ISA Transactions 158 (2025) 130–140

Contents lists available at ScienceDirect

ISA Transactions

journal homepage: www.elsevier.com/locate/isatrans

Research article

Distributed multi-agent reinforcement learning for multi-objective optimal dispatch of microgrids

Xiaowen Wang [a], Shuai Liu [a,*], Qianwen Xu [b], Xinquan Shao [a]

[a] School of Control Science and Engineering, Shandong University, Jinan, 250012, China
[b] Electric Power and Energy Systems Division, KTH Royal Institute of Technology, Stockholm, 100 44, Sweden

ARTICLE INFO

Keywords:
Multi-agent system
Distributed consensus strategy
Task decomposition
Multi-agent reinforcement learning
Microgrid

ABSTRACT

The distributed microgrids cooperate to accomplish economic and environmental objectives, which have a vital impact on maintaining the reliable and economic operation of power systems. Therefore a distributed multi-agent reinforcement learning (MARL) algorithm is put forward incorporating the actor-critic architecture, which learns multiple critics for subtasks and utilizes only information from neighbors to find dispatch strategy. Based on our proposed algorithm, multi-objective optimal dispatch problem of microgrids with continuous state changes and power values is dealt with. Meanwhile, the computation and communication resources requirements are greatly reduced and the privacy of each agent is protected in the process of information interaction. In addition, the convergence for the proposed algorithm is guaranteed with the adoption of linear function approximation. Simulation results validate the performance of the algorithm, demonstrating its effectiveness in achieving multi-objective optimal dispatch in microgrids.

1. Introduction

The proposal for microgrids aims to achieve the transition from traditional power grids to smart grids, enabling the transmission of electricity to consumers in an optimal, flexible, and reliable way through energy dispatch [1–3]. Multi-microgrid is a particular case of microgrid clusters, where a group of electrically coupled microgrids operate in a coordinated way. This type of interconnected microgrids can improve stability, reliability and power quality due to the connection of multi-distributed generators to the distribution system [4]. The challenge of individual or single microgrid operation also extends to multi-microgrid. Consequently, the optimal dispatch problem of multi-microgrid with complex security constraints, as a challenging topic, has been widely concerned by academic and industrial circles [5,6].

The main purpose of optimal dispatch is to coordinate the physical layer through information layer to attain a state of balance in supply and demand. This coordination seeks to minimize costs, diminish environmental pollution, and satisfy additional criteria, all of which constitute a multi-objective optimization challenge [7,8]. The information layer is a bridge that connects the cyber and physical spaces, mainly realizing data transmission via a network [9]. The physical layer includes various controllable and uncontrollable devices, such as DGs, PVs, etc.

While single-objective optimization yields a unique answer, multi-objective optimization typically produces a suite of Pareto optimal solutions. Numerous algorithms have been proposed to obtain Pareto optimal solutions for optimizing multiple objectives in microgrids, such as genetic algorithms [10] and particle swarm algorithms [11]. However, these methods are not suitable for scenarios involving high-dimensional optimization problems and nonconvex Pareto fronts which are common in microgrid systems.

RL approach [12], which has been extensively utilized in the power grid for solving sequential decision making problems, is expected to be an effective way to tackle the intractable problem. Accordingly, RL-based approaches show a great adaption in multi-objective optimization problems of microgrid systems. In [13], an RL algorithm based on SARSA was presented as an integrated approach to resource provisioning and scheduling for responsive grids. According to multiple objectives in energy resource management, a multi-objective evolutionary RL algorithm was developed in [14].

Nevertheless, the above methods have been primarily investigated for discrete action spaces, limiting their capability for continuous control in real world applications. Considering security, economic and environmental factors, a hierarchical multi-objective deep deterministic policy gradient algorithm was proposed in [15] to dispatch the smart

* Corresponding author.
E-mail addresses: xiaowenwang@mail.sdu.edu.cn (X. Wang), lius0025@ntu.edu.sg (S. Liu), qianwenx@kth.se (Q. Xu), 202000171027@mail.sdu.edu.cn (X. Shao).

https://doi.org/10.1016/j.isatra.2025.01.009
Received 14 June 2024; Received in revised form 2 January 2025; Accepted 3 January 2025
Available online 22 January 2025

Nomenclature

DG	Diesel generations
PV	Photovoltaics
RL	Reinforcement learning
MARL	Multi-agent reinforcement learning
PCC	Point of common coupling
SOC	State of charge
$\mathcal{G}$	Graph of the communication network
D	Weight matrix of the graph
d_{ij}	Element of the weight matrix
t	Time slot
$\mathcal{T}$	Set of time slots
$\mathcal{N}$	Set of all microgrids
p_i^l	Load at node i
p_i^s	Power output of the PV at node i
p_i^d	Power generation of DG at node i
p_i^b	Power dispatch of battery at node i
p_i^g	Power purchased from the utility grid at node i
p_i^{cap}	Battery capacity at node i
ρ_{ch}	Efficiency of battery charging
ρ_{dis}	Efficiency of battery discharging
$p_{i,\min}^d$	Minimum power output of DG at node i
$p_{i,\max}^d$	Maximum power output of DG at node i
$p_{i,\min}^b$	Minimum power output of battery at node i
$p_{i,\max}^b$	Maximum power output of battery at node i
$p_{i,\text{ch}}^b$	Maximum charging power of battery at node i
$p_{i,\text{dis}}^b$	Maximum discharging power of battery at node i

grid. In [16], an approach for microgrid resources planning was introduced based on multi-objective MARL techniques which can figure out Pareto optimal solutions. However, most of existing MARL algorithms for multi-objective problems are centralized, which is with less privacy and scalability.

Compared with centralized manners, distributed RL algorithms can accomplish complex tasks by using local information. In [17], an RL algorithm without central controller was studied for economic dispatch. A distributed RL algorithm was introduced in [18] which combined value function approximation utilizing a distributed optimization framework that employs multiplier splitting for dynamic economic dispatch. Note that the above mentioned literature only discussed single-objective problems. It remains challenging to simultaneously fulfill the multiple objectives by a distributed method.

Therefore, we take inspiration from [19] and develop a distributed MARL algorithm based on an actor-critic architecture for multi-objective optimal dispatch in multi-microgrid system. The primary contributions can be summarized in three key areas as outlined below:

(1) *Optimization problem formulation:* The multi-objective optimal dispatch problem of multi-microgrid is transformed into a multi-objective optimization one. Different from [5,17,18], which only discussed single objective problem, multiple optimization objectives include reducing energy cost, diminishing emissions of atmospheric pollutants and maintaining stable operation are considered simultaneously. The optimization model introduced in this paper matches actual operational scenarios more closely.

(2) *Algorithm design:* The distributed MARL algorithm not only aims at obtaining Pareto optimal solutions but also completes the intricate continuous tasks solely through the utilization of local information. Each agent learns a dispatch policy based on an actor-critic architecture after decomposing the overall reward into several components according to the multiple objectives with neighbors information. Unlike the centralized learning and decentralized execution manner in [19], we introduce a consensus strategy which depends on communications between neighbors with advantages such as flexibility, reliability and privacy preservation.

(3) *Performance evaluation:* The optimality and convergence analysis of the proposed algorithms are given which theoretically guarantee the performance of the designed algorithm. The convergence of the proposed algorithms is established through two time-scale stochastic approximation techniques when linear function approximation is used.

The remainder of this paper is structured in the following manner. In Section 2, the formulation of the primary optimization challenge within multi-microgrid systems is presented. Section 3 gives a detailed account of the multi-objective optimal dispatch problem under the RL framework. In Section 4, we introduce the design of distributed cooperative MARL algorithms with multiple tasks. The optimality and convergence of the proposed algorithm are proved rigorously in Section 5. In Section 6, comprehensive case studies are discussed. Ultimately, the conclusions are summarized in Section 7.

2. Problem formulation

In normal circumstances, multi-microgrids connect to the main utility grid through the PCC. This paper considers that the relationship among microgrids is cooperative, as the satisfaction of the load of the entire system requires coordination among all microgrids. Each DC microgrid is equipped with PV, batteries, loads and DGs, where DGs are installed as controllable energy resources. The load $p_i^l(t)$ and the power output of the PV panel $p_i^s(t)$ are uncontrollable variables. To achieve the optimization objectives, we can regulate the power generation of DG $p_i^d(t)$, the power dispatch of battery $p_i^b(t)$ and the power purchased from the utility grid $p_i^g(t)$.

2.1. Objectives

For the operation requirements of microgrids, the multi-objective optimization model includes economic objective and environmental objective.

(i) Economic Objective

The cost function C_i^g of microgrid i for purchasing power from the utility grid is

$$C_i^g(p_i^g(t)) = \alpha^g(t)|p_i^g(t)|,$$

in which $\alpha^g(t)$ represents the real-time electricity price of the utility grid. The fuel cost of DG C_i^d on microgrid i is characterized as

$$C_i^d(p_i^d(t)) = \alpha_i^d(p_i^d(t))^2 + \beta_i^d p_i^d(t) + \gamma_i^d,$$

where $\alpha_i^d > 0$, $\beta_i^d \geq 0$ and $\gamma_i^d \geq 0$ denote the coefficients of the DG on microgrid i. As for the battery on microgrid i, the cost function C_i^b can be described as

$$C_i^b(p_i^b(t)) = \alpha_i^b(x_i^b(t))^2 + \beta_i^b x_i^b(t) + \gamma_i^b,$$

where $x_i^b(t) = p_i^b(t)+3p_{i,\max}^b(1-\text{SOC})$. $\alpha_i^b > 0$, $\beta_i^b \geq 0$ and $\gamma_i^b \geq 0$ represent the predetermined coefficients. The state of charge (SOC) of the battery is limited to $[0,1]$ and follows

$$\text{SOC}_i(t+1) = \text{SOC}_i(t) - \left(\text{I}(p_i^b(t))\rho_{\text{dis}} + \frac{1-\text{I}(p_i^b(t))}{\rho_{\text{ch}}}\right)p_i^b(t)/p_i^{\text{cap}}.$$

where $\text{I}(\cdot)$ is an indicator function that equals 1 when argument $(\cdot) > 0$ and equals 0 otherwise. Note that $p_i^b(t) < 0$ for a charging period and $p_i^b(t) \geq 0$ for a discharging period.

