A Preference-based Online Reinforcement Learning with Embedded Communication Failure Solutions in Smart Grid

Yibing Dang, *Student Member, IEEE*, Jiangjiao Xu, *Member, IEEE*, Dongdong Li, *Member, IEEE*, Hongjian Sun, *Senior Member, IEEE*

***Abstract*—Microgrids, marked by substantial renewable energy integration, have garnered significant attention. Traditional optimization methods face challenges in handling the unpredictability of renewable energy, market prices, and loads. Reinforcement Learning (RL) offers a solution by learning from historical data. However, Offline-RL models often encounter adaptability challenges in new environments, while traditional Online-RL faces stability issues in certain scenarios. To address these concerns, this paper proposes a preference-based Deep Deterministic Policy Gradient (PDDPG) algorithm. It guides the online learning process using expert experience based on expert rules and Offline-RL to enhance the model's performance. Moreover, unlike studies that assume perfect communication and ignore Random Communication Failures (RCF), the proposed real-time microgrid energy management (MEM) system tackles communication challenges by incorporating a Communication Detection and Data Supplement System (CDDSS), especially during extreme weather conditions. The results indicate that the incorporation of CDDSS, as opposed to zero value supplementation, previous moment data, and conventional predictive models, results in a remarkable reduction of microgrid losses by 88.3%, 80.5%, and 53.4%, respectively.**

***Index Terms*—Online microgrid energy management, online reinforcement learning, extreme weather conditions, communication failures.**

I. Introduction

As global climate change intensifies and fossil fuel resources gradually deplete, the importance of renewable energy is increasingly coming to the fore. According to forecasts by the International Renewable Energy Agency, renewable energy is projected to meet over 80% of the world's electricity demand by 2050, with solar and wind energy contributing to 52% of this capacity [1]. To meet this demand, microgrids, composed of local loads, Energy Storage Systems (ESS), Renewable Distributed Resources, and other Controllable Distributed Generations, are emerging as a promising solution. Microgrid energy management (MEM) usually involves multiple optimization variables and optimization objectives. In recent years, model-based optimization methods, such as Mixed Integer Linear Programming, and model-free heuristic approaches, like Genetic Algorithms, have achieved significant success [2].

These MEM methods primarily target scheduling a day ahead, where each unit's operation is predetermined based on predictions of future states. However, a key challenge in optimizing microgrid operations is the significant uncertainty and randomness in renewable energy generation, along with unpredictable market prices and electricity demand. Accurately forecasting these uncertainties is often impractical [3]. Moreover, there are shortcomings in modeling, parameter control, and dynamic adaptation [4]. Model-free reinforcement learning (RL) offers a solution that does not rely on prior state prediction. RL agents learn the optimal decisions corresponding to each possible state of the microgrid by training with historical data. Utilizing the knowledge and experience gained, they can make real-time decisions based on the current actual state of the microgrid.

Energy management strategies for microgrids based on RL have been extensively studied [5]. For example, Alabdullah et al. [6] and Xiao et al. [7] employed RL algorithms using Deep Q-Networks (DQN) and an improved sampling strategy DQN, respectively, for scheduling various energy sources within microgrids to minimize the costs of grid exchanges and conventional power generation. However, these discrete action-based RL methods cannot achieve continuous energy control in microgrids and face convergence difficulties when the number of actions increases, limiting their applicability in real-world microgrids [8]. In contrast, policy gradient-based RL methods optimize the policy directly, demonstrating significant advantages in handling continuous control, better sample efficiency, and a more stable training process, and have been widely explored in microgrid energy management [9]–[14]. Zhang et al. [9] formulated a multi agent RL method to manage multi-energy microgrids, enhancing the stability of privacy resilience and achieving exceptional performance in reducing energy costs. In real-time microgrid energy management systems, solutions involving multiple agents face challenges in real-time performance and convergence. Guo et al. [10] and Lee et al. [11] utilized Deep RL algorithms based on Proximal Policy Optimization (PPO) for microgrid energy scheduling, achieving continuous energy dispatch and validating the effectiveness of the proposed methods through case studies. Dong et al. [12] introduced a multi-energy flow

Manuscript received Jun 02, 2024; revised Oct 04, 2024; accepted Nov 14, 2024. Date of current version Nov 19, 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) (52377111). (Corresponding author: Jiangjiao Xu.)

Yibing Dang, Jiangjiao Xu and Dongdong Li are with the Department of Electrical Engineering, Shanghai University of Electric Power, Shanghai, China (e-mail: jiangjiao.xu@shiep.edu.cn).

Hongjian Sun is with the Department of Engineering, Durham University, Durham, DH1 3LE, UK (e-mail: hongjian.sun@durham.ac.uk).

coordination optimization scheduling system based on the Soft Actor-Critic (SAC) for optimal scheduling of renewable energy and flexible loads. Xu et al. [13] employed the Deep Deterministic Policy Gradient (DDPG) RL method to address energy management issues, achieving better real-time control performance than model predictive control methods. Liu et al. [14] used DDPG for real-time economic energy management in microgrids, effectively utilizing energy storage systems to reduce operating costs. However, these studies are all based on offline reinforcement learning (offline-RL), where the model is trained on offline data and then directly applied to real microgrid energy management without any further updates or adjustments. Real microgrid environments are typically dynamic, and fixed models based on offline-RL can suffer from performance degradation in new environments due to sample bias, making efficient energy management challenging [15].

Online reinforcement learning (online-RL) offers a solution that can update and adjust the model according to changes in the new environment. For example, Yi et al. [16] and Adibi et al. [17] applied online-RL methods to provide continuous frequency regulation services for virtual power plants and microgrids, respectively, continuously updating the trained control policies in real environments to improve tracking accuracy compared to offline-RL. In the field of energy management, Du et al. [18] proposed an online-RL control framework for vehicle energy management strategies, using adaptive optimization methods to update neural network weights, achieving better SOC control performance than the original policy network. Meng et al. [19] introduced an online RL-based microgrid energy management optimization model, continuously updating the model to adapt to uncertainties. However, both [18] and [19] utilized traditional Q-value-based RL methods, which are not suitable for continuous control in microgrid environments. Zhang et al. [20] proposed a DDPG-based online deep RL energy management strategy to further reduce fuel consumption and improve the adaptability of algorithms for series hybrid electric tracked vehicles. Although online-RL approaches have great promise in microgrid energy management scenarios, challenges remain to be addressed. One pitfall is that traditional online updating methods can sometimes lead to irreversible performance declines under specific settings, which could result from distribution shifts between offline and online phases and changes in learning dynamics due to algorithm transitions. Furthermore, there is a problem of low sample efficiency when exploring updates in new environments [21], [22]. Therefore, directly performing simple online updates on models for real-time energy management in microgrids lacks stability and efficiency. Expert experience is a general principled approach that can guide agent exploration and updates, enhancing the learning process [23]. Li et al. [24] fine-tuned DDPG-based autonomous driving agents using human-preferred driving trajectories to provide personalized control solutions for different types of users. Ying et al. [25] employed an expert-guided method to help agents find better policies more efficiently. These works demonstrated the guiding role of human-preferred expert knowledge in the RL agent updating process.

