Deep reinforcement learning-based prosumer aggregation bidding strategy in a hierarchical local electricity market

Deep reinforcement learning-based prosumer aggregation bidding strategy in a hierarchical local electricity market

Haoyang Zhang, Koen Kok, Nikolaos G. Paterakis
Electrical Energy Systems, Eindhoven University of Technology, The Netherlands, h.zhang2@tue.nl

***Abstract*—This paper investigates the application of deep reinforcement learning (DRL) algorithm for the decision-support of a prosumer aggregation in a hierarchical local electricity market (LEM) comprising a peer-to-peer (P2P) market and a corrective market. The agent first submits bids/asks to the P2P market where prosumer aggregations are able to trade electricity directly with each other. After that, the agent participates in the corrective market, where the market operator formulates the corrective market as an AC optimal power flow (OPF) problem to ensure the system is operated within its operational limits. A DRL algorithm, namely Twin Delayed Deep Deterministic Policy Gradient (TD3), is used to find the strategic bidding strategy. The algorithm is tested on a real medium-voltage distribution grid to evaluate the effectiveness of the strategic bidding method. The result of the case study demonstrates that the agent can derive trading strategies to obtain high profits based on the TD3 algorithm.**

***Index Terms*—AC OPF, Deep reinforcement learning, Peer-to-peer market, Strategic bidding**

I. Introduction

Over the last few years, the rapid growth of distributed energy resources (DERs), energy storage systems (ESSs), and information and communication technology (ICT) in distribution systems have motivated passive consumers to become active players [1]. This change also characterizes the transition to more decentralized prosumer-centric market structures such as local electricity markets (LEMs), flexibility markets, and hybrid LEMs and flexibility markets [2]. One typical LEM organizational option is peer-to-peer (P2P) energy markets, which empowers participants to engage in direct energy exchanges with one another and with different stakeholders [3].

Different approaches have been employed in the design of P2P energy markets. Double-sided auction is one of the most widely used market mechanisms based on auction theory that is under intense consideration for LEMs. Double-sided auctions can be classified into two types based on the clearing-time frame, namely discrete-time double auction and continuous-time double auction [4]. In the discrete-time double auction, asks and bids from players are collected over a specific time interval, and the price and quantity offers are then matched simultaneously at the end of the round. As for the continuous-time double auction, bids from buyers and asks from sellers are continuously received and matched [5]. A variety of auction-based LEMs have been presented and compared in [6] considering user preferences and willingness to pay a premium for heterogeneous energy quantities. In [7], an iterative double auction mechanism has been developed and implemented on the blockchain with the objective of achieving maximum social welfare. Nevertheless, the development of P2P markets in distribution systems faces certain challenges that warrant attention. Firstly, the impact of the P2P market on the overall market outcome remains unclear. Secondly, the execution of P2P transactions in the electricity market is subject to governance by grid constraints. Consequently, to effectively manage P2P transactions and address grid constraints, a hierarchical LEM is essential, which comprises both a P2P market and a corrective market. Finally, it is crucial to explore the strategic bidding behavior within this hierarchical LEM to examine the influence of market power.

Since energy trades in the LEMs should be physically executed through the grid, the latter imposes constraints while transporting the energy from the seller to the buyer, which makes P2P energy markets more complex than P2P markets on other types of commodities. To address this challenge, [8] has proposed a hierarchical LEM design by introducing an ancillary service market into the distribution system as a complement to the P2P energy market. Grid usage prices of each P2P transaction are calculated as price signals in such a way that voltage constraints are maintained, congestion is removed, and power losses are balanced.

The primary objective of electricity market liberalization is to foster competition and encourage firms to bid based on marginal costs. However, in practice, the market is not always perfect, and certain phenomena, such as market power, can hinder its efficiency. Market participants with substantial market power possess the ability to sustain prices above a competitive level for extended periods, thereby generating additional profits and leading to an inefficient market result. In the case of market power in hierarchical LEMs, complex strategies can be leveraged by combining different determinants such as prices, volume, product specifications, and grid constraints. To simulate the market power in electricity markets, deep reinforcement learning (DRL) algorithms have been applied in [9]–[12] to solve decision-making problems of agents who bid strategically in the wholesale market or LEM to earn extra profit. Based on the agent-based simulation model using RL algorithms that capture the strategic bidding behavior LEMs, [13] investigated the effect of market concentration on oligopolistic markups and the number of energy suppliers needed to ensure competitive prices. However, most of them neglected the physical operations of electric power systems, which are essential for ensuring the safe operation of the systems. Therefore, it is important to design a hierarchical LEM that effectively incorporates both P2P and corrective markets that takes grids constraints

This publication is part of the research program 'MEGAMIND – Enabling distributed operation of energy infrastructures through Measuring, Gathering, Mining and Integrating grid-edge Data', (partly) financed by the Dutch Research Council (NWO), through the Perspectief funding instrument under number P19-25.

979-8-3503-8106-1/23/$31.00 ©2023 IEEE

2023 Asia Meeting on Environment and Electrical Engineering (EEE-AM) | 979-8-3503-8106-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EEE-AM58328.2023.10395533

into account and to study the strategic bidding behavior that leverages arbitrage opportunities in the hierarchical LEM.

