Interpretable Deep Reinforcement Learning for Optimizing Heterogeneous Energy Storage Systems

Luolin Xiong, Graduate Student Member, IEEE, Yang Tang, Senior Member, IEEE, Chensheng Liu, Member, IEEE, Shuai Mao, Ke Meng, Senior Member, IEEE,

Zhaoyang Dong, Fellow, IEEE, Feng Qian, Senior Member, IEEE

Abstract—Energy storage systems (ESS) are pivotal component in the energy market, serving as both energy suppliers and consumers. ESS operators can reap benefits from energy arbitrage by optimizing operations of storage equipment. To further enhance ESS flexibility within the energy market and improve renewable energy utilization, a heterogeneous photovoltaic-ESS (PV-ESS) is proposed, which leverages the unique characteristics of battery energy storage (BES) and hydrogen energy storage (HES). For scheduling tasks of the heterogeneous PV-ESS, cost description plays a crucial role in guiding operator's strategies to maximize benefits. We develop a comprehensive cost function that takes into account degradation, capital, and operation/maintenance costs to reflect real-world scenarios. Moreover, while numerous methods excel in optimizing ESS energy arbitrage, they often rely on black-box models with opaque decision-making processes, limiting practical applicability. To overcome this limitation and enable transparent scheduling strategies, a prototype-based policy network with inherent interpretability is introduced. This network employs human-designed prototypes to guide decisionmaking by comparing similarities between prototypical situations and encountered situations, which allows for naturally explained scheduling strategies. Comparative results across four distinct cases underscore the effectiveness and practicality of our proposed pre-hoc interpretable optimization method when contrasted with black-box models.

Index Terms—Heterogeneous energy storage systems, deep reinforcement learning, pre-hoc interpretability.

I. INTRODUCTION

S one of the significant resource, energy storage system (ESS), characterized by their flexibility, are extensively integrated into power systems, and contribute to carbon emission reduction [1–4]. Flexible ESS serves a dual role in the energy market, functioning both as an energy supplier and consumer [5]. One noteworthy application lies in its capacity

This work was supported by National Natural Science Foundation of China (61988101, 62293502, 62293504), Fundamental Research Funds for the Central Universities. (*Corresponding author: Yang Tang, Feng Qian.*)

Luolin Xiong, Yang Tang, Chensheng Liu and Feng Qian are with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, and the Engineering Research Center of Process System Engineering, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mails: xiongluolin@gmail.com, tangtany@gmail.com, cliu@ecust.edu.cn, fqian@ecust.edu.cn).

Shuai Mao is with the Department of Electrical Engineering, Nantong University, Nantong 226019, China (e-mail: mshecust@163.com).

Ke Meng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney NSW 2052, Australia (e-mail: kemeng@ieee.org).

Zhaoyang Dong is with the School of Electrical and Electronics Engineering, Nanyang Technological University, 50 Nanyang Avenue 639798, Singapore (e-mail: zydong@ieee.org). to profit from participation in the energy market through energy arbitrage [6]. Simultaneously, ESS can also support the integration of random renewable energy sources into the energy market, fostering competition with conventional nonrenewable energy producers [7–9]. Thoughtful scheduling of ESS operations can enhance the profitability of ESS operators and maximize the utilization of renewable energy resources, thereby mitigating the inherent uncertainties associated with renewable energy sources [10].

Many research efforts have been devoted to various ESS, such as battery energy storage (BES), hydrogen energy storage (HES), compressed air energy storage, and pumped hydro energy storage [11–14]. These ESS variants exhibit diverse dynamic characteristics. For instance, lithium batteries offer rapid response capabilities, enabling swift charging and discharging, whereas hydrogen and pumped hydro energy storage systems require more extended response times [11]. Furthermore, lithium batteries have constrained capacities, especially when compared to compressed air energy storage and pumped hydro energy storage, which are better suited for large-scale applications [12]. Hydrogen energy storage systems excel in energy density and storage duration [13, 14]. Given the flexibility and variety within ESS, numerous studies have explored the combination of photovoltaic (PV) power stations with ESS to enhance overall energy efficiency [4, 11]. In contrast to prior configurations involving PV-battery storage systems and PV-compressed air energy storage systems, we propose a unique combination of the PV system with both BES and HES as a PV-ESS, which leverages the distinctive characteristics of these heterogeneous energy storage systems, and further augment the revenue potential for the PV-ESS operator.

To treat a PV-ESS as an entity within power systems and optimize the economic profitability for the PV-ESS operator, a critical step involves the development of a realistic cost function. Significant efforts have been dedicated to describing the costs associated with various types of energy storage systems [15–18]. For instance, the degradation cost of BES has garnered considerable attention [15, 16]. Factors such as depth of discharge (DoD), discharge rate, and state of charge (SoC) are recognized as pivotal in influencing battery degradation. [15] has introduced a BES cost model, accounting for degradation cost based on DoD and discharge rate, which is applicable to conventional electrochemical batteries. [16] has captured the intricacies of battery degradation mechanisms and presented a battery degradation model that considers both DoD and SoC. Additionally, the capital cost and the operation/maintenance cost hold substantial importance, particularly for large-scale energy storage equipment. [19] has highlighted the impact of factors such as start-up/shut-down cycles, rapid transients in operational conditions, and the number of working hours on electrolyzers and fuel cells within HES systems. However, different from traditional homogeneous ESS, our proposed heterogeneous ESS poses a unique challenge in comprehensively characterizing various cost components. In this paper, we establish a comprehensive cost function that simultaneously incorporates degradation, capital, as well as operation/maintenance costs to mirror real-world scenarios.

On the other hand, due to the remarkable flexibility of ESS and the inherent uncertainty associated with PV power generation, the development of optimal scheduling strategies for PV-ESS has garnered significant research attention, particularly with the well-defined cost function [20-23]. Traditional optimization approaches for solving the scheduling problem have focused on mathematical programming and heuristic techniques [16, 24-26]. For instance, [16] has implemented a mixed integer linear programming algorithm to address a day-ahead economic scheduling problem of BES systems, where a one-dimensional linearization technique has been used to linearize the two-variable function, reducing computational complexity without sacrificing accuracy. In [25], particle swarm optimization has been employed to obtain an optimal operation schedule for the ESS. Despite notable advancements in these traditional optimization methods, these approaches often hinge on precise models or assumptions about the distribution of random variables such as the PV power generation, which limits their applicability.

Recent advancements in artificial intelligence (AI) have led to the popularity of optimization methods based on reinforcement learning, which are particularly attractive for their independence on precise system models and strong performance, especially in uncertain systems [27-29] and multiagent systems [30, 31]. In [4], a proximal policy optimization (PPO) based deep reinforcement learning (DRL) method has been employed to address capacity scheduling in PV-battery storage systems. This approach has demonstrated adaptability to uncertain market signals as well as PV generation profiles. Meanwhile, [11] has introduced a model-free DRL technique for optimizing energy arbitrage, utilizing a hybrid model to forecast intermittent PV power generation. Nevertheless, the practical application of these AI-powered methods is somewhat constrained due to their opaque decision-making processes.

