A Multi-agent Robust Deep Reinforcement Learning Approach for Coordination of Power Distribution Networks and Microgrids with Limited Information Exchange

Jiahui Jin, *Graduate Student Member*, *IEEE*, Guoqiang Sun, *Member*, *IEEE*, Sheng Chen, *Senior Member*, *IEEE*, Yaping Li, Yingqi Liao, Wenbo Mao, and Lu Shen

***Abstract*—The coordination of power distribution networks (PDNs) and microgrids (MGs) is challenging due to the abundant resources and their dispersed geographical distribution, making centralized computation inefficient. To address this issue, we propose a coordination framework with single leader and multiple followers that allows limited information exchange. In this framework, the PDN operators act as leaders, while the MG operators act as followers. However, variations in load and renewable energy during MG scheduling intervals can cause variability in power transactions between PDNs and MGs. This variability can reduce the net revenue of MGs and increase the operation costs of PDNs, which makes it essential to consider the worst-case fluctuations. We introduce a multi-agent robust deep reinforcement learning (MARDRL) approach for coordination of PDNs and MGs, accounting for the worst-case scenarios. The numerical results on the test systems verify the effectiveness of the proposed approach in enhancing the coordination of PDNs and MGs.**

***Index Terms*—Power distribution network, microgrid, leader, follower, renewable energy, deep reinforcement learning, information exchange, agent.**

NOMENCLATURE

A. Indices, Sets, and Functions

θ	Index of neural network parameter
$\mathcal{N}$	Function of Gaussian distribution
π_{CS}^{m}	Index of policy for certain state in set Π_{CS}^{m}
$\pi_{US}^{*,m}$	Index of policy for the worst-case scenario in set $\Pi_{US}^{*,m}$
*	Index of the worst-case scenario
adjh	Index of hour of adjacent similar day
cu	Index of load curtailment
CS	Index of Markov decision process under certain state
e	Index of energy storage system in set *E*
$\underline{f_m}$	Lower bound function of uncertainty in microgrid (MG) *m*
$\overline{f_m}$	Upper bound function of uncertainty in MG *m*
gt	Index of gas turbine generator in set *GT*
h	Index of hour of current day
j	Index of bus in set *B*
lo	Index of load
m	Index of MG in set *M*
max	Function of the maximum value
min	Function of the minimum value
re	Index of renewable energy in set *RE*
t	Index of 5-min interval in set *T*
tr	Index of transaction
up	Index of upper-level power grid
US	Index of Markov decision process under uncertain state

B. Parameters

α^{tr}	Reward-penalty coefficient of transaction power
γ	Discount factor
Σ_{CS}^{m}	Preset covariance matrix of action under policy π_{CS}^{m}
$\Sigma_{US}^{*,m}$	Preset covariance matrix of action under policy $\pi_{US}^{*,m}$
Δt	5-min interval
η	Learning rate

Manuscript received: January 12, 2025; revised: March 23, 2025; accepted: May 30, 2025. Date of CrossCheck: May 30, 2025. Date of online publication: July 15, 2025.

This work was supported by State Grid Corporation of China (No. 5108-202318433A-3-2-ZN).

J. Jin, G. Sun, and S. Chen (corresponding author) are with the School of Electrical and Power Engineering, Hohai University, Nanjing 210098, China (e-mail: 230206030007@hhu.edu.cn; hhusunguoqiang@163.com; chenshenghhu@163.com).

Y. Li and W. Mao are with the China Electric Power Research Institute Co., Ltd., Nanjing 210028, China (e-mail: liyaping@epri.sgcc.com.cn; maowenbo@epri.sgcc.com.cn).

Y. Liao and L. Shen are with the Nanjing Power Supply Branch, State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210019, China (e-mail: 112144566@qq.com; lushenjsnj@outlook.com).

DOI: 10.35833/MPCE.2025.000035

η_{ch} Charging efficiency of energy storage system

η_{dis} Discharging efficiency of energy storage system

η^{re} Active power conversion efficiency of inverter

a^{gt}, b^{gt}, c^{gt} Cost coefficients of gas turbine generator *gt*

c_j^{cu} Load curtailment price of bus *j*

c^e Charging/discharging cost of unit power in energy storage system *e*

$C_t^{m,e}$ Rated capacity of energy storage system *e* in MG *m* at interval *t*

c^{re} Generation cost of unit power in renewable energy *re*

c^{up} Electricity price of upper-level power grid

$\overline{P_j^{cu}}$ The maximum load curtailment of bus *j*

$\overline{P_{ch}^{m,e}}$ Preset maximum charging power of energy storage system *e* in MG *m*

$\underline{P_{dis}^{m,e}}$ Preset minimum discharging power of energy storage system *e* in MG *m*

$\overline{Q^{m,re}}$ Preset maximum output reactive power of renewable energy *re* in MG *m*

$\overline{RD}^{gt}$ The maximum ramping down power of gas turbine generator *gt*

$\overline{RU}^{gt}$ The maximum ramping up power of gas turbine generator *gt*

$\overline{SOC^{m,e}}$ The maximum state-of-charge of energy storage system *e* in MG *m*

$\underline{SOC^{m,e}}$ The minimum state-of-charge of energy storage system *e* in MG *m*

$\overline{S^{m,re}}$ The maximum apparent power of renewable energy *re* in MG *m*

C. Model Variables

$\lambda_{t,ch}^{m,e}$ Charging state of binary variable of energy storage system *e* in MG *m* at interval *t*

$\lambda_{t,dis}^{m,e}$ Discharging state of binary variable of energy storage system *e* in MG *m* at interval *t*

$\omega_t^{m,e}$ Output power ratio of energy storage system *e* in MG *m* at interval *t*

$\omega_t^{m,re}$ Output power ratio of renewable energy *re* in MG *m* at interval *t*

c_t^m Hourly electricity price in MG *m* at interval *t*

F_t^m Net revenue in MG *m* at initial time point of interval *t*

$F_t^{tr,m}$ Reward-penalty of transaction power P_t^m

F_t^{pd} Operation cost of power distribution network (PDN) at initial time point of interval *t*

l_t^m Locational marginal price in MG *m* at initial time point of interval *t*

$P_{j,t}^{cu}$ Load curtailment of bus *j* at interval *t*

$P_t^{m,e}$ Output power of energy storage system *e* in MG *m* at interval *t*

P_t^{gt} Output power of gas turbine generator *gt* at interval *t*

P_t^m Committed transaction power between MG *m* and PDN at initial time point of interval *t*

$P_t^{m,av}$ Actual transaction power between MG *m* and PDN at initial time point of interval *t*

$P_t^{m,re}$ Output power of renewable energy *re* in MG *m* at interval *t*

$\overline{P_{ch,t}^{m,e}}$ The maximum charging power of energy storage system *e* in MG *m* at interval *t*

$\underline{P_{dis,t}^{m,e}}$ The minimum discharging power of energy storage system *e* in MG *m* at interval *t*

$P_t^{m,lo}$ Active load in MG *m* at interval *t*

$P_t^{m,re}$ Output active power of renewable energy *re* in MG *m* at interval *t*

P_t^{up} Purchased power from upper-level power grid at interval *t*

$Q_t^{m,lo}$ Reactive load in MG *m* at interval *t*

$Q_t^{m,re}$ Output reactive power of renewable energy *re* in MG *m* at interval *t*

$\overline{Q_t^{m,re}}$ The maximum output reactive power of renewable energy *re* in MG *m* at interval *t*

$\underline{Q_t^{m,re}}$ The minimum output reactive power of renewable energy *re* in MG *m* at interval *t*

$SOC_t^{m,re}$ State-of-charge of energy storage system *e* in MG *m* at interval *t*

