Energy Management of Networked Microgrids With Real-Time Pricing by Reinforcement Learning

Gaochen Cui, *Student Member, IEEE*, Qing-Shan Jia, *Senior Member, IEEE*, and Xiaohong Guan, *Life Fellow, IEEE*

***Abstract*—Coordinating the microgrids (MGs) in the distribution network is a critical task for the distribution system operator (DSO), which could be achieved by setting prices as incentive signals. The high uncertainty of loads and renewable resources motivates the DSO to adopt real-time prices. The MGs require reference price sequences for a long time horizon in advance to make generation plans. However, due to privacy concerns in practice, the MGs may not provide adequate information for the DSO to build a closed-form model. This causes challenges to the implementation of the conventional model-based methods. In this paper, the framework of the coordination system through real-time prices is proposed. In this bi-level framework, the DSO sets real-time reference price sequences as the incentive signals, based on which the MGs make the generation and charging plan. The model-free reinforcement learning (RL) is applied to optimize the pricing policy when the response behavior of the MGs is unknown to the DSO. To deal with the large action space of this problem, the reference policy is incorporated into the RL algorithm for efficiency improvement. The numerical result shows that the minimized cost obtained by the developed model-free RL algorithm is close to the model-based method while the private information is preserved.**

***Index Terms*—Reinforcement learning, microgrids, distribution network, energy management.**

I. Introduction

In a SMART distribution network, the energy management system is faced with a complex network with distributed energy resources and energy storage (ES) units [1]. MGs consist of loads, generators, ES units, and renewable energy sources such as photovoltaic (PV) units. An MG could be both an electricity consumer and prosumer [2]. The DSO normally has the decision authority to solve the energy management problem for the entire system. Coordinating the resources in these MGs is a critical task for the DSO to achieve high energy efficiency. However, the DSO and MGs are owned by different organizations, therefore the DSO usually has no authority to directly command the MGs. An effective way is to motivate the MGs through pricing signals [3], while the real-time prices could be applied to handle the high uncertainty of loads and renewable energy sources [4]. Moreover, the MGs usually require a reference price sequence for a long time horizon (e.g., 24 hours) divided into multiple time slots (e.g., 5 minutes) in advance to make the generation plan [5]. Therefore in this paper, the policy to set the real-time reference price sequence is optimized.

The DSO pricing task could be modeled as a bi-level optimization problem [3]. At the upper level, the DSO decides the reference price sequence. At the lower level, the MG makes its generation plan for the received reference prices. This bi-level optimization problem could be transformed into a mixed-integer linear programming (MILP) problem using the Karush–Kuhn–Tucker (KKT) conditions and solved by commercial solvers [6]. However, in practice, the MGs may not provide their private information about the response behavior toward the prices. In this circumstance, the model-based methods encounter challenges since the closed-form model is hard to build.

The above problem could be transformed into a Markov decision process (MDP) problem and solved by the model-free RL [7] that optimizes the DSO pricing policy by learning from experience. In recent years, RL has been successfully applied to power systems [8]. Online pricing algorithms of demand response and for MGs based on RL are developed in [4], [5]. However, in this paper, the DSO agent decides the price sequence for a much longer time horizon at each time instant. This results in a much larger action space and causes challenges to the training process.

In this paper, the following contributions are made for addressing the above problems. (i) The price-based real-time coordination framework is established and formulated as a bi-level optimization problem. In this framework, the DSO transmits prices and collects metering data only once during each time slot, and thus is more tolerable for short-term communication failures. Moreover, the MGs would receive a reference price sequence to consider the long-term profit when making plans for the resources. (ii) This problem is transformed into an MDP that incorporates the responding behavior of the MGs, which could be solved by the model-free RL method when the behavior is unknown to the DSO.

Manuscript received 14 November 2022; revised 21 March 2023; accepted 14 May 2023. Date of publication 1 June 2023; date of current version 26 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFA1004600; in part by the National Natural Science Foundation of China under Grant 62125304, Grant 62192751, and Grant 62073182; and in part by the 111 International Collaboration Project under Grant BP2018006. Paper no. TSG-01699-2022. *(Corresponding author: Qing-Shan Jia.)*

Gaochen Cui and Qing-Shan Jia are with the CFINS, Department of Automation, BNRist, Tsinghua University, Beijing 100084, China (e-mail: cgc19@mails.tsinghua.edu.cn; jiaqs@tsinghua.edu.cn).

Xiaohong Guan is with the CFINS, Department of Automation, BNRist, Tsinghua University, Beijing 100084, China, and also with the MOEKLINNS Laboratory, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China (e-mail: xhguan@xjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2023.3281935.

Digital Object Identifier 10.1109/TSG.2023.3281935

The reference policy-based RL algorithm is developed that improves the training efficiency, while the state-of-the-art RL methods fail to significantly improve the policy in finite time due to the large action space in practice. (iii) Numerical experiment shows that the developed RL algorithm is independent of the closed-form model at the expense of less than 5% economic cost compared with the conventional model-based method.

II. Literature Review

Coordinating the MGs in a distribution network is a hot topic in the last decade. The smart metering devices enable the operator to apply efficient dispatch with an accurate demand-side model [9]. One approach is to manage the power flow through the distributed scheme [10]. This approach usually heavily relies on the communication network [11], and the issues such as packet loss are compensated in [12]. The small-scale energy resources and load demand at each MG are modeled and solved by the particle swarm method in [13]. A convex multi-objective optimization problem for multi-MG is formulated and solved by an online algorithm in [14]. However, these MGs may belong to different owners in practice, so the DSO couldn't directly control them. In this circumstance, prices could be applied as incentive signals to coordinate the MGs. The pricing task is generally modeled as a bi-level optimization problem [3]. A hierarchical market structure is proposed in [15] to guarantee the goals of both DSO and MGs. A bi-level stochastic model is formulated and transformed into a MILP problem that could be solved by commercial solvers to maximize the DSO profit in [6]. This method is also utilized to coordinate the distributed resources in the virtual power plants [16], [17].

However, the above methods require the response behavior of the MGs to build the closed-form model, which is hard to acquire in practice. The responding behavior of the MGs could be approximated by a neural network [18], while the optimization problem is still hard to solve due to the non-linearity. An alternative method is that the DSO and MGs bargain through sequentially solving the local optimization problem [19], [20], which heavily relies on the communication network. And whether the MGs would participate in this process is questionable.

Real-time electricity pricing models potentially lead to high energy efficiency as the stochasticity increases [21]. The real-time pricing could be applied to cope with the discrepancy in the actual consumer behavior and the load levels from forecast and planning [22], where the model of demand elasticity is critical [23]. It is also applied to energy management for electric vehicles [24].

Model-free RL methods could address this problem by interacting with the system and learning from experience to improve the policy. With deep learning showing strong feature extraction ability in computer vision and natural language processing [25], RL is empowered to deal with complicated tasks from games [26] to robotics [27]. To handle the large state space, parametrized methods are developed to improve the efficiency [7]. Deep neural networks further improve the learning ability to defeat humans in hundreds of Atari games [28] by the strong feature extraction ability. Similarly, the policy is also parametrized to handle large action space and is trained with the actor-critic framework [29].

In recent years, RL has been widely applied to power systems [8]. A consensus transfer Q-learning for decentralized generation command dispatch of automatic generation control is developed in [30]. The multi-agent RL is applied to Volt-VAR control in power distribution networks in [31]. An RL-based online optimal control method is developed for the hybrid ES system in AC–DC MGs in [32]. The problem of setting the tap positions of load tap changers for voltage regulation in power distribution systems is solved by RL in [33]. Efficient economic dispatch for MGs is achieved by a cooperative RL algorithm in [34]. A fully distributed multi-agent RL method for optimal reactive power dispatch is developed in [35]. An RL-based distributed optimal power flow algorithm is developed that reduces the computational complexity of the conventional linear programming approach while addressing the stochastic nature of the energy resources and loads in [36]. A hierarchical MG model considering communication uncertainty is developed and solved using RL in [37].

To solve the pricing problem, the RL algorithm is applied to solve dynamic pricing and energy consumption scheduling problem in [38]. An RL-based method for online pricing of demand response is developed in [4]. An RL-based game-theoretic approach is developed to solve the pricing problem for networked MGs in [39]. However, these methods generate the prices for only a short time. In this paper, we deal with the circumstance where the MGs require a reference price sequence for a much longer time horizon to make the generation and charging plan. In this case, the large action space results in low efficiency in practice. Thus in this paper, a reference policy-based RL algorithm is developed to address this problem.

III. Nomenclature

Parameters

$\mathcal{N}$	The set of all nodes in the distribution network.
$\mathcal{N}_{MG}$	The set of MG nodes.
$\mathcal{E}$	The set of distribution network lines.
$r_{ij} \in \mathbb{R}$	The resistance of line (i, j).
$x_{ij} \in \mathbb{R}$	The reactance of line (i, j).
$i \in \mathbb{N}$	The node index.
$t \in \mathbb{N}$	The time slot index.
$\tau \in \mathbb{N}$	The time slot index in a sequence.
$T \in \mathbb{N}$	The number of time slots in a time horizon.
$\Delta t \in \mathbb{R}$	The time duration of a time slot.
$\lambda_t^{HV} \in \mathbb{R}$	The actual marginal price for buying electricity from the high voltage grid at time t.
$\tilde{\lambda}_t^{HV} \in \mathbb{R}^T$	The predicted marginal price for buying electricity from the high voltage grid at time t for the horizon from $t+1$ to $t+T$.
$\bar{P}_i^G \in \mathbb{R}$	The maximum active power output of the generator in the MG at node i.

