Optimal Power Dispatch of Active Distribution Network and P2P Energy Trading Based on Soft Actor-critic Algorithm Incorporating Distributed Trading Control

Yongjun Zhang, Jun Zhang, Guangbin Wu, Jiehui Zheng, Dongming Liu, and Yuzheng An

***Abstract*—Peer-to-peer (P2P) energy trading in active distribution networks (ADNs) plays a pivotal role in promoting the efficient consumption of renewable energy sources. However, it is challenging to effectively coordinate the power dispatch of ADNs and P2P energy trading while preserving the privacy of different physical interests. Hence, this paper proposes a soft actor-critic algorithm incorporating distributed trading control (SAC-DTC) to tackle the optimal power dispatch of ADNs and the P2P energy trading considering privacy preservation among prosumers. First, the soft actor-critic (SAC) algorithm is used to optimize the control strategy of device in ADNs to minimize the operation cost, and the primary environmental information of the ADN at this point is published to prosumers. Then, a distributed generalized fast dual ascent method is used to iterate the trading process of prosumers and maximize their revenues. Subsequently, the results of trading are encrypted based on the differential privacy technique and returned to the ADN. Finally, the social welfare value consisting of ADN operation cost and P2P market revenue is utilized as a reward value to update network parameters and control strategies of the deep reinforcement learning. Simulation results show that the proposed SAC-DTC algorithm reduces the ADN operation cost, boosts the P2P market revenue, maximizes the social welfare, and exhibits high computational accuracy, demonstrating its practical application to the operation of power systems and power markets.**

***Index Terms*—Optimal power dispatch, peer-to-peer (P2P) energy trading, active distribution network (ADN), distributed trading, soft actor-critic algorithm, privacy preservation.**

I. Introduction

WITH the increasing penetration of distributed energy resources (DERs), battery energy storage (BES), and adjustable loads, the distribution networks face operational problems such as overloading, voltage overruns, and network losses. Under the unified management of distribution system operator (DSO), the active distribution network (ADN) [1], [2] regulates the active and reactive power outputs of various types of discrete devices (such as on-load tap changers (OLTCs) and capacitor banks (CBs)) and continuous devices (such as DERs and static var generators (SVGs)). This is achieved through the implementation of reasonable energy management strategies, in order to ensure the safe and efficient operation of distribution network [3], [4].

Consequently, the optimal power dispatch problems for ADNs are usually formulated as mixed-integer nonlinear models [4], [5]. However, the solution is dependent on the accuracy of the network topology models and is not applicable to ADNs with rapidly changing structures [6]. Therefore, some scholars have adopted reinforcement learning methods to solve the ADN optimization problems in a time-efficient and model-free manner. References [7] and [8] employ the deep Q-network (DQN) and proximal policy optimization (PPO), respectively, to explore the optimal control strategies for discrete and continuous devices in ADNs. PPO is able to reduce the variance during the training process more effectively but requires multiple collections of the same data sample. The deep deterministic policy gradient (DDPG) improves the sample utilization efficiency and exploration effects by adopting an actor-critic framework and adding random noises [9], yet it is quite sensitive to hyperparameters [10], [11]. The soft actor-critic (SAC) provides a smoother training process and lower variance through its entropy regularization and dual Q-network structure, facilitating more stable and effective learning in the context of complex and dynamic environments in ADNs [12].

Some devices such as DER and BES in ADNs may belong to independent individuals with different interest claims [13]. In the case where the DSO centralizes the dispatch of all the energy resources, the optimal outcome from a benefit-optimization perspective for some individuals may not be

Manuscript received: May 19, 2024; revised: July 11, 2024; accepted: August 28, 2024. Date of CrossCheck: August 28, 2024. Date of online publication: September 19, 2024.

This work was supported by the National Natural Science Foundation of China (No. 52177085).

Y. Zhang, J. Zhang, J. Zheng (corresponding author), D. Liu, and Y. An are with the School of Electric Power, South China University of Technology, Guangdong Key Laboratory of Clean Energy Technology, Guangzhou 510641, China (e-mail: zhangjun@scut.edu.cn; 202220114243@mail.scut.edu.cn; zhengjh@scut.edu.cn; 15553181020@163.com; 14737658962@163.com).

G. Wu is with the Customer Service Center of Guangdong Power Grid Corporation, Foshan, China (e-mail: 2439407952@qq.com).

DOI: 10.35833/MPCE.2024.000471

achieved [14], and it will not be feasible to leverage the individual motivation to participate in the operation regulation. Fortunately, the emerging peer-to-peer (P2P) energy trading provides a solution to this challenge. In a fully incentivized P2P market, the owners of the assets (called prosumers) can actively participate in the energy regulation of the local distribution network by selling electricity or reducing demand, thereby maximizing their revenue and mitigating the peak demand and operating costs of the distribution network [15].

The effectiveness of P2P markets has been extensively studied and validated [16]. Depending on the manner of coordination among participants, the P2P market mechanism can be classified into centralized and decentralized schemes.

In the centralized scheme, a central entity (such as P2P operator or DSO) is responsible for coordinating energy trading and benefit distribution, with the advantage of maximizing the social welfare [17]. However, as the number of DERs and prosumers increases, the operator may face problems such as data pressure, computational curse of dimensionality, and user information leakage [15]. In the decentralized scheme, the prosumers are able to decide the transaction parameters by themselves and complete the information interaction and energy trading process, which has the advantage of decision independence and strong privacy protection [18] but may lead to non-optimal social welfare.

In recent years, there has been an increase in research on P2P markets. At the level of information interaction and market operation, most studies have primarily employed block chain [19], [20], auctions [21], and game theoretic approaches for pricing and trading energy. Specifically, to explore the competitive relationship between DSOs and prosumers, the models of non-cooperative game and auction strategies are employed to evaluate the profits of P2P energy trading [22]-[24]. Meanwhile, the research efforts [25]-[27] focus on developing methods to fairly distribute benefits within communities, utilizing cooperative game concepts and predefined rules.

Furthermore, the P2P markets encompass energy trading at the information layer, which requires secure transmission at the physical layer of the distribution network. A fully decentralized two-loop algorithm is proposed in [28] to coordinate P2P energy trading with voltage regulation capability. Similarly, considering the distribution network constraints, the method in [29] proposes a trading strategy based on an alternating direction multiplier method and bidding auction.

However, the existing studies generally need to consider the control of the device governed by DSOs. The lack of transparency regarding the respective behaviors of DSO and prosumers may result in problems such as voltage overruns and network loss increase in the distribution network [14]. As a result, DSOs may need to take more conservative and stringent measures to maintain grid security, leading to a further reduction in social welfare.

Although these studies provide valuable insights, they are constrained by several limitations, such as difficulties in privacy protection, ignoring distribution network constraints, and an insufficient consideration of the control of devices in ADN, as shown in Table I.

TABLE I
COMPARISONS OF CONSIDERED FACTORS IN DIFFERENT REFERENCES

Reference	Privacy protection	Distribution network constraints	Control of devices
[19], [20]	√	–	–
[22], [24]	–	√	–
[23], [25]-[27]	–	–	–
[24], [28], [29]	√	√	–
[4]-[7], [12], [30], [31]	–	√	√
This paper	√	√	√

Note: the symbol √ represents that the corresponding factor is considered; and the symbol – represents that the corresponding factor is not considered.

With all the above, this paper establishes a soft actor-critic algorithm incorporating distributed trading control (SAC-DTC) based on data-driven (deep reinforcement learning (DRL) algorithm) and physical modeling (information-driven distributed algorithm) [32]-[34], which can be applied to coordinate the ADN and P2P markets. The main contributions of this paper are as follows.

1) The coordinated optimization for the power dispatch of ADN and P2P energy trading is constructed as a Markov decision process (MDP) and formulated as a social welfare maximization problem. The agent can explore the dispatch strategy that minimizes the ADN operation cost and creates an environment conducive to conducting P2P energy trading under the stochastic and uncertain conditions.

2) This paper proposes an SAC-DTC algorithm based on data-driven and physical modeling to solve the above problems. This proposed SAC-DTC algorithm utilizes differential privacy noise to protect users' information and price signals to effectively guide users' behavior, thus coupling the coordinated optimization process of ADN and P2P markets, and ultimately reducing the ADN operation cost and increasing the P2P market revenue.

3) The proposed SAC-DTC algorithm is superior in real-time optimization and operation processes of power systems because of its fast computation speed and small node voltage error of the obtained results.

The remainder of this paper is organized as follows. Section II introduces the framework of distribution network that contains both ADN and P2P markets. Section III formulates the optimal power dispatch model of ADN and P2P energy trading model. The proposed SAC-DTC algorithm based on data-driven and physical modeling is presented in Section IV to coordinate the ADN and local P2P market. Section V conducts empirical case studies to evaluate the effectiveness of the proposed SAC-DTC algorithm. Finally, Section VI concludes this paper.

II. FRAMEWORK OF DISTRIBUTION NETWORK CONTAINING BOTH ADN AND P2P MARKETS

As shown in Fig. 1, the proposed framework are applied to distribution networks, where both DSO management areas and autonomous operation areas of prosumers exist. There exists a node set N_{Bus} and a branch set N_{Branch} in the distribution network. At each node i, there are two principal elements: local device managed by the DSO for network loss reduction and voltage control, and agents of prosumers who have been accredited by the DSO.