D. Reinforcement Learning Variables

$\mu_{CS,t}^m$ Mean of action under policy π_{CS}^m at interval *t*

$\mu_{US,t}^{*,m}$ Mean of action under policy $\pi_{US}^{*,m}$ at interval *t*

a_t^m Action in MG *m* at interval *t* in set A^m

co_t^m Observation in MG *m* at interval *t*

$d_{\pi_{CS}^m}$ Probability of reaching state under policy π_{CS}^m

$d_{\pi_{US}^{*,m}}$ Probability of reaching state under policy $\pi_{US}^{*,m}$

D_J Loss of expected discounted return

D_{KL} Model output of Kullback-Leibler (KL) divergence

$D_{KL,\theta}$ Neural network output of KL divergence

D_{TV} Total variation distance

J Expected discounted return under policy

$p(s_{t+1}^m|s_t^m, a_t^m)$ Probability of transitioning from state s_t^m to state s_{t+1}^m with action a_t^m

r_t^m Reward in MG *m* at interval *t* in set S^m

s_t^m State in MG *m* at initial time point of interval *t*

$\underline{s_t^m}$ The minimum state s_t^m at interval *t*

$\overline{s_t^m}$ The maximum state s_t^m at interval *t*

$s_t^{*,m}$ The worst-case scenario of state in MG *m* at interval *t*

$s_{t,US}^m$ Mapping state in Markov decision process under certain states by policy $\pi_{US}^{*,m}$

uo_t^m Uncertain observation in MG *m* at interval *t*

$V_{\pi_{US}^{*,m}}$ State value function of state under policy $\pi_{US}^{*,m}$

$V_{\pi_{CS}^m}$ State value function of state under policy π_{CS}^m

I. Introduction

RECENTLY, the share of renewable energy in power distribution networks (PDNs) has grown significantly [1]. However, the inherent uncertainty and variability of renewable energy generation pose challenges for the planning and operation of PDNs. To mitigate these challenges, many energy storage systems have been implemented, leading to the high operation cost of PDNs [2]. Furthermore, the increasing number of dispatchable devices and their frequent dispatches escalate the computational complexity of the operation of PDNs. Interconnecting PDNs with microgrids (MGs) allows some renewable energy to be shared with MGs, reducing the need for energy storage systems [3] and alleviating computational burdens [4]. Enhanced collaboration between PDNs and MGs is anticipated to address these issues effectively.

Therefore, the MGs considered in this paper are grid-connected. When PDNs and MGs interconnect, their transactions primarily involve power and pricing at the point of common coupling [2]. This interaction presents two main challenges. First, fluctuations in transaction power can result in stability or reliability issues in PDNs. Second, pricing is complex because PDNs and MGs are distinct entities with different system operators and optimization goals. As a result, effective economic incentives are essential to promote MG development and integration, safeguarding the interests of all parties involved.

Several approaches have been explored to coordinate the operations of PDNs and MGs. Reference [5] proposes a renewable energy buyback program with dynamic pricing to achieve smart grid energy efficiency targets. Reference [6] presents a dual-layer optimization model, incorporating demand response. Reference [7] uses probabilistic modeling for MG energy and load to optimize operations and minimize costs. Reference [8] introduces a two-layer model for comprehensive pricing of active and reactive power, focusing on electricity market interactions and virtual power plant profits. Reference [9] proposes a Stackelberg game framework for these operations with a dual-layer model for MG energy management in distribution markets. Reference [10] proposes a planning and operation model for MGs with pumped hydro storage, serving the PDNs exclusively. However, as the scale of PDNs and MGs grows, system modeling becomes more complex, and computational demands rise due to increasing variables and constraints in the optimization models. Model-based optimization approaches often demand substantial computation resources, which are difficult to satisfy the requirement of real-time applications in practice.

Deep reinforcement learning offers several advantages: the ability to handle highly complex nonlinear systems, adaptability to high-dimensional data, and high computational efficiency in forward propagation. Therefore, it is increasingly being used for the coordination of PDNs and MGs. Notable research works are as follows. Reference [11] applies deep deterministic policy gradients to manage wind power output. Reference [12] proposes an energy trading algorithm based on deep reinforcement learning to solve the supply and demand mismatch problem of smart grids with a large number of MGs without relying on power supply and demand models of other MGs. Reference [13] proposes a federated decentralized reinforcement learning algorithm addressing privacy and scalability. Reference [14] introduces individual attention mechanisms for agent-specific reward information. Reference [15] uses double-delay deep deterministic policy gradients with a nonlinear battery degradation model for MG energy management. Reference [16] develops a multi-stage reward mechanism incorporating expert decisions to avoid suboptimal strategies. Reference [17] personalizes demand fulfillment with weighted vector adjustments in reward functions. Reference [18] proposes a deep reinforcement learning approach for resilient MG partition models.

The existing literature has notable gaps. Firstly, in trading between PDNs and MGs, time-of-use (TOU) pricing [5], [6] can lead to excessive incentives and insufficient responsiveness to short-term fluctuations. Real-time pricing (RTP) [9] demands significant computation and communication resources. Second, addressing uncertainty within scheduling intervals, reinforcement learning with hard constraints [13], [17] requires high-frequency sampled data, but even then, discrete data may overlook the worst-case scenario. On the other hand, reinforcement learning with soft constraints [11], [14], [18] diminishes benefits and fails to address the worst-case scenarios. Additionally, practical applications prefer offline training with historical data for power system safety. Addressing these gaps, the primary contributions of this study are as follows.

1) We develop a coordination framework with single leader and multiple followers for PDNs and MGs with limited information exchange. We propose a step-wise optimal pricing approach suitable for reinforcement learning training processes, which is distinct from TOU and RTP. This iterative solution effectively balances the operations for both PDN and DG.

2) We propose a multi-agent robust deep reinforcement learning (MARDRL) approach using semi-centralized training and decentralized execution to enhance the coordination. We model the coordination issues of MGs as a Markov decision process (MDP) under uncertain state (US), i.e., MDP-US, while simulating the optimal power flow of PDNs and MGs together with historical data as the environment for offline training.

3) Our experimental results validate the impact of the proposed approach in the coordination of PDNs and MGs, demonstrating mutual benefits and adaptability.

The remainder of this paper is organized as follows. Section II presents the problem formulation. Section III details the proposed MARDRL approach. Section IV summarizes the computational results of the test systems. Section V concludes the findings of this paper.

II. Problem Formulation

This section initially introduces the Stackelberg game framework for coordination of PDNs and MGs. Operation models for the PDN operators and the MG operators are detailed in Sections II-B and II-C.

Two key assumptions are as follows.

1) In the operation model, there is no information sharing among MGs.

2) In the PDNs, the electricity price is sent every hour. In the MGs, the transaction power is updated every 5 min.

A. Conceptual Framework for Stackelberg Game

To ensure the responsiveness of MGs, electricity prices are transmitted from the PDNs to MGs before each hour. Figure 1 illustrates the Stackelberg Game framework for coordination of PDNs and MGs, where DLMP is short for distribution locational marginal price. The PDN operators and MG operators act as the leader and followers, respectively.

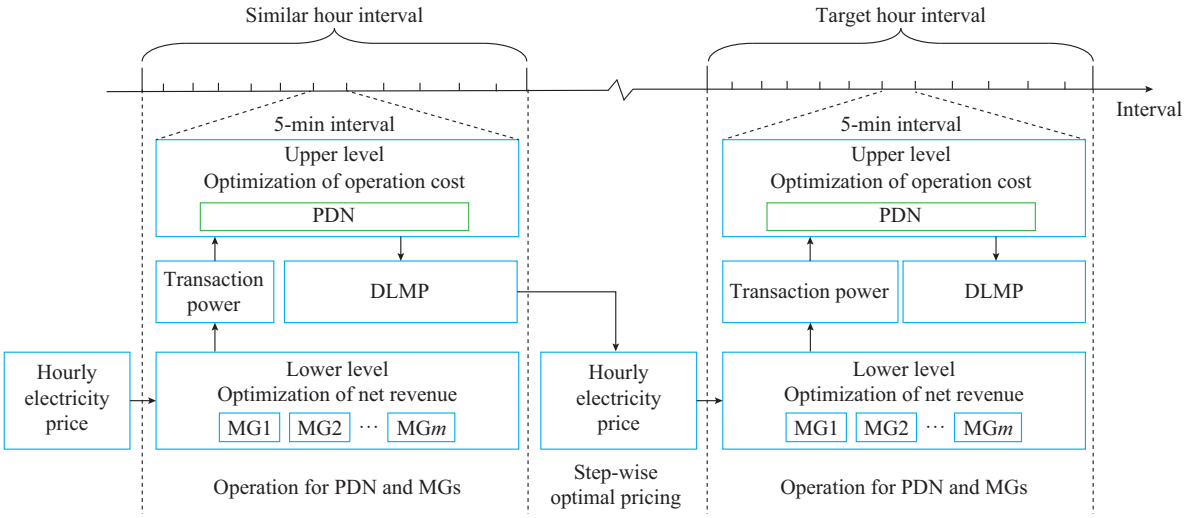

Fig. 1. Stackelberg game framework for coordination of PDNs and MGs.

The iterative optimization cycle continues until the training iteration limit is reached, at which point the game concludes. The process has two parts:

1) Operations for the PDNs and MGs are conducted at 5-min intervals. At the lower level, MGs independently determine their transaction power with the PDNs based on the received hourly electricity prices to maximize overall revenues. Subsequently, the PDNs minimize operation costs at the upper level, using the received transaction power to calculate DLMP for each MG.

