A Modified Maximum Entropy Inverse Reinforcement Learning Approach for Microgrid Energy Scheduling

Yanbin Lin*, Avijit Das†, and Zhen Ni*
*Florida Atlantic University, Boca Raton, FL, USA, 33431
{liny2020, zhenni}@fau.edu
†Pacific Northwest National Laboratory, Richland, WA, USA, 99352
avijit.das@pnnl.gov

2023 IEEE Power & Energy Society General Meeting (PESGM) | 978-1-6654-6441-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/PESGM52003.2023.10252933

***Abstract*—Increasing popularity of integrating distributed energy resources (DERs) into the power system brings a challenge to optimize the dispatch policy for microgrid energy scheduling. The reinforcement learning methods suffer from a long-time problem with the empirical assumption of the reward function for the microgrid system. Although the traditional inverse reinforcement learning (IRL) approaches can solve this problem to some extent, they encounter a limitation of extensive computations for state visitation frequency in the large and continuous state space. To alleviate this limitation, we propose a modified maximum entropy IRL (MMIRL) method to extract the reward function from the expert demonstrations for solving the microgrid energy scheduling problem. The computation of state visitation frequency is avoided by calculating the difference between the expert feature expectation and learner feature expectation. The microgrid optimization is suitable for using state-action (s, a) feature than state s feature only to recover the reward, and this setting drives the need for a computationally efficient method. To this end, the proposed MMIRL algorithm is designed to recover the reward function and learn the dispatch policy compared to the conventional approaches for microgrid energy scheduling. Case studies are performed in an energy arbitrage problem and a microgrid system with DERs, respectively. Results substantiate that the proposed MMIRL approach can learn the dispatch policy with more than 99% accuracy and outperforms other comparative methods in both cases.**

***Index Terms*—Distributed energy resources, reinforcement learning, maximum entropy inverse reinforcement learning, microgrid energy scheduling, and operation optimization.**

I. Introduction

The high variability of renewable energy sources (RESs) and different power supply units make the islanded microgrids challenging to optimize to maintain a minimum operational daily cost. It is often desired to control and coordinate the microgrid energy control center in an efficient and economical way [1]. Therefore, there is an increasingly attention to find a proper optimization approach in microgrid energy scheduling.

In recent years, both model-based optimization methods and reinforcement learning (RL) methods have been investigated by researchers to solve microgrid energy scheduling problems. There are extensive model-based online scheduling approaches having been proposed in the literature [2]–[4]. Although these model-based methods have been successfully applied in the aforementioned studies, they mainly depended on assumptive physical models of the microgrid system with accurate forecasts of uncertainties. This increases the operational difficulties of model-based microgrid energy scheduling methods in reality. The RL-based methods are learning-based methods aimed at maximizing the agent's total reward, including traditional RL methods and data-drive deep RL methods. Several adaptive energy storage charging and discharging strategy optimization models using traditional RL algorithms [5] [6] were proposed with the help of a modified deep learning approach to predict the photovoltaic power and the load demand [7]. However, these traditional RL methods were limited to the application of large and continuous state space due to the curse of dimensionality [8]. Hence, deep reinforcement learning (DRL) methods have been introduced to the microgrid energy scheduling to solve the problem of high dimensional state space in recent years [9]–[11]. Nevertheless, deep RL-based methods have the limitation that needs a lot of training data and samples. Besides, these current RL methods still suffer from a long-standing problem with the empirical assumption of the reward function for the microgrid system, while the actual reward function is usually unknown [12].

Fortunately, the inverse reinforcement learning (IRL) method has an advantage of extracting an agent's reward function from the expert demonstrations [13]. The individual preferences of a microgrid energy scheduling agent are hidden in its actual behaviors and can be extracted by the expert demonstrations. There is a study that optimized the economic operation of a microgrid with a variety of distributed energy resources (DERs) via the imitation learning method [8]. Although the imitation learning method can mimick the expert behavior in given tasks by learning a mapping between observations and actions, a direct reward function is not learned. In [12] and [14], the authors introduced a deep inverse reinforcement learning method to identify the individual reward functions of the bidding market and coupled multiple market through the historical bidding behaviors. However, to our knowledge, there have not been any research focused on restoring the reward function of the microgrid energy scheduling problem using the IRL method and avoiding

computation of the state visitation frequency at the same time.

In this paper, we work on the optimization of the microgrid energy scheduling problem using the inverse reinforcement learning method. Specifically, we propose a computationally efficient inverse reinforcement learning approach to extract the reward function from the expert demonstrations, called modified maximum entropy IRL (MMIRL) method. The contributions of this paper are provided as follows. First, a novel IRL-based framework with computational efficiency is proposed to identify the reward function related to the state-action pair for the microgrid energy scheduling for the first time. Moreover, the proposed MMIRL method avoids the calculation of state visitation frequency by calculating the difference between the expert feature expectation and learner feature expectation, which saves computational cost and makes the algorithm more efficient.