Consequently, the total economic cost $f_i^{\text{eco}}(\cdot)$ can be calculated as

$$f_i^{\text{eco}}(p_i^g(t),p_i^d(t),p_i^b(t)) = C_i^g\left(p_i^g(t)\right)+C_i^d\left(p_i^d(t)\right) + C_i^b\left(p_i^b(t)\right). \tag{1}$$

(ii) Environmental Objective

Given the increasing impact of the greenhouse effect, minimizing environmental pollution is of great importance. The objective is to reduce the amercement of non-renewable energy, which can be expressed as

$$f_i^{\mathrm{env}}(p_i^g(t), p_i^d(t)) = \sum_{l \in \mathcal{H}} c_l \times a_{il} \times (p_i^g(t) + p_i^d(t)), \tag{2}$$

where $\mathcal{H}$ is the set of atmospheric pollutant types. c_l and a_{il} denote the amercement and the emission rate of the lth type of atmospheric pollutant emission for microgrid i, respectively.

2.2. Constraints

Ensuring the power balance is important for the safety and stability operation of the system, and this constraint can be formulated as

$$\sum_{i \in \mathcal{N}} \left[p_i^s(t) + p_i^b(t) + p_i^d(t) + p_i^g(t)\right] = \sum_{i \in \mathcal{N}} p_i^l(t). \tag{3}$$

Each microgrid is connected to the bus, allowing power exchange to occur between all microgrids. Consequently, a balance between power supply and demand can be achieved by bidirectional power exchange among microgrids. Each microgrid will purchase electricity from the utility grid when the power supply is insufficient.

In this instance, the controllable variables must meet the operating ranges, i.e., capacity constraints. The capacity constraint of the DG is

$$p_{i,\min}^d < p_i^d(t) < p_{i,\max}^d. \tag{4}$$

The power output of battery is restricted within the following limit

$$p_{i,\min}^b < p_i^b(t) < p_{i,\max}^b. \tag{5}$$

To extend battery lifespan, extreme charging and discharging are restricted, hence $p_i^b(t)$ must be within specified limits

$$p_{i,\mathrm{dis}}^b(t) < p_i^b(t) < p_{i,\mathrm{ch}}^b(t). \tag{6}$$

2.3. Optimization problem

To facilitate the design and analysis, by combining the multiple objectives (1)–(2) and all constraints (3)–(6), the multi-objective optimal dispatch can be reformulated as

$$\begin{aligned} \min \quad & \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{N}} \left[f_i^{\mathrm{eco}}(x_i(t)), f_i^{\mathrm{env}}(x_i(t))\right] \\ s.t. \quad & (3) \sim (6), \end{aligned} \tag{7}$$

where $x_i(t) \in \{p_i^g(t), p_i^d(t), p_i^b(t)\}$ are the set of variables. Therefore, the goal of microgrids is to solve the above problems through mutual cooperation.

3. Multi-objective optimization under the RL framework

We model a microgrid as an agent within the RL framework, as depicted in Fig. 1, and manipulate a distributed MARL approach to tackle the aforementioned challenges. In Fig. 1, the agent only interacts with its neighboring microgrids to learn the scheduling strategy at the communication level. At the physical level, each microgrid is linked to the bus, enabling power exchange between all microgrids. Additionally, all microgrids interact with the main utility grid through PCC for power transfer.

3.1. Markov decision process (MDP)

We consider the MARL with N agents that can be described as $\langle \mathcal{S}_1, \dots, \mathcal{S}_N, \mathcal{A}_1, \dots, \mathcal{A}_N, \mathcal{R}_1, \dots, \mathcal{R}_N, p, \mathcal{G} \rangle$ in which $\mathcal{S}_i$ indicates the continuous state space. $\mathcal{A}_i$ represents the continuous action space. Further, $\mathcal{A} = \Pi_{i=1}^N \mathcal{A}_i$ signifies the joint action space and $\mathcal{R}_i : \mathcal{S}_i \times \mathcal{A}_i \to \mathbb{R}$ represents the reward function. Denote $\mathbb{E}(r_{i,t+1} | s_t, a_t)$ as the conditional expectation of agent i receiving reward $r_{i,t+1}$. Let $p(ds'|s,a)$ be the state transition kernel. The weighted graph $\mathcal{G}$ indicates the communication network for all agents.

Specifically, a compact set $\Theta_i \in \mathbb{R}^{p^i}$ is defined as the parameter space for the actor. Accordingly, define $\pi_{\theta_i}(a_i|s_i) : \mathcal{S}_i \to \mathcal{A}_i$ as the local policy of agent i parameterized by $\theta_i \in \Theta_i$ which represents the probability density for agent i selecting action a_i at state s_i. Further, the joint policy parameter is denoted by $\theta = [\theta_1^\top, \dots, \theta_N^\top]^\top \in \Theta$ with $\Theta = \prod_{i=1}^N \Theta^i$. The probability density is given by $\pi_\theta(a|s) = \Pi_{i=1}^N \pi_{\theta_i}(a_i|s_i)$. For convenience, we define all π_θ as θ in the subscript. The transition kernel of the Markov chain $\{s_t\}_{t \geq 0}$ is defined as $P_\theta(ds'|s) = \int_{\mathcal{A}} d\pi_\theta(a|s) P(ds'|s,a), \forall s, s' \in \mathcal{S}$, in which $d\pi_\theta(a|s) = \pi_\theta(a|s)\upsilon(da)$. The Markov chains $\{s_t\}_{t\geq 0}$ and $\{s_t, a_t\}_{t \geq 0}$ induced by π_θ are geometrically ergodic with unique invariant measures $\varsigma_\theta(ds)$ and $\hat{\varsigma}_\theta(ds, da) = \varsigma_\theta(ds) d\pi_\theta(a|s)$, respectively.

Assumption 1. For any $s_i \in \mathcal{S}_i$ and $\theta_i \in \Theta_i$, there holds $\pi_{\theta_i}(a_i|s_i) > 0$, $\forall i = 1, \dots, N$. Assume $\pi_{\theta_i}(a_i|s_i)$ possesses continuous differentiability with respect to θ_i throughout Θ_i.

All agents aim to maximize the globally expected time-average reward by utilizing local information, including the local instantaneous reward $r_{i,t+1}$ and action $a_{i,t}$, and the messages from neighbors via communication. Let $\bar{R}(s,a) = N^{-1} \sum_{i=1}^N R_i(s,a)$ be the globally averaged reward and $\bar{r}_t = N^{-1} \sum_{i=1}^N r_{i,t}$ be the instantaneous averaged reward, which yield that $\bar{R}(s,a) = \mathbb{E}[\bar{r}_{t+1} | s_t = s, a_t = a]$. The essential goal of an agent is to maximize the globally expected average reward, i.e.,

$$\max_\theta \mathcal{J}(\theta) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\Big(\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N r_{i,t+1}\Big) = \int_S \varsigma_\theta(ds) \int_{\mathcal{A}} d\pi_\theta(a|s) \bar{R}(s,a).$$

The relative action value function and the relative state value function can be given as $Q_\theta(s,a) = \sum_{t=0}^{T-1} \mathbb{E}\left[\bar{r}_{t+1} - \mathcal{J}(\theta) | s_0 = s, a_0 = a, \pi_\theta\right]$ and $V_\theta(s) = \sum_{a \in \mathcal{A}} \pi_\theta(s,a) Q_\theta(s,a)$, respectively. Therefore, define the local advantage function $A_\theta(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as $A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$. For agent i, the advantage function is given as

$$A_{i,\theta}(s,a) = Q_{i,\theta}(s,a) - V_{i,\theta}(s, a_{-i}), \tag{8}$$

where $V_{i,\theta}(s, a_{-i}) = \sum_{a_i \in \mathcal{A}_i} \pi_{\theta_i}(s_i, a_i) Q_{i,\theta}(s_i, a_i, a_{-i})$.

Lemma 1 (*[20]*)**.** *For MARL with N agents, the policy gradient is characterized as $\nabla_{\theta_i} \mathcal{J}(\theta) = \sum_{d=1}^K \mathbb{E}_{s \sim \rho_\theta, a \sim \pi_\theta}\left[\Psi_{i,t} A_{i,\theta}^d(s,a)\right]$ where $\Psi_{i,t} = \nabla_{\theta_i} \log \pi_{\theta_i}(s_i, a_i)$.*

To obtain the Pareto optimal solutions by the MARL approach, we introduce the section of multitask MARL as follows.

3.2. Multitask MARL

For a collaborative MARL involving K tasks, the optimization problem can be characterized as

$$\max_\theta \quad \mathcal{J}(\theta), \tag{9}$$

where $\mathcal{J}(\theta) = [\mathcal{J}^1(\theta), \dots, \mathcal{J}^d(\theta), \dots, \mathcal{J}^K(\theta)]$. $\mathcal{J}^d(\theta)$ represents the objective function of all agents for the dth task with $d = 1, \dots, K$.

Similarly, we can decompose the holistic reward function into K subreward functions, i.e., $R(s,a) = \sum_{d=1}^K R^d(s,a)$. The state and state–action value function can be given as $V_\theta(s) = \sum_{d=1}^K V_\theta^d(s)$ and $Q_\theta(s,a) = \sum_{d=1}^K Q_\theta^d(s,a)$, respectively. Meanwhile, we can get $A_\theta(s,a) = \sum_{d=1}^K A_\theta^d(s,a)$. For an MDP with multiple tasks, denote Π^* as the optimal policy set and Π^{d*} as dth optimal policy set, $\forall d = 1, \dots, K$.

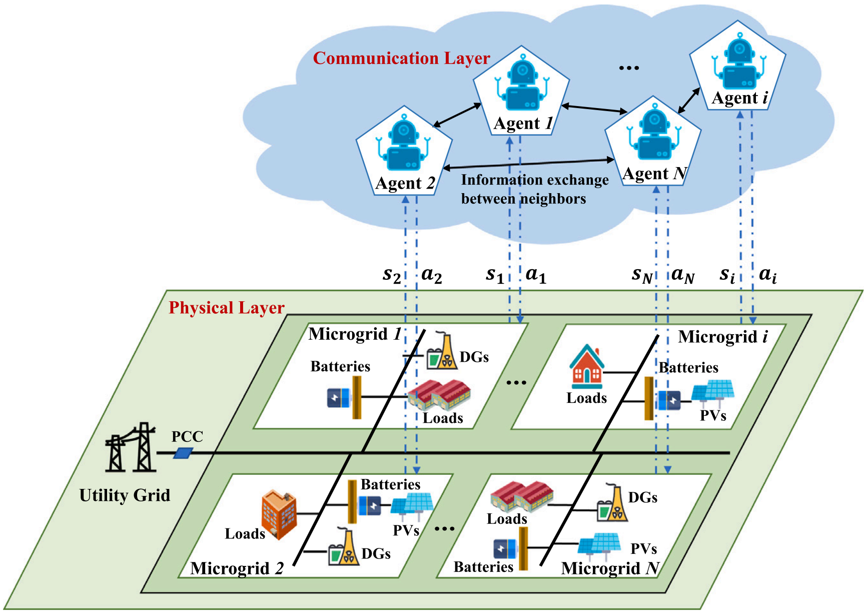


Fig. 1. The structure of distributed cooperative MARL.