It has been widely recognized that interpretability is a crucial factor to enhance the practical applicability of reinforcement learning. A few works have explored the interpretability of reinforcement learning when applied to forecasting/optimization problems in energy system [32, 33]. [32] has provided the post-hoc interpretability for the policy network, employing the Shapley value to uncover the significance of various input features in the decision-making process. However, it's worth noting that this post-hoc explanation method is primarily retrospective, offering insights about the black-box model after the decision-making process. Consequently, it does not empower operators to understand the agent's decisionmaking process.

Building upon the aforementioned motivation, this paper introduces an inherently interpretable DRL algorithm with prehoc interpretability tailored for addressing the heterogeneous PV-ESS scheduling problem. This pre-hoc explanation method relies on intuitive human-designed prototypes to guide the decision-making process by comparing similarities between prototypical situations and encountered situations [34]. To achieve this, a prototype-based policy network is trained and integrated with a pre-trained agent, which can shorten the gap between the well-performed black-box model and the interpretable strategies. Our main contributions are summarized as follows:

- A heterogeneous PV-ESS is proposed to leverage the distinctive characteristics of BES and HES, thereby enhancing flexibility within the energy market. The scheduling problem of this PV-ESS is formulated as a Markov Decision Process (MDP), with the primary objective of maximizing benefits of the operator through energy arbitrage. Compared with homogeneous ESS in [4, 11], our approach incorporates a comprehensive cost function for the heterogeneous PV-ESS accounting for degradation, capital, as well as operation and maintenance costs to provide more realistic and practical guidance for operator decision-making.
- An interpretable DRL method is developed to provide pre-hoc interpretability for agent's decision-making process. Different from post-hoc explanation methods outlined in [32], our approach involves training a prototypebased policy network, which enables the guidance of decision-making by assessing similarities between human-defined prototypical situations and encountered situations, thereby rendering the decision-making process transparent. Significantly, this method reveals the correlation between the decisions made by the agent and the comprehensible decisions made by human.
- A comprehensive assessment across four cases, each featuring different PV-ESS configurations, is conducted. The comparative results illustrate the effectiveness and applicability of the proposed pre-hoc interpretable DRL method compared with black-box models. Furthermore, we evaluate the revenue of the PV-ESS operator, considering various scenarios with heterogeneous energy storage devices, and investigate the impact of the learning rate on convergence and optimization.

The remainder of this paper is structured as follows: Section II provides a detailed formulation of the scheduling problem for the heterogeneous PV-ESS. Section III outlines our proposed interpretable DRL method, featuring a prototype-based policy network. We present the simulation results in Section IV, and finally, we conclude the paper in Section V.

II. PROBLEM FORMULATION

As depicted in Fig. 1, our study focuses on energy arbitrage through the coordinated operations of the heterogeneous PV-ESS. The goal is to maximize the revenue of the PV-ESS

operator by actively participating in electricity markets. In this section, we introduce dynamic models and the cost function of the heterogeneous PV-ESS, consisting of BES and HES. Subsequently, accounting for dynamic market prices and uncertainties associated with PV power generation, we formulate the operation scheduling task of the heterogeneous PV-ESS as a MDP. Within this framework, we optimize the charge and discharge operations for both BES and HES.



Fig. 1. Framework of the energy market with a heterogeneous PV-ESS.

A. BES

For the optimal scheduling of BES, the requisite dynamic model for the charge and discharge operations is defined as follows,

$$E_{t+1}^{\text{SoC}} = E_t^{\text{SoC}} + \eta^{\text{Bat}} P_t^{\text{Bat}} \Delta t, \qquad (1)$$

where $E_t^{\rm SoC}$ represents the current SoC of the battery. η^{Bat} signifies the charge/discharge efficiency, where $\eta^{Bat} = 0.9$ for charging and $\eta^{Bat} = 0.95$ for discharging. $P_t^{\rm Bat}$ denotes the charge/discharge power of the BES equipment, with $P_t^{\rm Bat} > 0$ indicating charging and $P_t^{\rm Bat} < 0$ indicating discharging at time t. It's important to note that we use a time interval of $\Delta t = 1$ hour, during which all values are assumed to remain constant.

In accordance with the dynamic model described above, the BES must adhere to the following operational constraints,

$$E_{\min}^{\text{SoC}} \le E_t^{\text{SoC}} \le E_{\max}^{\text{SoC}}, \tag{2a}$$

$$P_{\min}^{\text{Dat}} \le P_t^{\text{Dat}} \le P_{\max}^{\text{Dat}},\tag{2b}$$

where E_{\min}^{SoC} and E_{\max}^{SoC} represent the minimum and maximum energy states of the battery. P_{\min}^{Bat} and P_{\max}^{Bat} are the limits of the charge/discharge power per unit time.

To maximize the benefits of the PV-ESS operator, it is crucial to have an accurate and easily solvable battery cost function that accounts for coupled capital, degradation, and operation costs. Both degradation and operation costs are intertwined with the capital cost. Over time, the battery experiences degradation from its original state, with the degradation rate being dependent on operating characteristics and conditions. Thus, we employ degradation cost to encompass both capital and operation costs. As described in prior studies [15, 17], the degradation cost incurred during operation is significantly influenced by various factors, including battery capacity, SoC limits, environmental temperature, and current. In our BES cost function, we assume the presence of a temperature control system in the environment, thereby neglecting the impact of high temperatures. We also establish appropriate minimum and maximum values for SoC while not considering the effects of SoC limits. Instead, we focus on the degradation associated with DoD and discharge rate during periodic charge and discharge processes within the electricity market framework.

It is essential to highlight that the impact of DoD on degradation costs is influenced not only by the difference SoC at adjacent times but also by the initial and final levels of SoC during discharge process. For instance, discharging from 70% to 0% results in more significant degradation compared to discharging from 100% to 30%. Consequently, we employ SoC instead of DoD in the BES cost function.

Additionally, the discharge rate v_t^{DCR} related to the current can be computed as follows,

$$v_t^{\text{DCR}} = \frac{E_{t-1}^{\text{SoC}} - E_t^{\text{SoC}}}{\Delta t}.$$
(3)

Building on insights from [15] and taking into consideration the above mentioned influencing factors, the BES cost C^{Bat} is expressed as follows,

$$C_t^{\text{Bat}} = \frac{c_{\text{cc}}^{\text{Bat}}}{\text{Cap}\eta_{\text{r}}^2 \phi} ((1 - E_t^{\text{SoC}})^{\omega} - (1 - E_{t-1}^{\text{SoC}})^{\omega}), \quad (4)$$

where $c_{\rm cc}^{\rm Bat}$ represents the BES capital cost of the battery. Cap means the battery capacity, and $\eta_{\rm r}$ signifies the round trip efficiency of the BES. Coefficients ϕ and ω are used to capture the relationship between DoD and the number of cycles.

