Peer-to-Peer Energy Trading for Multi-microgrids via Stackelberg Game and Multi-agent Deep Reinforcement Learning

Pengjie Zhao, Junyong Wu, *Senior Member*, *IEEE*, *Senior Member*, *CSEE*, Fashun Shi, Lusu Li, Baoqing Li, and Yi Wang

***Abstract*—This paper proposes a novel framework based on the Stackelberg game and deep reinforcement learning for multi-microgrids (MGs) in achieving peer-to-peer (P2P) energy trading. A multi-leaders, multi-followers Stackelberg game is utilized to model the P2P energy trading process. Stackelberg equilibrium (SE) is regarded as a P2P optimal trading strategy. A two-stage privacy protection solution technique combining data-driven and model-driven is developed to obtain the SE. Specifically, energy storage scheduling problem in MGs is formulated as a Markov decision process with discrete periods, and a multi-action single-observation deep deterministic policy gradient (MASO-DDPG) algorithm is proposed to tackle optimal scheduling of energy storage in the first stage. According to optimal scheduling of energy storage, the closed-form expression for SE based on model-driven is derived, and distributed SE solution technique (DSET) is developed to obtain SE in the second stage. Case studies involving a 4-Microgrid demonstrate the P2P electricity price obtained by the two-stage method, as a novel pricing mechanism, can reasonably regulate microgrid operation mode and improve microgrid income participating in the P2P market, which verifies effectiveness and superiority of the proposed P2P energy trading model and two-stage solution method.**

***Index Terms*—Deep reinforcement learning, markov decision process, microgrid, peer-to-peer (P2P), stackelberg equilibrium.**

I. Introduction

To deal with the crisis of energy and environment, renewable energy resources such as wind and solar energy earn widespread attention. However, large integration of renewable energy resources has negative impacts on security and stability of power systems [1]. Microgrid (MG) technology is becoming an effective way to accommodate penetration of distributed energy resources (DERs) [2]. Meanwhile, intermittent and uncertainty of integrated DERs bring significant challenges to reliable and economic operation of MG. Participating in energy trading based on feed-in-tariff and peer to grid (P2G) schemes can ensure economic operation of MG [3]. In [4], [5], the authors proposed a transactive energy market for multi-MGs based on centralized optimization methods, in which a market operator directly optimizes DERs in the MG. Unfortunately, the benefit of MG to be involved in recent feed-in-tariff schemes has been marginal, which has affected the enthusiasm to participate in energy trading and further affected penetration of DERs [6].

There are also many references focusing on the transactive energy market, and building a peer-to-peer (P2P) trading energy market [7], [8]. In a P2P market, MGs communicate directly with each other and actively participate in the energy market either by selling their excess energy or by reducing energy demand [9]. The challenge of P2P energy trading is to model the energy trading process and determine rational trading principles when there are many entities in the energy trading market.

Gaming methods can be identified as important technology to the design of the P2P energy trading market, among which the Stackelberg game is the most widely used. In [9]–[11], the authors studied application of a Stackelberg game in a multi-MGs P2P energy trading market and existence of Stackelberg equilibrium (SE). In [12], [13], the authors studied a Stackelberg game model considering DERs, but microturbine (MT) was not considered. In [14], [15], the authors studied a Stackelberg game model considering DERs, MT, and demand response. In addition, heterogeneous DERs like batteries are ignored in the above literature.

In both the above gaming literature, model-driven energy modeling approaches and gaming methods are adopted in order to optimize energy trading of MGs. Applying model-free deep reinforcement learning (DRL) to make decisions has achieved remarkable success in recent years [16]. The characteristic of the DRL is the agent responsible for making decisions selects the best action (i.e., policy) by interacting with the environment sequentially and using the reward and observation from environmental feedback. Besides, the DRL takes into consideration expectation of future cumulative reward, not only immediate reward at the current time period. Especially when there are time-coupling variables in the environment, DRL decision-making has more advantages. For example, if the only immediate reward is considered in a single period, energy storage will not play the role of "arbitrage" and "peak cutting and valley filling".

Applying model-free DRL to optimize energy management

Manuscript received January 28, 2022; revised March 23, 2022; accepted May 12, 2022. Date of online publication June 27, 2023; date of current version August 17, 2024. This work was supported in part by the Fundamental Research Funds for the Central Universities (No. 2020YJS162).

P. J. Zhao, J. Y. Wu (corresponding author, email: wujy@bjtu.edu.cn), F. S. Shi, L. S. Li, B. Q. Li, and Y. Wang are with the Department of Electrical Engineering, Beijing Jiaotong University, Beijing 100084, China.

DOI: 10.17775/CSEEJPES.2022.00680

decisions has attracted growing interest [17], [18]. Literature converges so far on two main categories of approaches. The first category focuses on value-based function algorithm represented by Deep Q-Network (DQN), such as MG energy management [19], demand response [20], [21], building energy optimization [22], and battery energy storage system energy management [23]. There are some practical bottlenecks in the DQN algorithm: discretization of the action space for applications with continuous action variables leads to suboptimal policy. The second category solves the above bottlenecks and the deep deterministic policy gradient (DDPG) [24], asynchronous advantage actor-critic [25], proximal policy optimization [26] successfully applied with high-dimension and continuous action space.

Significant progress has been made in energy management optimization employing single-agent DRL (SADRL). Driven by success of the SADRL, multi-agent DRL (MADRL) method is emerging, but it is still relatively thin. Centralized training with decentralized execution (CTDE) framework is a major feature of MADRL [27]–[29], which eliminates non-stationarity of the DRL environment. A multi-agent deep Q-network (MADQN) framework is applied to generate independent MG market decisions and to keep balance of benefits [27]. MA-state-action-reward-state-action (MA-SARSA) [28] and multi-agent actor-critic (MAAC) [29] algorithms are proposed to realize home energy management, respectively.

This paper addresses the challenging multi-MG P2P energy trading market by combining model-driven and data-driven methods to develop a privacy-preserving two-stage method. Concretely, contributions of this paper are fourfold.

1) This paper considers multiple types of equipment in MG, including renewable DERs(e.g., solar PVs and WTs), micro-turbine, energy storage (ES) system, and price-based demand response, and fully considers energy regulation potential within the microgrid.

2) A multi-leaders, multi-followers (MLMF) Stackelberg game approach is utilized to model the P2P energy trading process among MGs. Closed-form expression for SE is proved.

3) An improved reinforcement learning algorithm based on multi-actions, single-observation DDPG (MASO-DDPG) is proposed, which is used to realize optimal scheduling of ES in the first stage. Inheriting the CTDE paradigm, MASO-DDPG protects privacy of the microgrid during centralized training. Furthermore, a distributed Stackelberg equilibrium (i.e., trading price and trading power) solution technique (DSET) is proposed in the second stage.

4) Case studies demonstrate superiority of the proposed DSET in protecting privacy and the proposed MASO-DDPG significantly outperforms other algorithms: MADDPG, DDPG, dueling DQN (DDQN). The obtained P2P energy trading scheme based on integration of data-driven and model-driven approaches effectively improves benefit of MG and stimulates enthusiasm to participate in the P2P energy trading market.

The rest of the paper is managed as follows: Section II provides detailed model-based mathematical formulation of MG. Section III details the P2P energy trading process based on the Stackelberg game. Section IV introduces the Markov game formulation of the multi-agent problem. Section V provides a two-stage privacy-preserving P2P energy trading solution technique combining MASO-DDPG based on data-driven and DSET based on the model-driven method. Section VI conducts numerical experiments. Finally, Section VII discusses conclusions and future extensions of this work.

II. Formulation Of Model-Based Peer To Peer Energy Trading In Multi-MGs

In this section, we introduce the structure of MG and the mathematical model that participates in P2P energy trading market. Structure of the microgrid that participates in the P2P energy trading market is illustrated in Fig. 1. Microgrids are included in a cluster, which can be configured with diverse DERs (solar PVs, WTs, and MTs), ES systems, and price-based demand response (DR). Not every microgrid contains all the above equipment. In the P2P energy trading market, all microgrids are connected with one another: if generated and stored energy is insufficient then it can buy energy from neighboring microgrids and the main grid, and other microgrids with excess energy can sell energy to neighboring microgrids and main grid. Meantime, P2P operators are used to collecting information of all producers and consumers in the market, which determine and provide a reasonable P2P electricity price for participants to conduct transactions and settle transaction results. In this case, the major function of the P2P operator is to determine price and amount of energy trading between producers and consumers.

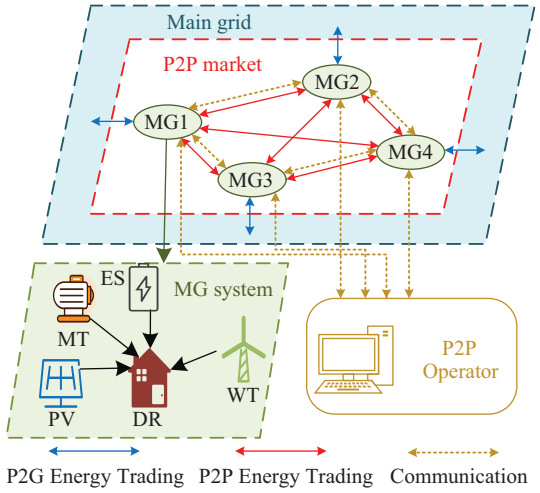

Fig. 1. The physical structure of P2P and P2G energy trading.