This paper delves into the implementation of DRL algorithms to develop and assess strategic bidding behavior in the hierarchical LEM. Through case studies, it is demonstrated that DRL algorithms can effectively derive bidding strategies in the hierarchical LEM, leading to increased profits. In summary, the key contributions of this paper are as follows:

- A hierarchical LEM framework comprising P2P markets and corrective markets is proposed, where P2P markets allow peers to trade with each other and corrective markets ensure the system is operated under grid constraints.
- A DRL-based strategic bidding method called the Twin Delayed Deep Deterministic Policy Gradient (TD3) is developed to simulate the strategic bidding behavior of the agent on the designed hierarchical LEM.

The remainder of this paper is organized as follows. Section II introduces the market design of the proposed hierarchical LEM in distribution grids. Section III describes the mathematical formulation of the strategic bidding method based on DRL. Case studies and numerical analysis are presented in Section IV. Finally, conclusions and perspectives for future work are drawn in Section V.

II. Hierarchical LEM Framework

In this section, the framework of a hierarchical LEM is introduced. Firstly, the market design of the P2P energy market based on the discrete-time double-sided auction market [11] is described. After that, the centralized corrective market is presented by solving an AC optimal power flow (OPF) problem, where the market operator can dispatch reserved capacities to ensure the P2P transactions can be delivered under the safe operation of the distribution grid.

A. Discrete-time Double-sided Auction Market

The discrete-time double-sided auction market [11] allows buyers and sellers to submit their bids and asks to the auctioneer, and the offers take the form of quantity-price pairs as $o_i^B = (q_i^{bid}, p_i^{bid})$ for buyer $i \in B$ and $o_j^S = (q_j^{ask}, p_j^{ask})$ for seller $j \in S$ within each trading time period $t \in T$. At the end of each trading period, buyers or sellers are not allowed to send new orders anymore and the auctioneer collects the orders from buyers and sellers into the public order book. The bids from buyers are sorted in descending order based on price as $O^B = \{o_1^B, ..., o_i^B, ...o_{N^B}^B\}$ and in ascending order for ask orders as $O^S = \{o_1^S, ..., o_j^S, ...o_{N^S}^S\}$. The market is then cleared in a discriminatory-price policy that charges buyers different prices. An ask and a bid are matched if the ask price is lower than the bid price. The matched quantity equals the minimum quantity between the bid and the ask, and the cleared price is determined using a mid-pricing method according to to [4]. The iterative procedure continues until the lowest ask price is higher than the highest bid price or there is no unmatched quantity in bid or ask. Finally, the market clearing result of the matched quantity and price pair list o^{P2P} can be obtained. The discrete-time double-sided auction market is outlined in Algorithm 1.

Algorithm 1 Discrete-time Double-sided Auction Market

Set $i = j = 1$, $o^{P2P} = \varnothing$
while $p_j^{ask} \leq p_i^{bid}$ and $i \leq N^B$ and $j \leq N^S$ **do**
  Match quantity: $q_{ij}^{P2P} = min(q_i^{bid}, q_j^{ask})$
  Match price: $p_{ij}^{P2P} = (p_i^{bid} + p_j^{ask}))/2$
  Save result: o^{P2P} append $(q_{ij}^{P2P}, p_{ij}^{P2P})$
  Update buyer i: $q_i^{bid} = q_i^{bid} - q_{ij}^{P2P}$
  Update seller j: $q_j^{ask} = q_j^{ask} - q_{ij}^{P2P}$
  if $q_i^{bid} = 0$ **then**
    $i = i + 1$
  if $q_j^{ask} = 0$ **then**
    $j = j + 1$
return Market result o^{P2P}

B. Centralized corrective market based on AC OPF

In this section, a centralized corrective market is designed to ensure the transactions in the P2P market can be executed physically with the objective of minimizing the total operational cost of the system [14]. The formulation of the second-order cone programming (SOCP) relaxations for the AC OPF problem follows the works in [15] and [16]. Accordingly, the objective function of the OPF problem can be formulated as follows:

$$\begin{aligned} \min \quad & \sum_{jt} p_j^P g_{jt}^P + \sum_t (p_t^{ext,in} g_t^{ext,in} - \\ & p_t^{ext,out} g_t^{ext,out}) + \sum_{jt} p_j^{ESS} (e_{jt}^{ch} + e_{jt}^{dis}) \end{aligned} \tag{1}$$

where g_{jt}^P represents the active power from the energy source of seller j at time $t \in \{1, ..., T\}$ and its generation cost is denoted by p_j^P. $p_t^{ext,in}$ and $p_t^{ext,in}$ are the hourly electricity prices at time t in the corrective market for buying and selling, respectively. $g_t^{ext,in}$, and $g_t^{ext,out}$ represent the active power procured from and sold to the external grid. e_{jt}^{ch} and e_{jt}^{dis} denote the power charged and discharged by the ESS of seller j, and p_j^{ESS} is the operational cost of the ESS [17]. Consider a distribution power network $\mathcal{DG} = (\mathcal{B}, \mathcal{L})$, where $\mathcal{B}$ denotes the set of buses and $\mathcal{L}$ denotes the set of lines, the constraints related to the grid are formulated as follows:

$$g_j^{min} \leq g_{jt}^P \leq g_j^{max}, j \in \mathcal{J}, t \in \mathcal{T} \tag{2}$$

$$q_j^{min} \leq q_{jt}^P \leq q_j^{max}, j \in \mathcal{J}, t \in \mathcal{T} \tag{3}$$

$$g^{ext,min} \leq g_t^{ext,in} - g_t^{ext,out} \leq g^{ext,max}, t \in \mathcal{T} \tag{4}$$

$$q^{ext,min} \leq q_t^{ext,in} - q_t^{ext,out} \leq q^{ext,max}, t \in \mathcal{T} \tag{5}$$

Constraints (2) and (3) are active and reactive power generation limits of the energy sources, and (4) and (5) denote the active power and reactive power limit of the transformer at the slack bus. The constraints related to the ESSs are expressed as follows:

$$0 \leq e_{jt}^{ch} \leq e_j^{max}, j \in \mathcal{J}, t \in \mathcal{T} \tag{6}$$

$$0 \leq e_{jt}^{dis} \leq e_j^{max}, j \in \mathcal{J}, t \in \mathcal{T} \tag{7}$$

$$SOE_{jt} = SOE_{jt-1} + (\eta_j^{ch} e_{jt}^{ch} - e_{jt}^{ch}/\eta_j^{dis})\Delta t, \quad j \in \mathcal{J}, t \in \mathcal{T} \tag{8}$$

$$SOE_j^{min} \leq SOE_{jt} \leq SOE_j^{max}, j \in \mathcal{J}, t \in \mathcal{T} \tag{9}$$

$$SOE_{j0} = SOE_{jT}, j \in \mathcal{J} \tag{10}$$

where (6) and (7) represent the charging and discharging power limit of the ESS. η_j^{ch} and η_j^{dis} are ESS charging and discharging efficiency, while the state of energy (SOE)

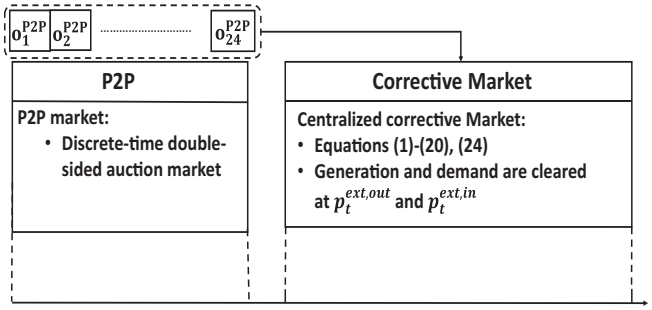

Fig. 1: Process of the hierarchical LEM

is updated using (8) and limited by the minimum and maximum SOE SOE_j^{min} and SOE_j^{max}. It is assumed that the SOE at the end of the day should be the same as the beginning of the day as shown in (10). The operational constraints of the grid are formulated as:

$$g_{nt}^P - g_{nt}^D + e_{nt}^{dis} - e_{nt}^{ch} = G_{nn}c_{nnt} + \sum_{k\in\delta(n)} (G_{nk}c_{nkt} - B_{nk}w_{nkt}) \ ,n \in \mathcal{B}, t \in \mathcal{T} \tag{11}$$

$$q_{nt}^P - q_{nt}^D = -B_{nn}c_{nnt} + \sum_{k\in\delta(n)} (-B_{nk}c_{nkt} - G_{nk}w_{nkt}) \ , n \in \mathcal{B}, t \in \mathcal{T} \tag{12}$$

$$g_{nkt}^L = -G_{nk}c_{nnt} + G_{nk}c_{nkt} - B_{nk}w_{nkt} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{13}$$

$$q_{nkt}^L = B_{nk}c_{nnt} - B_{nk}c_{nkt} - G_{nk}w_{nkt} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{14}$$

$$l_{nkt} = (B_{nk}^2 + G_{nk}^2)(c_{nnt} - 2c_{nkt} + c_{kkt}) \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{15}$$

$$0 \le l_{nkt} \le l_{nkt}^{max} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{16}$$

$$\underline{U}_n^2 \le c_{nnt} \le \bar{U}_n^2 \ , n \in \mathcal{B}, t \in \mathcal{T} \tag{17}$$

$$c_{nkt} = c_{knt} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{18}$$

$$w_{nkt} = -w_{knt} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{19}$$

$$c_{nkt}^2 + w_{nkt}^2 \le c_{nnt}c_{kkt} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{20}$$

where $\delta(n)$ denotes the buses connected with bus n. The variables c_{nnt}, c_{nkt}, and w_{nkt} for each bus $n \in \mathcal{B}$ and each branch $(nk) \in \mathcal{L}$ are defined as:

$$c_{nnt} = (U_{nt}^{real})^2 + (U_{nt}^{im})^2 \ , n \in \mathcal{B}, t \in \mathcal{T} \tag{21}$$