Assumption 2. For an MDP with K tasks that have no temporal relations with each other, we have $R(s,a) = \sum_{d=1}^{K} R^d(s,a)$ such that $\bigcap_{d=1}^{K} \Pi^{d*} \neq \emptyset$ holds.

Definition 1 (*Pareto Domination [21]*). For the optimization problem (9) of all agents, if $\mathcal{J}^d(\theta^1) \geq \mathcal{J}^d(\theta^2)$ holds $\forall d = 1, \ldots, K$, then policy θ^1 is Pareto dominate θ^2, i.e., $\theta^1 \succeq \theta^2$.

Definition 2 (*Pareto Optimal [21]*). For the optimization problem (9) of all agents, if $\theta^* \succeq \theta$ holds $\forall \theta \in \Theta$, policy θ^* is called Pareto optimal.

3.3. Formulation of the multitask MARL problem

Next, we redefine the variables in microgrids with elements in RL: a state set $\mathcal{S}_i$, an action set $\mathcal{A}_i$ and a sequence of local rewards r_i. The state includes the power generation of PV $p_i^s(t)$, the load $p_i^l(t)$, the SOC of battery $\text{SOC}_i(t)$ and the electricity selling price of the utility grid $\alpha^g(t)$

$$s_i(t) = \{p_i^s(t), p_i^l(t), \text{SOC}_i(t), \alpha^g(t)\} \in \mathcal{S}_i.$$

Further, the decision variables is considered as actions of agent i

$$a_i(t) = \{p_i^d(t), p_i^b(t)\} \in \mathcal{A}_i.$$

To fulfill the constraint (6), the action will be respectively enforced by $p_i^{b*}(t) = \text{clip}(p_i^b(t), p_{i,\text{dis}}^b(t), p_{i,\text{ch}}^b(t))$, where the clip function clamps the first input variable $p_i^b(t)$ into the latter range $[p_{i,\text{dis}}^b(t), p_{i,\text{ch}}^b(t)]$.

To strictly satisfy the power balance, we set the purchase electricity from the utility grid as a flexible variable which can be calculated by (3)

$$p_i^g(t) = \frac{1}{N} \sum_{i \in \mathcal{N}} \left(p_i^l(t) - p_i^s(t) - p_i^{b*}(t) - p_i^d(t)\right).$$

If the action does not meet the constraint, a penalty is imposed, i.e.,

$$L(t) = \sum_{i \in \mathcal{N}} \max\left\{\left(p_i^b(t) - p_{i,\text{dis}}^b(t), 0\right) + \max\left(-p_{i,\text{ch}}^b(t) - p_i^b(t), 0\right)\right\}.$$

Further, according to (7), we define the instantaneous holistic reward as

$$r_i(s_t, a_t) = \sum_{d=1}^{K} r_i^d(s_t, a_t) = -\left[\sum_{d=1}^{K} f_i^d(x_i(t)) + \lambda L(t)\right],$$

where $f^d(x_i(t))$ is the cost under the dth task and λ denotes the penalty coefficient.

Then, the aim of the MARL system with multiple tasks is to solve the problem as below

$$\max_{\theta} \mathcal{J}(\theta) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left(\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{K} r_{i,t}^d\right).$$

4. Distributed cooperative MARL for multi-objective optimization

We propose a distributed cooperative MARL algorithm. This algorithm combines the actor-critic architecture to learn Pareto optimal policy and consensus update methods to reduce the computational and communication resource requirements, protecting the privacy of each agent.

4.1. Graph theory

As depicted in Fig. 2, data transmission depends on information flows which are displayed by red line. The communication topology is an undirected graph denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices set $\mathcal{V} = \{1, \ldots, N\}$ and the edges set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Define $\mathcal{N}_i = \{j \in \mathcal{V} | (j, i) \in \mathcal{E}\}$ as the neighbors set of vertex i from which i can receive information. The weight matrix is denoted by $D = [d_{ij}] \in \mathbb{R}^{N \times N}$ with the entry $d_{ij} > 0$ if $(i, j) \in \mathcal{E}$ (including $j = i$) and $d_{ij} = 0$ otherwise, where d_{ij} is the weight. The weight matrix $D = [d_{ij}] \in \mathbb{R}^{N \times N}$ is doubly stochastic if $\sum_{j \in \mathcal{N}_i} d_{ij} = 1$ for all $i \in \mathcal{N}_j$.

Assumption 3. The communication topology about agents in microgrid systems is undirected and connected.

Assumption 4. The weight matrix $D = [d_{ij}] \in \mathbb{R}^{N \times N}$ is doubly stochastic. There exists a constant $\iota \in (0, 1)$ such that, for any $d_{ij} > 0$, we have $d_{ij} \geq \iota$.

Remark 1. The standard hypothesis in Assumption 4 guarantees that the update of each agent can converge to a consensual vector. The condition of the spectral norm is interrelated to the connectivity of the communication.

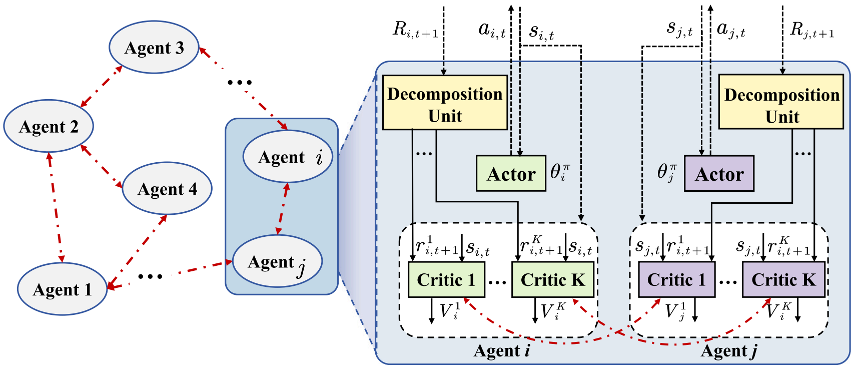


Fig. 2. The structure of distributed cooperative MARL with K tasks.

4.2. Algorithm design

Considering the PPO algorithm [22] is a cutting-edge algorithm with the ability to increase sample utilization, we are motivated to design a distributed PPO-based algorithm by applying the consensus update rule to replace the traditional MARL with centralized learning and decentralized execution. The architecture of the algorithm is shown in Fig. 2.

For the sake of handling continuous states and actions, we utilize a two-step actor-critic framework that operates across varying temporal scales. During training, actor receives the state $s_{i,t} \in \mathcal{S}_i$ and exports the local policy $\pi_{\theta_i}(a_i|s)$. Whereupon, critic evaluates the value of global policy by estimating V_θ since updating the global state value function only with reward $r_{i,t}$ is insufficient for achieving a local update. Motivated by all mentioned, the consensus update method is employed to estimate V_θ by communicating with neighbors. The state value function is approximated by $V(s;\omega) = \omega^\top \phi(s)$, in which $\phi(s) = [\phi_1(s), \dots, \phi_l(s), \dots, \phi_L(s)]^\top \in \mathbb{R}^L$ is the feature associated with state s. The computational complexity is $\mathcal{O}(N + K + p_i + L)$ where N and K are the number of agents and objectives, respectively. p_i and L denote the dimension of network θ_i and ω_i, respectively. However, if each agent must maintain a Q-table that has a dimensionality of $|\mathcal{S}| \cdot |\mathcal{A}| \times |\mathcal{S}| \cdot |\mathcal{A}|$ as in [23], which grows exponentially with the total number of agents.

The update rule of the critic network is introduced as follows. To avoid using global information, a scalar μ is used to estimate $\mathcal{J}^d(\theta)$ under the dth task

$$\hat{\mu}^d_{i,t} = (1 - \beta_{\omega,t})\mu^d_{i,t} + \beta_{\omega,t} r^d_{i,t+1}, \tag{10}$$

$$\mu^d_{i,t+1} = \sum_{j \in \mathcal{V}} d_{ij} \hat{\mu}^d_{j,t}, \tag{11}$$

where $\beta_{\omega,t} > 0$ represents the stepsize. Thus, the parameter ω_i^d for agent i under the dth task is updated by

$$\hat{\omega}^d_{i,t} = \omega^d_{i,t} + \beta_{\omega,t} \delta^d_{i,t} \nabla_\omega V^d_t(\omega^d_{i,t}), \tag{12}$$

$$\omega^d_{i,t+1} = \sum_{j \in \mathcal{V}} d_{ij} \hat{\omega}^d_{j,t}, \tag{13}$$

where the local TD-error $\delta^d_{i,t}$ is

$$\delta^d_{i,t} = r^d_{i,t+1} - \mu^d_{i,t} + V^d_{t+1}(\omega^d_{i,t}) - V^d_t(\omega^d_{i,t}). \tag{14}$$

Through Lemma 1, agent i improves its policy as follows

$$\theta_{i,t+1} = \theta_{i,t} + \beta_{\theta,t} A_{i,t} \Psi_{i,t}, \tag{15}$$

where $\beta_{\theta,t} > 0$ denotes the stepsize of the actor. Besides, the steps involved in executing the proposed algorithm are outlined in Algorithm 1.

Algorithm 1 the proposed distributed PPO-based algorithm for multi-objective optimal dispatch of multi-microgrid

```
Input: Agent number $N$, task number $M$, network architecture.
Initialization: Network parameters $\mu_i^d$, $\hat{\mu}_i^d$, $\omega_i^d$, $\hat{\omega}_i^d$, $\theta_i$, the stepsizes
    $\{\beta_{\omega,t}\}_{t \ge 0}$ and $\{\beta_{\theta,t}\}_{t \ge 0}$, the initial state $s_0$, the iteration counter $t \leftarrow 0$.
for iteration $c = 1, 2, \dots$ do
  Initialize an empty set $\mathcal{D}$;
  for all $i = 1$ to $N$ do
    for all $d = 1$ to $K$ do
      Get actions: $a_{i,t} = \pi_i(s_{i,t})$;
      Execute $a_{i,t}$, obtain $s_{i,t+1}$ and reward $r_{i,t+1} = [r^1_{i,t+1}, \dots, r^K_{i,t+1}]$;
      Update $\hat{\mu}^d_{i,t}$ according to (10);
      Update $\delta^d_{i,t}$ according to (14);
      Update critic network parameter $\hat{\omega}^d_{i,t}$ according to (12);
    end for
    Update $\Psi_{i,t} \leftarrow \nabla_{\theta_i} \log \pi_{\theta_i}(s_i, a_i)$;
    Update actor network parameter $\theta_{i,t+1}$ according to (15);
    Send $\hat{\mu}_i^d$ and $\hat{\omega}_i^d$ to neighbors $\{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$ over network
      $\mathcal{G}$;
  end for
  for all $i = 1$ to $N$ do
    Update consensus parameters $\mu^d_{i,t+1}$ and $\omega^d_{i,t+1}$ according to (11)
      and (13), respectively;
  end for
  Update the iteration counter $c \leftarrow c + 1$;
end for
```

Remark 2. As the accessibility of distributed renewable energy continues to improve, the scale of microgrids and their communication networks have also expanded significantly. Therefore, these systems are vulnerable to various forms of malicious attacks, including denial of service [24] and false data injection [25]. Due to the absence of a central controller in distributed algorithms, when a microgrid is attacked, the distributed algorithm can dynamically discard the attacked microgrid based on the actual situation, allowing other microgrids to continue working to maintain the stability of the system operation.