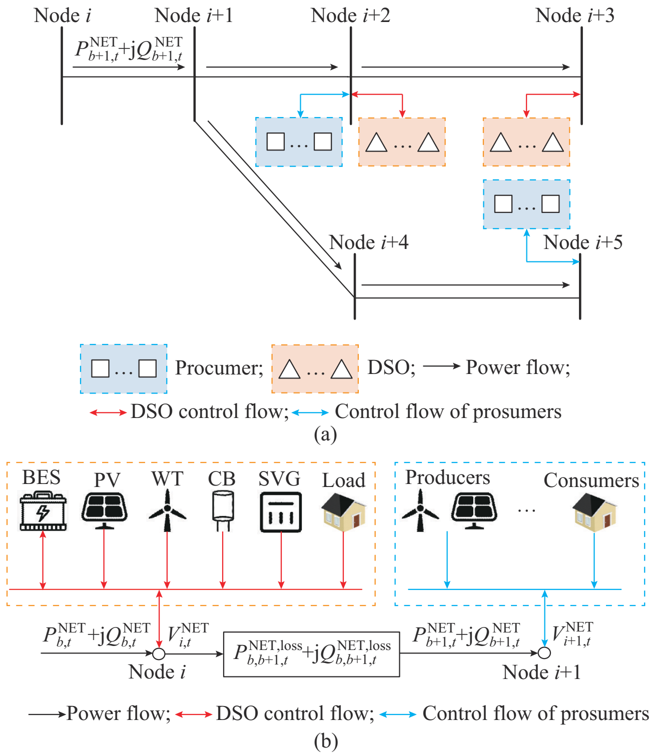

Node i
Node i+1
Node i+2
Node i+3
$P_{b+1,t}^{\text{NET}}$+j$Q_{b+1,t}^{\text{NET}}$
Node i+4
Node i+5
□…□ Procumer; △…△ DSO; ⟶ Power flow;
⟷ DSO control flow; ⟷ Control flow of prosumers
(a)
BES
PV
WT
CB
SVG
Load
Producers
Consumers
…
$P_{b,t}^{\text{NET}}$+j$Q_{b,t}^{\text{NET}}$
$V_{i,t}^{\text{NET}}$
$P_{b,b+1,t}^{\text{NET,loss}}$+j$Q_{b,b+1,t}^{\text{NET,loss}}$
$P_{b+1,t}^{\text{NET}}$+j$Q_{b+1,t}^{\text{NET}}$
$V_{i+1,t}^{\text{NET}}$
Node i
Node i+1
⟶Power flow; ⟷ DSO control flow; ⟷ Control flow of prosumers
(b)

Fig. 1. Proposed framework applied to distribution network. (a) Overall framework. (b) Control areas of DSO and prosumers.

1) DSO: as shown in the red part of Fig. 1(b), node i in this radial ADN contains various DERs such as wind turbines (WTs), photovoltaics (PVs), BESs, SVGs, CBs, and conventional loads. The DSO is tasked with managing the power equipment of ADN. It regulates the active and reactive power outputs to meet electricity demands at the minimum operation costs while ensuring system security.

2) Prosumers: as shown in the blue part of Fig. 1(b), node $i+1$ contains distributed prosumer agents. Prosumers are categorized into two distinct non-empty subsets: producers (with a total number of N_{S}) selling power and consumers (with a total number of N_{B}) purchasing power. Each prosumer coordinates its energy trading with other market participants in the P2P market of ADN, aiming to fulfill individual objectives, adhere to system security constraints, and maximize profits.

The behavior of both DSO and prosumers causes changes in the network losses and node voltages of ADN. Therefore, an efficient coordination and control process between DSOs and prosumers is required to avoid problems such as over-regulation. The optimization process of the whole system seeks to minimize the ADN operation costs of and maximize the profits of all individuals in the P2P market, which is ultimately regarded as a social welfare maximization problem.

III. Problem Formulation

In a radial ADN connected to the external grid, the DSO is responsible for regulating the device in the ADN to ensure that the ADN meets the needs of all users while maintaining a safe and stable operating condition. All two-way users constitute a local P2P energy trading market, where each user can trade electricity and transmit it through the distribution network subjected to safety constraints.

A. Optimal Power Dispatch Model of ADN

1) The objective function for the optimal power dispatch of ADN is to minimize the regulation costs of OLTC, CB, and BES, costs of network losses, and cost of wind power and PV power curtailment, which is formulated as:

$$\min C_t^{\text{ADN}} = \Bigg[C_{\text{DER}} \sum_{i=1}^{N_{\text{DER}}} (P_{i,t}^{\text{DER}} - P_{i,t}^{\text{DER, pre}})^2 + C_{\text{CB}} \sum_{i=1}^{N_{\text{CB}}} T_{i,t}^{\text{CB, loss}} + C_{\text{OLTC}} \sum_{i=1}^{N_{\text{OLTC}}} T_{i,t}^{\text{OLTC, loss}} + C_{\text{NET}} \sum_{i=1}^{N_{\text{line}}} P_{i,t}^{\text{NET, loss}} + C_{\text{BES}} \sum_{i=1}^{N_{\text{BES}}} (P_{i,t}^{\text{BES}})^2 \Bigg] \Delta t \tag{1}$$

$$\begin{cases} T_{i,t}^{\text{CB, loss}} = \left| T_{i,t}^{\text{CB}} - T_{i,t-1}^{\text{CB}} \right| \\ T_{i,t}^{\text{OLTC, loss}} = \left| T_{i,t}^{\text{OLTC}} - T_{i,t-1}^{\text{OLTC}} \right| \end{cases} \tag{2}$$

where C_t^{ADN} is the total ADN operation cost at time t; C_{DER} is the unit cost of wind power and PV power curtailment; C_{CB} and C_{OLTC} are the unit regulation costs of CB and OLTC, respectively; C_{NET} is the grid electricity price; C_{BES} is the unit loss cost of BES; $T_{i,t}^{\text{CB}}$ and $T_{i,t}^{\text{OLTC}}$ are the tap positions of CB and OLTC at node i at time t, respectively; $T_{i,t}^{\text{CB, loss}}$ and $T_{i,t}^{\text{OLTC, loss}}$ are the switching losses of CB and OLTC at node i at time t, respectively [35], [36]; $P_{i,t}^{\text{NET, loss}}$ is the loss of ADN at node i at time t; $P_{i,t}^{\text{BES}}$ is the active power of BES at node i at time t; $P_{i,t}^{\text{DER}}$ and $P_{i,t}^{\text{DER, pre}}$ are the active power output of DERs and its predicted value at node i at time t, respectively; and N_{CB}, N_{OLTC}, N_{BES}, N_{DRE}, and N_{line} are the numbers of CBs, OLTCs, BESs, DREs, and lines, respectively.

2) The following constraints must be included in the optimization model to ensure the safe operation of the ADN with the P2P energy trading process.

$$P_{b+1,t}^{\text{NET}} = P_{b,t}^{\text{NET}} + P_{i,t}^{\text{BES}} + P_{i,t}^{\text{DER}} + P_{i,t}^{\text{BS}} - P_{i,t}^{\text{load}} - P_{b,b+1,t}^{\text{NET, loss}} \tag{3}$$

$$Q_{b+1,t}^{\text{NET}} = Q_{b,t}^{\text{NET}} + Q_{i,t}^{\text{CB}} + Q_{i,t}^{\text{SVG}} + Q_{i,t}^{\text{DER}} + Q_{i,t}^{\text{BS}} - Q_{i,t}^{\text{load}} - Q_{b,b+1,t}^{\text{NET, loss}} \tag{4}$$

$$V_{\min} \le V_{i,t}^{\text{NET}} \le V_{\max} \tag{5}$$

$$V_{i,t}^{\text{NET}} - V_{i-1,t}^{\text{NET}} = (r_b P_{b,t}^{\text{NET}} + x_b Q_{b,t}^{\text{NET}}) / V_{\text{base}} \tag{6}$$

where $P_{b,t}^{\text{NET}}$ and $Q_{b,t}^{\text{NET}}$ are the inflow active and reactive power of branch b at time t, respectively; $P_{b,b+1,t}^{\text{NET, loss}}$ and $Q_{b,b+1,t}^{\text{NET, loss}}$ are the active and reactive power losses from branches b to $b+1$ at time t, respectively; $P_{i,t}^{\text{BS}}$ and $Q_{i,t}^{\text{BS}}$ are the active and reactive power of prosumers at node i at time t, respectively; $Q_{i,t}^{\text{CB}}$, $Q_{i,t}^{\text{SVG}}$, and $Q_{i,t}^{\text{DER}}$ are the reactive power of CB, SVG, and DER at node i at time t, respectively; $P_{i,t}^{\text{load}}$ and $Q_{i,t}^{\text{load}}$ are the active and reactive power of conventional loads at node i at time t, respectively; $V_{i,t}^{\text{NET}}$ is the voltage amplitude at node i at time t; $V_{\min}$ and $V_{\max}$ are the minimum and maximum voltage levels of ADN, respectively; r_b and x_b are the resistance and reactance of branch b, respectively; and V_{base} is the voltage reference value.

The OLTC, CB, SVG, DER, and ESS have their own constraints, which are depicted as (7)-(15), among which (7)-(9) are the operational constraints for the OLTC and CB; (10) and (11) are the operational constraints for the SVG and DER, respectively; and (12)-(15) are the constraints for the ESS.

$$\begin{cases} V_{1,t}^{\text{NET}} = V_{\text{base}}^{\text{OLTC}} + T_t^{\text{OLTC}} \Delta V^{\text{OLTC}} \\ T_{\min}^{\text{OLTC}} \le T_t^{\text{OLTC}} \le T_{\max}^{\text{OLTC}} \end{cases} \tag{7}$$

$$\begin{cases} Q_{i,t}^{\text{CB}} = T_{i,t}^{\text{CB}} \Delta Q^{\text{CB}} \\ T_{\min}^{\text{CB}} \le T_{i,t}^{\text{CB}} \le T_{\max}^{\text{CB}} \end{cases} \tag{8}$$

$$\begin{cases} \sum\limits_{t=1}^{24} \left| T_t^{\text{OLTC}} - T_{t-1}^{\text{OLTC}} \right| \le N_{\max}^{\text{OLTC}} \\ \sum\limits_{t=1}^{24} \left| T_{i,t}^{\text{CB}} - T_{i,t-1}^{\text{CB}} \right| \le N_{\max}^{\text{CB}} \end{cases} \tag{9}$$

$$Q_{\min}^{\text{SVG}} \le Q_{i,t}^{\text{SVG}} \le Q_{\max}^{\text{SVG}} \tag{10}$$

$$\begin{cases} 0 \le P_{i,t}^{\text{DER}} \le P_{i,t,\max}^{\text{DER}} \\ 0 \le Q_{i,t}^{\text{DER}} \le Q_{i,t,\max}^{\text{DER}} \end{cases} \tag{11}$$

