Real-time Operation Optimization in Active Distribution Networks Based on Multi-agent Deep Reinforcement Learning

Jie Xu, Hongjun Gao, Renjun Wang, and Junyong Liu

***Abstract*—The increasing integration of intermittent renewable energy sources (RESs) poses great challenges to active distribution networks (ADNs), such as frequent voltage fluctuations. This paper proposes a novel ADN strategy based on multi-agent deep reinforcement learning (MADRL), which harnesses the regulating function of switch state transitions for the real-time voltage regulation and loss minimization. After deploying the calculated optimal switch topologies, the distribution network operator will dynamically adjust the distributed energy resources (DERs) to enhance the operation performance of ADNs based on the policies trained by the MADRL algorithm. Owing to the model-free characteristics and the generalization of deep reinforcement learning, the proposed strategy can still achieve optimization objectives even when applied to similar but unseen environments. Additionally, integrating parameter sharing (PS) and prioritized experience replay (PER) mechanisms substantially improves the strategic performance and scalability. This framework has been tested on modified IEEE 33-bus, IEEE 118-bus, and three-phase unbalanced 123-bus systems. The results demonstrate the significant real-time regulation capabilities of the proposed strategy.**

***Index Terms*—Reconfiguration, active distribution network, distributed energy resource, real-time control, deep reinforcement learning, parameter sharing, scalability.**

I. Introduction

THE large-scale integration of intermittent distributed generation such as renewable energy sources (RESs) presents prospects for enhancing the energy decarbonization and flexibility of active distribution networks (ADNs). However, the uncertainty of RES output also challenges the operation of ADNs, including more frequent voltage violations and increased network loss [1]. Therefore, optimizing the ADN operation against the backdrop of high renewable energy penetration has emerged as a pressing issue that must be addressed.

Existing cutting-edge approaches in the field of ADN operation optimization mainly include model-based methods such as mathematical programming [2]-[4], and heuristic methods [5], [6] such as evolutionary methods. Nevertheless, the model-based methods rely heavily on precise global information to solve the optimal power flow (OPF) problem with poor computation time as the system complexity increases. The heuristic methods suffer from dimensionality curse, resulting in time-consuming calculations for ADN management. Considering these inherent drawbacks, it is critically important to propose an adaptive optimization strategy for the real-time control of ADNs.

With respect to the dynamic adaptive strategy for real-time control, the deep reinforcement learning (DRL) is a promising alternative algorithm with model-free characteristics [7]-[9]. DRL can make near-optimal decisions in a short time to address the dynamic transitions of an ADN by utilizing the knowledge extracted from historical data. Thus, the DRL-based methods can be applied to unseen scenarios without resolving the models [7], which is impossible using model-based methods. Reference [7] applies the proximal policy optimization (PPO) algorithm for the real-time control of an ADN [7], and the experimental results demonstrate the good generalization of DRL against unknown information as well as the rapid decision-making rate against environmental uncertainties.

Notably, the aforementioned literature predominantly utilizes centralized methods that require global system data for decision-making, resulting in possible single-point failures. For the centralized single-agent DRL, a high-dimensional action space may incur dimensionality curses [8]. Consequently, some scholars have explored the applicability of multi-agent DRL (MADRL) algorithms [8]-[14] in this domain. In [9], the multi-agent deep deterministic policy gradient (MADDPG) algorithm is leveraged to perform voltage/var control (VVC), alleviating the concerns of the aforementioned single-agent DRL. Moreover, it is worth noting that the operating timescales of different devices in ADNs are different [10]-[12]. Discrete electrical devices such as switches cannot operate randomly owing to risk concerns, making them more suitable for day-ahead-determined 1-hour interval deployment. In contrast, the flexible regulating function of

Manuscript received: April 6, 2023; revised: June 30, 2023; accepted: August 26, 2023. Date of CrossCheck: August 26, 2023. Date of online publication: October 5, 2023.

This work was supported by the National Natural Science Foundation of China (No. 52077146) and Sichuan Science and Technology Program (No. 2023NSFSC1945).

J. Xu, H. Gao, R. Wang (corresponding author), and J. Liu are with the College of Electrical Engineering, Sichuan University, Chengdu, China (e-mail: 3024040071@qq.com; gaohongjun@scu.edu.cn; 1328860755@qq.com; liujy@scu.edu.cn).

DOI: 10.35833/MPCE.2023.000213

continuous devices such as photovoltaics (PVs) must be fully exploited to handle real-time environmental fluctuations. Consequently, [10] and [11] propose the two-stage DRL-based VVC strategy with a day-ahead capacitor bank (CB), on-load tap-changer (OLTC) scheduling, and a real-time PV inverter control. Reference [12] integrates graph reinforcement learning into this two-stage VVC strategy to extract better topological information. However, [7]-[12] ignore the potential benefits of switch state transitions when designing the DRL-based ADN optimization strategies, highlighting the need for further research. Simultaneously, treating each distributed energy resource (DER) as an agent [9]-[11] poses scalability issues in ADNs that contain a large number of DERs. Reference [13] partitions the entire distribution network into multiple subnetworks and treats each of them as an agent to control internal DERs, thus significantly reducing the number of agents and promoting the scalability of the method. Reference [14] notes the unsatisfactory scalability of the existing DRL-based methods and proposes a scalable algorithm called distributed proximal policy optimization (DPPO) to address the issue of electric vehicle (EV) charging management. References [15]-[17] utilize MADRL to address the reactive power optimization, reconfiguration, and VVC issues in three-phase unbalanced systems. However, with the characteristics of multiphase and phasic coupling, the complexity of three-phase ADN optimization is aggravated. Thus, the scalability of the MADRL algorithm should be given more attention.

To address the limitations of prior studies, this paper proposes a novel MADRL-based real-time optimization strategy for ADN that fully harnesses switch state transitions while ensuring scalability. After adopting the optimal switch deployment calculated by the prior reconfiguration preliminary, the real-time management of DERs based on the MADRL-trained strategies can be executed to optimize the ADN operation. Unlike [18], which treats topology transitions as random variants, our framework leverages them as regulating mechanisms for mitigating loss and voltage violations, which is widely recognized as beneficial [19], [20], but rarely combined with DRL-based DER control. Furthermore, this paper integrates the parameter sharing (PS) [21] mechanism into the twin delayed deep deterministic policy gradient (TD3) algorithm [22] to overcome the non-stationarity of the multi-agent environment and promote scalability in larger systems [23]. The PS mechanism allows all agents to share the parameters of a single policy if the agents are homogenous or exhibit similar behaviors, enabling each agent to benefit from the episodic experiences and learned knowledge of other agents [24], [25].

The major contributions of this paper are summarized as follows.

1) A novel MADRL-based DER control method that integrates a preliminary model-based switch reconfiguration is proposed to optimize the ADN in real time. To the best of our knowledge, the existing DRL-based ADN control strategies are primarily combined with the day-ahead scheduling of CB and OLTC [10], with the minimal exploration of the potential benefits that could be derived from the switch reconfiguration.

2) The PS mechanism is integrated into the TD3 algorithm to solve the formulated problem. By sharing identical network parameters and samples gathered by all agents, this mechanism considerably enhances the algorithmic scalability in larger systems [22], [24]. In addition, the regional agent partitioning improves scalability at the model level by reducing the number of agents [13].

3) The proposed MADRL-based optimization strategy exhibits superior real-time decision-making capability and generalization performance against various unseen scenarios, which has been verified in several test systems.

The remainder of this paper is organized as follows. Section II presents the problem formulation of ADN optimization model. Section III formulates the proposed model within the decentralized partially observable Markov decision process (Dec-POMDP) framework. Section IV discusses the proposed parameter sharing−prioritized experience replay−independent twin delayed deep deterministic policy gradient (PS-PER-ITD3) algorithm. The simulation results and conclusions are presented in Sections V and VI, respectively.

II. Problem Formulation of ADN Optimization Model

A. Prior Reconfiguration Preliminary

To fully utilize the regulating function of switch state transitions in ADN operation optimization, we first calculate the optimal 24-hour switch states by solving a mixed-integer second-order-cone programming (MISOCP) reconfiguration problem [19] with the stipulation of a 1-hour switch scheduling interval.

$$\min_{\alpha_{ij,t}} \sum_{ij\in\Omega_{branch}} \sum_{t\in T} \alpha_{ij,t} l_{ij,t} r_{ij} \tag{1}$$

$$\begin{cases} \sum_{ij\in\Omega_{branch}} \alpha_{ij,t} = N_{bus} - N_{root,bus} \\ N_{root,bus} = 1 \end{cases} \tag{2}$$

$$\begin{cases} \beta_{ij,t} + \beta_{ji,t} = \alpha_{ij,t} & \forall ij \in \Omega_{branch}, t \in T \\ \sum_{j\in i} \beta_{ij,t} = 1 & \forall i \in \Omega_{bus} \backslash \Omega_{root,bus} \\ \beta_{ij,t} = 0 & \forall i \in \Omega_{root,bus}, j \in i \end{cases} \tag{3}$$

$$\begin{cases} P_{j,t}^{RES,MPPT} - P_{j,t}^{load} = \sum_{k\in j} P_{jk,t} - \sum_{j\in i} (P_{ij,t} - r_{ij} l_{ij,t}) \\ Q_{j,t}^{RES,MPPT} - Q_{j,t}^{load} = \sum_{k\in j} Q_{jk,t} - \sum_{j\in i} (Q_{ij,t} - x_{ij} l_{ij,t}) \end{cases} \tag{4}$$

$$U_{j,t}^{sqr} \le M(1-\alpha_{ij,t}) + U_{i,t}^{sqr} - 2(r_{ij}P_{ij,t} + x_{ij}Q_{ij,t}) + (r_{ij}^2 + x_{ij}^2) l_{ij,t} \quad \forall i \in \Omega_{bus}, ij \in \Omega_{branch}, t \in T \tag{5}$$

$$U_{j,t}^{sqr} \ge -M(1-\alpha_{ij,t}) + U_{i,t}^{sqr} - 2(r_{ij}P_{ij,t} + x_{ij}Q_{ij,t}) + (r_{ij}^2 + x_{ij}^2) l_{ij,t} \quad \forall i \in \Omega_{bus}, ij \in \Omega_{branch}, t \in T \tag{6}$$

$$\begin{cases} U_{i,t}^{sqr} = v_{i,t}^2 \\ l_{ij,t} = I_{ij,t}^2 \\ U_{\min}^{sqr} \le U_{i,t}^{sqr} \le U_{\max}^{sqr} \end{cases} \tag{7}$$

$$\left\| \begin{matrix} 2P_{ij,t} \\ 2Q_{ij,t} \\ l_{ij,t}-U_{j,t}^{sqr} \end{matrix} \right\|_2 \le U_{i,t}^{sqr}+l_{ij,t} \quad \forall ij \in \Omega_{branch} \tag{8}$$

where T is the operation period segment; Ω_{bus} and Ω_{branch} are the bus set and branch set of ADN, respectively; $\Omega_{root,bus}$ is the set of buses linked to substation; $N_{root,bus}$ is the number of buses linked to substations; N_{bus} is the number of buses of ADN; $\alpha_{ij,t}$ is the switch status of branch ij at time t (0 represents off and 1 represents on); $\beta_{ij,t}$ is the auxiliary variable to ensure connectivity; $j \in i$ indicates that j is the downstream bus of i; $P_{j,t}^{RES,MPPT}$ and $Q_{j,t}^{RES,MPPT}$ are the maximum active and reactive power that the RES device installed on bus j can output under external weather conditions at time t, respectively; $P_{j,t}^{load}$ is the active load demand of bus j; $P_{ij,t}$ is the active power flow on branch ij at time t; $Q_{ij,t}$ is the reactive power flow on branch ij at time t; r_{ij} and x_{ij} are the resistance and impedance of branch ij, respectively; $v_{ij,t}$ and $I_{ij,t}$ are the amplitudes of voltage and current phasors, respectively; the subscripts min and max represent the minimum and maximum values, respectively; and M is a huge relaxation coefficient.