$$c_{nkt} = U_{nt}^{real}U_{kt}^{real} + U_{nt}^{im}U_{kt}^{im} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{22}$$

$$w_{nkt} = U_{nt}^{real}U_{kt}^{im} - U_{kt}^{real}U_{nt}^{im} \ , (nk) \in \mathcal{L}, t \in \mathcal{T} \tag{23}$$

where U_{nt}^{real} and U_{nt}^{im} are the real and imaginary parts of the voltage at bus n. Given the information above, it can be seen (11) and (12) correspond to the conservation of active and reactive power flows at each bus. Constraints (13) and (14) are used to derive the active power flow g_{nkt}^L and reactive power flow q_{nkt}^L on branch $(nk) \in \mathcal{L}$. Constraint (16) and (17) restrict the squared magnitude of the bus voltage and the branch current. Since the cosine function is even and the sine function is odd, (18) and (19) are also considered.

C. Hierarchical LEM

The framework of the hierarchical LEM comprising P2P and corrective markets is summarized in Fig. 1. Prior to the day, each participant determines their bids or asks for each hour in the P2P market. When the P2P market is cleared at the end of the interval, the market result for each hour o_t^{P2P} is sent into the corrective market. For simplification the bid quantity q_i^{bid} for each buyer i is the energy consumed by the inflexible loads g_{it}^D, and the ask quantity q_j^{ask} for each seller j equals to the generation capacity minus its local inflexible loads $g_{jt}^G - g_{jt}^D$. The ESSs are used to provide reserve services in the corrective market to guarantee the P2P offers can be executed and delivered physically as shown in (24):

$$\sum_i q_{ij}^{P2P} \le g_{jt}^P - g_{jt}^D + e_{jt}^{dis} - e_{jt}^{ch} \ , j \in \mathcal{J}, t \in \mathcal{T} \tag{24}$$

III. Deep Reinforcement Learning Algorithm

In this section, a model-free DRL namely TD3 is introduced to simulate the strategic behavior of the agent by interacting with the hierarchical LEM.

A. Reinforcement Learning

The above-mentioned electricity trading problem can be formulated as a Markov decision process (MDP) and be solved effectively by RL algorithms. In contrast to Zero-Intelligence Agents (ZI) where the seller decides a set of prices of their asks in the P2P market randomly, the seller with RL method interacts with the hierarchical LEM at each time step t by choosing an action $a_t \in \mathcal{A}$ based on the observation $o_t \in \mathcal{O}$ of the current state $s_t \in \mathcal{S}$ under policy $\pi(s_t) = a_t$. In this paper, the action a_t of the seller is the price factor λ_t for each day, and the state $s_t = [g_{1:N_t}^D, p_{1:N_t}^{ext,in}] \in \mathcal{S}$ is a vector comprising the total demand of the distribution system and the electricity selling prices in the corrective market for each hour of the day. For the sake of simplicity, it is assumed the seller can perfectly predict the total demand and electricity price in the corrective market. By taking an action, the agent receives a profit (or reward) $r_t \in \mathcal{R}$ which equals the income in the P2P market plus income in the corrective market minus the total operational cost, as a function of state and action $\mathcal{S} \times \mathcal{A} \to \mathcal{R}$ and finds itself in a new state s_{t+1} under a probability $p(s_{t+1}, r_t|s_t, a)$ called dynamics of the MDP. The goal of the agent is to maximize its cumulative discounted profit (or return) $G_t = \sum_{i=1}^T \gamma^{i-1} r_{t+i}$, where $\gamma \in [0, 1]$ is the discounted factor to balance the profit between now and future. The action-value function (or Q-value function) $Q_\pi(s_t, a_t) = \mathbf{E}_\pi(G_t|s_t, a_t)$ denotes the expected discounted profit starting from s_t and taking the action a_t under bidding policy π. Also, an optimal bidding policy can be derived from the optimal Q-values $Q_*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t)$ by selecting the action corresponding to the highest Q-value in each state. The Q-value function can be expressed in a recursive format according to the Bellman equation as follows:

$$Q_\pi(s_t, a_t) = \mathbf{E}(r_t + \gamma Q_\pi(s_{t+1}, a_{t+1})) \tag{25}$$

As a result, the Q-value function can be updated by bootstrapping method, and the temporal difference (TD) method can be applied as follows:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \tag{26}$$

where $\alpha \in [0, 1]$ is the step size [18].