$$P_{i,t}^{\text{BES}} = \omega_{i,t}^{\text{BC}} P_{i,t}^{\text{BC}} + \omega_{i,t}^{\text{BD}} P_{i,t}^{\text{BD}} \tag{12}$$

$$\omega_{i,t}^{\text{BC}} + \omega_{i,t}^{\text{BD}} \le 1 \quad \omega_{i,t}^{\text{BC}}, \omega_{i,t}^{\text{BD}} \in \{0, 1\} \tag{13}$$

$$E_{i,t}^{\text{BES}} = E_{i,t-1}^{\text{BES}} + P_{i,t}^{\text{BC}} \eta - P_{i,t}^{\text{BD}} / \eta \tag{14}$$

$$\begin{cases} 0 \le P_{i,t}^{\text{BC}} \le P_{i,\max}^{\text{BC}} \\ 0 \le P_{i,t}^{\text{BD}} \le P_{i,\max}^{\text{BD}} \\ E_{i,t,\min}^{\text{BES}} \le E_{i,t}^{\text{BES}} \le E_{i,t,\max}^{\text{BES}} \end{cases} \tag{15}$$

where $V_{\text{base}}^{\text{OLTC}}$ is the base voltage of OLTC; ΔV^{OLTC} is the voltage change per tap of OLTC; $N_{\max}^{\text{OLTC}}$ is the maximum number of OLTC operations; T_t^{OLTC} is the tap position of OLTC at time t, and $T_{\min}^{\text{OLTC}}$ and $T_{\max}^{\text{OLTC}}$ are its lower and upper bounds, respectively; ΔQ^{CB} is the reactive power change per tap of CB; $N_{\max}^{\text{CB}}$ is the maximum number of CB operations; $T_{i,t}^{\text{CB}}$ is the tap position of CB at node i at time t, and $T_{\min}^{\text{CB}}$ and $T_{\max}^{\text{CB}}$ are its lower and upper bounds, respectively; $Q_{\min}^{\text{SVG}}$ and $Q_{\max}^{\text{SVG}}$ are the minimum and maximum reactive power of SVG, respectively; $P_{i,t,\max}^{\text{DER}}$ and $Q_{i,t,\max}^{\text{DER}}$ are the maximum active and reactive power of DER at node i at time t, respectively; $E_{i,t}^{\text{BES}}$ is the capacity of BES at node i at time t; η is the charging/discharging efficiency; $P_{i,t}^{\text{BC}}$ and $P_{i,t}^{\text{BD}}$ are the charging and discharging power of BES at node i at time t, respectively, and $\omega_{i,t}^{\text{BC}}$ and $\omega_{i,t}^{\text{BD}}$ are their Boolean variables; $E_{i,t,\max}^{\text{BES}}$ and $E_{i,t,\min}^{\text{BES}}$ are the upper and lower bounds of the capacity of BES at node i at time t, respectively; and $P_{i,\max}^{\text{BC}}$ and $P_{i,\max}^{\text{BD}}$ are the maximum charging and discharging power of BES at node i, respectively.

B. P2P Energy Trading Model

P2P energy trading entities need a model for maximizing revenue internally. Prosumers have increasing marginal costs of electricity generation when they act as producers and decreasing marginal benefits of electricity use when they act as consumers. Therefore, the producers' and sellers' electricity consumption behaviors can be characterized by a quadratic function [37]. The total revenue of prosumers is composed of three terms: the power utility benefit of prosumers, the active electricity cost, and the reactive electricity cost.

$$\max \sum_{i=1}^{N_{\text{B}}+N_{\text{S}}} U_{i,t}^{\text{P2P}} = \sum_{i=1}^{N_{\text{B}}+N_{\text{S}}} \left(u_{i,t} - \delta_{i,t}^{\text{PLMP}} P_{i,t}^{\text{BS}} - \delta_{i,t}^{\text{QLMP}} Q_{i,t}^{\text{BS}} \right) \tag{16}$$

$$u_{i,t} = \varepsilon_{i,t} (P_{i,t}^{\text{BS}})^2 + \beta_{i,t} P_{i,t}^{\text{BS}} + \tau_{i,t} (Q_{i,t}^{\text{BS}} - Q_{i,t-1}^{\text{BS}})^2 \tag{17}$$

where $U_{i,t}^{\text{P2P}}$ is the total revenue of prosumer at node i at time t; $u_{i,t}$ is the function of power utility benefits of prosumer at node i at time t; $\varepsilon_{i,t}$, $\beta_{i,t}$, and $\tau_{i,t}$ are the power utility parameters of prosumers, which are private information; and $\delta_{i,t}^{\text{PLMP}}$ and $\delta_{i,t}^{\text{QLMP}}$ are the marginal tariffs for active and reactive power at node i at time t, respectively.

In addition, the trading results need to satisfy the ADN security constraints as well as the market supply and demand balance constraints, which are shown as:

$$\sum_{i=1}^{N_{\text{B}}+N_{\text{S}}} \left(P_{i,t}^{\text{BS}} + P_{i,t}^{\text{BS, BES}} \right) - \sum_{b=1}^{N_{\text{line}}} P_{b,b+1,t}^{\text{P2P, loss}} = 0 \tag{18}$$

$$\sum_{i=1}^{N_{\text{B}}+N_{\text{S}}} Q_{i,t}^{\text{BS}} - \sum_{b=1}^{N_{\text{line}}} Q_{b,b+1,t}^{\text{P2P, loss}} = 0 \tag{19}$$

$$V_{\min} \le \Delta V_{i,t}^{\text{P2P}} \le V_{\max} \tag{20}$$

$$\begin{cases} \underline{P} \le P_{i,t}^{\text{BS}} \le \bar{P} \\ \underline{Q} \le Q_{i,t}^{\text{BS}} \le \bar{Q} \end{cases} \tag{21}$$

where $P_{i,t}^{\text{BS, BES}}$ is the active power of the prosumer's own BES; $P_{b,b+1,t}^{\text{P2P, loss}}$ and $Q_{b,b+1,t}^{\text{P2P, loss}}$ are the network active and reactive power losses from branches b to $b+1$ at time t caused by the P2P energy trading, respectively; $\Delta V_{i,t}^{\text{P2P}}$ is the amount of voltage amplitude change caused by the P2P energy trading; $\bar{P}$ and $\underline{P}$ are the upper and lower limits of active power regulation for prosumers, respectively; and $\bar{Q}$ and $\underline{Q}$ are the upper and lower limits of reactive power regulation for prosumers, respectively.

The ADN cannot access the specific power consumption information of prosumers for privacy protection and market fairness. Therefore, we decompose the original problem into multiple subproblems, thus facilitating the subsequent solution using a distributed approach.

The changes in active and reactive power for each prosumer impact the network losses and nodal voltages. Consequently, we incorporate all constraints into the electricity efficiency function for prosumer $w_{i,t}$ and differentiate it to determine the marginal tariffs for active and reactive power [33].

$$\begin{aligned} w_{i,t} = {} & u_{i,t} + \mu_{i,t}^{V\min} \left(\Delta V_{i,t}^{\text{P2P}} - V_{\min} \right) + \mu_{i,t}^{V\max} \left(V_{\max} - \Delta V_{i,t}^{\text{P2P}} \right) + \\ & \mu_{i,t}^{P} \left[\sum_{i=1}^{N_{\text{B}}+N_{\text{S}}} \left(P_{i,t}^{\text{BS}} + P_{i,t}^{\text{BS, BES}} \right) - \sum_{b=1}^{N_{\text{line}}} P_{b,b+1,t}^{\text{P2P, loss}} \right] + \\ & \mu_{i,t}^{Q} \left(\sum_{i=1}^{N_{\text{B}}+N_{\text{S}}} Q_{i,t}^{\text{BS}} - \sum_{b=1}^{N_{\text{line}}} Q_{b,b+1,t}^{\text{P2P, loss}} \right) + \mu_{i,t}^{P\max} \left(P_{i,t}^{\text{BS}} - \underline{P} \right) + \\ & \mu_{i,t}^{P\min} \left(\bar{P} - P_{i,t}^{\text{BS}} \right) + \mu_{i,t}^{Q\max} \left(Q_{i,t}^{\text{BS}} - \underline{Q} \right) + \mu_{i,t}^{Q\min} \left(\bar{Q} - Q_{i,t}^{\text{BS}} \right) \end{aligned} \tag{22}$$

$$\begin{cases}\delta_{i,t}^{\mathrm{PLMP}}=\partial w_{i,t}/\partial P_{i,t}^{\mathrm{BS}}\\ \delta_{i,t}^{\mathrm{QLMP}}=\partial w_{i,t}/\partial Q_{i,t}^{\mathrm{BS}}\end{cases} \tag{23}$$

where $\mu_{i,t}^{V\max}$ and $\mu_{i,t}^{V\min}$ are the dual variables corresponding to the upper and lower voltage constraints at node i at time t, respectively; $\mu_{i,t}^{P}$ and $\mu_{i,t}^{Q}$ are the dual variables corresponding to the active and reactive power balance constraints at node i at time t, respectively; $\mu_{i,t}^{P\max}$ and $\mu_{i,t}^{P\min}$ are the dual variables corresponding to the upper and lower active power constraints at node i at time t, respectively; and $\mu_{i,t}^{Q\max}$ and $\mu_{i,t}^{Q\min}$ are the dual variables corresponding to the upper and lower reactive power constraints at node i at time t, respectively.

IV. Proposed SAC-DTC Algorithm

During the ADN dispatching and P2P energy trading, if we do not consider the impact on the system, we may reach a trading and controlling result that violates the system operation constraints, ultimately leading to device failure or system instability. Therefore, we propose the SAC-DTC algorithm to coordinate the optimization process between the ADN and the P2P market to achieve the global optimum within a solution space that ensures the voltage levels safety. The objective is to minimize the ADN operation cost (including regulation costs of device and costs of network loss, etc.) and maximize the P2P market revenue, while ensuring the safe operation of the system.

The proposed SAC-DTC algorithm is a new type of algorithm by combining DRL algorithm and distributed control computing. The structure of the proposed SAC-DTC algorithm is shown in Fig. 2. It should be noted that the proposed SAC-DTC algorithm continues the learning process during the online operation.