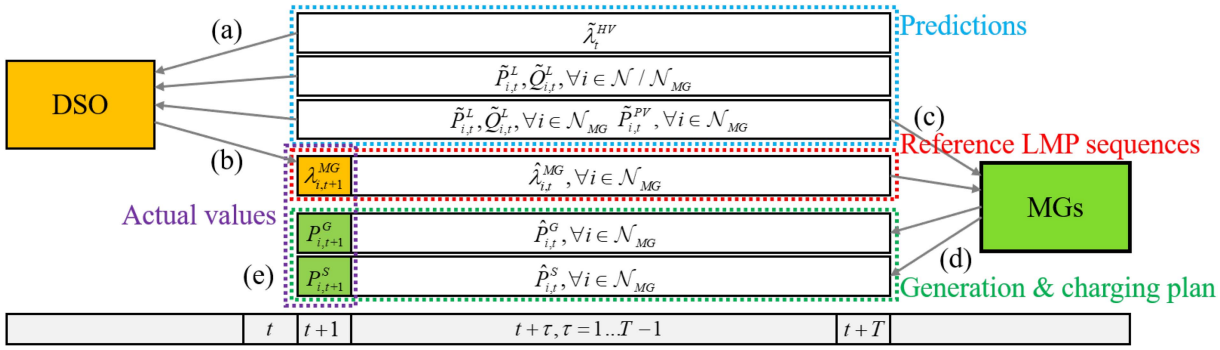

Fig. 1. The bi-level operational structure of the distribution network with multiple MGs.

$\underline{P}_i^G \in \mathbb{R}$ The minimum active power output of the generator in the MG at node i.
$\Delta\bar{P}_i^G \in \mathbb{R}$ The maximum ramping rate of the generator in the MG at node i.
$\bar{P}_i^S \in \mathbb{R}$ The maximum charging power of the ES unit in the MG at node i.
$\underline{P}_i^S \in \mathbb{R}$ The maximum discharging power of the ES unit in the MG at node i.
$\bar{S}_i \in \mathbb{R}$ The maximum energy storage of the ES unit in the MG at node i.
$\underline{S}_i \in \mathbb{R}$ The minimum energy storage of the ES unit in the MG at node i.
$\tilde{P}_{i,t}^{PV} \in \mathbb{R}^T$ The predicted PV active power at node i, time t for the horizon from $t+1$ to $t+T$.
$\tilde{P}_{i,t}^{L} \in \mathbb{R}^T$ The predicted load active power at node i, time t for the horizon from $t+1$ to $t+T$.
$\tilde{Q}_{i,t}^{L} \in \mathbb{R}^T$ The predicted load reactive power at node i, time t for the horizon from $t+1$ to $t+T$.
$P_{i,t}^{PV} \in \mathbb{R}$ The actual PV active power at node i, time t.
$P_{i,t}^{L} \in \mathbb{R}$ The actual active power of the loads at node i, time t.
$Q_{i,t}^{L} \in \mathbb{R}$ The actual reactive power of the loads at node i, time t.
$C_i^G \in \mathbb{R}$ The cost coefficient of the generator connected in the MG at node i.
$\bar{V}_i \in \mathbb{R}$ The maximum voltage magnitude at node i.
$\underline{V}_i \in \mathbb{R}$ The minimum voltage magnitude at node i.
$\Delta\bar{\lambda} \in \mathbb{R}$ The maximum price adjustment between two continuous time slots.
$\bar{\lambda}_i^{MG} \in \mathbb{R}$ The maximum price at node i.
$\underline{\lambda}_i^{MG} \in \mathbb{R}$ The minimum price at node i.

Variables

$P_{i,t} \in \mathbb{R}$ The active power injection at node i, time t.
$\hat{P}_{i,t} \in \mathbb{R}^T$ The planned active power injection at node i, time t for interval from $t+1$ to $t+T$.
$Q_{i,t} \in \mathbb{R}$ The reactive power injection at node i, time t.
$V_{i,t} \in \mathbb{R}$ The voltage magnitude at node i, time t.
$\hat{V}_{i,t} \in \mathbb{R}^T$ The planned voltage magnitude at node i, time t for interval from $t+1$ to $t+T$.
$P_{ij,t} \in \mathbb{R}$ The active power flow from node i to node j through line at time t.
$Q_{ij,t} \in \mathbb{R}$ The reactive power flow from node i to node j through line at time t.
$P_{i,t}^G \in \mathbb{R}$ The active power output of the generator in the MG connected at node i, time t.
$\hat{P}_{i,t}^G \in \mathbb{R}^T$ The planned active power output of the generator in the MG connected at node i, time t for the horizon from $t+1$ to $t+T$.
$P_{i,t}^S \in \mathbb{R}$ The (dis)charging power of the ES unit in the MG connected at node i, time t.
$\hat{P}_{i,t}^S \in \mathbb{R}^T$ The planned (dis)charging power of the ES unit in the MG connected at node i, time t for the horizon from $t+1$ to $t+T$.
$S_{i,t} \in \mathbb{R}$ The energy stored in the MG at node i, time t.
$\hat{S}_{i,t} \in \mathbb{R}^T$ The planned energy stored in the MG at node i, time t for the horizon from $t+1$ to $t+T$.
$\lambda_{i,t}^{MG} \in \mathbb{R}$ The price for the MG at node i, time t.
$\hat{\lambda}_{i,t}^{MG} \in \mathbb{R}^T$ The reference price for the MG at node i, time t for the horizon from $t+1$ to $t+T$.

IV. Problem Formulation

In this section, we first establish the real-time price-based coordination framework. In this bi-level system, the DSO aims at minimizing the cost of supplying electricity, and the MGs aim at maximizing their own profit by making electricity transactions with the DSO. Then, the pricing problem of the DSO is formulated as an optimization problem in closed form.

A. The Overall Structure

The operational framework of the DSO and MGs in the distribution network is shown in Fig. 1. Without losing generality, it is assumed that each MG owns a generator, a PV unit, an ES unit, and inflexible loads.

At each time t, the DSO and MGs consider the horizon of the next T time slots of time duration Δt. (a) The DSO has access to the predicted PV generation $\tilde{P}_{i,t}^{PV} \in \mathbb{R}^T$ at all MGs, loads $\tilde{P}_{i,t}^{L}, \tilde{Q}_{i,t}^{L} \in \mathbb{R}^T$ at all nodes, and the price $\tilde{\lambda}_t^{HV} \in \mathbb{R}^T$ for buying from the high voltage grid for the next T time slots. (b) Based on the prediction, the DSO decides the reference price sequences $\hat{\lambda}_{i,t}^{MG} \in \mathbb{R}^T$ for the next T time slots for each MG. (c) The MGs make local predictions and receive the reference price sequences. (d) Each MG solves its local lower-level problem to plan for its generator output $\hat{P}_{i,t}^G \in \mathbb{R}^T$ and ES (dis)charging power $\hat{P}_{i,t}^S \in \mathbb{R}^T$ for the next T time slots. (e) At time slot $t+1$, the first element of $\hat{\lambda}_{i,t}^{MG}$, i.e., $\hat{\lambda}_{i,t}^{MG}(1)$, is set as

the clearing price $\lambda^{MG}_{i,t+1} = \hat{\lambda}^{MG}_{i,t}(1)$. In our model of the strategy adopted by each MG, similarly, the first elements of $\hat{P}^{G}_{i,t}$ and $\hat{P}^{S}_{i,t}$ are applied, i.e., $P^{G}_{i,t+1} = \hat{P}^{G}_{i,t}(1)$ and $P^{S}_{i,t+1} = \hat{P}^{S}_{i,t}(1)$. Thus, the transacted electricity at node $i \in \mathcal{N}_{MG}$ during time slot $t+1$ is

$$P_{i,t+1}\Delta t = \left(P^{G}_{i,t+1} - P^{S}_{i,t+1} + P^{PV}_{i,t+1} - P^{L}_{i,t+1}\right)\Delta t.$$

In this framework, the DSO transmits the reference prices and receives the metering data once every Δt (several minutes). Thus, it is with greater tolerance for packet loss than the method which requires the DSO and MGs to repeatedly solve the local problems and transmit variables during each time slot [19], [20].