Remark 3. The proposed algorithm is scalable, which means that the number of agents can be increased according to actual needs, thereby improving the performance and capacity of the algorithm, and flexibly responding to the ever-growing data volume and computing demands. If agents increases, the overall computational load of the algorithm will increase, mainly because the interaction and collaboration between agents require more computational resources to handle. For example, the action space of an agent will expand rapidly as the number of agents increases. Therefore, the training time of reinforcement learning agents will increase accordingly.

5. Theoretical analysis

Next, optimality and convergence analysis will be discussed.

5.1. Optimality analysis

Our algorithm designs specific critics corresponding to their own reward to learn Pareto optimal solutions. The analysis is provided below.

Lemma 2. *The optimization for $\mathcal{J}(\theta)$ is equivalent to optimization for $\mathcal{J}^d(\theta)$, $\forall d = 1, \dots, K$ if and only if Assumption 2 holds.*

Proof. We first prove the necessity. If optimizing $\mathcal{J}(\theta)$ is equivalent to optimizing $\mathcal{J}^d(\theta)$, $\forall d = 1, \dots, K$ separately, one can obtain

$$\max_{\theta} \sum_{d=1}^{K} \mathcal{J}^d(\theta) = \sum_{d=1}^{K} \max_{\theta^d} \mathcal{J}^d(\theta^d).$$

Suppose the optimal policy is θ^* and the subtask optimal policy is θ^{d*}. By the above equation, we have $\mathcal{J}^d(\theta^*) = \mathcal{J}^d(\theta^{d*})$. Also, because of $\theta^* \in \Pi^*$ and $\theta^{d*} \in \Pi^{d*}$, it is easy to get $\theta^* \in \Pi^{d*}$, $\forall d = 1, \dots, K$. Therefore, it is deduced that $\bigcap_{d=1}^{K} \Pi^{d*} \neq \emptyset$, i.e., Assumption 2 holds.

Next, we prove the sufficiency. If Assumption 2 holds, which means that $\bigcap_{d=1}^{K} \Pi^{d*} \neq \emptyset$ and $\theta^* \in \bigcap_{d=1}^{K} \Pi^{d*}$, we can derive that $\mathcal{J}^d(\theta^*) = \max_\theta \mathcal{J}^d(\theta)$. Based on task decomposition, one can obtain

$$\mathcal{J}(\theta^*) = \sum_{d=1}^{K} \mathcal{J}^d(\theta^*) = \sum_{d=1}^{K} \max_{\theta} \mathcal{J}^d(\theta) = \max_{\theta} \mathcal{J}(\theta).$$

According to the above analysis, it can be concluded that optimizing $\mathcal{J}(\theta)$ is equivalent to optimizing $\mathcal{J}^d(\theta)$, $\forall d = 1, \dots, K$ separately.

Theorem 1. *For an MDP with K cooperative tasks, if an equivalent optimal solution θ^* satisfies $\mathcal{J}(\theta^*) = \sum_{d=1}^{K} \mathcal{J}^d(\theta^*)$, then it is the Pareto optimal solution.*

Proof. Since the policy θ^* is an optimal solution in Lemma 2, one has the objective function $\mathcal{J}^d(\theta^*) = \max_\theta \mathcal{J}^d(\theta) \geq \mathcal{J}^d(\theta)$, $\forall \theta \in \Pi$ and $d = 1 \cdots, K$. According to Definition 1, we can obtain $\theta^* \geq \theta$, $\forall \theta \in \Pi$. Through Definition 2, it can be referred that $\theta^* \in \mathcal{P}_\theta$, i.e., θ^* is a Pareto optimal solution.

5.2. Convergence analysis

Ensuring the convergence of our algorithm when handling multiple tasks is essential for the dependable operation of microgrids.

Assumption 5. For any state $s \in \mathcal{S}$, the $\phi(s)$ remains uniformly bounded. Moreover, the $\{\phi_l\}_{l \in [L]}$ are linearly independent. For any $u \neq 0 \in \mathbb{R}^L$, $u^\top \phi$ is not a constant.

Assumption 6. The time-varying stepsizes $\beta_{\omega,t}$ and $\beta_{\theta,t}$ satisfy

$$\sum_{t=0}^{\infty} \beta_{\omega,t} = \sum_{t=0}^{\infty} \beta_{\theta,t} = \infty, \quad \sum_{t=0}^{\infty} \left(\beta_{\omega,t}^2 + \beta_{\theta,t}^2\right) < \infty,$$

and

$$\beta_{\theta,t} = o(\beta_{\omega,t}), \quad \lim_{t\to\infty} \beta_{\omega,t+1}\beta_{\omega,t}^{-1} = 1.$$

The two time scale stochastic approximation technique is employed during the following analysis. Define an operator $\mathcal{K}_\theta$ over $V : \mathcal{S} \to \mathbb{R}$ as $(\mathcal{K}_\theta V)(s) = \int_{\mathcal{S}\times\mathcal{A}} V(s')P(ds'|s,a)$, and an operator $\mathcal{H}_\theta$ as $(\mathcal{H}_\theta V)(s) = \bar{R}(s,a) - \mathcal{J}(\theta) + (\mathcal{K}_\theta V)(s)$, $\forall s \in \mathcal{S}$, $a \in \mathcal{A}$. Further, we denote the inner product with functions $f_1 : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{q1\times p}$ and $f_2 : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{p\times q2}$ as $\langle f_1, f_2\rangle_\theta = \int_{\mathcal{S}\times\mathcal{A}} \varsigma(ds,da) f_1(s,a) f_2(s,a)$.

Let $x_{i,t}^d = [\mu_{i,t}^d, (\omega_{i,t}^d)^\top]^\top \in \mathbb{R}^{1+L}$. The stability of $\{x_{i,t}^d\}$ is shown in the following Lemma.

Lemma 3 ([20]). *Under Assumptions 1, 4–6, the sequence $x_{i,t}^d$ satisfies $\sup_{t\to\infty} \|x_{i,t}^d\| < \infty$ almost surely (a.s.), $\forall i \in \mathcal{N}$.*

Theorem 2. *Considering iterations (10)–(13) for any given π_θ under Assumptions 1, 4–6, we have $\lim_{t\to\infty} \mu_{i,t}^d = \mathcal{J}^d(\theta)$ and $\lim_{t\to\infty} \omega_{i,t}^d = \omega_\theta^d$ a.s. $\forall i \in \mathcal{N}$. Further, ω_θ^d has a unique solution which is characterized as*

$$\left\langle \phi^d, \mathbb{T}_{\theta^d}((\phi^d)^\top \omega^d) - (\phi^d)^\top \omega^d \right\rangle_\theta = 0. \tag{16}$$

Proof. Let $x_t^d = [(x_{1,t}^d)^\top, \dots, (x_{N,t}^d)^\top]^\top$. The updates of x_t^d about (10)–(13) can be given as

$$x_{t+1}^d = (D \otimes I)(x_t^d + \beta_{\omega,t} y_{t+1}^d),$$

where $\otimes$ is the Kronecker product. I denotes the identity matrix and $y_{t+1}^d = [(y_{1,t+1}^d)^\top, \dots, (y_{N,t+1}^d)^\top] \in \mathbb{R}^{(1+L)N}$ with $y_{i,t+1}^d = \left[r_{i,t+1}^d - \mu_{i,t}^d, \delta_{i,t}^d(\phi_{i,t}^d)^\top\right]^\top$.

For any x^d, we define the operator $\langle\cdot\rangle : \mathbb{R}^{(1+L)N} \to \mathbb{R}^{1+L}$ as

$$\langle x^d \rangle = N^{-1}(\mathbf{1}^\top \otimes I)x^d = N^{-1} \sum_{i\in\mathcal{N}} x_i^d.$$

Denote a projection operator $\mathbb{J} = (N^{-1}\mathbf{1}\mathbf{1}^\top) \otimes I$ and $\mathbb{J}x^d = \mathbf{1} \otimes \langle x^d \rangle$ with $\{\mathbf{1} \otimes f : f \in \mathbb{R}^{1+L}\}$. Consequently, we define $\check{\mathbb{J}} = I - \mathbb{J}$. As for the disagreement vector of x^d, we can show $\check{x}^d = \check{\mathbb{J}}x^d = x^d - \mathbf{1} \otimes \langle x^d \rangle$. Thus, one can rewrite x_t^d as $x_t^d = \check{x}_t^d + \mathbf{1} \otimes \langle x_t^d \rangle$. To prove the a.s. convergence of $\{x_t^d\}$, the proof consists of two stages as outlined below.