A. Energy Storage

At time period t, the relationship between state of charge $S_{\mathrm{soc},t}$ of ES with charging power and discharging power can be expressed as:

$$S_{\mathrm{soc},t+1} = S_{\mathrm{soc},t} - \eta_{\mathrm{c}} P_{\mathrm{c},t}/B_{\mathrm{esn}} - P_{\mathrm{d},t}/(\eta_{\mathrm{d}} B_{\mathrm{esn}}) \quad (1)$$

where $P_{\mathrm{c},t}$ and $P_{\mathrm{d},t}$ are charging power and discharging power, $P_{\mathrm{c},t} < 0$ and $P_{\mathrm{d},t} > 0$; η_{c} and η_{d} are efficiency of charging and discharging; B_{esn} is maximum capacity of the ES.

Energy storage is subject to the following constraints:

$$P_{\mathrm{ES},t} = P_{\mathrm{c},t} + P_{\mathrm{d},t} \quad (2)$$

$$P_{\mathrm{d},t}P_{\mathrm{c},t}=0 \tag{3}$$

$$-P_{\mathrm{c}}^{\max}\le P_{\mathrm{c},t}\le 0 \tag{4}$$

$$0\le P_{\mathrm{d},t}\le P_{\mathrm{d}}^{\max} \tag{5}$$

$$S_{\mathrm{soc}}^{\min}\le S_{\mathrm{soc},t}\le S_{\mathrm{soc}}^{\max} \tag{6}$$

where $P_{\mathrm{ES},t}$ is ES equivalent power; $P_{\mathrm{c}}^{\max}$ and $P_{\mathrm{d}}^{\max}$ are maximum charging and discharging power of ES; $S_{\mathrm{soc}}^{\min}$ and $S_{\mathrm{soc}}^{\max}$ are minimum and maximum of state of charge.

B. Distributed Energy Resources

DERs in this article are power sources that produce electricity and supply loads. Cost characteristics of these generators are different in terms of economy. MTs are generators with secondary cost characteristics, while PVs and WTs are generators without such costs. Power production cost of MTs is expressed as:

$$C(P_{\mathrm{MT},t})=a_{\mathrm{MT}}(P_{\mathrm{MT},t})^2+b_{\mathrm{MT}}(P_{\mathrm{MT},t})+c_{\mathrm{MT}} \tag{7}$$

where a_{MT}, b_{MT}, c_{MT} are cost curve coefficients of MTs; $P_{\mathrm{MT},t}$ is output of MTs.

From the perspective of dispatching, MTs are controllable generators and can be adjusted to meet requirements of dispatching, while PVs and WTs are uncontrollable generators. Hence, historical power data is used to describe timing and randomness of wind and solar output based on uncontrollable assumption.$P_{\mathrm{PV}}^{\max}$ and $P_{\mathrm{WT}}^{\max}$ are actual powers of PVs and WTs.

C. Price-based Demand Respond

We introduce a price-based DR program in which by changing microgrid internal electricity prices during different periods, customers are motivated to modify their consumption patterns. Payoff from microgrid supply price-based DR load at time period t can be expressed as:

$$E_t(P_{\Delta L,t})=\nu(\lambda_t+\Delta\lambda_t)(P_{\mathrm{L},t}-P_{\Delta L,t}) \tag{8}$$

where λ_t and $\Delta\lambda_t$ are retail electricity price and retail electricity price adjustment of a microgrid at time period t; $P_{\mathrm{L},t}$ and $P_{\Delta L,t}$ are load and load adjustment of MG's users at time period t; ν is preferential proportion of electricity price given to users.

Demand elasticity ε ($\varepsilon>0$) is defined as the correlation between price and load as (9).

$$\varepsilon=P_{\Delta L,t}\lambda_t/P_{\mathrm{L},t}\Delta\lambda_t \tag{9}$$

According to (9), equation (8) is equivalent to the quadratic function of load adjustment $P_{\Delta L,t}$, which can be expressed as (10):

$$E_t(P_{\Delta L,t})=(d_{\mathrm{L},t}(P_{\Delta L,t})^2+e_{\mathrm{L},t}P_{\Delta L,t}+g_{\mathrm{L},t}) \tag{10}$$

$$d_{\mathrm{L},t}=(-\nu\lambda_t)/(\varepsilon P_{\mathrm{L},t}) \tag{11}$$

$$e_{\mathrm{L},t}=\nu\lambda_t(1/\varepsilon-1) \tag{12}$$

$$g_{\mathrm{L},t}=\nu\lambda_t P_{\mathrm{L},t} \tag{13}$$

$$-\Delta\lambda_t P_{\mathrm{L},t}\varepsilon/\lambda_t\le P_{\Delta L,t}\le\Delta\lambda_t P_{\mathrm{L},t}\varepsilon/\lambda_t \tag{14}$$

where $d_{\mathrm{L},t}$, $e_{\mathrm{L},t}$ and $g_{\mathrm{L},t}$ are function coefficients. Constraint (14) is load adjustment limitation.

D. Model-based Coordination Optimization

P2P energy trading encourages MGs to exchange power with one another so MGs have the opportunity to achieve a higher payoff and lower cost. For determining the best scheme (trading price and trading quantity), addressing the MGs P2P energy trading, the model-based coordination optimization model for single MG i can be formulated as:

$$\begin{aligned}&\max\ \ U_{\mathrm{MG}.i}\\&=\sum_{t=1}^{T}E_t(P_{\Delta L,t})+(\lambda_{\mathrm{MG\_S},t}P_{\mathrm{MG\_S},t}-\lambda_{\mathrm{MG\_B},t}P_{\mathrm{MG\_B},t})\\&\quad+(\lambda_t^{\min}P_{\mathrm{DN\_S},t}-\lambda_t^{\max}P_{\mathrm{DN\_B},t})-C(P_{\mathrm{MT},t})\end{aligned} \tag{15}$$

$$\begin{aligned}&P_{\mathrm{L},t}-P_{\Delta L,t}+P_{\mathrm{MG\_S},t}+P_{\mathrm{DN\_S},t}\\&=P_{\mathrm{MG\_B},t}+P_{\mathrm{DN\_B},t}+P_{\mathrm{MT},t}+P_{\mathrm{PV},t}+P_{\mathrm{WT},t}+P_{\mathrm{ES},t}\end{aligned} \tag{16}$$

$$0\le P_{\mathrm{MT},t}\le P_{\mathrm{MT}}^{\max} \tag{17}$$

$$0\le P_{\mathrm{WT},t}\le P_{\mathrm{WT}}^{\max} \tag{18}$$

$$0\le P_{\mathrm{PV},t}\le P_{\mathrm{PV}}^{\max} \tag{19}$$

$$\lambda_t^{\min}\le\lambda_{\mathrm{MG\_S},t}\le\lambda_t^{\max} \tag{20}$$

$$\lambda_t^{\min}\le\lambda_{\mathrm{MG\_B},t}\le\lambda_t^{\max} \tag{21}$$

$$P_{\mathrm{MG\_S},t}P_{\mathrm{MG\_B},t}=0 \tag{22}$$

$$P_{\mathrm{DN\_S},t}P_{\mathrm{DN\_B},t}=0 \tag{23}$$

where $P_{\mathrm{MG\_S},t}$ and $P_{\mathrm{MG\_B},t}$ are selling power and purchasing power from other MGs at time period t; $P_{\mathrm{DN\_S},t}$ and $P_{\mathrm{DN\_B},t}$ are selling power and purchasing power from the main grid at time period t; $P_{\mathrm{MT},t}$, $P_{\mathrm{PV},t}$, $P_{\mathrm{WT},t}$ are output of MT, PV, and WT at time period t; $\lambda_t^{\min}$ and $\lambda_t^{\max}$ are main grid feed-in tariff and electricity tariff at time period t; $\lambda_{\mathrm{MG\_S},t}$ and $\lambda_{\mathrm{MG\_B},t}$ are selling price and purchasing price of P2P energy trading at time period t. Objective (15) is MG payoff and consists of four parts: payoff from MG supply price-based DR load, payoff from participating in P2P energy trading, payoff from participating in P2G energy trading, and MT production cost. Production cost of PV and WT is ignored.

It is worth mentioning for each MG, final value of the payoff not only depends on its parameters but on other MGs too. If i is sequence number for MG, and n is total number of MG in the P2P market, then final payoff can be expressed as:

$$\max\ \ \sum_{i=1}^{n}U_{\mathrm{MG}.i} \tag{24}$$

III. Stackelberg Game Formulation Of P2P Energy Trading Mechanism

Game theory is a mathematical tool that analyzes the decision-making process of multiple players in a competitive situation. In the P2P market, there is a game relationship about electricity price between producer and consumer.

First, we assume under the framework of game theory, the P2P energy trading market is perfectly competitive, and market participants are producers and consumers. Strategic space of the trading price is $[\lambda_t^{\min},\lambda_t^{\max}]$. If trading price is greater than $\lambda_t^{\max}$, consumers will buy power from the main grid rather than buying power from producers and vice versa. Participants

have individual rationality to maximize their own interests, but they have no ability to affect trading price alone.

Second, game process is analyzed. Producers should first publish P2P prices to the P2P market. After receiving P2P prices, consumers will respond according to demand and determine purchasing power under the current P2P price. The two sides negotiate many times until purchase and sale power is balanced. In the above game process, producers release the electricity price as leader and consumers respond to the electricity price as follower, forming the MLMF Stackelberg game pattern. When the game evolves to balance of supply and demand, the P2P price happens to be an equilibrium solution of the Stackelberg game.

Next, we will establish a producer and consumer model, analyze its operation mode under guidance of electricity price, and finally get P2P electricity price at game equilibrium.

