An Imitation Learning Method with Multi-Virtual Agents for Microgrid Energy Optimization under Interrupted Periods

Yanbin Lin, Zhen Ni, and Yufei Tang
Electrical Engineering and Computer Science
Florida Atlantic University, Boca Raton, FL, USA, 33431
{liny2020, zhenni, ytang}@fau.edu

2024 IEEE Power & Energy Society General Meeting (PESGM) | 979-8-3503-8183-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/PESGM51994.2024.10689197

***Abstract*—Existing computer analytic methods for the microgrid system, such as reinforcement learning (RL) methods, suffer from a long-term problem with the empirical assumption of the reward function. To alleviate this limitation, we propose a multi-virtual-agent imitation learning (MAIL) approach to learn the dispatch policy under different power supply interrupted periods. Specifically, we utilize the idea of generative adversarial imitation learning method to do direct policy mapping, instead of learning from manually designed reward functions. Multi-virtual agents are used for exploring the relationship of uncertainties and corresponding actions in different microgrid environments in parallel. With the help of a deep neural network, the proposed MAIL approach can enhance robust ability by minimizing the maximum crossover discriminators to cover more interrupted cases. Case studies show that the proposed MAIL approach can learn the dispatch policies as well as the expert method and outperform other existing RL methods.**

***Index Terms*—Imitation learning, interrupted power supply, deep neural networks, machine learning, multi-virtual agents, and microgrid energy scheduling.**

I. Introduction

The urgency of improving the microgrid energy scheduling has been highlighted in recent years since the threats of extreme weather incidents and natural disasters to the microgrid systems increased [1]. To enhance the robust ability of the microgrid systems, it is critical to optimize microgrid energy scheduling to cope with the uncertainties with lower costs.

Extensive model-based microgrid scheduling approaches have been proposed in the literature [2]–[4]. However, these methods generally rely on an accurate forecast of the uncertainties and a specific physical model of the microgrid system [5]. To reduce these dependencies of model-based microgrid scheduling approaches, machine learning (ML)-based methods have been proposed in recent years, which are referred to as learning-based methods [6].

Among these learning-based methods, single-agent and multi-agent reinforcement learning (RL) approaches have been widely applied in solving microgrid online decision-making and control problems. There are some single-agent deep reinforcement learning (DRL) methods for the microgrid providing open-source codes online. A novel microgrid model was proposed in [7] to coordinate among the different flexible sources using seven DRL methods. An adaptive emergency control scheme was designed to leverage the high dimensional feature extraction and non-linear generalization capabilities of DRL methods for complex power systems [8]. Nevertheless, these single-agent RL methods learned and made energy scheduling decisions according to individual information and environment, instead of considering overall profits and supply-demand balance of microgrids in different situations [9]. Multi-agent RL with a cooperative system is considered as a potential approach to enhance the performance of the microgrid system [10]. In [11], the deep deterministic policy gradients (DDPG) and multi-agent DDPG algorithms were investigated to conclude that multi-agent approaches can produce greater profits. Instead of increasing the number of states [12], asynchronous and synchronous RL approaches can save computational costs and handle complex tasks. An aggregating Q-learning approach was proposed in [13], utilizing multiple local Q-learning agents and aggregating the learned value functions to the global agent to learn different microgrid events. However, these RL methods rely deeply on the empirical assumption of the reward function for the microgrid system.

Generally, the inverse reinforcement learning (IRL) method and imitation learning (IL) method have the advantage of avoiding the empirical assumption of the reward function. A reinforcement and imitation learning approach was proposed to develop the pricing strategy of electricity retailers with customers' flexibility [14]. In [15], a data-driven approach based on imitation learning was presented to mimic the mixed-integer linear programming solver to optimize the operations of the microgrid. To avoid the computation of the state visitation frequency of the microgrid system and recover the reward function, our previous work [16] proposed a modified maximum entropy IRL method for solving the microgrid energy scheduling problem. However, it's not suitable for the microgrid under uncertainties. To our knowledge, there has not been any research focused on the microgrid energy scheduling problem using the multi-agent imitation learning method to improve the robust performance under the uncertainties.