The remainder of this paper is organized as follows. Section II states the main idea of the maximum entropy IRL method and the important improvements of our proposed MMIRL algorithms. Explanations about the microgrid system model is given in Section III. Specific simulations, problem formulation, and results are shown in Section IV. Finally, conclusions are provided in Section V.

II. Maximum Entropy Inverse Reinforcement Learning Approaches with Computational Efficiency

A. The Maximum Entropy IRL Method with Reward Related to the State

A Markov decision process (MDP) can be expressed by a tuple: $\{S, A, P, R, \gamma\}$, where S is a set of states s, A is a set of possible actions a, P is the transition probability, and R is the reward function [15].

In the maximum entropy IRL method [16], the path feature counts, $\mathbf{f}_\xi$, for a MDP path, ξ, is defined as the sum of the state features along the path. The feature for the state, s_i, is defined as $\mathbf{f}_{s_i}$. The relationship between the path feature counts and the state features can be formulated as

$$\mathbf{f}_\xi = \sum_{s_i \in \xi} \mathbf{f}_{s_i} \tag{1}$$

The goal for an inverse reinforcement learning agent is to extract the reward function from the expert demonstrations that linearly maps the features of each state, $\mathbf{f}_{s_i}$, to a state reward value,

$$R(\xi|\theta) = \theta^T \mathbf{f}_\xi = \sum_{s_i \in \xi} \theta^T \mathbf{f}_{s_i} \tag{2}$$

where θ is the reward weight applied to the path feature counts.

The expert feature expectation is formulated by the average path feature of all trajectories in the expert demonstrations,

$$\tilde{\mathbf{f}}_E = \frac{1}{N} \sum_{\xi \in \mathcal{D}} \mathbf{f}_\xi = \frac{1}{N} \sum_{\xi \in \mathcal{D}} \sum_{s_i \in \xi} \mathbf{f}_{s_i} \tag{3}$$

where N is the trajectory number of expert demonstrations, and $\mathcal{D}$ is the set of expert demonstrations.

The objective function for the maximum entropy IRL method is to maximize the entropy of the distribution over the demonstration from the observed data. The optimal weight θ^* is

$$\theta^* = \arg\max_\theta L(\theta) = \arg\max_\theta \sum_{\xi \in \mathcal{D}} logP(\xi|\theta) \tag{4}$$

where $L(\theta)$ is the log-likelihood function, and $P(\xi|\theta)$ is the probability of the trajectory ξ.

The gradient of $L(\theta)$ is expressed in terms of expected state visitation frequencies as

$$\nabla L(\theta) = \tilde{\mathbf{f}}_E - \sum_{s_i \in O} D_{s_i} \mathbf{f}_{s_i} \tag{5}$$

where D_{s_i} is the expected state visitation frequency of state s_i in [16] that is hard to be calculated in large-state space, and O is the set of learner trajectories.

In traditional maximum entropy IRL method, the feature of the trajectory is only related to the state, while there is a trend to apply the (s, a) pair to build features because sometimes the reward function is related to the (s, a) pair [17] [18]. Although these methods used the (s, a) pair to recover the reward, they still have some limitations of relying on the state visitation frequency or the regeneration of all the samples.

B. Proposed Modified Maximum Entropy IRL Method

For the majority of cases, the reward function is always associated with both state and action information, especially in the microgrid energy scheduling problems [19]. The maximum entropy IRL method requires the calculation of transition matrix and state visitation frequency, which brings computational challenges in the large and continuous state space. Based on the reality, we modify it using reward relate to the (s, a) pair called MMIRL method. This idea is inspired by calculating the margin of observed from learned feature expectations in [20]. It's more efficient than traditional maximum entropy IRL methods due to its feasibility in the large-state space.

The equation (1) of the path feature counts, $\mathbf{f}_\xi$, is modified as

$$\mathbf{f}_\xi = \sum_{(s_i, a_{i,j}) \in \xi} \mathbf{f}_{s_i, a_{i,j}} \tag{6}$$

The reward function is correspondingly changed as

$$R(\xi|\theta) = \theta^T \mathbf{f}_\xi = \sum_{(s_i, a_{i,j}) \in \xi} \theta^T \mathbf{f}_{s_i, a_{i,j}} \tag{7}$$

The expert feature expectation of (3) is reformulated as

$$\tilde{\mathbf{f}}_E = \frac{1}{N} \sum_{\xi \in \mathcal{D}} \sum_{(s_i, a_{i,j}) \in \xi} \mathbf{f}_{s_i, a_{i,j}} \tag{8}$$