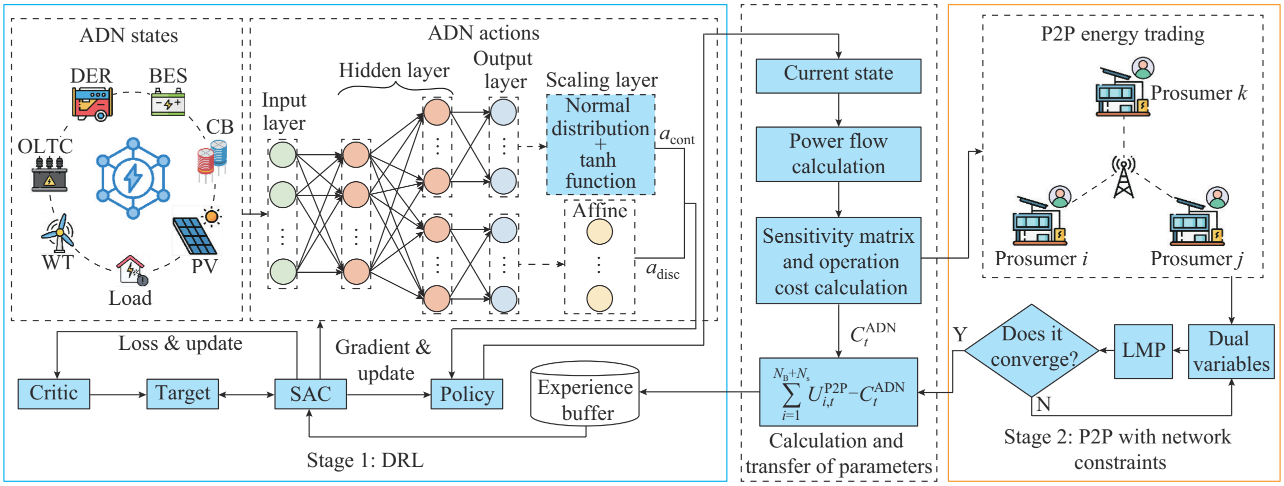


Fig. 2. Structure of proposed SAC-DTC algorithm.

The optimization process of ADN and P2P market can be modeled using the MDP, as shown in Fig. 3.

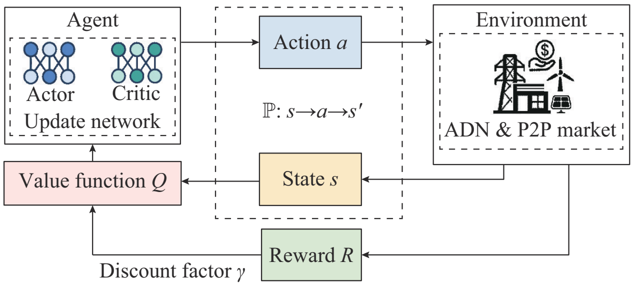


Fig. 3. Optimization process of ADN and P2P market using MDP.

First, the agent gives the optimal action a of each device in the ADN based on the local state s. Then, it calculates the network loss and node voltage in the ADN and issues the information to the P2P market. Subsequently, the prosumers adjust the output according to their interests and return the profits to ADN after differential privacy encryption processing. Finally, ADN calculates the reward value R based on (1) and (16), and then puts the data into the experience buffer pool $\mathcal{D}$ to update the network parameters.

$$R=\sum_{i=1}^{N_B+N_s} U_{i,t}^{\mathrm{P2P}}-C_t^{\mathrm{ADN}} \tag{24}$$

The MDP consists of five key elements: state space s, action space a, state transfer probability $\mathbb{P}$, reward function R, and discount factor γ, represented by $\langle s,a,\mathbb{P},R,\gamma\rangle$.

A. Continuous-discrete Hybrid SAC

For the reinforcement learning in continuous-discrete hybrid action space, assuming that there are n discrete devices, each with m_n actions, the output action dimension of the state-action value function Q will be $\prod_{i=1}^{n} m_i$. The action dimension will grow exponentially as the number of devices n increases. If a separate Q value is estimated for each possible combination of actions, the data required to be calculated and stored will grow rapidly and fall into a curse of dimensionality. Therefore, inspired by [38], we use a separate head for each discrete device. The head is only responsible for calculating the Q value associated with the device actions, which is expressed as:

$$Q(s,a)=c_0(s)+\sum_{i=1}^{n} c_i(s)Q_i(s,a) \tag{25}$$

where Q_i is the Q value of device i; and c_0 and c_i are the shared base value and the state parameters of device i, respectively.

During the training process, the formula for calculating the network target value y is:

$$y=R+\gamma\left[c_0(s')+\sum_{i=1}^{n}c_i(s')Q_{\text{target},i}^{\text{disc}}(s',a'_{\text{disc},i})+Q_{\text{target}}^{\text{cont}}(s',a'_{\text{cont}})-\right.$$
$$\left.\alpha\left(\log\pi_{\text{cont}}(a'_{\text{cont}}|s')+\sum_{i=1}^{n}\pi_{\text{disc},i}(a'_{\text{disc},i}|s')\log\pi_{\text{disc},i}(a'_{\text{disc},i}|s')\right)\right] \tag{26}$$

where s' is the new state; α is the temperature parameter used to control the contribution of entropy in the policy update; a'_{cont} and $a'_{\text{disc},i}$ are the continuous and discrete actions of the new state, respectively; $Q_{\text{target}}^{\text{cont}}$ and $Q_{\text{target},i}^{\text{disc}}$ are the state-action value functions; and π_{cont} and $\pi_{\text{disc},i}$ are the strategy functions.

The parameters of the critic network are updated by minimizing the mean square error J_Q between the predicted Q value of the critic network $Q_{\varphi_{\text{critic}}}$ and the target value y. Then, the parameters of actor network are updated by minimizing the loss function J_π:

$$J_Q=\mathop{\mathbb{E}}_{(s,a,R,s')\sim\mathcal{D}}\left(\frac{1}{2}(Q_{\varphi_{\text{critic}}}(s,a)-y)^2\right) \tag{27}$$

$$J_\pi=\mathop{\mathbb{E}}_{s\sim\mathcal{D},a\sim\pi_{\varphi_{\text{actor}}}}(\alpha\log\pi_{\varphi_{\text{actor}}}(a|s)-\min Q_{\varphi_{\text{critic}}}(s,a)) \tag{28}$$

where $\pi_{\varphi_{\text{actor}}}$ represents the probability of taking action a given state s under the policy parameterized by φ_{actor}.

Finally, the training network is slowly tracked by a soft update method:

$$\varphi_{\text{target}}\leftarrow\theta\varphi+(1-\theta)\varphi_{\text{target}} \tag{29}$$

where $\varphi=\varphi_{\text{actor}}$ or φ_{critic} is the training network parameter; φ_{target} is the target network parameter; and θ is the soft update rate.

B. Distributed Trading Control (DTC)

For the prosumers at each node, adjusting the active and reactive power during the energy trading process will bring changes to their benefits or costs as well as the node voltage and network loss. Therefore, in this paper, based on the dual ascent method of sensitivity calculation [33], we calculate the revenues of prosumers during the energy trading process and assess the impact on the ADN operation cost. The change of state of ADN is linearized as:

$$\Delta\boldsymbol{Z}(s_t)=\begin{bmatrix}\boldsymbol{g}_V\\ \boldsymbol{g}_P\\ \boldsymbol{g}_Q\end{bmatrix}[\Delta\boldsymbol{P}_t^{\text{BS}}\quad\Delta\boldsymbol{Q}_t^{\text{BS}}] \tag{30}$$

$$\boldsymbol{g}_V=\begin{bmatrix}\dfrac{\partial V_{1,t}}{\partial P_{1,t}} & \cdots & \dfrac{\partial V_{1,t}}{\partial P_{N_{\text{B}}+N_{\text{S}},t}} & \dfrac{\partial V_{1,t}}{\partial Q_{1,t}} & \cdots & \dfrac{\partial Q_{1,t}}{\partial Q_{N_{\text{B}}+N_{\text{S}},t}}\\ \vdots & & \vdots & \vdots & & \vdots\\ \dfrac{\partial V_{N_{\text{node}},t}}{\partial P_{1,t}} & \cdots & \dfrac{\partial V_{N_{\text{node}},t}}{\partial P_{N_{\text{B}}+N_{\text{S}},t}} & \dfrac{\partial V_{N_{\text{node}},t}}{\partial Q_{1,t}} & \cdots & \dfrac{\partial Q_{N_{\text{node}},t}}{\partial Q_{N_{\text{B}}+N_{\text{S}},t}}\end{bmatrix} \tag{31}$$

$$\boldsymbol{g}_P=\begin{bmatrix}\dfrac{\partial P_{1,t}}{\partial P_{1\text{oss},t}} & \cdots & \dfrac{\partial P_{N_{\text{B}}+N_{\text{S}},t}}{\partial P_{1\text{oss},t}} & \dfrac{\partial P_{1,t}}{\partial Q_{1\text{oss},t}} & \cdots & \dfrac{\partial P_{N_{\text{B}}+N_{\text{S}},t}}{\partial Q_{1\text{oss},t}}\end{bmatrix} \tag{32}$$

$$\boldsymbol{g}_Q=\begin{bmatrix}\dfrac{\partial Q_{1,t}}{\partial P_{1\text{oss},t}} & \cdots & \dfrac{\partial Q_{N_{\text{B}}+N_{\text{S}},t}}{\partial P_{1\text{oss},t}} & \dfrac{\partial Q_{1,t}}{\partial Q_{1\text{oss},t}} & \cdots & \dfrac{\partial Q_{N_{\text{B}}+N_{\text{S}},t}}{\partial Q_{1\text{oss},t}}\end{bmatrix} \tag{33}$$

$$\Delta\boldsymbol{P}_t^{\text{BS}}=[\Delta P_{1,t}^{\text{BS}}\quad\Delta P_{2,t}^{\text{BS}}\quad\cdots\quad\Delta P_{N_{\text{B}}+N_{\text{S}},t}^{\text{BS}}]^{\text{T}} \tag{34}$$

$$\Delta\boldsymbol{Q}_t^{\text{BS}}=[\Delta Q_{1,t}^{\text{BS}}\quad\Delta Q_{2,t}^{\text{BS}}\quad\cdots\quad\Delta Q_{N_{\text{B}}+N_{\text{S}},t}^{\text{BS}}]^{\text{T}} \tag{35}$$

where $\Delta\boldsymbol{Z}(s_t)$ is the state change matrix function; $\boldsymbol{g}_V\in\mathbb{R}^{N_{\text{node}}\times2(N_{\text{B}}+N_{\text{S}})}$, $\boldsymbol{g}_P\in\mathbb{R}^{1\times2(N_{\text{B}}+N_{\text{S}})}$, and $\boldsymbol{g}_Q\in\mathbb{R}^{1\times2(N_{\text{B}}+N_{\text{S}})}$ are the linear mapping functions for the node voltage, active power loss of ADN, and reactive power loss of ADN, respectively; and $\Delta\boldsymbol{P}_t^{\text{BS}}$ and $\Delta\boldsymbol{Q}_t^{\text{BS}}$ are the vectors of active power and reactive power adjustments during energy trading for the prosumers, respectively.

