behavior of the market (thereby substituting the high-dimensional vector with all agents' information) in the centralized training process. This substantially improves the scalability of the proposed MARL method and also protects the privacy of prosumers/consumers. To this effect, we approximate the joint Q-value function as:

$$Q_n(o_{1:N}, a_{1:N}|\theta_n^Q) \approx Q_n(o_n, a_n, \hat{k}_n^{da}|\theta_n^Q) \tag{7}$$

where $\hat{k}_n^{da} = \{\hat{k}_n^b, \hat{k}_n^s\} = \{k_{n-}^b, k_{n-}^s, \forall n- \in \mathcal{N} \setminus \{n\}\}$ denotes the combination of buy and sell order books other than agent n in DA market. It can be observed that $\hat{k}_n^{da}$ is an embedded function that not only abstracts all other agents' observations (e.g., P_{n-}^{inf}) as well as actions of the price strategies p_{n-} and the quantity strategies q_{n-} resulting from their ES energy schedules C_{n-}^{es} and D_{n-}^{es}, but also displays the DA market dynamics of P2P energy trading. As a result, this function provides a good approximation of agents' observations and actions as well as the DA market dynamics. Incorporating $\hat{k}_n^{da}$ in the critic estimation, each agent can make acquainted decisions on the basis of the impact of the actions of other agents in DA market, albeit not knowing their specific energy portfolios and usage activities. This averts the explosion of action and observation spaces and alleviates the environmental non-stationarity.

Given the collective Q-value function, the experience replay buffer $\mathcal{D}_n$ contains the tuples $(o_n, a_n, r_n, o_n', \hat{k}_n^{da})$, recording experiences for each agent n. The centralized action-value function Q_n is updated as:

$$\begin{aligned}\mathcal{L}(\theta_n^Q) &= \mathbb{E}_{o,a,r,o',\hat{k}^{da},\hat{k}^{da\prime}}\left[(y_n - Q_n(o_n, a_n, \hat{k}_n^{da}|\theta_n^Q))^2\right] \\ y_n &= r_n + \gamma Q_n'(o_n', a_n', \hat{k}_n^{da\prime}|\theta_n^{Q\prime})|_{a_n'=\mu_n'(o_n'|\theta_n^{\mu\prime})}\end{aligned} \tag{8}$$

where $\boldsymbol{\mu}' = \{\mu_{1:N}'\}$ is the set of target policies with soft updated parameters $\theta_n^{\mu'}$, $\hat{k}^{da\prime}$ is evaluated by $\boldsymbol{\mu}'$ given all agents' next observations. Note that the centralized Q function is only used during training. During decentralized execution, each policy $\mu_n(o_n|\theta_n^\mu)$ only takes its own observation o_n to produce the action.

Finally, it has to be discussed for this particular DA market applied problem that the DA market efficiency is completely represented by the public information of order books rather than the local observations and actions of all agents, since the market clearing outcomes (trading prices and clearing quantities) are directly calculated based on the order information. As a result, each agent in DA-MADDPG method receiving the market dynamics (i.e., order books) can make more acquainted decisions than the conventional MADDPG method.

5 Experiments and Analysis

5.1 Experiment Setup

Environment. The proposed MARL method is evaluated on a real-world open-source dataset recorded by the Australian distribution utility *Ausgrid* [Ratnam *et al.*, 2017]. We collect the corresponding electricity load and PV generation data of 8 residential households, and form them into 4 prosumers (with PV generation) and 4 consumers (electricity demand only). For each of prosumers and consumers, an ES is installed to provide the system flexibility potentials, where its operating parameters are derived from [Papadaskalopoulos and Strbac, 2016] and presented in Table 1. The grid prices are offered by *Ausgrid*, including the ToU tariff as the grid buy price varying for the time presented in Table 2 and the FiT as the grid sell price fixed at \$0.04/kWh during the whole day.