Moreover, we note that in RL-based MEM systems, due to the requirement for real-time state information to formulate the most appropriate control strategies, there is a high demand for communication, where data loss can easily lead to irrational decisions by the agent, resulting in significant energy waste [26]. Smart grid applications have diverse requirements for bandwidth, latency, and reliability, particularly in remote mountainous areas and isolated microgrids, where packet loss may occur when agents receive status information from distributed energy sources. Furthermore, in the harsh environments of Extreme Weather Events (EWE), communication systems are more susceptible to Random Communication Failures (RCF), highlighting the need for robust and adaptive control strategies in microgrid management [27]. Zhou et al. [28] proposed a belief-based correlated equilibrium strategy for joint action selection in multi-agent systems to manage microgrid energy during communication failures. However, the proposed methods rely on retrospective experiences to handle communication failures, making it challenging to address the randomness of the RCF problem in new environments, especially under extreme weather conditions. Moreover, the proposed approaches are not suitable for continuous control scenarios. Existing RL-based real-time energy management research often assumes perfect communication, which is risky in practice. To our knowledge, there is still a lack of exploration regarding the RCF problem in RL-based continuous MEM system.

In this paper, we design a novel preference-based online real-time microgrid energy management framework. This approach combines online RL with expert knowledge, aiming to integrate high-quality human-preferred expertise to more effectively guide the agent's online learning process and enhance the performance of real-time microgrid scheduling models. Additionally, we incorporate the possibility of communication failures into the agent's real-time scheduling operations, investigating and improving the performance of the energy management system under RCF conditions.

The main contributions of this paper are as follows:

- We model the MEM problem as a series of improved Markov decision processes, utilizing a simpler and more accurate Energy Storage System (ESS) model based on the linear approximation of energy content limits, as well as a precise continuous control model for various energies. The proposed Online-MEM achieves real-time scheduling independent of prediction.
- To overcome the adaptability constraints of traditional offline models, as well as the sample inefficiency and safety concerns associated with conventional online models in current MEM research, this paper proposes an enhanced preference-based online Deep Deterministic Policy Gradient (PDDPG) algorithm. The method leverages integrated expert knowledge, combining expert rules with high-quality historical trajectories to guide the agent's online learning process. This ensures more stable and safer model updates. Empirical results demonstrate that the proposed approach consistently surpasses both offline-DDPG and online-DDPG across multiple real-world scenarios.

- To address the challenges posed by communication failures (RCF) in real-time energy management systems, this paper introduces a Communication Detection and Data Supplement System (CDDSS). The system identifies RCF occurrences and compensates for missing data, proposing four data supplementation strategies tailored to RCF conditions. Among these, Strategy IV, which incorporates Transformer models and Transfer Learning, is shown to be the most effective, reducing additional losses in microgrids by 88.3% compared to conventional methods that use zero-value imputation. The proposed system significantly enhances the resilience and efficiency of microgrid operations under RCF scenarios.

The rest of this paper is organized as follows. Section II introduces the overall architecture of the microgrid system under study. Section III describes the process of modeling the microgrid as an MDP, followed by a detailed presentation of the proposed PDDPG and CDDSS. Section IV presents experimental setup discusses the experimental results. Conclusions are drawn in Section V.

II. Microgrid Modeling and Problem Formulation

In this study, a typical microgrid comprises uncontrollable renewable distributed generators (photovoltaics and wind turbines), controllable distributed generators (CDG), energy storage systems (ESS), and loads. The microgrid can exchange power with the public grid when connected. The objective of MEM is to minimize the real-time operational costs for each period, which include the generation costs of CDG, the costs of buying and selling electricity with the main grid, and the operational costs of ESS. This section first introduces the cost models for CDG and power trading, followed by a discussion on essential foundational constraints designed to ensure the safe and stable operation of the system.

A. System Modeling

1) System Costs: The electricity costs of the microgrid primarily include the fuel costs of CDG, the costs of purchasing electricity from the main grid, and the revenue from selling electricity. The expenditure of a CDG can usually be represented by a quadratic cost function:

$$C_{CDG}(t) = a + bP_{CDG}(t) + cP_{CDG}(t)^2 + S_{CDG}(t)H_{CDG} \tag{1}$$

where $C_{CDG}(t)$ is the generation cost, $P_{CDG}(t)$ is the output power of the generator at time t, and a, b, and c are coefficients of the cost function. H_{CDG} represents the start-up costs of the generator and $S_{CDG}(t)$ represents the state of the generator. When $S_{CDG}(t)$ equals 1, it means the generator has started. When $S_{CDG}(t)$ equals 0, it means the generator has not started.

The costs of purchasing and selling electricity can be described as:

$$C_{BS}(t) = (P_{buy}(t)p(t)) - (P_{sell}(t)p(t)) \tag{2}$$

where $C_{BS}(t)$ represents the net cost or revenue of electricity trading at the current moment for the microgrid. $P_{buy}(t)$ denotes the amount of electricity purchased from the main grid at time t, and $p(t)$ indicates the electricity price at time t. Similarly, $P_{sell}(t)$ represents the amount of electricity sold to the main grid at time t. This formula calculates the total net expense or revenue by assessing the difference between the costs of purchasing electricity and the income from selling electricity at different times.

B. The Constraints

The system's constraints primarily include the output constraint of the CDG, the power exchange constraint with the main grid, and the operational constraints of the ESS. These constraints ensure the safe and effective operation of the system. When the generator is operational, it is subject to output constraints:

$$P_{CDG}^{min} \leq P_{CDG}(t) \leq P_{CDG}^{max} \tag{3}$$

where $P_{CDG}(t)$ represents the DG's output at time t. P_{CDG}^{min} and P_{CDG}^{max} are the minimum and maximum output limits of the generator, respectively.

$$\frac{|P_{CDG}(t) - P_{CDG}(t-1)|}{\Delta t} \leq \frac{\Delta P_{\text{max}}^{R}}{\Delta t} \tag{4}$$

where $P_{CDG}(t)$ and $P_{CDG}(t-1)$ are the output powers of the generator at times t and $t-1$ respectively, Δt is the time interval between the two time points, and $\Delta P_{\text{max}}^{R}$ is the maximum allowable power change within the time interval Δt.

To avoid congestion in transmission lines, the energy exchange constraint with the main grid can be expressed as follows:

$$0 \leq P_{\text{buy}}(t) \leq S_{BS}(t)P_{\text{buy}}^{\text{max}} \tag{5a}$$
$$0 \leq P_{\text{sell}}(t) \leq (1 - S_{BS}(t))P_{\text{sell}}^{\text{max}} \tag{5b}$$

where $P_{buy}(t)$ is the amount of electricity the microgrid purchases from the main grid, and $P_{sell}(t)$ is the amount sold to the main grid. BS denotes the direction of power exchange, with $S_{BS}(t)$ equal to 1 indicating buying and BS equal to 0 indicating selling.

$$E_{SoC}^{min}(t) \leq E_{SoC}(t) \leq E_{SoC}^{max}(t) \tag{6a}$$
$$E_{SoC}(0) = E_{SoC}(T) \tag{6b}$$

where Eq. 6a outlines the desired range for the ESS's state of charge (SoC), where $E_{SoC}(t)$ indicates the current SoC. $E_{SoC}(t)^{min}$ and $E_{SoC}(t)^{max}$ are the desired minimum and maximum SoC, set to approximately 20% and 90% in this study. Eq. 6b ensures that the desired SoC is the same at the start and end of the dispatch cycle.

III. The Proposed Enhanced Framework for MEM

This section provides a detailed exposition of the proposed real-time energy management framework for microgrids. Building on the basic microgrid architecture introduced earlier, different microgrid settings are described as a series

of Markov Decision Processes (MDP). Subsequently, the online Preference-based Deep Deterministic Policy Gradient (PDDPG) model is introduced. Finally, strategies for addressing communication failures in real-time energy management for microgrids are presented.

A. Modeling Energy Management as MDPs

The energy management problem in microgrids can be conceptualized as an MDP. An MDP comprises states (S), actions (A), a reward function (R), and state transition probabilities (P). Within an MDP, the goal is to find a policy that maximizes the expected cumulative reward in a given microgrid environment. At a specific time t, S_t represents the description of the microgrid environment, A_t denotes the most valuable action taken by the agent given the state S_t, and R_t is the immediate reward or penalty received by the agent for taking action A_t in state S_t.

