Consensus-based Distributed Reinforcement Learning with Primal-Dual Update for Networked Microgrids On-Line Coordination

Gaochen Cui, Qing-Shan Jia, *Senior Member, IEEE*, Xiaohong Guan, *Fellow, IEEE*, Qiaozhu Zhai, *Member, IEEE*, Xianping Guo, *Senior Member, IEEE* and Qi Guo

***Abstract*—This paper develops a distributed reinforcement learning (RL) method to coordinate cooperative microgrids (MGs). The high uncertainty of power loads and renewable energy sources motivate the operator to dispatch in real time. On the one hand, the existing on-line methods usually utilize the approximate models that result in intractable constraint violation. A common method is to relax it as a chance constraint, while it is still hard to ensure that it is met in practice. On the other hand, some MGs may hope to preserve the private information on their local costs and states. To address these problems, we make the following contributions. First, the coordination problem is reformulated as a constrained multi-agent Markov decision process. Second, the distributed RL algorithm with a theoretical convergence guarantee is developed. Third, to further preserve the local private information and improve the performance, this algorithm is modified by adding a local feature extraction module for each agent. This module could also be regarded as an encryption module for the local state information. Fourth, numerical experiments are carried out to validate the effectiveness of the modified algorithm.**

***Index Terms*—Reinforcement learning, Microgrids, Distribution network, Constrained Markov decision processes, Multi-agent system.**

I. Introduction

In this paper, the MGs are considered to cooperatively minimize the total electricity cost while maintaining voltage safety. To deal with the high uncertainty, the on-line algorithms such as the on-line alternating direction method of multipliers (ADMM) are applied to dispatch the MGs in real time [1]. These on-line methods usually depend on the simplified mathematical models such as linear power flow equations and make one-step iteration for speed improvement. Thus, the optimized policy cannot be guaranteed to be optimal nor safe in the real system. Moreover, these methods are model-based that require the MGs and the operator to have an accurate closed-form model of the distribution network and the electrical devices. However, the model might be hard to acquire due to privacy concerns and information scarcity [2]. This issue also brings challenges to these conventional methods.

The economic dispatch problem belongs to the tertiary control in the power system [3]. The low-frequent violation of the voltage constraint could be tolerated in practice, which would be handled by the secondary control system [4], [5]. Thus, in this paper, we relax the voltage constraint as a chance constraint, which is a common way [6]. Then, the coordination problem could be reformulated as a constrained Markov decision process (CMDP) [7].

RL is a method to solve the single-agent CMDP problems through learning from experience without a closed-form model [8], [9]. For the multi-agent systems without constraints, distributed algorithms are developed to optimize the global objective while preserving private information on the local costs [10]. With deep learning [11], the parametrized RL methods achieve the state of the art in multiple areas [12] including the on-line energy management in power systems [2], [13]. However, with the current approach by adding a fixed penalizing term for voltage violation, it is still hard to directly regulate the frequency, since this term also requires tuning.

To address the above problems, we make the following contributions in this paper. (i) The energy management problem of the multi-MG system is modeled as a constrained multi-agent Markov decision process (CMAMDP). (ii) A distributed RL algorithm with a convergence guarantee is developed to solve the general CMAMDP, which preserves private information about local costs. (iii) To improve the performance and further preserve the privacy of local states and actions, the feature extraction module is developed. And the modified algorithm which incorporates this module is developed. (iv) Numerical experiments on the IEEE 33-bus distribution network with 4 MGs are carried out. The experimental results demonstrate that the developed algorithm outperforms the existing methods including the interior point method and on-line ADMM in regulating the frequency of voltage violations. This regulation is at the expense of 1-2% higher economic cost in the experiments. Compared against the RL algorithm with fixed

Gaochen Cui and Qing-Shan Jia are with the CFINS, Department of Automation, BNRist, Tsinghua University, Beijing 100084, China (e-mail: zhuyh21@mails.tsinghua.edu.cn; cgc19@tsinghua.org.cn; liuab19@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, China (e-mail:xhguan@xjtu.edu.cn).

Qiaozhu Zhai is with the MOEKLINNS Lab Systems Engineering Institute, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qzzhai@sei.xjtu.edu.cn).

Xianping Guo is with School of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: mcsgxp@mail.sysu.edu.cn).

Qi Guo is with the China Southern power Grid Electric Power Research Institute, Gaungzhou 510630, Guangdong, China (e-mail:haler@qq.com).

This work is supported by the National Key Research and Development Program of China (2022YFA1004600), the National Natural Science Foundation of China (No. 62125304, 62192751, and 62073182), and the 111 International Collaboration Project (No. BP2018006).

penalizing term, our algorithm is adaptive to various preset limits.

II. Literature Review

The distributed MG systems have been well defined and studied in the past 10 years. A decentralized control architecture for MGs is built in [14] where the MGs cooperate for the overall system objective. A multi-agent system for distributed management of MGs with a market is built in [15]. A distributed multi-agent energy management system architecture is built with the non-cooperative game theory in [16]. The decentralized algorithms for optimal resource management for MGs in both grid-connected and islanded modes are developed in [17]. These off-line algorithms usually take a long time to converge so that the requirement of real-time solution would not be met. A two-level distributed dispatch algorithm based on the ADMM is developed in [18], where the upper level is the day-ahead dispatch and the lower level is the intra-day scheduling plan. A fully decentralized algorithm is developed for economic dispatch and a locally centralized-distributed algorithm is developed for multi-step large-scale MGs in [19]. These on-line algorithms usually use the approximate models such as linear power flow equations. The voltage violation brought by the approximation error is hard to derive and thus uncontrollable in the real scenarios. Thus, we propose to restrict the frequency of voltage violations since the violation is inevitable in practice and the low-frequent abnormal situations could be properly regulated with the secondary control system [4], [5].

To make real-time decisions, RL has been widely implemented to dispatch the MGs. The Lagrangian-based RL method is developed to solve the real-time OPF problem in [20], where the approximated deterministic gradient is applied instead of the critic function. The PPO algorithm with imitation learning to obtain an initial policy is developed to solve the real-time OPF problem in [21]. The bi-level RL framework to coordinate MGs is proposed in [22]. The OPF problem of distribution networks embedded with RES and ES units is solved with the PPO algorithm in [23]. The economic dispatch problem of MGs is solved with the developed cooperative MARL algorithm in [24]. An on-line dynamic dispatch scheme and algorithm are developed for autonomous operation especially when a dispatch center is unavailable or day-ahead planning is infeasible in [25]. A distributed hierarchical RL algorithm embedded with operation knowledge is developed in [26] for real-time MGs dispatch.

To ensure the safety of the optimized policy, an additional safe RL agent is trained to take over the controller when the action selected by the vanilla RL agent is determined to be unsafe [27]. The logarithmic barrier function is penalized on the objective function when the constraint is violated in [28]. A security neural network is trained to recognize the unsafe actions, and a safety layer is developed to correct the actions in [29]. Although this method is implemented on the multi-MG system, each agent is focused on its own objective. The primal-dual method is implemented to deal with the constraints, and the consensus method leverages the Lagrange multipliers of the global constraints to coordinate the policy optimization of the MGs in [30]. However, this method still requires all the MGs to observe the global state of the distribution network, and each MG minimizes its own economic cost.

In this article, we solve the real-time economic dispatch problem for multiple MGs in a cooperative game. Compared with the existing methods, each agent only observes the local cost to minimize the global cost such that the global chance constraint is satisfied. The consensus method is leveraged to enable each MG to approximate the global objective and thus calculate the local policy gradient.

III. Nomenclature

A. Parameters

$\mathcal{N}_{DN}$	The set of all buses in the distribution network.
$\mathcal{N}_{MG}$	The set of MG buses.
$\mathcal{E}$	The set of distribution network lines.
$i \in \mathcal{N}_{DN}$	The bus index.
$t \in \mathbb{N}$	The time slot index.
$T \in \mathbb{N}$	The number of time slots in the time horizon of the forecast.
$\Delta t \in \mathbb{R}$	The time duration of a time slot.
$\bar{S}_i^G \in \mathbb{R}$	The max apparent power output of the generator in the MG at bus i.
$\Delta\bar{S}_i^G \in \mathbb{R}$	The max ramping rate of the generator in the MG at bus i.
$\bar{\alpha}_i^G, \underline{\alpha}_i^G \in \mathbb{R}$	The max/min phase angle of the generator in the MG at bus i.
$\Delta\bar{\alpha}_i^G \in \mathbb{R}$	The max phase angle change of the generator in the MG at bus i.
$\Delta\bar{S}_i^G \in \mathbb{R}$	The max ramping rate of the generator in the MG at bus i.
$\bar{S}_i^S \in \mathbb{R}$	The max apparent power of the ES unit in the MG at bus i.
$\bar{\alpha}_i^S, \underline{\alpha}_i^S \in \mathbb{R}$	The max/min phase angle of the ES unit in the MG at bus i.
$\bar{E}_i^S, \underline{E}_i^S \in \mathbb{R}$	The max/min energy storage of the ES unit in the MG at bus i.
$\tilde{P}_{i,t}^{PV} \in \mathbb{R}^T$	The predicted PV active power at bus i, time t for time from $t+1$ to $t+T$.
$\tilde{P}_{i,t}^{L} \in \mathbb{R}^T$	The predicted load active power at bus i, time t for time from $t+1$ to $t+T$.
$\tilde{Q}_{i,t}^{L} \in \mathbb{R}^T$	The predicted load reactive power at bus i, time t for time from $t+1$ to $t+T$.
$P_{i,t}^{PV} \in \mathbb{R}$	The actual PV active power at bus i, time t.
$P_{i,t}^{L} \in \mathbb{R}$	The actual load active power at i, time t.
$Q_{i,t}^{L} \in \mathbb{R}$	The actual load reactive power at i, time t.
$a_i, b_i \in \mathbb{R}$	The cost coefficients of the generator connected in the MG at bus i.
$\bar{V}_i, \underline{V}_i \in \mathbb{R}$	The max/min voltage magnitude at bus i.
$\lambda_t^{HV} \in \mathbb{R}$	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 price for buying electricity from the high voltage grid at time t for time from $t+1$ to $t+T$.

B. Variables

$P_{i,t} \in \mathbb{R}$	The active power injection at bus i, time t.
$Q_{i,t} \in \mathbb{R}$	The reactive power injection at bus i, time t.
$V_{i,t} \in \mathbb{R}$	The voltage magnitude at bus i, time t.
$\alpha_{i,t} \in \mathbb{R}$	The phase angle of voltage at bus i, time t.
$\alpha_{ij,t} \in \mathbb{R}$	$\alpha_{i,t} - \alpha_{j,t}$.
$S^G_{i,t} \in \mathbb{R}$	The apparent power output of the generator in the MG connected at bus i, time t.
$P^G_{i,t} \in \mathbb{R}$	The active power output of the generator in the MG connected at bus i, time t.
$Q^G_{i,t} \in \mathbb{R}$	The reactive power output of the generator in the MG connected at bus i, time t.
$\alpha^G_{i,t} \in \mathbb{R}$	The phase angle of the generator in the MG connected at bus i, time t.
$S^S_{i,t} \in \mathbb{R}$	The apparent power output of the ES unit in the MG connected at bus i, time t.
$P^S_{i,t} \in \mathbb{R}$	The active power output of the ES unit in the MG connected at bus i, time t.
$Q^S_{i,t} \in \mathbb{R}$	The reactive power output of the ES unit in the MG connected at bus i, time t.
$\alpha^S_{i,t} \in \mathbb{R}$	The phase angle of the ES unit in the MG connected at bus i, time t.
$E^S_{i,t} \in \mathbb{R}$	The energy stored in the ES unit in the MG at bus i, time t.