Baselines. We compare the proposed DA-MADDPG with the conventional ZI strategy [Friedman, 2018], four RL algorithms with three independent methods, e.g., Independent Q-learning (IQL) [Watkins and Dayan, 1992], Independent DDPG (IDDPG) [Lillicrap *et al.*, 2016] and Independent PPO (IPPO) [Schulman *et al.*, 2017], and one state-of-the-art MARL method with centralised critics, e.g., MADDPG [Lowe *et al.*, 2017]. To further evaluate the benefit of P2P energy trading, we benchmark the performance against to the scenario that agents trade independently with the utility company using DDPG method without P2P energy trading (Grid).

Parameter	Value
$\underline{E}^{es}, \overline{E}^{es}$ (kWh)	2,10
η^{esc}/η^{esd}	0.95
$\overline{P}^{es}$ (kW)	2
E_0^{es} (kWh)	$\mathcal{TN}(6, 1^2, 2, 10)$

Table 1: ES operating parameters.

Time	ToU ($/kWh) time	value	FiT ($/kWh)
shoulder	9:00-16:00	0.13	
peak	17:00-20:00	0.18	0.04
off-peak	21:00-8:00 (next day)	0.08	

Table 2: The grid price structure.

Method	Internal (kWh)	External (kWh)	Net bills ($)
Grid	-	441.70	38.32
ZI	31.74	378.23	35.47
DA-MADDPG	65.42	310.86	30.78

Table 3: Sum of internal, external trading quantities and energy bills of 8 agents under Grid, ZI and DA-MADDPG methods.

5.2 Training Performance

The convergence curves of mean episodic total energy bills of 8 agents evaluated by different MARL methods are illustrated in Figure 4. As for baselines, we first show the performance of three independent learning algorithms (IQL, IPPO, IDDPG) with decentralized critics. They fail to learn an optimal policy, where IPPO and IDDPG exhibit high variances which may be due to the non-stationarity issue. Conversely, MADDPG can alleviate the non-stationarity and learn the more effective coordination policies via the centralized critic incorporating with all agents' local observations and actions, but its capability could be limited due to the lack of the DA market information. Instead, we can see that the proposed DA-MADDPG method converges to the lowest energy bills owing to its more sufficient representation of the DA market collective dynamics for the critic.

5.3 Trading Analysis

Having demonstrated the superiority of DA-MADDPG method over the state-of-the-art MARL algorithms, this section aims at analysing: 1) the ES flexibility potentials in reducing peak demands; 2) the energy exchanges in local trading via DA market; and 3) the trading strategies under the dynamic DA-MADDPG method w.r.t. the statistic ZI method.

ES Flexibility. Although Grid, ZI and DA-MADDPG differ with regard to the P2P energy trading (under the latter twos) or not (under the first one), they exhibit some common trends. Compared with the inflexible load (w/o ES) in Figure 5, the net generation during mid-day hours 9-16 and the net demand during peak hours 17-20 of 4 prosumers are both reduced, since the abundant (free) PV generation during mid-day hours is locally absorbed by flexible ES, and the demand at peak hours is shaved through discharging of ES (Figure 6).

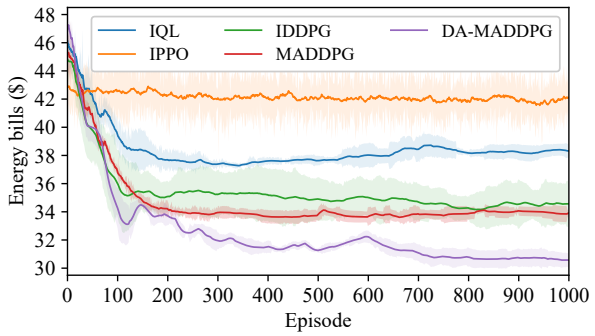


Figure 4: The episodic mean ± std of 8 agents' total energy bills for 10 random runs over 1000 episodes.

On the other hand, 4 consumers cannot make use of free PV generation locally, but still perform the same demand reduction effect during peak hours 17-20, either through charging ES from grid at cheap ToU during off-peak hours 1-8 (Grid, ZI) or buy energy locally from 4 prosumers in DA market during mid-day hours 9-16 (ZI, DA-MADDPG) in Figure 6. In this context, we can observe that agents under all methods are incentived to use their ES flexibility potentials to discharge to supply demand during peak hours through charging free PV generation during mid-day hours or cheap energy during off-peak hours.