First, we prove that $\lim_{t\to\infty} \check{x}_t^d = 0$ a.s. According to Assumption 4, $\check{x}_{t+1}^d$ can be rewritten as

$$\check{x}_{t+1}^d = \check{\mathbb{J}}[(D \otimes I)(x_t^d + \beta_{\omega,t} y_{t+1}^d)] = \check{\mathbb{J}}[(D \otimes I)(\check{x}_t^d + \beta_{\omega,t} y_{t+1}^d)].$$

According to the definition of $\mathbb{J}$ and $\check{\mathbb{J}}$, we can obtain $\check{x}_{t+1}^d = [(I - \mathbf{1}\mathbf{1}^\top/N)D \otimes I](\check{x}_t^d + \beta_{\omega,t} y_{t+1}^d)$. Define $\mathcal{P}_{t,1} = \sigma(r_\tau, x_\tau, s_\tau, a_\tau, \tau \leq t)$ as the filtration. Then, we get

$$\begin{aligned}
&\mathbb{E}(\|\beta_{\omega,t+1}^{-1}\check{x}_{t+1}^d\|^2|\mathcal{P}_{t,1}) \\
=\ & \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\mathbb{E}\left[(\beta_{\omega,t}^{-1}\check{x}_t^d + y_{t+1}^d)^\top[D^\top(I - \mathbf{1}\mathbf{1}^\top/N)D \otimes I](\beta_{\omega,t}^{-1}\check{x}_t^d + y_{t+1}^d)|\mathcal{P}_{t,1}\right] \\
\leq\ & \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\rho\mathbb{E}\left[(\beta_{\omega,t}^{-1}\check{x}_t^d + y_{t+1}^d)^\top(\beta_{\omega,t}^{-1}\check{x}_t^d + y_{t+1}^d)|\mathcal{P}_{t,1}\right] \\
\leq\ & \frac{\rho\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\left[\ \mathbb{E}\left(\left\|\beta_{\omega,t}^{-1}\check{x}_t^d\right\|^2|\mathcal{P}_{t,1}\right) + 2\mathbb{E}\left(\left\|\beta_{\omega,t}^{-1}\check{x}_t^d\right\||\mathcal{P}_{t,1}\right)[\mathbb{E}(\|y_{t+1}^d\|^2|\mathcal{P}_{t,1})]^{\frac{1}{2}}\right. \\
&\left. +\mathbb{E}(\|y_{t+1}^d\|^2|\mathcal{P}_{t,1})\ \right],
\end{aligned}$$

in which $\rho \in [0,1)$ denotes the spectral norm of $\mathbb{E}\left[D^\top(I - \mathbf{1}\mathbf{1}^\top/N)D\right]$. Through the definition of y_{t+1}^d, one can obtain

$$\mathbb{E}(\|y_{t+1}^d\|^2|\mathcal{P}_{t,1}) = \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left(\left\|r_{i,t+1}^d - \mu_{i,t}^d\right\|^2 + \left\|\delta_{i,t}^d(\phi_{i,t}^d)^\top\right\|^2\right)\Big|\mathcal{P}_{t,1}\right]. \tag{17}$$

Given Assumption 5, Lemma 3 and uniform boundedness of reward $r_{i,t}^d$, when there exists $H_1 < \infty$, we have

$$\mathbb{E}(\|y_{t+1}^d\|^2\mathbb{I}_{\{\sup_{\tau\leq t}\|x_\tau^d\|\leq M\}}|\mathcal{P}_{t,1}) \leq H_1.$$

Let $\mathbb{I}_{(\cdot)}$ be the indicator function and $\varpi_t^d = \|\beta_{\omega,t+1}^{-1}\check{x}_{t+1}^d\|^2\mathbb{I}_{\{\sup_{\tau\leq t}\|x_\tau^d\|\leq M\}}$. Then, we have

$$\mathbb{E}(\varpi_{t+1}^d) \leq \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\rho\left[\mathbb{E}(\varpi_t^d) + 2\sqrt{\mathbb{E}(\varpi_t^d)}\sqrt{H_1} + H_1\right].$$

For any $1 - \rho > \epsilon > 0$, there exists a sufficiently large t_0 such that for any $t > t_0$, $\beta_{\omega,t}^2\beta_{\omega,t+1}^{-2}\rho \leq 1 - \epsilon$. Consequently, for any $t \geq t_0$, there exists constant $h > 0$ such that

$$\begin{aligned}
\mathbb{E}(\varpi_{t+1}^d) &\leq (1-\epsilon)\left[\mathbb{E}(\varpi_t^d) + 2\sqrt{\mathbb{E}(\varpi_t^d)}\sqrt{H_1} + H_1\right] \\
&\leq (1-\epsilon/2)^{t-t_0}\mathbb{E}(\varpi_{t_0}^d) + 2h/\epsilon.
\end{aligned}$$

Thus, it is easy to know $\sup_{t\to\infty}\mathbb{E}(\varpi_t^d) < \infty$. Due to $\mathbb{I}_{\{\sup_{t\to\infty}\|x_t^d\|\le M\}} \le \mathbb{I}_{\{\sup_{\tau\le t}\|x_\tau^d\|\le M\}}$, it leads to

$$\sup_{t\to\infty}\mathbb{E}\Big(\|\beta_{\omega,t}^{-1}\check{x}_t^d\|^2\mathbb{I}_{\sup_{t\to\infty}\|x_t^d\|\le M}\Big) < \infty,$$

i.e., there exists a constant $H_2 < \infty$ such that $\mathbb{E}(\|\check{x}_t^d\|^2) \le H_2\beta_{\omega,t}^2$ on $\sup_{t\to\infty}\|x_t^d\| \le M$, for any $M > 0$ and $t \ge 0$.

Owing to Fubini's theorem, we further obtain

$$\sum_{t=0}^{\infty}\mathbb{E}\Big(\|\check{x}_t^d\|^2\mathbb{I}_{\sup_{t\to\infty}\|x_t^d\|\le M}\Big) < \infty,$$

which yields that $\sum_{t=0}^{\infty}\|\check{x}_t^d\|^2\mathbb{I}_{\{\sup_{t\to\infty}\|x_t^d\|\le M\}} < \infty$ a.s. This is equivalent to $\lim_{t\to\infty}\check{x}_t^d\mathbb{I}_{\{\sup_{t\to\infty}\|x_t^d\|\le M\}} = 0$ a.s. According to Lemma 3, we get

$$\lim_{t\to\infty}\check{x}_t^d = 0 \quad a.s. \tag{18}$$

Next, we establish the convergence of $\mathbf{1}\otimes\langle x_t^d\rangle$. The iteration of $\langle x_{t+1}^d\rangle$ is given as follows

$$\begin{aligned}\langle x_{t+1}^d\rangle &= N^{-1}(\mathbf{1}^\top\otimes I)(D\otimes I)(\mathbf{1}\otimes\langle x_t^d\rangle + \check{x}_t^d + \beta_{\omega,t}y_{t+1}^d)\\ &= \langle x_t^d\rangle + \beta_{\omega,t}\langle (D\otimes I)(y_{t+1}^d + \beta_{\omega,t}^{-1}\check{x}_t^d)\rangle.\end{aligned}\tag{19}$$

By denoting $\langle y_{t+1}^d\rangle = [\bar{r}_{t+1}^d - \langle\mu_t^d\rangle, \langle\delta_t^d\rangle(\phi_t^d)^\top]^\top$, (19) can be rewritten as

$$\langle x_{t+1}^d\rangle = \langle x_t^d\rangle + \beta_{\omega,t}\mathbb{E}[\langle y_{t+1}^d\rangle|\mathcal{P}_{t,1}] + \beta_{\omega,t}\zeta_{t+1}^d, \tag{20}$$

in which $\zeta_{t+1}^d = \langle (D\otimes I)(y_{t+1}^d + \beta_{\omega,t}^{-1}\check{x}_t^d)\rangle - \mathbb{E}[\langle y_{t+1}^d\rangle|\mathcal{P}_{t,1}]$.

Through simple manipulation, we have

$$\langle\check{x}_t^d\rangle = N^{-1}(\mathbf{1}^\top\otimes I)(x_t^d - \mathbf{1}\otimes\langle x_t^d\rangle) = \langle x_t^d\rangle - ((N^{-1}\mathbf{1}^\top\mathbf{1})\otimes\langle x_t^d\rangle) = 0,$$

and

$$\mathbb{E}[\langle (D\otimes I)(y_{t+1}^d + \beta_{\omega,t}^{-1}\check{x}_t^d)\rangle|\mathcal{P}_{t,1}] = \mathbb{E}[\langle (D\otimes I)y_{t+1}^d\rangle|\mathcal{P}_{t,1}] = \mathbb{E}[\langle (y_{t+1}^d)\rangle|\mathcal{P}_{t,1}].$$

Thus, it can be concluded that ζ_t^d is martingale difference sequence.

Moreover, we can obtain that

$$\mathbb{E}(\|\zeta_{t+1}^d\|^2|\mathcal{P}_{t,1}) \le 2(\mathbb{E}\|y_{t+1}^d + \beta_{\omega,t}^{-1}\check{x}_t^d\|_{\mathcal{Z}}^2|\mathcal{P}_{t,1}) + 2\|\mathbb{E}(\langle y_{t+1}^d\rangle|\mathcal{P}_{t,1})\|^2,$$

where $\mathcal{Z} = N^{-2}D^\top\mathbf{1}\mathbf{1}^\top D\otimes I$ has a bounded spectral norm. According to Assumption 4, there exists constants $H_3, H_4 < \infty$ such that

$$\begin{aligned}&\mathbb{E}(\|y_{t+1}^d + \beta_{\omega,t}^{-1}\check{x}_t^d\|_{\mathcal{Z}_t}^2|\mathcal{P}_{t,1})\mathbb{I}_{\{\sup_{t\to\infty}\|x_t^d\|\le M\}}\\ \le& H_3\mathbb{E}(\|y_{t+1}^d\|^2 + \|\beta_{\omega,t}^{-1}\check{x}_t^d\|_{\mathcal{Z}_t}^2|\mathcal{P}_{t,1})\mathbb{I}_{\{\sup_{t\to\infty}\|x_t^d\|\le M\}} < H_4.\end{aligned}\tag{21}$$

Then, according to the boundedness of $r_{i,t+1}^d$ and ϕ_t^d, we can get

$$\begin{aligned}\|\mathbb{E}(\langle y_{t+1}^d\rangle|\mathcal{P}_{t,1})\|^2 &\le \mathbb{E}(\|\langle y_{t+1}^d\rangle\|^2|\mathcal{P}_{t,1})\\ &\le H_5(1 + \|\langle\mu_t^d\rangle\|^2 + \|\langle\omega_t^d\rangle\|^2)\\ &= H_5(1 + \|\langle x_t^d\rangle\|^2).\end{aligned}\tag{22}$$

According to (21) and (22), there exists $H_6 < \infty$ such that

$$\mathbb{E}(\|\zeta_{t+1}^d\|^2|\mathcal{P}_{t,1}) \le H_6(1 + \|\langle x_t^d\rangle\|^2)$$

over $\{\sup_{t\to\infty}\|x_t^d\| \le M\}$.