2) Step-wise optimal pricing occurs at the end of each hour. The PDN operators update MG prices using DLMPs and the corresponding transaction power from the similar hour.

The interaction between PDNs and MGs includes hourly electricity pricing and penalties for discrepancies between actual and committed transaction power. The details are given as:

$$c_t^m = \frac{\sum \left| l_t^m P_t^m \right|_{t \in adjh}}{\sum \left| P_t^m \right|_{t \in adjh}} \tag{1}$$

$$F_t^{tr,m} = \alpha^{tr}\left(P_t^{m,av} - P_t^m \right) \tag{2}$$

Equation (1) represents the hourly electricity price updates. Similar hours are selected using the approach outlined in [19]. Equation (2) represents transaction penalties [20].

The pricing model proposed in this paper uses temporal decoupling and progressive relaxation to balance price responsiveness and computational complexity. Additionally, it aligns with the dynamic interaction mechanisms of reinforcement learning.

B. Operation Model for PDN Operators

To pursue tractability, the operation is formulated using conic relaxation. The PDN flow constraints are based on [21]. Details of the DLMP are provided in [9]. The remaining detailed model of the PDN operators is outlined as:

$$\min F_t^{pd} = \sum_{gt \in GT} a^{gt}\left(P_t^{gt} \right)^2 + b^{gt} P_t^{gt} + c^{gt} + c^{up} P_t^{up} + \sum_{j \in B} c_j^{cu} P_{j,t}^{cu} + \sum_{m \in M} c_t^m P_t^m \tag{3}$$

$$0 \le P_{j,t}^{cu} \le \overline{P_j^{cu}} \tag{4}$$

$$P_t^{gt} - P_{t-1}^{gt} \le \overline{RU}^{gt} \tag{5}$$

$$P_{t-1}^{gt} - P_t^{gt} \le \overline{RD}^{gt} \tag{6}$$

Equation (4) specifies the noncritical load curtailment to keep electrical parameters within limits. Equations (5) and (6) cover the ramping limits of gas turbine generators.

C. Operation Model for MG Operators

To improve the stability and efficiency of reinforcement learning, the power of the MG action devices, including the energy storage system and renewable energy inverters, will be standardized. The energy storage system uses lithium-ion batteries. The MG flow constraints are based on [6]. The remaining detailed MG operator model is expressed as:

$$\max F_t^m = c_t^m P_t^m - \sum_{e \in E} c^e \left| P_t^{m,e} \right| - \sum_{re \in RE} c^{re} P_t^{m,re} \tag{7}$$

$$\overline{P_{ch,t}^{m,e}} = \min\left(\frac{\overline{SOC^{m,e}} - SOC_t^{m,e} C^{m,e}}{\lambda_{t,ch}^{m,e} \eta_{ch} \Delta t}, \overline{P_{ch}^{m,e}} \right) \tag{8}$$

$$\underline{P_{dis,t}^{m,e}} = \max\left(\frac{SOC_t^{m,e} \eta_{dis} C^{m,e} - \underline{SOC^{m,e}}}{\lambda_{t,dis}^{m,e} \Delta t}, \underline{P_{dis}^{m,e}} \right) \tag{9}$$

$$P_t^{m,e} = \underline{P_{dis,t}^{m,e}} + \omega_t^{m,e}\left(\overline{P_{ch,t}^{m,e}} - \underline{P_{dis,t}^{m,e}} \right) \tag{10}$$

$$\lambda_{t,ch}^{m,e}+\lambda_{t,dis}^{m,e}\le 1 \tag{11}$$

$$SOC_{t+1}^{m,e}=SOC_{t}^{m,e}+\left(\lambda_{t,ch}^{m,e}\eta_{ch}+\frac{\lambda_{t,dis}^{m,e}}{\eta_{dis}}\right)\frac{P_{t}^{m,e}}{C^{m,e}}\Delta t \tag{12}$$

$$0\le\omega_{t}^{m,e}\le 1 \tag{13}$$

$$\overline{Q_{t}^{m,re}}=\min\left(\sqrt{\left(\overline{S^{m,re}}\right)^{2}-\left(\eta^{re}P_{t}^{m,re}\right)^{2}},\overline{Q^{m,re}}\right) \tag{14}$$

$$\underline{Q_{t}^{m,re}}=\max\left(-\sqrt{\left(\overline{S^{m,re}}\right)^{2}-\left(\eta^{re}P_{t}^{m,re}\right)^{2}},-\overline{Q^{m,re}}\right) \tag{15}$$

$$Q_{t}^{m,re}=\underline{Q_{t}^{m,re}}+\omega_{t}^{m,re}\left(\overline{Q_{t}^{m,re}}-\underline{Q_{t}^{m,re}}\right) \tag{16}$$

$$0\le\omega_{t}^{m,re}\le 1 \tag{17}$$

The set of constraints (8) - (13) describes the state-of-charge (SOC) and the charging/discharging power in the energy storage system, where (8) and (9) define the real-time charging and discharging power limits, respectively; (10) represents the output power of the system based on power ratios within these limits; (11) specifies the operation states for charging and discharging; (12) indicates SOC changes under charging and discharging; and (13) specifies the upper and lower power ratio limits for the energy storage system. The set of constraints (14)-(17) describes the reactive power output of inverters applied in renewable energy generation, where (14) and (15) set the upper and lower limits of the reactive power of the inverter; (16) governs the reactive power output of the inverter based on its ratio and thresholds mentioned earlier; and (17) specifies the upper and lower power ratio limits for the inverter.

III. Proposed MARDRL Approach

In this section, we introduce two following key assumptions in the proposed MARDRL approach.

1) In the MDP-US, optimal policies are adjusted for the worst-case scenarios. However, the environment still transitions from the current state to the next state, not the worst-case scenario.

2) US intervals are deterministic functions depending solely on current states, remaining consistent over time.

A. Formulation of MDP-US

The energy management in MGs is described as MDP-US. Each scheduling interval initiates decisions on the power ratios of energy storage systems and inverters based on current MG states. Key elements of the MDP-US formulation in MGs are as follows.

1) States: the MG states at each interval, including certain and uncertain observations are given as:

$$s_{t}^{m}=\left\{co_{t}^{m},uo_{t}^{m}\right\} \tag{18}$$

$$co_{t}^{m}=\left\{t,c_{t}^{m},SOC_{t}^{m,e}\right\} \tag{19}$$

$$uo_{t}^{m}=\left\{P_{t}^{m,re},P_{t}^{m,lo},Q_{t}^{m,lo}\right\} \tag{20}$$

The three elements in (19) are designated as certain observations due to their consistency throughout the scheduling interval. t serves as supplementary information to handle the non-stationary environments of MGs, and c_{t}^{m} directly impacts total revenue transactions. Considering the time-dependent constraints of energy storage system, $SOC_{t}^{m,e}$ indirectly restricts action ranges. The three elements in (20) are considered uncertain observations, directly impacting transaction power and varying within the scheduling interval.

2) State interval: to handle the uncertainty in net demands, the MG state is expanded to state interval, expressed as:

$$\left(\underline{s_{t}^{m}},\overline{s_{t}^{m}}\right)=\left(co_{t}^{m},\underline{f_{m}}\left(uo_{t}^{m}\right),\overline{f_{m}}\left(uo_{t}^{m}\right)\right) \tag{21}$$

The bound of uncertain observations is constructed by uncertainty upper bound function and uncertainty lower bound function, which are provided in [22]. By introducing the uncertain function, this model extends MDP-CS to MDP-US. The uncertain function operates independently of the actor network, enhancing adaptability and simplifying gradient calculations during decentralized training for MDP-US.

3) Actions: the standardized actions taken by MG operators at the beginning of each scheduling interval are defined as:

$$a_{t}^{m}=\left\{\omega_{t}^{m,e},\omega_{t}^{m,re}\right\} \tag{22}$$

4) Reward: it aligns with optimal energy goals and includes a penalty term to enforce MG energy constraints [17]. For a single time step, it is defined as:

$$r_{t}^{m}=r_{t}^{m}\left(s_{t}^{m},a_{t}^{m}\right)=F_{t}^{m}+F_{t}^{tr,m} \tag{23}$$

5) Policy: the policy π_{CS}^{m} is established for the MDP-CS in MG m. With the addition of uncertain functions, π_{CS}^{m} is changed to π_{US}^{m} for the MDP-US in MG m. The policy $\pi_{US}^{*,m}$ is selected within π_{US}^{m}. To address the worst-case scenario in MG m, since π_{CS}^{m}, π_{US}^{m}, and $\pi_{US}^{*,m}$ are generated from the same state space to produce the same action space, their sets are equivalent, i.e., $\Pi_{CS}^{m}=\Pi_{US}^{m}=\Pi_{US}^{*,m}$.