IV. Problem Formulation

A. The Overall Structure

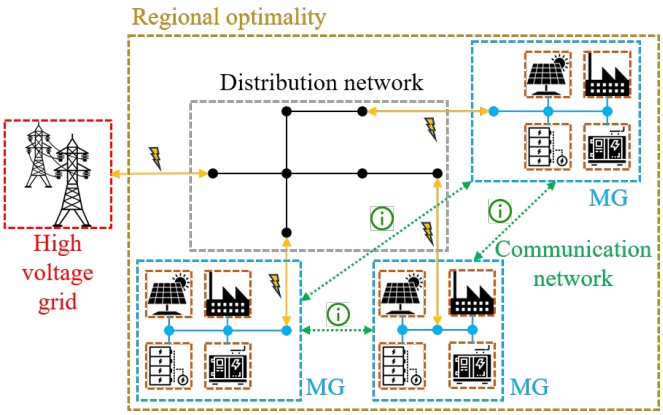

Fig. 1. The framework of the multi-MG system.

The overall structure of the multi-MG system is shown in Fig. 1. The MGs are connected to the distribution network and the power flow is bi-directional. The distribution also exchanges power with the high voltage grid. Each MG owns power loads and distributed energy resources. The MGs cooperatively minimize the total electricity cost of the entire distribution system. Meanwhile, the MGs need to ensure the safety of the power grid, which is to limit the nodal voltage within the safe range. Moreover, we put restrictions on the information acquisition, which is that each MG does not know the cost and state of other MGs due to the privacy preservation. With this restriction, we build a communication network for the MGs to transmit information to locally approximate the global objectives.

B. The Optimization Problem

We consider the dispatch problem where the MGs cooperate to minimize the global electricity cost of the entire distribution network. Without loss of generality, we assume that each MG owns a gas turbine generator, an ES unit, a PV unit, and power loads. The problem is formulated as

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

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

$$\{S^G_{i,t+1}, \alpha^G_{i,t+1}, S^S_{i,t+1}, \alpha^S_{i,t+1}\}_{i\in\mathcal{N}_{MG}} = h(\Omega_t), \tag{1c}$$

$$f^G(P^G_{i,t}) = a_i {P^G_{i,t}}^2 + b_i P^G_{i,t}, i \in \mathcal{N}_{MG}, \tag{1d}$$

$$P^G_{i,t} = S^G_{i,t}\cos\alpha^G_{i,t}, Q^G_{i,t} = S^G_{i,t}\sin\alpha^G_{i,t}, i \in \mathcal{N}_{MG}, \tag{1e}$$

$$-\Delta\bar{S}^G_i \le S^G_{i,t} - S^G_{i,t-1} \le \Delta\bar{S}^G_i, i \in \mathcal{N}_{MG}, \tag{1f}$$

$$-\Delta\bar{\alpha}^G_i \le \alpha^G_{i,t} - \alpha^G_{i,t-1} \le \Delta\bar{\alpha}^G_i, i \in \mathcal{N}_{MG}, \tag{1g}$$

$$\underline{S}^G_i \le S^G_{i,t} \le \bar{S}^G_i, i \in \mathcal{N}_{MG}, \tag{1h}$$

$$\underline{\alpha}^G_i \le \alpha^G_{i,t} \le \bar{\alpha}^G_i, i \in \mathcal{N}_{MG}, \tag{1i}$$

$$P^S_{i,t} = S^S_{i,t}\cos\alpha^S_{i,t}, Q^S_{i,t} = S^S_{i,t}\sin\alpha^S_{i,t}, i \in \mathcal{N}_{MG}, \tag{1j}$$

$$E^S_{i,t} = E^S_{i,t-1} - \eta^S(P^S_{i,t-1})P^S_{i,t-1}\Delta t, i \in \mathcal{N}_{MG}, \tag{1k}$$

$$\underline{S}^S_i \le S^S_{i,t} \le \bar{S}^S_i, i \in \mathcal{N}_{MG}, \tag{1l}$$

$$\underline{\alpha}^S_i \le \alpha^S_{i,t} \le \bar{\alpha}^S_i, i \in \mathcal{N}_{MG}, \tag{1m}$$

$$\underline{E}^S_i \le E^S_{i,t} \le \bar{E}^S_i, i \in \mathcal{N}_{MG}, \tag{1n}$$

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

$$Q_{i,t} = -Q^L_{i,t}, i \in \mathcal{N}_{DN}/(\mathcal{N}_{MG}\cup\{0\}), \tag{1p}$$

$$P_{i,t} = P^G_{i,t} + P^S_{i,t} + P^{PV}_{i,t} - P^L_{i,t}, i \in \mathcal{N}_{MG}, \tag{1q}$$

$$Q_{i,t} = Q^G_{i,t} + Q^S_{i,t} - Q^L_{i,t}, i \in \mathcal{N}_{MG}, \tag{1r}$$

$$P_{i,t} = V_{i,t}\sum_{(i,j)\in\mathcal{E}} V_{j,t}(G_{ij}\cos\alpha_{ij,t} + B_{ij}\sin\alpha_{ij,t}) \tag{1s}$$

$$Q_{i,t} = V_{i,t}\sum_{(i,j)\in\mathcal{E}} V_{j,t}(G_{ij}\sin\alpha_{ij,t} - B_{ij}\cos\alpha_{ij,t}) \tag{1t}$$

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

where $\Omega_t \in \Omega$ is the set of the system information to make dispatch decision at t. It includes the system state of the distribution network, $\{P_{i,t}, Q_{i,t}, V_{i,t}\}_{i\in\mathcal{N}_{DN}}$, the state of the MGs, $\{S^G_{i,t}, \alpha^G_{i,t}, S^S_{i,t}, \alpha^S_{i,t}, E^S_{i,t}\}_{i\in\mathcal{N}_{MG}}$, and the forecast sequences $\{\tilde{P}^L_{i,t}, \tilde{Q}^L_{i,t}\}_{i\in\mathcal{N}_{DN}} \cup \{\tilde{P}^{PV}_{i,t}\}_{i\in\mathcal{N}_{MG}} \cup \{\tilde{\lambda}^{HV}_t\}$. Constraint (1b) is the average electricity cost of the entire distribution network, where the first term is the cost of the generator and the second term is the cost for buying electricity from the high voltage grid. Constraint (1c) is the policy taken by the MGs, where $h(\cdot) : \Omega \to \mathbb{R}^{4\times|\mathcal{N}_{MG}|}$ is the policy function to optimize. The MGs observe the system state and the forecast, and then dispatch their generators and ES units by $h(\cdot)$. Constraint (1d) is the quadratic cost function of the generator in each MG. Constraints (1e)-(1i) are for the generators and constraints (1j)-(1n) are for the ES units. In constraint (1j), the power electronic device of the ES unit outputs/inputs both active and reactive power. The energy stored in the ES unit is calculated by constraint (1k), where $P^S_{i,t-1}\Delta t$ is the energy exchanged with the distribution network during time $t-1$ and $\eta^S(\cdot)$

is the (dis)charging efficiency. Constraints (1o)-(1r) are the power balance equations at each bus. Constraint (1s)-(1t) is the nonlinear power flow equations. Constraint (1u) is to restrict the nodal voltage within its limit, which is the safety constraint.

Since the uncertainties of power loads and renewable resources cause violations of voltage constraint in practice, to relax the safety constraint as the chance constraint is a typical method to incorporate the uncertainties [6]. In the following subsection, we reformulate this problem as a CMAMDP that could be solved by the model-free RL methods.

C. CMAMDP problem

A CMAMDP is a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, c, g \rangle$, where $\mathcal{N}$ is the agent set with $|\mathcal{N}| = N$ as its cardinality, i.e. the number of agents, $\mathcal{S}$ is the finite state space, $\mathcal{A} = \prod_{i=1}^{N} \mathcal{A}^i$ is the finite joint action space with $\mathcal{A}^i$ as agent i's action space and $|\mathcal{A}^i| = 4$, $\mathcal{P}(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the state transition probability from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ determined by action $a = \prod_{i=1}^{N} a^i \in \mathcal{A}$, and $c^i(s,a^i) : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}$ and $g^i(s,a^i) : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}$ are two local cost functions received by agent i. Each agent executes a local policy $\pi^i(s,a^i) : \mathcal{S} \times \mathcal{A} \to [0,1]$ with $\sum_{a^i \in \mathcal{A}^i} \pi^i(s,a^i) = 1$, where $\pi^i(s,a^i)$ is the probability of choosing action a^i at state s. At each time t, the system state is denoted as s_t and each agent instantly gains costs $c_t^i = c^i(s_t, a_t^i)$ and $g_t^i = g^i(s_t, a_t^i)$ after taking action a_t^i at s_t.

The distribution network is divided into N regions. Each MG agent i at bus $n_i \in \mathcal{N}_{MG}$ regulates the region of the bus set $\mathcal{N}_i$, that $n_i \in \mathcal{N}_i \subset \mathcal{N}_{DN}$, and $\bigcup_{i \in \mathcal{N}} \mathcal{N}_i = \mathcal{N}_{DN}/\{0\}$. At time t, each MG i observes its local state $s_t^i = \{P_{j,t}, Q_{j,t}, V_{j,t}, \tilde{P}_{j,t}^L, \tilde{Q}_{j,t}^L\}_{j \in \mathcal{N}_i} \cup \{S_{j,t}^G, \alpha_{j,t}^G, S_{j,t}^S, \alpha_{j,t}^S, E_{j,t}^S, \tilde{P}_{j,t}^{PV}\}_{j=n_i} \cup \{\tilde{\lambda}_t^{HV}\}$ and send it to the central collector. The collector will then broadcast the system state $s_t = \bigcup_{i \in \mathcal{N}} s_t^i$ to all MGs. In Section VI, we develop the algorithm that each agent sends its encrypted information instead of s_t^i so that the privacy is preserved. Each MG agent i takes action $a_t^i = \{S_{j,t+1}^G, \alpha_{j,t+1}^G, S_{j,t+1}^S, \alpha_{j,t+1}^S\}_{j=n_i}$ that follows the distribution $\pi^i(s_t, \cdot)$. Each MG agent receives the local cost c_t^i and g_t^i, where

$$c_t^i = \frac{1}{N} \lambda_t^{HV} P_{0,t} + a_i {P_{i,t}^G}^2 + b_i P_{i,t}^G \tag{2}$$

and

$$g_t^i = |\mathcal{N}_i| - \sum_{j \in \mathcal{N}_i} I(\underline{V}_j \le V_{j,t} \le \bar{V}_j) \tag{3}$$

with $I(\cdot)$ is the indicating function. Thus, g_t^i equals the number of nodes with abnormal voltage. Then the system steps to s_{t+1} following $\mathcal{P}(\cdot|s_t, a_t)$. So and so forth. Probability $\mathcal{P}$ is determined by the following factors: the stochastic variation of the power loads, the renewable energy resources, and the real-time electric price of the high voltage network; the stochastic error of measurement and prediction; the state transition of the facilities such as the generator and the ES units; the power flow in the network. These transition rules are modeled as Eqs. (1f)-(1t). The closed form of $\mathcal{P}$ is difficult to acquire in practice, and thus we implement the model-free RL algorithm in the following sections to solve this MDP problem.

We apply parametrized functions to generate the distribution of randomized policy $\pi_{\theta^i}^i$ with parameter $\theta^i \in \Theta^i$ for each agent i, where $\Theta^i \subset \mathbb{R}^{m_i}$ is a convex and compact set. Let $\pi_\theta = \prod_{i=1}^{N} \pi_{\theta^i}^i$ denote the joint policy, where $\theta = ((\theta^1)^\top, (\theta^2)^\top, ..., (\theta^N)^\top) \in \Theta$ and $\Theta = \prod_{i=1}^{N} \Theta^i$. Let $\mathcal{P}^\theta$ denote the transition probability of the Markov chain $\{s_t\}_{t \ge 0}$ induced by the joint policy π_θ, i.e.,

$$\mathcal{P}^\theta(s'|s) = \sum_{a \in \mathcal{A}} \pi_\theta(s,a) \mathcal{P}(s'|s,a), \forall s, s' \in \mathcal{S}. \tag{4}$$

Assumption 1. *For the Markov chain,*
(1.1) $\forall i \in \mathcal{N}, s \in \mathcal{S}, a^i \in \mathcal{A}^i$, and $\forall \theta^i \in \Theta^i$, $\pi_{\theta^i}^i(s,a^i) > 0$.
(1.2) $\pi_{\theta^i}^i(s,a^i)$ is continuously differentiable with respect to (w.r.t.) θ^i over Θ^i.
(1.3) The Markov chain $\{s_t\}_{t \ge 0}$ induced by any policy π_θ is irreducible and aperiodic. The randomized actions chosen by all agents are statistically independent.
(1.4) $c^i(s,a^i)$ and $g^i(s,a^i)$ are bounded $\forall s \in \mathcal{S}, a^i \in \mathcal{A}^i, i \in \mathcal{N}$.