1) The State in Microgrid Management: We define the system state as:

$$S_t = \{t, P_{t,PV}, P_{t,Wind}, P_{t,load}, E_{t,ess}, p_t\} \tag{7}$$

where t represents the current time step (a day is divided into 24 hours), $P_{t,PV}$ and $P_{t,Wind}$ denote the output power of PV and wind power generation, respectively, $P_{t,load}$ indicates the total power consumption of the load. $E_{t,ess}$ signifies the remaining electric energy in the ESS, p_t represents the real-time prices for selling to and buying from the main grid, respectively.

2) Precision-Enhanced Continuous Action: This paper utilizes an improved continuous action space for more precise energy control in microgrid, ensuring that the number of action variables grows linearly with the number of batteries and generators, enhancing practicality. The action space is defined as follows:

$$\begin{aligned} A &= \{\xi_t^1, \xi_t^2 \mid t \in T, \xi_t \in (-1, 1)\} \\ &= \{\Delta P_{t,ch}, \Delta P_{t,dc}, \Delta P_{t,CDG}\} \end{aligned} \tag{8}$$

where ξ_t^1 and ξ_t^2 represent the control variables for ESS and CDG, $\xi_t^1 > 0$ denotes charging, and $\xi_t^1 < 0$ denotes discharging. $\Delta P_{t,ch}, \Delta P_{t,dc}, \Delta P_{t,CDG}$ represent the specific ESS charging power, ESS discharging power and generating power, respectively. Moreover, to accommodate the minimum exchange power constraints, a small constant ϵ is introduced to clip the continuous action values. At each time t, the charging and discharging power of the batteries can be calculated using a transformation coefficient a as follows:

$$\begin{cases} \Delta P_{t,dc} = \xi_t \cdot a, \text{ if } -1 < \xi_t^1 \leq -\epsilon \\ \Delta P_{t,dc} = \Delta P_{t,ch} = 0, \text{ if } -\epsilon < \xi_t^1 < \epsilon \\ \Delta P_{t,ch} = \xi_t^1 \cdot a, \text{ if } \epsilon \leq \xi_t^1 < 1 \end{cases} \tag{9}$$

The power of the CDG can be calculated using a transformation coefficient b as follows:

$$\begin{cases} \Delta P_{t,CDG} = 0, \text{ if } \xi_t^2 < \epsilon \\ \Delta P_{t,ch} = \xi_t^2 \cdot b, \text{ if } \epsilon \leq \xi_t^2 < 1 \end{cases} \tag{10}$$

3) Improved ESS State Transition: Most existing MEM research utilizes a simplistic ESS model, primarily considering basic state transitionst [10], [13]. However, as ESS plays a vital role in MEM as a critical control component, a more comprehensive model is required for a better approximation of the charging and discharging processes. Therefore, this study employs a more detailed lithium-ion battery as the ESS model, aiming to enhance the representation and operational dynamics of energy storage within MEM systems. It maintains the simplicity of a linear model while offering a more accurate approximation of energy limits [29]. The energy state of the battery for each interval can be calculated as follows:

$$E(t) = E(t-1) + \Delta E(t) \tag{11}$$

$$\Delta E(t) = \begin{cases} \eta_{ch} P(t) \Delta t & \text{if } P(t) \geq 0 \\ \eta_{dc} P(t) \Delta t & \text{if } P(t) < 0 \end{cases} \tag{12}$$

where $E(t)$ represents the estimated energy content of the battery at the end of the current interval t, $E(t-1)$ denotes the estimated energy content at the end of the previous interval, and $\Delta E(t)$ signifies the change in energy within the battery during interval t, calculated based on power and charging/discharging efficiencies. $\Delta P(t)$ indicates the power applied to the battery during interval t, with η_{ch} and η_{dc} representing the efficiencies of charging and discharging, respectively (assumed constant here). A positive value of $\Delta P(t)$ indicates charging, while a negative value indicates discharging.

The battery must adhere to maximum charging and discharging power constraints, denoted as $P(t)_{ch}$ and $P(t)_{dc}$, respectively. These constraints are as given by the formula:

$$\begin{aligned} &P(t)_{dc} \leq P(t) \leq P(t)_{ch} \\ &P(t)_{dc} = n\alpha_{dc} V(t) \\ &P(t)_{ch} = n\alpha_{ch} V(t) \\ &V(t) = \begin{cases} V_{nom,ch} & \text{if } P(t) \geq 0 \\ V_{nom,dc} & \text{if } P(t) < 0 \end{cases} \end{aligned} \tag{13}$$

where $V(t)$ represents the current voltage approximation of the battery, determined based on its charging or discharging state. $V_{nom,ch}$ and $V_{nom,dc}$ denote the nominal voltages during charging and discharging, respectively. Additionally, the battery must satisfy constraints on energy content limits as functions of the current. The formula is as follows:

$$u_1 I(t) + v_1 n < E(t) < u_2 I(t) + v_2 n \tag{14}$$

where $I(t) = \frac{P(t)}{V(t))}$ denotes the current during the battery's charging/discharging, and $v1, v2, u1, u2$ are constant parameters associated with the battery's energy content constraints.

4) Comprehensive Reward Function Design: An effective reward function can prompt the agent to make prudent decisions regarding the stability and efficiency of the microgrid, thereby optimizing overall system performance. In this study, the optimization objectives encompass not only the minimization of energy costs and maximization of renewable energy utilization but also the optimization of energy storage system operations. Based on these objectives and the constraints

outlined in the preceding section, a novel three-part reward function has been devised:

$$R_t = r_{1t} + r_{2t} \tag{15a}$$
$$r_{1t} = -(C_{CDG}(t) + C_{BS}(t)) \tag{15b}$$

where r_{1t} represents the total cost of the system (a negative value is used since the goal in MDP is to maximize rewards, thereby minimizing costs). $C_{CDG}(t)$ and $C_{BS}(t)$ respectively denote the cost of CDG power generation and the expenses for buying and selling electricity, as indicated in Eqs. 1 and 2. To reduce training complexity, a linear function is used to approximate the cost $C_{CDG}(t)$ [2], as shown below:

$$C_{CDG}(t) = \begin{cases} a_1 \cdot P_{CDG}(t) + b_1 & \text{if } 0 \le P_{CDG}(t) \le P_1 \\ a_2 \cdot P_{CDG}(t) + b_2 & \text{if } P_1 < P_{CDG}(t) \le P_2 \\ a_3 \cdot P_{CDG}(t) + b_3 & \text{if } P_2 < P_{CDG}(t) \le P_{\text{max}} \end{cases}$$

$$r_{2t,1} = \begin{cases} -\xi_1 |E_{SoC}(t) - E_{SoC}^{min}(t)|, \text{if } E_{SoC}(t) < E_{SoC}^{min}(t) \\ -\xi_2 |E_{SoC}(t) - E_{SoC}^{max}(t)|, \text{if } E_{SoC}(t) > E_{SoC}^{max}(t) \\ \xi_3, \text{if } E_{SoC}^{min}(t) \le E_{SoC}(t) \le E_{SoC}^{max}(t) \end{cases}$$

$$r_{2t,2} = \begin{cases} 0, & \text{if } t \neq 24 \\ -\xi_4 |E_{SoC}(T) - E_{SoC}(0)|, & \text{if } t = 24 \end{cases} \tag{16}$$

where $r_{2t,1}$ and $r_{2t,2}$ signify the reward and penalty associated with the operational status of the energy storage system, corresponding to Eqs. 6a and 6b. ξ_1 , ξ_2, ξ_3 and ξ_4 represent the coefficients for rewards and penalties, both being positive constants.