Formulas (2) and (3) restrict the topology of ADN at any time to be radial and connective, where (2) ensures the connectivity and (3) prevents the appearance of "island" [19]; and (4) represents the network power flow constraints. Formulas (5) and (6) are the adjusted voltage droop constraints to which the big-M relaxation is added to address the reconfiguration issue. The meanings of $U_{j,t}^{sqr}$ and $l_{ij,t}$ are expressed in (7). Formula (8) is a second-order cone relaxation constraint. The optimal switch deployment results can be easily obtained by solving (1)-(8).

Subsequently, whether it is during offline training or online execution, the MADRL-based DER control of ADN will be used under the topology with the optimal 24-hour switch state, which has often been neglected in previous studies [7]-[17] but is considered in our study.

B. Objective Function

The objective function of the real-time optimization model for ADN operation is composed of the cost of network loss and the voltage violation penalty, as given by:

$$F = \min_{P_{i,t}^{DER}} \sum_{t \in T} (\kappa_1 c_t^{loss} P_t^{loss} + \kappa_2 c_t^{pen} V_t^{vio}) \tag{9}$$

$$V_t^{vio} = \sum_{j \in \Omega_{bus}} (ReLU(v_{j,t}-v_{\max}) + ReLU(v_{\min}-v_{j,t})) \tag{10}$$

$$P_t^{loss} = \sum_{i \in \Omega_{bus}} \sum_{j \in \Omega_{bus}} e_{i,t} e_{j,t} (G_{ij,t} \cos(f_{i,t}-f_{j,t}) + B_{ij,t} \sin(f_{i,t}-f_{j,t})) \tag{11}$$

where F is the objective of ADN optimization in the entire period; $P_{i,t}^{DER}$ is the active power output of DER installed on bus i at time t that can be adjusted; c_t^{loss} is the unit cost of active power loss P_t^{loss}; c_t^{pen} is the penalty factor of voltage violation V_t^{vio}; κ_1 and κ_2 are the target coefficients of loss cost and voltage penalty, respectively; $ReLU(x)=\max(0,x)$ is the function for depicting nodal voltage violation [26]; $v_{j,t}$ is the amplitude of the voltage phasor $\boldsymbol{v}_{j,t}$; $[v_{\min}, v_{\max}]$ is the acceptable range of nodal voltage; $e_{i,t}$ is the real component of the voltage phasor $\boldsymbol{v}_{i,t}$ on bus i at time t; $f_{i,t}$ is the complex component of the voltage phasor $\boldsymbol{v}_{i,t}$ on bus i at time t; and $G_{ij,t}$ and $B_{ij,t}$ are the real and complex components of admittance for branch ij at time t, respectively. Note that the optimization of ADN is under the topology with the optimal 24-hour switch states calculated by the preliminary reconfiguration. Therefore, the admittance matrix of branch changes dynamically according to the varying ADN structures.

C. Constraints

1) Power Flow Constraints

$$P_{i,t}^{DER} - P_{i,t}^{load} = e_{i,t} \sum_{j \in i} (G_{ij,t} e_{j,t} - B_{ij,t} f_{j,t}) - f_{i,t} \sum_{j \in i} (G_{ij,t} f_{j,t} + B_{ij,t} e_{j,t}) \tag{12}$$

$$Q_{i,t}^{DER} - Q_{i,t}^{load} = f_{i,t} \sum_{j \in i} (G_{ij,t} e_{j,t} - B_{ij,t} f_{j,t}) - e_{i,t} \sum_{j \in i} (G_{ij,t} f_{j,t} + B_{ij,t} e_{j,t}) \tag{13}$$

where $Q_{i,t}^{DER}$ is the reactive power output of the DER installed on bus i at time t. DERs consist of RESs and energy storage systems (ESSs).

2) Security Operation Constraints

Ensuring the secure operation of the ADN necessitates maintaining nodal voltages within predetermined ranges and preventing the RES device from surpassing its maximum power output. As the inverter-based RESs have not yet been widely adopted, we employ the traditional method of curtailing the active power output of RES to optimize the ADN operation.

$$v_{\min} \le v_{i,t} \le v_{\max} \quad \forall i \in \Omega_{bus}, t \in T \tag{14}$$

$$0 \le P_{i,t}^{RES} \le P_{i,t}^{RES,MPPT} \quad \forall i \in \Omega_{bus}, t \in T \tag{15}$$

$$0 \le Q_{i,t}^{RES} \le Q_{i,t}^{RES,MPPT} \quad \forall i \in \Omega_{bus}, t \in T \tag{16}$$

where $P_{i,t}^{RES}$ and $Q_{i,t}^{RES}$ are the active and reactive power outputs of the RES installed on bus i at time t, respectively.

Formula (14) represents the nodal voltage constraint and it is achieved by adding penalty terms in this paper.

3) ESS Operation Constraints

ESSs are introduced into our model to further enhance the regulatory effect in ADN optimization. The mathematical model of ESS can be expressed as:

$$SOC_{i,t} = SOC_{i,t-1} + \eta_{char} \Delta t \cdot \max(P_{i,t}^{ESS}, 0) + \Delta t \cdot \min(P_{i,t}^{ESS}, 0)/\eta_{disc} \tag{17}$$

$$SOC_i^{\min} \le SOC_{i,t} \le SOC_i^{\max} \tag{18}$$

where $SOC_{i,t}$ is the state of charge (SOC) of ESS installed on bus i at time t; η_{disc} and η_{char} are the discharging and charging efficiencies, respectively; $P_{i,t}^{ESS}$ is the output power of ESS installed on bus i at time t; and $SOC_t^{\min}$ and $SOC_t^{\max}$ are the minimum and maximum SOCs of ESS, respectively.

III. Dec-POMDP Modeling for Optimization Strategy

In this section, we formulate the proposed optimization strategy within the MADRL framework. First, the basic concepts of the Markov decision process (MDP) and Dec-POMDP are briefly explained to facilitate the modeling procedure. Second, the optimization problem is constructed as a

Dec-POMDP model that focuses on determining the state variables, action variables, and other essential factors.

A. MDP and Dec-POMDP

1) MDP

First, the concept of an MDP in single-agent DRL is introduced. An MDP can be modeled as a tuple $M=(S,A,p,r,\gamma)$ consisting of state space S, action space A, state transition probability $p(s'|s,a)$:$\forall s',s\in S,\forall a\in A$, reward function $r(s,a)$: $S\times A\to R,\forall s\in S,\forall a\in A$, and the discount factor γ. In the MDP, an agent will make action $a_t\in A$ at each time step $t\in T$ based on the environmental observation $s_t\in S$; then, it will obtain a reward $r_t=r(s_t,a_t)$. Meanwhile, the state is transmitted to the next new state s_{t+1} according to the state transition probability $p(s'|s,a)$.

We define $\tau=\{s_0,a_0,s_1,a_1,\ldots,s_{T-1},a_{T-1},s_T\}$ as the trajectory of MDP and $\pi(\cdot|s)$ as the mapping of the action probability distribution for each state. The objective of the agent is to find a control policy π that can maximize the cumulative reward $J(\pi)=E_{\tau\sim\pi}\left(\sum_{t\geq 0}\gamma^t r_{t+1}\right)$.

2) Dec-POMDP

Dec-POMDP is a variant of MDP under a multi-agent full cooperation mode, which indicates that each agent shares an identical target and reward. It can be described by a tuple (K,S,O,A,R,T,γ), including K agents, global state variable $s\in S$, the local observation of agent k ($o_k\in O_{1:K}$), the action of agent k ($a_k\in A_{1:K}$), reward $r_1=r_2=\ldots=r_K=r$, state transition function $T(s,o_{1:K},a_{1:K})$, and discount factor γ. The interaction process of POMDP is similar to that of the MDP; thus, it is not repeated here.

B. Dec-POMDP Modeling

1) Brief Introduction

A schematic of the complete framework of the proposed optimization strategy is shown in Fig. 1. This strategy comprises two processes: offline centralized training and online decentralized execution. The offline centralized training process is described as follows. After adopting the 24-hour optimal ADN topologies calculated by the preliminary reconfiguration, each divided regional agent obtains the optimal control strategy through continuous interactions with the virtual distribution network environment.

The online decentralized process is described as follows. After receiving the trained policies, each regional agent can achieve online local DER control without the need for information exchange, which has been learned in the offline centralized training process.

2) Setup of Dec-POMDP Model

This part describes the construction of the DER control problem as a Dec-POMDP model that contains several fundamental elements.

1) Observation space: the observation of agent k is expressed as:

$$o_{k,t}=(SOC_{j,t},v_{j,t},P_{j,t}^{load},Q_{j,t}^{load})\quad\forall j\in\Omega_k \tag{19}$$

where $j\in\Omega_k$ represents the buses that belong to region k.