B. Twin Delayed DDPG

The Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [19] is an actor-critic DRL algorithm using advanced deep neural networks to approximate the Q-value and to select optimal bidding actions, improved from the Deep Deterministic Policy Gradient (DDPG) [20]. TD3 and DDPG are actor-critic methods with both a critic neural network $Q(s, a|\theta^Q)$ that determines the Q-value and an actor neural network $\mu(s|\theta^\mu) = \text{argmax}_a\, Q(s, a)$ that determines the greedy bidding action to take, which also extends the finite discrete action space of Deep Q Learning (DQN) into infinite continuous action space. Apart from the regular critic and actor networks shown above, there are also target critic networks $Q'(s, a|\theta^{Q'})$ and target actor networks $\mu'(s|\theta^{\mu'})$ which have the same network structure as the regular networks and are copied over from the regular network every fixed number of steps (soft update) to stabilize the learning process as follows:

$$\theta^{Q'} = \tau\theta^Q + (1-\tau)\theta^{Q'} \tag{27}$$

$$\theta^{\mu'} = \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{28}$$

where τ is a learning rate close to 1. A reply buffer $\mathcal{R}$ of a size $N^{\mathcal{R}}$ is created to store the training data in the form of (s_t, a_t, r_t, s_{t+1}). When training the networks, a mini-batch of size N^m is sampled uniformly from the reply buffer, and the agent updates the critic network by minimizing the Smooth L1 Loss function $L(\theta^Q)$ defined as:

$$L(\theta^Q) = \begin{cases} \frac{1}{N^m}\sum_{n=1}^{N^m} 0.5\zeta_n^2 & \text{if } |\zeta_n| \le 1, \\ \frac{1}{N^m}\sum_{n=1}^{N^m}(|\zeta_n| - 0.5) & \text{otherwise,} \end{cases} \tag{29}$$

$$\zeta_n = r_t + \gamma Q'(s_{t+1}, a_{t+1}|\theta^{Q'}) - Q(s_t, a_t|\theta^Q) \tag{30}$$

The objective is to minimize the TD error ζ_n defined as the difference of the estimated Q-value between target networks and regular networks. The actor network is updated with the objective of maximizing the expected profit return as follows:

$$J(\theta^\mu) = \mathbf{E}(Q(s, a(\theta^\mu)|\theta^Q) \tag{31}$$

$$\nabla_{\theta^\mu} J(\theta^\mu) = \frac{1}{N^m}\sum_{n=1}^{N^m}(\nabla_a Q(s_n, a_n(\theta^\mu)|\theta^Q) \nabla_{\theta^\mu} a_n(\theta^\mu)) \tag{32}$$

TD3 trains a deterministic policy in an off-policy way, and Gaussian noise $\mathcal{N}_t^\epsilon$ is added to the action for exploring the environment and finding the global optimum as follows:

$$a_t = clip(\mu(s_t|\theta^\mu) + \mathcal{N}^\epsilon(0, \sigma^\epsilon), a^{min}, a^{max}) \tag{33}$$

where σ^ϵ is the standard deviation of the Gaussian noise. In order to reduce actions exploration over time and to help convergence, σ^ϵ is decaying linearly with episodes by multiplying with a decaying factor κ at the end of each episode.

Compared with the DDPG algorithm, the TD3 algorithm improves in three aspects. Firstly, two regular Q-value networks and corresponding target Q-value networks are established to compute the value of the same next state and the minimum of the two values is selected as the target Q-value to offset the overestimation of Q-value to compute the TD error ζ_n for each regular Q-value network as follows:

$$\zeta_n = r_t + \gamma \min(Q'(s_{t+1}, a_{t+1}|\theta_1^{Q'}), Q'(s_{t+1}, a_{t+1}|\theta_2^{Q'})) - Q(s_t, a_t|\theta^Q) \tag{34}$$

The second improvement of TD3, namely delayed target networks and policy update, updates the target networks and policy network with a lower frequency than the value network in order to provide a stable objective with smaller variance and allow better coverage of the training data. Thirdly, a concern with deterministic policies is they can overfit to narrow peaks in the value estimate. As a result, TD3 uses a so-called target policy smoothing regularization method by fitting the value of a small area around the target action a_{t+1} as follows:

$$a_{t+1} = clip(\mu'(s_{t+1}|\theta^{\mu'}) + \epsilon', a^{min}, a^{max}) \tag{35}$$

$$\epsilon' = clip(\mathcal{N}^{\epsilon'}(0, \sigma^{\epsilon'}), \epsilon^{min}, \epsilon^{max}) \tag{36}$$

where $\mathcal{N}^{\epsilon'}(0, \sigma^{\epsilon'})$ represents the Gaussian noise. Moreover, a simple warm-up method is applied for the first N^w episodes: the actions are randomly selected as a ZI agent to interact with the environment, and the tuples (s_t, a_t, r_t, s_{t+1}) are stored in the reply buffer. The objective is to have more valued and informed learning samples in the experience pool, before offline training, so as to give guidance on the exploration in offline training [21].