To address the challenges mentioned above, we work on the optimization of the microgrid energy scheduling problem under interrupted periods. Specifically, we propose a multi-virtual-agent imitation learning (MAIL) approach to learn the dispatch policy under different power supply interrupted peri-

979-8-3503-1637-7/24/$31.00 ©2024 IEEE

ods. The contributions of this paper are provided as follows. First, different from cooperative multi-agent RL methods, we utilize the idea of the generative adversarial imitation learning (GAIL) method [17] combined with multi-virtual agents in parallel to bypass the assumptive reward function. Moreover, to improve the robustness, we optimize the upper bound of the objective function using N maximum crossover discriminators, which can cover more interrupted microgrid cases. The performance of the proposed MAIL method is slightly better than the corresponding expert policy and outperforms other existing microgrid energy scheduling methods.

The remainder of this paper is organized as follows. Section II states the main idea and framework of our proposed multi-virtual-agent imitation learning method. Explanations about the microgrid system model are given in Section III. Specific case studies of power supply interruptions happening during extreme weather events and results are shown in Section IV. Finally, conclusions are provided in Section V.

II. Multi-virtual-agent Imitation Learning Approach

A. Markov Decision Process for Microgrid Systems

A Markov decision process (MDP) [18] can be expressed by a tuple: $\{S, A, \mathcal{P}(s|s', a'), R, \gamma, s_0\}$, where S is a set of states s, A is a set of possible actions a, $\mathcal{P}(s|s', a')$ is the transition probability, s' ans a' is the next state and the next action, R is the reward function, γ is the discounted rate, and s_0 is the initial state where the agent will depart from. The occupancy measure $\rho_\pi(s,a) = \pi(a|s)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s|\pi, \mathcal{P})$ is the state-action distribution induced by policy π. The state distribution is $\mu_\pi(s) = \sum_{t=0}^{\infty}\gamma^t Pr(s_t = s|\pi, \mathcal{P})$, and the initial state distribution is denoted as $\mu_0(s)$. In the microgrid system, the cost function is the negative value of the reward function.

We consider a collection of MDPs for the microgrid system, where the state and action spaces are the same while the cost function is unknown to the agent. The microgrid system can be varied among these MDPs due to different interrupted periods. We sample N microgrid systems with different power supply interrupted periods and denote N environments with $E_1, E_2, \cdots, E_N$. Each expert policy π_E^j will generate corresponding expert demonstrations set $\mathcal{D}_E^j$, where $j \in (1, 2, \cdots, N)$. ξ_E^j stands for the expert demonstration belonging to $\mathcal{D}_E^j$, and A_E^j is the expert actions set. In our assumptions, the virtual agents can interact with all sampled environments. The learner policy of the imitation learning network for each sampled environment is π^i, and the corresponding imitated actions set is A^i, where $i \in (1, 2, \cdots, N)$. We assume that we can only acquire expert demonstrations set $\mathcal{D}_E^j$, while the expert policy π_E^j is unknown.

B. Proposed MAIL Method and Framework

To solve the microgrid energy optimization problem under different interrupted periods caused by uncertain extreme weather events, we consider the occupancy matching technique used in the GAIL method [17].

In a single environment, the occupancy measure $\rho_\pi(s,a)$ satisfies the Bellman flow constraint as

$$\rho_\pi(s,a) = \mu_0(s)\pi(a|s) + \gamma\int_{(s',a')}\mathcal{P}(s|s',a')\rho_\pi(s',a')\pi(a|s) \tag{1}$$

The GAIL method reproduces the expert's policy according to the objective function,

$$\begin{aligned}&\min_\pi D_{JS}(\bar{\rho}_\pi, \bar{\rho}_{\pi_E})\\ &= \min_\pi \max_D \mathbb{E}_{\rho_{\pi_E}}[logD(s,a)] + \mathbb{E}_{\rho_\pi}[log(1 - D(s,a))]\end{aligned} \tag{2}$$

where $\bar{\rho}_\pi$ is the normalized learner occupancy measure, $\bar{\rho}_{\pi_E}$ is the normalized expert occupancy measure, D_{JS} is the Jensen-Shannon (JS) divergence [19], and D is the discriminator used to distinguish whether a given pair (s,a) is from the expert or not.