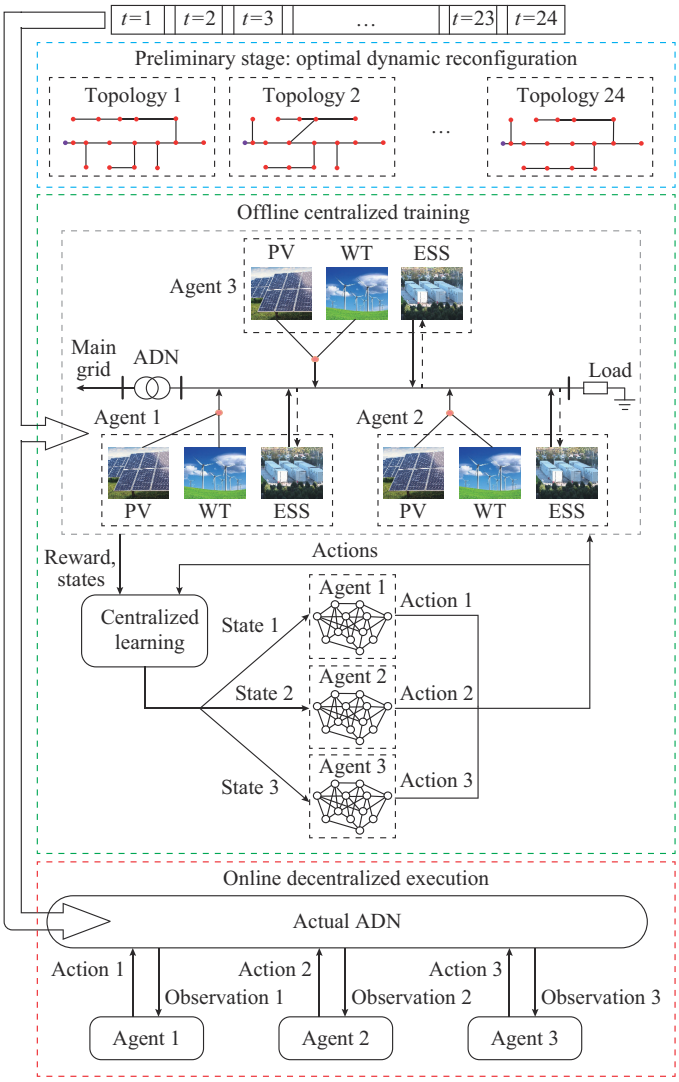

Fig. 1. Complete framework of proposed optimization strategy.

2) Action space: $a_{k,t}\in A_{k,t}$ is the action of the regional agent k at time t, which refers to the active power output of DER that regional agent k can control:

$$a_{k,t}=(P_{j,t}^{PV},P_{j,t}^{WT},P_{j,t}^{ESS})\quad\forall j\in\Omega_k \tag{20}$$

where $P_{j,t}^{PV}$ is the active power output of the PV installed on bus j; and $P_{j,t}^{WT}$ is the active power output of the wind turbine (WT) installed on bus j.

3) Constraints: the action space and observation space must satisfy the RES output and SOC constraints:

$$\begin{cases}P_{j,t}^{PV}\in[0,P_{j,t}^{PV,MPPT}]\\ Q_{j,t}^{PV}=\lambda_{j,t}^{PV}P_{j,t}^{PV}\end{cases} \tag{21}$$

$$\begin{cases}P_{j,t}^{WT}\in[(1-\chi)P_{j,t}^{WT,MPPT},P_{j,t}^{WT,MPPT}]\\ Q_{j,t}^{WT}=\lambda_{j,t}^{WT}P_{j,t}^{WT}\quad\forall j\in\Omega_k\end{cases} \tag{22}$$

where $\lambda_{j,t}^{PV}$ and $\lambda_{j,t}^{WT}$ are the power factors of the PV and WT installed on bus j, respectively.

Formulas (21) and (22) represent the boundaries of the active and reactive power outputs of the RESs, respectively. Here, the active power output of WT is restricted to partial

curtailable only, with a maximum curtailment range of χ times the predicted output.

$$P_{j,t}^{ESS,pre} \in [-P_{j,t}^{ESS,\max}, P_{j,t}^{ESS,\max}] \quad \forall j \in \Omega_k \tag{23}$$

$$P_{j,t}^{ESS} = \begin{cases} \min(P_{j,t}^{ESS,pre}, (SOC_j^{\max} - SOC_{j,t-1})) & P_{j,t}^{ESS,pre} > 0 \\ \max(P_{j,t}^{ESS,pre}, (SOC_j^{\min} - SOC_{j,t-1})) & P_{j,t}^{ESS,pre} < 0 \end{cases} \tag{24}$$

$$P_{j,t}^{DER} = P_{j,t}^{PV} + P_{j,t}^{WT} + P_{j,t}^{ESS} \tag{25}$$

where $P_{j,t}^{ESS,pre}$ is the predicted active power output of ESS that has not been clipped within safe range.

Formulas (23)-(25) represent the constraints of ESS that can prevent its SOC from violating feasible regions regardless of the offline centralized training or online decentralized execution stages of the DRL-based optimization task [24], which is unavoidable if a penalty function is used to restrict the charging/discharging behavior [27], [28].

4) Reward function: as this paper constructs the optimization problem as a Dec-POMDP model under a full cooperative framework, each agent shares an identical reward as:

$$r_t = \kappa_1 c_t^{loss} P_t^{loss} + \kappa_2 c_t^{pen} V_t^{vio} \tag{26}$$

The reward function in (26) includes the penalty of the voltage violation and the cost of the network loss, as expressed in (9)-(11).

IV. Proposed PS-PER-ITD3 Algorithm

Figure 2 shows the centralized training framework of the proposed PS-PER-ITD3. At the offline training stage, the historical operation data of the regional agents are collected by the MADRL aggregator for training, and the centralized aggregator is discarded during online execution. Thus, real-time distributed decision-making can be implemented by sending the local observation of each agent to the well-trained policies. The details of the algorithm are as follows.

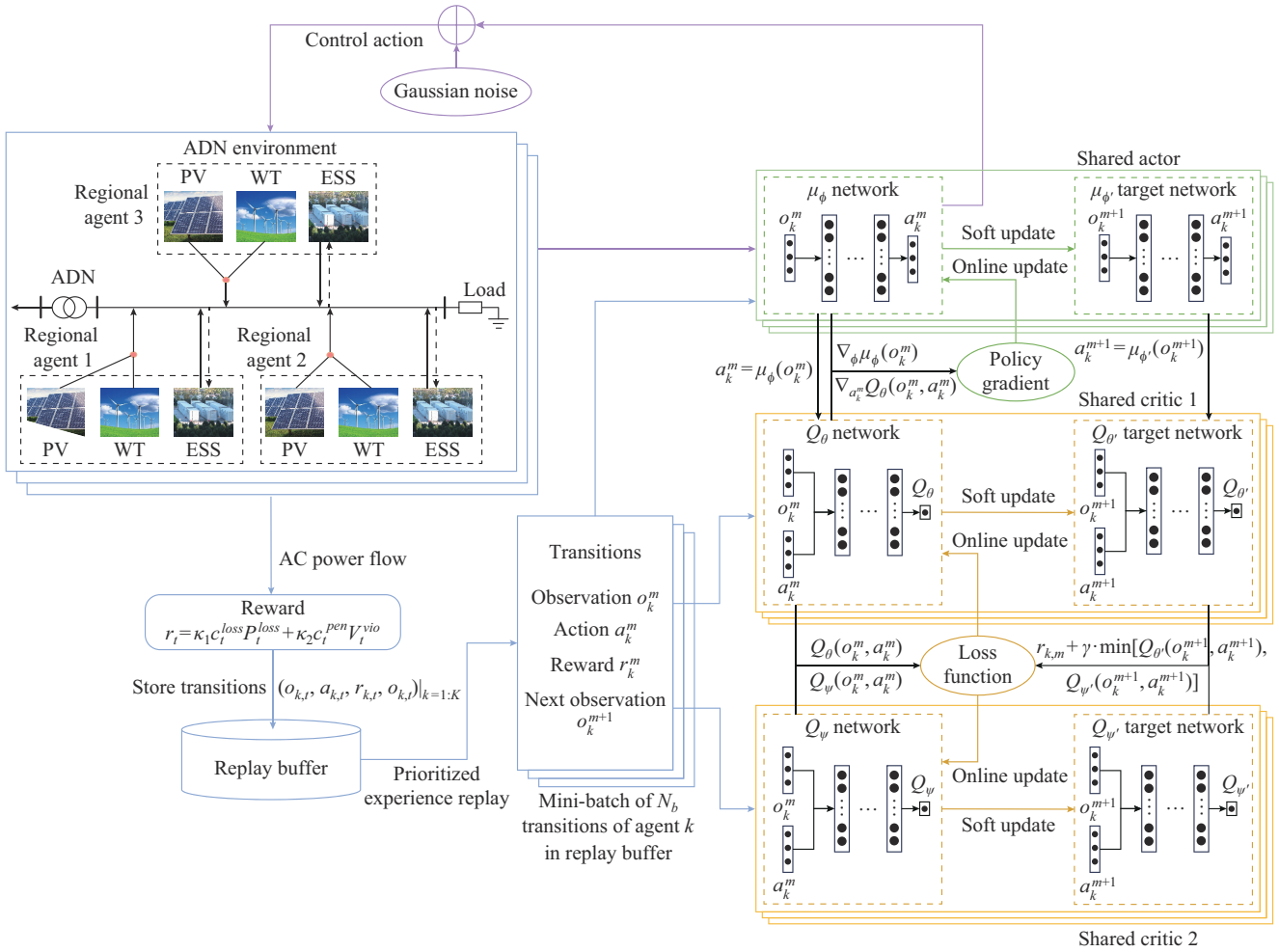

Fig. 2. Centralized training framework of proposed PS-PER-ITD3.

A. ITD3

Compared with DDPG, TD3 has been widely recognized for its effectiveness in alleviating the "bootstrapping" phenomenon, owing to the double-Q clipped learning technique [22] that it utilizes. In this study, we combine TD3 with an independent Q-learning [29] structure to obtain ITD3, which can be extended to multi-agent framework.

Under the ITD3 framework, each agent behaves as an independent TD3 learner that is unable to capture transitions gathered by other agents, leading to a nonstationary Markovian environment [23], [29]. To solve this problem, [30] proposes that each agent could utilize one global critic

$Q(o_{1,t}, o_{2,t}, \ldots, o_{K,t}, a_{1,t}, a_{2,t}, \ldots, a_{K,t})$ that can receive the joint action-state space $(o_{k,t}, a_{k,t})$ of all agents. Under this global critic framework, agents can learn experiences from each other, thus overcoming the nonstationary environment caused by independent training.

However, as global critics must receive global information, the algorithms such as MADDPG [30] or MATD3 [31] that rely on this face the challenge of a large agent size (network parameters), which leads to poor scalability with the rise in system scale [32]. To maintain a stationary environment and enhance the scalability of the proposed strategy, a PS mechanism is introduced.

B. PS

PS [21] is a mechanism that allows homogeneous agents to share identical network parameters. In this paper, all homogenous regional agents share the parameters of two critic networks $Q(o_{k,t}, a_{k,t})$ and one actor network $\mu(o_{k,t})$, which still allows agents to take different actions based on their different local observations at test time. The PS mechanism significantly curtails the network parameters that need to be updated during training, leading to the improvement in scalability [24].