B. The DSO Optimization Problem

At the upper level, the DSO agent sets the reference price sequences $\hat{\lambda}^{MG}_{i,t}$ during time slot t. The MGs respond to $\hat{\lambda}^{MG}_{i,t}$ according to their strategies at the lower level. The upper-level problem $\mathbf{P}^0$ is formulated in (1a)-(1n).

$$\min_{\Omega_0} \lim_{\mathcal{T}\to\infty} f(\Omega_0, \mathcal{T}) \tag{1a}$$

$$\text{s.t. } f(\Omega_0, \mathcal{T}) = \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\Big[\sum_{i\in\mathcal{N}_{MG}} \lambda^{MG}_{i,t}P_{i,t} + \lambda^{HV}_{t}P_{0,t}\Big], \tag{1b}$$

$$P_{i,t} + \sum_{(i,j)\in\mathcal{E}} P_{ij,t} = 0, i \in \mathcal{N}, \tag{1c}$$

$$Q_{i,t} + \sum_{(i,j)\in\mathcal{E}} Q_{ij,t} = 0, i \in \mathcal{N}, \tag{1d}$$

$$V_{i,t} - V_{j,t} - \frac{\left(r_{ij}P_{ij,t} + x_{ij}Q_{ij,t}\right)}{V^2_{0,t}} = 0, (i,j) \in \mathcal{E}, \tag{1e}$$

$$P_{ij,t} + P_{ji,t} = 0, (i,j) \in \mathcal{E}, \tag{1f}$$

$$Q_{ij,t} + Q_{ji,t} = 0, (i,j) \in \mathcal{E}, \tag{1g}$$

$$\underline{V}_i \le V_{i,t} \le \bar{V}_i, i \in \mathcal{N}, \tag{1h}$$

$$-\Delta\bar{\lambda} \le \hat{\lambda}^{MG}_{i,t}(\tau) - \hat{\lambda}^{MG}_{i,t-1}(\tau+1) \le \Delta\bar{\lambda}, \quad i \in \mathcal{N}_{MG}, \tau = 1, \ldots, T-1, \tag{1i}$$

$$\underline{\lambda}^{MG}_{i} \le \hat{\lambda}^{MG}_{i,t}(\tau) \le \underline{\lambda}^{MG}_{i}, i \in \mathcal{N}_{MG}, \tau = 1, \ldots, T, \tag{1j}$$

$$P_{i,t} = -P^{L}_{i,t}, i \in \mathcal{N}/(\mathcal{N}_{MG} \cup \{0\}), \tag{1k}$$

$$Q_{i,t} = -Q^{L}_{i,t}, i \in \mathcal{N}, \tag{1l}$$

$$P_{i,t} \in \mathbf{P}^{*}_{i,t-1}\left(\hat{\lambda}^{MG}_{i,t-1}\right), i \in \mathcal{N}_{MG}, \tag{1m}$$

$$\lambda^{MG}_{i,t} = \hat{\lambda}^{MG}_{i,t-1}(1), i \in \mathcal{N}_{MG}, \tag{1n}$$

where $\Omega_0 = \bigcup_{t=1}^{\mathcal{T}}\{\hat{\lambda}^{MG}_{i,t-1}, P_{i,t}, Q_{i,t}, P_{ij,t}, Q_{ij,t}, V_{i,t}\}$ is the set of decision variables, $\mathbf{P}^{*}_{i,t}(\hat{\lambda}^{MG}_{i,t-1})$ is the strategy adopted by the MG which yields $P_{i,t}$ with respect to the reference price sequence $\hat{\lambda}^{MG}_{i,t-1}$. Node 0 is connected to the high voltage grid and set as the voltage balance node. Index τ is the time index in the horizon of predictions and reference prices at each time instant t. T is the length of the horizon.

The first term in the objective function (1b) for the DSO is the cost of buying electricity from the MGs or the income by selling electricity to the MGs, while the second term is the cost of buying electricity from the high voltage grid. Δt is omitted since it is constant. We assume that the loads at $i \in \mathcal{N}/\mathcal{N}_{MG}$ are inflexible and the electric prices at these nodes are fixed. In this case, the objective function (1b) is equivalent to the profit of the DSO running the distribution network. Constraints (1c) and (1d) are the active and reactive power balance at each node. Constraint (1e) is the Distflow equation [40] which is widely applied to model the optimal power flow in the distribution network. Constraint (1h) is to guarantee that the voltage level of each node is within a predefined range for safety. Constraint (1i) guarantees that the variation of the reference price for the same time slot given at two continuous time slots does not exceed a certain range. This constraint is to prevent the DSO from tricking the MGs with a higher price into improving generator output in advance and then lowering the price when clearing the transacted electricity. Constraint (1j) is to limit the prices for the MGs. Constraints (1k) and (1l) are the power balance equations at nodes $i \in \mathcal{N}/\mathcal{N}_{MG}$. Constraint (1m) is the strategy adopted by the MGs which is unknown to the DSO in practice.

The strategy function $\mathbf{P}^{*}_{i,t-1}(\hat{\lambda}^{MG}_{i,t-1})$ is usually associated with the resources and states in the MG. Based on the strategy, the MG coordinates its generator, PV, and ES to optimize its objective, such as maximizing its generation profit when selling electricity to the DSO or minimizing its cost when buying from the DSO. A model of the strategy is given in the next subsection.

At the upper level, at time t, the DSO agent decides $\hat{\lambda}^{MG}_{i,t}$ based on the predictions to optimize the upper-level problem $\mathbf{P}_0$. The electricity transacted during time slot $t+1$ is cleared at price $\lambda^{MG}_{i,t+1} = \hat{\lambda}^{MG}_{i,t}(1)$. Then the DSO observes the predictions at time $t+1$ and decides $\hat{\lambda}^{MG}_{i,t+1}$. So and so forth.

Problem $\mathbf{P}^0$ is generally non-convex since the objective function (1b) is with bi-linear terms $\lambda^{MG}_{i,t}P_{i,t}$. Moreover, constraint (1m) may also result in a non-convex feasible region since $P_{i,t}$ is the solution to the lower-level optimization problem.

C. The MG Optimization Problem

In this subsection, we provide the MG optimization problem [13] as an instance of the response behaviors. We note that the developed method in this paper is not limited to this model, but is suitable for other MG strategies as well.[1] The linear programming problem $\mathbf{P}^{LP}_{i,t}$ is formulated for the MG at node $i \in \mathcal{N}_{MG}$ in (2a)-(2k). $\mathbf{P}^{*}_{i,t}(\hat{\lambda}^{MG}_{i,t})$ in (1m) is the set of solutions of $\mathbf{P}^{LP}_{i,t}$.

$$\min_{\Omega_{i,t}} g\left(\Omega_{i,t}\right) \tag{2a}$$

$$\text{s.t. } g\left(\Omega_{i,t}\right) = \sum_{\tau=1}^{T}\Big[-\hat{\lambda}^{MG}_{i,t}(\tau)\hat{P}_{i,t}(\tau) + C^{G}_{i}\hat{P}^{G}_{i,t}(\tau)\Big], \tag{2b}$$

$$\hat{P}_{i,t} = \hat{P}^{G}_{i,t} - \hat{P}^{S}_{i,t} - \tilde{P}^{L}_{i,t} + \tilde{P}^{PV}_{i,t}, \tag{2c}$$

$$-\Delta\bar{P}^{G}_{i} \le \hat{P}^{G}_{i,t}(\tau+1) - \hat{P}^{G}_{i,t}(\tau) \le \Delta\bar{P}^{G}_{i}, \quad \tau = 1, \ldots, T-1, \tag{2d}$$

$$-\Delta\bar{P}^{G}_{i} \le \hat{P}^{G}_{i,t}(1) - P^{G}_{i,t} \le \Delta\bar{P}^{G}_{i}, \tag{2e}$$

[1] The behavior that considers quadratic cost functions and (dis)charging loss is tested in Section VI.

$$\hat{S}_{i,t}(\tau+1)-\hat{S}_{i,t}(\tau)=\hat{P}^{S}_{i,t}(\tau+1)\Delta t,\quad \tau=1,\ldots,T-1, \tag{2f}$$

$$\hat{S}_{i,t}(1)-S_{i,t}=\hat{P}^{S}_{i,t}(1)\Delta t, \tag{2g}$$

$$\underline{P}^{G}_{i}\le\hat{P}^{G}_{i,t}(\tau)\le\bar{P}^{G}_{i},\ \tau=1,\ldots,T, \tag{2h}$$

$$\underline{P}^{S}_{i}\le\hat{P}^{S}_{i,t}(\tau)\le\bar{P}^{S}_{i},\ \tau=1,\ldots,T, \tag{2i}$$

$$\underline{S}_{i}\le\hat{S}_{i,t}(\tau)\le\bar{S}_{i},\ \tau=1,\ldots,T, \tag{2j}$$

$$P_{i,t+1}=\hat{P}^{G}_{i,t}(1)-\hat{P}^{S}_{i,t}(1)+P^{PV}_{i,t+1}-P^{L}_{i,t+1}, \tag{2k}$$

where $\Omega_{i,t}=\{P_{i,t+1},\hat{P}_{i,t},\hat{P}^{G}_{i,t},\hat{P}^{S}_{i,t},\hat{S}_{i,t}\}$ is the set of decision variables of the MG at node $i\in\mathcal{N}_{MG}$.