A. Consumer Model

We assume the P2P energy trading case where producers' excess power can supply consumers' load demand. Otherwise, consumers participate in the P2G market. For consumer, the payoff function at time period t can be rewritten from (15) to (25):

$$\max U_{\mathrm{MG,B},t} = E(P_{\Delta\mathrm{L,B},t}) - \lambda_{\mathrm{MG\_B},t} P_{\mathrm{MG\_B},t} - C(P_{\mathrm{MT,B},t}) \tag{25}$$

Subject to:

$$\text{(1)–(14), (17)–(21)}$$

$$P_{\mathrm{MG,B},t} = P_{\mathrm{L,B},t} - P_{\Delta\mathrm{L,B},t} - P_{\mathrm{MT,B},t} - P_{\mathrm{PV,B},t} - P_{\mathrm{WT,B},t} - P_{\mathrm{ES,B},t} \tag{26}$$

where $U_{\mathrm{MG,B},t}$ is the payoff function of consumer B. Constraint (26) is the consumer MG power balance.

We assume consumer B includes renewable DERs (e.g., solar PVs and WTs) and price-based DR load, consumer problem is restated as:

$$\max U_{\mathrm{MG\_B}} = E(P_{\Delta\mathrm{L,B},t}) - \lambda_{\mathrm{MG\_B},t}(P_{\mathrm{L,B},t} - P_{\Delta\mathrm{L,B},t} - P_{\mathrm{PV,B},t} - P_{\mathrm{WT,B},t}) \tag{27}$$

When renewable DERs power is completely consumed, $U_{\mathrm{MG\_B}}$ is only a function of variable $P_{\Delta\mathrm{L,B},t}$, first-order and second-order derivatives of (27) with respect to$P_{\Delta\mathrm{L,B},t}$ can be expressed as:

$$\frac{\partial U_{\mathrm{MG\_B}}}{\partial P_{\Delta\mathrm{L,B}}, t} = 2d_{\mathrm{L,B},t}(P_{\Delta\mathrm{L,B},t}) + e_{\mathrm{L,B},t} + \lambda_{\mathrm{MG\_B},t} \tag{28}$$

$$\frac{\partial^2 U_{\mathrm{MG\_B}}}{\partial (P_{\Delta\mathrm{L,B},t})^2} = 2d_{\mathrm{L,B},t} \tag{29}$$

According to (11), $d_{\mathrm{L,B},t}$ is a negative constant, so second-order derivatives of (29) are negative. Using the first-order optimality condition, we have the following:

$$\frac{\partial U_{\mathrm{MG\_B}}}{\partial P_{\Delta\mathrm{L,B}}, t} = 2d_{\mathrm{L,B},t}(P_{\Delta\mathrm{L,B},t}) + e_{\mathrm{L,B},t} + \lambda_{\mathrm{MG\_B},t} = 0 \tag{30}$$

Solving (30), consumer best response (i.e., best purchasing power) is obtained as (31)–(32):

$$P^*_{\Delta\mathrm{L,B},t} = \begin{cases} P^{\max}_{\Delta\mathrm{L,B},t} & \frac{-(e_{\mathrm{L,B},t}+\lambda_{\mathrm{MG\_B},t})}{2d_{\mathrm{L,B},t}} > P^{\max}_{\Delta\mathrm{L,B},t} \\ \frac{-(e_{\mathrm{L,B},t}+\lambda_{\mathrm{MG\_B},t})}{2d_{\mathrm{L,B},t}} & P^{\min}_{\Delta\mathrm{L,B},t} \le \frac{-(e_{\mathrm{L,B},t}+\lambda_{\mathrm{MG\_B},t})}{2d_{\mathrm{L,B},t}} \le P^{\max}_{\Delta\mathrm{L,B},t} \\ P^{\min}_{\Delta\mathrm{L,B},t} & \frac{-(e_{\mathrm{L,B},t}+\lambda_{\mathrm{MG\_B},t})}{2d_{\mathrm{L,B},t}} < P^{\min}_{\Delta\mathrm{L,B},t} \end{cases} \tag{31}$$

$$P^*_{\mathrm{MG\_B},t} = P_{\mathrm{L,B},t} - P^*_{\Delta\mathrm{L,B},t} - P^{\max}_{\mathrm{PV,B},t} - P^{\max}_{\mathrm{WT,B},t} \tag{32}$$

Considering existence of MT in a microgrid, cost function of MT is a quadratic function of output power, and there is marginal cost. When the P2P purchasing price (i.e., $\lambda_{\mathrm{MG\_B},t}$) is determined, optimal output of the MT is output when the marginal cost is equal to $\lambda_{\mathrm{MG\_B},t}$ or maximum output. Marginal cost of MT is first-order derivatives of (7) with respect to $P_{\mathrm{MT,B},t}$. Using first-order optimality condition, we have the following:

$$\partial C(P_{\mathrm{MT,B},t})/\partial P_{\mathrm{MT,B},t} = \lambda_{\mathrm{MG\_B},t} \tag{33}$$

Solving (33), optimal $P^*_{\mathrm{MT,B},t}$ is obtained as (34):

$$P^*_{\mathrm{MT,B},t} = \begin{cases} \frac{(\lambda_{\mathrm{MG\_B},t} - b_{\mathrm{MT,B}})}{2a_{\mathrm{MT,B}}} & (\lambda_{\mathrm{MG\_B},t} - b_{\mathrm{MT,B}}) \le 2a_{\mathrm{MT,B}} P^{\max}_{\mathrm{MT,B}} \\ P^{\max}_{\mathrm{MT,B}} & (\lambda_{\mathrm{MG\_B},t} - b_{\mathrm{MT,B}}) > 2a_{\mathrm{MT,B}} P^{\max}_{\mathrm{MT,B}} \end{cases} \tag{34}$$

Further, if MG is equipped with an ES system, we consider it is difficult for consumer to obtain ES optimal response directly. We ignore the key point that superposition of all single time period optimal responses is the optimal response of the complete dispatching period for no time-coupling variables, (i.e., $P_{\Delta\mathrm{L,B},t}$, $P_{\mathrm{MT,B},t}$, $P_{\mathrm{PV,B},t}$, $P_{\mathrm{WT,B},t}$), while the $P_{\mathrm{ES,B},t}$ is a time-coupling variable.

In this section, in order to completely analyze the P2P energy trading making process under the Stackelberg game, we assume $P^*_{\mathrm{ES,B},t}$ is the best response of ES time period t, solution method of the optimal response of ES will be introduced in Section IV.

Finally, the best response $P^*_{\mathrm{MG,B},t}$ is calculated as:

$$P^*_{\mathrm{MG,B},t} = P_{\mathrm{L,B},t} - P^*_{\Delta\mathrm{L,B},t} - P^*_{\mathrm{MT,B},t} - P^*_{\mathrm{ES,B},t} - P^{\max}_{\mathrm{PV,B},t} - P^{\max}_{\mathrm{WT,B},t} \tag{35}$$

B. Producer Model

Selling power to consumers in the P2P market is the behavior of producers. Payoff function of producers at time period t can be rewritten from (15) to (36):

$$\max U_{\mathrm{MG,S},t} = E(P_{\Delta\mathrm{L,S},t}) + \lambda_{\mathrm{MG\_S},t} P_{\mathrm{MG,S},t} - C(P_{\mathrm{MT,S},t}) \tag{36}$$

Subject to:

$$\text{(1)–(14), (17)–(21)}$$

$$P_{\mathrm{MG,S},t} = P_{\mathrm{MT,S},t} + P_{\mathrm{PV,S},t} + P_{\mathrm{WT,S},t} + P_{\mathrm{ES,S},t} - (P_{\mathrm{L,S},t} - P_{\Delta\mathrm{L,S},t}) \tag{37}$$

where $U_{\mathrm{MG,S},t}$ is payoff function of producer S. Constraint (37) is producer power balance. Similar to consumer model, we can analyze producer's best optimal response in which

$\lambda_{\mathrm{MG\_S},t}$ is a parameter. First, optimal load adjustment for producer S at a given $\lambda_{\mathrm{MG\_S},t}$ can be expressed as:

$$P^*_{\Delta\mathrm{L,S},t} = \begin{cases} P^{\max}_{\Delta\mathrm{L,S},t} & \frac{-(e_{\mathrm{L,S},t}+\lambda_{\mathrm{MG\_S},t})}{2d_{\mathrm{L,S}},t} > P^{\max}_{\Delta\mathrm{L,S},t} \\ \frac{-(e_{\mathrm{L,S},t}+\lambda_{\mathrm{MG\_S},t})}{2d_{\mathrm{L,S}},t} & P^{\min}_{\Delta\mathrm{L,S},t} \le \frac{-(e_{\mathrm{L,S},t}+\lambda_{\mathrm{MG\_S},t})}{2d_{\mathrm{L,S}},t} \le P^{\max}_{\Delta\mathrm{L,S},t} \\ P^{\min}_{\Delta\mathrm{L,S},t} & \frac{-(e_{\mathrm{L,S},t}+\lambda_{\mathrm{MG\_S},t})}{2d_{\mathrm{L,S}},t} < P^{\min}_{\Delta\mathrm{L,S},t} \end{cases} \tag{38}$$

Similar to (34), optimal output of MT and ES are expressed as:

$$P^*_{\mathrm{MT,S},t} = \begin{cases} \frac{(\lambda_{\mathrm{MG\_S},t}-b_{\mathrm{MT,S}})}{2a_{\mathrm{MT,S}}} & (\lambda_{\mathrm{MG\_S},t}-b_{\mathrm{MT,S}}) \le 2a_{\mathrm{MT,S}}P^{\max}_{\mathrm{MT,S}} \\ P^{\max}_{\mathrm{MT,S}} & (\lambda_{\mathrm{MG\_S},t}-b_{\mathrm{MT,S}}) > 2a_{\mathrm{MT,S}}P^{\max}_{\mathrm{MT,S}} \end{cases} \tag{39}$$

$$P^*_{\mathrm{ES,S},t} = f(\lambda_{\mathrm{MG\_S},t}) \tag{40}$$

where (40) indicates optimal output of ES is a function of $\lambda_{\mathrm{MG\_S},t}$, but the function expression is unknown at this time.