Furthermore, the shared networks can be updated based on the experiences collected by all agents [25]. This updating mode based on experience sharing among multiple agents enables the learned behavior of one agent to be influenced by the experiences of other agents. Consequently, the integrated PS mechanism can enhance the algorithmic scalability of ITD3 without breaking the stationarity of the RL environment, owing to its sharing structure [21], [24], [25]. In addition, integrating this algorithm with the ADN partitioning approach can lead to improvements in scalability at both the algorithmic and model levels as the number of agents decreases significantly [13].

A comparison between the PS-integrated MADRL algorithm and the conventional global critic based algorithm (eg., MADDPG) is presented in Fig. 3, where π_k represents the policy of regional agent k.

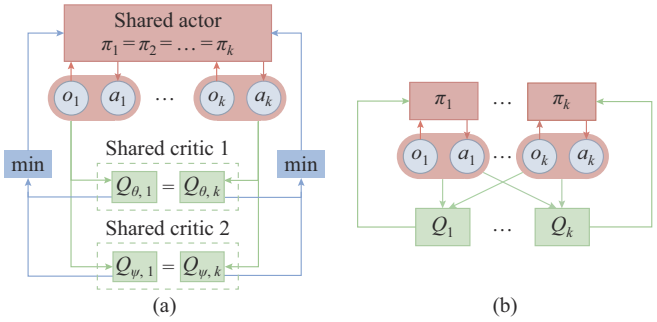

Fig. 3. Comparison between PS-integrated MADRL algorithm and conventional global critic based algorithm. (a) PS-integrated MADRL algorithm. (b) Global critic based algorithm.

C. PER

The TD3 algorithm is an improvement of DDPG [22], and PS indicates that all agents share the policy and replay buffer [24]. Thus, the PS-integrated ITD3 (PS-ITD3) remains an off-policy algorithm. Consequently, any behavioral policy can be utilized to collect experiences for model training. As conventional experience replay fails to distinguish the importance of various transitions, a PER [33] mechanism is added to the current algorithm and PS-PER-ITD3 is formulated. The calculation process of PER is presented as follows.

1) Setting the absolute value of the temporal difference (TD) error $|\delta_m|$ of the m^{th} experience transition $(o_k^m, a_k^m, r_k^m, o_k^{m+1})$ as its priority $p(m)$, where m represents the index of transition.

$$p(m) = |\delta^m| = |Q_\theta(o_k^m, a_k^m) - y| = Q_\theta(o_k^m, a_k^m) - |(r_k^m + \gamma \cdot \min(Q_{\theta'}(o_k^{m+1}, a_k^{m+1}), Q_{\psi'}(o_k^{m+1}, a_k^{m+1}))| \quad (27)$$

where $Q_{\theta'}$ and $Q_{\psi'}$ are the target networks of twin critic networks Q_θ and Q_ψ, respectively.

2) Computing the sampling probability P for each experience in the replay buffer based on their priority:

$$P(m) = p_\alpha(m) \Big/ \sum_{m=1}^{N_M} p_\alpha(m) \quad (28)$$

where N_M is the number of transitions stored in the replay buffer; and α is the priority exponent that needs to be adjusted.

3) Sampling a minibatch of N_b transitions stored in the replay buffer according to their computed probability P.

4) Computing the importance sampling weights ω for the sampled transitions as:

$$\omega(m) = (N_b P(m))^{-\beta} \quad (29)$$

where β is the importance sampling exponent that needs to be adjusted. The importance sampling weight ω and TD error will be used to calculate the critic loss.

The PER mechanism deviates from uniform sampling by assigning higher sampling weights to transitions with higher learning values, as indicated by the absolute TD error. This error is positively correlated with the extent to which the critic has been inadequately trained for the corresponding experience. Therefore, sampling these experiences during the updating process can effectively enhance the model performance.

D. Training Procedure

The complete training framework of the proposed PS-PER-ITD3 is illustrated in Fig. 2. Each agent shares an actor network μ_ϕ, twin critic networks Q_θ and Q_ψ, their target networks $Q_{\theta'}$ and $Q_{\psi'}$, and the replay buffer.

The regional agent k firstly receives the local observation $o_{k,t}$ at time t and generates an action $a_{k,t}$ according to the local observation and the shared policy network μ_ϕ. Subsequently, the output action $a_{k,t}$ will be added with a random Gaussian noise $N(\sigma)$ to promote exploration. The aggregated action set $\{a_{1,t}, a_{2,t}, \ldots, a_{K,t}\}$ is then implemented on the virtual ADN, whose topology X_t at time t is calculated using the optimal reconfiguration preliminary (1) - (8). Subsequently, each agent receives the same reward $r_t = r_{k,t}$ and its next observation $o_{k,t+1}$ by solving the power flow issues. Finally, the transitions $\{o_{k,t}, a_{k,t}, r_{k,t}, o_{k,t+1}\}_{k=1:K}$ of all regional agents at time t are formulated and sent to the shared experience re-

play buffer. During each episode of the centralized offline training, the MADRL aggregator will choose each agent k and sample N_b transitions $\{(o_k^m, a_k^m, r_k^m, o_k^{m+1})\}_{m \in \Omega_b}$ from the shared buffer to update the model parameters, where Ω_b is the set of minibatch samples. A more concise training framework of the proposed optimization strategy is shown in Fig. 4.

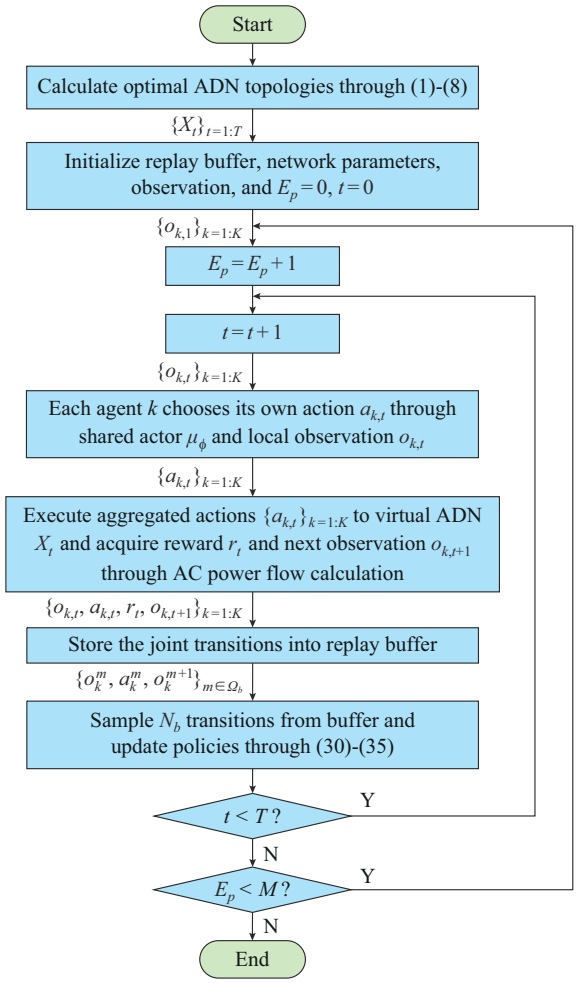

Fig. 4. Training framework of proposed PS-PER-ITD3.

After sampling transitions from the buffer, the mean-squared TD error δ_1 of the sampled transitions is then calculated as the loss of critic Q_θ:

$$\delta_1 = L(\theta) = N_b^{-1} \sum_{m \in \Omega_b} \omega(m)(Q_\theta(o_k^m, a_k^m) - y_k^m)^2/2 \tag{30}$$

The target value y_k^m is calculated using the predicted next action a_k^{m+1} that is approximated by the target actor network $\mu_{\phi'}(o_k^{m+1})$:

$$y_k^m = r_k^m + \gamma \cdot \min(Q_{\theta'}(o_k^{m+1}, \mu_{\phi'}(o_k^{m+1})), Q_{\psi'}(o_k^{m+1}, \mu_{\phi'}(o_k^{m+1}))) \tag{31}$$

Similar operations are executed again with another critic Q_ψ:

$$\delta_2 = L(\psi) = N_b^{-1} \sum_{m \in \Omega_b} \omega(m)(Q_\psi(o_k^m, a_k^m) - y_k^m)^2/2 \tag{32}$$

Subsequently, the two loss functions are utilized to update the weights of corresponding shared twin critics through the gradient descent algorithm:

$$\begin{cases} \theta \leftarrow \theta - \alpha^\theta \nabla_\theta L(\theta) \\ \psi \leftarrow \psi - \alpha^\psi \nabla_\psi L(\psi) \end{cases} \tag{33}$$

where α^θ is the learning rate of Q_θ; and α^ψ is the learning rate of Q_ψ.

As for the actor network $\mu_\phi(\cdot)$, the policy gradient for updating can be expressed as:

$$\nabla_\phi J(\mu_\phi) = N_b^{-1} \sum_{N_b} \nabla_\phi \mu_\phi(o_k^m) \cdot \nabla_{a_{k,m}} Q_\theta(o_k^m, a_k^m) \tag{34}$$

Finally, the weights of the three shared target networks are updated softly according to their corresponding networks with the fixed frequency τ:

$$\begin{cases} \theta' \leftarrow \tau\theta + (1-\tau)\theta' \\ \psi' \leftarrow \tau\psi' + (1-\tau)\psi \end{cases} \tag{35}$$

V. Numerical Analysis

In this paper, the modified IEEE 33-bus system [34], IEEE 118-bus system [35], and three-phase unbalanced 123-bus system [36] are utilized to evaluate the performance of the proposed strategy.

A. Corresponding Setting

The tanh(·) function is used as the terminal activation function of the actor network to limit its output to [−1, 1], which can be linearly scaled back to the output power of the DER device. Both the MADRL algorithm and power flow calculations are run in MATLAB 2022b with an Intel Core i9 CPU and Nvidia RTX4090 GPU. The topology, DER installation, and subnetwork division results for each test system are presented in Appendix A. The operation optimization model formulated for the three-phase unbalanced 123-bus system is presented in Appendix B. Detailed parameters such as load ratio and branch impedance can be learned from MATPOWER [34].

B. Strategic Performance on Modified IEEE 33-bus System

First, the reconfiguration issue of the modified IEEE 33-bus system is solved using GUROBI to acquire 24-hour optimal switch deployment, which is presented in Appendix A. Subsequent MADRL training and execution occur under the calculated topologies with the optimal 24-hour switch states.

1) Training Comparison Among Different Algorithms

To explore the optimal operation control strategies, several DRL algorithms are applied, i. e., the single-agent DDPG (SADDPG), MADDPG, PS-ITD3, and PS-PER-ITD3. The training performances of these algorithms in the reconfigured IEEE 33-bus system are presented in Fig. 5.