Algorithm 2 TD3 Algorithm

```
Set hyperparameters and reply buffer R
Randomly initialize regular (target) critic and actor networks with θ_1^Q (θ_1^Q'), θ_2^Q (θ_2^Q') and θ^μ (θ^μ')
while episode ≤ N^m do
    t = 1 and agent observe the state s_t
    while episode ≤ N^w and t ≤ T do
        Agent observes the state s_t
        Take uniform-random action a_t
        Obtain reward r_t and observe next state s_{t+1}
        Store (s_t, a_t, r_t, s_{t+1}) in R
        s_t ← s_{t+1}
        t ← t + 1
    while N^w < episode and t ≤ T do
        Agent observes the state s_t
        Execute action a_t
        Obtain reward r_t and observe next state s_{t+1}
        Store (s_t, a_t, r_t, s_{t+1}) in R
        Update critic network by (29)
        if episode mod d then
            Update actor network by (32)
            Update target networks by (27) and (28)
        s_t ← s_{t+1}
        t ← t + 1
    σ ← κσ
    episode ← episode + 1
```

IV. Case Study

A. Test Environment and Simulation Setting

The test system is a 10 kV 31-bus MV distribution network of a Dutch DSO as shown in Fig. 2. Bus 1 is defined as the slack bus of the system and the base for power is 10 MVA. The switches at branch $(9, 10)$ and $(19, 20)$ are disconnected

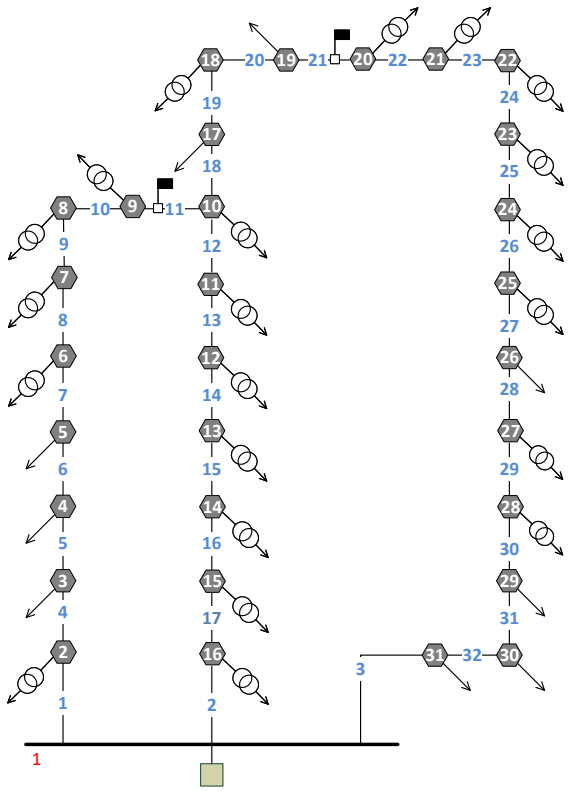

Fig. 2: Test system

so that the system is operated in a radial configuration. [22]. There are 8 sellers (prosumer aggregation as an energy community comprising one energy source, one ESS, and inflexible loads) and 22 buyers (consumer aggregation with inflexible loads) in the system as shown in Table I and II. The time duration of the case study is 7 days with hourly-resolution data. The hourly prices were collected from ENTSO-E for the period from 1^{st} December 2021 to 7^{th} December 2021 in the Netherlands [23]. Based on the work in [24], a price-spread factor with 0.9 is be selected ($p^{ext,out} = 0.9p^{ext,in}$). The operational cost p^{ESS} of the ESS is 0.1 €$/MWh$, and the ESS charging efficiency η_j^{ch} and discharging efficiency η_j^{dis} are 0.9. Seller 1 is selected to be the strategic bidder while the rest of the participants are assumed to behave competitively by revealing their true characteristics to the market. The state vector is composed of the hourly aggregated loads and the hourly prices as $s = [g^D_{1:24}, p^{ext,in}_{1:24}] \in \mathcal{S}$, and the action a is the price factor λ between $\lambda^{min} = 0$ and $\lambda^{max} = 2$ based on the work in [10]. The reward r is defined as the profit of Seller 1 for each day.

TABLE I: Parameters of sellers

Seller j	1	2	3	4	5	6	7	8
Location (bus)	3	4	5	17	19	26	29	31
p_j^P(€/MWh)	104.7	46.6	92.9	184.8	293.9	338.7	253.8	241.7
g_j^{min}(MW)	0	0	0	0	0	0	0	0
g_j^{max}(MW)	0.73	0.67	0.93	0.73	0.87	0.87	0.73	0.73
q_j^{min}(MVar)	-0.44	-0.4	-0.56	-0.44	-0.52	-0.52	-0.44	-0.44
q_j^{max}(MVar)	0.44	0.4	0.56	0.44	0.52	0.52	0.44	0.44
e_j^{max}(MW)	0.59	0.53	0.75	0.59	0.69	0.69	0.59	0.59
SOE_j^{min}(MWh)	0.22	0.2	0.28	0.22	0.26	0.26	0.22	0.22
SOE_j^{max}(MWh)	1.1	1	1.4	1.1	1.3	1.3	1.1	1.1
$SOE_{j,0}$(MWh)	0.55	0.50	0.70	0.55	0.65	0.65	0.55	0.55

TABLE II: Parameters of buyers

Buyer i	1	2	3	4	5	6	7	8
Location (bus)	2	6	7	8	9	10	11	12
p_i^{bid}(€/MWh)	286.5	299.6	232.3	223.9	279.2	199.4	230.0	238.5
Buyer i	9	10	11	12	13	14	15	16
Location (bus)	13	14	15	16	18	20	21	22
p_i^{bid}(€/MWh)	266.6	196.6	182.7	182.4	283.9	200.6	181.6	275.1
Buyer i	17	18	19	20	21	22	-	-
Location (bus)	23	24	25	27	28	30	-	-
p_i^{bid}(€/MWh)	267.2	286.8	268.2	200.8	201.8	285.5	-	-