In the multiple environments, the occupancy measure $\rho_\pi(s,a)$ becomes the mixture with N environments,

$$\begin{aligned}\rho_\pi(s,a) = \frac{1}{N}\sum_{i=1}^{N}\rho_\pi^i(s,a) = \mu_0(s)\pi(a|s)\\ + \gamma\frac{1}{N}\sum_{i=1}^{N}\int_{(s',a')}\mathcal{P}^i(s|s',a')\rho_\pi^i(s',a')\pi(a|s)\end{aligned} \tag{3}$$

where $\mathcal{P}^i$ is the transition probabilities of the i^{th} environment, and ρ_π^i is the i^{th} environment's occupancy measure.

Therefore, considering the multiple environments with uncertainties [20], the objective function can be changed as

$$\min_\pi \mathbb{E}_{s\sim\frac{1}{N}\sum_{i=1}^{N}\mu_\pi^i}[\sum_{j=1}^{N}\lambda_j(s)D(\pi(\cdot|s), \pi_E^j(\cdot|s))] \tag{4}$$

where $\lambda_j(s)$ is the weight to determine how much $\pi_E^j(\cdot|s)$ is imitated, and $\sum_{j=1}^{N}\lambda_j(s) = 1$.

In order to handle $\lambda_j(s)$, we can replace $\sum_{j=1}^{N}\lambda_j(s)$ with $\max_j$ to yield the upper bound as (5).

$$\min_\pi \mathbb{E}_{s\sim\frac{1}{N}\sum_{i=1}^{N}\mu_\pi^i}\max_j D(\pi^i(\cdot|s), \pi_E^j(\cdot|s)) \tag{5}$$

which can be regarded as the mapping between the expert action set A_E^j and the imitated actions set A^i as (6).

$$\min_\pi \frac{1}{N}\sum_{i=1}^{N}\max_j D_{ij}(A^i, A_E^j) \tag{6}$$

We denote (7) as the loss function of the imitation learning network. The discriminator D_{ij} uses the mean squared error (MSE) to calculate the discrimination of A^i and A_E^j.

$$\mathcal{L}(w) = \frac{1}{N}\sum_{i=1}^{N}\max_j D_{ij}(A^i, A_E^j) \tag{7}$$

Fig.1 shows the framework of the implementation of our proposed multi-virtual-agent imitation learning method in the microgrid optimization problem.

We use N virtual environments and N corresponding expert policies π_E^j to generate N expert demonstrations sets $\mathcal{D}_E^j$

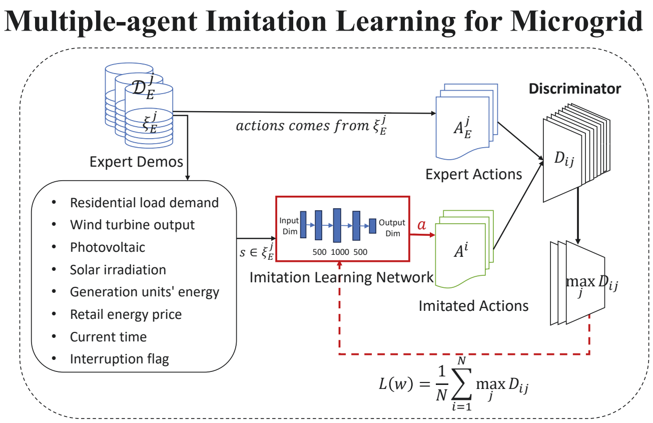

Fig. 1. The framework of our proposed multi-virtual-agent imitation learning approach for the microgrid system. Several virtual environments' expert demonstrations are used to imitate parallel, and the maximum crossover discriminators are calculated as the loss function.

and N expert actions sets A_E^j, where $j \in (1, 2, \cdots, N)$. Meanwhile, we initial the imitation learning network with random weights and use 8-dimensional expert states as the input data to imitate the actions in A_E^j as the imitated actions set A^i, where $i \in (1, 2, \cdots, N)$. Then we use N^2 crossover discriminators to compare the difference of A^i and A_E^j and select N maximum discriminators to calculate the loss function $\mathcal{L}(w)$ as (7). $\mathcal{L}(w)$ is used to update the weights of the imitation learning network by calculating the gradient. The updated imitation learning network will output new imitated actions to do new comparations. After the data aggregation and imitation learning network update, we will get a final imitation learning network that can generate robust actions against the uncertainties of the environment.