1) All the DRL algorithms converge terminally, except for SADDPG. The optimization objective is to adjust the output of the nine DERs in the IEEE 33-bus system to promote security and economy. However, the dimensionality of the action space in SADDPG, which is nine, is too large to learn a stable control strategy, thus resulting in severe fluctuations. Similarly, other single-agent DRL approaches exhibit poor scalability in the scenarios with numerous DERs.

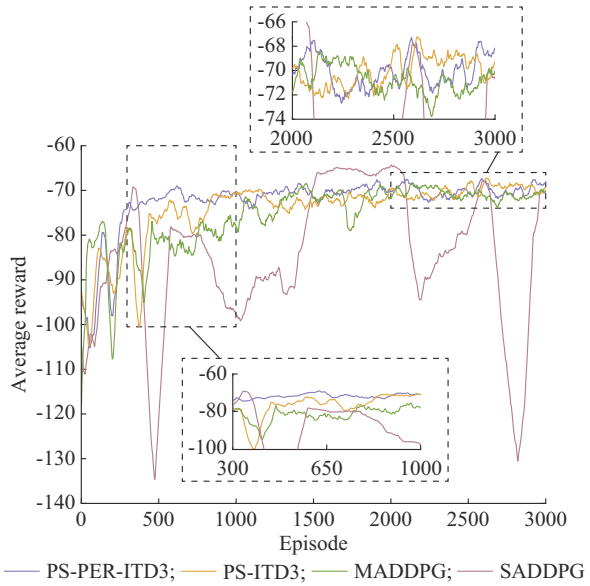

Fig. 5. Training performances of different algorithms in reconfigured IEEE 33-bus system.

2) Comparing the volatility of the three MADRL algorithms during the 300th to 1000th episodes, we can observe that PS-PER-ITD3 converges the fastest, reaching a stable reward of approximately −70 by the 300th episode; PS-ITD3 attains this value until the 800th episode, whereas MADDPG fails to achieve this reward within 1000 episodes. This illustrates the effectiveness of the PER mechanism in improving the sampling efficiency and accelerating convergence.

3) By analyzing the convergence performances of the three MADRL algorithms during the 2000th to 3000th episodes, it is found that all of them eventually reach convergence with an approximate reward of −70. However, the convergence time of MADDPG (the 1700th episode) is significantly longer than those of PS-ITD3 and PS-PER-ITD3, which verifies the superiority of the PS-based algorithms over the conventional MADDPG.

The variations in the average loss cost and voltage violation penalty of the PS-PER-ITD3 are shown in Fig. 6. This illustrates that agents successfully learn to mitigate losses and violations, with the average loss cost from \$145 to \$125 and average voltage violation penalty from 55 to 10.

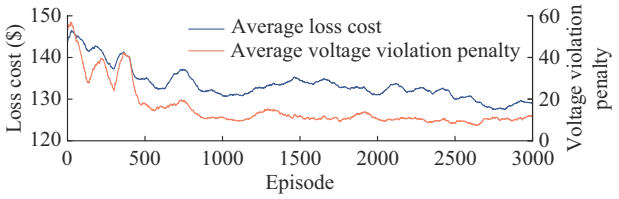

Fig. 6. Average loss cost and voltage violation penalty of PS-PER-ITD3 in reconfigured IEEE 33-bus system.

2) Evaluation of Decision-making Effects

As shown in Table I, the proposed PS-PER-ITD3 exhibits the best performance compared with the other four DRL-trained algorithms utilized for online decision-making. Its total loss cost on the test day outperforms other algorithms and nearly reaches the theoretical optimum calculated by MISOCP.

TABLE I
COMPARISONS OF DECISION-MAKING EFFECTS IN RECONFIGURED IEEE 33-BUS SYSTEM

Algorithm	Voltage violation rate (%)	Loss cost ($)
MISOCP	0	117.72
SADDPG	0	127.89
MADDPG	0	126.77
PS-ITD3	0	124.65
PS-PER-ITD3	0	121.84

However, none of the four DRL-based algorithms violate the voltage limits. This illustrates that the optimization task in the IEEE 33-bus system is not difficult; thus, a larger system is required to test the methodological scalability rigorously.

3) Scenario Comparison Analysis

Subsequently, a comparative experiment is conducted to verify the significance of the reconfiguration prior to the proposed optimization strategy. This involved two scenarios: one without reconfiguration (scenario 1) and the other with reconfiguration (scenario 2). The comparisons of average training rewards are presented in Fig. 7, and the comparisons of the real-time online decision-making effects are presented in Table II.

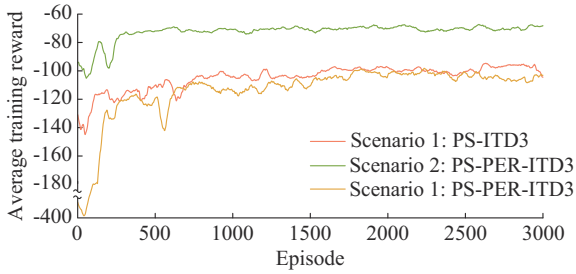

Fig. 7. Scenario comparisons of training in IEEE 33-bus system.

TABLE II
COMPARISONS OF DECISION-MAKING EFFECTS IN IEEE 33-BUS SYSTEM

Scenario	Algorithm	Voltage violation rate (%)	Loss cost ($)
1	MISOCP	0	140.03
	PS-ITD3	1.340	142.87
	PS-PER-ITD3	1.091	143.90
2	PS-PER-ITD3	0	121.84

1) The terminal rewards of PS-ITD3 and PS-PER-ITD3 in scenario 1 both converge at −100, which is 30 less than that of PS-PER-ITD3 in scenario 2.

2) The online execution results indicate that both the PS-ITD3 and the PS-PER-ITD3 in scenario 1 cannot strictly restrict the nodal voltage within a safe range, with approximately 1% of the buses violating the limits over 24 hours. Conversely, an identical PS-PER-ITD3 in scenario 2 rectifies this issue. Furthermore, the loss cost in scenario 1 is \$20 more than that in scenario 2. These observations demonstrate

the necessity of a preliminary reconfiguration in ADN optimization.

C. Strategic Performance on IEEE 118-bus System

Similar operations are implemented to calculate the optimal reconfiguration deployment for an IEEE 118-bus system. A 24-hour reconfiguration period is stipulated because of the concerns that more frequent switch operations could promote potential risks.

1) Training Comparison of Different Algorithms

The aforementioned MADRL algorithms are tested in a reconfigured IEEE 118-bus system to evaluate their scalability. The reward convergence curves of different algorithms in the reconfigured IEEE 118-bus system are presented in Fig. 8.

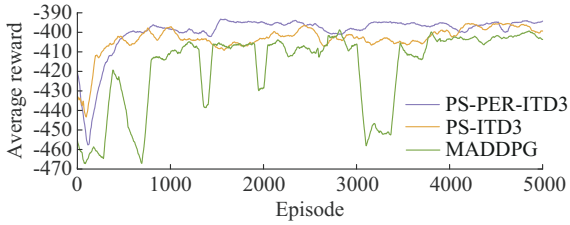

Fig. 8. Algorithmic training comparisons in reconfigured IEEE 118-bus system.

1) MADDPG in the reconfigured IEEE 118-bus system shows considerable volatility, with several collapses and surges during training, and converges until the 3700th episode. Conversely, the PS-ITD3 and PS-PER-ITD3 converge approximately during the 500th episode and remain stable during subsequent episodes. Additionally, all the algorithms show a larger fluctuation extent than the IEEE 33-bus system, which can be attributed to the increased complexity and scale of the optimization issue.

2) The terminal rewards of PS-PER-ITD3, PS-ITD3, and MADDPG are approximately −395, −400, and −405, respectively. Compared with the IEEE 33-bus system, the PS-PER-ITD3 exhibits significant superiority over MADDPG in terms of training reward and convergence rate in the reconfigured IEEE 118-bus system. This demonstrates that the performance gap between the PS-integrated algorithms and the conventional MADRL algorithms increases with the system scale.

2) Scenario Comparison Analysis

To enable better reward convergence, the target coefficients κ_1 and κ_2 in the original IEEE 118-bus system without reconfiguration are set distinct from the reconfigured IEEE 118-bus system. Owing to this stipulation, a comparison of the comprehensive rewards of two scenarios, as shown in Fig. 7, is meaningless. Therefore, the loss cost $c_t^{loss}P_t^{loss}$ and voltage violation penalty $c_t^{pen}V_t^{vio}$ of the two scenarios are separately compared during training. The results are presented in Fig. 9.

1) The loss cost of the reconfigured IEEE 118-bus system converges at approximately \$475, whereas that of the original IEEE 118-bus system converges at \$750.

2) The voltage violation penalty of the reconfigured IEEE 118-bus system converges at 32, whereas that of the original IEEE 118-bus system converges at 185.

3) In summary, the PS-PER-ITD3 in scenario 2 exhibits superior mitigation effects on both the loss cost and voltage violation, which preliminarily confirms the necessity and effectiveness of the integrated reconfiguration.

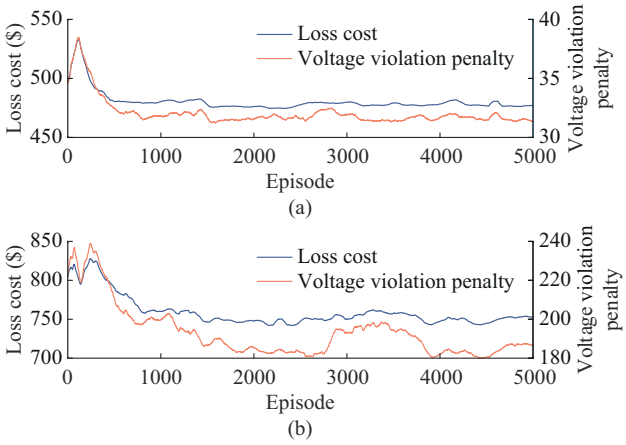

Fig. 9. Loss cost and voltage violation penalty of PS-PER-ITD3 in IEEE 118-bus system. (a) Reconfigured IEEE 118-bus system (scenario 2). (b) Original IEEE 118-bus system without reconfiguration (scenario 1).

3) Evaluation of Decision-making Effects

The online decision-making effects of different algorithms in the reconfigured and original IEEE 118-bus systems are compared in Table III.