In objective function (2b), the first term is to minimize the cost of buying electricity when $\hat{P}_{i,t}(\tau)<0$ or to maximize the profit for selling electricity when $\hat{P}_{i,t}(\tau)>0$. Constraint (2c) is the active power balance of the node where the MG is located. The active power injection equals the sum of the generator output, ES unit (dis)charging, PV unit output, and loads. It is assumed that only the load produces reactive power, and thus only the active power is the decision variable in this model. Constraints (2d) and (2e) are to restrict the ramping rate of the generator. Constraints (2f) and (2g) are the energy balance of the ES. Constraints (2h), (2i), and (2j) are to restrict the generator power output, the ES (dis)charging power, and the stored energy, respectively. Constraint (2k) calculates the actual active power injection at node i to the distribution network, which is the feedback to the DSO.

The lower-level problems $\mathbf{P}^{LP}_{i,t}$ of the MGs are a part of the upper-level problem $\mathbf{P}_0$. At each time t, the DSO agent decides the reference price $\hat{\lambda}^{MG}_{i,t}$ to minimize the long-term average cost. Then, the MGs solve the lower-level problem $\mathbf{P}^{LP}_{i,t}$ to determine the generation $P^{G}_{i,t+1}=\hat{P}^{G}_{i,t}(1)$ and the (dis)charging power $P^{S}_{i,t+1}=\hat{P}^{S}_{i,t}(1)$, and thus feed back the power injection $P_{i,t+1}$ to the DSO agent in (1m). So and so forth.

V. The Proposed Methodology

In this section, we first transform the pricing problem into an MDP problem. Then, a model-free RL algorithm is developed to solve this problem without the knowledge of the MGs' response behavior. To deal with the large state space, a deep neural network structure is developed to decrease the feature space. To address the difficulty of low exploration efficiency caused by the large action space, the reference policy is incorporated into the algorithm.

A. The MDP Problem

For the DSO, the system described in the previous section is modeled as an MDP, which is characterized by a tuple $\langle\mathcal{S},\mathcal{A},P,c\rangle$, where $\mathcal{S}$ is the finite state space with cardinality $|\mathcal{S}|$, $\mathcal{A}$ is the finite action space with cardinality $|\mathcal{A}|=T\cdot|\mathcal{N}_{MG}|$, $P(s'|s,a):\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ is the state transition probability from $s\in\mathcal{S}$ to $s'\in\mathcal{S}$ determined by action $a\in\mathcal{A}$, and $c(s,a):\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is the cost function received by the agent. The agent executes policy $\pi(s,a):\mathcal{S}\times\mathcal{A}\to[0,1]$ with $\sum_{a\in\mathcal{A}}\pi(s,a)=1$, where $\pi(s,a)$ is the probability of choosing action a at state s. The physical system is continuous in $\mathbb{R}$, while $\mathcal{S}$ is considered to be finite since the measurement accuracy is limited and the values are bounded. Action space $\mathcal{A}$ is considered to be finite because the price is also bounded and is quoted in increments of \$0.01 in this paper. Requirement of state and action spaces to be finite is standard in most studies on MDPs [7], which makes it convenient to define the MDP and the corresponding objective function.

At each time t, the DSO agent stays at state $s_t\in\mathcal{S}$, where s_t is $\{\tilde{P}^{L}_{i,t},\tilde{Q}^{L}_{i,t},i\in\mathcal{N}\}\cup\{\tilde{P}^{PV}_{i,t},i\in\mathcal{N}_{MG}\}\cup\{\tilde{\lambda}^{HV}_{t}\}\cup\{P_{i,t},Q_{i,t},V_{i,t},i\in\mathcal{N}\}$, i.e., the predictions and current system state. Then, it takes action $a_t\sim\pi(s_t,a)$, where a_t is $\{\hat{\lambda}^{MG}_{i,t},i\in\mathcal{N}_{MG}\}$, i.e., the reference prices. The cost

$$c_t=\sum_{i\in\mathcal{N}_{MG}}\lambda^{MG}_{i,t}P_{i,t}+\lambda^{HV}_{t}P_{0,t}+Pen(V_t), \tag{3}$$

will be received by the agent, where $Pen(V_t)$ is a large number if $\exists V_{i,t}$ that violates the safe constraint and equals 0 otherwise. Then the agent steps to state s_{t+1} $P(s|s_t,a_t)$, i.e., the predictions and system state at time $t+1$, where $P(s|s_t,a_t)$ is the state transition probability.

Probability $P(s'|s,a)$ is implicit in the bi-level system defined in Fig. 1. It is determined by the following factors: variation of the load, the renewable energy, and the real-time electric price of the high voltage network; the measurement and prediction error; the response behavior of the MGs that determines the power injection by the MGs during the next time slot; the probability distribution of the facility state which are unobservable to the DSO such as the generator output and the state of charge of the ES units. The model-free RL algorithm does not require the closed-form of transition probability P. Since the cardinality of the state space is large in practice, methods that store a distribution over actions for each state are impractical. Thus, we apply parametrized functions to generate the distribution of randomized policy π_θ with parameter $\theta\in\Theta$ for the agent, where $\Theta\subset\mathbb{R}^m$ is a convex and compact set. Normal distribution $Normal(\mu_\theta,\Sigma)$ is applied to represent pricing policy π_θ, where $\mu_\theta:\mathcal{S}\to\mathbb{R}^{|\mathcal{A}|}$ is the parametrized function mapping the state space $\mathcal{S}$ to the mean vector. The covariance matrix Σ is a diagonal matrix with fixed elements. When the DSO agent takes an action, a vector is sampled from π_θ and then rounded to 2 decimal places to produce a_t. The probability density of $Normal(\mu_\theta(s),\Sigma)$ is taken as $\pi_\theta(s,a)$ for computing simplicity.

We then make the following assumption.

Assumption 1: The MDP satisfies

(1.1) $\forall s\in\mathcal{S},a\in\mathcal{A}$, and $\forall\theta\in\Theta$, $\pi_\theta(s,a)>0$;

(1.2) $\pi_\theta(s,a)>0$ is continuously differentiable with respect to θ over Θ;

(1.3) the Markov chain $\{s_t\}_{t\ge0}$ is irreducible and aperiodic induced by any policy π_θ.

Assumption 1 implies that the MDP is with a stationary distribution $d_\theta(s)$ induced by policy π_θ. The objective of the agent is to find the optimal policy to minimize the expected long-term average cost, which is given by

$$\begin{aligned}\min_{\theta\in\Theta}J(\theta)&=\min_{\theta\in\Theta}\lim_{\mathcal{T}\to\infty}\frac{1}{\mathcal{T}}\mathbb{E}\left[\sum_{t=0}^{\mathcal{T}-1}c_{t+1}\right]\\&=\min_{\theta\in\Theta}\sum_{s\in\mathcal{S}}d_\theta(s)\sum_{a\in\mathcal{A}}\pi_\theta(s,a)c(s,a),\end{aligned} \tag{4}$$

B. Reference Policy-Based RL Algorithm

Given a policy π_θ with parameter θ, the relative action-value function is defined by

$$Q_\theta(s,a)=\mathbb{E}\left[\sum_{t=0}^{\infty}(c_{t+1}-J(\theta))\middle|s_0=s,a_0=a,\pi_\theta\right], \quad (5)$$

and the state-value function is defined by

$$V_\theta(s)=\sum_{a\in\mathcal{A}}\pi_\theta(s,a)Q_\theta(s,a), \quad (6)$$

which satisfies the Poisson equation [7]

$$J(\theta)+Q_\theta(s,a)=c(s,a)+\sum_{s'\in\mathcal{S}}P(s'|s,a)V_\theta(s'). \quad (7)$$

To optimize objective (4) over θ, it is needed to calculate the gradient. The result in [7] shows that the gradient of $J(\theta)$ w.r.t. θ is given by

$$\nabla_\theta J(\theta)=\mathbb{E}_{s\sim d_\theta,a\sim\pi_\theta}[Q_\theta(s,a)\nabla_\theta\ln\pi_\theta(s,a)] \quad (8)$$

when Assumption 1 holds.

This means that the policy gradient could be estimated under the current stationary distribution of state and action. However, in practice, $Q_\theta(s,a)$ is hard to obtain, so a parametrized function $Q_\phi(s,a)$ with parameter $\phi\in\mathbb{R}^n$ is utilized to approximate $Q_\theta(s,a)$. The update of ϕ is called the critic step. In the critic step, it is aimed at minimizing the error of approximating Q_θ induced by parametrization. For each state s and action a, the residual is defined as

$$\delta(s,a,\phi)=c(s,a)+\mathbb{E}_{s'\sim P,a'\sim\pi_\theta}\left[Q_\phi(s',a')\right]-Q_\phi(s,a), \quad (9)$$

The objective function is given by (10)

$$\min_\phi F(\phi)=\mathbb{E}_{s\sim d_\theta,a\sim\pi_\theta}\left[\delta(s,a,\phi)^2\right]. \quad (10)$$

The update of θ is the step. In the actor step, we replace $Q_\theta(s,a)$ in (8) with $Q_\phi(s,a)$ to estimate the policy gradient, and push θ in the opposite direction of the policy gradient to minimize $J(\theta)$.