B. Preference-based DDPG Model

Unlike traditional offline-to-online reinforcement learning approaches, the preference-based Deep Deterministic Policy Gradient (PDDPG) proposed in this paper utilizes integrated expert experience data to guide agent learning, aiming to enhance the agent's performance in real-world environments, as illustrated in Fig. 1.

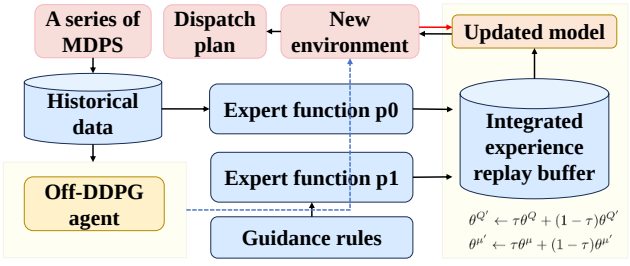


Fig. 1: The overall architecture of PDDPG. Blue arrows represent the process of traditional Offline RL, while red arrows represent the process of traditional Online RL.

1) Integrated Expert Experience Pools: Inspired by [30], to accelerate the online-DDPG agent for real-time microgrid energy management systems, we propose an enhanced framework that leverages expert experience to guide the online learning process. This includes expert knowledge obtained from offline RL and human preference-based expert experience derived from expert rules.

Due to the complexity of system dynamics and the adequacy of data, expert experiences acquired through offline RL and those derived from expert rules tend to execute differently. The accuracy of their Q-value estimates can vary across different states based on the data distribution in the historical dataset. Therefore, combining both is a natural idea. Expert experiences obtained through offline RL are aimed at more effectively utilizing historical data, filtered and generated through the expert function p_0 to form the expert experience pool $\mathcal{M}^{p_0}$:

$$\mathcal{M}^{p_0} = \{Traj_1, Traj_2, ..., Traj_N | Z_t(Traj_i) \ge Z_t(p_0), i \in N\} \tag{17}$$

where $Traj_i$ represents a trajectory in the offline data, comprising a series of information from the scheduling initial time to the ending time, Z_t denotes the cumulative return function designed to filter high-quality historical data for reuse.

The expert experience pool $\mathcal{M}^{p_1}$ is described by expert rules $R_{rule}(s)$, typically set by domain experts or grid operators based on past experiences and domain expertise. These rules do not provide optimal control actions for a given state, but rather offer suggestions based on human preferences that can be considered as guidance or constraints. They are taken into account as sequences of state-action pairs. For instance, actions such as increasing generation, reducing purchase, increasing sales, or decreasing sales might be suggested depending on specific circumstances. For each state, expert rules can generate a set of action candidates:

$$Y_s^{p_1} = \{a | a \in R_{rule}(s), a \in A, s \in S, t \in T\} \tag{18}$$

Based on the action candidate set $Y_s^{p_1}$ for each state s, the corresponding expert experience pool $\mathcal{M}^{p_1}$ is obtained through the constraints and state transition rules in the MDP:

$$\mathcal{M}_s^{p_1} = \{(R_s^{p_1}, D_s^{p_1}, S_{next,s}^{p_1}) | Y_s^{p_1}\} \tag{19}$$

where $R_s^{p_1}$ represents the reward function set, $D_s^{p_1}$ denotes the set indicating whether the current scheduling cycle is ending, and $S_s'^{p_1}$ is the set of next state values. By dynamically adjusting the mixing coefficient, control over the weights from the expert experience pool is exercised to enhance the updating performance of the DDPG agent within a narrowed MDP scope.

2) Online Updating Process based on Expert Experience for PDDPG: Online updates utilize the integrated expert experiences pool to guide the learning process of the agent, achieving stable and secure performance improvements. During each update, the agent randomly selects a batch of experiences from the integrated expert experience pool $\mathcal{M}^{p_0}$ and $\mathcal{M}^{p_1}$ for model updating. In the proposed PDDPG framework, the update process includes the Actor network update guided by mixed expert experience, the Critic network update, and the target network update [25].

$$Traj_i = \{(s_i, a_i, r_i, s_{i+1}) | (s_i, a_i, r_i, s_{i+1}) \in \mathcal{M}^{p_1} \cup \mathcal{M}^{p_2}\}$$
$$y_i = r_i + \gamma Q'(s_{t+1}, \mu'(s_{t+1}) | \theta^{\mu'}) \tag{20}$$

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$
$$L = \frac{1}{N} \sum_i (y_i - Q(s_i, a_i|\theta^Q))^2 \tag{21}$$

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{22}$$

where θ^Q and θ^μ represent the parameters of the Critic and Actor networks respectively, $\theta^{Q'}$ and $\theta^{\mu'}$ denote the parameters of the target Critic and Actor networks, the discount factor γ is used to calculate the current value of future rewards, and the soft update parameter τ is employed to smoothly update the parameters of the target networks.

C. Communication Detection and Data Supplement System (CDDSS)

Random Communication Failures (RCF), including data loss, can easily lead to irrational or incorrect decisions, resulting in significant energy wastage. For example, RCF can transform a data packet $x(t)$ into $p[bx(t) + (1-b)x(t-T)]$, where p represents the probability of successful transmission, b is the packet loss rate, and T is the sampling time [31].

The core of the Communication Detection System (CDS) focuses on ensuring the timeliness and integrity of data. Initially, the system checks for consistency between the data and its corresponding time tags. Normally, each data packet should arrive on time, accompanied by an accurate time stamp. To counteract the effects of communication delays on the system, a maximum communication delay threshold Δt_{max} has been established. Let t_{arr} denote the data packet's arrival time and t_{exp} the anticipated arrival time. Hence, the communication delay Δt can be defined as:

$$\Delta t = t_{arr} - t_{exp} \tag{23}$$

If Δt surpasses the predefined maximum delay threshold Δt_{max}, that is, $\Delta t > \Delta t_{max}$, the system identifies this as a communication failure. In such scenarios, to ensure the microgrid's stable operation, data supplementation measures are taken. Moreover, the CDS evaluates the quality of incoming data. Should the received data packets be blank or damaged during transit (e.g., incomplete or conspicuously incorrect packets), the CDS also treats these instances as communication failures, implementing data supplementation strategies.

The Data Supplement System (DSS) aims to mitigate the impact of communication problems on microgrid operations by supplementing missing data in the event of RCF detected by the CDS. It is important to note that wind energy systems exhibit greater output fluctuations, especially under extreme weather conditions. Consequently, the DSS focuses particularly on developing solutions for communication failures specific to wind energy data, addressing the challenges posed by the variability in renewable energy sources. We propose four data supplementation methods:

1) Strategy I-Zero Value Strategy (ZVS): When an RCF is detected, missing data points can be assigned a zero value. This method is simple and effective for situations where missing data has a minor impact on the system, allowing temporary zero-value settings to avoid further disruption.

2) Strategy II-Previous Value Strategy (PVS): Microgrid control systems often store data over time. In cases of RCF, the data from the last successful transmission can serve as a substitute for current missing data [32]. This approach assumes stability in data over short intervals and is valid for systems with infrequent or minor short-term data changes.