To simplify the BES cost function and make it readily integrable into a comprehensive cost function of the heterogeneous PV-ESS, the cost function from Eq. (4) can be linearized as follows,

$$C_t^{\text{Bat}} = w_1 E_t^{\text{SoC}} + w_2 E_{t-1}^{\text{SoC}} + w_3 v_t^{\text{DCR}} + w_4, \qquad (5)$$

where w_1 and w_2 are the coefficients of the cost related to DoD. w_3 is the coefficient of the cost related to discharge rate. w_4 is related to battery capacity and serves as a linearization offset term within the degradation cost function [15]. Indeed, it's important to emphasize that the proposed degradation cost function for the BES is time-dependent, as it takes into account factors such as DoD, SoC, and discharge rate, all of which vary with time. These parameters collectively enable a comprehensive and time-sensitive representation of the degradation cost model.

B. HES

In the pursuit of optimizing the scheduling of the HES, we first introduce the composition of the HES and elucidate the energy conversion processes associated with each component. Subsequently, the dynamic model of the HES that govern these energy conversion processes is presented. Finally, we delve into the HES cost function. The HES comprises hydrogen proton exchange membrane fuel cell stacks (FCs), an electrolyzer (EL), and a hydrogen storage reservoir, encompassing the conversion of hydrogen into electricity and vice versa [13]. More specifically, the hydrogen FCs serve as power generation equipment capable of converting the chemical energy stored in hydrogen into electric energy, while the EL can perform the reverse operation. We quantify these conversion relationships using the molar flow of hydrogen and the power output of the EL and FCs, in accordance with Faraday's law [35],

$$F_t^{\mathrm{EL}_{\mathrm{H}_2}} = \eta^{\mathrm{EL}} P_t^{\mathrm{EL}} / \mathrm{NCV}_{\mathrm{H}_2}, \tag{6a}$$

$$F_t^{\rm FC_{H_2}} = P_t^{\rm FC} / \eta^{\rm FC} \rm NCV_{H_2}, \tag{6b}$$

where $F_t^{\rm EL_{H_2}}$ and $F_t^{\rm FC_{H_2}}$ represent the hydrogen molar flow in the EL and FCs, respectively. $P_t^{\rm EL}$ and $P_t^{\rm FC}$ denote the power output of the EL and FCs, while $\eta^{\rm EL}$ and $\eta^{\rm FC}$ characterize the energy conversion rates of the EL and FCs, respectively. $\rm NCV_{H_2}$ represents the net calorific value, which is the effective calorific value obtained by subtracting the heat of water vaporization from the full combustion calorific value.

Based on the above conversion relationship, the state of the hydrogen storage reservoir at the previous time step and the change in the hydrogen molar flow at current time step can be employed to calculate the current state of the hydrogen storage reservoir in the following manner [13],

$$E_{t}^{\text{LoH}} = \left(1 - \eta^{\text{HES}}\right) E_{t-1}^{\text{LoH}} + \frac{\mathcal{R}T_{\text{H}_{2}}}{V_{\text{H}_{2}}} \left(F_{t}^{\text{EL}_{\text{H}_{2}}} - F_{t}^{\text{FC}_{\text{H}_{2}}}\right),$$
(7)

where E_t^{LoH} denotes the pressure of the hydrogen storage reservoir at time t. η^{HES} represents the self-consumption rate of the hydrogen storage equipment. \mathcal{R} , T_{H_2} , and V_{H_2} correspond to the gas constant, mean temperature of the hydrogen storage reservoir, and the reservoir volume, respectively. Similar to the BES, we still assume the existence of a temperature control system, so the mean temperature remains constant.

Additionally, the HES must adhere to the following operational constraints,

$$E_{\min}^{\text{LoH}} \le E_t^{\text{LoH}} \le E_{\max}^{\text{LoH}},$$
 (8a)

$$P_{\min}^{\text{EL}} \le P_t^{\text{EL}} \le P_{\max}^{\text{EL}},$$
 (8b)

$$P_{\min}^{\text{FC}} \le P_t^{\text{FC}} \le P_{\max}^{\text{FC}}, \tag{8c}$$

$$P_t^{\rm EL} P_t^{\rm FC} = 0, \tag{8d}$$

where E_{\min}^{LoH} and E_{\max}^{LoH} are the lower and upper limits for the pressure of the hydrogen storage reservoir. P_{\min}^{EL} and P_{\max}^{EL} impose constraints on the power output of the EL, while P_{\min}^{FC} and P_{\max}^{FC} are limits for the power output of FCs. The final constraint specifies that the EL and FCs cannot operate simultaneously at time t.

Following the depiction of the state transition of the hydrogen storage reservoir, we now turn our attention to the comprehensive cost function of the HES. Much like the BES, the cost of HES incorporates capital, degradation, and operation costs. In contrast to the BES, HES incurs a higher capital cost. It's evident that the latter two costs are intricately tied to the operational scheduling of the HES, which includes factors like runtime, state switching frequency, power output, and current. Inspired by [19], the cost function for the HES can be formulated as follows,

$$C_t^{\text{HES}} = \sum_{i=\text{EL,FC}} \left(\left(\frac{c_{\text{cc}}^i}{\nu^i} + c_{\text{op}}^i \right) \sigma^i + c_{\text{st}}^i \zeta^i + c_{\text{de}}^i \kappa^i \right), \quad (9)$$

where $\sigma^i \in {\sigma^{\text{EL}}, \sigma^{\text{FC}}}$ are binary variables associated with the on/off-status of EL and FCs, where 0 indicates off-status and 1 indicates on-status. ζ^i represent logical variables that account for the start-up state. κ^i is defined as the power variation at instances when EL/FCs are active. c_{cc}^i and ν^i denote the capital acquisition cost for the EL/FCs devices and the total number of working hours. c_{op}^i is the hourly operation cost associated with the maintenance of EL/FCs devices. c_{st}^i and c_{de}^i are utilized to formulate the degradation cost resulting from start-up cycles and high current values during the charge/discharge processes.

C. MDP Formulation

As for the heterogeneous PV-ESS scheduling framework developed in this paper shown in Fig. 1, it comprises an operator and a heterogeneous ESS integrated with PV, which can serve as both an energy supplier and an energy consumer in the energy market. On the supply side, the framework primarily includes the traditional centralized main grid, which relies on thermal power generation. The market electricity prices in this framework are determined by the main grid. On the demand side, there are various users with diverse power requirements, such as municipalities, factories, and individual households. We assume a continuous electricity demand scenario, ensuring that users are constantly in need of electricity from a whole perspective.

It's crucial to emphasize that, to enhance the competitiveness of the PV-ESS in the energy market, its transaction prices consistently remain below market prices. This allows users prioritize purchasing electricity at a lower price from the PV-ESS. The revenue of the PV-ESS operator is derived from the sale of PV power and energy arbitrage. Energy arbitrage entails storing excess PV power or procuring electricity from the main grid when market prices are low and subsequently selling it at a lower price than the market rate when prices rise and electricity demand is high. Typically, the selling price is often higher than the price at which the PV-ESS initially bought electricity from the market.

The operator has access to energy market information, including electricity prices, as well as internal status information about the PV-ESS. This internal status information covers PV power generation, the SoC of the BES, the hydrogen storage level of the HES, and the operational status of each equipment. This framework forms the basis for optimizing operations of the heterogeneous PV-ESS and maximizing its economic profitability.