B. Analysis of Proposed MARDRL Approach

Traditional robust optimization uses a max-min framework, where the worst-case scenario employs robust strategies to maximize safety redundancy, leading to the lowest revenues. Based on the assumption 1) in Section III, which considers robust strategies within a single time segment, the state value function and expected discounted return are derived as:

$$V_{\pi_{US}^{*,m}}\left(s_{t}^{m}\right)=\sum_{a_{t}^{m}\in A^{m}}\pi_{US}^{*,m}\left(a_{t}^{m}|s_{t}^{m}\right)\cdot\left(r_{t}^{m}+\gamma\sum_{\substack{a_{t}^{m}\in A^{m}\\ s_{t}^{m},s_{t+1}^{m}\in S^{m}}}p\left(s_{t+1}^{m}|s_{t}^{m},a_{t}^{m}\right)V_{\pi_{CS}^{m}}\left(s_{t+1}^{m}\right)\right) \tag{24}$$

$$J\left(\pi_{US}^{*,m}\right)=\sum_{s_{t}^{m}\in S^{m}}d_{\pi_{US}^{*,m}}\left(s_{t}^{m}\right)V_{\pi_{US}^{*,m}}\left(s_{t}^{m}\right) \tag{25}$$

Since robust strategies yield the lowest revenues, the loss in expected discounted return between robust and deterministic strategies is maximized. To simplify the computation of robust strategies, the problem of minimizing revenue is transformed into maximizing the loss. Based on (25), the loss of expected discounted return is calculated as:

$$D_{J}\left(\pi_{US}^{*,m}\middle\|\pi_{CS}^{m}\right)=J\left(\pi_{CS}^{m}\right)-J\left(\pi_{US}^{*,m}\right)=\sum_{s_{t}^{m}\in S^{m}}d_{\pi_{CS}^{m}}\left(s_{t}^{m}\right)V_{\pi_{CS}^{m}}\left(s_{t}^{m}\right)-\sum_{s_{t}^{m}\in S^{m}}d_{\pi_{US}^{*,m}}\left(s_{t}^{m}\right)V_{\pi_{US}^{*,m}}\left(s_{t}^{m}\right) \tag{26}$$

Due to the nonlinearity of power flow, identifying the worst-case scenario under US is challenging. However, it is possible to demonstrate that the difference is bounded. By the triangle inequality and the total variation distance, an upper bound for the loss can be derived as:

$$D_J\left(\pi_{US}^{*,m}\middle\|\pi_{CS}^{m}\right)\le 2\max_{\substack{a_t^m\in A^m\\ s_t^m,s_{t+1}^m\in S^m}}\left|r_t^m\right|\max_{s_t^m\in S^m}\left\{D_{TV}\left(\pi_{US}^{*,m}\middle\|\pi_{CS}^{m}\right)\left[s_t^m\right]\right\} \quad (27)$$

Based on the derivations in (25)-(27), the worst-case strategy maximizes the total variation distance. As a result, under MDP-US, the agent can update uncertain policies based on the certain policies under MDP-CS. In Fig. 2, the information exchange is illustrated in the semi-centralized training and decentralized execution architecture, combining centralized training [23] for MDP-CS and decentralized training for MDP-US. For MDP-CS, the input to the actor network is the local CS, as shown in (18). The input to the critic network consists of global CSs and actions. For MDP-US, the input to the actor network is the local state interval, as shown in (21), which includes certain observations along with the maximum and minimum values of uncertain observations. The critic network takes the local state interval and policy as inputs.

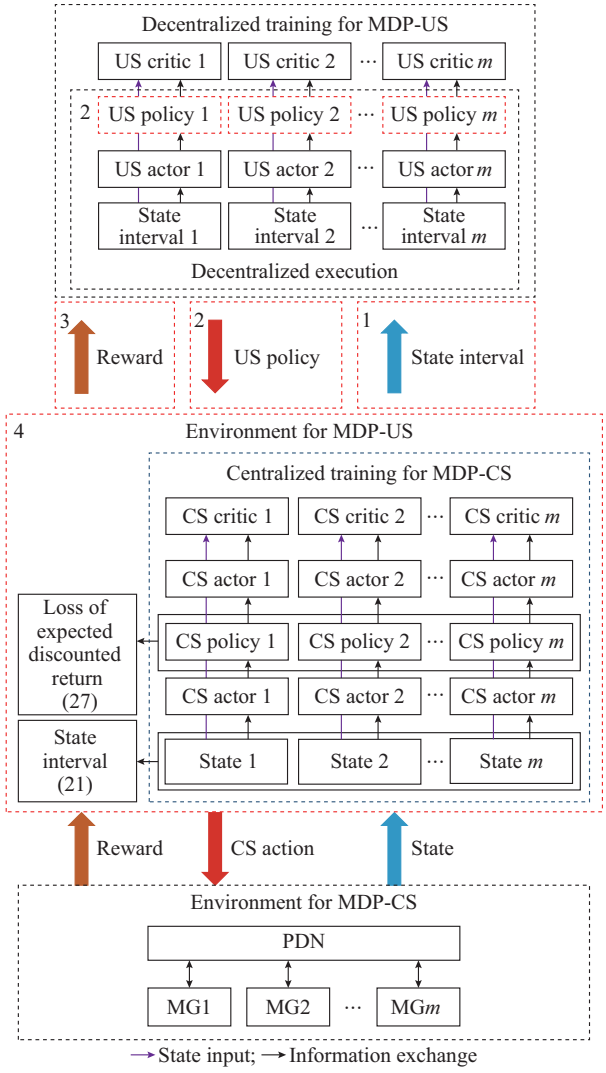

Fig. 2. Training and execution architecture of proposed approach.

In Fig. 2, MDP-US is illustrated, which incorporates four key modifications to the current architecture [24]. First, the actor network processes the individual state intervals instead of individual states. Second, the critic network and environment input individual US policies, unlike the previous approach that uses sampled actions from these policies. Third, the output of the critic network is the reward for a single time step rather than the expected discounted return. Fourth, the environment for MDP-US uses a reinforcement learning model instead of a real-world model. In Fig. 2, decentralized training and execution for MDP-US involve no information sharing, while centralized training for MDP-CS allows limited sharing. The proposed approach does not directly employ the data encryption algorithm but still offers multiple privacy protections. Shared states are normalized as per-unit values, and the rated values hide the state details (de-identified).

The unique mapping of uncertain functions allows for partial modification of shared states (dynamic obfuscation). Shared actions are standardized by power ratios, with consistent value ranges masking action types (anonymization), and dynamic boundaries safeguarding action details (differential privacy).

C. Algorithm Implementation

Due to the high training stability, multi-agent proximal policy optimization (MAPPO) algorithm is selected as the reference model for improvement, using the traditional MAPPO algorithm for centralized training [24]. Continuous action spaces in this algorithm are typically represented by Gaussian distributions. However, the total variation distance for multivariate Gaussian distributions is computationally complex and lacks non-negativity and symmetry. Consequently, Kullback-Leibler (KL) divergence [25] is preferred. We employ the Bretagnolle-Huber inequality [26] to delineate the relationship between KL divergence and total variation distance.

$$D_{TV}\left(\pi_{US}^{*,m}\middle\|\pi_{CS}^{m}\right)\left[s_t^m\right]\le 1-\frac{1}{2}\mathrm{e}^{-D_{KL}\left(\pi_{US}^{*,m}\middle\|\pi_{CS}^{m}\right)\left[s_t^m\right]} \quad (28)$$

The Gaussian policy distribution for MDP-CS is denoted as $\pi_{CS}^{m}\left(a_t^m\middle|s_t^m\right)\mathcal{N}\left(\mu_{CS,t}^{m},\boldsymbol{\Sigma}_{CS,t}^{m}\right)$, and for MDP-US, it is denoted as $\pi_{US}^{*,m}\left(a_t^m\middle|s_t^m\right)\mathcal{N}\left(\mu_{US,t}^{*,m},\boldsymbol{\Sigma}_{US,t}^{*,m}\right)$. To simplify computations, we adopt a fixed covariance matrix $\boldsymbol{\Sigma}_{CS,t}^{m}=\boldsymbol{\Sigma}_{US,t}^{*,m}$. The multivariate KL divergence is illustrated in [25].

Based on the derivations in (28) and [25], the worst-case strategy maximizes KL divergence. This change eliminates the need for computations in (24)-(28) and the buffering of $\max\left|r_t^m\right|$ in (27), thereby significantly enhancing computational efficiency.