The mapping function can be fitted based on a neural network, but this requires a separate neural network for each variable and constraint, which will also fall into the curse of dimensionality. Therefore, in this paper, we utilize the sensitivity matrix as an equivalent alternative to the mapping function and validate the accuracy of the solution. The original problem (16)-(23) in the P2P market is transformed into a quadratic programming problem as:

$$\begin{cases}\min U_t^{\text{P2P}}=-\dfrac{1}{2}\boldsymbol{x}^{\text{T}}\boldsymbol{Mx}-\boldsymbol{H}^{\text{T}}\boldsymbol{x}\\ \text{s.t. }\boldsymbol{Ax}\le\boldsymbol{B}\end{cases} \tag{36}$$

$$\begin{cases}\boldsymbol{x}=\begin{bmatrix}\boldsymbol{P}_t^{\text{BS}}+\Delta\boldsymbol{P}_t^{\text{BS}}\\ \boldsymbol{Q}_t^{\text{BS}}+\Delta\boldsymbol{Q}_t^{\text{BS}}\end{bmatrix}\\ \boldsymbol{M}=[\boldsymbol{m}_P\quad\boldsymbol{m}_Q]^{\text{T}}\\ \boldsymbol{H}=[\boldsymbol{h}_P\quad\boldsymbol{h}_Q]^{\text{T}}\\ \boldsymbol{A}=[\boldsymbol{g}_V\quad-\boldsymbol{g}_V\quad\boldsymbol{g}_P\quad\boldsymbol{g}_Q\quad\boldsymbol{1}_{1\times4(N_{\text{B}}+N_{\text{S}})}]^{\text{T}}\\ \boldsymbol{B}=[\bar{\boldsymbol{V}}\quad\underline{\boldsymbol{V}}\quad\boldsymbol{0}\quad\boldsymbol{0}\quad\bar{\boldsymbol{P}}\quad\underline{\boldsymbol{P}}\quad\bar{\boldsymbol{Q}}\quad\underline{\boldsymbol{Q}}]^{\text{T}}\end{cases} \tag{37}$$

where the matrix parameters $\boldsymbol{m}_P$, $\boldsymbol{m}_Q$, $\boldsymbol{h}_P$, and $\boldsymbol{h}_Q$ are extracted from the objective function for prosumers shown in (17); $\bar{\boldsymbol{V}}$ and $\underline{\boldsymbol{V}}$ are the upper and lower matrices of node voltages, respectively; $\bar{\boldsymbol{P}}$ and $\underline{\boldsymbol{P}}$ are the matrices of upper and lower active power for prosumers, respectively; and $\bar{\boldsymbol{Q}}$ and $\underline{\boldsymbol{Q}}$ are the matrices of upper and lower reactive power for prosumers, respectively.

The dual function is:

$$\inf_{\boldsymbol{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\boldsymbol{x}^{\text{T}}\boldsymbol{Mx}+\boldsymbol{Hx}+\boldsymbol{\mu}^{\text{T}}(\boldsymbol{Ax}-\boldsymbol{B})\right\} \tag{38}$$

The lower definitive bound for this problem is taken at $\boldsymbol{x}=-\boldsymbol{M}^{-1}(\boldsymbol{H}+\boldsymbol{A}^{\text{T}}\boldsymbol{\mu})$. By disregarding the constant term and changing the sign of the objective function, the maximization problem is transformed into a minimization problem to obtain the dyadic problem as:

$$\begin{cases}\min d=\dfrac{1}{2}\boldsymbol{\mu}^{\text{T}}\boldsymbol{AM}^{-1}\boldsymbol{A}^{\text{T}}\boldsymbol{\mu}+(\boldsymbol{B}+\boldsymbol{AM}^{-1}\boldsymbol{H})^{\text{T}}\boldsymbol{\mu}\\ \text{s.t. }\boldsymbol{\mu}=[\mu_{i,t}^{V\max},\mu_{i,t}^{V\min},\mu_{i,t}^{P},\mu_{i,t}^{Q},\mu_{i,t}^{P\max},\mu_{i,t}^{P\min},\mu_{i,t}^{Q\max},\mu_{i,t}^{Q\min}]^{\text{T}}\ge\boldsymbol{0}\end{cases} \tag{39}$$

where $\boldsymbol{\mu}$ is the vector of Lagrangian multipliers associated with the constraints.

When the original problem is convex, we can find the gradient of $\boldsymbol{A}$ for its dual problem and obtain:

$$\nabla d = \boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{B} + \boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{H} \tag{40}$$

$$d(\boldsymbol{\mu}') \le d(\boldsymbol{\mu}) + (\nabla d(\boldsymbol{x}))^{\mathrm{T}}(\boldsymbol{\mu}' - \boldsymbol{\mu}) + \frac{L}{2}\|\boldsymbol{\mu}' - \boldsymbol{\mu}\|^2 \tag{41}$$

For any two points in the dual function d, the value of function d is at least the linear approximation minus a quadratic term, which depends on the distance between the two points and Lipschitz constant. Lipschitz constant should be the largest eigenvalue of $\boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{A}^{\mathrm{T}}$. This is because in the quadratic functions, the largest eigenvalue of the matrix determines the maximum curvature. According to [33], the generalized Lipschitz constant matrix $\boldsymbol{L} \succcurlyeq \boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{A}^{\mathrm{T}}$ is used to determine the step size in the dual ascent method, where $\boldsymbol{L}$ is set to be a diagonal matrix. By minimizing the trace of $\boldsymbol{L}$, the semi-positive definite programming problem can be solved.

After solving the dual problem (40), the optimal power for each prosumer is obtained, which is then substituted into (16) to obtain the maximum welfare for each prosumer $U_{i,t}^{\mathrm{P2P*}}$. To protect the privacy of prosumers, a differential privacy technique is used. This involves adding random noise to the data through Laplace-distributed sampling $Lap(\cdot)$, as expressed in (42). The noise is then returned to the agent for learning as part of the reward.

$$U_t^{\mathrm{P2P}} = \sum_{i=1}^{N_{\mathrm{B}}+N_{\mathrm{S}}} \left(U_{i,t}^{\mathrm{P2P*}} + Lap\left(\frac{\Delta f}{\epsilon}\right) \right) \tag{42}$$

where Δf is the sampling sensitivity, representing the maximum variation that $U_{i,t}^{\mathrm{P2P*}}$ may experience; and ϵ is the privacy strength parameter, whose value is smaller for stronger privacy protection.

Since the noise is random and its mathematical expectation is 0, the effects of the noise are canceled when aggregating large amounts of data. This ensures that the statistical estimation of total P2P market revenues remains accurate.

The detailed calculation procedure of the proposed SAC-DTC algorithm is explained in Algorithm 1.

Algorithm 1: detailed calculation procedure of proposed SAC-DTC algorithm

```
S1: Initialize $\varphi_{\mathrm{actor}}$, $\varphi_{\mathrm{critic}}$, $\varphi_{\mathrm{target}}$, $\theta$, $\mathcal{D}=\varnothing$, $\boldsymbol{\mu}=\boldsymbol{0}$, time step $\Delta t=1$ hour, and
    the maximum time step $T=24$ hours
S2: Repeat
S3: for $t=1:\Delta t:T$ do
S4:     $a \sim \pi(a|s)$
S5:     Calculate power flow
S6:     Release $V_{i,t}^{\mathrm{NET}}$, $\boldsymbol{g}_V$, $\boldsymbol{g}_P$, $\boldsymbol{g}_Q$, and locational marginal price (LMP) to prosumers
S7:     Solve (16) for each prosumer
S8:     Update $\boldsymbol{\mu}$ and LMP
S9:     Update $a$, $s'$, and $R$, and store $[s, a, R, s']$ in $\mathcal{D}$
S10: end for
S11: Update $\varphi_{\mathrm{actor}}$ and $\varphi_{\mathrm{critic}}$ using (25)-(27)
S12: Update $\varphi_{\mathrm{target}}$ using (28)
S13: end
```

V. Case Studies

A. System Setting

This paper evaluates the proposed SAC-DTC algorithm using the IEEE 33-node system. We assumes that five prosumers participate in the P2P energy trading, and the basic parameters of the utility function can be found in [33]. The training process involves base loads at all nodes of this distribution network, with load data values originating from a regional grid in southern China over a time span of 1000 randomly selected days. According to [39], the upper and lower limits of node voltage amplitude are set to be 1.06 and 0.94 p.u., respectively.

Three operation models are set up to compare the effectiveness in reducing the ADN operation cost and improving the P2P market revenue.

Model 1: without considering voltage constraints, the ADN operation cost is minimized as the objective function for optimization, and the P2P market is optimized with the objective function of maximizing the operation revenue.

Model 2: based on Model 1, the system voltage constraints are further considered, and the P2P market is optimized for operation based on the method in [33].

Model 3: as illustrated in Section III, the voltage constraints are considered and the total social welfare of the sum of P2P market revenue and ADN operation cost is taken as the objective function, the joint optimization is run using the proposed SAC-DTC algorithm.