Under Assumption 1, we formulate the following CMAMDP problem

$$\begin{aligned} \min_{\theta \in \Theta} J(\theta) &= \lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \mathbb{E}\left[\sum_{t=0}^{\mathcal{T}-1} \frac{1}{N} \sum_{i=1}^{N} c_{t+1}^i \right] \\ &= \sum_{s \in \mathcal{S}} d_\theta(s) \sum_{a \in \mathcal{A}} \pi_\theta(s,a) \bar{c}(s,a), \end{aligned} \tag{5}$$

$$\begin{aligned} \text{s.t. } G(\theta) &= \lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \mathbb{E}\left[\sum_{t=0}^{\mathcal{T}-1} \frac{1}{N} \sum_{i=1}^{N} g_{t+1}^i \right] \\ &= \sum_{s \in \mathcal{S}} d_\theta(s) \sum_{a \in \mathcal{A}} \pi_\theta(s,a) \bar{g}(s,a) \le \bar{G}, \end{aligned} \tag{6}$$

where

$$\bar{c}(s,a) = \frac{1}{N} \sum_{i=1}^{N} c^i(s,a^i), \tag{7a}$$

$$\bar{g}(s,a) = \frac{1}{N} \sum_{i=1}^{N} g^i(s,a^i) \tag{7b}$$

are the global average cost functions. With fixed θ, $d_\theta(s)\pi_\theta(s,a)$ is the joint probability that the system stays at state s and takes action a. Thus, after plugging Eq. (2) in (5), we could verify that $\lim_{\mathcal{T} \to \infty} f(h, \mathcal{T})$ is almost surely equal to $J(\theta)$. In the same way, with g_t^i equal to the number of buses with abnormal voltage at time t, we could regard $N \cdot \mathbb{E}[\bar{g}]$ as the frequency of voltage off-limits in the distribution network, which is restricted by constraint (6).

In this article, each MG only observes its local costs, and thus power system is also equipped with a communication network summarized as graph $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$. Each agent is on a vertex and is connected with other agents through edges in the edge set $\mathcal{E}_t \subseteq \{(i,j) | i,j \in \mathcal{N}, i \ne j\}$. Let W_t denote the communication weight matrix at time t, where $W_t(i,j) = 0$ if $(i,j) \notin \mathcal{E}_t$. The communication network is to assist the agents to estimate the global action-value in the Section V. Let $\mathbb{1}$ denote an all-one vector with appropriate dimension in the rest of this paper. We impose the following assumption.

Assumption 2. *The sequence of nonnegative random matrices* $\{W_t \in \mathbb{R}^{N\times N}\}_{t\geq 0}$ *satisfies*
(2.1) $W_t\mathbb{1} = \mathbb{1}$ *and* $\mathbb{1}^\top\mathbb{E}[W_t] = \mathbb{1}^\top$. *There exists a constant* ϵ *such that for any entry* $W_t(i,j) > 0$, $W_t(i,j) \geq \epsilon$.
(2.2) $W_t(i,j) = 0$ *if* $(i,j) \notin \mathcal{E}_t$.
(2.3) The spectral norm of $\mathbb{E}[W_t^\top(\mathbf{I} - \mathbb{1}\mathbb{1}^\top/N)W_t]$ *is strictly smaller than 1, where* $\mathbf{I}$ *is an identity matrix of size* N.
(2.4) Given the σ*-algebra* $\mathcal{F}_t$ *generated by the random variables before time* t, $\Pr\{W_t|\mathcal{F}_t\}\Pr\{c^i_{t+1}, g^i_{t+1}|\mathcal{F}_t\} = \Pr\{W_t, c^i_{t+1}, g^i_{t+1}|\mathcal{F}_t\}, \forall i \in \mathcal{N}$.

V. Distributed RL algorithm for CMAMDP

In this section, we develop the distributed RL algorithm to solve the problem (5)-(6). Then, the theoretical result is derived to show the convergence of this algorithm.

A. The Developed Algorithm

We first define the global action-value functions under policy π_θ as

$$Q^c_\theta(s,a) = \mathbb{E}\left[\sum_t^\infty (\bar{c}_{t+1} - J(\theta))\middle| s_0 = s, a_0 = a, \pi_\theta\right], \quad (8a)$$

$$Q^g_\theta(s,a) = \mathbb{E}\left[\sum_t^\infty (\bar{g}_{t+1} - G(\theta))\middle| s_0 = s, a_0 = a, \pi_\theta\right], \quad (8b)$$

where $\bar{c}_t = \bar{c}(s_t, a_t)$ and $\bar{g}_t = \bar{g}(s_t, a_t)$. For the constrained problem (5)-(6), the dual problem is defined as

$$\begin{aligned} \max_\lambda \min_\theta L(\theta,\lambda) &= J(\theta) + \lambda(G(\theta) - \bar{G}) \\ \text{s.t. } \lambda &\geq 0, \end{aligned} \quad (9)$$

We extend the multi-agent policy gradient theorem in [10] to characterize the gradient of $L^\lambda(\theta)$, i.e. the Lagrangian function $L(\theta,\lambda)$ with λ fixed.

Lemma 1. *Under Assumption 1, the gradient of* $L^\lambda(\theta)$ *w.r.t.* θ^i *is given by*

$$\nabla_{\theta^i} L^\lambda(\theta) = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[A^\lambda_\theta(s,a)\nabla_{\theta^i}\ln\pi^i_{\theta^i}(s,a^i)\right], \quad (10)$$

where $A^\lambda_\theta(s,a) = Q^c_\theta(s,a) + \lambda Q^g_\theta(s,a)$.

Proof. Follows Theorem 3.1 in [10] directly. □

Based on the fact that the forecast is expected to be more accurate for the closer time instant, we take place Q^c_θ and Q^g_θ with the discounted cumulative cost

$$\tilde{Q}^c_\theta(s,a) = \mathbb{E}\left[\sum_t^\infty \gamma^t \bar{c}_{t+1}\middle| s_0 = s, a_0 = a, \pi_\theta\right], \quad (11a)$$

$$\tilde{Q}^g_\theta(s,a) = \mathbb{E}\left[\sum_t^\infty \gamma^t \bar{g}_{t+1}\middle| s_0 = s, a_0 = a, \pi_\theta\right], \quad (11b)$$

where $\gamma \in [0,1)$ is the discount factor. This would allow us to reduce variance by down-weighting costs corresponding delayed effects [31]. The values of $\tilde{Q}^c_\theta(s,a)$ and $\tilde{Q}^g_\theta(s,a)$ $\forall s \in \mathcal{S}, a \in \mathcal{A}$ could be obtained by solving the Bellman equations [32]

$$\tilde{Q}^c_\theta(s,a) = \bar{c}(s,a) + \gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)V^c_\theta(s'), \quad (12a)$$

$$\tilde{Q}^g_\theta(s,a) = \bar{g}(s,a) + \gamma\sum_{s'\in\mathcal{S}} P(s'|s,a)V^g_\theta(s'), \quad (12b)$$

where

$$V^c_\theta(s') = \sum_{a'\in\mathcal{A}} \pi_\theta(a'|s')\tilde{Q}^c_\theta(s',a'), \quad (13a)$$

$$V^g_\theta(s') = \sum_{a'\in\mathcal{A}} \pi_\theta(a'|s')\tilde{Q}^g_\theta(s',a'), \quad (13b)$$

However, due to the large state and action space in our problem, it is not applicable to store a table for each (s,a) pair. Thus, we use linear functions $Q^c_{\omega^i}(s,a) = \phi(s,a)^\top\omega^i$ and $Q^g_{\psi^i}(s,a) = \phi(s,a)^\top\psi^i$, where $\phi(\cdot,\cdot) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^K$ is the feature function and $\omega^i \in \mathbb{R}^K$ and $\psi^i \in \mathbb{R}^K$ are the local parameter vectors stored at agent i to approximate functions $\tilde{Q}^c_\theta$ and $\tilde{Q}^g_\theta$. For the feature matrix of the entire state-action space $\mathcal{S}\times\mathcal{A}$, $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times K}$, we make the following assumption.

Assumption 3. *The feature vector* $\phi(s,a)$ *is bounded,* $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. $K \leq |\mathcal{S}||\mathcal{A}|$ *and* Φ *is with full column rank.*

Algorithm 1 The Distributed RL Algorithm

1: **repeat**
2: **for all** agent $i \in \mathcal{N}$ **do**
3: observe state s_{t+1}
4: take action $a^i_{t+1} \sim \pi^i_{\theta^i}(s_{t+1})$
5: **end for**
6: observe joint actions $a_{t+1} = (a^1_{t+1}, ..., a^N_{t+1})$
7: **for all** agent $i \in \mathcal{N}$ **do**
8: **Critic Step**
9: $\tilde{\mu}^i_t \leftarrow (1-\beta_{1,t})\mu^i_t + \beta_{1,t}g^i_t$
10: $\delta^i_t \leftarrow c^i_{t+1} + \gamma Q^c_{\omega^i_t}(s_{t+1}, a_{t+1}) - Q^c_{\omega^i_t}(s_t, a_t)$
11: $\Delta^i_t \leftarrow g^i_{t+1} + \gamma Q^g_{\psi^i_t}(s_{t+1}, a_{t+1}) - Q^g_{\psi^i_t}(s_t, a_t)$
12: $\tilde{\omega}^i_t \leftarrow \omega^i_t + \beta_{1,t}\delta^i_t\phi(s_t, a_t)$
13: $\tilde{\psi}^i_t \leftarrow \psi^i_t + \beta_{1,t}\Delta^i_t\phi(s_t, a_t)$
14: **Actor Step**
15: $A^i_t \leftarrow Q^c_{\omega^i_t}(s_t, a_t) + \lambda^i_t Q^g_{\omega^i_t}(s_t, a_t)$
16: $\theta^i_{t+1} \leftarrow \Gamma_{\theta^i}\left[\theta^i_t - \beta_{2,t}A^i_t\nabla_{\theta^i}\ln\pi^i_{\theta^i}(s_t, a_t)\right]$
17: **Dual Step**
18: $\tilde{\lambda}^i_{t+1} \leftarrow \Gamma_\lambda\left[\lambda^i_t + \beta_{3,t}(\mu^i_t - \bar{G})\right]$
19: **end for**
20: **Consensus Step**
21: **for all** $i \in \mathcal{N}$ **do**
22: $\mu^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\mu}^j_t$
23: $\omega^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\omega}^j_t$
24: $\psi^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\psi}^j_t$
25: $\lambda^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\lambda}^j_t$
26: **end for**
27: **until** Max loop number

With the above analysis, we develop Algorithm 1, which is composed of four main steps. In the critic step, μ^i is to estimate the expectation of the restrictive global average cost, i.e. $\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}[\bar{g}(s,a)]$, which is updated in lines 9 and 22. The local TD-errors, δ^i_t and Δ^i_t, are calculated in lines 10-11 to update the temporal local parameters, $\tilde{\omega}^i$ and $\tilde{\psi}^i$, in lines

12-13. In the actor step, the local policy parameter is driven towards the direction to minimize $L^{\lambda_t^i}(\theta)$ in lines 15-16 where Γ_{θ^i} is a projection to the convex and compact set Θ^i. In the dual step, the agent processes a dual ascent for $\tilde{\lambda}^i$ based on the local estimator μ_t^i in line 18 where Γ_λ is a projection to $[0, \lambda_{max}]$. In the consensus step, each agent receives $\tilde{\mu}^j$, $\tilde{\omega}^j$, $\tilde{\psi}^j$, and $\tilde{\lambda}^j$ from the agents in $\{j|(j,i) \in \mathcal{E}_t, \forall j \in \mathcal{N}\}$ and update its local μ^i ω^i ,ψ^i, and λ^i with the communication weight matrix W_t in lines 22-25. We then make the following assumption for the step sizes, $\beta_{1,t}$, $\beta_{2,t}$, and $\beta_{3,t}$.