The modified MAPPO algorithm, i.e., the proposed algorithm, addresses the worst-case scenarios through a min-max process. Initially, centralized training for MDP-CS identifies policies that maximize the expected discounted return for individual states. Then, decentralized training for MDP-US searches for individual states that minimize the expected discounted return, maximizing the KL divergence within US intervals. The pseudo-code of the proposed algorithm is given in

Supplementary Material A Algorithm SA1. The parameters of the proposed algorithm for MDP-CS remain fixed. Only the parameters of the actor network in the proposed algorithm are initialized with pre-trained weights based on the input state interval, generating MAPPO policies. Other network parameters are randomly initialized. This pre-training ensures that the decoder keeps the worst-case states within the uncertain interval during early training, accelerating the search for the worst-case scenario through non-zero actor gradients. Three key modifications to conventional MAPPO algorithm are highlighted within the blue font. First, the actor network utilizes an autoencoder, with the decoder reversely computing from policy to state, ensuring the maximization of KL divergence under the unknown worst-case scenarios. Second, the actor network initializes parameters based on the input state interval and generates MAPPO policies. This setup accelerates the search for the worst-case scenario through non-zero actor gradients. Third, the gradient updates of the actor and critic network are determined by whether the decoder outputs belong to US intervals.

IV. Case Studies

This section summarizes the findings using four MGs and the IEEE 33-bus system as a PDN [27], as shown in Fig. 3. The dataset includes daily system demand, as well as photovoltaic and wind power generation data spanning six months. Subsequently, we divide it into two parts: the data for the first five months serve as training data, while the data for the last month is reserved for testing the proposed approach. Each training episode represents one day, consisting of 288 steps. The US data are obtained by randomly sampling within intervals formed by CS data and uncertain functions. Detailed data of the test system are available online [1]. Our model is implemented in Python 3.7.16 and executed on a personal computer with an Intel Core i9 processor (6.0 GHz) and 64 GB RAM, using packages Gurobi and PyTorch.

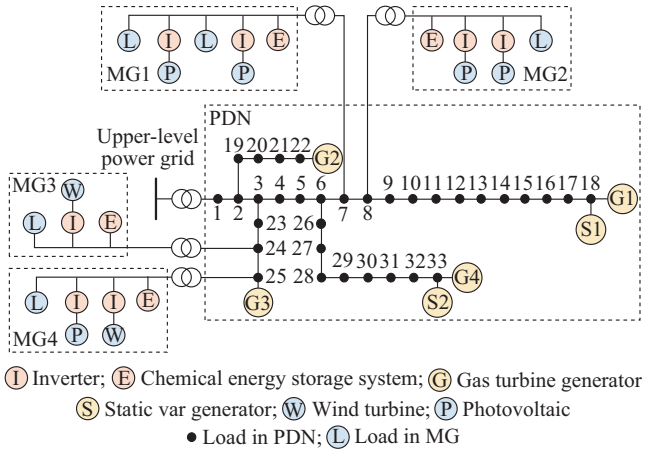

Fig. 3. Topologies of power system (example).

A. Economic Performance of Proposed Algorithm

We compared the proposed algorithm with four benchmarks, which include one non-robust algorithm and three robust algorithms, to evaluate its performance. The details of each algorithm are as follows.

1) CS-No Soft: training with CS data without soft penalties (non-robust).

2) CS-SOC Soft: training with CS data including a 20% soft penalty [14] on the SOC of energy storage systems in MG (robust).

3) CS-Action Soft: training with CS data including a 20% soft penalty [14] on the charging and discharging power of energy storage systems in MG (robust).

4) US-No Soft: training with US data without soft penalties (robust).

Figure 4 presents the episodic average reward of MGs for five examined algorithms. After 153 training rounds, the state and environment in the US-No Soft become non-stationary due to the continuous fluctuations of USs, leading to unstable and non-convergent learning behaviors. In contrast, convergence is consistently achieved under CSs. The uncertainties in MARDRL approach add extra training burden, which causes the proposed algorithm to converge or stabilize more slowly than the comparison algorithm under CSs. Additionally, the proposed algorithm achieves the maximum episodic average reward in each MG, significantly outperforming the benchmarks. Specifically, it demonstrates a relative average growth of 33.41%, 16.86%, and 20.64% compared with CS-No Soft, CS-SOC Soft, and CS-Action Soft, respectively.

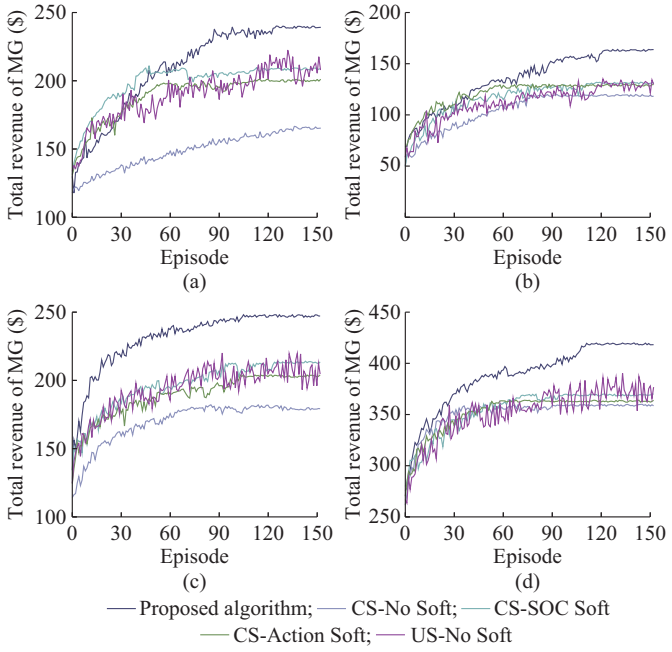

Fig. 4. Episodic average reward of MGs for five examined algorithms. (a) MG1. (b) MG2. (c) MG3. (d) MG4.

Considering the instability and non-convergence in the US-No Soft, the test results obtained from the four examined algorithms in MG1 are summarized in Table I. The proposed algorithm achieves the highest total revenue for MG. The CS-No Soft, which excludes operation redundancy, tends to reduce demand power and increase supply power during

transactions with the PDN. Consequently, CS-No Soft achieves the lowest transaction costs ($53.09) and the highest transaction revenues ($253.98). However, ignoring system uncertainties diminishes the stability of transaction power in the 5-min time intervals, resulting in the highest transaction penalties ($57.35) and the lowest transaction rewards ($37.42). In contrast, the proposed algorithm dynamically adjusts redundancy, unlike the fixed redundancies of the CS-SOC Soft and the CS-Action Soft. Therefore, the proposed algorithm achieves the lower transaction costs and higher transaction revenues compared with both. Moreover, the proposed algorithm designed for the worst-case scenarios achieves the lowest transaction penalties (0) and highest transaction rewards ($128.68). These results verify the performance of the proposed algorithm in the total revenue of MG.

TABLE I
AVERAGE DAILY VALUES OF TRANSACTION COST, TRANSACTION REVENUE, TRANSACTION PENALTY, TRANSACTION REWARD, OTHER COST, AND TOTAL REVENUE OF MG OVER 31 TEST DAYS FOR FOUR EXAMINED ALGORITHMS IN MG1

Algorithm	Average daily value ($)					Total revenue ($)
	Transaction cost	Transaction revenue	Transaction penalty	Transaction reward	Other cost	
Proposed algorithm	67.43	195.84	0	128.68	16.80	240.29
CS-No Soft	53.09	253.98	57.35	37.42	16.98	163.98
CS-SOC Soft	57.54	217.22	10.37	72.03	16.20	205.14
CS-Action Soft	56.77	215.02	10.85	66.07	16.41	197.06

To further explore the performance of the proposed algorithm in the coordination of PDNs and MGs, Table II summarizes the test results obtained from four examined algorithms in PDN. The proposed algorithm achieves the highest total revenue of MGs by adaptive redundancy, thereby maximizing transaction costs of PDN. Moreover, through adaptive redundancy, the proposed algorithm ensures supply stability, resulting in minimal load curtailment costs. Since the load curtailment price is higher than the reward-penalty coefficient, the proposed algorithm not only reduces the operation costs of the PDN but also increases the revenue of the MG. By considering future periods with the discount factor of the agent, it adjusts MG supply to mitigate PDN demand peaks, thereby reducing the generation cost of gas turbine generators and purchased costs from the upper-level power grid. Overall, the proposed algorithm achieves the lowest total operation cost in PDN, reducing it by 2.89%, 1.59%, and 1.89%, compared with the CS-No Soft, the CS-SOC Soft, and the CS-Action Soft, respectively, which demonstrates significant outperformance over the three examined algorithms.