To design an explainable scheduling strategy for the heterogeneous PV-ESS, the charge and discharge operation scheduling problem can be formulated as a MDP. In this formulation, state transitions depend solely on the previous one step state and not on any memory. The MDP framework comprises four



Fig. 2. Structure of interpretable DRL method with a prototype-based policy network.

key elements: a set of states ($s \in \mathbb{S}$), a set of actions ($a \in \mathbb{A}$), a reward function r, and transition probabilities p from state s and action a to state s' [36]. For the operation scheduling problem, these elements are defined as follows:

1) The state: The state s_t serves as a representation of the current situation of the heterogeneous PV-ESS. In this study, the state encompasses the following elements,

$$s_t = \{ \operatorname{Pr}_t, P_t^{\operatorname{PV}}, E_t^{\operatorname{SoC}}, E_t^{\operatorname{LoH}}, \sigma^{\operatorname{EL}}, \sigma^{\operatorname{FC}} \}, \qquad (10)$$

where \Pr_t represents the dynamic electricity price, and the $P_t^{\rm PV}$ signifies the power output from PV generation. In order to ensure that the EL and FCs do not operate simultaneously, we impose the constraint $\sigma^{\rm EL}\sigma^{\rm FC} = 0$. The observation is denoted as $\{\Pr_t, P_t^{\rm PV}, E_t^{\rm SoC}, E_t^{\rm LoH}\}$.

2) *The action:* Based on the definition of the system state, the actions are defined as follows,

$$a_t = \{P_t^{\text{Bat}}, P_t^{\text{EL}}, P_t^{\text{FC}}\},\tag{11}$$

where P_t^{Bat} , P_t^{EL} , and P_t^{FC} are continuous variables in the action space \mathbb{A} . It is important to take into account the constraints imposed by the battery capacity, hydrogen storage reservoir pressure, the charge/discharge power limitations, and

the power output of the EL/FCs. Consequently, the actual actions are constrained as follows,

$$P_t^{\text{Bat}} = \begin{cases} \min\{P_t^{\text{Bat}}, \frac{1 - E_t^{\text{SoC}}}{\eta^{\text{Bat}} \Delta t}\}, & \text{if } P_t^{\text{Bat}} > 0, \\ \max\{P_t^{\text{Bat}}, \frac{-E_t^{\text{SoC}}}{\eta^{\text{Bat}} \Delta t}\}, & \text{if } P_t^{\text{Bat}} < 0, \end{cases}$$
(12a)

$$\begin{cases} P_t^{\text{EL}} = \min\{P_t^{\text{EL}}, \frac{\Delta E_t^{\text{LoH}} V_{\text{H}_2}}{\mathcal{R} T_{\text{H}_2}}\}, & \text{if } P_t^{\text{EL}} > 0, \\ P_t^{\text{FC}} = \min\{P_t^{\text{FC}}, \frac{\Delta E_t^{\text{LoH}} V_{\text{H}_2}}{\mathcal{R} T_{\text{H}_2}}\}, & \text{if } P_t^{\text{FC}} > 0, \end{cases}$$
(12b)

$$\Delta E_t^{\text{LoH}} = \begin{cases} E_{\text{max}}^{\text{LoH}} - \left(1 - \eta^{\text{HES}}\right) E_t^{\text{LoH}}, & \text{if } P_t^{\text{EL}} > 0, \\ \left(1 - \eta^{\text{HES}}\right) E_t^{\text{LoH}}, & \text{if } P_t^{\text{FC}} > 0. \end{cases}$$
(12c)

Eq. (12a) serves to ensure that the charge and discharge power of the battery do not breach the maximum/minimum capacity limits. Additionally, Eq. (12b) ensures that the hydrogen produced by electrolysis does not exceed the maximum remaining capacity of the hydrogen storage tank, and it also ensures that the hydrogen demand of fuel cell stacks does not exceed the available hydrogen reserve. Eq. (12c) calculates the permissible pressure state change while taking into account the impact of equipment self-consumption.

3) State transition: The system transition at time t can be depicted as Eq. (1) and Eq. (7).

4) *The reward function:* The reward function is designed to quantify benefits of the PV-ESS operator at time *t*, aligning with the optimization objective,

$$r_t = \rho \Pr_t P_t^{\text{sell}} - C_t^{\text{Bat}} - C_t^{\text{HES}}, \qquad (13a)$$

$$P_t^{\text{sell}} = P_t^{\text{PV}} + P_t^{\text{FC}} - P_t^{\text{Bat}} - P_t^{\text{EL}}, \quad (13b)$$

where P_t^{sell} represents the electricity sold to customers. $\rho \in (0, 1]$ signifies the discount rate applied to the market electricity price \Pr_t . For this analysis, ρ is set to 0.95. This choice indicates that it is more advantageous for customers to engage in energy transactions with the PV-ESS operator rather than with the power grid, primarily due to the more favorable electricity prices offered by the PV-ESS operator. Participation in the energy market with the PV-ESS operator clearly leads to improved economic performance for prosumers.

III. PROPOSED APPROACH

In this section, we provide a detailed introduction to the proposed interpretable DRL method, which includes a prototypebased policy network designed for pre-hoc interpretability. We will first delve into the prototype-based policy network, and then offer insights into the human-friendly interpretable DRL method, which demonstrates how the prototype-based policy network enhances the transparency and understandability of the agent's decision-making process.

A. Prototype-based Policy Network

The rapid advancement of DRL across various domains has led to the emergence of interpretable methods to facilitate its real-world application. Currently, the prevalent methods are post-hoc interpretation techniques that provide insights into model predictions over time [32]. While these methods are widely adopted, they may not provide a complete understanding of the agent's decision-making process, as it remains concealed.

Motivated by this, we introduce a prototype-based policy network that transforms a DRL agent from a black-box model into an interpretable model [34]. This approach compels the agent to generate policies that are comprehensible in a humanfriendly manner. The structure of the prototype-based policy network, as applied to the PV-ESS scheduling problem, is depicted in Fig. 2. It comprises a pre-trained agent serving as a coding network, several transformation networks along with their corresponding prototypical states. The similarity score is derived by comparing the prototypes transformed from prototypical states with the actual potential representation of the states. This score is then employed to guide the agent's decision-making process. Notably, the method's interpretability is derived from prototypes based on human experience, which incorporate intuitive and easily understandable actions in the prototypical states. These prototypes, in turn, provide guidance for the actual actions within each dimension.

Remark 1: It's important to note that a pre-trained agent can be acquired using a black-box approach, which can achieve commendable performance. The primary purpose of the prototype-based policy network is to assist the pre-trained agent in rendering its decision-making process transparent and understandable, thereby enhancing the pre-hoc interpretability of the algorithm. Consequently, within the prototype-based policy network, we fully leverage the capabilities of the pretrained agent.