Grid vs. DA. When P2P energy trading is allowed in DA market, prosumers/consumers with energy surplus/deficiency are incentived to trade locally among themselves. Thereby, we can observe that compared with the Grid case without P2P energy trading, the generation of 4 prosumers and the demand of 4 consumers during mid-day hours (when both energy surplus and deficiency exist) in Figure 5 are both reduced under ZI and DA-MADDPG methods, since a amount of energy is balanced locally in DA market.

ZI vs. DA-MADDPG. The final comparison is regarding statistic ZI and dynamic DA-MADDPG methods in the same DA market environment. On the one hand, we can observe in Figure 6 that consumers under ZI method charge ES from grid during off-peak hours 1-8 at low ToU and discharge ES to supply demand during peak hours 17-20 when ToU is high. However, consumers under ZI method do not exhibit any charging behavior during mid-day hours when prosumers' PV generation is abundant, this is because the energy schedules of ES are pre-optimized given the grid buy and sell prices in day-ahead planning without considering P2P energy trading. In other words, the quantity availability traded in DA market is the net demand/generation that assumes to trade with the grid in day-ahead (i.e., Grid case). On the other hand, we can observe in Figure 6 that consumers under DA-MADDPG method stop charging ES from grid during off-peak hours 1-8 but start charging ES during mid-day hours 9-16 with PV generation, this is driven by the availability of DA market allowing consumers to buy energy directly from prosumers with PV generation in real-time. Under such method, consumers learn the dynamics of DA market to save energy during off-peak hours and buy energy in DA market during mid-day hours, while prosumers learn to sell energy in DA market instead of selling their energy surplus (PV generation) back to the grid at the unfavorable FiT.

5.4 Economic Benefits

The final section aims at 1) evaluating the agents' economic benefits from DA market and 2) comparing the proposed DA-MADDPG with the benchmark ZI strategy for DA market.

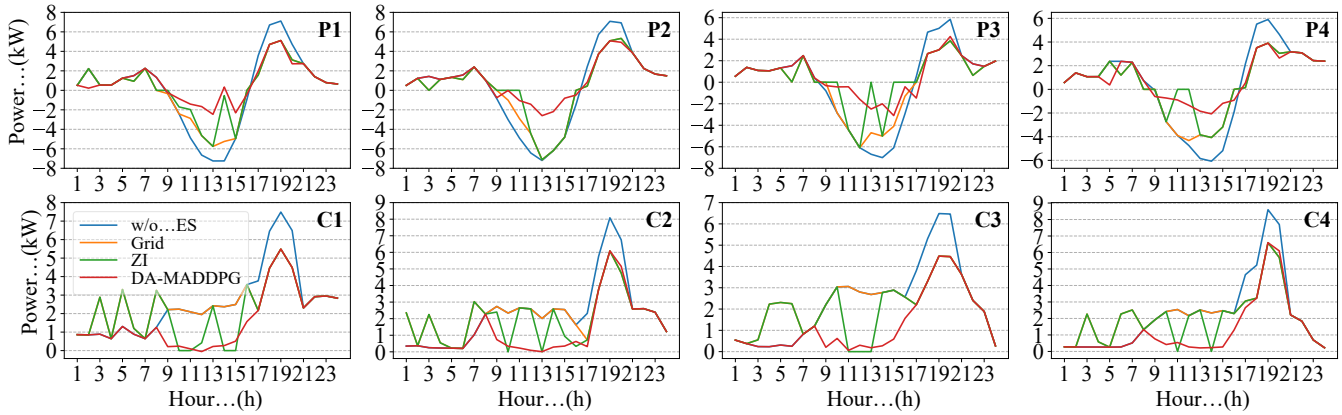


Figure 5: Net load (positive for consumption, negative for generation) of 4 prosumers (P1-4) and 4 consumers (C1-4) under different methods, including for comparison purposes the inflexible load without ES.