III. Model Description and Problem Formulation

In this paper, we consider a grid-connected microgrid consisting of four units from the perspective of energy generation and load demand in [13], including the battery energy storage system (BESS), the distributed generations, the main grid, and the residential load. The connection with the main grid makes the microgrid flexibly export/import power to/from the utility network.

In this problem, we aim to schedule the generation units efficiently with the lowest cost over a time period of T (24 hours), based on the defined probabilistic power supply interruptions. The generation units' decisions/actions include the power output of diesel generators (DG) $p_{t,d}$, the power $p_{t,p}$ purchased from (positive) or sold to (negative) the main grid, and the charging or discharging power of BESS $p_{t,b}$, which is positive when discharging.

The state at the hour t is defined as:

$$s_t = (s_{t,b}, s_{t,d}, s_{t,g}, s_{t,p}, s_{t,l}, s_{t,pr}) \tag{8}$$

where $s_{t,b}$ is the state of charge (SOC) of the BESS, $s_{t,d}$ is the binary variable that indicates the ON/OFF status of DG, $s_{t,g}$ is the output of renewable generations (RG), $s_{t,p}$ is the retail energy price, $s_{t,l}$ is the residential load demand, and $s_{t,pr}$ is the probability of the power supply interruption happening.

As mentioned in the Sec.II-A and (6), A^i and A_E^j stand for the imitated actions set and expert actions set for the corresponding environment, respectively. They are composed of several individual actions a_t, which is defined as:

$$a_t = (p_{t,b}, p_{t,d}, p_{t,p}) \tag{9}$$

The transition function of SOC is given as

$$s_{t+1,b} = s_{t,b} - \frac{\Delta s_{t,b}}{E_b} \tag{10}$$

where E_b is the rated energy capacity, and $\Delta s_{t,b}$ is the energy changing amount of BESS at hour t,

$$\Delta s_{t,b} = \begin{cases} \eta_b^- p_{t,b} \Delta t, & \text{if } p_{t,b} \leq 0 \\ \dfrac{p_{t,b}}{\eta_b^+} \Delta t, & \text{otherwise} \end{cases} \tag{11}$$

where Δt is the time step size, η_b^- and η_b^+ stand for the charging and discharging efficiencies respectively.

The SOC of the battery is constrained by $s_{b-} \leq s_{t,b} \leq s_{b+}$, which are the lower and upper bounds of SOC, respectively. The value of $p_{t,b}$ needs to be in a scope of maximum discharging power P_b^+ and maximum charging power P_b^-.

The power balance of the microgrid can be expressed as

$$p_{t,b} + p_{t,d} + p_{t,p} + s_{t,g} = s_{t,l} \tag{12}$$

The objective function is to $\min_{a_t} \sum_{t=1}^{T} C(s_t, a_t)$. The cost function $C(s_t, a_t)$ is defined as

$$C(s_t, a_t) = p_{t,p} s_{t,p} + s_{t,d}(a_d p_{t,d}^2 + b_d p_{t,d} + c_d) \tag{13}$$

where a_d, b_d, and c_d are the coefficients of the quadratic function. Note, that the cost function is only used for performance comparisons in our proposed MAIL method.

IV. Simulation Results and Analysis

To investigate the impact of power supply interruptions on the microgrid operations, we assume during extreme weather-related events, the utility grid goes down for a certain period, and the microgrid operates in an isolated mode during this period [13]. The RG outputs are taken from the system advisory model by the National Renewable Energy Laboratory for the city of Phoenix, AZ [21]. For the load-demand, a small residential community load-demand data is collected from [22]. All the comparison experiments are based on the same assumptions for other existing approaches, including dynamic programming (DP) method, aggregated Q-learning method [13], and cooperative Q-learning method [23].