Assumption 4. *The step size sequences* $\{\beta_{1,t}\}_{t\geq 0}$, $\{\beta_{2,t}\}_{t\geq 0}$, *and* $\{\beta_{3,t}\}_{t\geq 0}$ *satisfy*

$$\sum_{t=0}^{\infty} \beta_{1,t} = \sum_{t=0}^{\infty} \beta_{2,t} = \sum_{t=0}^{\infty} \beta_{3,t} = \infty \tag{14}$$

$$\sum_{t=0}^{\infty} \beta_{1,t}^2 + \beta_{2,t}^2 + \beta_{3,t}^2 < \infty \tag{15}$$

$$\lim_{t\to\infty} \frac{\beta_{2,t}}{\beta_{1,t}} = \lim_{t\to\infty} \frac{\beta_{3,t}}{\beta_{2,t}} = 0 \tag{16}$$

$$\lim_{t\to\infty} \frac{\beta_{1,t+1}}{\beta_{1,t}} = \lim_{t\to\infty} \frac{\beta_{2,t+1}}{\beta_{2,t}} = \lim_{t\to\infty} \frac{\beta_{3,t+1}}{\beta_{3,t}} = 1 \tag{17}$$

B. Theoretical Results

We use similar techniques in [9] and [10] to analyze the convergence of Algorithm 1. A standard method to analyze this algorithm with three time scales is to derive from the fastest one [33], [34], which is the critic step.

For the critic step, we consider θ_t with fixed value θ due to Assumption 4 as $t \to \infty$. We first introduce several lemmas that could be directly derived from [10].

Lemma 2. *Under Assumptions 1, 2, and 4, for any policy* π_θ, *with* $\{\mu_t^i\}$ *generated by Algorithm 1, we have* $\forall i \in \mathcal{N}, \lim_{t\to\infty} \mu_t^i = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}[\bar{g}(s,a)]$.

Proof. See Theorem 4.10 in [10]. □

Lemma 2 implies that all the local μ_t^i achieve consensus and track the global restrictive cost $\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}[\bar{g}(s,a)]$ of the current policy π_θ. Then we analyze the convergence of ω_t^i. The convergence of ψ_t^i follows directly since they are updated in the same manner.

We define $\omega_t = [(\omega_t^1)^\top, (\omega_t^2)^\top, ..., (\omega_t^N)^\top]^\top \in \mathbb{R}^{KN}$, $\delta_t = [\delta_t^1, \delta_t^2, ..., \delta_t^N]^\top \in \mathbb{R}^N$, $\bar{\omega}_t = N^{-1}\sum_{i\in\mathcal{N}} \omega_t^i$ and

$$\omega_t^\perp = \omega_t - \mathbb{1} \otimes \bar{\omega}_t,$$

where $\mathbb{1}$ is the all-one vector of dimension N, and $\otimes$ is the Kronecker product.

Lemma 3. *Under Assumptions 1, 2, and 4, we have a.s.*

$$\sup_t \mathbb{E}[||\beta_{1,t}^{-1}\omega_t^\perp||^2] < \infty,$$

Proof. Follows the proof of Lemma 5.3 in [10]. □

We then define $D_\theta = diag[d_\theta(s)\pi_\theta(s,a), s \in \mathcal{S}, a \in \mathcal{A}] \in \mathbb{R}^{|S||A|\times|S||A|}$. $R^c = [\bar{c}(s,a), s \in \mathcal{S}, a \in \mathcal{A}]^\top \in \mathbb{R}^{|S||A|}$. $P_\theta = [P(s'|s,a)\pi_\theta(s',a'), (s,a) \in \mathcal{S}\times\mathcal{A}, (s',a') \in \mathcal{S}\times\mathcal{A}] \in \mathbb{R}^{|S||A|\times|S||A|}$ is the transition probability matrix. Operator

$$T_\theta^c(\Phi\omega) = R^c + \gamma P_\theta \Phi\omega.$$

Theorem 1. *Under Assumptions 1, 2, 3, and 4, we have a.s.* $\lim_{t\to\infty} \omega_t^i = \omega_\theta$, *where* ω_θ *is the unique solution to*

$$\Phi^\top D_\theta[T_\theta^c(\Phi\omega) - \Phi\omega] = 0. \tag{18}$$

Proof. Lemma 3 implies that $\lim_{t\to\infty} \omega_t^\perp = 0$, so we only need to show the convergence of $\{\bar{\omega}_t\}$. We define the operator $\langle x\rangle = \frac{1}{N}(\mathbb{1}^\top \otimes \mathbf{I})x$, where $\mathbf{I} \in \mathbb{R}^{K\times K}$ is the identity matrix. Thus, $\bar{\omega}_t = \langle\omega_t\rangle$. According to the update in lines 10, 12, and 23 in Algorithm 1, the iteration of $\bar{\omega}_t$ is with the form

$$\bar{\omega}_{t+1} = \langle(W_t \otimes \mathbf{I})(\mathbb{1} \otimes \bar{\omega}_t + \omega_t^\perp + \beta_{1,t}y_{t+1})\rangle \tag{19a}$$

$$= \bar{\omega}_t + \beta_{1,t}\langle(W_t \otimes \mathbf{I})(y_{t+1} + \beta_{1,t}^{-1}\omega_t^\perp)\rangle \tag{19b}$$

$$= \bar{\omega}_t + \beta_{1,t}\langle(W_t \otimes \mathbf{I})y_{t+1}\rangle, \tag{19c}$$

where $y_{t+1} = \delta_t \otimes \phi_t$ and $\phi_t = \phi(s_t, a_t)$ for simplicity. The third equality holds because $\langle\omega_t^\perp\rangle = 0$ Then, we have

$$\bar{\omega}_{t+1} = \bar{\omega}_t + \beta_{1,t}\mathbb{E}[\langle\delta_t\rangle\phi_t|\mathcal{F}_{1,t}] + \beta_{1,t}\zeta_{1,t+1}, \tag{20}$$

where $\zeta_{1,t+1} = \langle(W_t \otimes \mathbf{I})y_{t+1}\rangle - \mathbb{E}[\langle\delta_t\rangle\phi_t|\mathcal{F}_{1,t}]$, which could be easily verified as a martingale difference sequence since $\mathbb{1}^\top\mathbb{E}[W_t] = \mathbb{1}^\top$. Also,

$$\mathbb{E}[||\zeta_{1,t+1}||^2|\mathcal{F}_{1,t}] \leq L(1 + ||\bar{\omega}_t||^2) \tag{21}$$

for some constant L, which is proven in [10]. With the definition in Lemma 5 shown in Appendix A,

$$h_\infty(\omega) = \Phi^\top D_\theta(\gamma P_\theta - \mathbf{I})\Phi\omega. \tag{22}$$

It could be verified that $\Phi^\top D_\theta(\gamma P_\theta - \mathbf{I})\Phi$ is negative definite [9]. Thus, $\dot{\omega} = h_\infty(\omega)$ is with the origin as its unique globally asymptotically stable equilibrium. We conclude that the ODE

$$\dot{\omega} = \Phi^\top D_\theta(\gamma P_\theta - \mathbf{I})\Phi\omega + \Phi^\top D_\theta R^c \tag{23}$$

captures the asymptotic behavior of $\{\bar{\omega}_t\}_{t\geq 0}$ by Lemma 5, which is with the unique globally asymptotically stable equilibrium ω^* such that

$$\Phi^\top D_\theta[T_\theta^c(\Phi\omega^*) - \Phi\omega^*] = 0. \tag{24}$$

□

Theorem 1 concludes that all the local ω_t^i achieve consensus to the same limiting point which is the solution to

$$\min_\omega ||\Phi\omega - \Pi T_\theta^c(\Phi\omega)||_{D_\theta}^2, \tag{25}$$

where Π is the operator that projects a vector to the space spanned by the columns of Φ and $||\cdot||_{D_\theta}^2$ denotes the Euclidean norm weighted by the matrix D_θ [10]. This result is also suitable for $\{\psi_t^i\}_{t\geq 0}$.

For the actor step, we consider that λ_t is fixed as λ. We also define $Q_{\omega_\theta}^c(s,a) = \phi(s,a)^\top\omega_\theta$ where ω_θ is the solution of (18), and so is for $Q_{\psi_\theta}^g$. Then, we have the following lemma.

Lemma 4. *Under Assumptions 1, 3, and 4, suppose that there is a compact set ${\Theta_\lambda^i}^*$ as the asymptotically stable equilibrium of ODE*

$$\dot{\theta}^i = \hat{\Gamma}_\theta \left[-\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta} \left[A_\theta^\lambda(s,a) \nabla_{\theta^i} \ln \pi_{\theta^i}(s,a) \right] \right], \quad (26)$$

where $A_\theta^\lambda(s,a) = Q_{\omega_\theta}^c(s,a) + \lambda Q_{\psi_\theta}^g(s,a)$ for each agent i. Then the policy parameters $\{\theta_t^i\}_{t\geq 0}$ generated by Algorithm 1 converges almost surely to ${\Theta_\lambda^i}^*$, $\forall i \in \mathcal{N}$[1].

Proof. Follows Theorem 4.7 in [10] directly. □

Lemma 4 implies that the policy parameter converges to a stationary point which is a local minimum in sense of $\hat{\Gamma}_\theta \left[-\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta} \left[A_\theta^\lambda(s,a) \nabla_{\theta^i} \ln \pi_{\theta^i}(s,a) \right] \right] = 0$ with λ fixed. In the following, we denote θ_λ as the stationary point which corresponds to policy π_λ^i and $d_\lambda = d_{\theta_\lambda}$ as the stationary distribution for any λ. We then have the following results for the dual step, which can be extended from [9].

Theorem 2. *Under Assumption 4, $\forall i \in \mathcal{N}$, with $\{\lambda_t^i\}_{t\geq 0}$ generated by Algorithm 1, $\lim_{t\to\infty} \lambda_t^i = \bar{\lambda}_t$ a.s. with $\bar{\lambda}_t = \frac{1}{N}\mathbb{1}^\top \lambda_t$*

Proof. See Appendix B. □

Theorem 2 shows that all the local dual variables will converge to consensus. With Lemma 2 showing that all the local μ_t^i converge to consensus, we have the following theorem.

Theorem 3. *Under Assumptions 1, 3, and 4, suppose that for any λ there is a compact set Λ^* as the asymptotically stable equilibrium of ODE*

$$\dot{\lambda} = \hat{\Gamma}_\lambda (\mathbb{E}_{s\sim d_\lambda, a\sim\pi_\lambda} [\bar{g}(s,a)] - \bar{G}), \quad (27)$$

Then the dual variable $\{\bar{\lambda}_t\}_{t\geq 0}$ generated by Algorithm 1 converges almost surely to Λ^*.

Proof. See Appendix C. □

Theorem 2 and 3 show that all the local sequences $\{\lambda_t^i\}_{t\geq 0}$ converge to the same local maximum of the "dual problem" where the primal problem reaches a local minimum. We also have the following propositions.

VI. Practical Implementation

In this section, we develop a modified version of Algorithm 1 to optimize the feature function $\phi(\cdot,\cdot)$ based on deep neural networks, which also preserves the private information about local state and action.

A. The Modified Algorithm

We first show the information flow among the MGs in Fig 2. The MGs communicate with the central feature collector to send (through the blue lines) local information and acquire (through the red lines) the global information on the states and actions. In the follows, we develop the algorithm to encrypt the local information. The MGs also share their local parameters through the communication network that satisfies Assumption 2, i.e. the green lines.

[1]The convergence to a set is the same as Theorem 2.1, Chapter 5 in [33]. And so is the result in rest of this paper.