Our method calculates the gradient with the approximation of expert feature expectation $\tilde{\mathbf{f}}_E$ and learner feature expectation $\tilde{\mathbf{f}}_{learn}$, instead of calculating the state visitation frequency. The equation (5) is modified as

$$\nabla L(\theta) = \tilde{\mathbf{f}}_E - \tilde{\mathbf{f}}_{learn} \tag{9}$$

978-1-6654-6441-3/23/$31.00 ©2023 IEEE

where similar to (8), the learner feature expectation $\tilde{\mathbf{f}}_{learn}$ can be defined as

$$\tilde{\mathbf{f}}_{learn} = \frac{1}{M} \sum_{\hat{\xi} \in O} \sum_{(s_i, a_{i,j}) \in \hat{\xi}} \mathbf{f}_{s_i, a_{i,j}} \tag{10}$$

where M is the trajectory number of learner trajectories.

This design saves the computational cost in both the discrete and continuous state-space problems by avoiding the computation of state visitation frequency. Moreover, considering the large or continuous state space, our MMIRL method can also be involved with deep RL methods to learn the optimal policy according to the recovered reward function, which avoids the computation and storage of large table in tabular methods.

In the microgrid energy scheduling problem, the actual reward intuitively depends to both the current state and the current action [21]. Therefore, our goal is to find out the reward function $R(s, a)$ through the expert demonstrations, then use the recovered reward function to learn the optimal microgrid energy scheduling policy.

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 shown in Fig. 1. The four units are the battery energy storage system (BESS), the distributed generations, including diesel generator (DG) and renewable generations (RG), the main grid, and the residential load. The optimization problem is to make hourly dispatch decisions over a time period of T (24 hours).

The state of our microgrid system at the end of hour t is defined as:

$$s_t = (s_{t,b}, s_{t,d}, s_{t,g}, s_{t,p}, s_{t,l}) \tag{11}$$

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 RG, $s_{t,p}$ is the retail energy price, and $s_{t,l}$ is the residential load demand.

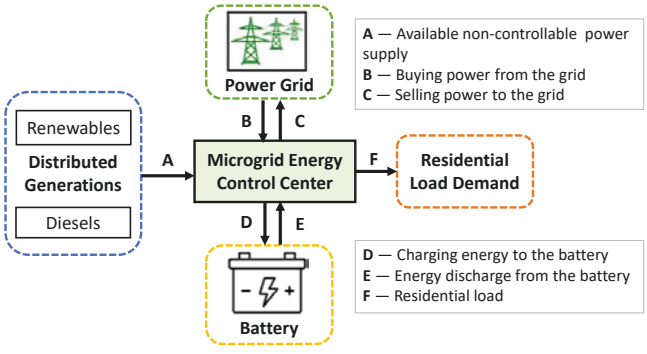

Fig. 1. A microgrid system consists of a battery unit, a distributed generations unit, a power grid unit, and a residential load demand unit. This is a modified figure from [22].

The transition function of SOC is given as

$$s_{t+1,b} = s_{t,b} - \frac{\Delta s_{t,b}}{E_b} \tag{12}$$

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

$$\Delta s_{t,b} = \begin{cases} \eta_b^- p_{t,b} \Delta t, & \text{if } p_{t,b} \leq 0 \\ \dfrac{p_{t,b}}{\eta_b^+} \Delta t, & \text{otherwise} \end{cases} \tag{13}$$

where Δt is the time step size, η_b^- and η_b^+ stand for the charging and discharging efficiencies respectively, and $p_{t,b}$ is the charging or discharging power of BESS, which is positive when discharging.

Besides, the SOC of the battery is constrained by

$$s_{b-} \leq s_{t,b} \leq s_{b+} \tag{14}$$

where s_{b-} and s_{b+} stand for the lower and upper bounds of SOC, respectively.

The $p_{t,b}$ should satisfy the limitations of the maximum discharging power P_b^+ and the maximum charging power P_b^-,

$$-P_b^- \leq p_{t,b} \leq P_b^+ \tag{15}$$

The action of our microgrid system at the end of hour t is defined as:

$$a_t = (p_{t,b}, p_{t,d}, p_{t,p}) \tag{16}$$

where $p_{t,d}$ is the power output of DG, and $p_{t,p}$ is the power purchased from (positive) or sold to (negative) the main grid.

The power balance of the microgrid can be expressed as

$$p_{t,b} + p_{t,d} + p_{t,p} + s_{t,g} = s_{t,l} \tag{17}$$

The DG status updated by the transitions in [23] is applied for determining the next-state of DG status. We uses the Q-learning method and the dynamic programming (DP) method to generate the expert demonstrations with the microgrid cost setting in [21]. For the simulations, we consider two case studies. One is an energy arbitrage problem with a microgrid system only involved with the units of the battery and the main grid, the other is the microgrid system involved with all four units.