We define the policy derived from the pre-trained agent, based on the black-box model, as π' and assume that this policy can be decomposed into an encoder network \mathcal{F} and a linear layer, implying that $\pi' = W' \mathcal{F}(s) + b'$ [34]. To fully utilize the well-performing black-box model, within the prototype-based policy network, we initially input the state s into the pre-trained encoder network \mathcal{F} to obtain the latent representation $z = \mathcal{F}(s)$. Subsequently, to elucidate the action generation process clearly, separate transformation networks \mathcal{H}_k are introduced for each action dimension, which map the latent representation z of the state s to specific representations z_k for different action dimension k. In particular, for the PV-ESS scheduling problem, the action encompasses three dimensions: the charge/discharge power of the BES, the power output of the EL, and the power output of the FCs. However, to distinguish between the charge and discharge behavior of the BES, the first dimensional action P^{Bat} is divided into separate components for charging and discharging, resulting in a total of four action dimensions. This division enhances the ease of prototype design and facilitates a better understanding of the agent's decision-making process.

$$z_k = \mathcal{H}_k(z), k \in \{1, 2, 3, 4\}.$$

With the networks described above, a set of prototypical states S_k are designed for prototypical actions within each dimension, which are intuitive and human-friendly. These prototypical states are then fed into both the original encoder network \mathcal{F} and the transformation networks \mathcal{H}_k , and used as prototypes $p_k = \mathcal{H}_k(\mathcal{F}(\mathcal{S}_k))$ for the k-th dimension. For the PV-ESS scheduling problem, we design four prototypical states, as illustrated in Fig. 3. One prototypical state represents a typical charging scenario for the BES in an environment characterized by low electricity prices, ample PV power generation, and a low level of battery energy. Conversely, another prototypical state signifies an obvious and intuitive profitable operation for the BES, which is discharging in an environment featuring high market electricity prices, insufficient PV power generation, and a high SoC. Similar situations apply to the EL and FCs within the HES as well.

Utilizing the prototypes p_k mentioned earlier, the similarity between specific representations z_k and prototypes p_k are calculated as outlined in [37]. Subsequently, we introduce a human-defined linear weight matrix W that is employed in combination with the similarity scores to generate actions. This weight matrix W encapsulates the relationship between prototypes and actions, and it provides an intuitive explanation for W_k , which signifies how the prototype p_k should influence the action a_k . This approach ensures that each prototype is associated with an action that is intuitively comprehensible.

$$sim(z_k, p_k) = log\left(\frac{(z_k - p_k)^2 + 1}{(z_k - p_k)^2 + \epsilon}\right),$$
 (13a)

$$a_{t,k} = W_k \sin(z_k, p_k), \tag{13b}$$



Fig. 3. Four prototypical states.

where $\epsilon = 1e^{-5}$ is a hyperparameter utilized to characterize similarity.

In the scheduling problem of the PV-ESS, where P_t^{Bat} is the charge/discharge power, we set $W_1 = 1$ and $W_2 = -1$. This configuration signifies that the charge/discharge power should be equal to the difference between the similarity to prototypical charge and discharge actions. Here, we illustrate with a straightforward example that when the current state of the BES closely resembles the typical BES charging state as shown in Fig. 3, the action P_t^{Bat} becomes strongly associated with charging. Likewise, if the actual state bears similarity to the typical BES discharge state, the learned action leans towards discharge. Consequently, the decision-making process becomes more interpretable as it naturally explains why a specific action is chosen. This approach can be likened to a case-based reasoning strategy, where the decision to take action a is made because the current situation bears similarity to a prior prototypical situation in which action a was also chosen [34].

$$P_t^{\text{Bat}} = W_1 \sin(z_1, p_1) + W_2 \sin(z_2, p_2)$$

Remark 2: It's important to emphasize that, in contrast to previous approaches where prototypes are learned [37], the prototypes in our method are human-defined. This choice is in line with the idea that involving humans in the learning loop can be beneficial, as suggested by [38]. Similarly, the weight matrix W is manually defined rather than learned. Learning W could lead to each prototype corresponding to multiple undesirable actions, which is why we opt for manual specification.

In the prototype-based policy network, only the transformation networks \mathcal{H}_k with parameters ψ can be trained. These networks build upon the pre-trained encoder network \mathcal{F} , the provided prototypical states \mathcal{S}_k , and a manually specified weight matrix W. The training process involves minimizing the loss between the output of the prototype-based policy network a and the action $a' \in \pi'$ obtained from the blackbox model in specific states. The parameters ψ of the transformation networks \mathcal{H}_k are updated using the gradient descent method. The pseudocode for training the prototype-based policy network is presented in the following Algorithm 1.

Algorithm 1 Training the prototype-based policy network								
Input:	А	pre-trained	agent	with	encoder	network	\mathcal{F}	and
pol	icy	π' .						

Output: A well-trained prototype-based policy network.

- 1: Initialize the prototype-based policy network with a manually specified weight-matrix W.
- Sample n state-action pairs from Dataset collected by the pre-trained agent D ← {(s, π'(s))}ⁿ_{i=0}.
- 3: Choose Human-Interpretable Prototypical States $S_k \in \mathbb{S}$.
- 4: for batch $(s, a') \in \mathcal{D}$ do
- 5: $z = \mathcal{F}(s);$
- 6: for $k \in \{1, 2, 3, 4\}$ do
- 7: $a_k = 0;$
- 8: **for** each S_k **do**
- 9: $p_k = \mathcal{H}_k(\mathcal{F}(\mathcal{S}_k))$
- 10: end for
- 11: $z_k = \mathcal{H}_k(z)$
- 12: $a_k = a_k + W_k \sin(z_k, p_k)$
- 13: Minimize Loss $\mathcal{L}(a|a', \mathcal{F}, \psi, W)$ with gradient descent, updating only ψ .
- 14: end for

15: Cache all $p_k = \mathcal{H}_k(\mathcal{F}(\mathcal{S}_k))$ for testing time inference. 16: end for

17: return trained prototype-based policy network.

B. Interpretable DRL Algorithm

Before training a prototype-based policy network, a welltrained black-box model is required. This black-box model is used, in part, as an encoder network \mathcal{F} for the interpretable policy network. The PPO algorithm is employed for pretraining the agent. In this section, we will provide a brief introduction to the PPO algorithm as applied to solve the scheduling problem of the PV-ESS.

The PPO algorithm employs a neural network architecture with shared parameters θ for predicting the policy function and the value function. The loss function used to train the shared network encompasses error terms from the policy surrogate and the value function. Additionally, an entropy term is incorporated into the loss function to promote exploration in the action space. The loss function can be expressed as follows,

$$L_t(\theta) = \hat{\mathbb{E}}_t \left[L_t^{\mathcal{C}}(\theta) - m_1 L_t^{\mathcal{V}}(\theta) + m_2 L_t^{\mathcal{S}} \left[\pi_{\theta} \right| (s_t) \right] \right], \quad (15)$$

where $L_t^{\rm C}(\theta)$ represents the policy surrogate error term, $L_t^{\rm V}(\theta) = (V_{\theta}(s_t) - V_t^{\rm targ})^2$ is the error of the value function, and $L_t^{\rm S}[\pi_{\theta}|(s_t)]$ is the entropy bonus used for exploration. m_1 and m_2 are coefficients for $L_t^{\rm V}$ and $L_t^{\rm S}$, respectively. $L_t^{\rm C}(\theta)$ can be calculated as follows,

$$L_t^{\rm C}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\Upsilon_t \hat{A}_t, \operatorname{clip} \left(\Upsilon_t, 1 - \xi, 1 + \xi \right) \hat{A}_t \right) \right],$$
(16)

where $\hat{A}_t = \delta_t + (\gamma \lambda) \delta_{t+1} + \dots + (\gamma \lambda)^{T-t+1} \delta_{T-1}$ is the advantage function estimated with $\delta_t = r_t + \gamma V(s_{t+1}) - 1$ $V(s_t)$. $\Upsilon_t = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio, and ξ is a hyperparameter.