3) Strategy III-Multi-Head Self-Attention Based Prediction Strategy (MSPS): When microgrid systems experience significant data fluctuations and ongoing communication failures, the previously mentioned data supplementation strategies may not adequately reflect real conditions. Under these circumstances, a prediction-based approach can be utilized to estimate missing data [33]. This method, more intricate than its predecessors, offers more precise estimations, particularly in scenarios of high data variability. Thus, we introduce a prediction model founded on a multi-head self-attention mechanism. It enables parallel processing of different data segments and captures long-range dependencies present in the input data. The training process unfolds as follows: Initially, the raw training data is formatted and dimensionally tailored for the Input Layer. Post-adjustment, it proceeds to the Transformer Encoder Layer:

$$A(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{24}$$

where Q, K, V denote the query, key, and value matrices, respectively, with d_k representing the key's dimension. Subsequent to this, the pooling layer conducts average pooling on the multi-head attention's output. The Fully Connected Layer (FCL) then further manipulates the data:

$$f(x) = ReLU(Wx + b) \tag{25}$$

where W and b symbolize the weights and biases of the layer, respectively. Following this, the Flatten Layer converts the input data into a one-dimensional array. The Concatenate Layer amalgamates this flattened input with the Transformer layer's output. This merged data is then processed through an additional FCL, culminating in an output.

4) Strategy IV-Targeted Prediction Strategies for Extreme Weather Events (TPS-EWE): Data prediction supplementation under extreme weather conditions, particularly in the context of volatile wind energy forecasting, confronts a notable challenge: data scarcity. Extreme weather events, owing to their infrequency, often result in limited data for training prediction models. Conventional prediction models may underperform in these extreme scenarios as they are generally trained on data from more typical and milder climates, potentially failing to capture the intricate dynamics of wind energy output during extreme weather [34]. Hence, we introduce the TPS-EWE, grounded in transfer learning. This approach adapts a model from a generic domain to a specific target domain-extreme weather conditions, modifying existing models to better align with the data features and patterns prevalent in extreme weather.

The TPS-EWE training involves two stages: initial pre-training of the model on a comprehensive dataset encompassing all source data, followed by fine-tuning the pre-trained model using a smaller, target domain-specific dataset (rare data under extreme weather). The pre-training and fine-tuning stages share several model parameters. More precisely, the inaugural training phase aligns with the methodology of Strategy III. During the second phase of training, all layers of the model, barring the ultimate FCL, are designated as non-trainable (effectively freezing these layers). Consequently, their weights remain unaltered in subsequent training sessions. Freezing the majority of the model's layers ensures the retention of the acquired feature representations. Simultaneously, training the final layer facilitates the model's adaptation to the novel dataset specific to the target domain. The update process in the target domain is as follows:

$$L_{tar}(y^t, \hat{y}^t) = \frac{1}{M}\sum_{j=1}^{M}(y_j^t - \hat{y}_j^t)^2 \tag{26}$$

where L_{tar} represents the loss on the target domain dataset, with y^t and $\hat{y}^t$ denoting the actual labels and model predictions in the target domain dataset, respectively. M is the number of samples in the target domain.

$$w_t = w_s - \alpha\frac{\partial L_{tar}}{\partial w_s} \tag{27a}$$

$$b_t = b_s - \alpha\frac{\partial L_{tar}}{\partial b_s} \tag{27b}$$

where w_t and b_t represent the updated weights and biases in the target domain, respectively, while w_s and b_s denote the weights and biases obtained from pre-training in the source domain. α is the learning rate. The CDDSS with TPS-EWE Data Supplement system is illustrated in Fig. 2. The left segment of the Fig. 2 delineates the Strategy IV-TPS-EWE model's training process, while the right segment conveys the operational schematic of CDDSS during instances of RCF. For the sake of brevity, only scenarios devoid of RCF and those involving RCF occurrences within 1 and 2 time steps are presented.

IV. Results and Discussions

In this section, we present the settings of various test experiments, showcase the results, and engage in detailed discussions. Both training and testing data are derived from a year's worth of photovoltaic, wind, and load operation data from a European location, with a time interval of 1 hour, totaling 8760 data points. For training scenario setups, we consider four typical photovoltaic output scenarios, four typical load consumption scenarios, and three typical electricity price scenarios. In each training step, a scenario is randomly selected as the current training task. The comparative experiments include several RL algorithms: Offline Deep Deterministic Policy Gradient (Offline DDPG) [13], which is trained on a fixed historical dataset without any model updates; and traditional Online DDPG (On-DDPG) [20], which interacts with the environment in real-time, continuously updating the policy using current state information. Additionally, advanced

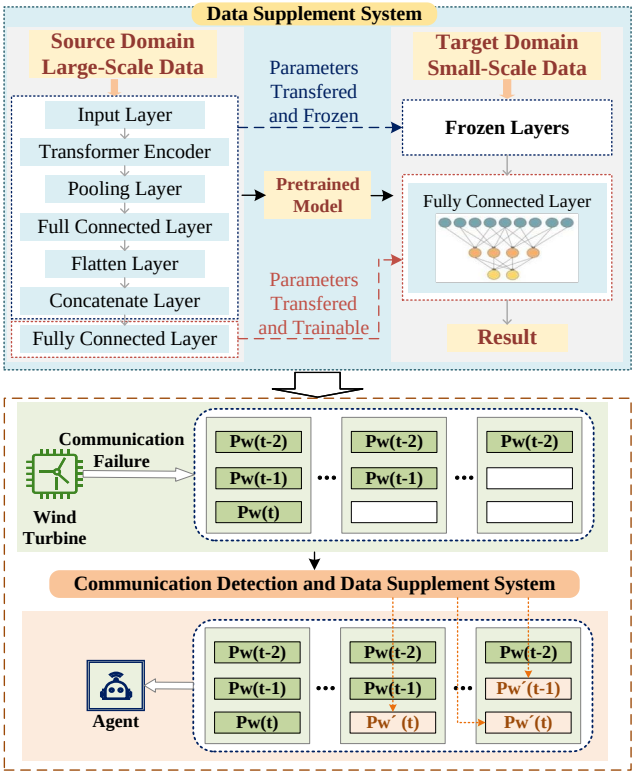


Fig. 2: The framework of CDDSS with Strategy IV.

RL algorithms such as Soft Actor-Critic (SAC) [12] and Proximal Policy Optimization (PPO) [10] are included. SAC enhances performance in complex environments by maximizing expected returns and promoting exploration, while PPO maintains policy stability by limiting the magnitude of each update. Finally, the experiments also introduce Preference-SAC (P-SAC) and Preference-PPO (P-PPO) as comparative methods. For the CDDSS, training data for strategies III and IV consist of 2000 wind turbine operation data points, each containing meteorological data such as wind energy data, temperature, pressure, humidity, etc. Furthermore, the target domain of strategie IV TPS-EWE (Extreme Weather Events) constitutes approximately 5% of the source domain data volume.

A. Online Microgrid Energy Management with PDDPG

We randomly select four 24-hour new scenarios as testing scenarios, each with different scales of renewable energy outputs, load consumption, and electricity price fluctuations. At each time step t (one hour), the agent makes corresponding actions based on the microgrid's current state, without prediction information. The microgrid costs incurred by the four agents in the four scenarios are presented in Table I, with the minimum cost data highlighted in bold. Firstly, we observe that compared to off-DDPG, on-DDPG, based on traditional online updates, does improve performance in certain scenarios. For instance, in scenario 2, it reduces costs by approximately 30%. However, in scenarios 1 and 4, it actually incurs higher costs, indicating that simplistic online updates sometimes lead to worse outcomes. In contrast, PDDPG, guided by expert experience-based updates, consistently exhibits stable performance enhancements. Compared

TABLE I: Cost in Four New Scenarios ($)

Scenario	Method						
	Off-DDPG	On-DDPG	**PDDPG**	SAC	*P-SAC*	PPO	*P-PPO*
1	84.47	104.08	**64.20**	85.82	*71.70*	80.64	*69.23*
2	80.99	67.34	**60.16**	81.34	*64.57*	75.39	*68.69*
3	50.01	51.70	**39.69**	58.78	*45.15*	58.59	*43.25*
4	67.77	167.87	**53.33**	77.87	*57.91*	70.72	*61.32*

to off-DDPG, it reduces costs by 23%, 25%, 20%, and 21% in the four scenarios, respectively. Similarly, PDDPG also demonstrated higher performance compared to the PPO agent and SAC agent. Furthermore, by comparing the scheduling results of the preference-based algorithms P-SAC and P-PPO with those of SAC and PPO agents, it is demonstrated that the preference-based updates of P-SAC and P-PPO result in better performance in terms of cost reduction than the SAC and PPO agents. This further confirms that the proposed preference information can effectively guide the learning process, optimize decision quality, and further reduce energy consumption and operational costs. In summary, our proposed PDDPG agent achieved the best performance in all test scenarios, proving its effectiveness in practical energy management applications for microgrids.