Finally, best optimal excess power $P^*_{\mathrm{MG,S},t}$ is calculated as:

$$P^*_{\mathrm{MG,S},t} = P^*_{\mathrm{MT,S},t} + P^*_{\mathrm{ES,S},t} + P^{\max}_{\mathrm{PV,S},t} + P^{\max}_{\mathrm{WT,S},t} - (P_{\mathrm{L,S},t} - P^*_{\Delta\mathrm{L,S},t}) \tag{41}$$

According to [11], [15], each producer's selling price will be the same and equal to purchasing price of each consumer when the perfectly competitive P2P energy trading market is in Stackelberg equilibrium. Two important equations in market equilibrium are stated as:

$$\lambda_{\mathrm{MG},t} = \lambda_{\mathrm{MG\_B},t} = \lambda_{\mathrm{MG\_S},t} \tag{42}$$

$$\sum_S P^*_{\mathrm{MG,S},t} = \sum_B P^*_{\mathrm{MG,B},t} \tag{43}$$

where $\lambda_{\mathrm{MG},t}$ is P2P energy trading price in market equilibrium. S, B is index of producer and consumer. $\boldsymbol{S}$, $\boldsymbol{B}$ is set of producers and consumers.

By substituting (35), (41), (42) into (43), the function of P2P energy trading price can be formulated as:

$$\lambda_{\mathrm{MG},t} = \frac{\sum_{i\in\Omega_{\mathrm{PV}}} P^{\max}_{\mathrm{PV},i,t} + \sum_{i\in\Omega_{\mathrm{WT}}} P^{\max}_{\mathrm{WT},i,t} + \sum_{i\in\Omega_{\mathrm{ES}}} P^*_{\mathrm{ES},i,t} - \sum_{i\notin\Omega_1\cup\Omega_2} \frac{e_{\mathrm{L},i,t}}{2d_{\mathrm{L},i,t}} - \sum_{i\notin\Omega_3} \frac{b_{\mathrm{MT},i}}{2a_{\mathrm{MT},i}} - \sum_{i\in\Omega_{\mathrm{L}}} P_{\mathrm{L},i,t} + \sum_{i\in\Omega_3} P^{\max}_{\mathrm{MT},i} + \sum_{i\in\Omega_1} P^{\max}_{\Delta L,i,t} + \sum_{i\in\Omega_2} P^{\min}_{\Delta L,i,t}}{\sum_{i\notin\Omega_1\cup\Omega_2} \frac{1}{2d_{\mathrm{L},i,t}} - \sum_{i\notin\Omega_3} \frac{1}{2a_{\mathrm{MT},i}}} \tag{44}$$

where Ω_{PV}, Ω_{WT}, Ω_{L}, Ω_{ES} are the set of PV, WT, basic load, and ES. Ω_1, Ω_2 are the set of load adjustments that reach maximum and minimum limits. Ω_3 is the set of MT where output power reaches the maximum limits.

Equation (44) is the closed-form expression for Stackelberg equilibrium, which can be utilized to obtain optimal solution $\lambda_{\mathrm{MG},t}$ to guide MGs P2P energy trading process. However, this expression requires all participants' (MGs) relevant privacy parameters and $P^*_{\mathrm{ES},i,t}$ is still unknown.

After the above analysis, it is essential we obtain the optimal response of multi-ES ($P^*_{\mathrm{ES,B},t}$ and $P^*_{\mathrm{ES,S},t}$) in any case.

IV. Markov Decision Process Of Multi Energy Storage

In this section, optimal response of the multi-ES problem is formulated as a Markov decision process (MDP) with unknown transition probability considering uncertainty of the P2P energy trading price. In the MDP, the optimal response problem is represented by a 4-tuple (o, a, Pr, R), where o is set of observation, a is set of actions; Pr denotes transition probability from observation o to new observation o'; R is reward function.

If there are n participants in the P2P market, each agent represents ES in the MG, and environment of all agents is equal to P2P market environment. Modular design enables the model to be extended arbitrarily. Detailed MDP formulation is outlined as follows.

A. Observation

Observation serves as a feedback signal for the ES agent that represents impact of its action on states of the MG environment. Observation for each agent i at time period t is defined as a 5-dimensional vector:

$$o_{i,t} = [S_{\mathrm{soc},t}, P_{\mathrm{L},t}, P_{\mathrm{DER},t}, t, \lambda_{\mathrm{MG},t}] \tag{45}$$

where $P_{\mathrm{DER},t}$ is sum of PV and WT output power. t is time period at present. Note $S_{\mathrm{soc},t}, P_{\mathrm{L},t}, P_{\mathrm{DER},t}$ are internal state of MG but $t, \lambda_{\mathrm{MG},t}$ is state of global P2P energy trading market. Taking local observations into account, P2P market observation (i.e., global observation) can be expressed as:

$$o_t = [o_{1,t}, o_{2,t}, \cdots, o_{n,t}] \tag{46}$$

B. Action

Action of each agent i at time period t is charge and discharge of ES. For continuous action, $a_{i,t} \in [-1, 1]$ is used to represent size of charging and discharging power of ES. Taking local actions of all agents into account, the P2P market actions can be expressed as:

$$a_t = [a_{1,t}, a_{2,t}, \cdots, a_{n,t}] \tag{47}$$

C. State Transition

After execution of actions, the P2P market environment maps actions to charging (negative) and discharging (positive) decisions of ES. Considering time-varying constraints of energy storage, action is not completely consistent with actual charging and discharging power. Therefore, we propose a practical mapping method for ES as follows:

$$P_{\mathrm{d},t} = \min(\max(v_{\mathrm{d}} a_{i,t}/\eta_{\mathrm{d}}, 0), B_{\mathrm{es},t}) \tag{48}$$

$$P_{\mathrm{c},t} = \max(\min(v_{\mathrm{c}} a_{i,t}, 0), B_{\mathrm{es},t} - B_{\mathrm{esn}})\eta_{\mathrm{c}} \tag{49}$$

$$B_{\mathrm{es},t} = S_{\mathrm{soc},t} B_{\mathrm{esn}} \tag{50}$$

where v_{d}, v_{c} are rate of discharging and charging.$B_{\mathrm{es},t}$ is current capacity of ES at time period t. Transition of$S_{\mathrm{soc},t}$ can be expressed as:

$$S_{\mathrm{soc},t+1} = S_{\mathrm{soc},t} - P_{\mathrm{c},t}/B_{\mathrm{esn}} - P_{\mathrm{d},t}/B_{\mathrm{esn}} \tag{51}$$

The agent observation vector we set contains 5 variables. State of charge is closely related to ES action, which has been

discussed above. The DER output and load power state are not obviously related to the action. Its next state is determined by internal uncertainty of time series. Time t is an independent state variable. The most critical state transition is the P2P energy trading price subjected to variability and uncertainty of source-load output and actions of all ES. Fortunately, through the MLMF model in Section III, we obtain the closed-form expression of P2P energy trading price. As long as all actions are input in the environment, we think energy storage action is $P^*_{\mathrm{ES},i,t}$ and (44) can be solved directly, which greatly reduces complexity of calculating state transition of $\lambda_{\mathrm{MG},t}$.

Equation (44) is a centralized analysis method, which involves private parameters of each MG and does not meet requirements of multi-agent architecture to protect privacy. Therefore, we propose a DSET, which can obtain results completely consistent with the centralized method. Privacy protection is reflected in change of information shared by participants from internal parameters to equivalent power (i.e., $P^*_{\mathrm{MG,B},t}$ or $P^*_{\mathrm{MG,S},t}$). Flow diagram of the proposed DSET is shown in Fig. 2. σ is convergence adjustment factor. δ is a small positive number. Transition of$\lambda_{\mathrm{MG},t}$ can be realized by DSET. Comparison between DSET and the centralized method will be shown in the case studies.

D. Reward

The basic reward of each agent i at time period t is payoff of microgrid and can be expressed as:

$$r_{i,t} = P_{\mathrm{ES},i,t}\lambda_{\mathrm{MG},t} + \delta_1|v_{\mathrm{d}}a_{i,t}/\eta_{\mathrm{d}} - B_{\mathrm{es}},t|+ \\ \delta_2|v_{\mathrm{c}}a_{i,t} - B_{\mathrm{es},t} + B_{\mathrm{esn}}| \tag{52}$$

where δ_1 and δ_2 are penalty coefficients. The reward consists of three parts, namely, payoff of ES, penalty for discharging power deviation, penalty for charging power deviation. It is worth noting during DRL training, appropriate rewards can accelerate the training process of agents, and we will scale rewards.

Considering impact of basic reward on future reward, final reward for an agent is equal to discounted reward as follow:

$$R_{i,t} = r_{i,t+1} + \gamma r_{i,t+2} + \gamma^2 r_{i,t+3} + \cdots = \sum_{k=0}^{T-1}\gamma^k r_{i,t+k+1} \tag{53}$$

where $\gamma \in [0,1]$ is the discount factor.

V. Multi-agent Reinforcement Learning For Optimal Response of Energy Storage

Deep reinforcement learning (DRL) is an effective method to deal with MDP problems. Single-agent RL relies heavily on the centralized framework, which not only requires complete communication links and expensive communication equipment but also is difficult to be applied in the decentralized P2P market because of requirements of privacy protection. Based on the CTDE framework, the most advanced multi-agent algorithm is the MADDPG algorithm [30], [31]. In this section, MADDPG is introduced first. Second, according to the object of this paper, a more scalable MASO-DDPG is proposed.