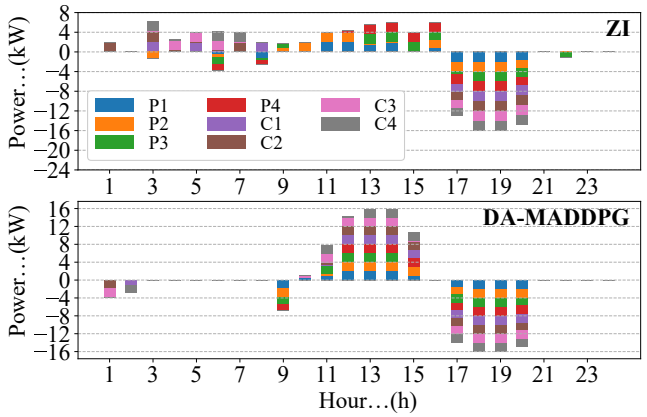


Figure 6: ES charging (positive) and discharging (negative) schedules of 4 prosumers (P1-4) and 4 consumers (C1-4) under ZI and DA-MADDPG methods.

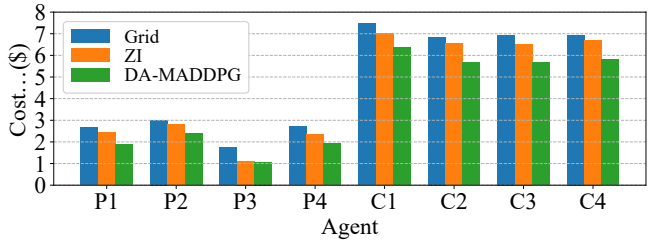


Figure 7: Energy bills of 4 prosumers (P1-4) and 4 consumers (C1-4) under Grid, ZI and DA-MADDPG methods.

It can be observed in Figure 7 that all 8 agents receive the highest energy bills under Grid case without P2P energy trading. After participating in DA market, the energy bills lower down in ZI method. Furthermore, with DA-MADDPG modeling the dynamics of DA market, the energy bills achieve the lowest. The above economic trends can be also validated in Table 3: a) there is no internal trading under Grid case, so the net demand and generation (441.70kWh in total) are bought at high ToU or sold at low FiT with the utility company; b) ZI achieves $2.85 total energy bills saving by 31.74kWh internal trading within DA market; c) DA-MADDPG achieves the least total energy bills by making the highest internal trading at 65.42kWh. In relative terms, DA-MADDPG achieves approximately 17.81% and 29.62% lower external trading with the utility company as well as 13.22% and 16.69% lower total energy bills over ZI and Grid, respectively.

6 Conclusion and Future Work

We introduce a P2P energy trading problem among energy prosumers and consumers evaluated on a highly efficient DA market mechanism. For this specific problem, we propose a novel MARL algorithm namely DA-MADDPG. Based on MADDPG, we construct a representative Q-value function for each agent by abstracting the other agents' observations and actions through the DA market public information. The proposed DA-MADDPG is evaluated on a real-world dataset and compared with five baselines (e.g., ZI, IQL, IDDPG, IPPO, MADDPG). The experimental results demonstrate the benefit of more internal energy exchange among agents through DA market. Consequently, it leads to higher economic benefits in reducing the cost of energy bills.

The future works are in two directions. The first one lies in modelling a more realistic P2P energy trading problem by increasing the population of agents and forms them into a large-scale multi-agent system and further demonstrate the scalability of the proposed DA-MADDPG method under this setting. Secondly, the economic model in this paper neglects the physical operations of electricity systems. To this end, future work will adopt the introduced DA market to an environment with network constrained microgrid.

Acknowledgments

This work was supported by the UK EPSRC project "Integrated Development of Low-Carbon Energy Systems (IDLES): A Whole-System Paradigm for Creating a National Strategy" (project code: EP/R045518/1).