One difficulty in solving this MDP problem is the high action dimension. In our case study, the price sequences for 4 MGs are generated for the next 24 hours divided into 5 minutes, then the cardinality of the action space, $|\mathcal{A}|$, would be $24\times12\times4=1152$, which causes the agent hard to reach an optimized policy. The experiment shows that if the initial point of θ is random, the regular RL algorithm will converge to the policy with poor performance. To address this problem, we propose to incorporate a reference policy into the regular RL algorithm as shown in Algorithm 1 and Fig. 2. The reference policy could assist the agent to generate a reasonable policy at the beginning of training.

We use deep neural networks to represent function $Q_\phi(s,a)$ and $\mu_\theta(s)$. It is not suitable to apply a fully connected layer as the first layer, because the large space of $\mathcal{S}$ and $\mathcal{A}$ would result in an extremely large weighting matrix. To address this problem, we develop the neural network structure as shown in Fig. 3, where the prediction sequences $\{\tilde{P}^L_{i,t},\tilde{Q}^L_{i,t},i\in\mathcal{N}\}\cup\{\tilde{P}^{PV}_{i,t},i\in\mathcal{N}_{MG}\}$ (also with the action a_t, i.e., $\{\hat{\lambda}^{MG}_{i,t},i\in\mathcal{N}_{MG}\}$ in the $Q_\phi(s_t,a_t)$ network) are first fed into the block that consists of 1D convolution layers and max-pooling layers, which could reduce the feature space. Second, the extracted feature vectors are flattened and concatenated with $\{P_{i,t},Q_{i,t},V_{i,t},i\in\mathcal{N}\}$. Third, the fully connected layers are fed with the concatenated vector and generate the final

Algorithm 1 Reference Policy-Based RL Algorithm

$t=0$
The DSO observes state s_0
The DSO takes action $a_0\sim\pi_{\theta_0}(s_0)$
repeat
  The DSO observes state s_{t+1}
  The DSO takes action $a_{t+1}\sim\pi_{\theta_t}(s_{t+1})$
  The DSO receives c_{t+1}
  Critic Step
  $\mu^c_{t+1}\leftarrow(1-\alpha_t)\mu^c_t+\alpha_t c_{t+1}$
  $\delta_t\leftarrow c_{t+1}-\mu^c_t+Q_{\phi_t}(s_{t+1},a_{t+1})-Q_{\phi_t}(s_t,a_t)$
  $\phi_{t+1}\leftarrow\phi_t+\alpha_t\delta_t\nabla_\phi Q_\phi(s_t,a_t)$
  Actor Step
  $A_t\leftarrow Q_t\nabla_\theta\ln\pi_\theta(s_t,a_t)$
  $B_t\leftarrow2(\mu_{\theta_t}(s_t)-\tilde{\mu}^*(s_t))^\top\nabla_\theta\mu_\theta(s_t)$
  $\theta_{t+1}\leftarrow\theta_t-\beta_t(A_t+\gamma_t B_t)$
  $t\leftarrow t+1$
until Max loop number

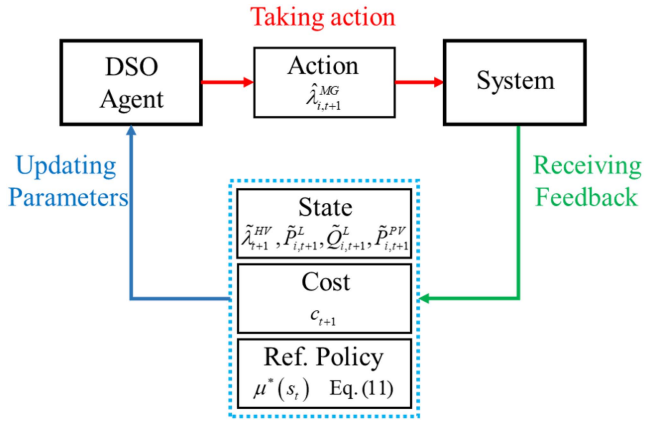

Fig. 2. The flowchart of training the DSO agent.

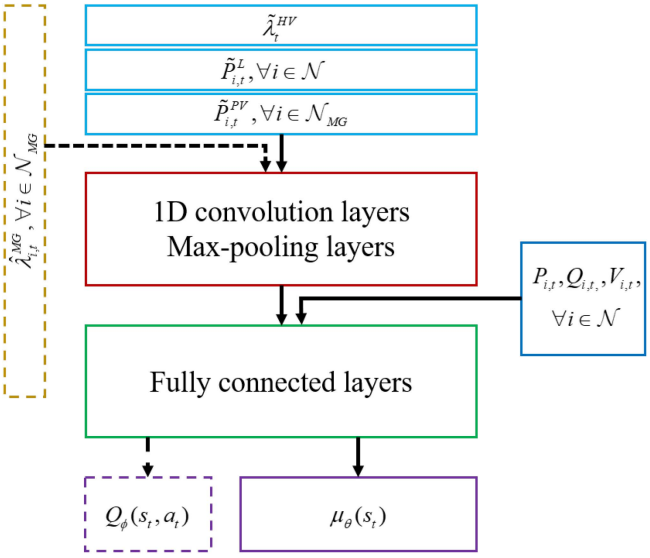

Fig. 3. The neural network structure for $Q_\phi(s,a)$ and $\mu_\theta(s)$. Compared with $\mu_\theta(s)$, the $Q_\phi(s,a)$ network needs the action as additional input, which is $\hat{\lambda}^{MG}_{i,t}$.

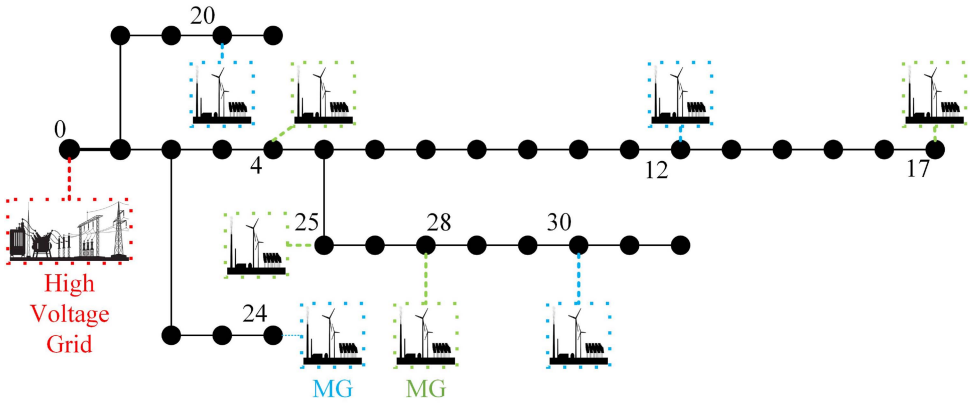

Fig. 4. The IEEE 33-bus distribution network with multiple MGs.

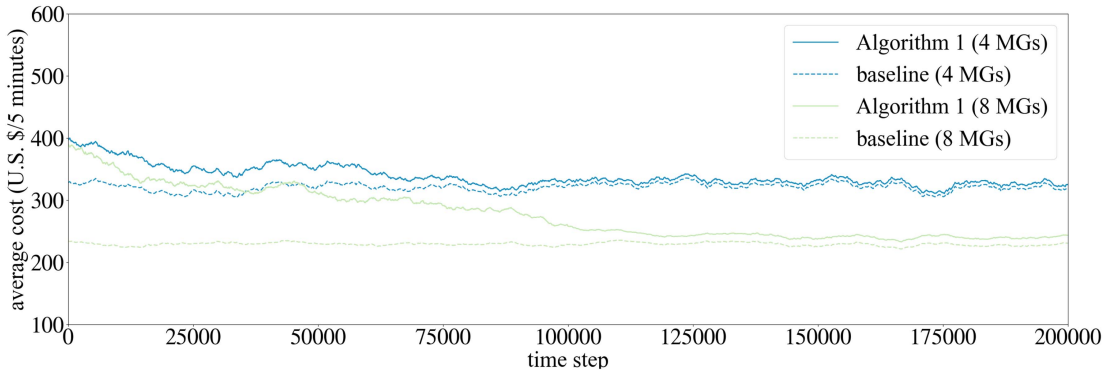

Fig. 5. The cost of the DSO during the training process with $\mathbf{P}_{i,t}^{LP}$ as the lower-level problem (the response behavior of the MGs).

output. This deep neural network structure is similar to the structures in [41], [42], which are applied for the analysis of the aggregated load power. Detailed information on the 1D convolution and fully connected layers is omitted in this paper due to space limitations.