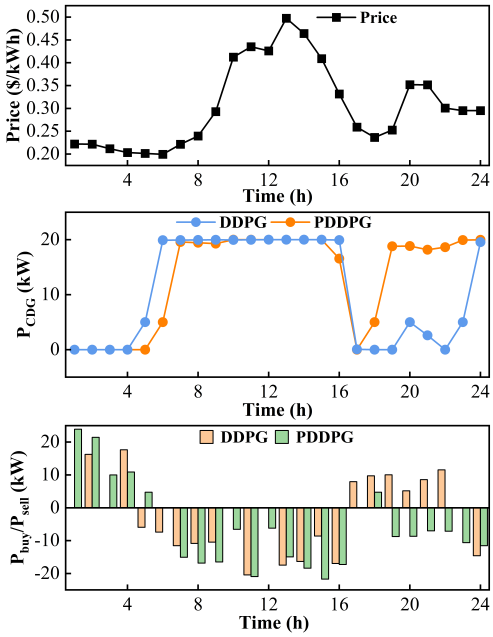


Fig. 3: Comparison of specific strategies of the two agents.

In addition, we compared the specific operational differences between DDPG and PDDPG in energy management. Taking Scenario 1 as an example, the real-time electricity prices over 24 hours, as well as the generation dispatch plan P_{CDG} and the buy/sell electricity plan P_{buy}/P_{sell} for both agents, are illustrated in Fig. 3.

In the real-time generation dispatch plan, during the second peak of electricity prices (from 19 to 24 hours), PDDPG, benefiting from excellent expert guidance, initiated the controllable generators at 17 hours to reduce the amount of purchased electricity during high-price periods, thus saving costs for the microgrid. On the other hand, DDPG's real-time generation plan in the new scenario was evidently inferior, failing to initiate generators during periods of higher electricity prices. Similarly, in the real-time buy/sell electricity plan, PDDPG also demonstrated better performance compared to DDPG: it purchased more electricity during lower price periods (from 1 to 5 hours), sold more electricity during higher price periods (from 9 to 16 hours), and chose a strategy of generating more and selling less electricity rather than generating less and buying more electricity during the higher price periods (from 20 to 24 hours).

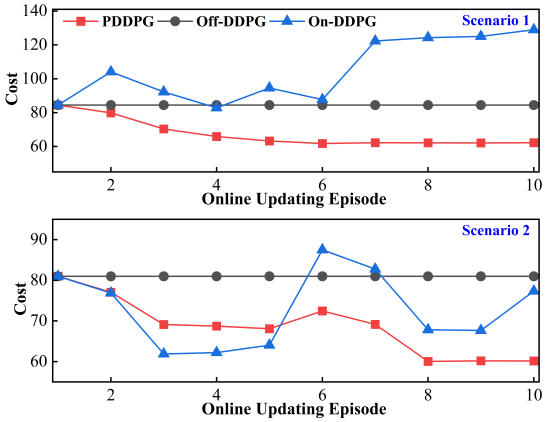


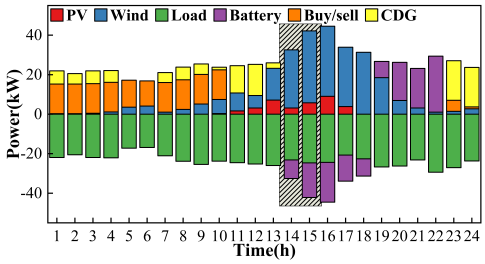
Fig. 4: Performance of three methods in new scenarios.

To provide a clearer analysis and demonstration of the performance of three methods, Fig. 4 illustrates the cost variation during the updating process. Off-DDPG, due to its offline nature, maintains a constant cost. In comparison to PDDPG, it exhibits significant differences, mainly attributed to the offline learning relying on historical data for training, which may introduce sample bias and affect the model's generalization ability in new environments. Additionally, we observe considerable fluctuations in the cost of the On-DDPG agent with real-time updating capability. While it achieves performance improvements in certain cases, it occasionally exhibited poorer results and had lower sample efficiency [35]. The proposed PDDPG shows a notable decrease in cost during the online updating phase while maintaining stability. By integrating expert knowledge with online learning, PDDPG ensures a more robust and effective model update process. It exhibits superior stability and lower costs across multiple test scenarios, validating the effectiveness of leveraging expert knowledge to guide online updates in reinforcement learning.

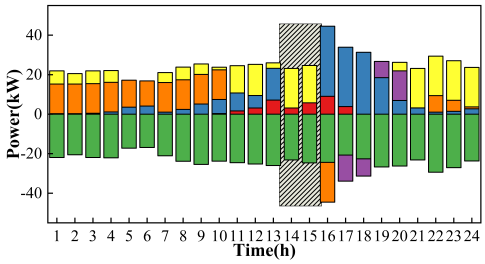
B. Online Microgrid Energy Management with Communication Failures

In this section, we discuss how CDDSS assists the PDDPG agent in scenarios involving RCF. Fig. 5(a) displays the agent's

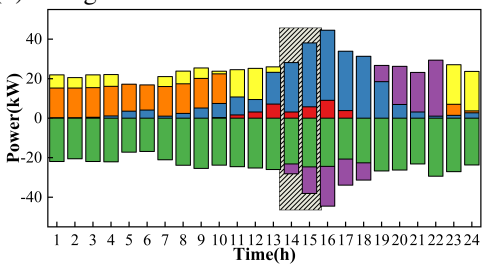
performance of the agent in load-side energy scheduling under perfect communication, where the agent makes various operations based on the current state of the microgrid for rational energy management. In Fig. 5(b), we assume a communication failure at 14:00, resulting in the loss of two time steps of wind energy data, and employ Strategy I-ZVS for data supplementation, setting the wind energy data for these two steps to zero. It is observed that the agent's actions change with the state: in Fig. 5(a), without communication failure, the agent chooses to use part of the renewable energy for load and store the rest in the ESS, which is the desired outcome. However, in Fig. 5(b), after data supplementation with ZVS, the agent, observing insufficient renewable energy for the load demand, opts to use CDG generation unnecessarily, increasing microgrid generation costs and leading to significant economic losses. Furthermore, these RCFs also impact subsequent actions: due to RCFs during times of sufficient renewable energy, the agent did not store excess energy in the ESS, resulting in the inability to use ESS at 21:00 and 22:00, thus incurring additional costs by relying on CDG generation. Fig. 5(c) shows the operation after data supplementation using Strategy IV-TPS-EWE. It is evident that after data supplementation, the agent's energy allocation is almost identical to the previous cases, with only minor differences in quantity, resulting in smaller additional costs.