B. Convergence Performance

There have been several studies applying DRL algorithms to the power system domain. In this subsection, we focus on comparing the SAC algorithm with the widely-used DDPG and PPO algorithms. All the three DRL algorithms utilize an actor-critic architecture. The DDPG algorithm employs a deterministic strategy network (actors) to directly predict actions and evaluates the expected returns of these actions through a value network (critics). In contrast, the PPO algorithm ensures the stability and convergence of policy updates by introducing a clip loss function that limits the magnitude of these updates, while the SAC algorithm encourages broader exploration by increasing policy entropy. The hyperparameters are shown in Tables II-IV. The Ornstein-Uhlenbeck noises are provided in [10] and [41].

TABLE II
Common Hyperparameters for Three DRL Algorithms

Hyperparameter	Value	Hyperparameter	Value
Architecture of actor and critic networks	[256, 256]	Activation function	ReLU
Optimizer	Adam	Discount factor	0.99
Actor learning rate	1×10^{-3}	T	24 hours
Critic learning rate	5×10^{-4}	Δt	1 hour
Minibatch size	64	Evaluation frequency	3

Figure 4 demonstrates the training performance using SAC-DTC, DDPG-DTC, and PPO-DTC algorithms in the IEEE 33-node system, where the shaded area represents the range of fluctuation of these algorithms over the course of multiple training sessions. It can be observed that the proposed SAC-DTC algorithm performs better in reducing the ADN operation cost. As for the P2P market revenue, all the three algorithms show similar convergence, mainly attributed to the effectiveness of DTC. Overall, the proposed SAC-DTC algorithm outperforms both DDPG-DTC and PPO-DTC algorithms regarding the training speed and final results, indicating its potential advantages in power system optimization.

TABLE III
INDEPENDENT HYPERPARAMETERS FOR SAC AND DDPG ALGORITHMS

Hyperparameter	Value	
	SAC algorithm	DDPG algorithm
Target network update rate	0.005	0.005
Replay buffer size	5×10^5	5×10^5
Entropy coefficient	Auto	
Noise type		Ornstein-Uhlenbeck

TABLE IV
INDEPENDENT HYPERPARAMETERS FOR PPO ALGORITHMS

Hyperparameter	Value
Value function coefficient	0.5
Generalized advantage estimation Lambda	0.95
Clip ratio	0.2
Number of epochs	3
Gradient clipping	0.1

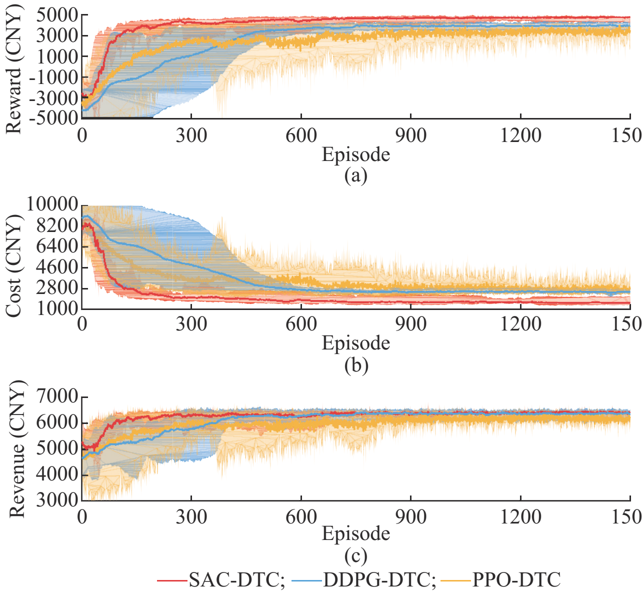


Fig. 4. Training performance using SAC-DTC, DDPG-DTC, and PPO-DTC algorithms in IEEE 33-node system. (a) Total reward. (b) ADN operation cost. (c) P2P market revenue.

C. AND Operation Costs and P2P Market Revenue

The results of the three operation models are presented in Table V.

Figure 5 shows the node voltage comparisons of the three operation models. Figure 6 illustrates the comparison of ADN operation costs and P2P market revenues with the three operation models.

TABLE V
RESULTS OF THREE OPERATION MODELS

Model	ADN operation cost (CNY)	P2P market revenue (CNY)	Number of voltage violations	The maximum voltage difference (p.u.)
Model 1	1054	6491	188	0.12350
Model 2	1615	5561	0	0.07934
Model 3	1481	6283	0	0.07254

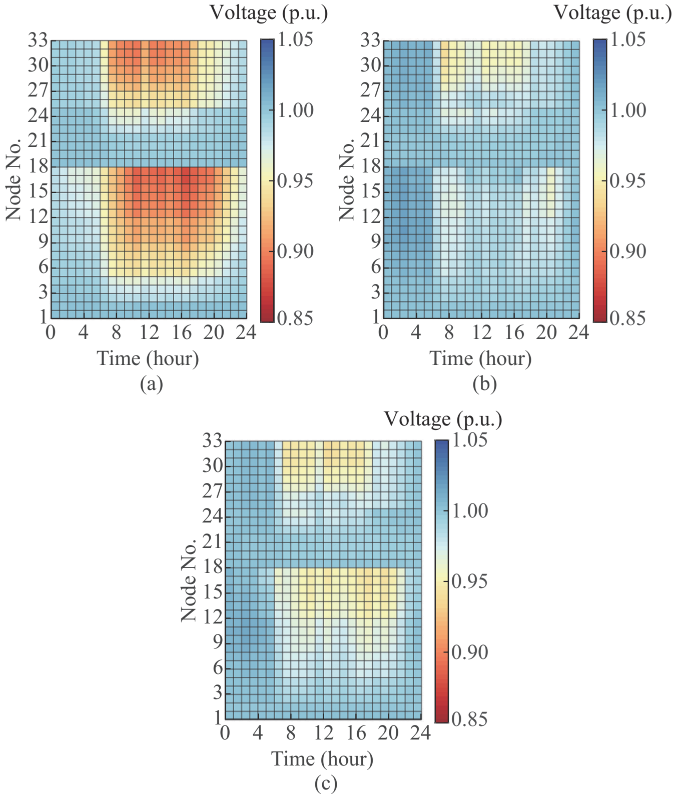


Fig. 5. Node voltage comparison of three operation models. (a) Model 1. (b) Model 2. (c) Model 3.

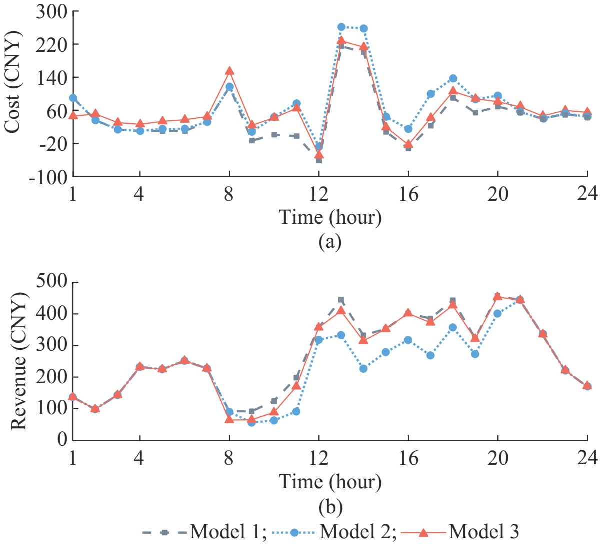
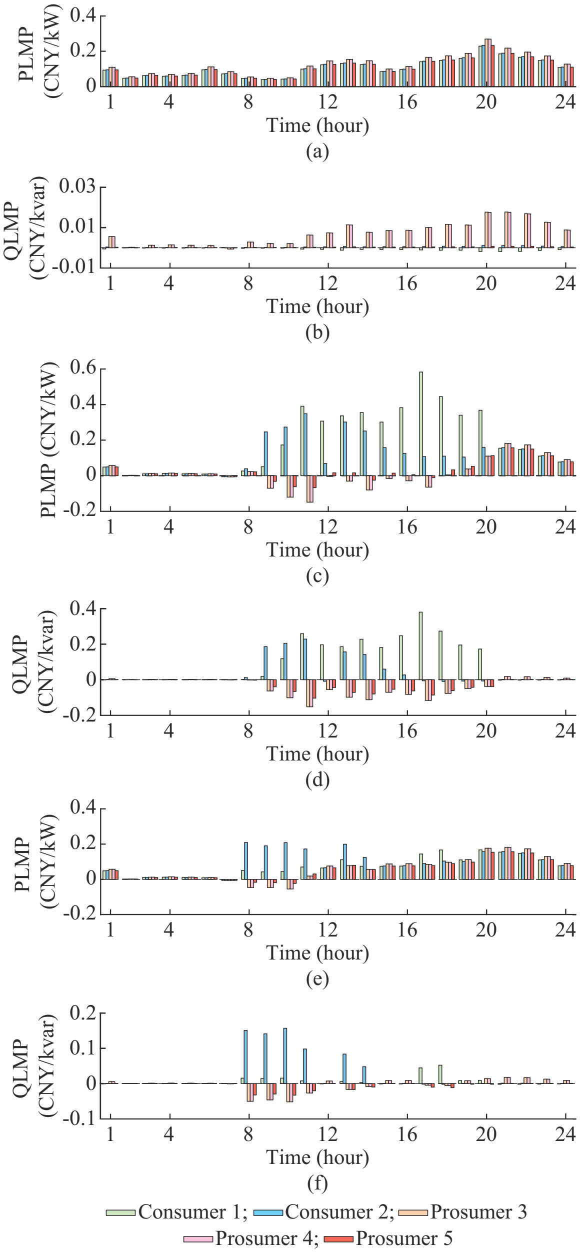


Fig. 6. Comparison of ADN operation costs and P2P market revenues.(a) ADN operation cost. (b) P2P market revenue.