TABLE II
AVERAGE DAILY VALUES OF GENERATION COST, PURCHASED COST, LOAD CURTAILMENT COST, TRANSACTION COST, AND TOTAL OPERATION COST OVER 31 TEST DAYS FOR FOUR EXAMINED ALGORITHMS IN PDN

Algorithm	Average daily value ($)				Total operation cost ($)
	Generation cost	Purchased cost	Load curtailment cost	Transaction cost	
Proposed algorithm	12384.82	2909.36	283.81	1132.10	16710.09
CS-No Soft	10839.35	2785.71	2827.39	754.23	17206.68
CS-SOC Soft	11838.74	2873.71	1281.94	985.08	16979.47
CS-Action Soft	11394.73	2838.19	1834.72	963.93	17031.57

B. Adaptability Performance of Proposed Algorithm

We evaluated the proposed algorithm in two scenarios to test its adaptability. The details are as follows:

1) Non-deployed agent for MDP-CS: no centralized or decentralized training has been completed.

2) Deployed agent for MDP-CS: centralized training is completed, but no decentralized training has been completed.

Figure 5 presents the episodic average reward of MGs for two examined scenarios. Table III summarizes the average daily values of total revenue of MGs over 31 test days for two examined scenarios in each MG. The deployed agent for MDP-CS reduces the number of training episodes required to achieve 99% of the maximum reward by an average of 84.41% compared with the non-deployed agent for MDP-CS. The mean absolute relative error percentage between the two agents in the total revenue of the MG is 0.22%, which is below 0.5%. Since the deployed agent has already completed centralized training, with known actor and critic parameters for MDP-CS, it only requires decentralized training for MDP-US. This results in faster convergence and stability compared with the non-deployed agent. These results highlight the adaptability of the proposed algorithm, allowing the transition from MDP-CS to MDP-US with decentralized updates, avoiding re-centralized training and speeding up the training process.

C. Comparison of TOU and RTP Models with Proposed Pricing Model

We compare the proposed pricing model with two benchmarks to evaluate the performance.

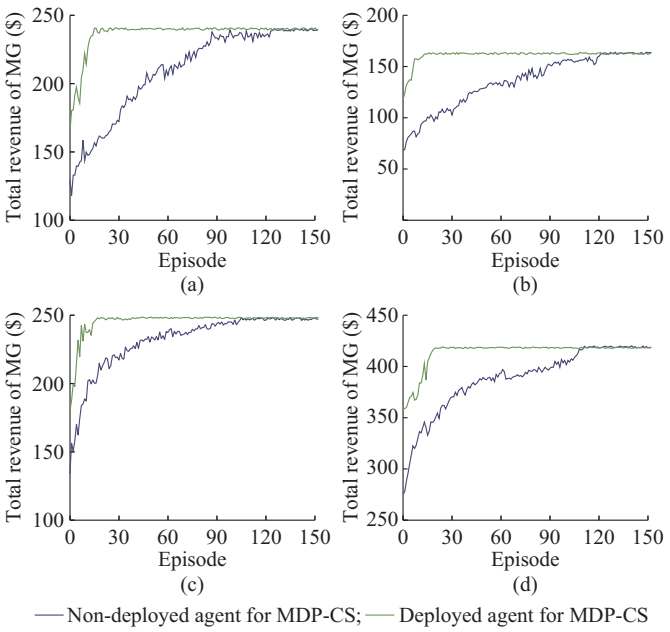

Fig. 5. Episodic average reward of MGs for two examined scenarios. (a) MG1. (b) MG2. (c) MG3. (d) MG4.

TABLE III
AVERAGE DAILY VALUES OF TOTAL REVENUE OF MGS OVER 31 TEST DAYS FOR TWO EXAMINED SCENARIOS IN EACH MG

Scenario	Average daily value of total revenue of MGs ($)			
	MG1	MG2	MG3	MG4
Non-deployed agent for MDP-CS	240.06	163.65	248.00	420.02
Deployed agent for MDP-CS	240.69	163.99	248.45	419.01

1) TOU model: 0.080 $/kWh (00:00-08:00); 0.117 $/kWh (08:00-16:00); and 0.160 $/kWh (16:00-24:00).

2) RTP model: in each reinforcement learning episode, the MG optimizes its net revenue to send transaction power from MG to PDN. Then, the PDN optimizes operation costs to send DLMP from PDN to MG. This iterative process continues until the error of DLMPs between adjacent cycles is less than 0.001 $/kWh.

Figure 6 summarizes the hourly electricity price of MGs for the three examined models on the 12th test day by (1). Table IV summarizes daily net transaction revenue and average 5-min interval computation time of MGs for three examined models on the 12th test day. Due to the high similarity in net demands between similar and target hours, the mean percentage errors between the proposed pricing model and RTP model in hourly electricity prices are 1.18% in MG1, 1.21% in MG2, 1.36% in MG3, and 1.32% in MG4, all of which are below 2%. Since RTP model fully reflects real-time demand in the PDN, it achieves the highest daily net transaction revenue, as shown in Table IV. Compared with RTP model, the proposed pricing model and TOU model reduce daily net transaction revenue by 1.25% and 29.96%, respectively, as shown in Table IV. Since TOU model only requires one PDN optimization and one MG optimization per step, it achieves the shortest computation time, as shown in Table IV. The proposed pricing model and RTP model increase average computation time by 0.19% and 7837.14%, respectively, compared with TOU model, as shown in Table IV. We conclude from these results that the proposed pricing model properly balances economic efficiency and computation time.

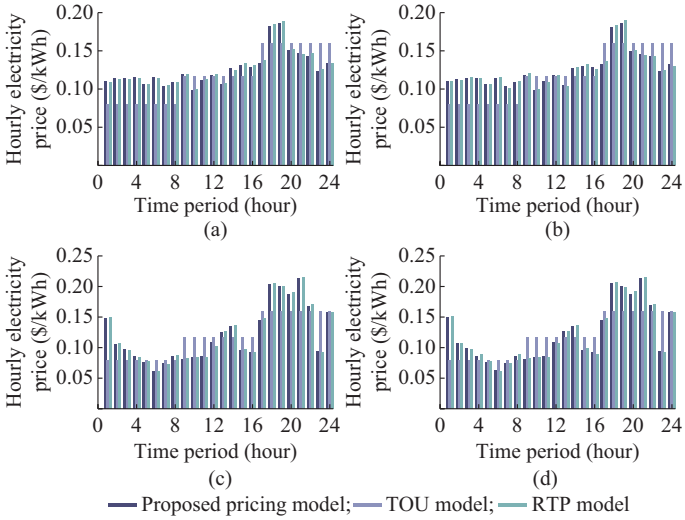

Fig. 6. Hourly electricity price of MGs for three examined models on the 12th test day. (a) MG1. (b) MG2. (c) MG3. (d) MG4.

TABLE IV
DAILY NET TRANSACTION REVENUE AND AVERAGE 5-MIN INTERVAL COMPUTATION TIME OF MGS FOR THREE EXAMINED MODELS ON THE 12TH TEST DAY

Model	MG1		MG2		MG3		MG4	
	Transaction revenue ($)	Time (ms)	Transaction revenue ($)	Time (ms)	Transaction revenue ($)	Time (ms)	Transaction revenue ($)	Time (ms)
Proposed pricing model	129.34	404.90	87.78	404.66	131.58	407.05	220.60	410.17
TOU model	94.45	404.89	63.80	404.73	92.44	405.28	149.36	408.84
RTP model	130.86	31732.87	88.41	31837.92	133.23	32437.34	224.81	32874.92

D. Impact of MG Demand Level

To further explore the interactions between PDN and MGs using the proposed pricing model, we consider the case previously described as the base case and four additional cases in which the MG demands are increased by 10% with respect to those of the base case. Figure 7 illustrates how MG demand affects PDN prices during operations. Each subplot illustrates that higher MG demand results in increased hourly electricity prices. Among the scenarios, the MG1+10% case shows a larger price increase than the MG2+10% case,

while the MG3+10% case shows a larger price increase than the MG4+10% case. This is because higher MG net demands compel the PDN to rely on more expensive energy sources and exacerbate network congestion. Furthermore, the hourly electricity prices of MG1 and MG4 are higher than those of MG2 and MG3, respectively. This is because MG1 and MG4 are closer to high-cost distributed power sources.