TABLE III
COMPARISONS OF DECISION-MAKING EFFECTS IN RECONFIGURED AND ORIGINAL IEEE 118-BUS SYSTEMS

System	Algorithm	Voltage violation rate (%)	Loss cost (\$)
Original IEEE 118-bus	MISOCP	4.926	748.30
	PS-PER-ITD3	5.525	736.43
Reconfigured IEEE 118-bus	MISOCP	0.812	427.39
	MADDPG	0.983	494.87
	PS-ITD3	1.017	478.12
	PS-PER-ITD3	0.983	470.70

1) In contrast to the IEEE 33-bus system, the task difficulty of the IEEE 118-bus system increases significantly, as reflected by the higher voltage violation rates.

2) With the observation that both the voltage violation penalty and loss cost in scenario 1 are markedly higher than those in scenario 2, the necessity of the reconfiguration prior to the DRL-based ADN operation optimization is again verified in the larger IEEE 118-bus system.

3) In both scenarios 1 and 2, the voltage violation penalty and loss cost of the proposed PS-PER-ITD3 are close to the values calculated by MISOCP, demonstrating its superiority in approaching the theoretical optimum. Notably, the PS-PER-ITD3 abnormally outperforms MISOCP in scenario 1, which originates from the random noise added to the renewable predicted output during the MADRL training.

4) Comparisons of Decision-making Time

The decision-making time for different optimization algorithms in the reconfigured IEEE 118-bus system and reconfigured IEEE 33-bus system are compared in Table IV. In this context, all references to decision-making time refer to single time-step values.

TABLE IV
COMPARISONS OF DECISION-MAKING TIME IN RECONFIGURED IEEE 33-BUS SYSTEM AND RECONFIGURED IEEE 118-BUS SYSTEM

System	Algorithm	Decision-making time (s)
Reconfigured IEEE 33-bus	MISOCP	1.1070
	PS-ITD3	0.0118
	PS-PER-ITD3	0.0099
Reconfigured IEEE 118-bus	MISOCP	1.2320
	PS-ITD3	0.0137
	PS-PER-ITD3	0.0124

Table IV lists notable observations regarding the online decision-making rates of different algorithms for the two test systems. Specifically, the MADRL algorithms exhibit faster decision-making rates (millisecond level) by leveraging experiences extracted from training. This renders them well suited for addressing almost any short-term control requirement involving ADN fluctuation mitigation.

However, with the decision-making time of 1.107 s and 1.232 s for MISOCP in the reconfigured IEEE 33-bus system and reconfigured IEEE 118-bus systems, respectively, there is no overwhelming superiority in terms of speed for the MADRL algorithms regarding online real-time control. This is likely due to the simplicity of the test system. Therefore, further assessment of larger systems is necessary.

5) Algorithmic Generalization on Test Day Sets

Based on the stipulated 10% uncertainty in the renewable predicted output, 50 test days are generated to verify the generalization of the proposed PS-PER-ITD3 against unseen scenarios (unknown renewable predicted output). All the renewable predicted output data from the test days are excluded from the training process. The performance of the proposed PS-PER-ITD3 on the test day set is compared with that of MISOCP, as shown in Fig. 10, and the identical test objectives are adopted in the MADRL and MISOCP with $\kappa_1 = \kappa_2 = 1$.

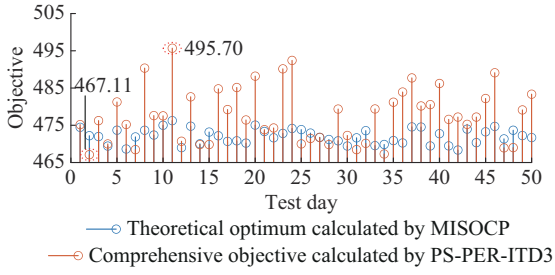

Fig. 10. Generalization validation on test day set in reconfigured IEEE 118-bus system.

1) The proposed PS-PER-ITD3 shows significant decision-making effects on the test day set because the gap from the optimum calculated by MISOCP is small, which confirms its superior generalization in similar but unseen scenarios.

2) Although the proposed PS-PER-ITD3 does not acquire the same effects as the MISOCP algorithm, it still exhibits nearly unique and overwhelming generalization and decision-making rates that are well suited for the real-time control of ADNs. However, the model-based algorithms require re-computation whenever the ADN scenario changes.

D. Strategic Performance on Three-phase 123-bus System

The proposed PS-PER-ITD3 is then tested on a three-phase unbalanced 123-bus system to evaluate its scalability and decision-making effects. With three phases per bus, the scale and complexity of the 123-bus system far exceed those like IEEE 118-bus system. Invalid load nodes and vacant branches are omitted so only 114 valid buses remain. The corresponding numerical results are presented in Table V and Fig. 11. In Fig, 11, lines with different colors correspond to different hours.

TABLE V
COMPARISONS OF DECISION-MAKING EFFECTS IN THREE-PHASE UNBALANCED 123-BUS SYSTEM

Algorithm	Decision-making time (s)	Loss (kWh)
MISOCP	5.240	1197
PS-PER-ITD3	0.049	1334

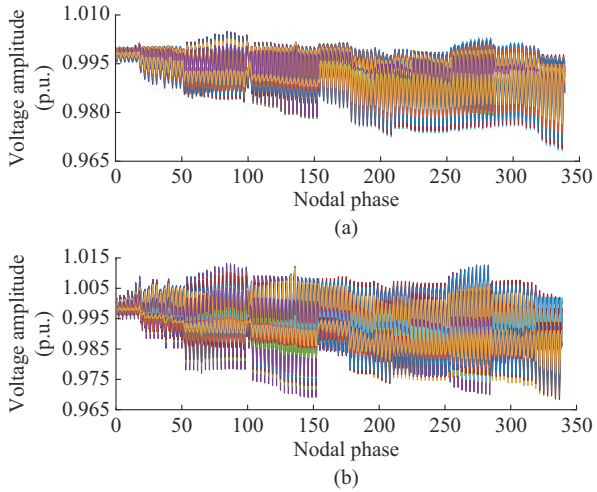

Fig. 11. 24-hour voltage distribution comparisons in three-phase unbalanced 123-bus system. (a) MISOCP. (b) PS-PER-ITD3.

1) Both the MISOCP [37] and the PS-PER-ITD3 eliminate voltage violations, as shown in Fig. 11. However, a loss gap occurs between the two algorithms in an unbalanced 123-bus system, which originated from:

① Existing model-based algorithms (such as MISOCP) for solving three-phase optimal power flow issues commonly assume that the voltage phasors are nearly balanced [37]; otherwise, the nonconvex mathematical model is intractable. This assumption induces a computational error in the optimal loss calculated using MISOCP.

② MISOCP aims to discover a unique optimal solution for a given scenario, whereas DRL is prone to exploring a policy that can be utilized to acquire near-optimal decision-

making effects in numerous unseen scenarios. This normally incurs a weak sacrifice of solution optimality in the DRL-based algorithm, which further aggregates owing to the multiphase coupling nature of the 123-bus system and the multi-agent learning mode of the PS-PER-ITD3.

2) In contrast to the IEEE 33-bus and IEEE 118-bus systems, the decision-making time of MISOCP in the three-phase unbalanced 123-bus system significantly exceeds that of the PS-PER-ITD3. Consequently, for large-scale three-phase unbalanced distribution systems, the conventional MISOCP is limited to day-ahead or intraday optimizations with minute-level decision intervals. However, PS-PER-ITD3 still exhibits online millisecond-level decision-making capabilities.

In conclusion, compared with the conventional model-based MISOCP, the proposed PS-PER-ITD3 exhibits overwhelming generalization and decision-making rates in a three-phase unbalanced 123-bus system. In addition, the integrated PS mechanism retains sufficient optimization capability and scalability in the three-phase unbalanced test system because its gap from MISOCP is acceptable, even with the aforementioned unavoidable inherent sacrifices.

VI. Conclusion

An MADRL-based real-time optimization strategy of ADN is proposed to mitigate the voltage violations and network losses. Adopting the optimal switch deployment calculated by the prior reconfiguration preliminary, the ADN is then partitioned into multiple parallel regional agents and trained by the proposed PS-PER-ITD3. The PS mechanism is integrated into the ITD3-based MADRL algorithm, which significantly enhances its scalability and stability in larger systems by substituting the conventional global critic mechanism with shared network parameters and replay buffer.

In the numerical studies, the PS-PER-ITD3 and several other algorithms are tested on IEEE 33-bus, IEEE 118-bus, and three-phase unbalanced 123-bus systems. The simulation results confirm the scalability and superiority of the proposed PS-PER-ITD3 for real-time operation control of ADN. Moreover, a scenario-based comparative experiment demonstrates the necessity and effectiveness of the preliminary reconfiguration in the proposed PS-DER-ITD3. Based on the aforementioned experiments, the proposed PS-DER-ITD3 outperforms others on its convergence rapidity, online decision-making rate, excellent generalization, and scalability.

Further studies are required to explore better coordination modes between reconfiguration and DER control under the MADRL-based framework and to improve strategic scalability in large-scale systems.

Appendix A

A. Information of Test Distribution Systems

The information of the test distribution systems is shown in Tables AI-AIII and Figs. A1-A4. In Fig. A2, under the original topology without reconfiguration, lines with red triangle are closed, and dashed lines are open; under the reconfigured topology, lines with red triangle are open, and dashed lines are closed (remain unchanged for 24 hours).

TABLE AI
Regional Agent Partitioning Results

System	Partitioning result			
	Region 1	Region 2	Region 3	Region 4
IEEE 33-bus	Buses 1-11	Buses 12-22	Buses 23-33	-
IEEE 118-bus	Buses 1-29	Buses 30-58	Buses 59-87	Buses 88-116
123-bus	Buses 1-28	Buses 29-56	Buses 57-85	Buses 85-112

TABLE AII
DER Installations in Test Systems

System	PV	WT	ESS
IEEE 33-bus	Buses 6, 13, and 27	Buses 10, 16, and 30	Buses 4, 15, and 29
IEEE 118-bus	Buses 9, 55, 80, and 114	Buses 23, 51, 67, and 104	Buses 10, 39, 60, and 91
123-bus	Buses 6, 14, 28, 35, 46, 52, 57, 62, 75, 92, 101, 107 (only inverter-based PV)		

TABLE AIII
24-hour Optimal Switch Deployment of Reconfigured IEEE 33-bus System

Time (hour)	Open branch	Time (hour)	Open branch
1	14, 28, 33, 34, 37	13	33, 34, 35, 36, 37
2	9, 14, 28, 34, 37	14	5, 14, 33, 35, 36
3	9, 14, 28, 34, 37	15	9, 14, 34, 36, 37
4	5, 14, 33, 35, 36	16	9, 34, 35, 36, 37
5	5, 14, 28, 34, 35	17	9, 34, 35, 36, 37
6	14, 28, 33, 34, 35	18	14, 33, 34, 35, 36
7	9, 14, 28, 34, 37	19	9, 34, 35, 36, 37
8	9, 14, 28, 34, 37	20	9, 14, 28, 34, 37
9	9, 28, 34, 35, 37	21	9, 28, 33, 34, 35
10	9, 14, 28, 34, 37	22	5, 9, 14, 28, 34
11	14, 33, 34, 36, 37	23	5, 9, 14, 28, 35
12	9, 19, 28, 34, 35	24	5, 9, 14, 34, 37

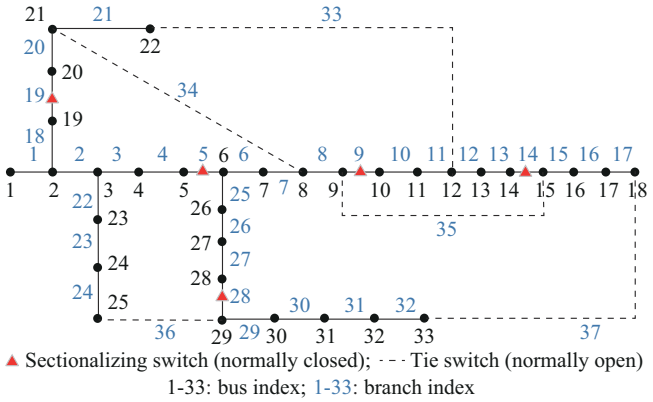

Fig. A1. Topology of IEEE 33-bus system.