For the simulations, we consider three virtual environments that the power supply interruptions happen at the time-period 10^{th}-13^{th} hours, 11^{th}-14^{th} hours, and 12^{th}-15^{th} hours. Three local virtual agents are applied to imitate the corresponding virtual environment's actions. We use the DP method and the aggregated Q-learning method as the expert policies to validate our proposed method. Note that the DP approach is an offline optimization technique. Besides, we conduct two case studies with low probability (30%) and high probability (70%) rates of weather-related events and test the performance on the microgrid systems for all methods at a new interrupted period 10^{th}-15^{th} hours.

A. *Comparative Performance with Different Methods*

To validate our proposed MAIL method, we first consider the power supply interruptions happening with the probability of 100%. Our proposed imitation learning approach can output the microgrid scheduling decisions online, given the current state of the microgrid system.

The imitation learning network we apply has three hidden layers with the sizes of $(500, 1000, 500)$. The input states of the network are shown in Fig.1, including

- Residential load demand (kWh)
- Wind turbine output (kWh)
- Photovoltaic (kWh)
- Solar irradiation (kWh/m^2)
- Generation units' energy (kWh)
- Retail energy price ($/kWh)
- Current time (hour)
- Interruption flag

To apply the DP and the aggregated Q-learning expert methods, discretized battery SOC and DG ON/OFF status must be used to define the states. On one hand, we analyze the influence of the SOC state sizes from 7 to 51. On the other hand, we test the proposed method and other existing approaches at a new interrupted period 10^{th}-15^{th} hours, different from three training time slots 10^{th}-13^{th} hours, 11^{th}-14^{th} hours, and 12^{th}-15^{th} hours. The performance comparison is presented in Tab.I.

TABLE I
TOTAL OPERATIONAL COST ($) USING DIFFERENT METHODS AND DISCRETIZED-STATE SIZES WHEN INTERRUPTIONS HAPPEN AT THE TIME-PERIOD 10^{th}-15^{th} HOURS.

SOC Sizes	DP	MAIL(DP)	Aggregated Q	MAIL(AQ)	Cooperative Q
7	19.7	**19.4**	21.3	21.5	19.9
11	19.3	**18.8**	21.2	20.7	20.4
21	18.8	**18.7**	20.7	20.1	22.1
31	18.8	**18.7**	20.5	20.1	21.5
41	18.7	**18.6**	20.5	20.1	21.7
51	18.7	**18.6**	20.4	20.0	22.5

In Tab.I, MAIL (DP) means the proposed MAIL method using DP expert demonstrations, and MAIL (AQ) means the proposed MAIL method using aggregated Q-learning expert demonstrations. From the results, we can see that our proposed imitation learning method can achieve a lower operational cost, slightly better than the corresponding expert methods without the need for reward functions. This is because our proposed MAIL method decides to charge extra energy and sell it when the energy price is higher. The performances of the MAIL method and the corresponding expert method get closer when the discretized-state size increases. While the cooperative Q-learning method suffers from overfitting when the discretized-state size increases. Note that to compare the performance fairly, we discretize our imitated results to corresponding discretized-state size. For practical applications, our proposed method has the natural advantage of providing continuous power decisions. Besides, our MAIL method using DP demonstrations can achieve similar performance using a smaller expert SOC discretized state size, which saves the computational cost of collecting expert demonstrations. For example, when the SOC state size $= 11$, our proposed method has a total cost of $18.8 almost the same as $18.6 with the discretized size $= 51$. Overall, the proposed MAIL algorithm can achieve performance comparable to expert policies, no matter whether using expert demonstrations from the DP method or the aggregated Q-learning method. It also outperforms other existing RL methods with the lowest total operational cost.