(a) Without Communication Failure



(b) Using ZVS under Communication Failure



(c) Using TPS-EWE under Communication Failure

Fig. 5: Scheduling strategies without communication failure and with data supplementation during communication failure.

To clearly demonstrate the differences in detailed data, we conducted tests to compare the specific microgrid costs incurred under no RCF and RCF conditions using four data supplementation strategies, as shown in Table II. The first row indicates the perfect communication scenario without RCF, while the first and second columns represent the data values supplemented by the four strategies after RCF occurrences at 14:00 and 15:00. The third column shows the actual costs of the microgrid under various scenarios, and the fourth column details the additional loss costs caused by the four data supplementation strategies under RCF conditions. It is evident that using ZVS for data supplementation is unreliable, leading to the highest costs. Although using PVS results in somewhat lesser additional costs, it can cause even greater losses when there is a significant difference between consecutive data values. However, the proposed MSPS and TPS-EWE strategies result in relatively lower cost losses, especially TPS-EWE, which only causes an additional 3.94% in costs. Compared to the other three methods, it reduces the additional costs by 88.3%, 80.5%, and 53.4%, respectively, verifying the effectiveness of our proposed approach in handling RCFs.

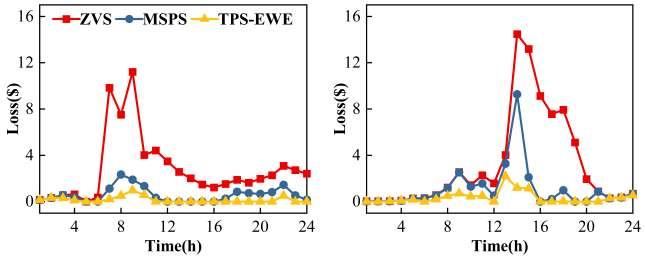
TABLE II: Comparison of Different Strategies Meets RCF

Methods	Pw(t-1)(kW)	Pw(t)(/kW)	Cost($)	Loss($)
No-RCF	29.39	36.41	65.34	-
ZVS	0	0	85.59	20.25
PVS	16.09	16.09	77.54	12.15
MSPS	20.76	28.94	70.45	5.11
TPS-EWE	**25.01**	**32.33**	**67.71**	**2.37**

In reality, due to the unpredictability of communication failures at any given moment, these failures are inherently random, and their duration is also uncertain. Consequently, we assumed the possibility of communication failures throughout the entire 24-hour scheduling period to ensure the randomness of RCF. We tested the additional costs incurred when assuming communication failures at various times within the 24-hour period and using different strategies for data supplementation. Moreover, we found that while PVS could handle some RCF at certain times, it is a risky strategy when data fluctuates significantly, potentially leading to substantial erroneous actions. Therefore, we only considered using Strategy I-ZVS, Strategy III-MSPS, and Strategy IV-TPS-EWE to address RCF. We randomly selected two scenarios for our experiment: a normal weather scenario with relatively stable energy fluctuations, and an extreme weather scenario with more severe energy fluctuations. For simplicity, we assume that the RCF only lasts for one time step. The experimental results are shown in Fig. 6.

As we can see in both scenarios, relying solely on ZVS can lead to significant losses when RCF occurs at each time step. To be more specific, the substantial loss in Fig. 6(a) (between 7-12 hours) is due to high electricity prices coinciding with RCF, leading to the agent mistakenly purchasing electricity to meet load demands. In contrast, the high losses in Fig. 6(b) (between 13-18 hours) are caused by RCF during energy fluctuations. We found that using the MSPS can mitigate most losses due to RCF under normal weather conditions, but some losses are still inevitable under Extreme Weather Events. Comparatively, the proposed TPS-EWE performs best, significantly reducing the additional costs incurred by RCF in both scenarios to a more acceptable level, demonstrating

higher applicability in addressing the RCF problem in real-time microgrid energy management systems.



(a) Normal Weather (b) Extreme Weather

Fig. 6: Comparison of additional cost losses in microgrids under RCF for normal and extreme weather conditions.

To explain the feasibility of the results from an implementation and computational perspective, a detailed analysis of the computational complexity of the proposed method was conducted. For the computational analysis during the real-time scheduling cycle, the strategy model update time for a single scenario and the strategy model testing time during a real-time scheduling cycle (24 hours) were reported. All model training and testing were conducted on a computer equipped with an AMD Ryzen 7 5800H CPU (3.20 GHz), 16 GB RAM, and an Nvidia 3050 Ti GPU. The average computational cost for strategy model updates based on PDDPG, on-DDPG, and off-DDPG was 1.45s, 0.24s, and 0s, respectively, while the average computational cost for online real-time scheduling for all three agents was 0.08s ± 0.01s. The time for data supplementation during communication failures was less than 0.1s. Therefore, from a computational cost perspective, the proposed method demonstrates high feasibility.

V. Conclusion

In this study, we have successfully developed an Online-MEM framework based on the Preference-based DDPG model to guide agent learning through two types of expert experience. Testing results in new scenarios have shown that, compared to traditional Offline-DDPG and Online-DDPG agents, PDDPG can devise more rational scheduling strategies, thereby reducing microgrid expenses. Additionally, a Communication Detection and Data Supplement System (CDDSS) was integrated to detect and address communication failure events in real-time MEM. A data supplementation strategy specifically designed for extreme weather events, utilizing Transformer and Transfer Learning algorithms, was also developed. Experimental results demonstrate that CDDSS effectively mitigates the impact of communication failures, significantly reducing additional losses in microgrids under RCF scenarios. Overall, this study provides a new perspective in the field of online microgrid energy management, with potential for future research to further enhance the adaptability of online energy management systems.

References

[1] P. Ralon, M. Taylor, A. Ilas, H. Diaz-Bone, and K. Kairies, "Electricity storage and renewables: Costs and markets to 2030," *International Renewable Energy Agency: Abu Dhabi, United Arab Emirates*, vol. 164, 2017.

[2] G. S. Thirunavukkarasu, M. Seyedmahmoudian, E. Jamei, B. Horan, S. Mekhilef, and A. Stojcevski, "Role of optimization techniques in microgrid energy management systems—a review," *Energy Strategy Reviews*, vol. 43, p. 100899, 2022.

[3] H. Shuai and H. He, "Online scheduling of a residential microgrid via monte-carlo tree search and a learned model," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1073–1087, 2021.

[4] B. She, F. Li, H. Cui, J. Zhang, and R. Bo, "Fusion of microgrid control with model-free reinforcement learning: Review and vision," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3232–3245, 2023.

[5] D. Vamvakas, P. Michailidis, C. Korkas, and E. Kosmatopoulos, "Review and evaluation of reinforcement learning frameworks on smart grid applications," *Energies*, vol. 16, no. 14, p. 5326, 2023.

[6] M. H. Alabdullah and M. A. Abido, "Microgrid energy management using deep q-network reinforcement learning," *Alexandria Engineering Journal*, vol. 61, no. 11, pp. 9069–9078, 2022.

[7] H. Xiao, X. Pu, W. Pei, L. Ma, and T. Ma, "A novel energy management method for networked multi-energy microgrids based on improved dqn," *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4912–4926, 2023.

[8] B. She, F. Li, H. Cui, J. Zhang, and R. Bo, "Fusion of microgrid control with model-free reinforcement learning: Review and vision," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3232–3245, 2023.

[9] D. Qiu, T. Chen, G. Strbac, and S. Bu, "Coordination for multienergy microgrids using multiagent reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 4, pp. 5689–5700, 2022.

[10] C. Guo, X. Wang, Y. Zheng, and F. Zhang, "Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning," *Energy*, vol. 238, p. 121873, 2022.