In this system, each MG agent maintains two deep neural networks with parameter ν^i and υ^i to represent the local feature functions $\phi_{\nu^i}^{c,i}(s^i,a^i)$ and $\phi_{\upsilon^i}^{g,i}(s^i,a^i) : \mathcal{S}^i \times \mathcal{A}^i \to K^i$. These local feature functions extract features and also encrypt the local states and actions. Moreover, each agent has a local lossless encryption function $f^i(\cdot) : \mathcal{S}^i \to \mathcal{S}^i$ that encrypts the local state s_t^i. At each time t, all the agents upload their local vectors $\phi_{\nu^i,t}^{c,i} = \phi_{\nu^i}^{c,i}(s_t^i,a_t^i)$, $\phi_{\upsilon^i,t}^{g,i} = \phi_{\upsilon^i}^{c,i}(s_t^i,a_t^i)$, and $f_t^i = f^i(s_t^i)$ to the central feature collector through the blue lines. Then, the feature collector combine them to form the global feature vector $\phi_t^c(s_t,a_t) = \bigcup_{i\in\mathcal{N}} \phi_{\nu^i,t}^{c,i}$, $\phi_t^g(s_t,a_t) = \bigcup_{i\in\mathcal{N}} \phi_{\upsilon^i,t}^{g,i}$ and $f_t(s_t) = \bigcup_{i\in\mathcal{N}} f_t^i$, which are broadcast to all agents through the red lines. In the following, we show that ϕ_t^c and ϕ_t^g will be used to estimate the action-value functions, while f_t will be fed into the policy network. Except for the communication between the agents and the central collector, there is also a communication network among the agents, which is the one demanded by Algorithm 1. Since that the locations of the MGs do not change and the communication network is stable, we use time-invariant communication matrix, i.e. $W_t = W$, which satisfies Assumption 2. The MGs transmit local parameters $\tilde{\mu}_t^i$, $\tilde{\omega}_t^i$, $\tilde{\psi}_t^i$, and $\tilde{\lambda}_t^i$ through the green lines which is represented by matrix W.

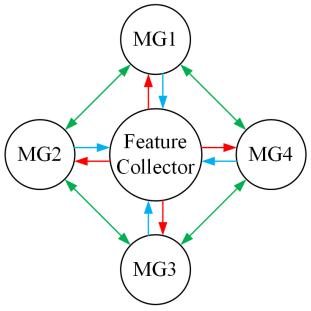

Fig. 2. The information flow among the MGs.

The key issue is that the global action-value could be represented through a linear function of the extracted feature ϕ_t. Thus, we propose to add a feature training process to Algorithm 1. Moreover, to update the parameters at every time step t is with low efficiency. So, we modify Algorithm 1 to collect data and update the parameters every M time steps to reduce variance. The modified algorithm is shown in Algorithm 2.

Algorithm 2 is mainly composed of four parts. At each time t, each agent stores the current data sample $\langle \phi_t, f_t, s_t^i, a_t^i, c_{t+1}^i, g_{t+1}^i, \phi_{t+1}, f_{t+1} \rangle$ into the local buffer. After every M time steps, each agent updates the parameters with the stored M data samples and then clears the local buffer. In the feature step, each agent updates its local parameters ν_k^i and υ_k^i in lines 23-28. The key point of the feature vector is that it could be applied to estimate the action-value function through linear functions, i.e. $\phi^c(s,a)^\top \omega^i$ and $\phi^g(s,a)^\top \psi^i$. Thus, we tune ν^i and υ^i by connecting $\phi_{\nu^i}^{c,i}(s,a)$ and $\phi_{\upsilon^i}^{g,i}(s,a)$ to a linear function with weight vectors $\nu^{i'}$ and $\upsilon^{i'}$ in (29) to estimate the local cumulative discounted costs $\mathbb{E}[\sum_{\tau=1}^{M'} \gamma^\tau c_{t+\tau}^i | s_t = s, a_t = a]$ and $\mathbb{E}[\sum_{\tau=1}^{M'} \gamma^\tau g_{t+\tau}^i | s_t =$

Algorithm 2 Modified RL Algorithm

for $k = 1 : Maxloop$ **do**
  for $t = (k-1)M : kM$ **do**
    for all agent $i \in \mathcal{N}$ **do**
      observes local state s^i_{t+1} and upload f^i_{t+1}
      takes action $a^i_{t+1} \sim \pi^i_{\theta^i_k}(f_{t+1})$
      uploads $\phi^{c,i}_{\nu^i_k,t}$ and $\phi^{g,i}_{\upsilon^i_k,t}$
      receives instant costs c^i_{t+1} and g^i_{t+1}
      stores data sample $\langle \phi^c_t, \phi^g_t, f_t, s^i_t, a^i_t, c^i_{t+1}, g^i_{t+1}, \phi^c_{t+1}, \phi^g_{t+1}, f_{t+1}\rangle$
      $\tilde{\mu}^i_t \leftarrow (1-\beta_{1,t})\mu^i_t + \beta_{1,t} g^i_{t+1}$
    end for
    for all $i \in \mathcal{N}$ **do**
      $\mu^i_{t+1} \leftarrow \sum_{j=1}^N W(i,j)\tilde{\mu}^j_t$
    end for
  end for
  for all $i \in \mathcal{N}$ **do**
    Feature Step
    $\nabla_\nu \leftarrow \sum_{m=1}^{M-M'} h^{c,i}_k(s^i_m)\nabla_{\nu^i} h^{c,i}_k$
    $\nabla_{\nu'} \leftarrow \sum_{m=1}^{M-M'} h^{c,i}_k(s^i_m)\nabla_{\nu^{i\prime}} h^{c,i}_k$
    $\nabla_\upsilon \leftarrow \sum_{m=1}^{M-M'} h^{g,i}_k(s^i_m)\nabla_{\upsilon^i} h^{g,i}_k$
    $\nabla_{\upsilon'} \leftarrow \sum_{m=1}^{M-M'} h^{g,i}_k(s^i_m)\nabla_{\upsilon^{i\prime}} h^{g,i}_k$
    $\nu^i_k \leftarrow \nu^i_k + \frac{\beta_{1,k}}{M-M'}\nabla_\nu, \nu^{i\prime}_k \leftarrow \nu^{i\prime}_k + \frac{\beta_{1,k}}{M-M'}\nabla_{\nu'}$
    $\upsilon^i_k \leftarrow \upsilon^i_k + \frac{\beta_{1,k}}{M-M'}\nabla_\upsilon, \upsilon^{i\prime}_k \leftarrow \upsilon^{i\prime}_k + \frac{\beta_{1,k}}{M-M'}\nabla_{\upsilon'}$
    Critic Step
    $\delta^i_k \leftarrow \frac{1}{M}\sum_{m=1}^M (c^i_{m+1} + \gamma {\omega^i_k}^\top \phi^c_{m+1} - {\omega^i_k}^\top \phi^c_m)\phi^c_m$
    $\Delta^i_k \leftarrow \frac{1}{M}\sum_{m=1}^M (g^i_{m+1} + \gamma {\psi^i_k}^\top \phi^g_{m+1} - {\psi^i_k}^\top \phi^g_m)\phi^g_m$
    $\tilde{\omega}^i_k \leftarrow \omega^i_k + \beta_{1,k}\delta^i_k$
    $\tilde{\psi}^i_k \leftarrow \psi^i_k + \beta_{1,k}\Delta^i_k$
    Actor Step
    $A^i_m \leftarrow {\omega^i_k}^\top \phi^c_m + \lambda^i_k {\psi^i_k}^\top \phi^g_m, \forall m = 1, 2, ..., M$
    $\theta^i_{t+1} \leftarrow \Gamma_\theta\left[\theta^i_k - \frac{\beta_{2,k}}{M}\sum_{m=1}^M A^i_m \nabla_{\theta^i} \ln \pi^i_{\theta^i}(f_m, a^i_m)\right]$
    Dual Step
    $\tilde{\lambda}^i_{k+1} \leftarrow \Gamma_\lambda\left[\lambda^i_k + \beta_{3,t}(\mu^i_t - \bar{G})\right]$
  end for
  Consensus Step
  for all $i \in \mathcal{N}$ **do**
    $\omega^i_{k+1} \leftarrow \sum_{j=1}^N W(i,j)\tilde{\omega}^j_k$
    $\psi^i_{k+1} \leftarrow \sum_{j=1}^N W(i,j)\tilde{\psi}^j_k$
    $\lambda^i_{k+1} \leftarrow \sum_{j=1}^N W(i,j)\tilde{\lambda}^j_k$
  end for
end for

$s, a_t = a]$. During the kth iteration, the local objective of the feature step is

$$\min_{\nu^i,\nu^{i\prime},\upsilon^i,\upsilon^{i\prime}} \frac{1}{2}\sum_{m=1}^{M-M'} (h^{c,i}_k(m))^2 + (h^{g,i}_k(m))^2, \tag{28}$$

where

$$h^{c,i}_k(m) = \phi^{c,i}_{\nu^i_k}(s_m, a_m)^\top \nu^{i\prime} - \sum_{\tau=1}^{M'} \gamma^\tau c^i_{m+\tau} \tag{29a}$$

$$h^{g,i}_k(m) = \phi^{g,i}_{\upsilon^i_k}(s_m, a_m)^\top \upsilon^{i\prime} - \sum_{\tau=1}^{M'} \gamma^\tau g^i_{m+\tau}. \tag{29b}$$

$\phi^{c,i}_{\nu^i_k}(s_m, a_m)^\top \nu^{i\prime}$ and $\phi^{g,i}_{\upsilon^i_k}(s_m, a_m)^\top \upsilon^{i\prime}$ are linear regression for the cumulative discounted costs and M' is the finite time horizon which is large enough to omit the approximation error against infinite cumulative discounted costs. The critic, actor, and dual step are similar with Algorithm 1, while the parameters are updated with a batch of data instead of one data sample at each time step.

We set the local policy $\pi^i_{\theta^i}(f(s))$ as the normal distribution $Normal(\mu^i_{\theta^i}(f(s)), \sigma\mathbf{I})$, where $\mu^i_{\theta^i}(\cdot) : f(\mathcal{S} \to \mathbb{R}^{|\mathcal{A}^i|})$ determines the mean value and $\sigma\mathbf{I}$ is the covariance matrix and $\mathbf{I}$ is the identity matrix of dimension $|\mathcal{A}^i|$.

All the local encryption functions $f^i(\cdot), \forall i \in \mathcal{N}$ are selected as functions by adding a time-invariant noise to the state s^i, which could not be recovered from $f^i(s^i)$ by other agents.

B. Neural Network Structure

To deal with the large state space $\mathcal{S}$, we propose the neural network structure with 1D convolution layers. The structures of the neural networks that represent $\phi^i_{\nu^i}(s^i, a^i)(\phi^i_{\upsilon^i}(s^i, a^i))$ and $\mu^i_{\theta^i}(f)$ are shown in Fig. 3, which also shows the flow of the tensors of Algorithm 2.

For the local feature neural network parametrized by ν^i and υ^i (the red box), the local state time series are first fed into the module composed of 1D convolution layers and max-pooling layers [11]. 1D convolution layer could effectively extract features for sequences with small parameter dimension [35]. Then, these features are fed into the fully connected layers together with the non-sequence local state and action. The local features $\phi^{c,t}_t$ and $\phi^{g,i}_t$ are the output.

For the local policy neural network parametrized by θ^i (the green box), each agent first adds noise to local state s^i with the local encryption function f^i such that s^i and $f^i(s^i)$ are in the same shape. Then, the sequence part of the global encrypted feature $f(s)$ is fed into the 1D convolution layers and max-pooling layers. The extracted features are fed into the fully connected layers together with the non-sequence part of $f(s)$. The mean vector $\mu^i_{\theta^i}$ is the output.

VII. Numerical Results

In this section, we carry out comparison experiments to validate the effectiveness of our algorithm and show the advantage over the existing methods.