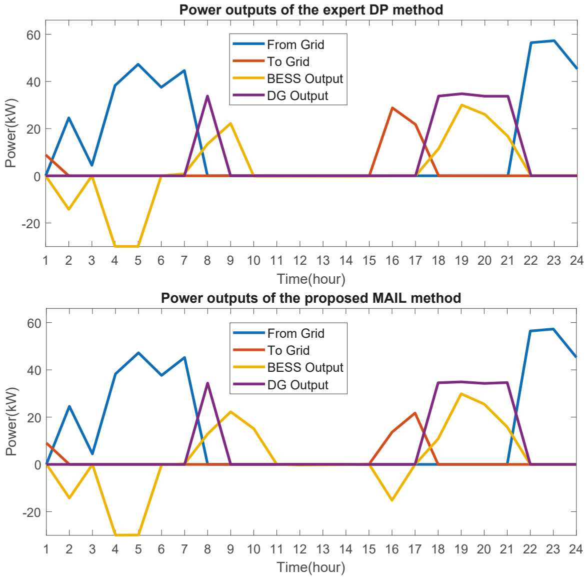

Fig. 2. The power outputs for DP method and proposed MAIL method using DP expert demonstrations with a discretized-state size=1000. The upper figure is the decisions of the DP method (expert), and the lower figure is the decisions of the proposed MAIL method.

We plot the power outputs of the microgrid resources obtained from the DP method and the MAIL approach using DP expert demonstrations shown in Fig.2. The discretized-state size used in this figure is 1000. We can see that the decisions of the generation units for these two methods are almost the same. There is a little difference where the BESS output of the MAIL policy is negative at time = 16^{th} hour. This is because the net load at this time is negative, so it is better to charge the power of BESS and sell extra energy to the grid. We also compare with the imitation learning method using all the expert demonstrations of the three virtual training environments. We find that the MAIL method performs slightly better than the IL method using all expert demonstrations aggregated.

B. *Two Case Studies with Different Interrupted Probabilities*

We consider two case studies with a low (30%) probability and a high (70%) probability of interruptions and apply the

MAIL method using DP expert demonstrations with the SOC size $= 11$. The comparison results are shown in Tab.II.

TABLE II
TOTAL OPERATIONAL COST (\$) WITH A LOW/HIGH INTERRUPTED PROBABILITY AND A DISCRETIZED-STATE SIZE=11 AT THE TIME-PERIOD 10^{th}-15^{th} HOURS.

Probability	MAIL	DP	Aggregated Q	Cooperative Q
30%	**18.7**	18.8	20.2	22.3
70%	**18.7**	19.1	20.7	22.0

From the results, we can conclude that our proposed MAIL method achieves performance with fewer variations when the interrupted probability is changed. Its cost \$18.7 is the lowest one, no matter with a low probability or a high probability.

V. Conclusion and Discussion

In this paper, an imitation learning approach with multi-virtual agents called the MAIL method, is proposed to learn the microgrid system's decisions under different interrupted periods. The main contribution of our work is the new solution for microgrid energy optimization using imitation learning with multi-virtual agents. By direct mapping of the policy with a deep neural network, our approach can learn the dispatch policy for a new interrupted period, avoiding the assumptive reward function and the direct expert policy of the new environment. Besides, the framework with multiple virtual agents improves the robust ability through the utilization of maximum discriminators to cover more interrupted cases. The case studies with different interrupted probabilities are conducted to validate the effectiveness of the proposed MAIL approach. Our experiments show that the MAIL algorithm can match the performance of expert policies and outperform other existing methods without the need for reward functions.

The proposed MAIL method presents a potential solution for microgrid applications based on MDP formulations while incorporating resiliency factors. For example, it's suitable for the microgrid stochastic optimization problem in [24] to accommodate the uncertain extreme events systematically. The proposed MAIL method can employ local learning agents to interact with pertinent microgrid events in a distributed way and minimize the maximum crossover discrimination. We will apply our proposed method for this application in the future.

Acknowledgment

This work is partially supported by the National Science Foundation under grants # 1949921, # 2047064, and # 2320972. We also acknowledge Dr. Avijit Das for guiding the microgrid model and results discussion.

References

[1] Y. Lin, D. Duan, X. Hong, X. Cheng, L. Yang, and S. Cui, "Very-short-term solar forecasting with long short-term memory (lstm) network," in *2020 Asia Energy and Electrical Engineering Symposium (AEEES)*, 2020, pp. 963–967.

[2] L. Luo, S. S. Abdulkareem, A. Rezvani, M. R. Miveh, S. Samad, N. Aljojo, and M. Pazhoohesh, "Optimal scheduling of a renewable based microgrid considering photovoltaic system and battery energy storage under uncertainty," *Journal of Energy Storage*, vol. 28, p. 101306, 2020.