From a system security perspective, during hours 8-20, Model 1 exhibits the largest voltage fluctuation deviation, with several node voltages crossing the lower limit at various time points. However, during other periods, the system does not experience voltage crossings. Model 2 and 3 are able to operate safely throughout all periods because the voltage constraints are considered in the optimization process of the ADN and P2P markets. In Model 2, the optimization process of ADN and P2P markets operates independently, and the lower bound of system voltage is generally higher than that in Model 3, but the maximum voltage variation is greater.

Additionally, Fig. 7 shows the comparison results of LMP.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 7. Comparison results of LMPs. (a) PLMP in Model 1. (b) QLMP in Model 1. (c) PLMP in Model 2. (d) QLMP in Model 2. (e) PLMP in Model 3. (f) QLMP in Model 3.

From Figs. 6, 7(a), and 7(b), it can be observed that when the voltage constraints are not considered in Model 1, ADN does not need to regulate the actions. Each prosumer only needs to fine-tune its output value according to the active and reactive power balance constraints of P2P market. Hence, the differences in the LMPs for the active power and reactive power, i.e., PLMP and QLMP, respectively, of each node are minor, and all PMLPs are positive. Producers 1 and 2 have negative active power values, absorb energy from the P2P market, and pay for the cost of electricity consumption. Consumers 3-5 have positive active power values, supply energy to the P2P market, and receive revenues from electricity sales. At this point, ADN has the lowest cost, and P2P market has the highest revenue.

As shown in Fig. 7(c) and (d), during hours 0-7 and 21-24, the energy trading between producers and consumers is constrained by the voltage limitations in Model 2. The PMLP decreases, which reduces the size of the energy trading between producers and consumers. During hours 8-20, the PLMP and QLMP of consumers increase significantly due to the voltage limitation constraints, leading consumers to reduce their power consumption. The PLMP and QLMP of producers decrease dramatically to negative values due to the reduced power consumption of consumers. To maintain the power balance, the PLMP guides the producer to reduce the amount of electricity sale using a negative price signal. The effectiveness of the LMP mechanism can be illustrated by comparing the changes in PLMP and QLMP in Model 1 and Model 2. The P2P market can utilize economic instruments to efficiently dispatch the active and reactive power for each prosumer, thereby mitigating the voltage crossing the lower limit. The results of Model 1 and Model 2 in Fig. 6(a) and (b) are almost the same during hours 0-7 and 21-24. This is because the network constraints are met in both models. However, in Model 2, when there is a voltage overrun during hours 8-20, due to the lack of complete information in the ADN and P2P markets, the two parties can only supervise their internal devices independently to ensure the safe operation. This leads to an increase in the ADN operation cost by 53.2% and an increase in the P2P regulation cost by 14.3% compared with Model 1.

In Model 3, based on the proposed SAC-DTC algorithm, the encrypted information can be shared between the ADN and the prosumers. The system security regulation cost can be effectively shared with the ADN and each prosumer. As can be observed in Fig. 7(e) and (f), the PLMP and QLMP changes of prosumers in Model 3 are much less drastic than those in Model 2, which are similar during hours 0-7 and 21-24. However, during hours 8-20, the PLMP and QLMP of consumers in Model 3 are overall lower than those in Model 2, indicating that consumers are able to purchase electricity in the P2P market at a lower cost. Producers, on the other hand, have an overall increase in PLMP and QLMP, indicating that producers can supply electricity to the P2P market at a higher price and make higher profits. The ADN operation costs increase during certain time periods due to the earlier adjustment of device. On the premise of ensuring the system safe operation, compared with Model 2, the ADN operation cost of Model 3 is reduced by 8.3%, the P2P market revenue increases by 12.9%, and the maximum voltage difference is minimized, making the system operate more stable. The accumulated savings in ADN operation costs for the whole year amount to 49000 CNY, and the P2P market revenue increases by 264000 CNY.

Overall, the joint optimization of ADN and P2P markets can reduce the feeder voltage drop and avoid violating the voltage constraints. Meanwhile, the economic cost paid by the market members to ensure system security in Model 3 is much smaller than that in Model 2 and close to that in Model 1. For all members in the ADN, the system security status should be the primary. Therefore, this paper concludes that trading a smaller economic cost for safer system operation is reasonable.

D. Comparison of Proposed SAC-DTC Algorithm with Centralized Algorithm

In order to verify the accuracy and scalability of the proposed SAC-DTC algorithm, its computational results are compared with those of the mixed-integer second-order cone programming (MISOCP) based centralized algorithm in IEEE 33-, 69-, and 136-node systems, with the specific settings shown in Table VI. The MISOCP model is a mathematical optimization technique that integrates integer variables into the second-order cone programming (SOCP) model, making it particularly suitable for complex power system applications. The validity and accuracy of the MISOCP model have been widely demonstrated, with a speedup ratio of about six times that of the traditional optimal power flow (OPF) model for small-scale test systems [9]. Since the SOCP model is convex, the global optimal solution can be guaranteed, further enhancing the reliability of the MISOCP model in practical applications. The comparison in ADN operation costs and P2P market revenues calculated by the proposed SAC-DTC algorithm and the MISOCP-based centralized algorithm are shown in Table VII. From the perspective of ADN operation costs and P2P market revenues, while the strategy optimization of the proposed SAC-DTC algorithm for complex systems may converge toward the global optimum [10], it exhibits some discrepancies compared with the results from MISOCP-based centralized algorithm. However, with the increase in the number of devices and prosumers in P2P market and ADN, the resources and computation time using MISOCP-based centralized algorithm increase exponentially, and it cannot meet the requirements of timeliness in the electricity market. In addition, all prosumers must communicate bi-directionally and share sufficient information with the central organization, which places high demands on the communication system and does not guarantee data privacy [28].

TABLE VI
SPECIFIC SETTING OF IEEE 33-, 69-, AND 136-NODE SYSTEMS

System	Number of prosumers	Number of CBs	Number of SVGs	Number of DERs	Number of ESSs
IEEE 33-node	5	2	2	2	1
IEEE 69-node	28	4	5	2	1
IEEE 136-node	40	6	8	8	2

TABLE VII
COMPARISON IN ADN OPERATION COSTS, P2P MARKET REVENUES, AND COMPUTATION TIME

Algorithm	System	ADN operation cost (CNY)	P2P market revenue (CNY)	Computation time (s)
MISOCP-based centralized algorithm	IEEE 33-node	5752	24556	20.90
	IEEE 69-node	26248	73784	143.00
	IEEE 136-node	45380	207560	501.00
Proposed SAC-DTC algorithm	IEEE 33-node	6072	24508	4.27
	IEEE 69-node	27108	73743	14.50
	IEEE 136-node	47937	207440	33.80

In the IEEE 33-, 69-, and 136-node systems, the ADN operation costs obtained by the proposed SAC-DTC algorithm are slightly higher than those by the MISOCP-based centralized algorithm, while the P2P market revenues are almost the same. Based on the characteristics of distributed computation, the proposed SAC-DTC algorithm can effectively protect the privacy information, and the computation speed is 4.9, 9.8, and 14.8 times faster than that of MISOCP-based centralized algorithm in IEEE 30-, 69-, 136-node systems, respectively.

In addition, the linearization of voltage mapping in the proposed SAC-DTC algorithm may introduce some errors in the final results. Therefore, we perform power flow calculations using the proposed SAC-DTC algorithm and MISOCP-based centralized algorithm, and compare the node voltages. As shown in Table VIII and Fig. 8, the maximum error in voltage magnitude is within 0.3% and the average error is not larger than 0.08% for the power flow calculation, which indicates that the proposed SAC-DTC algorithm has a higher computation accuracy.

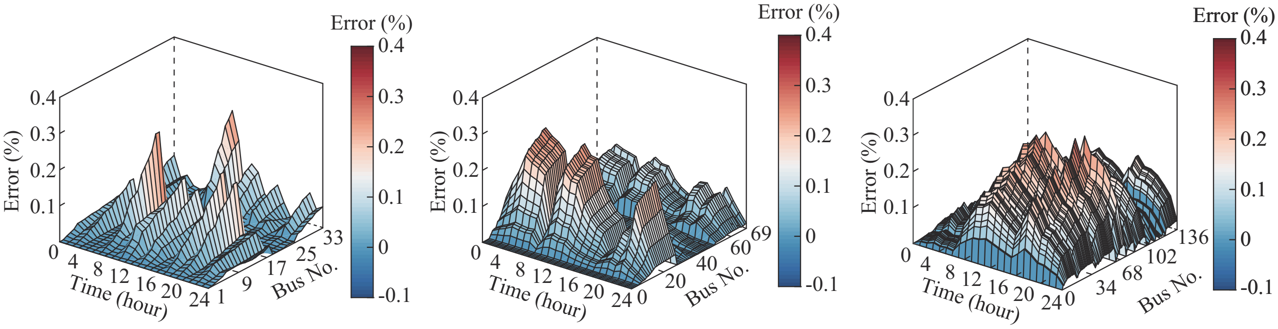
TABLE VIII
ANALYSIS OF ERROR IN VOLTAGE MAGNITUDE

System	Error in voltage magnitude (%)		
	Maximum	Minimum	Average
IEEE 33-node	0.249	1.10×10^{-4}	0.0281
IEEE 69-node	0.269	1.60×10^{-4}	0.0528
IEEE 136-node	0.258	1.51×10^{-3}	0.0735

Therefore, the proposed SAC-DTC algorithm is more suitable for the fast-changing operation of ADN and P2P markets to meet the real-time demand.

VI. CONCLUSION

In this paper, an SAC-DTC algorithm based on data-driven and physical modeling is proposed to tackle the coordinated optimization problem of ADN and P2P energy trading, which is analyzed via simulation based on the real-world dataset. The results show that the proposed SAC-DTC algorithm can effectively reduce the ADN operation cost and increase the P2P market revenue under the network security constraints. Specifically, the conclusions can be summarized as follows.



(a)

(b)

(c)

Fig. 8. Error in voltage magnitude. (a) IEEE 33-node system. (b) IEEE 69-node system. (c) IEEE 136-node system.

1) Compared with mainstream DDPG algorithms with the same network structure, the agents trained by the proposed SAC-DTC algorithm perform better in terms of the training speed and convergence results.