[11] S. Lee, J. Seon, Y. G. Sun, S. H. Kim, C. Kyeong, D. I. Kim, and J. Y. Kim, "Novel architecture of energy management systems based on deep reinforcement learning in microgrid," *IEEE Transactions on Smart Grid*, vol. 15, no. 2, pp. 1646–1658, 2024.

[12] Y. Dong, H. Zhang, C. Wang, and X. Zhou, "Soft actor-critic drl algorithm for interval optimal dispatch of integrated energy systems with uncertainty in demand response and renewable energy," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107230, 2024.

[13] G. Xu, J. Shi, J. Wu, C. Lu, C. Wu, D. Wang, and Z. Han, "An optimal solutions-guided deep reinforcement learning approach for online energy storage control," *Applied Energy*, vol. 361, p. 122915, 2024.

[14] D. Liu, C. Zang, P. Zeng, W. Li, X. Wang, Y. Liu, and S. Xu, "Deep reinforcement learning for real-time economic energy management of microgrid system considering uncertainties," *Frontiers in Energy Research*, vol. 11, p. 1163053, 2023.

[15] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[16] Z. Yi, Y. Xu, X. Wang, W. Gu, H. Sun, Q. Wu, and C. Wu, "An improved two-stage deep reinforcement learning approach for regulation service disaggregation in a virtual power plant," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2844–2858, 2022.

[17] M. Adibi and J. van der Woude, "Secondary frequency control of microgrids: An online reinforcement learning approach," *IEEE Transactions on Automatic Control*, vol. 67, no. 9, pp. 4824–4831, 2022.

[18] G. Du, Y. Zou, X. Zhang, L. Guo, and N. Guo, "Energy management for a hybrid electric vehicle based on prioritized deep reinforcement learning framework," *Energy*, vol. 241, p. 122523, 2022.

[19] Q. Meng, S. Hussain, F. Luo, Z. Wang, and X. Jin, "An online reinforcement learning-based energy management strategy for microgrids with centralized control," *IEEE Transactions on Industry Applications*, 2024.

[20] B. Zhang, Y. Zou, X. Zhang, G. Du, F. Jiao, and N. Guo, "Online updating energy management strategy based on deep reinforcement learning with accelerated training for hybrid electric tracked vehicles," *IEEE Transactions on Transportation Electrification*, vol. 8, no. 3, pp. 3289–3306, 2022.

[21] A. Nair, A. Gupta, M. Dalal, and S. Levine, "Awac: Accelerating online reinforcement learning with offline datasets," *arXiv preprint arXiv:2006.09359*, 2020.

[22] I. Uchendu, T. Xiao, Y. Lu, B. Zhu, M. Yan, J. Simon, M. Bennice, C. Fu, C. Ma, J. Jiao *et al.*, "Jump-start reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 34 556–34 583.

[23] J. Ramírez, W. Yu, and A. Perrusquía, "Model-free reinforcement learning from expert demonstrations: a survey," *Artificial Intelligence Review*, pp. 1–29, 2022.

[24] N. Li and P. Chen, "Research on a personalized decision control algorithm for autonomous vehicles based on the reinforcement learning

from human feedback strategy," *Electronics*, vol. 13, no. 11, p. 2054, 2024.
[25] F. Ying, H. Liu, R. Jiang, and M. Dong, "Extensively explored and evaluated actor-critic with expert-guided policy learning and fuzzy feedback reward for robotic trajectory generation," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7749–7760, 2022.
[26] B. Huang, L. Liu, H. Zhang, Y. Li, and Q. Sun, "Distributed optimal economic dispatch for microgrids considering communication delays," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 8, pp. 1634–1642, 2019.
[27] S. Marzal, R. Salas, R. González-Medina, G. Garcerá, and E. Figueres, "Current challenges and future trends in the field of communication architectures for microgrids," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 3610–3622, 2018.
[28] H. Zhou, A. Aral, I. Brandić, and M. Erol-Kantarci, "Multiagent bayesian deep reinforcement learning for microgrid energy management under communication failures," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 11 685–11 698, 2022.
[29] F. Kazhamiaka, C. Rosenberg, and S. Keshav, "Tractable lithium-ion storage models for optimizing energy systems," *Energy Informatics*, vol. 2, pp. 1–22, 2019.
[30] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, "Efficient online reinforcement learning with offline data," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1577–1594.
[31] B. Zhang, C. Dou, D. Yue, Z. Zhang, and T. Zhang, "A cyber-physical cooperative hierarchical control strategy for islanded microgrid facing with random communication failure," *IEEE Systems Journal*, vol. 14, no. 2, pp. 2849–2860, 2020.
[32] J. Xu, H. Sun, and C. J. Dent, "Admm-based distributed opf problem meets stochastic communication delay," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5046–5056, 2018.
[33] K. Qu, G. Si, Z. Shan, X. Kong, and X. Yang, "Short-term forecasting for multiple wind farms based on transformer model," *Energy Reports*, vol. 8, pp. 483–490, 2022.
[34] J. Xu, K. Li, and D. Li, "Multioutput framework for time-series forecasting in smart grid meets data scarcity," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 9, pp. 11 202–11 212, 2024.
[35] S. Lee, Y. Seo, K. Lee, P. Abbeel, and J. Shin, "Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble," in *Conference on Robot Learning*. PMLR, 2022, pp. 1702–1712.

Yibing Dang (S'24) received the B.E. degree in electrical engineering and automation from Shandong University of Science and Technology, Shandong, China, in 2021. She is currently working toward the M.E. degree in electrical engineering with the College of Electrical Engineering, Shanghai University of Electric Power, Shanghai, China. Her research interests include smart grid, machine learning, and the operation and control of power systems.

Jiangjiao Xu (S'15–M'19) received the M.Sc. in Electric Power, School of Engineering, University of Newcastle, Newcastle upon Tyne, U.K., in 2013, the Ph.D. degree in electronic and electrical engineering from the University of Durham, U.K., in 2019 and then took post-doctoral positions with University of Exeter, U.K. (From 2019 to 2022). Since 2022, he has been with the Shanghai University of Electric Power, China, as a Assistant Professor in Smart Grid. His research mainly focuses on smart grid: demand side management and demand response, and renewable energy forecasting, and machine learning and applications in smart grid.

Dongdong Li (M'08) received his B.S. and Ph.D. degrees from Zhejiang University and Shanghai Jiao Tong University both in electrical engineering in 1998 and 2005, respectively. He is currently a professor and dean of College of Electric Engineering in Shanghai University of Electric Power, Shanghai, China. His current research interests include analysis of electric power system, renewable energy system, and smart grid.

Hongjian Sun (S'07–M'11–SM'15) received the Ph.D. degree in electronic and electrical engineering from The University of Edinburgh, U.K., in 2011. He held post-doctoral positions with King's College London, U.K., and Princeton University, USA. Since 2013, he has been with the University of Durham, U.K., as a Professor (Chair) in Smart Grid, where he was an Assistant Professor from 2013 to 2017 and an Associate Professor (Reader) from 2017 to 2020. He has authored or coauthored over 120 articles in refereed journals and over 120 papers in international conferences. He has made contributions to and coauthored the IEEE 1900.6a-2014 Standard. He has authored or coauthored five book chapters and edited two books: *Smarter Energy: From Smart Metering to the Smart Grid* (IET) and *From Internet of Things to Smart Cities: Enabling Technologies* (CRC). His research mainly focuses on smart grid communications and networking, demand-side management and demand response, and renewable energy sources integration. He also served as a Guest Editor for the *IEEE Communications Magazine* and the *IEEE Transactions on Industrial Informatics* for several feature topics. He is an Editor-in-Chief for the *IET Smart Grid* journal and an Editor for *Journal of Communications and Networks*.