A. Experimental Settings

Case I. We select the IEEE 33-bus distribution network with 4 MGs shown in Fig. 4 as the test bed. Each MG owns a GT generator, ES unit, PV unit, and non-dispatchable power loads. The simulator is built with rules represented by the constraints in (1d)-(1t). The prediction sequences such as $\tilde{P}^L_{i,t}$ are simulated by adding an error vector in which τth element obeys Gaussian distribution $Normal(0, \sigma^e_\tau)$ to the actual values, where σ^e_τ increases as τ increases. The time horizon of predictions and reference LMP sequences

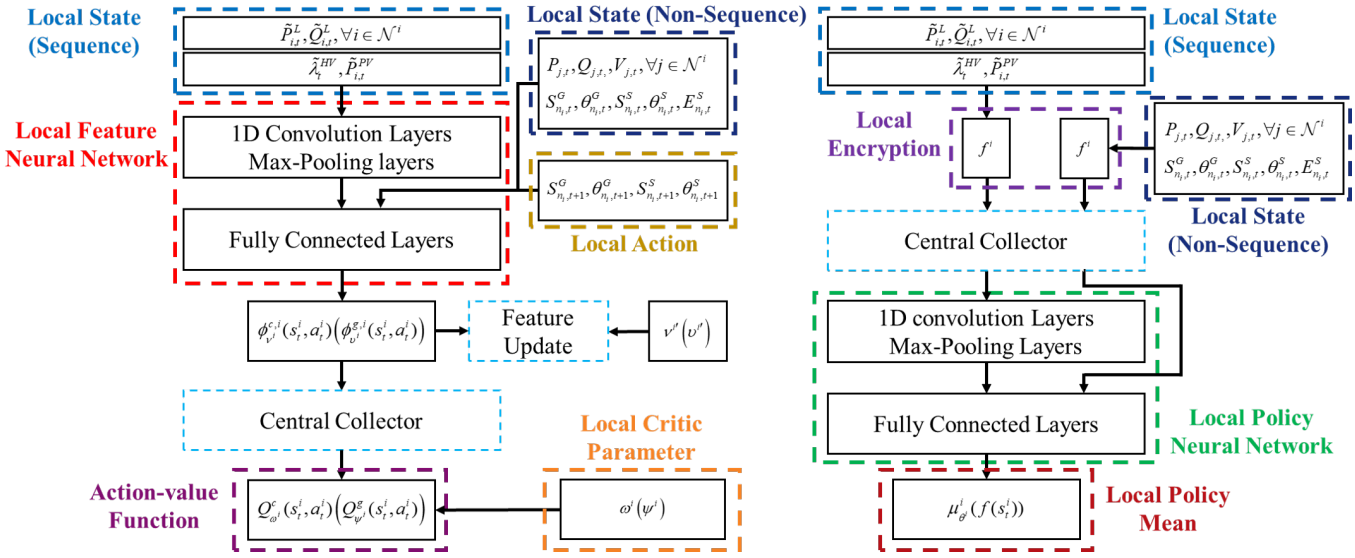

Fig. 3. The structure of the deep neural networks and the information graph.

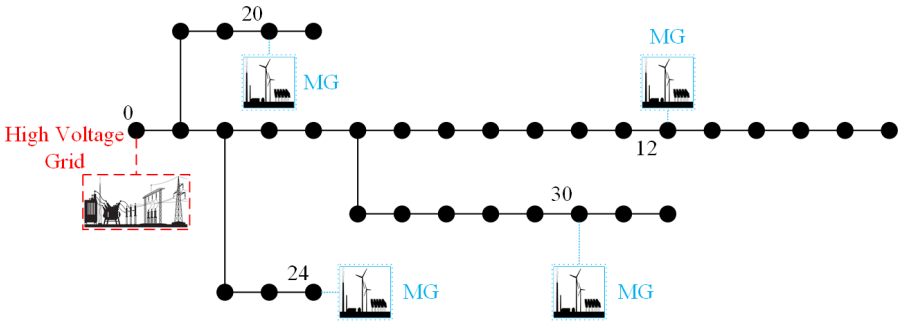

Fig. 4. The 33-bus distribution network with 4 MGs.

is 24 hours and divided into time slots of 5 minutes. Thus, $T = 24 \times 12 = 288$ and $\Delta t = 5$minutes. The load data, PV data, and real-time electric price of the high voltage network are from PJM[2].

We test Algorithm 2 with $4\bar{G} = 0.02/0.03/0.05$, respectively. Then, Algorithm 2 is compared against two model-based optimization methods. We first replace the non-linear equality constraints (1s)-(1t) with the Distflow model. Then we transform the original problem (1a)-(1u) to an optimization problem with receding horizon $\mathcal{T} = T$. The centralized interior point method [36] and the on-line ADMM [1] are respectively applied to solve this problem as baselines. Moreover, we compare against the method that deals the voltage violation with a fixed penalizing term [24] instead of the dual process.

Case II. The distribution network in the metropolitan area of Caracas [37] is selected as the test bed. We add 8 MGs to this network. Different from *Case I*, each MG includes several buses with the GT generator, PV unit, ES unit, and loads connected to different buses as shown in Fig. 5. We test Algorithm 2 with $8\bar{G} = 0.03/0.1/0.15$. The model-based optimization methods in *Case I* are also set as the baselines. Other settings and data are the same with *Case I*.

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

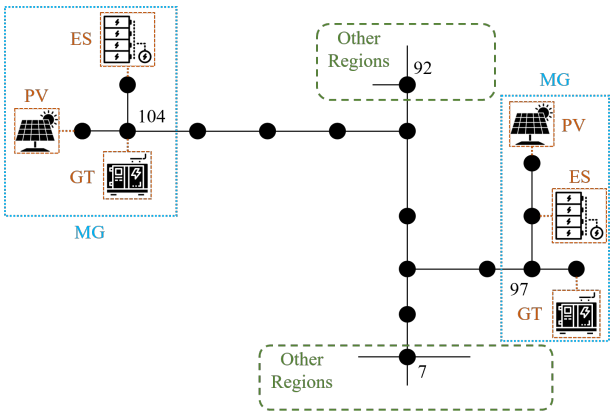

Fig. 5. Part of the 141-bus distribution network with 8 MGs.

B. Experimental Results

For *Case I*, the experimental results of Algorithm 2 and the method with fixed penalizing term are shown in Figs. 6-8. Fig. 6 shows that Algorithm 2 could optimize the global cost to a local minimum. While with the fixed penalizing

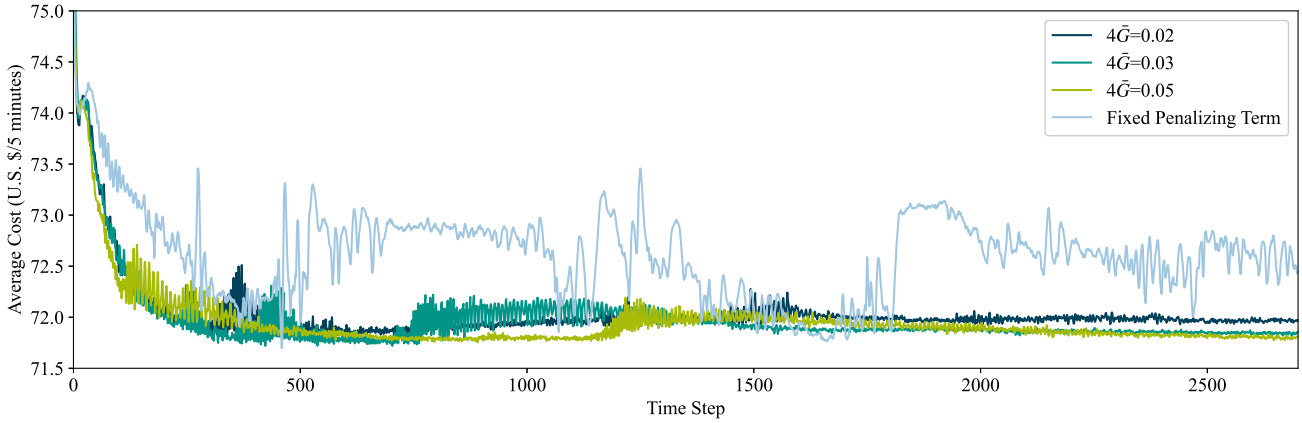

Fig. 6. The average global cost during the training process in *Case I*.

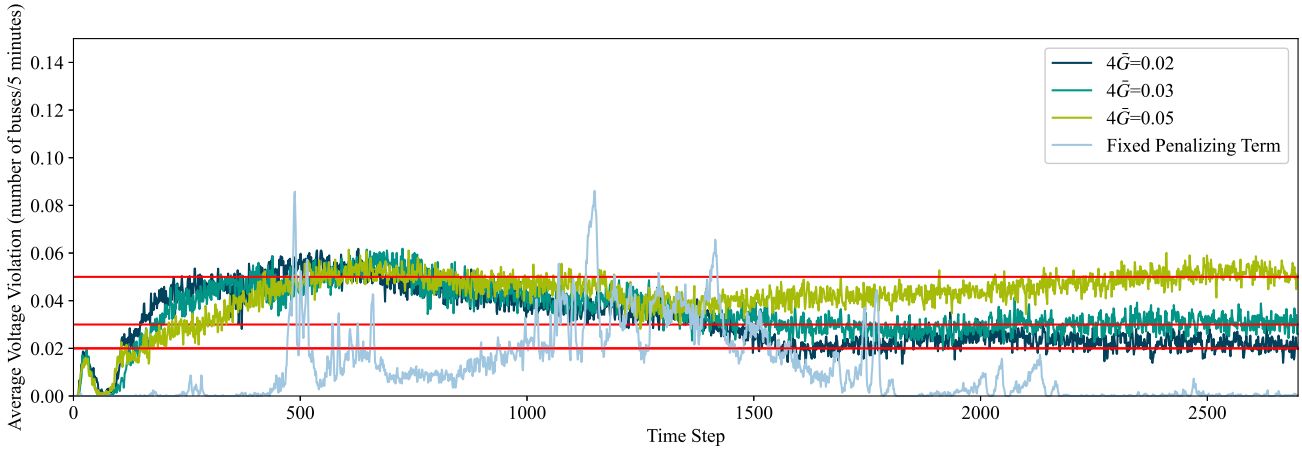

Fig. 7. The frequency of voltage violations during the training process in *Case I*.

term [24], the training process is extremely unstable. This is because when the voltage constraint is violated, the agents will be penalized with an extremely large cost that causes them hard to evaluate the action-value function. The advantage of the fixed penalizing term is that the voltage constraint is satisfied from Fig. 7. However, this would sacrifice the economic benefits and the trade-off is hard to control. With Algorithm 2, we only need to decide the limit of frequency of violations, $\bar{G}$, in advance. Then, the primal-dual update would tune the dual variable to the proper value that penalizes the constraint violation. Fig. 7 also shows that $G(\theta) \approx \bar{G}$ and constraint (6) is activated. We also conclude that when we set smaller $\bar{G}$, the value of the converged dual variable would be greater to penalize the violation, resulting in a larger electricity cost. Because when the bus voltage reaches the limit, the MGs should regulate the facilities that the generated active and reactive power deviate from the most economical plan.

The comparison against the two conventional methods is shown in TABLE I. The centralized method is achieved when the DSO has full control authority of all power equipment in the MGs. The on-line ADMM method proceeds incomplete iterations at each time step. Thus, the centralized method outperforms the on-line ADMM method in both aspects of economy and safety. However, these two methods are based on linear power flow model, which results in intractable influence on the practical nodal voltage. Compared with the two baselines, Algorithm 2 (the first three lines) could regulate the voltage violation frequency around the preset limit in the real scenario. Compared with the centralized method, our algorithm also preserves the privacy of each MG.