With the loss function as Eq. (15), the shared network with parameters θ for both policy and value functions can be trained as shown in [39].

IV. EXPERIMENTS

This section outlines the experiments to assess the effectiveness of the proposed interpretable DRL method for solving the heterogeneous PV-ESS scheduling problem. Meanwhile, the revenue of the PV-ESS operator in various cases is also analyzed, considering heterogeneous energy storage devices. The following subsections describe the preliminaries of experiments. Then, the performance of the proposed interpretable DRL method compared with other approaches is verified and corresponding discussions about the pre-hoc interpretability are provided. Lastly, we examine the revenue of the PV-ESS operator with heterogeneous energy storage devices in various scenarios and assess the impact of the learning rate on convergence and optimality.

A. Preliminaries of Case Studies

The experiments are conducted using PV power generation data and time-varying electricity prices data from [40]. We focus on the PV-ESS scheduling problem during a single day, with each hour representing one time slot, resulting in the time horizon T = 24 hours. Considering the one-hour time scale, there are a total of 8760 data sets in a year. As for the heterogeneous PV-ESS, the initial SoC is set to a random value between 0.25 and 1, while the initial level of hydrogen storage (LoH) is set to a random value between 5 and 35. Various parameters used in the experiments are detailed in Table I.

We design four distinct cases, as summarized in Table II, to analyze the impact of the heterogeneous PV-ESS on the operator's revenue, while considering various characteristics and cost structures. The experiments are conducted using Python 3.6.13 with the machine learning library PyTorch 1.8.1.

 TABLE I

 PARAMETERS USED IN THE EXPERIMENTS.

Parameters	Value	Parameters	Value
a	-36.23	$ m_1$	0.5
b	34.80	m_2	0.01
c	2.77	η^{EL}	0.725
d	-2.45	$\eta^{\rm FC}$	0.6
ξ	0.2	η^{HES}	0.05
$\tilde{\lambda}$	0.95	V _{H2}	35 Nm ³
γ	0.99	$T_{\rm H_2}$	313 K
\mathcal{R}	8.314 J/mol K	NCV_{H_2}	240 MJ/kmol

B. Performance of the Prototype-based Policy Network

Here, we first present several baseline methods employed for comparison with the proposed interpretable DRL method using the prototype-based policy network in the scheduling

TABLE II DESCRIPTION OF EXPERIMENTAL CASES.

Cases	BES		HES	
Cases	existence	cost	existence	cost
Case 1	 ✓ 	\checkmark	\checkmark	\checkmark
Case 2	\checkmark	×	\checkmark	×
Case 3	\checkmark	\checkmark	×	×
Case 4	×	×	\checkmark	\checkmark

problem of the heterogeneous PV-ESS, following the presentation of four human-designed prototypes. Then the performance and interpretability of each method are presented, and the results are analyzed to provide insights into the benefits of the proposed interpretable DRL approach. It is remarkable that the pre-trained agent is based on PPO, and only the design of the policy network and prototypes is changed in different baselines.

Below, we provide an overview of the baseline methods:

- Prototype-based policy network*: This variant uses a single transformation network H for all prototypes, rather than individual transformation networks for each prototype dimension. Besides, it learns prototypes as training parameters, which are then mapped to the most recent training example. In contrast, our proposed prototype-based policy network employs manually defined prototypes. The purpose of this variant is also to conduct an ablation test, exploring the impact of these specific design choices, such as separate transformation networks and manually defined prototype-based policy network.
- K-Means: This method obtains the prototypes through the clustering and the mapping process. The clustering process aims to identify centroids that match the number of prototypes used in the proposed prototype-based policy network. When clusters in space *z* with the same number as the prototypes are obtained, each centroid is mapped to the most recent training sample, essentially associating each centroid with a specific state from the training data. These states serve as the prototypical states. Besides, K-Means are allowed to learn the weight parameters of the last layer, which suggests that K-Means clusters are not only used to identify prototypes but also contribute to the network's final decision-making process through weight parameters.

To enhance the pre-hoc interpretability of the method proposed in this paper, we illustrate the four prototypical states and their corresponding intuitive actions in Fig. 4. These prototypical states are designed based on common human intuition, considering the charge and discharge actions of BES and HES, which aid in elucidating the agent's policy. For instance, in Prototype 1, under conditions where the energy market price is high and PV power generation is minimal, the SoC reaches its maximum level, whereas the hydrogen storage reservoir remains empty. Drawing from common human experience, the optimal action for the BES in this scenario is to discharge and sell previously stored electricity at the elevated market price to maximize revenue. Likewise, in a scenario where the

Methods	Metrics	Case 1	Case 2	Case 3	Case 4
Prototype-based Policy Network	Reward MSE	$ \begin{array}{c} \textbf{7.43} \pm \textbf{ 3.25} \\ \textbf{4.92} \ \pm \ \textbf{1.80} \end{array} $	$\begin{array}{c} \textbf{23.63} \pm \textbf{1.01} \\ \textbf{4.97} \pm \textbf{0.51} \end{array}$	$\begin{array}{c} 15.69 \ \pm 0.78 \\ 3.17 \ \pm \ 0.50 \end{array}$	7.48 ± 0.67 0.13 \pm 0.07
Prototype-based Policy Network*	Reward MSE	-7737.81 ± 1627.94 58.26 \pm 6.49	$-6979.15 \pm 4963.04 \\ 45.46 \pm 8.92$	-5054.70 ± 1746.51 138.61 ± 37.65	-5943.58 ± 3174.58 10.60 ± 2.92
K-Means	Reward MSE	-98.36 ± 67.68 24.74 ± 3.99	$\begin{array}{c} 14.72 \pm 3.46 \\ 25.73 \pm 3.12 \end{array}$	$\begin{array}{c} 13.94 \pm 0.72 \\ 23.37 \pm 1.25 \end{array}$	9.38 ± 3.49 8.44 ± 2.55
Reward of Black-box		12.42	24.44	14.70	7.66

 TABLE III

 Results of the Prototype-based Policy Network compared with Various baselines.

Results in bold and cells colored gray denote the best and the second best, respectively.

market price is lower and PV power generation is sufficient, the most advantageous action for the BES with $E_t^{\rm SoC} = 0$ is to charge and store energy. This readies the system to sell electricity when prices increase and PV power generation becomes inadequate. Analogous situations also apply to the EL and FCs within the PV-ESS.



Fig. 4. The output of the prototype-based policy network for four prototypical states.