In Algorithm 1, α_t and β_t are the step sizes. The distribution network system stays at the initial state s_0, and the parameters of the DSO agent are initialized as ϕ_0 and θ_0. At each time $t+1$, the DSO agent observes the state s_{t+1} and takes action $a_{t+1} \sim \pi_{\theta_t}(s_{t+1})$ in line 5 and 6. Then the agent executes the critic step and the actor step. In the critic step, the estimator, μ_{t+1}^c, for $J(\theta_t)$ is updated in line 9. The TD error, which is the residual value of equation (7), is calculated in line 10. The parameter of the action-value function, ϕ_t, is updated to minimize the TD error in line 11. In the actor step, A_t is the gradient of $J(\theta)$ and B_t is the gradient of $(\mu_{\theta_t}(s_t) - \tilde{\mu}^*(s_t))^\top (\mu_{\theta_t}(s_t) - \tilde{\mu}^*(s_t))$. The truncated $\tilde{\lambda}_{i,t}^{HV}$, i.e.,

$$\tilde{\mu}^*(s_t)(i,\tau) = \begin{cases} \underline{\lambda}_i^{MG}, \tilde{\lambda}_{i,t}^{HV}(\tau) < \underline{\lambda}_i^{MG} \\ \tilde{\lambda}_{i,t}^{HV}(\tau), \underline{\lambda}_i^{MG} \le \tilde{\lambda}_{i,t}^{HV}(\tau) \le \bar{\lambda}_i^{MG} \\ \bar{\lambda}_i^{MG}, \bar{\lambda}_i^{MG} < \tilde{\lambda}_{i,t}^{HV}(\tau) \end{cases} \tag{11}$$

is set as the reference policy. This reference policy could be easily generated when the prediction for the real-time prices $\tilde{\lambda}_{i,t}^{HV}$ of the high voltage network is acquired. The policy parameter is updated to the direction of $-\beta_t(A_t + \gamma_t B_t)$ in line 15, which is a stochastic gradient descent step with γ_t as the weighting parameter that decays to 0 as $t \to \infty$. The initial value γ_0 balances the convergence to the reference value and the policy improvement through exploration. Since $\gamma_t \to 0$, a large γ_0 would not deteriorate the final optimized policy. However, if γ_0 is too large, the policy will be locked around the reference policy for a long time, which delays further policy improvement. In the following experiment, γ_0 is set as 100 with a decay coefficient 0.97, i.e., $\gamma_{t+1} = 0.97 \cdot \gamma_t$. These values are determined by experimental experience and the scale of the action-value Q is a critical factor. The entire training process is repeated until the policy converges or the user terminates the program.

VI. Numerical Results

A. Experimental Settings

In this section, we apply the developed Algorithm 1 to coordinate the 4 MGs in the IEEE 33-node distribution network as shown in Fig. 4. Each MG owns inflexible loads, a generator, an ES unit, and a PV unit. A simulator for this system is built, where each MG locally solves $\mathbf{P}_{i,t}^{LP}$ at time t to decide the power output of its generator and ES unit at time $t+1$. The load data, PV data, and real-time electric price of the high voltage network are downloaded from PJM.[2] The prediction sequences such as $\tilde{P}_{i,t}^{L}$ are simulated by adding an error vector in which the τth element obeys Gaussian distribution $N(0, \sigma_\tau^e)$ to the actual values, where σ_τ^e increases as τ increases. The time horizon of predictions and reference price sequences is 24 hours and divided into time slots of 5 minutes. Thus, $T = 24 \times 12 = 288$. The simulator is built with Python [43] and the power flow is calculated using

[2]http://dataminer2.pjm.com/

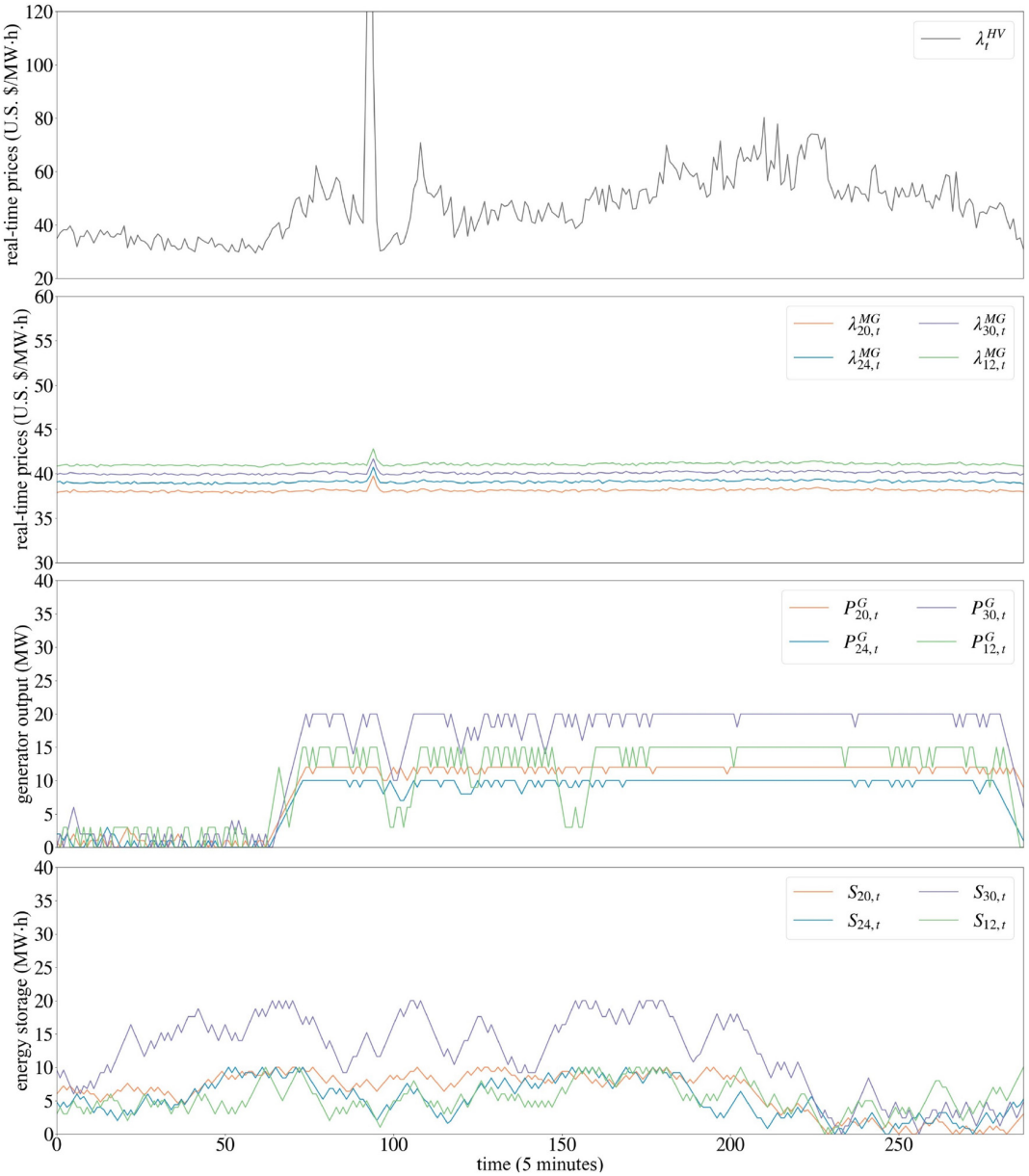

Fig. 6. The curves of $\lambda_{i,t}^{MG}$, $P_{i,t}^{G}$, and $S_{i,t}$ of the MGs with $\mathbf{P}_{i,t}^{LP}$ as the lower-level problem.

pandapower [44]. Moreover, to illustrate the convergence of Algorithm 1, we also add 4 more MGs to the distribution network at buses $i \in \{4, 17, 25, 28\}$ and test the algorithm.

To further evaluate Algorithm 1, the lower-level problem is replaced by $\mathbf{P}_{i,t}^{QP}$, where the cost functions of the generators are quadratic w.r.t. $P_{i,t}^{G}$. Moreover, constraint (2f) is replaced by

$$\hat{S}_{i,t}(\tau+1) - \hat{S}_{i,t}(\tau) = \left(\eta_i^c \hat{P}_{i,t}^c(\tau+1) + \eta_i^d \hat{P}_{i,t}^d(\tau+1)\right)\Delta t,$$
$$\hat{P}_{i,t}^S = \hat{P}_{i,t}^c + \hat{P}_{i,t}^d,$$

where η_i^c and η_i^d are the charging and discharging efficiency factors and so is constraint (2g).