To capture the asymptotic behavior of (20), it can be characterized by the following ordinary diffrential equation (ODE) [26],

$$\langle\dot{x}^d\rangle = \begin{pmatrix}-1 & 0\\ -\langle\phi^d, 1\rangle_\theta & \langle\phi^d, \mathcal{K}_\theta((\phi^d)^\top) - (\phi^d)^\top\rangle_\theta\end{pmatrix}\begin{pmatrix}\langle\mu^d\rangle\\ \langle\omega^d\rangle\end{pmatrix} + \begin{pmatrix}\mathcal{J}^d(\theta)\\ \langle\phi^d, \bar{R}^d\rangle_\theta\end{pmatrix}. \tag{23}$$

Denote the right hand side of the above ODE as $f(\langle x^d\rangle)$ and $\mathbb{A} = \begin{pmatrix}-1 & 0\\ -\langle\phi^d, 1\rangle_\theta & \langle\phi^d, \mathcal{K}_\theta((\phi^d)^\top) - (\phi^d)^\top\rangle_\theta\end{pmatrix}$. Apparently, $f(\langle x^d\rangle)$ is Lipschitz continuous in $\langle x^d\rangle$.

Then we suppose that $x_\theta^d = (\mu_\theta^d, (\omega_\theta^d)^\top)^\top$ is a solution for $f(\langle x_\theta^d\rangle) = 0$ and construct the Lyapunov function $L(\langle x^d\rangle) = \frac{1}{2}f(\langle x^d\rangle)^\top f(\langle x^d\rangle)$ of (23) which follows that $\frac{dL(\langle x^d\rangle)}{dt} = f(\langle x^d\rangle)^\top\mathbb{A}f(\langle x^d\rangle)$.

On account of the Jensen's inequality, we can claim that

$$\begin{aligned}f(\langle x^d\rangle)^\top\langle\phi^d, \mathcal{K}_\theta((\phi^d)^\top)\rangle_\theta f(\langle x^d\rangle) &\le \|f(\langle x^d\rangle)^\top\phi^d\|_\theta\|\mathcal{K}_\theta((\phi^d)^\top)f(\langle x^d\rangle)\|_\theta\\ &\le \|f(\langle x^d\rangle)^\top\phi^d\|_\theta^2,\end{aligned}$$

which means that $f(\langle x^d\rangle)^\top\langle\phi^d, \mathcal{K}_\theta((\phi^d)^\top) - (\phi^d)^\top\rangle_\theta f(\langle x^d\rangle) < 0$.

Note that A is a Lower triangular block matrix and $f(\langle x^d\rangle)^\top\langle\varphi^d, (\varphi^d)^\top\rangle_\theta f(\langle x^d\rangle) < 0$. Moreover, for any $f(\langle x^d\rangle) \ne 0$, the matrix A is negative definite, i.e., $\frac{dL(\langle x^d\rangle)}{dt} < 0$ and $\frac{dL(\langle x^d\rangle)}{dt} = 0$ only for the unique x_θ^d such that $f(\langle x^d\rangle) = 0$. The solution x_θ^d, i.e. the unique globally asymptotically stable equilibrium of (23), satisfies

$$\begin{aligned}-\mu_\theta^d + \mathcal{J}^d(\theta) &= 0,\\ \Big\langle\phi^d, \bar{R}^d - \mu_\theta^d + \mathcal{K}_\theta((\phi^d)^\top\omega_\theta^d) - (\phi^d)^\top\omega_\theta^d\Big\rangle_\theta &= 0.\end{aligned}$$

Moreover, we obtain that $\{x_t^d\}$ is bounded a.s. by Lemma 3 and so is the $\{\langle x_t^d\rangle\}$. According to Corollary 8 and Theorem 9 in [26], for any $M > 0$, we can get $\lim_{t\to\infty}\langle\mu_t^d\rangle = \mathcal{J}^d(\theta)$ and $\lim_{t\to\infty}\langle\omega_t^d\rangle = \omega_\theta^d$ over $\{\sup_{t\to\infty}\|x_t^d\| \le M\}$. Based on Lemma 3 and (18), for any $i \in \mathcal{N}$, we further obtain $\lim_{t\to\infty}\mu_{i,t}^d = \mathcal{J}^d(\theta)$ and $\lim_{t\to\infty}\omega_{i,t}^d = \omega_\theta^d$ a.s. Hence, the proof is complete.

Assumption 7. There exists a local projection operator $\mathbb{Y}_i : \mathbb{R}^{p_i} \to \Theta_i \subset \mathbb{R}^{p_i}$ in the update of $\theta_{i,t}$, that projects any $\theta_{i,t}$ onto the compact set Θ_i. And $\Theta = \prod_{i=1}^N\Theta_i$ is large enough to include at least one local minimum of $\mathcal{J}(\theta)$.

Additionally, we define a vector as $\check{\mathbb{Y}}_i[h(\theta)] = \lim_{\eta\to 0^+}\{\mathbb{Y}_i[\theta_i+\eta h(\theta)] - \theta_i\}/\eta$, for any $\theta \in \Theta$ and continuous function $h : \Theta \to \mathbb{R}^{\sum_{i\in\mathcal{N}}p_i}$.

Theorem 3. *Under Assumptions 1, 4–7, the $\theta_{i,t}$ in (15) converges a.s. to a point in the set of asymptotically stable equilibria of*

$$\dot{\theta}_i = \check{\mathbb{Y}}_i\Big[\int_{\mathcal{S}\times\mathcal{A}_{-i}}\varsigma_\theta(ds_t)d\pi_{\theta_{-i}}(a_{-i,t}|s_t)A_{i,t}\Psi_{i,t}\Big],\ \forall i \in \mathcal{N}. \tag{24}$$

Proof. Let $\mathcal{P}_{t,2} = \sigma(\theta_\tau, \tau \le t)$ be the σ-algebra. Besides, denote $\xi_{i,t+1}^{(1)} = A_{i,t}\Psi_{i,t} - \mathbb{E}_{s_t\sim\varsigma_{\theta_t}, a_{-i,t}\sim\pi_{\theta_{-i,t}}}\big(A_{i,t}\Psi_{i,t}|\mathcal{P}_{t,2}\big)$ and $\xi_{i,t+1}^{(2)} = \mathbb{E}_{s_t\sim\varsigma_{\theta_t}, a_{-i,t}\sim\pi_{\theta_{-i,t}}}\big[(A_{i,t} - A_{i,\theta_t})\Psi_{i,t}|\mathcal{P}_{t,2}\big]$, where A_{i,θ_t} is defined as $A_{i,t}$ by setting $\theta = \theta_t$. Hence, the update of actor (15) can be rewritten as

$$\theta_{i,t+1} = \mathbb{T}_i\Big\{\theta_{i,t} + \beta_{\theta,t}\big[\mathbb{E}(A_{i,t}\Psi_{i,t}|\mathcal{P}_{t,2}) + \xi_{i,t+1}^{(1)} + \xi_{i,t+1}^{(2)}\big]\Big\}. \tag{25}$$

First, it is easy to know $A_{i,t} \to A_{i,\theta_t}$, i.e. $\lim_{t\to\infty}\xi_{i,t+1}^{(2)} = 0$ by the critic step converges in Theorem 2. Then, we denote $\{Z_{i,t}\}$ as a martingale sequence with $Z_{i,t} = \sum_{\tau=0}^t\beta_{\theta,\tau}\xi_{i,\tau+1}^{(1)}$. According to Lemma 3, Assumptions 1 and 5, the sequence $\{\xi_{i,t}^{(1)}\}$ is bounded. By Assumption 6, it follows that a.s.,

$$\sum_{t=0}^{\infty}\mathbb{E}\big(\|Z_{i,t+1} - Z_{i,t}\|^2|\mathcal{P}_{t,2}\big) = \sum_{t\ge 1}^{\infty}\Big\|\beta_{\theta,t}\xi_{i,t+1}^{(1)}\Big\|^2 < \infty.$$

From the martingale convergence theorem [27], we obtain the martingale sequence $\{Z_{i,t}\}$ converges a.s. Also, from implicit function theorem, it follows that $\mathbb{E}(A_{i,t}\Psi_{i,t}|\mathcal{P}_{t,2})$ is continuous in θ_t. Ultimately, according to Lemma 4 (see in Appendix), it can be obtained that (25) converges a.s. to the set of asymptotically stable equilibria of (24). Thus, the proof is established.

6. Algorithm implementation

Next, simulation is conducted for the system with 3 microgrids whose schematic is shown in Fig. 3. We first reveal the relevant parameter settings.

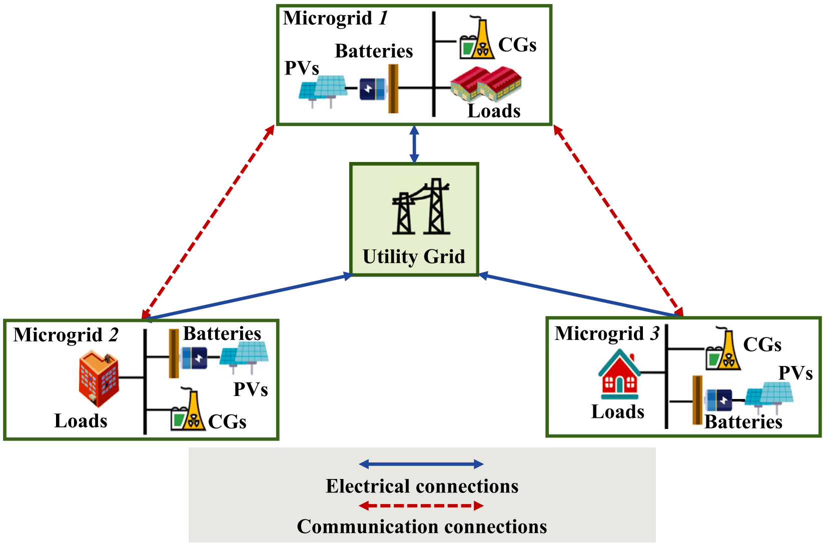


Fig. 3. Schematic diagram of multi-microgrid system.

Table 1
Physical parameters setting of DGs and Batteries.

Label	Microgrid 1		Microgrid 2		Microgrid 3	
	DG1	Battery1	DG2	Battery2	DG3	Battery3
α_i^d	0.0083	0.0152	0.0079	0.0158	0.0096	0.0172
β_i^d	5.76	5.61	5.67	5.63	5.79	5.72
γ_i^d	64	24	357	34	104	37
$p_{i,\min}$	0	−100	0	−100	0	−100
$p_{i,\max}$	225	100	260	100	200	100

Table 2
Emission parameters of pollution gas.

Type	Amercement	Emission rate		
		DGs	BESS	PVs
CO_2	0.0210	72.46	0	0
SO_2	1.4827	0.0004	0	0
NO_x	6.3456	0.0200	0	0

6.1. Experimental setup

The red dotted line presented in Fig. 3 determines the communication topology, i.e., the corresponding Laplacian matrix D in the algorithm. The time series data, from 2015 to 2019, in Open Power System Data [28] is utilized to simulate the load demand, PV generation power and electricity prices. Note that we use the first three years (2016–2019) data as the training set and the last year (2019) as the test set. Specifically, all the data are adjusted to an appropriate scale in accordance with the requirements of the simulation system.