Leveraging our human-friendly prototypes, we conduct a performance evaluation comparing the proposed prototypebased policy network against the aforementioned baseline methods in the operation optimization of the heterogeneous PV-ESS. We evaluate performance using two key metrics: average reward and mean-squared error (MSE). The reward metric is based on the average of five trials, with each trial comprising 30 simulations. We calculate a cumulative average reward across these trials and subsequently determine the average reward and standard error over the five trials. The second metric, MSE, quantifies the dissimilarity between actions generated by the black-box model and the interpretable model during each iteration. It provides insights into how closely these methods approximate the oracle. The prototypebased policy network serves a dual purpose. First, it aims to align with the pre-trained black-box model to achieve comparable performance, and MSE serves as a tool to assess this alignment. The second purpose is to integrate human-defined prototypes, thus incorporating pre-hoc human experience. It's essential to note that the output of our prototype-based policy network is not expected to precisely replicate that of the pretrained agent. In fact, we intentionally seek some divergence, with the goal of the network learning a new policy grounded in interpretability, informed by reasoning with prototypes and the manually specified weight matrix W.



Fig. 5. The examples of interpretable decision-making by the agent.

The results are presented in Table III with the bestperforming results highlighted in bold, and the second-best results shaded in gray. Analyzing these results, it's evident that the prototype-based policy network we introduced excels in achieving optimal performance in cases 1, 2, and 3. However, in case 4, it attains sub-optimal results in terms of reward. A comparison between the prototype-based policy network and prototype-based policy network* reveals that prototypes designed with the integration of human experience outperform learned prototypes. Having multiple prototypes proves advantageous in guiding agent selection strategies, as each can extract key information relevant to their respective actions.

Furthermore, when compared to the K-Means method with the same number of clusters, human-designed prototypes offer more valuable insights than prototypes generated through self-classification of samples. This enriched knowledge aids in the development of superior strategies. In case 4, which exclusively involves HES, both K-Means and the prototypebased policy network achieve similar average rewards to the black-box model. Notably, the prototype-based policy network yields predictions that closely align with the black-box model, evident in the smallest MSE observed in case 4. In terms of interpretability, we also present examples of decision-making by the agent, as illustrated in Fig. 5, which unveils the correlation between the decisions made by the agent and the comprehensible decisions made by human.

C. Performance comparison between different cases with different learning rate

To further elucidate the influence of heterogeneity in the PV-ESS on the operator's revenue, we extend our analysis beyond the baseline comparisons and consider the performance of the pre-trained black-box model across different cases, as depicted in Fig. 6. Notably, in case 4, where only the HES is involved, the operator's profit is significantly reduced and can even result in a loss. This is primarily attributed to the fact that HES incorporates three types of equipment: the EL, the FCs, and the hydrogen storage reservoir, leading to considerably higher capital cost than other ESS. Additionally, it entails increased degradation and operation/maintenance expenses.

The noticeable increase in reward observed in case 2 compared to case 1 can be explained by the absence of any ESSrelated costs in case 2. When comparing case 3 with case 1, which only accounts for the BES cost, it becomes evident that the presence of the HES significantly reduces operator's profitability.

The occurrence of negative rewards in case 4, while seldom encountered in real-world scenarios, reflects situations where users may be required to pay rental or participation fees to access the energy market with the PV-ESS. In such market, users benefit from lower-priced electricity. Additionally, the PV-ESS operator might be eligible for government incentives aimed at encouraging the use of renewable energy.

TABLE IV Performance comparison of four cases with heterogeneous energy storage devices.

Cases	Reward	Loss	Description
Case 1	151.84	194.28	Both BES and HES are considered and their costs are included.
Case 2	182.88	300.13	Both BES and HES are considered, but their costs are not included.
Case 3	336.60	23.01	Only BES and its cost are considered.
Case 4	-32.17	1186.37	Only HES and its cost are considered.

The learning rate is another critical factor influencing performance. In the above experiments, we employ an adaptive learning rate, initially set at $1e^{-4}$ with an initial attenuation coefficient of $\alpha = 1$. The attenuation coefficient gradually decreases with the number of simulations, following the formula $\alpha = 1 - \frac{\text{step}}{\text{total_{step}}}$. To evaluate the impact of different learning rates on convergence and optimality, we design three alternative learning rate schemes: constant $1e^{-2}$, constant $1e^{-4}$, and a gradually declining learning rate with a constant attenuation coefficient of 0.95 and carry out experiments on case 1. The results are presented in Fig. 7. As illustrated in the figure, we can observe that agents with a constant learning rate tend to exhibit slower convergence rates and are more susceptible to getting stuck in local optimization. In contrast, agents with a gradually declining learning rate demonstrate improved convergence performance. Furthermore,



Fig. 6. Results of four different cases.

the adaptive attenuation coefficient proves more effective in ensuring both convergence and optimal performance compared to a constant attenuation coefficient. These findings underscore the importance of choosing an appropriate learning rate for reinforcement learning tasks.



Fig. 7. Results considering different learning rates.

V. CONCLUSION

In this paper, a heterogeneous PV-ESS is proposed to leverage the unique characteristics of BES and HES for scheduling tasks, with the primary objective of maximizing benefits of the PV-ESS operator through energy arbitrage. To provide more precise guidance for the operator in real-world scenarios, we present a comprehensive cost function that accounts for degradation, capital, as well as operation/maintenance costs. Additionally, in an effort to enhance the interpretability of strategies based on black-box models, we introduce a prototype-based policy network. This network utilizes humandesigned prototypes to guide decision-making by comparing similarities between prototypical situations and encountered situations, leading to natural explanations of scheduling strategies. Comparative results across four distinct cases underscore the effectiveness and practicality of our proposed pre-hoc interpretable optimization method when contrasted with black-box models. Looking ahead to our future work, we plan to extend scheduling tasks to more intricate large-scale ESS featuring multiple uncertainties and heterogeneity. Furthermore, we aim to combine pre-hoc interpretable DRL with these post-hoc interpretable methods to further promote the interpretability of scheduling strategies within energy systems.