To evaluate the trained policy, we set up the baseline solved by the model-based method. We formulate a finite horizon bi-level problem $\mathbf{P}_t^0$ by replacing (1a) with

$$\min_{\Omega_0} f(\Omega_0, T)$$

and replacing the predictions $\tilde{P}$ and $\tilde{Q}$ in $\mathbf{P}_{i,t}^{LP}$ ($\mathbf{P}_{i,t}^{QP}$) with their actual values. $\mathbf{P}_t^0$ is solved every T time steps. The optimal solutions are set as the baseline. This problem could be transformed into a MILP (MIQP) problem and solved by commercial solvers. The outline is as follows and the whole process is similar to [6]. First, the optimal solution of each problem $\mathbf{P}_{i,t}^{LP}$ ($\mathbf{P}_{i,t}^{QP}$) is represented by the KKT conditions since the problem is linear and the strong duality holds. In this way, the constraint (1m) could be replaced by these KKT conditions. Second, the complementary slackness conditions, which include bi-linear terms, are transformed into mixed-integer linear constraints. Third, the bi-linear term of $\lambda_{i,t}^{MG} P_{i,t}$ in the objective function is transformed into a linear term w.r.t. the dual function of $\mathbf{P}_{i,t}^{LP}$ ($\mathbf{P}_{i,t}^{QP}$) since the strong duality holds. Thus, $\mathbf{P}_t^0$ is transformed into a MILP problem. Gurobi [45] is implemented to the global optimal solution. In addition, TD3 [46] and SAC [47], which are state-of-the-art

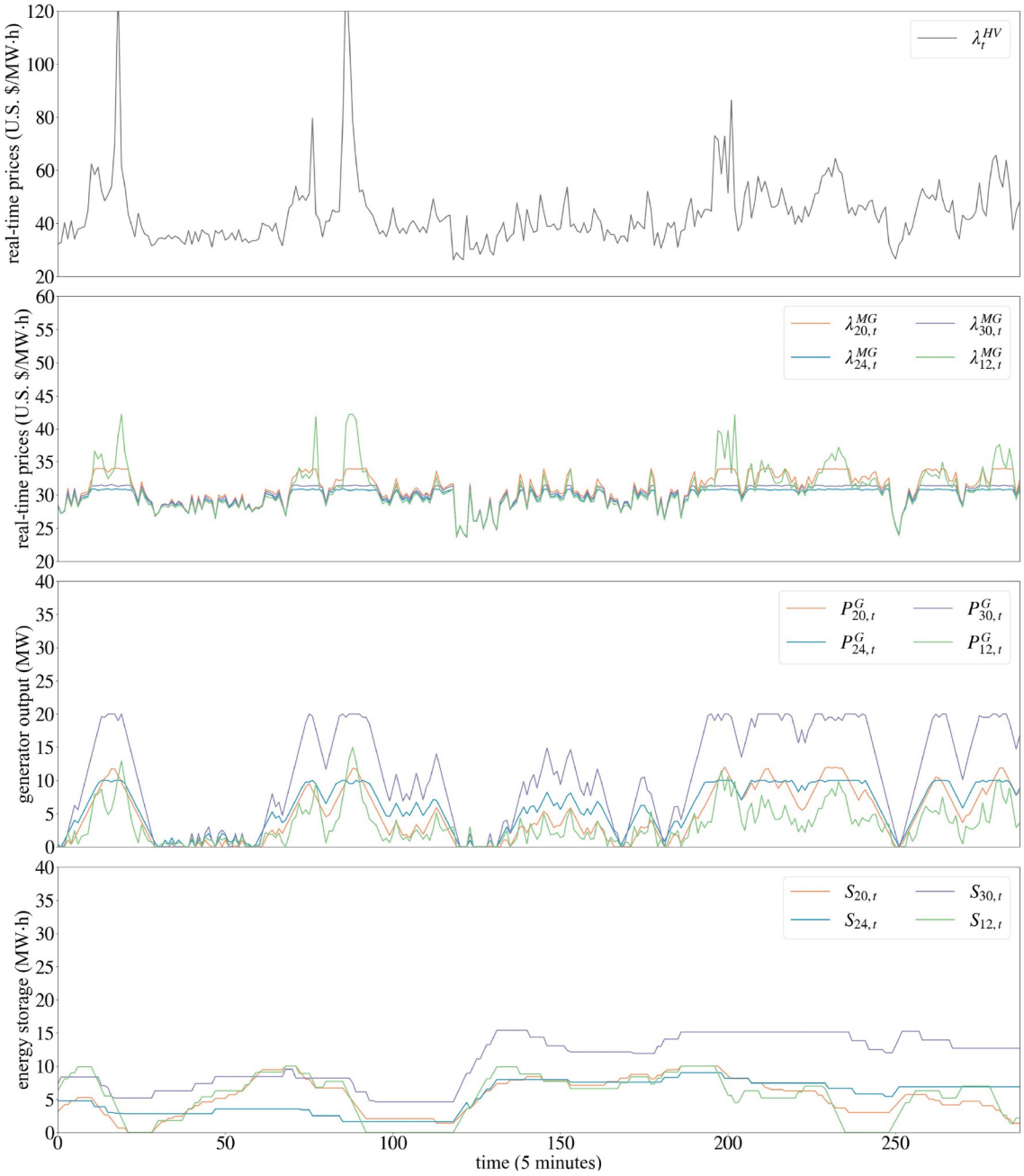

Fig. 7. The curves of $\lambda_{i,t}^{MG}$, $P_{i,t}^{G}$, and $S_{i,t}$ of the MGs with $\mathbf{P}_{i,t}^{QP}$ as the lower-level problem.

RL methods, are directly applied to solve this MDP problem, while they both fail to significantly reduce the cost in finite time.[3]

This baseline is hard to achieve since the DSO does not know $\mathbf{P}_{i,t}^{LP}(\hat{\lambda}_{i,t}^{MG})$ ($\mathbf{P}_{i,t}^{QP}(\hat{\lambda}_{i,t}^{MG})$) in practice. Algorithm 1 is intended to manage this situation by iteratively evaluating and improving the current policy. In the following subsection, we show the experimental results.

B. Experimental Results

We first show the results with $\mathbf{P}_{i,t}^{LP}$ as the lower-level problem. The costs of the DSO during the training process are shown in Fig. 5 as the solid lines and the baseline is shown as the dashed lines. It could be found that Algorithm 1 finally converges to a locally optimal policy.

The comparison of the pricing policies is shown in TABLE I, where "Initial policy" is the policy with randomized θ_0 and "Reference policy" is $\tilde{\mu}^*$ defined in (11). It is shown that the initial randomized policy is improved from the average cost of \$400.06/5min to \$323.32/5min in the 4-MG system and \$404.04/5min to \$240.35/5min in the 8-MG system, while the baseline is \$317.13/5min and \$228.14/5min, respectively. The result shows that Algorithm 1 could generate an optimized pricing policy which is close to the model-based method under incomplete information. It could be also found that the number of MGs affects the convergence of Algorithm 1 since the

[3]This result is not exhibited since it is not the main contribution.

TABLE I
COMPARISON OF COSTS (U.S. $/5 MINUTES) OF THE DSO WITH $\mathbf{P}_{i,t}^{LP}$ AS THE LOWER-LEVEL PROBLEM

MG number	Initial policy	Reference policy	Algorithm 1	Baseline
4	400.06	352.18	323.32	317.13
8	404.04	286.57	240.35	228.14

TABLE II
COMPARISON OF COSTS (U.S. $/5 MINUTES) OF THE DSO WITH $\mathbf{P}_{i,t}^{QP}$ AS THE LOWER-LEVEL PROBLEM

MG number	Initial policy	Reference policy	Algorithm 1	Baseline
4	409.67	373.45	346.38	338.71
8	315.57	284.59	255.82	245.26

action and state space of the system grows larger as the MG number increases.

In Fig. 6, the states of the coordinated facilities in each MG are shown. In general, when the value of λ_t^{HV} is high, the DSO would raise the price to motivate the MG to generate more active power. The resulting output of the generators and E.S. units also increases as λ_t^{HV} does. Moreover, the DSO agent is adapted to the local cost C_i^G. In the setting of the system, $C_{20}^G < C_{24}^G < C_{30}^G < C_{12}^G$. The DSO agent sets the local price $\lambda_{i,t}^{MG}$ around C_i^G to maximize its benefit. This results in the divergence of the behaviors of different MGs. For example, the generator at node 20 (orange line) works at its maximum for the longest time, while the one at node 12 (green line) is for the shortest time. This is compatible with the intuition that the generator with the highest cost should be the last choice to generate active power.

The costs with $\mathbf{P}_{i,t}^{QP}$ as the lower-level problem are shown in TABLE II. The results show that Algorithm 1 is also effective for the response behavior that the MG considers quadratic cost functions and (dis)charging loss. Compared with the initial policy and the reference policy, the optimized costs by Algorithm 1 are closer to the baselines which is optimized by the model-based method in both systems with 4 and 8 MGs.

The states of the coordinated facilities in each MG are shown in Fig. 7. According to the results, the quadratic cost functions of the MGs lead to more variable $\lambda_{i,t}^{MG}$. This is because, without the quadratic term, the MGs would maximize the generator output for any positive return, i.e., $\lambda_{i,t}^{MG} > C_i^G$. The quadratic cost also leads to more fluctuations in $P_{i,t}^G$ since the optimal output continuously varies as $\lambda_{i,t}^{MG}$ changes. In addition, the (dis)charging loss causes the energy storage more stable since these operations result in energy loss.

During each time slot, the DSO agent only spends around 0.002 seconds computing the reference price sequences for all MGs, which is much smaller than $\Delta t = 5$ minutes. This is because the DSO agent only needs to perform a forward propagation of the neural network at each time. Thus, Algorithm 1 is suitable for online dispatch.