Without loss of generality, each day is divided into 24 time intervals, i.e., $\mathcal{T} = 24$. The parameters of DGs and Batteries including cost coefficients, device capacity and pollutant emission are listed in Tables 1–2, respectively. To avoid battery damage caused by deep charging and discharging, we limit SOC_i to $[0.2, 0.8]$ in the experiment. It is also assumed that the battery has an efficiency of $\rho_{ch} = \rho_{dis} = 0.98$.

As for RL, the stepsizes are set to $\beta_{\omega,t} = 1/t^{0.65}$ and $\beta_{\theta,t} = 1/t^{0.85}$, which satisfy Assumption 6. We employ a discount rate of $\gamma = 0.99$. The penalty coefficient λ is set to 50. Both actor and critic have two hidden layers with the Tanh function and Relu function for activation of the output layer, respectively. The simulation is executed on a computing platform with an Intel Xeon Gold 6226R Processor 22M Cache 2.90 GHz. All the algorithms are implemented in Python 3.11 with the deep learning package Pytorch 2.2.1.

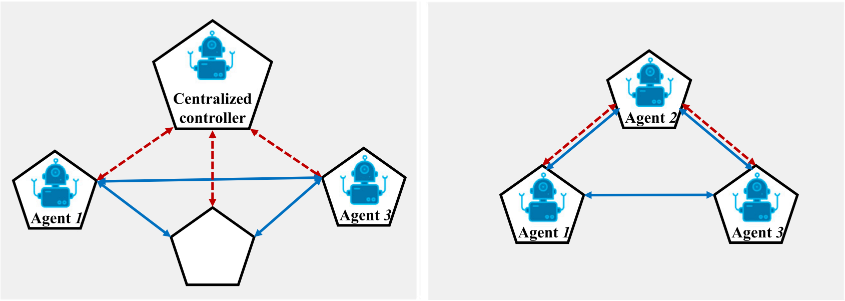
6.2. Performance evaluation of the distributed MARL algorithm

Within the framework of distributed learning, each agents utilize their local data to train the model, engaging solely in communication with neighbors. The topological diagrams for both centralized and distributed algorithms are depicted in Fig. 4. Additionally, we delve into the trade-off that exists between centralized and distributed algorithmic approaches. The centralized algorithm can manage and dispatch specific goals to achieve better cooperation and optimization of the entire system. Nevertheless, centralized architecture mainly relies on a centralized controller, which requires collecting information from all nodes and incurs expensive communication costs. Not only is information privacy easily threatened during data transmission, but the entire system will crash when the controller is attacked. However, if agents utilize local information to independently make decisions, the environment is non-stationary from the perspective of each agent. This implies that even if an agent takes the same action in the same state, the distribution of state transitions and rewards obtained may constantly change, making it difficult to learn the global optimal scheduling strategy. A distributed structure for communicating with neighbors can achieve coordinated global goals while reducing communication costs. The interaction information used in this algorithm is the parameter of neural network, which not only protects the privacy of microgrid data but also avoid using a centralized controller, making it have a certain degree of anti-attack ability.

To contrast the performance of the above topology, Fig. 5 reveals the total rewards of the two structures respectively. The reward curve shows that the centralized algorithm has the best performance due to its ability to obtain global information for model training. In addition, our proposed distributed algorithm exhibits comparable performance to the centralized algorithm, but the total reward is slightly lower than that of centralized algorithms. The reason is that the centralized algorithm can directly obtain global information, while the distributed algorithm can only indirectly obtain global information through information exchange between neighbors. However, distributed algorithms not only can reduce communication costs but also solve privacy protection issues between microgrids. As a consequence, there is a trade-off between centralized and distributed situations.

6.3. Analysis of dispatch results

The polices learned through our algorithm are applied to ensure the optimal dispatch of each microgrid. As illustrated in Fig. 6(a), the



(a) Centralized structure.

(b) Distributed structure.

Fig. 4. The graph of centralized and distributed algorithms.

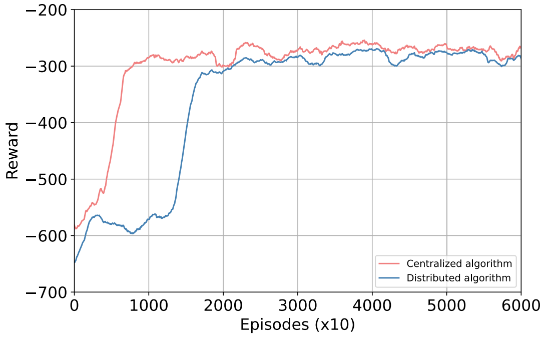


Fig. 5. The total reward of centralized and distributed algorithms.

output power of DG 1 would shift with the change of load 1. Although the load demands of microgrid 1 are higher than the capacity of DG 1, i.e., the demands are unsatisfied during the whole day. Owing to microgrid 1 working in a power self-insufficient circumstance, external power from other microgrids and the main utility grid is needed. The agent learns that energy interactions are more economical than generating power through its own DG and batteries, which is why it does not fully operate its devices all the time. Fig. 6(c) and 6(e) show the scheduling policies of DG 2 and 3 which are similar to DG 1. The difference is that the capacity of DG 2 can fulfill its load while the capacity of DG 3 only satisfies the load 3 in the most time. Due to the maximum capacity of DG 2 and its ability to fully meet its own load, it often generates more power to supply other microgrids. DG 3 generates more power when its load demand is low to help other microgrids. When its load demand exceeds its capacity, it prioritizes the use of excess energy from other microgrids.

In Fig. 6(b), 6(d) and 6(f), the charging and discharging strategies of batteries 1, 2 and 3 are presented. From Table 1, it is clear that the economic cost of batteries is higher than that of power generation in all microgrids. Therefore, agents do not always fully operate their batteries but only use them during the PV power generation period. Fig. 7(a) reveals that the SOC of all batteries can be strictly maintained within the range to avoid over charging and discharging. The power price and quantity purchased from the main utility grid are shown in Fig. 7(b), and we can see that agents purchase power when prices are low and sell excess power when prices are high.

6.4. Comparative algorithm

To exhibit the performance capabilities of our algorithm, other algorithms for comparison in this simulation are given as follows. Algorithm 1: the distributed MARL algorithm with task decomposition. Algorithm 2: the distributed MARL algorithm without task decomposition. Algorithm 3: the centralized MARL algorithm with task decomposition. The comparison of convergence results of each algorithm is illustrated in Fig. 8. It is worth noting that the algorithm we proposed converges after approximately 20000 episodes and outperforms the Algorithm 2 in terms of total rewards. As the proposed algorithm does not require setting weight values, it can be seen from Fig. 8(a) that the designed actor-critic architecture enables the agent to obtain higher rewards.

Table 3
Scheduling results of the environmental objective.

Algorithm	Amercement ($)	Pollutant gas emission (t)		
		CO_2	SO_2	NO_x
Algorithm 1	22955.38	1008.60	0.0056	0.2784
Algorithm 2	24531.79	1077.91	0.0060	0.2975
Algorithm 3	21300.34	935.92	0.0052	0.2583

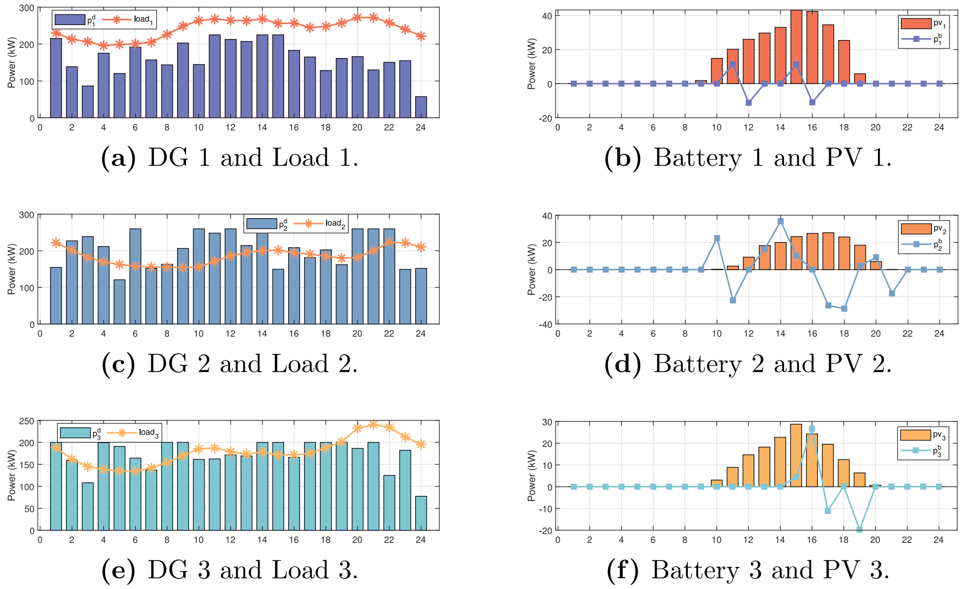
Table 4
The cost on one testing days of the comparative algorithms.

Algorithms	Economic cost ($)	Environmental cost ($)	All cost ($)
Algorithm 1	608847.45	22955.38	631802.83
Algorithm 2	–	–	635571.77
Algorithm 3	606054.54	21300.34	627354.88
ADMM	–	–	702471.46

Moreover, the atmospheric pollutant emissions and environmental objective costs of the scheduling plans generated by all algorithms are presented in Table 3. It can be observed that Algorithm 1 achieves a 6.43% lower cost compared to Algorithm 2 and a 7.78% higher cost compared to Algorithm 3, respectively. Additionally, the atmospheric pollutant emissions can also be reduced by 6.43% when compared with Algorithm 2.

The costs of our proposed algorithms, along with three existing methods, i.e., Algorithm 2, Algorithm 3, and the Alternating Direction Method of Multipliers (ADMM), on a single testing day are compared in Table 4. Specifically, our algorithm achieves 5.93% and 10.06% lower cost compared to Algorithm 2 and the ADMM method, respectively. Furthermore, Algorithm 1 exhibits a 7.09% higher cost relative to Algorithm 3. Results indicate that, while our proposed method falls slightly behind Algorithm 3 in performance, it surpasses the ADMM method. Centralized algorithms have access to larger datasets, enabling them to learn from a wide range of examples. Simultaneously, centralized models utilize more computing resources to efficiently process large amounts of data. However, distributed algorithms address issues of computing and communication costs, as well as privacy concerns, by allocating data and computational responsibilities to multiple agents. Consequently, there is a trade-off between centralized and distributed structures.