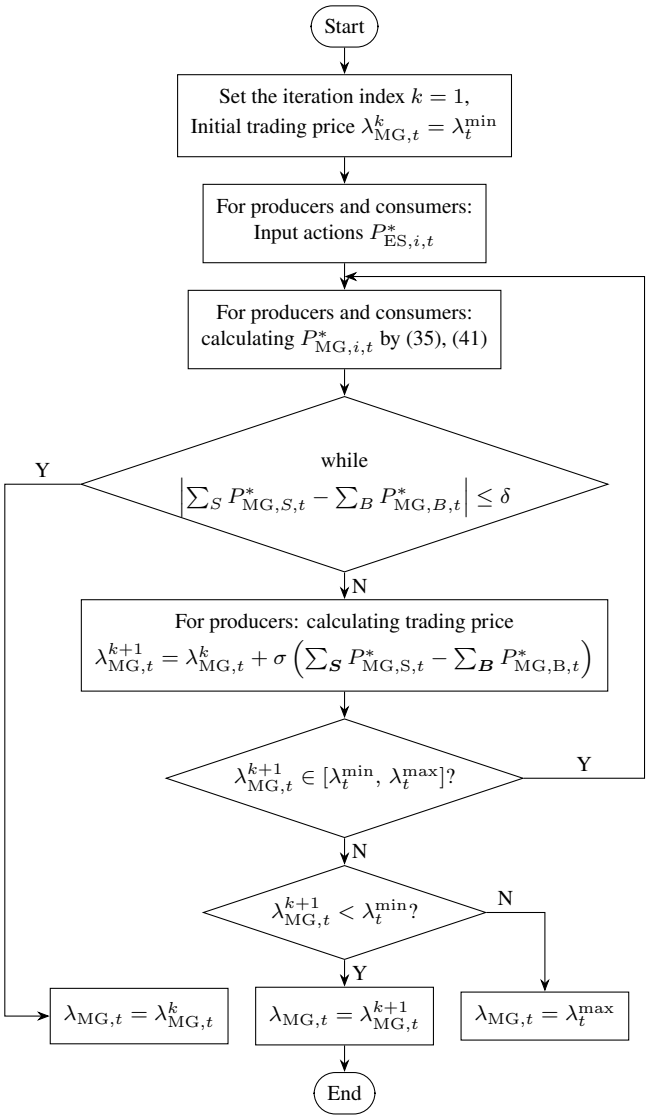

Fig. 2. Flow chart of distributed P2P energy trading price solution method.

A. MADDPG Method

Centralized training with decentralized execution (CTDE) framework is a major feature of MADDPG. During training, the agent can be served as the actor (policy) network, which outputs continuous action based on current local observation and policy. Critic network performs policy evaluation, improves policy by producing estimated Q-value function. Remarkably, the critic is a centralized network, which can use observations and actions of all agents to realize centralized training. But actor only needs agent's local observation to make decisions in testing.

Considering a RL task with n agents with policies parameterized by $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_n]$, let $\boldsymbol{\pi} = [\pi_1, \cdots, \pi_n]$ be the set of all agents' policies. For agent i, the RL main task is to adjust parameters of policy to maximize the objective $J(\theta_i) = \mathbb{E}(R_i)$, gradient of the expected reward can be expressed as:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{o\sim p^{\mu}, a_i\sim\pi_i}[\nabla_{\theta_i}\log\pi_i(a_i|o_i)Q^{\pi}_i(\mathrm{o}, a_1, \cdots a_n)] \tag{54}$$

where $Q_i^{\pi}(\mathrm{o}, a_1, \cdots a_n)$ is a centralized action-value function that takes as input actions of all agents, in addition to observations of all agents. p^{μ} is state distribution. o is observations of all agents, $\mathrm{o} = [o_1, \cdots, o_n]$.

In order to obtain deterministic policies μ_{θ_i} (optimal action of ES must be deterministic rather than random), gradient can be written as:

$$\nabla_{\theta_i} J(\mu_{\theta_i}) = \mathbb{E}_{\mathrm{o},a\sim D}\left[\nabla_{\theta_i}\mu_{\theta_i}(a_i|o_i)\nabla_{a_i}Q_i^{\mu_{\theta_i}}(\mathrm{o}, a_1, \cdots a_n)\big|_{a_i=\mu_{\theta_i}(o_i)}\right] \tag{55}$$

where D is experience replay buffer, which contains tuples recording experience of all agents. Here, MADDPG algorithm follows experience replay buffer mechanism of DDPG and DQN.

During training, actor is updated by gradient descent, and centralized critic network (i.e., action-value function) is updated by using backpropagation to minimize loss; loss function is expressed as:

$$\mathbb{L}(\theta_i) = \mathbb{E}_{\mathrm{o},a,r,\mathrm{o}'}\left[\left(Q_i^{\mu}(\mathrm{o}, a_1, \cdots a_n) - y\right)^2\right] \tag{56}$$

$$y = r_i + \gamma Q_i^{\mu'}(\mathrm{o}', a_1', \cdots a_n')|_{a_j'=\mu_j'(o_j)} \tag{57}$$

where $\mu' = [\mu_{\theta_1'}, \cdots \mu_{\theta_n'}]$ is the set of target policies with delayed parameters.

B. MASO-DDPG Method

Application of MADDPG [30], [31] is very different from the environment of this paper (P2P market), which is mainly reflected in: observation of each agent is completely different, but $\lambda_{\mathrm{MG},t}$ is a global observation, which is calculated by MLMF model in this paper. For agent i, $\lambda_{\mathrm{MG},t}$ has implied observation information of other agents, or we can understand that $\lambda_{\mathrm{MG},t}$ is an abstract representation of all MG states.

Therefore, we propose the MASO-DDPG algorithm by improving the centralized critic network, taking as input all agent actions and single-agent local observation, and outputting the Q-value function.

Critic network comparison between MADDPG and MASO-DDPG is shown in Fig. 3. Critic is a neural network including an input layer, hidden layer, and output layer. Relu and Linear are used as the activation function of a neural network. The most important improvement is the input dimension of the critic is reduced from $6n$ to $5 + n$, which greatly improves scalability and reduces computational complexity.

C. Train and Execution of MASO-DDPG Method

MASO-DDPG uses some skills of the DDPG algorithm.

1) Exploration

When DDPG agent selects actions, a random Gaussian noise to achieve an appropriate balance between exploration and exploitation. Exploration encourages agent to perform different actions to obtain sufficient information. Gaussian noise can be expressed as:

$$\mathscr{N} = N(0, \varpi\chi^2) \tag{58}$$

where χ is the initial value of the standard deviation of Gaussian distribution, ϖ is the decline rate less than 1.

2) Target network

Target networks sharing the same structure and parameters with the online networks, are introduced to improve the stability of the training process.

Algorithm 1 outlines the training process of agents applying MASO-DDPG.

When data-driven agent outputs optimal action of ES, each MG uses the model-driven method to obtain optimal response of P2P energy trading price, excess power, purchasing power, and other internal equipment output.

In a word, we establish a P2P energy trading strategy based on MLMF Stackelberg game and a two-stage SE solution method combining data-driven and model-driven. For a specific day, Stackelberg equilibrium is obtained by Algorithm 2.

VI. Cases Studies

A. Cases Studies Setup

The proposed model and algorithm are applied to a 4-Microgrid P2P energy trading case. Technical characteristics of the Ess and MT in 4-Microgrid are in Appendix Tables I andII. Electricity price is listed in Appendix Table III. For demand response, parameters ν are 0.9, 0.8, 0.8, 0.7; parameters ε are 0.17, 0.51, 0.43, 0.41, respectively.

When implementing the DRL algorithm for off-line training, source and load data in Zhejiang, China is selected. The dataset provides basic load, PV generation, and WT generation data from 2019 with an hourly resolution. The first 270 days are used for training the DRL agents, and remaining days are selected for evaluation.

Our performance evaluation is divided into two parts. The first part evaluates convergence performance of DSET, and

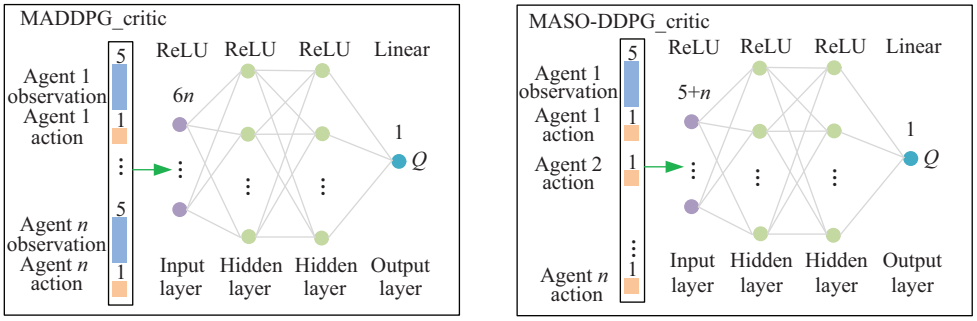

Fig. 3. Critic evaluation under MADDPG and MASO-DDPG.