REFERENCES

- Z. Chen, X. Yu, W. Xu, and G. Wen, "Modeling and control of islanded DC microgrid clusters with hierarchical eventtriggered consensus algorithm," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 376–386, 2021.
- [2] L. Xiong, Y. Tang, C. Liu, S. Mao, K. Meng, Z. Dong, and F. Qian, "Meta-reinforcement learning-based transferable scheduling strategy for energy management," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 4, pp. 1685–1695, 2023.
- [3] L. Xiong, S. Mao, Y. Tang, K. Meng, Z. Dong, and F. Qian, "Reinforcement learning based integrated energy system management: A survey," *Acta Automatica Sinica*, vol. 47, no. 10, pp. 2321–2340, 2021.
- [4] B. Huang and J. Wang, "Deep-reinforcement-learning-based capacity scheduling for pv-battery storage system," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2272–2283, 2021.
- [5] Y. Levron and D. Shmilovitz, "Optimal power management in fueled systems with finite storage capacity," *IEEE Transactions* on Circuits and Systems I: Regular Papers, vol. 57, no. 8, pp. 2221–2231, 2010.
- [6] D. Krishnamurthy, C. Uckun, Z. Zhou, P. R. Thimmapuram, and A. Botterud, "Energy storage arbitrage under day-ahead and real-time price uncertainty," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 84–93, 2018.
- [7] Y. Levron, D. Shmilovitz, and L. Martínez-Salamero, "A power management strategy for minimization of energy storage reservoirs in wireless systems with energy harvesting," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 3, pp. 633–643, 2011.
- [8] X. Xiong, C. K. Tse, and X. Ruan, "Bifurcation analysis of standalone photovoltaic-battery hybrid power system," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 5, pp. 1354–1365, 2013.
- [9] B. A. Abdelmagid, M. H. K. Hmada, and A. N. Mohieldin, "An adaptive fully integrated dual-output energy harvesting system with mppt and storage capability," *IEEE Transactions* on Circuits and Systems I: Regular Papers, vol. 70, no. 2, pp. 593–606, 2023.
- [10] K.-B. Kwon and H. Zhu, "Reinforcement learning-based optimal battery control under cycle-based degradation cost," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4909–4917, 2022.
- [11] A. Dolatabadi, H. Abdeltawab, and Y. A.-R. I. Mohamed, "Deep reinforcement learning-based self-scheduling strategy for a caes-pv system using accurate sky images-based forecasting," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1608– 1618, 2023.
- [12] B. Cleary, A. Duffy, A. OConnor, M. Conlon, and V. Fthenakis, "Assessing the economic benefits of compressed air energy storage for mitigating wind curtailment," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 3, pp. 1021–1028, 2015.
- [13] M. Shi, H. Wang, P. Xie, C. Lyu, L. Jian, and Y. Jia, "Distributed energy scheduling for integrated energy system clusters with peer-to-peer energy transaction," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 142–156, 2023.

- [14] X. Sun, X. Cao, B. Zeng, Q. Zhai, and X. Guan, "Multistage dynamic planning of integrated hydrogen-electrical microgrids under multiscale uncertainties," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3482–3498, 2023.
- [15] N. Padmanabhan, M. Ahmed, and K. Bhattacharya, "Battery energy storage systems in energy and reserve markets," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 215–226, 2020.
- [16] C. Liu, H. Ma, H. Zhang, X. Shi, and F. Shi, "A MILP-based battery degradation model for economic scheduling of power system," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 2, pp. 1000–1009, 2023.
- [17] Y. Li, Y. Gu, G. He, and Q. Chen, "Optimal dispatch of battery energy storage in distribution network considering electrothermal-aging coupling," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3744–3758, 2023.
- [18] Y. Yang, D. Xu, T. Ma, and X. Su, "Adaptive cooperative terminal sliding mode control for distributed energy storage systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 434–443, 2021.
- [19] F. Garcia-Torres, C. Bordons, J. Tobajas, R. Real-Calvo, I. Santiago, and S. Grieu, "Stochastic optimization of microgrids with hybrid energy storage systems for grid flexibility services considering energy forecast uncertainties," *IEEE Transactions* on Power Systems, vol. 36, no. 6, pp. 5537–5547, 2021.
- [20] Y. Ma, Z. Hu, and Y. Song, "Hour-ahead optimization strategy for shared energy storage of renewable energy power stations to provide frequency regulation service," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 4, pp. 2331–2342, 2022.
- [21] B. Zhou, J. Fang, X. Ai, S. Cui, W. Yao, Z. Chen, and J. Wen, "Storage right-based hybrid discrete-time and continuous-time flexibility trading between energy storage station and renewable power plants," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 1, pp. 465–481, 2023.
- [22] P. Aaslid, M. Korpås, M. M. Belsnes, and O. B. Fosso, "Stochastic optimization of microgrid operation with renewable generation and energy storages," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 3, pp. 1481–1491, 2022.
- [23] B. Zhang, C. Dou, D. Yue, J. H. Park, Y. Xue, Z. Zhang, Y. Zhang, and X. Ding, "Event-triggered hierarchical multimode management strategy for source-load-storage in microgrids," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 5, pp. 2201–2214, 2023.
- [24] C. Zhang, Z. Dong, and L. Yang, "A feasibility pump based solution algorithm for two-stage robust optimization with integer recourses of energy storage systems," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1834–1837, 2021.
- [25] Z. Rostamnezhad, N. Mary, L.-A. Dessaint, and D. Monfet, "Electricity consumption optimization using thermal and battery energy storage systems in buildings," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 251–265, 2023.
- [26] F. Li, B. Sun, and C. Zhang, "Operation optimization for integrated energy system with energy storage," *Science China Information Sciences*, vol. 61, pp. 1–3, 2018.
- [27] L. Xiong, Y. Tang, S. Mao, H. Liu, K. Meng, Z. Dong, and F. Qian, "A two-level energy management strategy for multimicrogrid systems with interval prediction and reinforcement learning," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 4, pp. 1788–1799, 2022.
- [28] Z. Liu, H. Gao, X. Yu, W. Lin, J. Qiu, J. J. Rodríguez-Andina, and D. Qu, "B-spline wavelet neural-network-based adaptive control for linear-motor-driven systems via a novel gradient descent algorithm," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 2, pp. 1896–1905, 2024.
- [29] L. Xiong, Y. Tang, C. Liu, S. Mao, K. Meng, Z. Dong, and F. Qian, "A home energy management approach using decoupling value and policy in reinforcement learning," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, pp. 1261–1272, 2023.

- [30] Y. Tang, H. Gao, W. Zou, and J. Kurths, "Distributed synchronization in networks of agent systems with nonlinearities and random switchings," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 358–370, 2013.
- [31] J. Wang, Y. Hong, J. Wang, J. Xu, Y. Tang, Q. L. Han, and J. Kurths, "Cooperative and competitive multi-agent systems: From optimization to games," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 5, pp. 763–783, 2022.
- [32] X. Li, H. Liu, C. Li, G. Chen, C. Zhang, and Z. Y. Dong, "Deep reinforcement learning based explainable pricing policy for virtual storage rental service," *IEEE Transactions on Smart Grid*, pp. 1–1, 2023.
- [33] C. Li, Z. Dong, L. Ding, H. Petersen, Z. Qiu, G. Chen, and D. Prasad, "Interpretable memristive lstm network design for probabilistic residential load forecasting," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2297– 2310, 2022.
- [34] E. M. Kenny, M. Tucker, and J. Shah, "Towards interpretable deep reinforcement learning with human-friendly prototypes," in *The Eleventh International Conference on Learning Repre*sentations, 2022.
- [35] G. Cau, D. Cocco, M. Petrollese, S. K. Kær, and C. Milan, "Energy management strategy based on short-term generation scheduling for a renewable microgrid using a hydrogen storage system," *Energy Conversion and Management*, vol. 87, pp. 820– 831, 2014.
- [36] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- [37] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.
- [38] A. Bontempelli, S. Teso, K. Tentori, F. Giunchiglia, and A. Passerini, "Concept-level debugging of part-prototype networks," arXiv preprint arXiv:2205.15769, 2022.
- [39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [40] Microgrid/distribution network level energy market managed by an RL agent. [Online]. Available: https://github.com/ utkarshapets/microgrid-RL