Meanwhile, it is apparent that the convergence speed is slightly lower than Algorithm 2 and 3 in Table 5. These results indicate that in the multi-objective optimal dispatch problem of microgrid systems with strong uncertainties, the improved algorithm achieves a fine balanced trade-off between training speed and strategy performance among agents.



(a) DG 1 and Load 1.

(b) Battery 1 and PV 1.

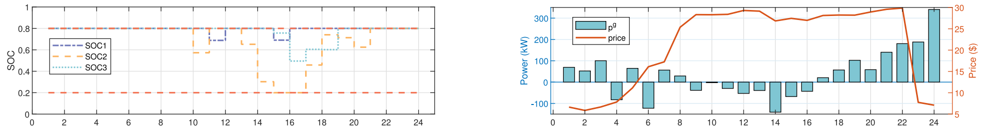
(c) DG 2 and Load 2.

(d) Battery 2 and PV 2.

(e) DG 3 and Load 3.

(f) Battery 3 and PV 3.

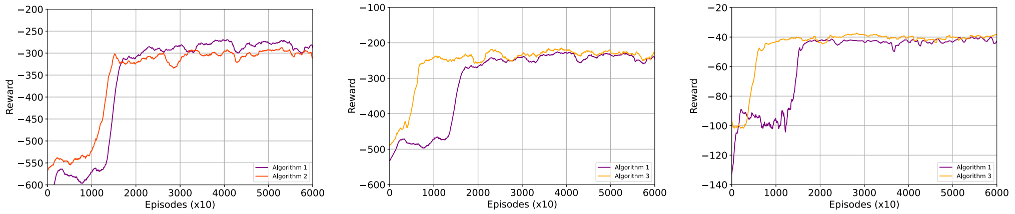
Fig. 6. Dispatch strategy of DGs and batteries obtained by the proposed algorithm.



(a) SOC.

(b) Purchase power and price.

Fig. 7. Scheduling of SOC and purchase power.



(a) Algorithm 1 and 2.

(b) Algorithm 1 and 3.

(c) Algorithm 1 and 3.

Fig. 8. Convergence performance comparison of the algorithms during training.

Table 5
Computational performance of the comparative algorithms.

	Algorithm 1	Algorithm 2	Algorithm 3
Structure	Distributed	Distributed	Centralized
Number of episodes	60 000	60 000	60 000
Average training time per episode (s)	0.96	0.92	0.9

7. Conclusion

In this paper, the multi-objective optimal dispatch of multi-microgrid with multiple heterogeneous power supply devices is re-formulated as a distributed multi-objective optimization problem that takes the economic and environmental objectives into consideration. Afterward, in light of the requirements of achieving multiple objectives, the distributed MARL algorithm with task decomposition is put forward, which shows a novel paradigm to solve this kind of problem. With the help of the consensus strategy in two steps, privacy and convergence are capable to be guaranteed, simultaneously. Ultimately, both theoretical analysis and performance simulation experiments have been conducted to validate the practicability of the designed algorithms. Further, we will concentrate on the distributed optimal dispatch of microgrids with competitive agents over the network, which results in a game theory-based formulation of MARL. Additionally, we aim to explore the scalability and efficiency of the algorithm, particularly in the context of larger-scale systems.

CRediT authorship contribution statement

Xiaowen Wang: Writing – original draft. **Shuai Liu:** Conceptualization. **Qianwen Xu:** Investigation, Formal analysis. **Xinquan Shao:** Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is funded by National Natural Science Foundation of China (NSFC) under Grants 62133008, 62373226.

Appendix. Kushner-Clark lemma

Define an operator $\Gamma : \mathbb{R}^N \rightarrow \mathbb{R}^N$ that projects a vector onto a compact set $\mathcal{S} \subseteq \mathbb{R}^N$. Let $\hat{\Gamma}$ be

$$\hat{\Gamma}[g(s)] = \lim_{\eta \rightarrow 0^+} \left\{ \frac{\Gamma[s + \eta g(s)] - s}{\eta} \right\},$$

for any $s \in \mathcal{S}$ and with $g : \mathcal{S} \rightarrow \mathbb{R}^N$ continuous. We consider the recursion as follows

$$s_{t+1} = \Gamma \left\{ s_t + \gamma_t \left[g(s_t) + \xi_t^{(1)} + \xi_t^{(2)} \right] \right\}. \tag{26}$$

Consequently, the ODE associated with (26) is given by

$$\dot{s} = \hat{\Gamma}[g(s)]. \tag{27}$$

Assumption 8. For the recursion (26), there holds

1. The function $h(\cdot)$ is continuous.
2. The stepsize γ_t satisfies $\sum_{t\rightarrow\infty} \gamma_t = \infty$ and $\sum_{t\rightarrow\infty} \gamma_t^2 < \infty$.
3. The sequence $\{\xi_t^{(1)}\}$ is a bounded random sequence with $\xi_t^{(1)} \rightarrow 0$ a.s. as $t \rightarrow \infty$.
4. The sequence $\{\xi_t^{(2)}\}$ is convergent a.s.

The Kushner-Clark Lemma [29] will be introduced as follows.

Lemma 4. *Under Assumption 8, if ODE (27) has a compact set $\mathcal{S}^*$ which comprises its asymptotically stable equilibria, it follows that s_t in (26) converges to $\mathcal{S}^*$ a.s. as $t \rightarrow \infty$.*

References

[1] Zhao D, Zhang C, Sun Y, Li S, Sun B, Li Y. Distributed robust frequency restoration and active power sharing for autonomous microgrids with event-triggered strategy. IEEE Trans Smart Grid 2021;12(5):3819–34.

[2] Liu S, Han S, Zhu S. Reinforcement learning based energy trading and management of regional interconnected microgrids. IEEE Trans Smart Grid 2022;14(3):2047–59.

[3] Lin C, Hu B, Shao C, Niu T, Cheng Q, Li C, et al. An analysis of delay-constrained consensus-based optimal algorithms in virtual power plants. ISA Trans 2022;125:189–97.

[4] Yan S, Gu Z, Park JH, Xie X, Sun W. Distributed cooperative voltage control of networked islanded microgrid via proportional-integral observer. IEEE Trans Smart Grid 2024;15(6):5981–91.

[5] Li D, Yu L, Li N, Lewis F. Virtual-action-based coordinated reinforcement learning for distributed economic dispatch. IEEE Trans Power Syst 2021;36(6):5143–52.

[6] Lin L, Guan X, Peng Y, Wang N, Maharjan S, Ohtsuki T. Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy. IEEE Internet Things J 2020;7(7):6288–301.

[7] Gazijahani FS, Ravadanegh SN, Salehi J. Stochastic multi-objective model for optimal energy exchange optimization of networked microgrids with presence of renewable generation under risk-based strategies. ISA Trans 2018;73:100–11.

[8] Zhang H, Yue D, Dou C, Hancke GP. PBI based multi-objective optimization via deep reinforcement elite learning strategy for micro-grid dispatch with frequency dynamics. IEEE Trans Power Syst 2023;38(1):488–98.

[9] Gu Y, Shen M, Park JH, Wang Q-G, Wang ZH. Dynamic guaranteed cost event-triggered-based anti-disturbance control of T-S fuzzy wind-turbine systems subject to external disturbances. IEEE Trans Fuzzy Syst 2024;32(12):7063–72.

[10] Kampouropoulos K, Andrade F, Sala E, Espinosa AG, Romeral L. Multiobjective optimization of multi-carrier energy system using a combination of ANFIS and genetic algorithms. IEEE Trans Smart Grid 2018;9(3):2276–83.

[11] Azaza M, Wallin F. Multi objective particle swarm optimization of hybrid micro-grid system: A case study in Sweden. Energy 2017;123(15):108–18.

[12] Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge: MIT Press; 1998.

[13] Perez J, Germain-Renaud C, Kégl B, Loomis C. Multi-objective reinforcement learning for responsive grids. J Comput 2010;8(3):473–92.

[14] Leite GMC, Jiménez-Fernández S, Salcedo-Sanz S, Marcelino CG, Pedreira CE. Solving an energy resource management problem with a novel multi-objective evolutionary reinforcement learning method. Knowl-Based Syst 2023;280:111027.

[15] Yang N, Li X, Huang Y, Xiao M, Wang Z, Song X, et al. Hierarchical multi-agent deep reinforcement learning for multi-objective dispatching in smart grid. In: China automation congress. 2021, p. 4714–9.

[16] Abid MS, Apon HJ, Hossain S, Ahmed A, Ahshan R, Lipu MSH. A novel multi-objective optimization based multi-agent deep reinforcement learning approach for microgrid resources planning. Appl Energy 2024;353:122029.

[17] Liu W, Zhuang P, Liang H, Peng J, Huang Z. Distributed economic dispatch in microgrids based on cooperative reinforcement learning. IEEE Trans Neural Netw Learn Syst 2018;29(6):2192–203.

[18] Dai P, Yu W, Wen G, Baldi S. Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions. IEEE Trans Ind Inf 2020;16(4):2258–67.

[19] Sun C, Liu W, Dong L. Reinforcement learning with task decomposition for cooperative multiagent systems. IEEE Trans Neural Netw Learn Syst 2021;32(5):2054–65.

[20] Zhang K, Yang Z, Liu H, Zhang T, Başar T. Fully decentralized multi-agent reinforcement learning with networked agents. In: International conference on machine learning. 2018, p. 5872–81.

[21] Veldhuizen DAV, Lamont GB. Evolutionary computation and convergence to a pareto front. In: Late-breaking papers at the genetic programming conference. 1998, p. 221–8.

[22] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, arXiv:1707.06347.

[23] Kar S, Moura JM, Poor HV. QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus+innovations. IEEE Trans Signal Process 2013;61:1848–62.

[24] Deng C, Guo F, Wen C, Yue D, Wang Y. Distributed resilient secondary control for DC microgrids against heterogeneous communication delays and DoS attacks. IEEE Trans Ind Electron 2022;69(11):11560–8.

[25] Zhang H, Yue D, Dou C, Xie X, Li K, Hancke GP. Resilient optimal defensive strategy of TSK fuzzy-model-based microgrids' system via a novel reinforcement learning approach. IEEE Trans Neural Netw Learn Syst 2023;34(4):1921–31.

[26] Borkar VS. Stochastic approximation: A dynamical systems viewpoint. Cambridge University Press; 2008.

[27] Neveu J. Discrete-parameter martingales. Elsevier; 1975.

[28] Open power system data platform. 2020, https://data.open-power-system-data.org/. [Accessed October 2020].

[29] Kushner HJ, Clark DS. Stochastic approximation methods for constrained and unconstrained systems. Springer Science & Business Media; 1978.