Algorithm 1: Training of MASO-DDPG

Input: replay buffer D_i size, batch size $\mathscr{M}$, exploration noise $\mathscr{N}$, initial critic, and actor network with weights

for *episode* $=$ *1 to* M **do**
  Reset environment ($t=0$) to get observation $o_{i,0}$ for each agent i
  for $t = 0$ *to* T **do**
    Select action $a_{i,t} = \mu_{\theta_i}(o_{i,t}) + \mathscr{N}$ for each agent i
    execute actions $\mathrm{a}_t = [a_{1,t}, a_{2,t}, \cdots, a_{n,t}]$ and obtain reward $r_{i,t}$ and next observation $o_{i,t+1}$ for agent i
    store experience $(o_{i,t}, \mathrm{a}_t, r_{i,t}, o_{i,t+1})$ in the replay buffer D_i
    $o_{i,t} \leftarrow o_{i,t+1}$
    for *agent* i $=$ *1 to* n **do**
      sample a random minibatch of $\mathscr{M}$ experience from D_i
      set $y^j = r_i^j + \gamma Q_i^{\mu'}(o_i'^j, a_1', \cdots a_n')|_{a_k'=\mu_k'(o_k^j)}$
      update critic by minimizing the loss $L(\theta_i) = \frac{1}{\mathscr{M}} \sum_j (y^j - Q_i^{\mu}(o^j, a_1^j, \cdots a_n^j))^2$
      update actor using the sample policy gradient: $\nabla_{\theta_i} J = \frac{1}{\mathrm{M}} \sum_j \nabla_{\theta_i} \mu_{\theta_i}(o_i^j)$ $\nabla_{a_i} Q_i^{\mu_{\theta_i}}(o^j, a_1^j, \cdots, a_i, \cdots a_n^j)|_{a_i=\mu_{\theta_i}(o^j)}$
    end
    update target network parameters for each agent i
  end
end

Algorithm 2: Stackelberg equilibrium solution solving process

Input: Load online actor network for each agent i

Reset environment ($t=0$) to get observation $o_{i,0}$ for each agent i
for $t = 0$ *to* T **do**
  Select action $a_{i,t} = \mu_{\theta_i}(o_{i,t})$ for each agent i by online actor
  each energy storage execute actions $\mathrm{a}_t = [a_{1,t}, a_{2,t}, \cdots, a_{n,t}]$
  each MG calculate $\lambda_{\mathrm{MG},t}$ by using DSET
  Other devices (DR, MT) perform optimal response actions according to (31), (34), (38), (39) for each MG
  return next observation $o_{i,t+1}$ for agent i
end

second part evaluates effectiveness of DRL agent offline training and online application. Our simulation environment is Python 3.7 on a desktop with 24 GB RAM and an i7-8700.

B. Performance Evaluation of DSET

Case I: We assume optimal response of ES at any time is 0 (i.e., ES is off) and P2P energy trading in a 4-Microgrid system. Centralized solution technology (i.e., closed-expression (44)) is utilized to compare with DSET to verify effectiveness of DSET. Fig. 1 in the Appendix shows actual output of equipment in each microgrid in case I.

Figure 4 shows P2P trading price continues the iterative trends until getting to consensus point. X/Y and Z axes indicate the time period in one hour, number of iterations, and the P2P trading price value, respectively. Here σ is equal to 0.001. It is obvious from the figure that over iterations 1–10, P2P trading price converges to a steady value in all time periods. In the period 1–6 and 23–24, the P2P trading price converged after iteration 1.

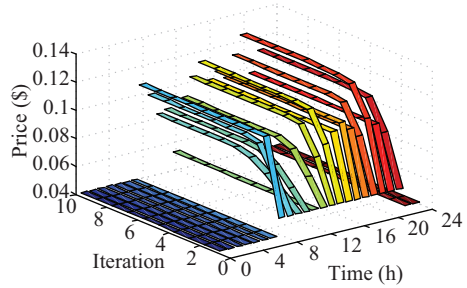

Fig. 4. The P2P trading price iterative trends in the market.

As shown in Fig. 5, in each period, P2P trading price is between retail electricity tariff and feed-in tariff, which proves rationality of the P2P trading price. In addition, the P2P trading price obtained by DSET, and centralized solution technology is completely consistent. In DSET, each MG only needs to submit its insufficient/excess power quantity, which effectively protects personal privacy of participants. Although computational complexity of DSET is higher than of centralized technology, reasonable selection of σ can greatly reduce computational complexity. In short, DSET takes into account requirements of calculation accuracy and privacy protection and has advantages.

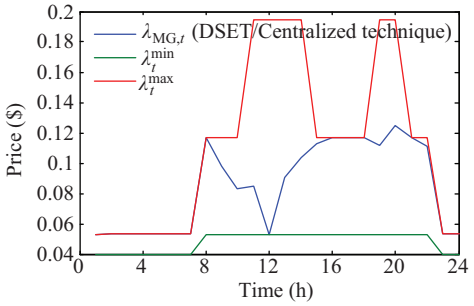

Fig. 5. Comparison of the P2P trading price in DSET and centralized technique.

Convergence times of the DSET algorithm are closely related to selection of σ. Fig. 6 shows the relationship between σ of different orders of magnitude, convergence times, and calculation accuracy. Specifically, the P2P electricity price obtained by centralized technology is used as the benchmark, and calculation error is defined as the error between P2P electricity price obtained by DSET and the benchmark.

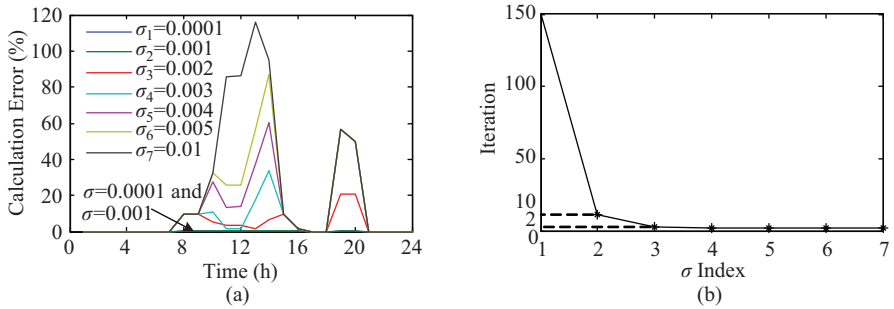

Fig. 6. Relationship between calculation error, iteration, and factor in DSET. (a) Relationship between calculation error and factor. (b) Relationship iteration error and factor.

Calculation error of DSET is 0 when σ is equal to 0.0001 and 0.001, but the number of iterations is 150 and 10, respectively. Calculation error increases with increase of σ. At the same time, we find the increase of calculation error is related to number of iterations. If iteration converges twice and the error is large, it just shows selection ofσ is unreasonable. Therefore, in this paper, we choose 0.001 to give consideration to accuracy and computational complexity.

C. Performance Evaluation of MASO-DDPG

Case II: There are ES systems in the 4-Microgrid system, and source and load output of the microgrid come from historical data in 2019. We compare 4 data-driven DRL algorithms: MASO-DDPG, MADDPG, DDPG, and DDQN to verify superiority of the proposed MASO-DDPG algorithm.

Table I summarizes input and output dimensions of DNNs that need to be trained. In order to apply DDQN, we discretize action in five integer values representing -100%, -50%, 0, 50%, 100% respectively, reflecting charging and discharging power of ES. Because of the discretization action, current output dimension of the DDQN algorithm is 625, which has faced a dimension disaster on the Q network. The neural network input/output dimension of the DDPG agent is higher than MADDPG because DDPG uses one agent to control all ES and its neural network needs all observation and action information. The difference between MASO-DDPG and MADDPG lies only in input dimension of the critic network, which is 9 and 24, respectively. With increase of the number of agents, the critic network of MADDPG may face dimensional disaster. Considering scalability, ranking of the four algorithms is MASO-DDPG $>$ MADDPG $>$ DDPG $>$ DDQN.

TABLE I
COMPARISON OF NEURAL NETWORK DIMENSION

Method	Actor/Q-network dimensions	Critic dimensions
MASO-DDPG	(5/1)	(9/1)
MADDPG	(5/1)	(24/1)
DDPG	20/4	(24/4)
DDQN	20/625	–

For the case of the 4-Microgrid system, we select MASO-DDPG, MADDPG, and DDPG algorithms to train agents to further verify advantages of MASO-DDPG. Minibatch size and the replay buffer size are set as 64 and 10^4, respectively. To assess generality, 5 different random seeds are generated, and each algorithm is trained for 10^4 episodes for each seed, where each episode is randomly selected from the training set including 270-day historical data. Average reward is defined as average reward of 200 episodes. Please note that in the following experimental results, reward scaling factor is 1/100. The change process of average reward of the three algorithms is shown in Fig. 7.

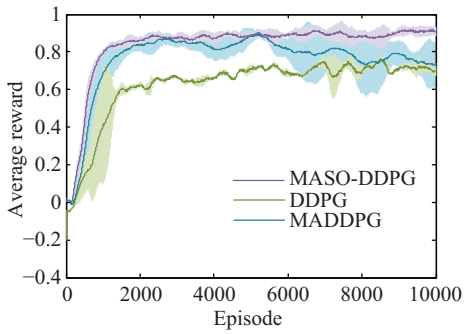

Fig. 7. The average reward of training process under different methods.

Mean and standard deviation of average reward over the 5 seeds are illustrated through solid lines and shaded areas, respectively, in Fig. 7. Comparing MASO-DDPG and MADDPG, we can find average reward of the two algorithms is close at the beginning of training. With number of training times greater than 5000, it can be seen from average reward that MADDPG is not able to stably learn the policy of high reward. On the contrary, average reward of MASO-DDPG is relatively stable and the standard deviation is very small in the whole training process.

After the training procedure, we deploy the model to assess the training set and testing set. Table II shows the mean/standard deviation of average reward. The average reward represents optimization performance of the algorithm, and standard deviation represents stability of the algorithm. It can be seen that MASO-DDPG outperforms DDPG and

TABLE II
MEAN AND STANDARD DEVIATION OF AVERAGE REWARD

Method	Average reward (mean/standard deviation)	
	Testing set	Training set
MASO-DDPG	0.84 ± 0.14	0.853 ± 0.103
MADDPG	0.69 ± 0.16	0.691 ± 0.131
DDPG	0.78 ± 0.14	0.731 ± 0.116

MADDPG towards optimization performance both training and testing set, and MASO-DDPG achieves a 21.7%/7.69% higher average reward over MADDPG/DDPG. The disadvantage of both MADDPG and DDPG in the training set is the standard deviation larger than MASO-DDPG, which indicates the training process of different random seeds has great differences and randomness.