References

[Aitzhan and Svetinovic, 2018] Nurzhan Zhumabekuly Aitzhan and Davor Svetinovic. Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams. *IEEE Transactions on Dependable and Secure Computing*, 15(5):840–852, September – October 2018.

[Alam *et al.*, 2019] Muhammad Raisul Alam, Marc St-Hilaire, and Thomas Kunz. Peer-to-peer energy trading among smart homes. *Applied Energy*, 238:1434–1443, March 2019.

[Friedman, 2018] Daniel Friedman. *The double auction market: institutions, theories, and evidence*. Routledge, March 2018.

[Guerrero *et al.*, 2019] Jaysson Guerrero, Archie C Chapman, and Gregor Verbič. Decentralized p2p energy trading under network constraints in a low-voltage network. *IEEE Transactions on Smart Grid*, 10(5):5163–5173, September 2019.

[Haller *et al.*, 2012] Markus Haller, Sylvie Ludig, and Nico Bauer. Decarbonization scenarios for the eu and mena power system: Considering spatial distribution and short term dynamics of renewable generation. *Energy policy*, 47:282–290, August 2012.

[Jiayi *et al.*, 2008] Huang Jiayi, Jiang Chuanwen, and Xu Rong. A review on distributed energy resources and microgrid. *Renewable and Sustainable Energy Reviews*, 12(9):2472–2483, December 2008.

[Lillicrap *et al.*, 2016] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–14, San Juan, Puerto Rico, May 2016.

[Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6379–6390, Long Beach, CA, USA, December 2017.

[Morstyn *et al.*, 2018] Thomas Morstyn, Niall Farrell, Sarah J Darby, and Malcolm D McCulloch. Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants. *Nature Energy*, 3(2):94–101, February 2018.

[Nicolaisen *et al.*, 2001] James Nicolaisen, Valentin Petrov, and Leigh Tesfatsion. Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. *IEEE Transactions on Evolutionary Computation*, 5(5):504–523, October 2001.

[Papadaskalopoulos and Strbac, 2016] Dimitrios Papadaskalopoulos and Goran Strbac. Nonlinear and randomized pricing for distributed management of flexible loads. *IEEE Transaction on Smart Grid*, 7(2):1137–1146, March 2016.

[Parag and Sovacool, 2016] Yael Parag and Benjamin K Sovacool. Electricity market design for the prosumer era. *Nature energy*, 1(4):1–6, March 2016.

[Pedasingu *et al.*, 2020] Bala Suraj Pedasingu, Easwar Subramanian, Yogesh Bichpuriya, Venkatesh Sarangan, and Nidhisha Mahilong. Bidding strategy for two-sided electricity markets: A reinforcement learning based framework. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 110–119, Virtual Event, Japan, November 2020. Association for Computing Machinery.

[Qiu *et al.*, 2020] Dawei Qiu, Yujian Ye, and Dimitrios Papadaskalopoulos. Exploring the effects of local energy markets on electricity retailers and customers. *Electric Power Systems Research*, 189:106761, December 2020.

[Ratnam *et al.*, 2017] Elizabeth L Ratnam, Steven R Weller, Christopher M Kellett, and Alan T Murray. Residential load and rooftop pv generation: an australian distribution network dataset. *International Journal of Sustainable Energy*, 36(8):787–806, October 2017.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, July 2017.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, December 2008.

[Sun *et al.*, 2015] Yannan Sun, Abhishek Somani, and Thomas E Carroll. Learning based bidding strategy for hvac systems in double auction retail energy markets. In *Proceedings of 2015 American Control Conference (ACC)*, pages 2912–2917, Chicago, IL, USA, July 2015. IEEE.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, Cambridge, Massachusetts, 2018.

[Watkins and Dayan, 1992] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, May 1992.

[Yang *et al.*, 2019] Yaodong Yang, Jianye Hao, Yan Zheng, and Chao Yu. Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 630–636, Macao, China, August 2019.

[Zhang *et al.*, 2018] Dongxia Zhang, Xiaoqing Han, and Chunyu Deng. Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE Journal of Power and Energy Systems*, 4(3):362–370, September 2018.