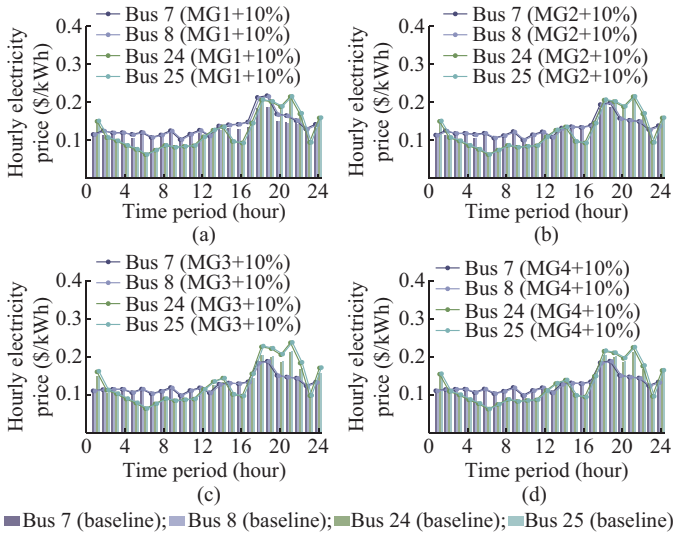

Fig. 7. Hourly electricity price of MGs under baseline and higher power. (a) MG1. (b) MG2. (c) MG3. (d) MG4.

E. Comparison of Model Optimization with Proposed Algorithm

We compare the proposed algorithm with four benchmarks to evaluate the uncertainty part of this paper.

1) DO [28]: deterministic optimization for a single scenario within certain parameters.

2) SO [29]: stochastic optimization for expected values within full probability distribution.

3) RO [30]: robust optimization for the worst-case scenario within the uncertain set.

4) DRO [31]: distributionally robust optimization for the worst-case scenario within the ambiguity set of distributions.

The test results of the proposed algorithm and four benchmarks in MG1 are summarized in Table V. DO ignores uncertainty, achieving the highest net transaction revenue but the lowest reward due to penalties, resulting in the lowest total revenue. DRO partially mitigates penalties by optimizing over the ambiguity set of distributions, but residual risk from its conservative scenario selection limits total revenue compared with the robust strategy. The proposed algorithm and RO both eliminate penalties by accounting for the worst-case scenarios. However, the proposed algorithm enhances robustness (via scaling in (27) and (28)), sacrificing net revenue but maximizing rewards to achieve the highest total revenue. For execution tractability, the proposed algorithm shifts computational load to offline training. During the online operation, it requires only neural network inference, reducing the average 5-min interval computation time by 99.999%, 99.997%, and 99.995%, compared with SO, RO, and DRO, respectively. For training tractability, the proposed algorithm shifts uncertainty part to the reinforcement learning formulation, thereby replacing RO with DO in training environment. This reduces the training time by 98.77% compared with RO.

TABLE V
AVERAGE DAILY VALUES OF TRANSACTION COST, TRANSACTION REVENUE, TRANSACTION PENALTY, TRANSACTION REWARD, OTHER COST, TOTAL REVENUE OF MG, AND AVERAGE 5-MIN INTERVAL COMPUTATION TIME OF MG1 FOR PROPOSED ALGORITHM AND FOUR EXAMINED BENCHMARKS ON THE 12TH TEST DAY

Item	Transaction cost ($)	Transaction revenue ($)	Transaction penalty ($)	Transaction reward ($)	Other cost ($)	Total revenue ($)	Time (ms)
Proposed algorithm	62.06	194.73	0	128.35	17.29	243.73	0.98
DO	52.15	249.06	62.92	39.29	17.98	155.30	404.91
SO	54.64	209.38	37.53	71.04	17.49	170.76	81191.96
RO	61.51	195.42	0	126.76	17.25	243.42	32805.59
DRO	49.54	202.94	10.29	91.40	17.36	217.14	18018.08

F. Scalability Performance of Proposed Algorithm

This subsection further demonstrates the scalability of the proposed algorithm using the modified IEEE 123-bus system and PNNL 329-bus taxonomy feeder [32]. New MGs are added to the existing four MGs from the IEEE 33-bus system, as shown in Fig. 8. Their points of common coupling can be found in Table VI. Table VII presents the average 5-min interval computation time over 31 test days for examined systems and feeder, i.e., three PDNs, in each MG. The average 5-min interval computation time of MGs over 31 test days is 1.01 ms, 1.02 ms, and 1.04 ms, for the three PDNs, respectively. During operation, MGs rely solely on local states and avoid coordinated global computations with PDNs. Therefore, the average 5-min interval computation time of MGs depends on the scale of the MGs, not that of the PDNs, showing the scalability of the proposed algorithm within the PDNs. Additionally, the calculation times for modified IEEE 123-bus system and PNNL 329-bus taxonomy feeder increase by 0.69% and 3.31%, respectively, compared with that of the IEEE 33-bus system. This is because, in the actor networks of reinforcement learning, matrix operations during forward propagation are mainly influenced by the size of the intermediate layers, further demonstrating the scalability of the proposed algorithm within the MGs.

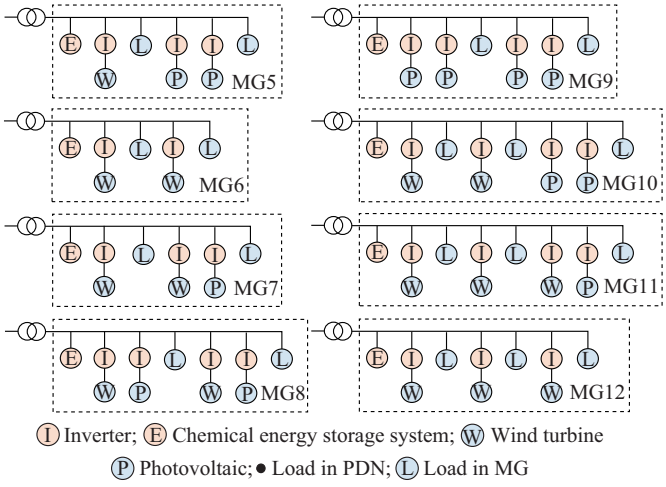

Fig. 8. Topologies of additional MGs.

TABLE VI
POINTS OF COMMON COUPLING FOR MODIFIED IEEE 123-BUS SYSTEM AND PNNL 329-BUS TAXONOMY FEEDER

MG	Point of common coupling	
	Modified IEEE 123-bus system	PNNL 329-bus taxonomy feeder
1	19	15
2	22	33
3	36	103
4	47	121
5	64	132
6	76	133
7	87	143
8	106	158
9		193
10		201
11		204
12		205

TABLE VII
AVERAGE 5-MIN INTERVAL COMPUTATION TIME OVER 31 TEST DAYS FOR THREE EXAMINED PDNS IN EACH MG

MG	Average 5-min interval computation time (ms)		
	IEEE 33-bus system	Modified IEEE 123-bus system	PNNL 329-bus taxonomy feeder
1	0.98	1.05	0.95
2	0.99	0.95	0.98
3	0.98	1.07	0.97
4	1.10	0.99	1.02
5		0.94	0.99
6		1.12	1.10
7		1.01	1.11
8		1.02	1.11
9			1.01
10			1.09
11			1.11
12			1.10

V. CONCLUSION

This paper presents an MARDRL approach for coordination of PDNs and MGs with limited information exchange. Comparative evaluation demonstrates that the proposed algorithm improves MG operation profits and reduces the operation costs of PDNs, outperforming the existing reinforcement learning approaches. Additionally, the proposed algorithm supports adaptive training deployment in diverse real-world scenarios via serial computation. The proposed pricing model effectively balances price responsiveness and computational complexity by embedding iterative power-pricing loops between MGs and PDNs into the reinforcement learning framework.

The findings underscore the importance of implementing robust strategies for coordination of PDNs and MGs to enhance system economics. The proposed approach provides a decision-making agent for each MG, enabling optimal scheduling under uncertainty. Furthermore, the coordination model, which operates under limited information exchange (MGs send power and PDNs send electricity prices), could facilitate closer coordination of PDNs and MGs, particularly in the scenarios with higher renewable energy penetration. Lastly, integrating the pricing model into the reinforcement learning framework optimizes dispatch decisions, highlighting the critical role of pricing strategies in ensuring stability and efficiency in real-time coordination of PDNs and MGs.

REFERENCES

[1] A. Ghasemi, A. Shojaeighadikolaei, and M. Hashemi, "Combating uncertainties in smart grid decision networks: multiagent reinforcement learning with imperfect state information," *IEEE Internet of Things Journal*, vol. 11, no. 13, pp. 23985-23997, Jul. 2024.

[2] Z. Guo, W. Wei, L. Chen *et al*., "Operation of distribution network considering compressed air energy storage unit and its reactive power support capability," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 2954-2965, Jul. 2020.

[3] A. Alamolhoda, R. Ebrahimi, M. S. Moghaddam *et al*., "Integrated load and energy management in active distribution networks featuring prosumers based on PV and energy storage systems," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 6, pp. 1869-1879, Nov. 2024.