TABLE I
COMPARISON WITH THE MODEL-BASED METHODS IN *Case I*

Method	Average Cost (U.S. $/5 minutes)	Voltage Violation Frequency (buses/5 minutes)
$4\bar{G} = 0.02$	71.97	0.0198
$4\bar{G} = 0.03$	71.86	0.0299
$4\bar{G} = 0.05$	71.81	0.0503
On-line ADMM	71.71	0.0786
Centralized method	70.76	0.0625

For *Case II*, the training processes of Algorithm 2 are shown in Figs. 9 and 10. The results in Fig. 10 show that Algorithm 2 converges to numerically feasible policies in all three cases with "$8\bar{G} = 0.03/0.1/0.15$". In *Case II*, we set lager upper bounds for $G(\theta)$ in *Case II* because the bus number in *Case II* is more than 4 times larger than that in *Case I*. The

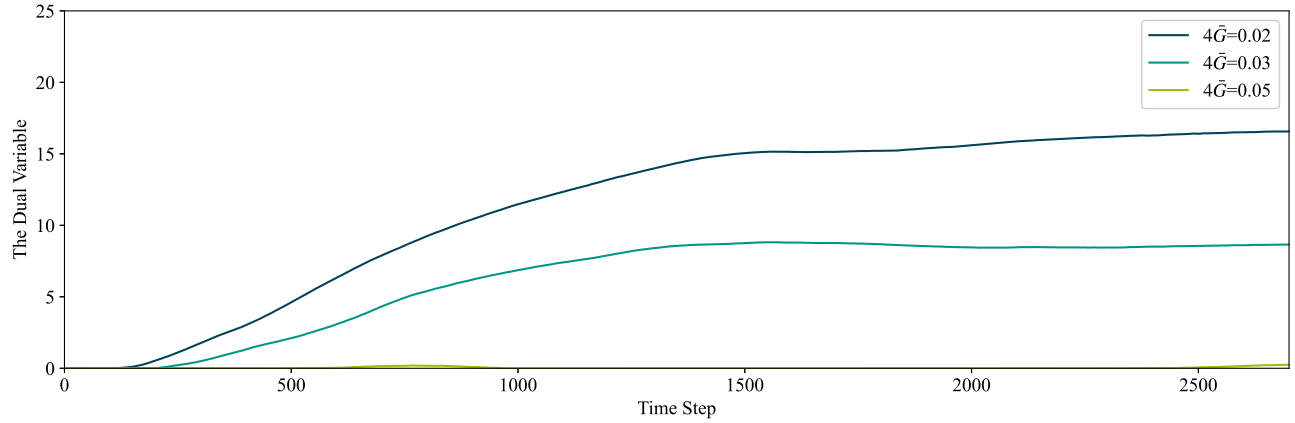

Fig. 8. The average dual value during the training process in *Case I*.

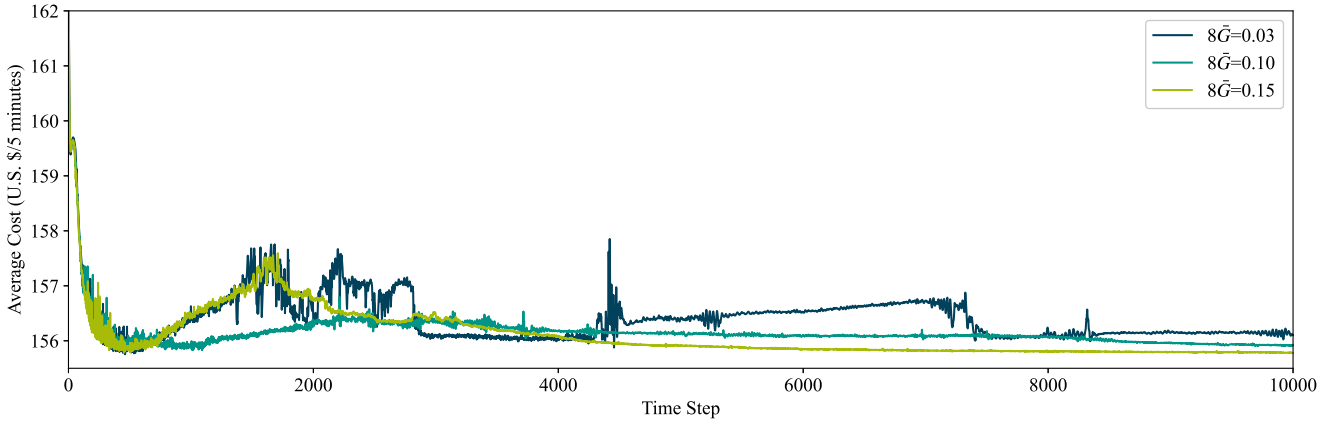

Fig. 9. The average global cost during the training process in *Case II*.

voltage constraint is on the number of abnormal buses, and thus to restrict the policy in *Case II* by the safety standard in *Case I* is too strict. This is reflected by the comparison result of the training processes shown in Fig. 10. In the cases with "$8\bar{G} = 0.1/0.15$", it takes around 6,000 time steps for Algorithm 2 to find numerically feasible policies, while in the case with "$8\bar{G} = 0.03$", it takes around 10,000 time steps. Moreover, compared with *Case I*, the larger system with more MGs induces the longer training time in general, because the problem size of *Case II* is larger than *Case I*. It numerically shows that the problem complexity affects the convergence speed.

For *Case II*, we also compare Algorithm 2 against the two conventional methods with the results shown in TABLE II. It could be found that the frequencies of voltage violation with the conventional methods are higher because the approximate power flow model is utilized. The frequencies with Algorithm 2 are around the preset upper bounds in all the three cases with 1-2% higher economic costs. This finding is similar with that in *Case I*.

In addition, during each time slot, it takes each MG agent only around 0.002 seconds to decide the local action after the global encrypted information f_t is received, which is much smaller than $\Delta t = 5$ minutes. This is because the agent only needs to perform a forward propagation of the neural network at each time. Thus, Algorithm 2 is suitable for on-line dispatch.

TABLE II
COMPARISON WITH THE MODEL-BASED METHODS IN *Case II*

Method	Average Cost (U.S. $/5 minutes)	Voltage Violation Frequency (buses/5 minutes)
$8\bar{G} = 0.03$	156.10	0.0309
$8\bar{G} = 0.10$	155.92	0.0972
$8\bar{G} = 0.15$	155.78	0.1472
On-line ADMM	154.29	0.3055
Centralized method	152.11	0.2743

VIII. CONCLUSION

In this paper, we first develop a distributed RL algorithm to solve the general CMAMDP problems. To further preserve local privacy and improve performance in the distribution system with MGs, we develop the modified algorithm by adding feature modules which also encrypts local state-action pair. The experimental results show that the primal-dual update in our modified algorithm could handle the chance constraint of nodal voltage better than the fixed penalizing term. Moreover, compared against the on-line ADMM method, our algorithm

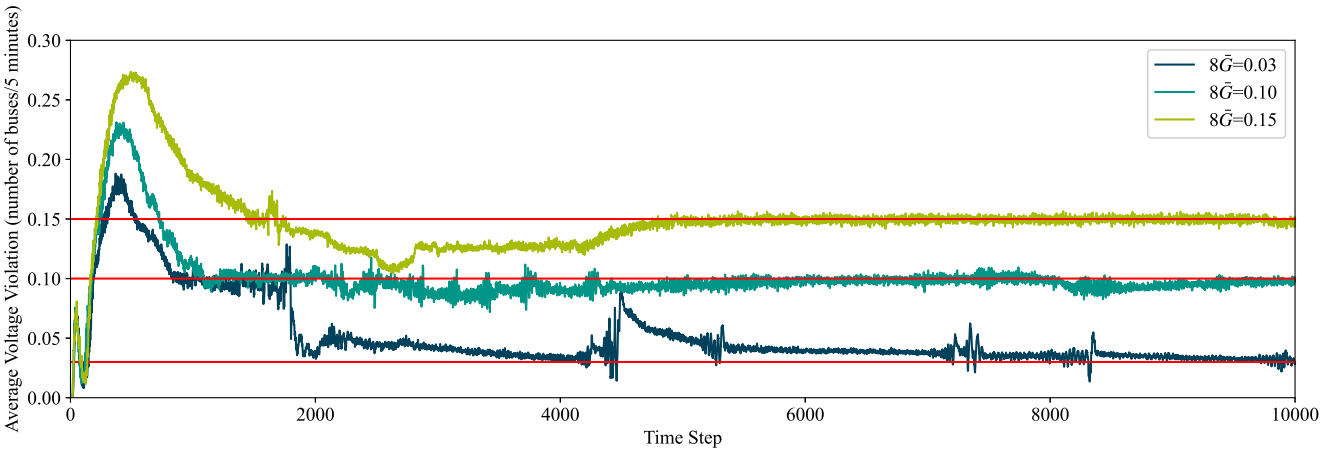

Fig. 10. The frequency of voltage violations during the training process in *Case II*.

could control the frequency of voltage violations while the economic target is only slightly sacrificed. In addition, our algorithm could preserve the local privacy of each MG, and thus is more applicable in the real system than the centralized method.

APPENDIX A
STOCHASTIC APPROXIMATION

The following lemma could be found in [33] and [34]. Consider the n-dimensional stochastic approximation iteration

$$x_{t+1} = x_t + \beta_t[h(x_t, Z_t) + \xi_{1,t+1} + \xi_{2,t+1}], t \geq 0, \quad (30)$$

where $\beta_t > 0$ and $\{Z_t\}_{t\geq 0}$ is a Markov chain on a finite set $\mathcal{Z}$.

Assumption 5. *We make the following assumptions*
(5.1) $h : \mathbb{R}^n \to \mathbb{R}^n$ *is Lipschitz continuous w.r.t.* x_t.
(5.2) $\{\beta_t\}_{t\geq 0}$ *satisfies* $\sum_t \beta_t = \infty$ *and* $\sum_t \beta_t^2 < \infty$
(5.3) $\{\xi_{1,t}\}_{t\geq 0}$ *is a martingale difference sequence, i.e.* $\mathbb{E}[\xi_{1,t+1}|\mathcal{F}_t] = 0$*, satisfying that for some* K_1 *and* $t \geq 0$*,*

$$\mathbb{E}[\xi_{1,t+1}|\mathcal{F}_t] \leq K_1 \cdot (1 + ||x_t||^2).$$

(5.4) The sequence $\{\xi_{2,t}\}_{t\geq 0}$ *is a bounded random sequence with* $\xi_{2,t} \to 0$ *a.s. as* $t \to \infty$.
(5.5) $\{Z_t\}_{t\geq 0}$ *is an irreducible Markov chain with stationary distribution* d.

Then the asymptotic behavior of the iteration (30) is related to the behavior of the solution to the ODE

$$\dot{x} = \bar{h}(x) = \sum_{z\in\mathcal{Z}} d(z)h(x, z). \quad (31)$$

Suppose (31) is with a unique globally asymptotically stable equilibrium x^*, we then have the following lemma.

Lemma 5. *Under Assumption 5, with* $h_\infty(x) = \lim_{c\to\infty} \frac{\bar{h}(cx)}{c}$*, if the ODE* $\dot{y} = h_\infty(y)$ *is with origin as the unique globally asymptotically stable equilibrium, then we have a.s.* $x_t \to x^*$.

Proof. See [33] and [34]. □

APPENDIX B
PROOF OF THEOREM 2

Proof. The proof is extended from Lemma 5.3 in [10]. The difference is that there is a projection in our dual update. We first define

$$\lambda_t^\perp = \lambda_t - \bar{\lambda}_t \mathbb{1}, \quad (32)$$

where $\lambda_t = [\lambda_t^1, ..., \lambda_t^N]^\top$. Then, with

$$h_t = [\mu_t^1 - \bar{G}, ..., \mu_t^N - \bar{G}]^\top \quad (33)$$

we have

$$\begin{aligned}\lambda_{t+1}^\perp &= \lambda_{t+1} - \bar{\lambda}_{t+1}\mathbb{1} \\ &= \Gamma_\lambda[W_t(\lambda_t + \gamma_t h_t)] - \frac{1}{N}\mathbb{1}\mathbb{1}^\top\Gamma_\lambda[W_t(\lambda_t + \gamma_t h_t)] \\ &= (\mathbf{I} - \frac{1}{N}\mathbb{1}\mathbb{1}^\top)\Gamma_\lambda[W_t(\lambda_t + \gamma_t h_t)],\end{aligned} \quad (34)$$