[3] A. Moradmand, M. Dorostian, and B. Shafai, "Energy scheduling for residential distributed energy resources with uncertainties using model-based predictive control," *International Journal of Electrical Power & Energy Systems*, vol. 132, p. 107074, 2021.

[4] F. Garcia-Torres, C. Bordons, J. Tobajas, J. J. Márquez, J. Garrido-Zafra, and A. Moreno-Muñoz, "Optimal schedule for networked microgrids under deregulated power market environment using model predictive control," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 182–191, 2020.

[5] Y. Ji, J. Wang, J. Xu, and D. Li, "Data-driven online energy scheduling of a microgrid based on deep reinforcement learning," *Energies*, vol. 14, no. 8, p. 2120, 2021.

[6] W. Dong, Q. Yang, W. Li, and A. Y. Zomaya, "Machine-learning-based real-time economic dispatch in islanding microgrids in a cloud-edge computing environment," *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13 703–13 711, 2021.

[7] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, p. 100413, 2021.

[8] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive power system emergency control using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1171–1182, 2020.

[9] X. Fang, J. Wang, G. Song, Y. Han, Q. Zhao, and Z. Cao, "Multi-agent reinforcement learning approach for residential microgrid energy scheduling," *Energies*, vol. 13, no. 1, p. 123, 2019.

[10] D. Chen, K. Chen, Z. Li, T. Chu, R. Yao, F. Qiu, and K. Lin, "Powernet: Multi-agent deep reinforcement learning for scalable powergrid control," *IEEE Transactions on Power Systems*, vol. 11, pp. 1007–1017, 2021.

[11] D. J. Harrold, J. Cao, and Z. Fan, "Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning," *Applied Energy*, vol. 318, p. 119151, 2022.

[12] J. Di, S. Chen, P. Li, X. Wang, H. Ji, and Y. Kang, "A cooperative-competitive strategy for autonomous multidrone racing," *IEEE Transactions on Industrial Electronics*, 2023.

[13] A. Das, Z. Ni, and X. Zhong, "Aggregating learning agents for microgrid energy scheduling during extreme weather events," in *2021 IEEE Power & Energy Society General Meeting (PESGM)*, 2021, pp. 01–05.

[14] Y. Zhang, Q. Yang, D. Li, and D. An, "A reinforcement and imitation learning method for pricing strategy of electricity retailer with customers' flexibility," *Applied Energy*, vol. 323, p. 119543, 2022.

[15] S. Gao, C. Xiang, M. Yu, K. T. Tan, and T. H. Lee, "Online optimal power scheduling of a microgrid via imitation learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 861–876, 2021.

[16] Y. Lin, A. Das, and Z. Ni, "A modified maximum entropy inverse reinforcement learning approach for microgrid energy scheduling," in *2023 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2023, pp. 1–5.

[17] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[18] M. L. Puterman, "Markov decision processes," *Handbooks in operations research and management science*, vol. 2, pp. 331–434, 1990.

[19] B. Fuglede and F. Topsoe, "Jensen-shannon divergence and hilbert space embedding," in *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, 2004.

[20] J. Chae, S. Han, W. Jung, M. Cho, S. Choi, and Y. Sung, "Robust imitation learning against variations in environment dynamics," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2828–2852.

[21] J. Freeman, N. Blair, D. Guittet, M. Boyd, B. Mirletz *et al.*, "System Advisor Model," Available: https://sam.nrel.gov/.

[22] OpenEI, "OpenEI Data," Available: https://openei.org/wiki/Data.

[23] F. Luo, Y. Chen, Z. Xu, G. Liang, Y. Zheng, and J. Qiu, "Multiagent-based cooperative control framework for microgrids' energy imbalance," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1046–1056, 2017.

[24] A. Das, Z. Ni, and X. Zhong, "Microgrid energy scheduling under uncertain extreme weather: Adaptation from parallelized reinforcement learning agents," *International Journal of Electrical Power & Energy Systems*, vol. 152, p. 109210, 2023.