According to performance of training set and testing set, MASO-DDPG is better than MADDPG and DDPG. In the current data set, optimization performance and stability of DDPG are difficult to be directly compared with MADDPG, but privacy protection performance of DDPG is relatively backward.

D. Value of P2P Energy Trading

The purpose of this section is to further verify performance of the proposed P2P energy trading strategy based on the Stackelberg game and its two-stage solution technique by comparing it with the model-based optimization method. Two cases are proposed to compare P2P transaction results, and the test set is used as cases data.

Case III: P2P energy trading mode is not applied to the MGs, thus each MG can only trade with the main grid. The solution technique for transaction strategy (i.e., trading power) is obtained by using the YALMIP toolbox.

Case IV: P2P energy trading strategy based on the MLMF Stackelberg game is applied to MGs, and algorithm 2 in Section V is used to obtain trading strategy (i.e., P2P trading price and trading power).

A comparison is made between performances of the Stackelberg game and its two-stage solution technique and of the model-based optimization method. This verifies performance of the proposed P2P energy trading strategy. Average payoff of each MG and average payoff of all MGs in different cases are given and compared in Table III. In case III, each MG sells surplus power to the main grid at a feed-in tariff, or purchases power at a retail price in case of shortage. Further, each MG can only trade with the main grid. In case IV, all MGs have obtained higher benefits through P2P energy trading power and achieved 12.9%, 30.1%, 38.6%, 26.8% higher average payoff compared with case III.

TABLE III
MEAN PAYOFF IN DIFFERENT CASES

Case	Mean Payoff ($)				
	MG1	MG2	MG3	MG4	Total
III	522.7	412.3	315.4	283.9	1534.3
IV	589.9	536.6	437.2	360.1	1923.8

Figures 8 and 9 show trading power and P2P trading price on the 8th day of the test set. In Fig. 8, positive power indicates the MG purchases power and negative power indicates the MG sells power. We can find when the P2P trading price reaches maximum limit, supply in the whole market is less than demand (e.g., periods 1–8 and 23- 24); when P2P trading price reaches minimum limit (e.g., periods 10 and 12–15), market supply exceeds demand; only when P2P trading price is between $\lambda_t^{\min}$ and $\lambda_t^{\max}$, the Stackelberg equilibrium exists and market supply and demand equalize.

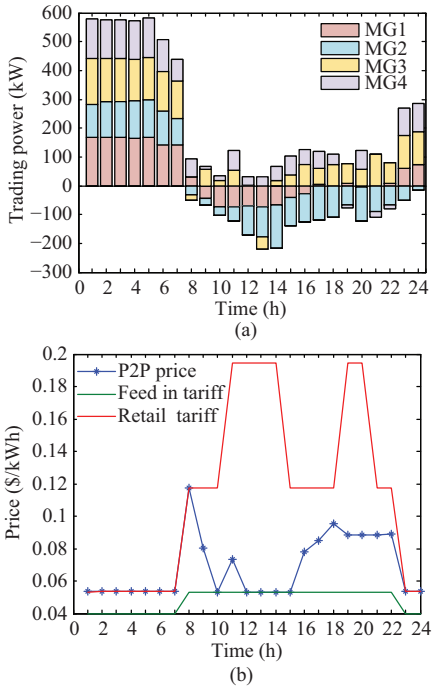

Fig. 8. Trading power and trading price in case IV. (a) Trading power in case IV. (b) Trading Price in case IV.

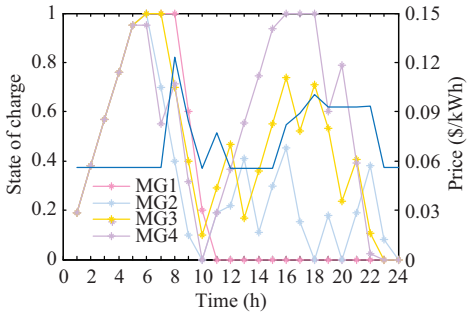

Fig. 9. Charging and discharging schedule of ES in case IV.

Figure 9 shows scheduling results of each ES obtained by Algorithm 2. Increase of state of charge (SOC) indicates ES is charged, and vice versa indicates ES is discharged. By learning the potential relationship of P2P trading prices from historical data, ES can predict future electricity price trend. The common feature is ES stores power in time period 1–5. In the subsequent period of high P2P trading price (there is a power shortage), the ES supplies the power shortage, but actions of different ESs are not completely consistent.

Comparing case I and case IV can highlight the advantages of ES system. Table IV shows the impact of ES on average power purchasing cost of each microgrid in the testing set. Through the arbitrage action of ES, power purchasing cost of each microgrid has been reduced by 4.0%, 8.4%, 1.5% and 3.5% respectively.

Energy self-sufficiency of P2P market is defined as the ratio between supply and demand balance period and total period of P2P market. In self-sufficiency period, P2P market does not need to conduct any transactions with the main grid and operates independently. Table V shows comparison of P2P

TABLE IV
COMPARISON OF AVERAGE POWER PURCHASING COST OF DIFFERENT CASES

Case	Mean Cost ($)				
	MG1	MG2	MG3	MG4	Total
I	202.7	47.5	152.1	193.4	595.7
IV	194.9	43.8	149.9	186.8	575.4

TABLE V
SELF-SUFFICIENCY RATE OF DIFFERENT CASES

Description	Case I	Case IV
Balance period	1226	1282
Total period	2160	2160
self-sufficiency rate (%)	56.7	59.3

market self-sufficiency rate in the test set. After ES regulation, market self-sufficiency rate increased by 2.6%. Specifically, 56 supply-demand balance periods have been added in 90 days, which can promote independence of P2P market.

All in all, the P2P energy trading model based on the MLMF Stackelberg game can provide an effective method to analyze P2P energy trading process. Integration of data-driven and model-driven approaches can effectively determine P2P trading price and exchange energy among participants connected in the P2P market.

E. Performance Evaluation of Scalability

To further demonstrate advantages in large-scale cases, we expand real data of 4-Microgrids to 8 and 16-Microgrids and compare computational performance of the algorithm in Table VI. Average training time per episode of DDPG agent is the smallest, but DDPG does not converge in the face of 8-Microgrids and 16-Microgrids. It can be observed as the number of agents increases, test time increases synchronously, but all DRL agents exhibit a similar average test time at the millisecond level, implying DRL can be applied to real-time management. Overall, MASO-DDPG exhibits higher computational efficiency.

Another aspect of scalability is learning performance, i.e., able to learn policies with high rewards. In the 8-Microgrid and 16-Microgrid systems, the DDPG fails, and the optimal response of ES cannot be found. With increase of the number of agents, the disadvantage of the unstable training process of MADDPG becomes more prominent In Fig. 10.

TABLE VI
COMPUTATIONAL PERFORMANCE OF DIFFERENT ALGORITHM

Number of MG	Average training time per episode (s)			Average test time per episode (ms)		
	①	②	③	①	②	③
4	0.53	0.57	0.26	67.7	69.7	17.5
8	1.36	1.40	0.33*	122.3	128.6	56.1
16	3.34	3.55	0.33*	248.5	256.6	59.3

* Failure to attain the reward
① represents the MASO-DDPG algorithm, ② represents the MADDPG algorithm and ③ represents the DDPG algorithm.

VII. CONCLUSION

In this paper, the P2P energy trading method based on the MLMF Stackelberg game is proposed. SE is defined as a novel P2P energy trading strategy. Experimental results show the validity of the proposed P2P energy trading model in stimulating MGs to participate in energy transactions.

A two-stage distributed method of MASO-DDPG and DSET is proposed for the first time to solve the P2P energy trading model. The two-stage method has two advantages. First, each stage scheme is distributed, and privacy protected, which is suitable for decentralized P2P trading market structure. Second, application of the DRL algorithm shows great potential in computing real-time and scalability and is expected to be applied to the real-time market.

However, further research is still needed. Disregarding power network constraints might lead to infeasible or untrue trading strategies. Our work will be focused on how to consider impact of different power network topologies in the P2P energy trading model and solution technology.

APPENDIX

TABLE AI
PARAMETERS OF ENERGY STORAGE

Number	Capacity (kWh)	Charging/discharging efficiency	Maximum discharging/ charging rate (%)
1	200	0.95/0.95	¥0.4/0.2
2	100	0.95/0.95	¥0.3/0.2
3	100	0.95/0.95	¥0.3/0.2
4	150	0.95/0.95	¥0.4/0.2

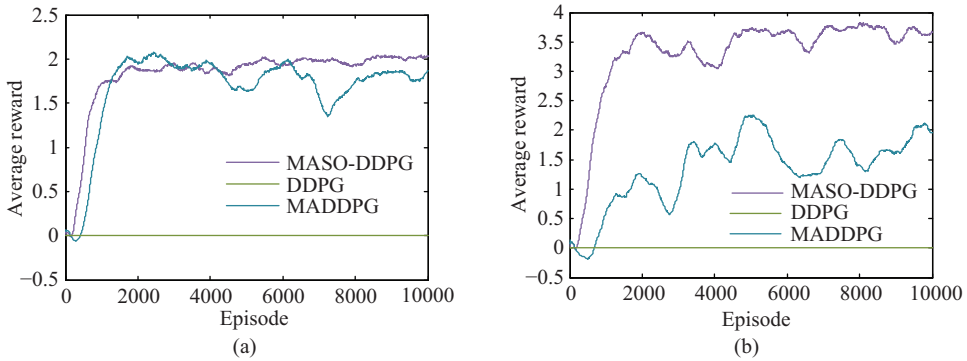

Fig. 10. The average reward of training process under different number of agents. (a) Average reward of training process under 8 MGs. (b) Average reward of training process under 16 MGs.