1) To simplify the issue and reduce risks, the period of reconfiguration in the IEEE 118-bus system is 24-hour whereas that in the IEEE 33-bus system is 1-hour.

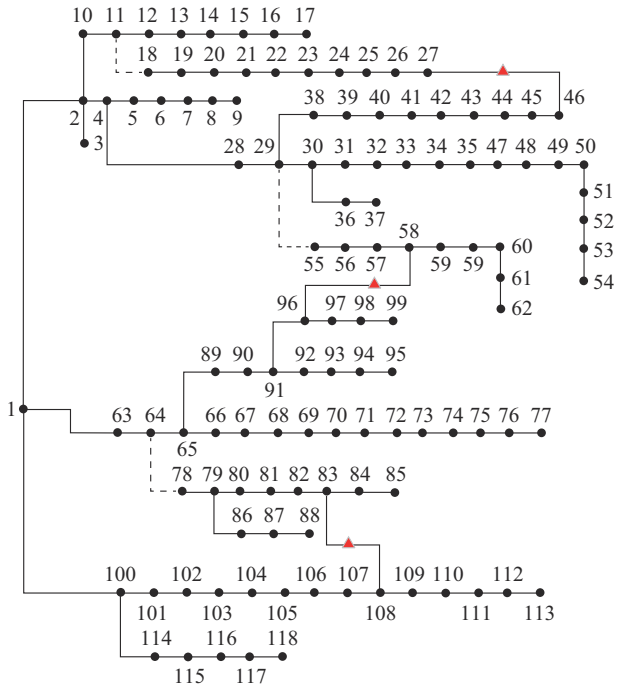

Fig. A2. Topology of IEEE 118-bus system.

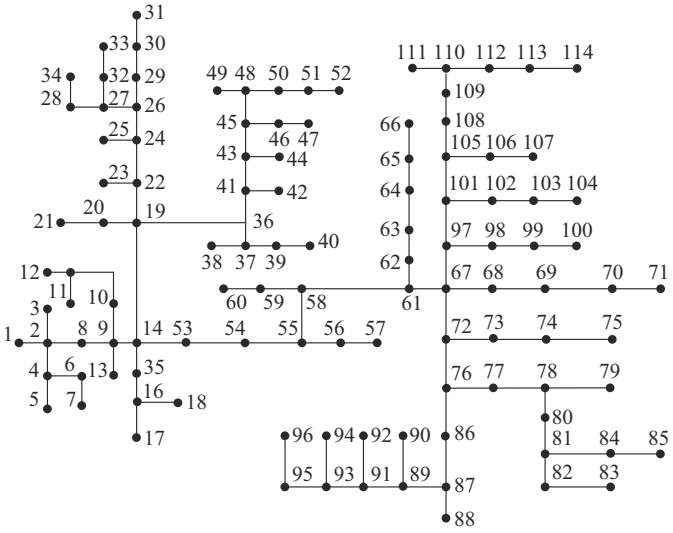

Fig. A3. Topology of three-phase unbalanced 123-bus system.

2) All the vacant branches and invalid buses of the three-phase unbalanced 123-bus system are omitted, and hence, there are only 114 valid buses.

3) As discussed in [13], the conventional partitioning of regional agents is based on a voltage-sensitivity matrix to optimize the regulatory effects. However, the proposed strategy is based on a preliminary reconfiguration; therefore, the topology of ADN during MADRL training is dynamic, which leads to a dynamic voltage sensitivity matrix. Therefore, the method in [13] has little effect, and we directly partition the ADN according to the bus sequence.

B. Operation Optimization Model of Three-phase Unbalanced 123-bus Network

The operation optimization model of three-phase unbalanced 123-bus network is shown in Fig. A4.

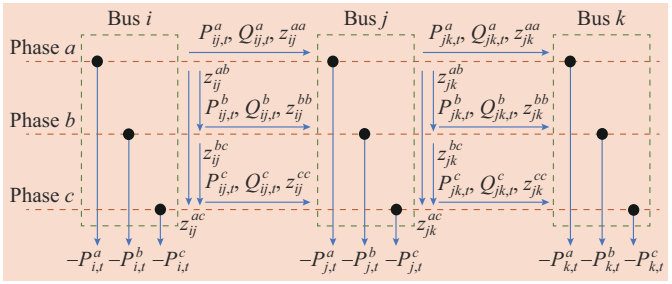

Fig. A4. Operation optimization model of three-phase unbalanced 123-bus system.

1) Objective Function

An optimization model for the three-phase unbalanced ADN is constructed to minimize the comprehensive objective, which is composed of active network loss and voltage violation penalty.

The detailed mathematical form is given as:

$$Obj = \min\left(\sum_{t=1}^{T} P_t^{loss} + \gamma \sum_{j\in\Omega_{bus}} \sum_{\phi\in\{a,b,c\}} \sum_{t=1}^{T} V_{j,t}^{\phi,pen}\right) \tag{A1}$$

where γ, $V_{j,t}^{\phi,pen}$, and ϕ represent the multiplying coefficient of voltage violation penalty, voltage violation penalty, and phases in $\{a,b,c\}$, respectively. The corresponding two compositions are calculated by

$$P_t^{loss} = \sum_{ij\in\Omega_{branch}} \sum_{\phi\in\{a,b,c\}} \left(\frac{(P_{ij,t}^{\phi})^2 + (Q_{ij,t}^{\phi})^2}{(v_{i,t}^{\phi})^2} r_{ij}^{\phi}\right) \tag{A2}$$

$$V_{j,t}^{\phi,pen} = ReLU(v_{\max} - v_{j,t}^{\phi}) + ReLU(v_{j,t}^{\phi} - v_{\min}) \tag{A3}$$

where $P_{ij,t}^{\phi}$ and $Q_{ij,t}^{\phi}$ are the real-time active and reactive power flows of branch ij at phase ϕ, respectively; r_{ij}^{ϕ} is the self-resistance of branch ij at phase ϕ; $V_{j,t}^{\phi,pen}$ is the penalty to voltage violation with $ReLU(x)=\max(0,x)$; $(v_{\min}, v_{\max})$ is the acceptable range of voltage at each phase; and $v_{j,t}^{\phi}$ is the voltage amplitude of bus j at phase ϕ.

2) Constraints

1) Three-phase unbalanced power flow equation

$$\boldsymbol{P}_{i,t} = |\boldsymbol{v}_{i,t}| \odot \sum_{j\in\Omega_{bus}} (\boldsymbol{G}_{ij} \odot C(\boldsymbol{\theta}_{ij,t}) + \boldsymbol{B}_{ij} \odot S(\boldsymbol{\theta}_{ij,t}))\,|\boldsymbol{v}_{j,t}| \tag{A4}$$

$$\boldsymbol{Q}_{i,t} = |\boldsymbol{v}_{i,t}| \odot \sum_{j\in\Omega_{bus}} (\boldsymbol{G}_{ij} \odot S(\boldsymbol{\theta}_{ij,t}) - \boldsymbol{B}_{ij} \odot C(\boldsymbol{\theta}_{ij,t}))\,|\boldsymbol{v}_{j,t}| \tag{A5}$$

where $\boldsymbol{P}_{i,t}=[P_{i,t}^a, P_{i,t}^b, P_{i,t}^c]^{\mathrm{T}}$ represents the active power injections on bus i; $\boldsymbol{Q}_{i,t}=[Q_{i,t}^a, Q_{i,t}^b, Q_{i,t}^c]^{\mathrm{T}}$ represents the reactive power injections on bus i; $\boldsymbol{G}_{ij}$ and $\boldsymbol{B}_{ij}$ are the real and imaginary submatrices of branch ij in the admittance matrix $\boldsymbol{Y}\in\mathbb{R}^{3n_{\#}\times 3n_{\#}}$, respectively, $n_{\#}$ is the number of bus set Ω_{bus}; $C(\cdot)$ denotes the cosine calculation; $S(\cdot)$ denotes the sine calculation; $\boldsymbol{v}_{i,t}=[v_{j,t}^a, v_{j,t}^b, v_{j,t}^c]^{\mathrm{T}}$; and $\boldsymbol{\theta}_{ij,t}$ is denoted as:

$$\boldsymbol{\theta}_{ij,t} = \begin{bmatrix} \theta_{i,t}^a - \theta_{j,t}^a & \theta_{i,t}^a - \theta_{j,t}^b & \theta_{i,t}^a - \theta_{j,t}^c \\ \theta_{i,t}^b - \theta_{j,t}^a & \theta_{i,t}^b - \theta_{j,t}^b & \theta_{i,t}^b - \theta_{j,t}^c \\ \theta_{i,t}^c - \theta_{j,t}^a & \theta_{i,t}^c - \theta_{j,t}^b & \theta_{i,t}^c - \theta_{j,t}^c \end{bmatrix} \tag{A6}$$

$$\begin{cases} P_{j,t}^{\phi} = P_{j,t}^{PV,\phi} - P_{j,t}^{load,\phi} \\ Q_{j,t}^{\phi} = Q_{j,t}^{PV,\phi} - Q_{j,t}^{load,\phi} \end{cases} \quad j\in\Omega_{bus}\backslash ref \tag{A7}$$

$$\begin{cases} P_{j,t}^{\phi} = P_{j,t}^{root,\phi} + P_{j,t}^{PV,\phi} \\ Q_{j,t}^{\phi} = Q_{j,t}^{root,\phi} + Q_{j,t}^{PV,\phi} \end{cases} \quad j \in ref \tag{A8}$$

where $j \in ref$ represents the root bus of the unbalanced distribution system.