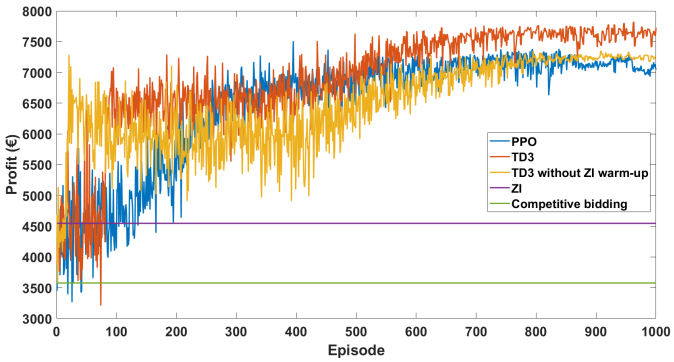

Fig. 3: Agent's performance in the hierarchical LEM

B. Simulation Results

Seller 1 interacts with the market for 1000 episodes by DRL algorithms TD3, TD3 without ZI warm-up, and Proximal Policy Optimization (PPO). Figure 3 shows the performance of the DRL techniques in the hierarchical LEM by comparing with the case where the agent bids with the ZI method and competitive behavior by bidding true marginal costs. It can be seen that the TD3 gradually finds the optimal solution and converges to the most profitable solution which is higher than the PPO method and the TD3 without ZI warm-up, and is significantly higher than the ZI method and the competitive behavior. More specifically, the profit of Seller 1 reaches 7730 € by TD3 at the final episode in the hierarchical LEM with auction-based P2P and corrective markets, which is 8.6% higher than the PPO (7117 €), 6.5% higher than the TD3 without ZI warm-up (7259 €), 70.0% higher than ZI (4547 €), and 116% higher than competitive bidding method (3577 €).

Figure 4 shows the action and income of the TD3 agent in the hierarchical LEM in the last episode. It can be seen that the agent selects price factors (λ) between 1.7 and 1.75 to increase the clearing price $p_{i,j}^{P2P}$ and obtain high income in the auction-based P2P market for most of the time. When the average electricity price in the corrective market goes higher on Saturday, the agent takes a high price factor (close to the upper limit 2) in order not to be matched in the auction-based P2P market and participate in the corrective market to gain a higher profit.

Hyperparameters of the DRLs play an important role in their performances. The standard deviation σ^ϵ of the Gaussian noise in (33) and the entropy coefficient are important hyperparameters of the TD3 and the PPO, respectively, which are used to control the exploration and exploitation trade-off between immediate and future profits. Figure 5 presents a sensitivity analysis of the impact of the hyperparameters on the performance of the DRLs in the market, where the

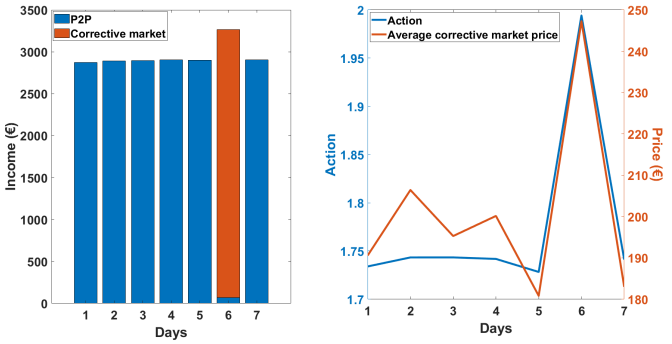

Fig. 4: Agent's income and action in the hierarchical LEM

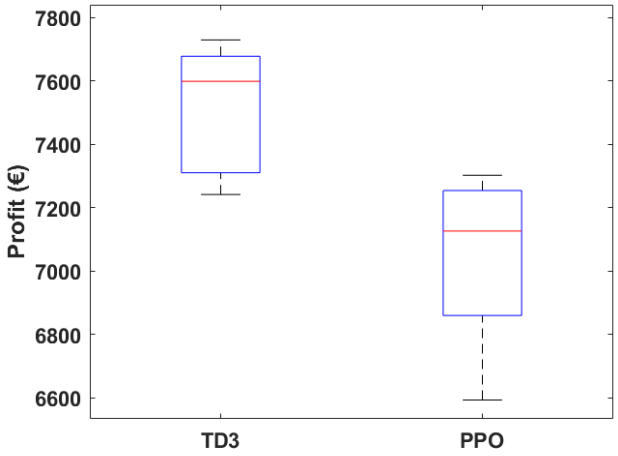

Fig. 5: Sensitivity analysis of the DRL algorithms

standard deviation σ^{ϵ} of TD3 ranges from 0.1 to 1 and the entropy coefficient of PPO ranges from 0.001 to 0.01 in the box charts. It can be seen both TD3 and PPO are sensitive to the hyperparameters related to exploration, where the profits of TD3 and PPO range from 7242 € to 7729 € and from 6594 € to 7256 €, respectively. Even though both TD3 and PPO select sub-optimal actions with improper hyperparameters, TD3 has a higher chance to find a better result than PPO since the lowest profit of TD3 (7242 €) is almost the same with the highest profit of PPO (7256 €).