2) Considering the network security constraints, the proposed SAC-DTC algorithm for coordinated optimization can reduce the ADN operation cost by 8.3% and increase the P2P market revenue by 12.9% on average.

3) In the IEEE 33-, 69-, and 136-node systems, the proposed SAC-DTC algorithm effectively protects the privacy of prosumers although the ADN operation cost is slightly higher compared with the traditional MISOCP-based centralized algorithm. The computation speed is 4.9, 9.8, and 14.8 times faster, and the voltage magnitude error is no more than 0.08% on average.

Future work will investigate additional scenarios, including the integration of electrical, thermal, and cooling energy systems for consumers. Moreover, efforts will be made to deploy larger-scale networks utilizing multiple agents to manage complex coordination tasks involving both discrete and continuous actions. Additionally, there will be a focus on optimizing the linearization process to further enhance accuracy.

References

[1] K. H. M. Azmi, N. A. M. Radzi, N. A. Azhar *et al*., "Active electric distribution network: applications, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 134655-134689, Dec. 2022.

[2] Z. Yang, H. Li, and H. Zhang, "Dynamic collaborative pricing for managing refueling demand of hydrogen fuel cell vehicles," *IEEE Transactions on Transportation Electrification*, vol. PP, no. 99, pp. 1-1, Mar. 2024.

[3] S. Gorbachev, A. Mani, L. Li *et al*., "Distributed energy resources based two-layer delay-independent voltage coordinated control in active distribution network," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 1220-1230, Feb. 2024.

[4] Z. Deng, M. Liu, H. Chen *et al*., "Optimal scheduling of active distribution networks with limited switching operations using mixed-integer dynamic optimization," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4221-4234, Jul. 2019.

[5] H. Zhu and H. Liu, "Fast local voltage control under limited reactive power: optimality and stability analysis," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3794-3803, Sept. 2016.

[6] H. Liu and W. Wu, "Online multi-agent reinforcement learning for decentralized inverter-based volt-var control," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2980-2990, Jul. 2021.

[7] Q. Yang, G. Wang, A. Sadeghi *et al*., "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313-2323, May 2020.

[8] W. Shi, D. Zhang, X. Han *et al*., "Coordinated operation of active distribution network, networked microgrids, and electric vehicle: a multi-agent PPO optimization method," *CSEE Journal of Power and Energy Systems*, doi: 10.17775/CSEEJPES.2022.05640

[9] M. Mansourlakouraj, M. Gautam, H. Livani *et al*., "Multi-stage volt/var support in distribution grids: risk-aware scheduling with real-time reinforcement learning control," *IEEE Access*, vol. 11, pp. 54822-54838, May 2023.

[10] A. R. Sayed, C. Wang, H. I. Anis *et al*., "Feasibility constrained online calculation for real-time optimal power flow: a convex constrained deep reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 38, no. 6, pp. 5215-5227, Nov. 2023.

[11] D. Cao, W. Hu, X. Xu *et al*., "Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1101-1110, Sept. 2021.

[12] H. Liu, W. Wu, and Y. Wang, "Bi-level off-policy reinforcement learning for two-timescale volt/var control in active distribution networks," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 385-395, Jan. 2023.

[13] K. Schmitt, R. Bhatta, M. Chamana *et al*., "A review on active customers participation in smart grids," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 1, pp. 3-16, Jan. 2023.

[14] W. Tushar, T. K. Saha, C. Yuen *et al*., "Peer-to-peer trading in electricity networks: an overview," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3185-3200, Jul. 2020.

[15] W. Tushar, C. Yuen, T. K. Saha *et al*., "Peer-to-peer energy systems for connected communities: a review of recent advances and emerging challenges," *Applied Energy*, vol. 282, p. 116131, Jan. 2021.

[16] Y. Zou, Y. Xu, X. Feng *et al*., "Transactive energy systems in active distribution networks: a comprehensive review," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 5, pp. 1302-1317, Sept. 2022.

[17] D. Han, L. Wu, X. Ren *et al*., "Calculation model and allocation strategy of network usage charge for peer-to-peer and community-based energy transaction market," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 1, pp. 144-155, Jan. 2023.

[18] T. AlSkaif, J. L. Crespo-Vazquez, M. Sekuloski *et al*., "Blockchain-based fully peer-to-peer energy trading strategies for residential energy systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 1, pp. 231-241, Jan. 2022.

[19] F. Luo, Z. Y. Dong, G. Liang *et al*., "A distributed electricity trading system in active distribution networks based on multi-agent coalition and blockchain," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 4097-4108, Sept. 2019.

[20] X. Yang, G. Wang, H. He *et al*., "Automated demand response framework in ELNs: decentralized scheduling and smart contract," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 1, pp. 58-72, Jan. 2020.

[21] J. Zheng, Z. Liang, Y. Li *et al*., "Multi-agent reinforcement learning with privacy preservation for continuous double auction-based P2P energy trading," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 6582-6590, Apr. 2024.

[22] L. Chen, N. Liu, and J. Wang, "Peer-to-peer energy sharing in distribution networks with multiple sharing regions," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 6760-6771, Nov. 2020.

[23] L. Wang, Y. Zhang, W. Song *et al*., "Stochastic cooperative bidding strategy for multiple microgrids with peer-to-peer energy trading," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1447-1457, Mar. 2022.

[24] J. Li, C. Zhang, Z. Xu *et al*., "Distributed transactive energy trading

framework in distribution networks," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7215-7227, Nov. 2018.

[25] W. Tushar, B. Chai, C. Yuen *et al.*, "Energy storage sharing in smart grid: a modified auction-based approach," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1462-1475, May 2016.

[26] W. Lee, L. Xiang, R. Schober *et al.*, "Direct electricity trading in smart grid: a coalitional game analysis," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 7, pp. 1398-1411, Jul. 2014.

[27] N. Liu, X. Yu, C. Wang *et al.*, "Energy sharing management for microgrids with PV prosumers: a Stackelberg game approach," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1088-1098, Jun. 2017.

[28] Y. Liu, C. Sun, A. Paudel *et al.*, "Fully decentralized P2P energy trading in active distribution networks with voltage regulation," *IEEE Transactions on Smart Grid*, vol. 14, no. 2, pp. 1466-1481, Mar. 2023.

[29] Y. Jia, C. Wan, and B. Li, "Strategic peer-to-peer energy trading framework considering distribution network constraints," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 3, pp. 770-780, May 2023.

[30] Y. Zhou, B. Zhang, C. Xu *et al.*, "A data-driven method for fast AC optimal power flow solutions via deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1128-1139, Nov. 2020.

[31] D. Cao, J. Zhao, W. Hu *et al.*, "Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4137-4150, Sept. 2021.

[32] P. Giselsson, "Improved dual decomposition for distributed model predictive control," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 1203-1209, Oct. 2014.

[33] C. Feng, B. Liang, Z. Li *et al.*, "Peer-to-peer energy trading under network constraints based on generalized fast dual ascent," *IEEE Transactions on Smart Grid*, vol. 14, no. 2, pp. 1441-1453, Mar. 2023.

[34] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183-202, Jan. 2009.

[35] Y. Zhang and Z. Ren, "Optimal reactive power dispatch considering costs of adjusting the control devices," *IEEE Transactions on Power Systems*, vol. 20, no. 3, pp. 1349-1356, Aug. 2005.

[36] Z. Li, L. Wu, and Y. Xu, "Risk-averse coordinated operation of a multi-energy microgrid considering voltage/var control and thermal flow: an adaptive stochastic approach," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 3914-3927, Sept. 2021.

[37] X. Chang, Y. Xu, H. Sun *et al.*, "Privacy-preserving distributed energy transaction in active distribution networks," *IEEE Transactions on Power Systems*, vol. 38, no. 4, pp. 3413-3426, Jul. 2023.

[38] P. Sunehag, G. Lever, A. Gruslys *et al.* (2017, Jun.). Value-decomposition networks for cooperative multi-agent learning. [Online]. Available: https://arxiv.org/abs/1706.05296

[39] Z. Zhang, C. Dou, D. Yue *et al.*, "Regional coordinated voltage regulation in active distribution networks with PV-BESS," *IEEE Transactions on Circuits and Systems II*: *Express Briefs*, vol. 70, no. 2, pp. 596-600, Feb. 2023.

[40] Y. Zhang, Y. Han, D. Liu *et al.*, "Low-carbon economic dispatch of electricity-heat-gas integrated energy systems based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 6, pp. 1827-1841, Nov. 2023.

[41] R. S. Sutton and A. G. Barto, (2024, Apr.). Reinforcement learning: an introduction. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=uWV0DwAAQBAJ&oi=fnd&pg=PR7&dq=info:t8N5xiW9bXoJ:scholar.google.com&ots=mjoHs_Z0k1&sig=CKvWTrZ0FoBPRC-mO4-Yoo4uv5z0

Yongjun Zhang received the Ph. D. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2004. Currently, he is a Professor with the School of Electric Power, South China University of Technology. His research interests include reactive power optimization, smart energy, and high-voltage direct current (HVDC) transmission.

Jun Zhang received the B. E. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2022. He is now a master's student at South China University of Technology, Guangzhou, China. His research interests include distribution network optimization, peer-to-peer energy trading, and shared energy storage system.

Guangbin Wu received the B. E. degree from South China University of Technology, Guangzhou, China, in 2008. He is currently working in the Customer Service Center of Guangdong Power Grid Corporation, Foshan, China. His research interests include specialty of power marketing customer service.

Jiehui Zheng received the B. E. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2012, and the Ph.D degree from the South China University of Technology, Guangzhou, China, in 2017, all in electrical engineering. He is currently an Associate Professor with the School of Electric Power Engineering, South China University of Technology. His research interests include optimization algorithm, decision-making method, and their application on integrated energy system.

Dongming Liu received the B. E. degree in electrical engineering from Shanghai University of Electric Power, Shanghai, China, in 2022. He is now pursuing the master degree at South China University of Technology, Guangzhou, China. His research interest includes distribution network optimization.

Yuzheng An received the B.S. degree in electrical engineering from Shanghai Electric Power University, Shanghai, China, in 2021. He received the M.S. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2024. His research interest includes deep reinforcement learning.