2) Operation constraints of inverter-based PV

$$0 < P_{j,t}^{PV,\phi} < P_{j,t}^{PV,MPPT,\phi} \quad \forall j \in \Omega_{bus}, \forall t \tag{A9}$$

$$-\sqrt{(S_j^{inv,\phi})^2 - (P_{j,t}^{PV,\phi})^2} < Q_{j,t}^{PV,\phi} < \sqrt{(S_j^{inv,\phi})^2 - (P_{j,t}^{PV,\phi})^2} \quad \forall j \in \Omega_{bus}, \forall t \tag{A10}$$

where $P_{j,t}^{PV,MPPT,\phi}$ is the maximum output active power of the inverter-based PV installed on phase ϕ of bus j; $S_j^{inv,\phi}$ is the nominal capacity of the PV inverter; and $Q_{j,t}^{PV,\phi}$ is the actual output reactive power of PV inverter installed on phase ϕ of bus j, which can be adjusted in quadrants I and IV of the P-Q coordinate system dynamically. Note that only the unbalanced test system utilizes inverter-based PV in this paper.

3) POMDP Modeling

This paper formulates the operation optimization of three-phase unbalanced distribution system as a Dec-POMDP model.

1) Observation of the k^{th} agent

$$o_{k,t} = (P_{j,t}^{PV,\phi}, v_{j,t}^{\phi}, P_{k,t}^{load}, Q_{k,t}^{load}) \quad \forall j \in \Omega_k, \phi \in \{a,b,c\} \tag{A11}$$

where $P_{k,t}^{load}$ is the total active load in regional agent k; and $Q_{k,t}^{load}$ is the total reactive load in regional agent k.

2) Action of agent k

$$a_{k,t} = Q_{j,t}^{PV,\phi} \quad \forall j \in \Omega_k, \phi \in \{a,b,c\} \tag{A12}$$

3) Constraints of agent k

$$0 < P_{j,t}^{PV,\phi} < P_{j,t}^{PV,MPPT,\phi} \quad \forall j \in \Omega_k, \phi \in \{a,b,c\} \tag{A13}$$

$$-\sqrt{(S_j^{inv,\phi})^2 - (P_{j,t}^{PV,\phi})^2} < Q_{j,t}^{PV,\phi} < \sqrt{(S_j^{inv,\phi})^2 - (P_{j,t}^{PV,\phi})^2} \quad \forall j \in \Omega_k, \phi \in \{a,b,c\} \tag{A14}$$

4) Reward

$$r_t = -P_t^{loss} - \gamma \sum_{j \in \Omega_{bus}} \sum_{\phi \in \{a,b,c\}} V_{j,t}^{\phi,pen} \tag{A15}$$

REFERENCES

[1] G. Švenda, I. Krstić, S. Kanjuh *et al*., "Volt var watt optimization in distribution network with high penetration of renewable energy sources and electric vehicles," in *Proceedings of 2022 IEEE PES Innovative Smart Grid Technologies Conference Europe* (*ISGT-Europe*), Novi Sad, Serbia, Oct. 2022, pp. 1-5.

[2] S. H. Low, "Convex relaxation of optimal power flow – part I: formulations and equivalence," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 15-27, Mar. 2014.

[3] S. H. Low, "Convex relaxation of optimal power flow – part II: exactness," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 2, pp. 177-189, Jun. 2014.

[4] Y. Fan, L. Feng, and G. Li, "Dynamic optimal power flow in distribution networks with wind/PV/storage based on second-order cone programming," in *Proceedings of 2020 5th Asia Conference on Power and Electrical Engineering* (*ACPEE*), Chengdu, China, Jun. 2020, pp. 1136-1142.

[5] M. Niu, C. Wan, and Z. Xu, "A review on applications of heuristic optimization algorithms for optimal power flow in modern power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 2, no. 4, pp. 289-297, Dec. 2014.

[6] Y. Ai, M. Du, Z. Pan *et al*., "The optimization of reactive power for distribution network with PV generation based on NSGA-III," *CPSS Transactions on Power Electronics and Applications*, vol. 6, no. 3, pp. 193-200, Sept. 2021.

[7] 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.

[8] H. Liu and W. Wu, "Two-stage deep reinforcement learning for inverter-based volt-var control in active distribution networks," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2037-2047, May 2021.

[9] D. Cao, W. Hu, J. Zhao *et al*., "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.

[10] X. Sun and J. Qiu, "Two-stage volt/var control in active distribution networks with multi-agent deep reinforcement learning method," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 2903-2912, Jul. 2021.

[11] D. Hu, Z. Ye, Y. Gao *et al*., "Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4873-4886, Nov. 2022.

[12] H. Wu, Z. Xu, M. Wang *et al*., "Two-stage voltage regulation in power distribution system using graph convolutional network-based deep reinforcement learning in real time," *International Journal of Electrical Power & Energy Systems*, vol. 151, p. 109158, Sept. 2023.

[13] H. Liu, C. Zhang, Q. Chai *et al*., "Robust regional coordination of inverter-based volt/var control via multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5420-5433, Nov. 2021.

[14] J. Zhang, Y. Guan, L. Che *et al*., "EV charging command fast allocation approach based on deep reinforcement learning with safety modules," *IEEE Transactions on Smart Grid*, doi: 10.1109/TSG.2023.3281782

[15] Y. Zhang, X. Wang, J. Wang *et al*., "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 361-371, Jan. 2021.

[16] Z. Yin, S. Wang, and Q. Zhao, "Sequential reconfiguration of unbalanced distribution network with soft open points based on deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 1, pp. 107-119, Jan. 2023.

[17] J. Zhang, M. Cui, and Y. He, "Dual timescales voltages regulation in distribution systems using data-driven and physics-based optimization," *IEEE Transactions on Industrial Informatics*, doi: 10.1109/TII.2023.3274216

[18] Y. Pei, J. Zhao, Y. Yao *et al*., "Multi-task reinforcement learning for distribution system voltage control with topology changes," *IEEE Transactions on Smart Grid*, vol. 14, no. 3, pp. 2481-2484, May 2023.

[19] M. R. Dorostkar-Ghamsari, M. Fotuhi-Firuzabad, M. Lehtonen *et al*., "Value of distribution network reconfiguration in presence of renewable energy resources," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1879-1888, May 2016.

[20] R. A. Jabr, R. Singh, and B. C. Pal, "Minimum loss network reconfiguration using mixed-integer convex programming," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 1106-1115, May 2012.

[21] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," *Proceedings of International Conference on Autonomous Agents and Multiagent Systems*, vol. 10642, pp. 66-83, Nov. 2017.

[22] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of International Conference on Machine Learning*, Stockholm, Sweden, Jun. 2018, pp. 1587-1596.

[23] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 895-943, Apr. 2022.

[24] D. Qiu, Y. Ye, D. Papadaskalopoulos *et al*., "Scalable coordinated management of peer-to-peer energy trading: a multi-cluster deep reinforcement learning approach," *Applied Energy*, vol. 292, p. 116940, Jun. 2021.

[25] N. Yang, B. Ding, P. Shi *et al*., "Improving scalability of multi-agent reinforcement learning with parameters sharing," in *Proceedings of 2022 IEEE International Conference on Joint Cloud Computing* (*JCC*), Fremont, USA, Aug. 2022, pp. 37-42.

[26] 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.

[27] X. Liu, S. Li, and J. Zhu, "Optimal coordination for multiple network-constrained VPPs via multi-agent deep reinforcement learning," *In IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3016-3031, Jul.

2023.
[28] Y. Tao, J. Qiu, S. Lai *et al*., "A data-driven agent-based planning strategy of fast-charging stations for electric vehicles," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 3, pp. 1357-1369, Jul. 2023.
[29] A. Tampuu, T. Matiisen, D. Kodelja *et al*. (2015, Nov.). Multiagent cooperation and competition with deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1511.08779
[30] R. Lowe, Y. Wu, A. Tamar *et al*. (2017, Jun.). Multi-agent actor-critic for mixed cooperative-competitive environments. [Online]. Available: https://arxiv.org/abs/1706.02275
[31] J. Ackermann, V. Gabler, T. Osa *et al*. (2011, Oct.). Reducing overestimation bias in multi-agent domains using double centralized critics. [Online]. Available: https://arxiv.org/abs/1910.01465
[32] D. Qiu, J. Wang, Z. Dong *et al*., "Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading," *IEEE Transactions on Power Systems*, vol. 38, no. 5, pp. 4853-4866, Sept. 2023.
[33] T. Schaul, J. Quan, I. Antonoglou *et al.* (2015, Nov.). "Prioritized experience replay. [Online]. Available: https://arxiv.org/abs/1511.05952
[34] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12-19, Feb. 2011.
[35] D. Zhang, Z. Fu, and L. Zhang, "An improved TS algorithm for loss-minimum reconfiguration in large-scale distribution systems," *Electric Power Systems Research*, vol. 77, no. 5-6, pp. 685-694, Apr. 2007.
[36] W. H. Kersting, "Radial distribution test feeders," *IEEE Transactions on Power Systems*, vol. 6, no. 3, pp. 975-985, Aug. 1991.
[37] R. R. Nejad and W. Sun, "Distributed load restoration in unbalanced active distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5759-5769, Sept. 2019.

Jie Xu received the B. E. degree in electrical engineering from the China University of Petroleum (East China), Qingdao, China, in 2022. He is currently pursuing the M.Sc. degree in electrical engineering in Sichuan University, Chengdu, China. His research interests include operation optimization of distribution network, deep reinforcement learning, and voltage control.

Hongjun Gao received the B.S., M.S., and Ph.D. degrees in electrical engineering from Sichuan University, Chengdu, China, in 2011, 2014, and 2017, respectively. From 2015 to 2016, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of Wisconsin – Milwaukee, Milwaukee, USA. He is currently an Associate Professor with the College of Electrical Engineering, Sichuan University. His research interests include active distribution system planning and operation, economic dispatch, distributed generation integration, and multi-energy system optimization.

Renjun Wang received the M.S. degree in electrical engineering from Sichuan University, Chengdu, China, in 2021. He is currently pursuing the Ph.D. degree in electrical engineering in Sichuan University. His research interests include active distribution system planning and operation and deep reinforcement learning.

Junyong Liu received the Ph.D. degree in electrical engineering from Brunel University, Uxbridge, U.K., in 1998. He is currently a Professor with the College of Electrical Engineering, Sichuan University, Chengdu, China, where he is the Director with Sichuan Province Key Smart Grid Laboratory, Chengdu, China. His current research interests include power system planning, operation, and computer applications.