TABLE AII
COST CURVE COEFFICIENTS OF MT

Number	Maximum power (kWh)	a_{MT_i} ($/kWh)	b_{MT_i} ($/kWh)	c_{MT_i} ($/kWh)
1	200	0.00026	0.067	1.72
2	100	0.00021	0.067	1.72
3	200	0.00018	0.079	2.01
4	–	–	–	–

TABLE AIII
THE ELECTRIC PRICE IN CASES

Time	Purchase price $\lambda_t^{\max}$ ($/kWh)	Feed-in tariff $\lambda_t^{\min}$ ($/kWh)
1:00–7:00, 23:00–24:00	0.05	0.04
8:00–10:00, 15:00–18:00, 21:00–22:00	0.12	0.05
11:00–14:00, 19:00–20:00	0.19	0.05

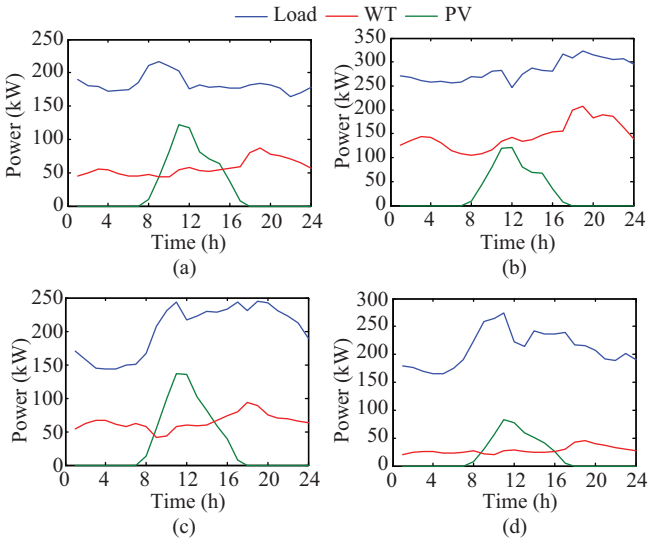

Fig. A1. The actual output of equipment in each microgrid in case I. (a) MG1. (b) MG2. (c) MG3. (d) MG4.

REFERENCES

[1] T. T. H. Pham, Y. Besanger, and N. Hadjsaid, "New challenges in power system restoration with large scale of dispersed generation insertion," *IEEE Transactions on Power Systems*, vol. 24, no. 1, pp. 398–406, Feb. 2009.

[2] S. Parhizi, H. Lotfi, A. Khodaei, and S. Bahramirad, "State of the art in research on microgrids: a review," *IEEE Access*, vol. 3, pp. 890–925, Jun. 2015.

[3] L. C. Ye, J. F. D. Rodrigues, and H. X. Lin, "Analysis of feed-in tariff policies for solar photovoltaic in China 2011–2016," *Applied Energy*, vol. 203, pp. 496–505, Oct. 2017.

[4] Y. K. Renani, M. Ehsan, and M. Shahidehpour, "Optimal transactive market operations with distribution system operators," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6692–6701, Nov. 2018.

[5] W. J. Liu, J. P. Zhan, and C. Y. Chung, "A novel transactive energy control mechanism for collaborative networked microgrids," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 2048–2060, May 2019.

[6] W. Tushar, B. Chai, C. Yuen, D. B. Smith, K. L. Wood, Z. Y. Yang, and H. V. Poor, "Three-party energy management with distributed energy resources in smart grid," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 2487–2498, Apr. 2015.

[7] C. H. Zhang, J. Z. Wu, Y. Zhou, M. Cheng, and C. Long, "Peer-to-Peer energy trading in a Microgrid," *Applied Energy*, vol. 220, pp. 1–12, Jun. 2018.

[8] M. J. Thompson, H. Sun and J. Jiang, "Blockchain-based Peer-to-Peer Energy Trading Method," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 5, pp. 1318–1326, Sep. 2022.

[9] W. Q. Hua, H. Xiao, W. Pei, W. Y. Chiu, J. Jiang, and P. Matthews, "Transactive Energy and Flexibility Provision in Multi-microgrids Using Stackelberg Game," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 2, pp. 505–515, Mar. 2023.

[10] J. Lee, J. Guo, J. K. Choi, and M. Zukerman, "Distributed energy trading in microgrids: a game-theoretic model and its equilibrium analysis," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3524–3533, Jun. 2015.

[11] K. Anoh, S. Maharjan, A. Ikpehai, Y. Zhang, and B. Adebisi, "Energy peer-to-peer trading in virtual microgrids in smart grids: a game-theoretic approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1264–1275, Mar. 2020.

[12] A. M. Jadhav, N. R. Patne, and J. M. Guerrero, "A novel approach to neighborhood fair energy trading in a distribution network of multiple microgrid clusters," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1520–1531, Feb. 2019.

[13] A. M. Jadhav and N. R. Patne, "Priority-based energy scheduling in a smart distributed network with multiple microgrids," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 3134–3143, Dec. 2017.

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

[15] F. Wei, Z. X. Jing, P. Z. Wu, and Q. H. Wu, "A Stackelberg game approach for multiple energies trading in integrated energy systems," *Applied Energy*, vol. 200, pp. 315–329, Aug. 2017.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[17] D. Cao, W. H. Hu, J. B. Zhao, G. Z. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029–1042, Nov. 2020.

[18] Z. D. Zhang, D. X. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: an overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, Mar. 2020.

[19] Y. Ji, J. H. Wang, J. C. Xu, X. K. Fang, and H. G. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, pp. 2291, Jun. 2019.

[20] Y. K. Liu, D. X. Zhang, and H. B. Gooi, "Optimization strategy based on deep reinforcement learning for home energy management," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 3, pp. 572–582, Sep. 2020.

[21] Y. K. Liu, D. X. Zhang, and H. B. Gooi, "Data-driven decision-making strategies for electricity retailers: a deep reinforcement learning approach," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 358–367, Mar. 2021.

[22] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.

[23] V. H. Bui, A. Hussain, and H. M. Kim, "Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457–469, Jan. 2020.

[24] B. Zhang, W. H. Hu, J. H. Li, D. Cao, R. Huang, Q. Huang, Z. Chen, and F. Blaabjerg, "Dynamic energy conversion and management strategy for an integrated electricity and natural gas system with renewable energy: deep reinforcement learning approach," *Energy Conversion and Management*, vol. 220, pp. 113063, Sep. 2020.

[25] 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, pp. 100413, Mar. 2021.

[26] S. Y. Zhou, Z. J. Hu, W. Gu, M. Jiang, M. Chen, Q. T. Hong, and C. Booth, "Combined heat and power system intelligent economic

dispatch: a deep reinforcement learning approach," *International Journal of Electrical Power & Energy Systems*, vol. 120, pp. 106016, Sep. 2020.
[27] X. H. Fang, Q. Zhao, J. K. Wang, Y. H. Han, and Y. C. Li, "Multi-agent deep reinforcement learning for distributed energy management and strategy optimization of microgrid market," *Sustainable Cities and Society*, vol. 74, pp. 103163, Nov. 2021.
[28] S. Aladdin, S. El-Tantawy, M. M. Fouda, and A. S. T. Eldien, "MARLA-SG: multi-agent reinforcement learning algorithm for efficient demand response in smart grid," *IEEE Access*, vol. 8, pp. 210626–210639, Nov. 2020.
[29] H. M. Chung, S. Maharjan, Y. Zhang, and F. Eliassen, "Distributed deep reinforcement learning for intelligent load scheduling in residential smart grids," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2752–2763, Apr. 2021.
[30] D. N. Liu, Y. Gao, W. Y. Wang, and Z. X. Dong, "Research on bidding strategy of thermal power companies in electricity market based on multi-agent deep deterministic policy gradient," *IEEE Access*, vol. 9, pp. 81750–81764, Jun. 2021.
[31] S. Y. Wang, J. J. Duan, D. Shi, C. L. Xu, H. F. Li, R. S. Diao, and Z. W. Wang, "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4644–4654, Nov. 2020.

Pengjie Zhao received the Ph.D. degree at Beijing Jiaotong University, Beijing, China, in 2023. He is currently working at the Department of Electric Power Control Center, State Grid Shanxi Electric Power Company. His research interests includedeep reinforcement learning, game theory, and deeplearning.

Junyong Wu received the B.S., M.S., and Ph.D. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 1987, 1989, and 1993, respectively. Since 2004, he has been a Professor, Ph.D. Supervisor in the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include power system analysis, smart grid, and AI technology in power systems.

Fashun Shi received the B.S. degree in Electrical Engineering from North China University of Water Resources and Electric Power, Zhengzhou, China, in 2019. He is currently pursuing the Ph.D. degree at Beijing Jiaotong University, Beijing, China. His research interests include transient stability analysis and control of power system, deep learning.

Lusu Li received the B.S. and M.S. degrees from the Science and Technology College, North China Electric Power University, Hebei, China, in 2015, and from the School of Automation Engineering, Shanghai University of Electric Power, Shanghai, China, in 2020. He is currently pursuing the Ph.D. degree at Beijing Jiaotong University, Beijing, China. His research interests are frequency safety analysis and assessment of power systems, frequency stability control, deep learning.

Baoqin Li received the B.S. degree in Electrical Engineering from Beijing Jiaotong University, Beijing, China, in 2018, where she is currently pursuing the Ph.D. degree. Her research interests include transient stability assessment of power system, data mining, and deep learning.

Yi Wang received the B.S. degree in Electrical Engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include distribution network optimization planning, data mining, and deep learning.