V. Conclusion and Outlook

In this paper, a hierarchical LEM framework comprising the P2P and corrective markets together was proposed. In addition, a strategic bidding method based on the DRL algorithm namely TD3 was tested on the proposed hierarchical LEM. The result of the case study demonstrated that the proposed strategic bidding method based on TD3 can achieve significantly higher profits than the ZI method and competitive bidding method in the hierarchical LEM. Compared with the PPO, the TD3 has achieved approximately 8.6% higher profit in the hierarchical LEM by participating in the P2P market and the corrective market flexibly according to the demands and the price signals. Also, simulation results indicated that the TD3 can achieve a higher profit by using the ZI warm-up method. However, there are still limitations in this work that can be improved in the next stage. Firstly, the test case is a radial distribution grid and phase angle constraint in a loop is not considered. In addition, only one strategic bidder is considered in this work. According to the limitations above, future work can be carried out as follows. Firstly, the method can be tested on a meshed distribution system. Also, multi-agent DRL algorithms should be implemented to simulate the interaction of the market participants.

References

[1] M. Khorasany, Y. Mishra, and G. Ledwich, "A decentralized bilateral energy trading system for peer-to-peer electricity markets," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 6, pp. 4646–4657, 2020.

[2] G. Tsaousoglou, J. S. Giraldo, and N. G. Paterakis, "Market mechanisms for local electricity markets: A review of models, solution concepts and algorithmic techniques," *Renewable and Sustainable Energy Reviews*, vol. 156, p. 111890, 2022.

[3] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-peer trading in electricity networks: An overview," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3185–3200, 2020.

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

[5] X. Vilajosana, D. Lázaro, J. M. Marquès, and A. A. Juan, "Dymra: A decentralized resource allocation framework for collaborative learning environments," in *Intelligent Collaborative e-Learning Systems and Applications*. Springer, 2009, pp. 147–169.

[6] M. Zade, S. D. Lumpp, P. Tzscheutschler, and U. Wagner, "Satisfying user preferences in community-based local energy markets—auction-based clearing approaches," *Applied Energy*, vol. 306, p. 118004, 2022.

[7] C. Zhang, T. Yang, and Y. Wang, "Peer-to-peer energy trading in a microgrid based on iterative double auction and blockchain," *Sustainable Energy, Grids and Networks*, vol. 27, p. 100524, 2021.

[8] K. Zhang, S. Troitzsch, S. Hanif, and T. Hamacher, "Coordinated market design for peer-to-peer energy trade and ancillary services in distribution grids," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 2929–2941, 2020.

[9] H. Zang and J. Kim, "Reinforcement learning based peer-to-peer energy trade management using community energy storage in local energy market," *Energies*, vol. 14, no. 14, p. 4131, 2021.

[10] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep reinforcement learning for strategic bidding in electricity markets," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343–1355, 2019.

[11] D. Qiu, J. Wang, J. Wang, and G. Strbac, "Multi-agent reinforcement learning for automated peer-to-peer energy trading in double-side auction market," *IJCAI*, pp. 2913–2920, 2021.

[12] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, p. 100413, 2021.

[13] P. Staudt, J. Gärttner, and C. Weinhardt, "Assessment of market power in local electricity markets with regards to competition and tacit collusion," *Tagungsband Multikonferenz Wirtschaftsinformatik*, vol. 2018, pp. 912–923, 2018.

[14] G. De Zotti, S. A. Pourmousavi, J. M. Morales, H. Madsen, and N. K. Poulsen, "A control-based method to meet tso and dso ancillary services needs by flexible end-users," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1868–1880, 2019.

[15] B. Kocuk, S. S. Dey, and X. A. Sun, "Strong socp relaxations for the optimal power flow problem," *Operations Research*, vol. 64, no. 6, pp. 1177–1196, 2016.

[16] A. G. Expósito and E. R. Ramos, "Reliable load flow technique for radial distribution networks," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 1063–1069, 1999.

[17] T. Chen and S. Bu, "Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning," in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. IEEE, 2019, pp. 1–5.

[18] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[19] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[20] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," pp. 387–395, 2014.

[21] A. Stooke and P. Abbeel, "Accelerated methods for deep reinforcement learning," *arXiv preprint arXiv:1803.02811*, 2018.

[22] M. Grond, "Computational capacity planning in medium voltage distribution networks," *Eindhoven University of Technology*, 2016.

[23] ENTSO-E, "day-ahead prices," https://transparency.entsoe.eu/, 2022, accessed: 7 April 2022.

[24] I. Dukovska, H. J. Slootweg, and N. G. Paterakis, "Decentralized optimization and power flow analysis for a local energy community," pp. 1–6, 2021.