VII. CONCLUSION

This paper studies the optimization problem of the pricing policy to coordinate multiple MGs in the distribution network. In practice, the MGs may not provide their response behavior for the DSO due to privacy concerns. Thus, this bi-level system is transformed into an MDP, where the DSO is the agent. The pricing policy is optimized by the developed model-free RL algorithm. The numerical result shows that the policy optimized by our algorithm performs almost as well as the conventional model-based method, while the former is more practical by privacy preservation for the MGs. The number of MGs affects the convergence. And the optimized pricing policy encourages the generator with lower cost to generate more power. Moreover, it shows that the developed algorithm is also effective when the MGs consider quadratic cost functions and (dis)charging loss.

REFERENCES

[1] G. T. Heydt, "The next generation of power distribution systems," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 225–235, Dec. 2010.

[2] F. Farzan, S. Lahiri, M. Kleinberg, K. Gharieh, F. Farzan, and M. Jafari, "Microgrids for fun and profit: The economics of installation investments and operations," *IEEE Power Energy Mag.*, vol. 11, no. 4, pp. 52–58, Jul./Aug. 2013.

[3] M. N. Alam, S. Chakrabarti, and A. Ghosh, "Networked microgrids: State-of-the-art and future perspectives," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1238–1250, Mar. 2019.

[4] X. Kong, D. Kong, J. Yao, L. Bai, and J. Xiao, "Online pricing of demand response based on long short-term memory and reinforcement learning," *Appl. Energy*, vol. 271, 2020, Art. no. 114945. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920304578

[5] Q. Zhang, K. Dehghanpour, Z. Wang, and Q. Huang, "A learning-based power management method for networked microgrids under incomplete information," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1193–1204, Mar. 2020.

[6] A. N. Toutounchi, S. Seyedshenava, J. Contreras, and A. Akbarimajd, "A stochastic bilevel model to manage active distribution networks with multi-microgrids," *IEEE Syst. J.*, vol. 13, no. 4, pp. 4190–4199, Dec. 2019.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[8] M. Glavic, "(Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annu. Rev. Control*, vol. 48, pp. 22–35, Nov. 2019.

[9] H. Çimen, N. Çetinkaya, J. C. Vasquez, and J. M. Guerrero, "A microgrid energy management system based on non-intrusive load monitoring via multitask learning," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 977–987, Mar. 2021.

[10] M. Hosseinzadeh, L. Schenato, and E. Garone, "A distributed optimal power management system for microgrids with plug & play capabilities," *Adv. Control Appl.*, vol. 3, no. 1, p. e65, 2021.

[11] T. S. Ustun and R. H. Khan, "Multiterminal hybrid protection of microgrids over wireless communications network," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2493–2500, Sep. 2015.

[12] J. Duan and M.-Y. Chow, "Robust consensus-based distributed energy management for microgrids with packet losses tolerance," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 281–290, Jan. 2020.

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

[14] M. R. Sandgani and S. Sirouspour, "Priority-based microgrid energy management in a network environment," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 980–990, Apr. 2018.

[15] Y. Du and F. Li, "A hierarchical real-time balancing market considering multi-microgrids with distributed sustainable resources," *IEEE Trans. Sustain. Energy*, vol. 11, no. 1, pp. 72–83, Jan. 2020.

[16] A. Baringo and L. Baringo, "A stochastic adaptive robust optimization approach for the offering strategy of a virtual power plant," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3492–3504, Sep. 2017.

[17] Z. Yi, Y. Xu, H. Wang, and L. Sang, "Coordinated operation strategy for a virtual power plant with multiple DER aggregators," *IEEE Trans. Sustain. Energy*, vol. 12, no. 4, pp. 2445–2458, Oct. 2021.

[18] D. Peng, H. Xiao, W. Pei, and L. Kong, "Interactive pricing optimization of multi-microgrid based on deep learning," in *Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, 2021, pp. 82–85.

[19] D. Papadaskalopoulos and G. Strbac, "Nonlinear and Randomized pricing for distributed management of flexible loads," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1137–1146, Mar. 2016.
[20] D. A. Quijano, M. Vahid-Ghavidel, M. S. Javadi, A. Padilha-Feltrin, and J. P. S. Catalão, "A price-based strategy to coordinate electric springs for demand side management in microgrids," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 400–412, Jan. 2023.
[21] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Optimal residential load control with price prediction in real-time electricity pricing environments," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 120–133, Sep. 2010.
[22] S.-J. Kim and G. B. Giannakis, "An online convex optimization approach to real-time energy pricing for demand response," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2784–2793, Nov. 2017.
[23] R. Yu, W. Yang, and S. Rahardja, "A statistical demand-price model with its application in optimal real-time price," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 1734–1742, Dec. 2012.
[24] T. M. Aljohani, A. F. Ebrahim, and O. A. Mohammed, "Dynamic real-time pricing mechanism for electric vehicles charging considering optimal microgrids energy management system," *IEEE Trans. Ind. Appl.*, vol. 57, no. 5, pp. 5372–5381, Sep./Oct. 2021.
[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[26] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
[27] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 3389–3396.
[28] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.
[29] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 179–186.
[30] X. S. Zhang, Q. Li, T. Yu, and B. Yang, "Consensus transfer *Q*-learning for decentralized generation command dispatch based on virtual generation tribe," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2152–2165, May 2018.
[31] Y. Gao, W. Wang, and N. Yu, "Consensus multi-agent reinforcement learning for volt-VAR control in power distribution networks," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 3594–3604, Jul. 2021.
[32] J. Duan, Z. Yi, D. Shi, C. Lin, X. Lu, and Z. Wang, "Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC–DC microgrids," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5355–5364, Sep. 2019.
[33] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1990–2001, May 2020.
[34] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2192–2203, Jun. 2018.
[35] Y. Xu, W. Zhang, W. Liu, and F. Ferrese, "Multiagent-based reinforcement learning for optimal reactive power dispatch," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1742–1751, Nov. 2012.
[36] M. Al-Saffar and P. Musilek, "Distributed optimization for distribution grids with stochastic DER using multi-agent deep reinforcement learning," *IEEE Access*, vol. 9, pp. 63059–63072, 2021.
[37] D. Fang, X. Guan, Y. Peng, H. Chen, T. Ohtsuki, and Z. Han, "Distributed deep reinforcement learning for renewable energy accommodation assessment with communication uncertainty in Internet of Energy," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8557–8569, May 2021.
[38] B.-G. Kim, Y. Zhang, M. van der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2187–2198, Sep. 2016.
[39] V.-H. Bui, A. Hussain, and W. Su, "A dynamic internal trading price strategy for networked microgrids: A deep reinforcement learning-based game-theoretic approach," *IEEE Trans. Smart Grid*, vol. 13, no. 5, pp. 3408–3421, Sep. 2022.
[40] Z. Yang, K. Xie, J. Yu, H. Zhong, N. Zhang, and Q. Xia, "A general formulation of linear power flow models: Basic theory and error analysis," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1315–1324, Mar. 2019.
[41] J. Kelly and W. Knottenbelt, "Neural NILM: Deep neural networks applied to energy disaggregation," in *Proc. 2nd ACM Int. Conf. Embedded Syst. Energy-Efficient Built Environ.*, 2015, pp. 55–64.
[42] G. Cui, B. Liu, W. Luan, and Y. Yu, "Estimation of target appliance electricity consumption using background filtering," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 5920–5929, Nov. 2019.
[43] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA, USA: CreateSpace, 2009.
[44] L. Thurner et al., "Pandapower—An open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6510–6521, Nov. 2018.
[45] "Gurobi optimizer reference manual." Gurobi Optimization, LLC. 2021. [Online]. Available: https://www.gurobi.com
[46] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[47] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

Gaochen Cui (Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Tianjin University, Beijing, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in control science and engineering with Tsinghua University, Tianjin, China. His research interest is non-intrusive load monitoring, economic dispatch, machine learning, and reinforcement learning.

Qing-Shan Jia (Senior Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2002 and 2006, respectively.

He was a Visiting Scholar with Harvard University, Cambridge, MA, USA, the Hong Kong University of Science and Technology, Hong Kong, and the Massachusetts Institute of Technology, Cambridge, in 2006, 2010, and 2013, respectively. He is currently a Professor with the Center for Intelligent and Networked Systems, Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University. His research interests include theories and applications of discrete-event dynamic systems and simulation-based optimization of cyber-physical systems.

Xiaohong Guan (Life Fellow, IEEE) received the B.S. and M.S. degrees in control engineering from Tsinghua University, Beijing, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and systems engineering from the University of Connecticut in 1993.

He is currently a Professor with the Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, China. He was appointed as a Cheung Kong Professor of Systems Engineering in 1999, and the Dean of the Faculty of Electronic and Information Engineering in 2008. He has been the Director of the Center for Intelligent and Networked Systems, Tsinghua University since 2001, and served as the Head of the Department of Automation from 2003 to 2008. His research interests include economics and security of networked systems, optimization-based planning and scheduling of power and energy systems, manufacturing systems, and cyber-physical systems, including smart grid and sensor networks. He is a member of Chinese Academy of Science.