where $\mathbf{I}$ is the identity matrix of size N, and

$$\begin{aligned}&\mathbb{E}[\|\gamma_{t+1}^{-1}\lambda_{t+1}^\perp\|^2|\mathcal{F}_t^\lambda] \\ &\overset{(a)}{=} \frac{\gamma_t^2}{\gamma_{t+1}^2}\cdot\frac{1}{\gamma_t^2}\mathbb{E}[\Gamma_\lambda[W_t(\lambda_t + \gamma_t h_t)]^\top(\mathbf{I} - \frac{1}{N}\mathbb{1}\mathbb{1}^\top) \\ &\quad \Gamma_\lambda[W_t(\lambda_t + \gamma_t g_t)]|\mathcal{F}_t^\lambda] \\ &\overset{(b)}{\leq} \frac{\gamma_t^2}{\gamma_{t+1}^2}\cdot\frac{1}{\gamma_t^2}\mathbb{E}[(W_t(\lambda_t + \gamma_t g_t))^\top(\mathbf{I} - \frac{1}{N}\mathbb{1}\mathbb{1}^\top) \\ &\quad W_t(\lambda_t + \gamma_t h_t)|\mathcal{F}_t^\lambda] \\ &\overset{(c)}{=} \frac{\gamma_t^2}{\gamma_{t+1}^2}\cdot\frac{1}{\gamma_t^2}\mathbb{E}[\|(\mathbf{I} - \frac{1}{N}\mathbb{1}\mathbb{1}^\top)W_t(\lambda_t^\perp + \gamma_t h_t)\|^2|\mathcal{F}_t^\lambda] \\ &\overset{(d)}{\leq} \frac{\gamma_t^2}{\gamma_{t+1}^2}\rho(\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\|^2|\mathcal{F}_t^\lambda] + 2K_1\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\||\mathcal{F}_t^\lambda] + K_1^2),\end{aligned} \quad (35)$$

where (a) is by plugging (34) into $\lambda_t^\perp$ and $(\mathbf{I} - N^{-1}\mathbb{1}\mathbb{1}^\top)^\top(\mathbf{I} - N^{-1}\mathbb{1}\mathbb{1}^\top) = (\mathbf{I} - N^{-1}\mathbb{1}\mathbb{1}^\top)$; (b) is because that $(\mathbf{I} - N^{-1}\mathbb{1}\mathbb{1}^\top)$ is positive definite matrix; (c) is because that $(\mathbf{I} - N^{-1}\mathbb{1}\mathbb{1}^\top)\bar{\lambda}_t\mathbb{1} = 0$; (d) is by Assumption (4.3) where $\rho < 1$ is the upper bound of the spectral norm of $(\mathbf{I} - \frac{1}{N}\mathbb{1}\mathbb{1}^\top)W_t$ and K_1 is the upper bound of $\|\gamma_t h_t\|$ since $\|h_t\|$ is upper

bounded by the assumption of that the cost $g^i(s,a)$ is bounded. Since $\lim_{t\to\infty}\gamma_t^2\gamma_{t+1}^{-2}=1$ and $\rho<1$, there exist some $\varepsilon\in(0,1)$ and large enough t_0 such that $\gamma_t^2\gamma_{t+1}^{-2}\rho\leq(1-\varepsilon)$, for $t\geq t_0$. Thus, we have

$$\begin{aligned}&\mathbb{E}[\|\gamma_{t+1}^{-1}\lambda_{t+1}^{\perp}\|^2|\mathcal{F}_t^{\lambda}]\\ \leq&(1-\varepsilon)(\mathbb{E}[\|\gamma_t^{-1}\lambda_t^{\perp}\|^2|\mathcal{F}_t^{\lambda}]+2K_1\mathbb{E}[\|\gamma_t^{-1}\lambda_t^{\perp}\||\mathcal{F}_t^{\lambda}]+K_1^2)\\ \leq&(1-\frac{\varepsilon}{2})\mathbb{E}[\|\gamma_t^{-1}\lambda_t^{\perp}\|^2|\mathcal{F}_t^{\lambda}]+K_2.\end{aligned}\tag{36}$$

Then, by taking expectation on both sides and induction, we have

$$\mathbb{E}\|\gamma_t^{-1}\lambda_t^{\perp}\|^2\leq(1-\frac{\varepsilon}{2})^{(t-t_0)}\mathbb{E}\|\gamma_{t_0}^{-1}\lambda_{t_0}^{\perp}\|^2+\frac{2K_2}{\varepsilon}<\infty.\tag{37}$$

Therefore, $\sup_t\mathbb{E}\|\gamma_t^{-1}\lambda_t^{\perp}\|^2<\infty$ and under Assumption 4, $\sum_{t=0}^{\infty}\|\lambda_t^{\perp}\|^2$ is finite which yields $\lim_{t\to\infty}\lambda_t^{\perp}=0$, and thus concludes the proof. □

APPENDIX C
PROOF OF THEOREM 3

Proof. The proof is based on Theorem 2.1, in Chapter 5, [33]. With

$$h_t=[\mu_t^1-\bar{G},...,\mu_t^N-\bar{G}]^{\top},\tag{38}$$

recall the dual update

$$\begin{aligned}\lambda_{t+1}=&W_t\Gamma_{\lambda}[\lambda_t+\gamma_t h_t]\\ =&W_t(\lambda_t+\gamma_t h_t+\gamma_t Z_t),\end{aligned}\tag{39}$$

where $Z_t=(Z_t^1,...,Z_t^N)^{\top}$ and $\gamma_t Z_t^i$ is a vector of the shortest Euclidean length needed to take $\lambda_t+\gamma_t h_t$ back to the constraint set $[0,\lambda_{max}]^N$. We then have

$$\begin{aligned}\bar{\lambda}_{t+1}=&\frac{1}{N}\mathbb{1}^{\top}W_t\Gamma_{\lambda}[\lambda_t+\gamma_t h_t]\\ =&\frac{1}{N}\mathbb{1}^{\top}W_t(\lambda_t+\gamma_t h_t+\gamma_t Z_t)\\ =&\bar{\lambda}_t+\gamma_t\mathbb{E}[\bar{g}(s,a)-\bar{G}]+\gamma_t\bar{Z}_t+\gamma_t\zeta_{2,t}+\gamma_t\zeta_{3,t},\end{aligned}\tag{40}$$

where $\gamma_t\bar{Z}_t$ is a scalar of the minimal absolute value needed to take $\bar{\lambda}_t+\gamma_t\bar{h}_t$ back to the constraint set $[0,\lambda_{max}]$

$$\zeta_{2,t}=\frac{1}{N}\mathbb{1}^{\top}(h_t-\mathbb{E}[\bar{g}(s,a)-\bar{G}]\mathbb{1})\tag{41a}$$

$$\zeta_{3,t}=\frac{1}{N}\mathbb{1}^{\top}(Z_t-\bar{Z}_t).\tag{41b}$$

By Lemma 2, we have $\lim_{t\to\infty}\zeta_{2,t}=0$ a.s. Next, we show that $\lim_{t\to\infty}\zeta_{3,t}=0$. From Theorem 2, we have that $\sup_t\mathbb{E}\|\gamma_t^{-1}\lambda_t^{\perp}\|^2<\infty$, which means that $\lambda_t^{\perp}$ decreases at the same speed as γ_t. Also, $\lim_{t\to\infty}\mu_t^i=\mathbb{E}_{s\sim d_\theta,a\sim\pi_\theta}[\bar{g}(s,a)]$. We consider two circumstances. One is that if $h_t>>\gamma_t$ as $t\to\infty$, we have that $|\lambda_t^{\perp}|_{\infty}<<h_t$. So, $\zeta_{3,t}\to 0$ as $t\to\infty$. The other is that if $h_t\sim\gamma_t$ or $h_t=o(\gamma_t)$, we have $\bar{Z}_t\to 0$ directly since the limits of $\bar{Z}_t$ and Z_t are both 0. Then the proof follows that in [33]. □

REFERENCES

[1] W.-J. Ma, J. Wang, V. Gupta, and C. Chen, "Distributed energy management for networked microgrids using online admm with regret," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 847–856, 2018.

[2] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for selective key applications in power systems: Recent advances and future challenges," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2935–2958, 2022.

[3] B. Chen, J. Wang, X. Lu, C. Chen, and S. Zhao, "Networked microgrids for grid resilience, robustness, and efficiency: A review," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 18–32, 2021.

[4] B. Fan, Q. Li, W. Wang, G. Yao, H. Ma, X. Zeng, and J. M. Guerrero, "A novel droop control strategy of reactive power sharing based on adaptive virtual impedance in microgrids," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 11, pp. 11 335–11 347, 2022.

[5] R. Xu, C. Zhang, Y. Xu, Z. Dong, and R. Zhang, "Multi-objective hierarchically-coordinated volt/var control for active distribution networks with droop-controlled pv inverters," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 998–1011, 2022.

[6] T. Chen, Y. Song, D. J. Hill, and A. Y. S. Lam, "Chance-constrained opf in droop-controlled microgrids with power flow routers," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2601–2613, 2022.

[7] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[8] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] S. Bhatnagar, "An actor–critic algorithm with function approximation for discounted cost constrained markov decision processes," *Systems & Control Letters*, vol. 59, no. 12, pp. 760–766, 2010.

[10] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5872–5881.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[13] M. Glavic, "(deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annual Reviews in Control*, vol. 48, pp. 22–35, 2019.

[14] C. M. Colson and M. H. Nehrir, "Comprehensive real-time microgrid power management and control with distributed agents," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 617–627, 2013.

[15] Y. S. Foo. Eddy, H. B. Gooi, and S. X. Chen, "Multi-agent system for distributed management of microgrids," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 24–34, 2015.

[16] M. R. Basir Khan, R. Jidin, and J. Pasupuleti, "Multi-agent based distributed control architecture for microgrid energy management and optimization," *Energy Conversion and Management*, vol. 112, pp. 288–307, 2016.

[17] T. Zhao and Z. Ding, "Distributed finite-time optimal resource management for microgrids based on multi-agent framework," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6571–6580, 2018.

[18] X. Zhou and Q. Ai, "An integrated two-level distributed dispatch for interconnected microgrids considering unit commitment and transmission loss," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 2, p. 025504, 2019.

[19] X. Zhou, Q. Ai, and H. Wang, "Consensus-based distributed economic dispatch incorporating storage optimization and generator ramp-rate constraints in microgrids," *Journal of Renewable and Sustainable Energy*, vol. 10, no. 4, p. 045501, 2018.

[20] Z. Yan and Y. Xu, "Real-time optimal power flow: A lagrangian based deep reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3270–3273, 2020.

[21] Y. Zhou, B. Zhang, C. Xu, T. Lan, R. Diao, D. Shi, Z. Wang, and W.-J. Lee, "A data-driven method for fast ac optimal power flow solutions via deep reinforcement learning," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1128–1139, 2020.

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

[23] D. Cao, W. Hu, X. Xu, Q. Wu, Q. Huang, Z. Chen, and F. Blaabjerg, "Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1101–1110, 2021.
[24] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, 2018.
[25] R. Hao, T. Lu, Q. Ai, Z. Wang, and X. Wang, "Distributed online learning and dynamic robust standby dispatch for networked microgrids," *Applied Energy*, vol. 274, p. 115256, 2020.
[26] R. Hao, T. Lu, Q. Ai, and H. He, "Distributed online dispatch for microgrids using hierarchical reinforcement learning embedded with operation knowledge," *IEEE Transactions on Power Systems*, pp. 1–1, 2021.
[27] G. Ceusters, L. R. Camargo, R. Franke, A. Nowé, and M. Messagie, "Safe reinforcement learning for multi-energy management systems with known constraint functions," *Energy and AI*, vol. 12, p. 100227, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666546822000738
[28] Y. Ye, H. Wang, P. Chen, Y. Tang, and G. Strbac, "Safe deep reinforcement learning for microgrid energy management in distribution networks with leveraged spatial–temporal perception," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3759–3775, 2023.
[29] Y. Wang, D. Qiu, M. Sun, G. Strbac, and Z. Gao, "Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach," *Applied Energy*, vol. 335, p. 120759, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030626192300123X
[30] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao, "Multi-agent safe policy learning for power management of networked microgrids," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1048–1062, 2021.
[31] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *4th International Conference on Learning Representations, ICLR 2016*, May 2-4 2016.
[32] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
[33] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*. Stochastic approximation algorithms and applications, 1997.
[34] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.
[35] G. Cui, B. Liu, W. Luan, and Y. Yu, "Estimation of target appliance electricity consumption using background filtering," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 5920–5929, 2019.
[36] Y. Ye, *Interior point algorithms: theory and analysis*. John Wiley & Sons, 2011.
[37] H. Khodr, F. Olsina, P. D. O.-D. Jesus, and J. Yusta, "Maximum savings approach for location and sizing of capacitors in distribution systems," *Electric Power Systems Research*, vol. 78, no. 7, pp. 1192–1203, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779607002143