[4] G. Song, C. Ma, H. Ji *et al*., "Bi-level supply restoration method for active distribution networks considering multi-resource coordination," *Journal of Modern Power Systems and Clean Energy*, vol. 13, no. 3, pp. 967-979, May 2025.

[5] T. C. Chiu, Y. Y. Shih, A. C. Pang *et al*., "Optimized day-ahead pricing with renewable energy demand-side management for smart grids," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 374-383, Apr. 2017.

[6] M. Jalali, K. Zare, and H. Seyedi, "Strategic decision-making of distribution network operator with multi-microgrids considering demand response program," *Energy*, vol. 141, pp. 1059-1071, Dec. 2017.

[7] N. Nikmehr and S. N. Ravadanegh, "Optimal power dispatch of multi-microgrids at future smart distribution grids," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1648-1657, Jul. 2015.

[8] Z. Yi, Y. Xu, J. Zhou *et al*., "Bi-level programming for optimal operation of an active distribution network with multiple virtual power plants," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2855-2869, Oct. 2020.

[9] R. Zhang, X. Li, L. Fu *et al*., "Network-aware energy management for microgrids in distribution market: a leader-followers approach," *Applied Energy*, vol. 332, p. 120522, Feb. 2023.

[10] O. M. Adeyanju, P. Siano, and L. N. Canha, "Dedicated microgrid planning and operation approach for distribution network support with pumped-hydro storage," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 7, pp. 8229-8241, Jul. 2023.

[11] J. Zhu, W. Hu, X. Xu *et al*., "Optimal scheduling of a wind energy dominated distribution network via a deep reinforcement learning ap-

proach," *Renewable Energy*, vol. 201, pp. 792-801, Dec. 2022.
[12] X. Lu, X. Xiao, L. Xiao *et al*., "Reinforcement learning-based microgrid energy trading with a reduced power plant schedule," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10728-10737, Dec. 2019.
[13] H. Liu and W. Wu, "Federated reinforcement learning for decentralized voltage control in distribution networks," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3840-3843, Sept. 2022.
[14] S. Li, D. Cao, W. Hu *et al*., "Multi-energy management of interconnected multi-microgrid system using multi-agent deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 11, no. 5, pp. 1606-1617, Sept. 2023.
[15] D. Domínguez-Barbero, J. García-González, M. Á. Sanz-Bobi, "Energy management of a microgrid considering nonlinear losses in batteries through deep reinforcement learning," *Applied Energy*, vol. 368, p. 123435, Aug. 2024.
[16] H. H. Goh, Y. Huang, C. S. Lim *et al*., "An assessment of multistage reward function design for deep reinforcement learning-based microgrid energy management," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4300-4311, Nov. 2022.
[17] P. Chen, M. Liu, C. Chen *et al*., "A battery management strategy in microgrid for personalized customer requirements," *Energy*, vol. 189, p. 116245, Dec. 2019.
[18] Y. Huang, G. Li, C. Chen *et al*., "Resilient distribution networks by microgrid formation using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4918-4930, Nov. 2022.
[19] Y. Chen, P. B. Luh, C. Guan *et al*., "Short-term load forecasting: similar day-based wavelet neural networks," *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 322-330, Feb. 2010.
[20] R. Ghorani, M. Fotuhi-Firuzabad, and M. Moeini-Aghtaie, "Main challenges of implementing penalty mechanisms in transactive electricity markets," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3954-3956, Sept. 2019.
[21] S. Chen, H. Cheng, S. Lv *et al*., "Learning-aided collaborative optimization of power, hydrogen, and transportation networks," *Journal of Modern Power Systems and Clean Energy*, vol. 13, no. 2, pp. 475-487, Mar. 2025.
[22] X. Serrano-Guerrero, M. Briceño-León, J.-M. Clairand *et al*., "A new interval prediction methodology for short-term electric load forecasting based on pattern recognition," *Applied Energy*, vol. 297, p. 117173, Sept. 2021.
[23] J. Zhang, L. Che, and M. Shahidehpour, "Distributed training and distributed execution based Stackelberg multi-agent reinforcement learning for EV charging scheduling," *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4976-4979, Nov. 2023.
[24] Y. Guan, S. Zou, H. Peng *et al*., "Cooperative UAV trajectory design for disaster area emergency communications: a multiagent PPO method," *IEEE Internet of Things Journal*, vol. 11, no. 5, pp. 8848-8859, Mar. 2024.
[25] S. Ji, Z. Zhang, S. Ying *et al*., "Kullback-Leibler divergence metric learning," *IEEE Transactions on Cybernetics*, vol. 52, no. 4, pp. 2047-2058, Apr. 2022.
[26] A. B. Tsybakov, "Inequalities for distances," in *Introduction to Nonparametric Estimation*. New York: Springer, pp. 86-89.
[27] S. H. Dolatabadi, M. Ghorbanian, P. Siano *et al*., "An enhanced IEEE 33 bus benchmark test system for distribution system studies," *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 2565-2572, May 2021.
[28] Z. Li, Y. Xu, S. Fang *et al*., "Multiobjective coordinated energy dispatch and voyage scheduling for a multienergy ship microgrid," *IEEE Transactions on Industry Applications*, vol. 56, no. 2, pp. 989-999, Apr. 2020.
[29] L. Wu, M. Shahidehpour, and T. Li, "Stochastic security-constrained unit commitment," *IEEE Transactions on Power Systems*, vol. 22, no. 2, pp. 800-811, May 2007.
[30] W. Wei, F. Liu, S. Mei *et al*., "Robust energy and reserve dispatch under variable renewable generation," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 369-380, Jan. 2015.
[31] C. Zhao and Y. Guan, "Data-driven stochastic unit commitment for integrating wind generation," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2587-2596, Jul. 2016.
[32] R. R. Jha, A. Dubey, C. C. Liu *et al*., "Bi-level volt-var optimization to coordinate smart inverters with voltage control devices," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 1801-1813, May 2019.

Jiahui Jin received the B.S. degree from the School of Electrical and Engineering, Southwest Jiaotong University, Sichuan, China, in 2019, and the M.S. degree from the School of Electrical and Engineering, Shanghai DianJi University, Shanghai, China, in 2023. He is currently pursuing the Ph.D. degree from the School of Electrical and Power Engineering, Hohai University, Nanjing, China. His research interests include coordinated operation of smart power distribution networks and microgrids.

Guoqiang Sun received the B.S., M.S., and Ph.D. degrees in electrical engineering from Hohai University, Nanjing, China, in 2001, 2005, and 2010, respectively. He was a Visiting Scholar with North Carolina State University, Raleigh, USA, from 2015 to 2016. He is currently a Professor with the College of Energy and Electrical Engineering, Hohai University. His research interests include power system analysis and economic dispatch and optimal control of integrated energy systems.

Sheng Chen received the B.S. and Ph.D. degrees from the College of Energy and Electrical Engineering, Hohai University, Nanjing, China, in 2014 and 2019, respectively. From January 2018 to January 2019, he was a Visiting Scholar at The Ohio State University, Columbus, USA. He is currently a Professor in the College of Energy and Electrical Engineering, Hohai University. He serves as an Associate Editor of the Journal of Modern Power Systems and Clean Energy. His research interests include integrated energy system and electricity market.

Yaping Li received the B.S. degree from Sichuan University, Chengdu, China, in 2003, the M.S. degree from Nanjing Automation Research Institute, Nanjing, China, in 2006, and the Ph.D. degree from Hohai University, Nanjing, China, in 2017, all in power system and its automation. She is a Senior Engineer with China Electric Power Research Institute, Beijing, China. Her research interests include demand response, power system simulation, and artificial intelligence in power network dispatch.

Yingqi Liao received the B.S. degree from Northeast Electric Power University, Jilin, China, in 2003, and the M.S. degree from Hohai University, Nanjing, China, in 2014, all in electrical engineering and its automation. He is currently working for Nanjing Power Supply Branch, State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China, as a Senior Engineer. His research interests include power system optimization and distribution power network dispatch.

Wenbo Mao received the M.S. degree in electrical engineering from Wuhan University, Wuhan, China, in 2012. He is currently an Engineer with China Electric Power Research Institute, Nanjing, China. His research interests include power system optimization and distribution power network dispatch.

Lu Shen received the B.S. degree from Hohai University, Nanjing, China, in 2017, and the Ph.D. degree from Southeast University, Nanjing, China, in 2022, all in electrical engineering. She is currently working for Nanjing Power Supply Branch, State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China, as an Engineer. Her research interests include power system optimization and distribution power network dispatch.