IV. Simulation Results and Analysis

In this section, we conduct two case studies to evaluate the performance of our proposed MMIRL method. We first compare our approach with two expert policies, including the Q-learning method and the DP method. Then we also compare with the maximum entropy IRL method [16] and the imitation learning method [8] to justify the performance improvements.

A. Case 1: an energy arbitrage problem

The problem is how to schedule battery charging and discharging so that we can maximize the revenue on the basis of the fluctuations of the power price. The state s_t of case 1 can be simplified as $(s_{t,b}, s_{t,p})$, and the action a_t is the BESS charging/discharging power $p_{t,b}$. Therefore, the objective function of case 1 is maximizing the total reward over the time period of T,

$$\max_{p_{t,b}} \sum_{t=1}^{T} R(s_t, a_t) = \max_{p_{t,b}} \sum_{t=1}^{T} [s_{t,p} p_{t,b}] \tag{18}$$

978-1-6654-6441-3/23/$31.00 ©2023 IEEE

We use the DP method as our expert policy to implement our MMIRL method. To compare the difference of using the reward related to s and (s, a), we also implement the maximum entropy IRL method using (s, a) features. Besides, we discretize the battery SOC $s_{t,b}$ into 11 states from 0.1 to 0.9. The initial SOC is defined as 0.1 and the energy storage parameters are set as

- $E_b = 4$ MWh, $T = 24$ hours
- $\eta_b^+ = \eta_b^- = 87\%$
- $s_{b+} = 0.9, s_{b-} = 0.1$
- $P_b^+ = P_b^- = 1$ MW

We conduct our MMIRL algorithm to recover the reward function of $R(s, a)$ and generate the learner's optimal policy by two RL methods using the recovered reward, including the Q-learning method and the deep Q-network method. It is worth mentioning that our MMIRL algorithm recovers the reward function related to the (s, a) pair. The simulation results are shown in Table I. For case 1, the higher reward is better since the goal is to maximize the reward.

TABLE I
SIMULATION RESULTS FOR CASE 1

Methods	Total Reward ($)	Reward Accuracy
Expert: DP	46.66	-
Max Entropy IRL (s)	44.84	96.1%
Imitation Learning	44.94	96.3%
Max Entropy IRL (s, a)	**46.66**	**100%**
MMIRL (s, a)	**46.66**	**100%**

As we can see from the value of recovered policy's total reward, all these methods can achieve more than 96% accuracy compared with the DP expert policy. From the results of maximum entropy IRL method using s feature and (s, a) feature, we can conclude that recovering reward related to the (s, a) pair can have better performance with 100% accuracy in this energy arbitrage problem. Moreover, our proposed MMIRL algorithm can precisely recover the reward and the optimal learner policy the same as the expert with 100% accuracy. This performance is 4.1% better than the maximum entropy IRL method using s feature and 3.8% better than the imitation learning method in the reward accuracy metric.

B. Case 2: a microgrid system with DERs

The state and action of this case are defined as (11) and (16). This microgrid energy scheduling problem is to minimize the total operational cost during the time T,

$$\min_{a_t} \sum_{t=1}^{T} C(s_t, a_t) = \min_{a_t} \sum_{t=1}^{T} [C_{t,d}(p_{t,d}) + C_{t,p}(p_{t,p})] \quad (19)$$

where $C_{t,d}(p_{t,d})$ is the operational cost of DG expressed as

$$C_{t,d}(p_{t,d}) = s_{t,d}(a_d p_{t,d}^2 + b_d p_{t,d} + c_d) \quad (20)$$

where $p_{t,d}$ is the power output of DG, a_d, b_d, and c_d are the coefficients of the quadratic function. The power purchasing cost or selling revenue $C_{t,p}(p_{t,p})$ is expressed as

$$C_{t,p}(p_{t,p}) = \begin{cases} p_{t,p} s_{t,p} \eta \Delta t, & \text{if } p_{t,p} \geq 0 \\ \dfrac{p_{t,p} s_{t,p} \Delta t}{\eta}, & \text{otherwise} \end{cases} \quad (21)$$

where η is used to capture the network losses.

The cost function value equals to the opposite value of the reward function. And the parameters of the dispatchable DG and the BESS for this microgrid are adopted from [21]. The rated power of wind for RG is defined as 100 kW, and the power output profiles are generated using the System Advisory Model [24] for the city of Phoenix, Arizona. Similarly, the residential load profiles and power prices are from Phoenix with a peak of 150 kW in a time period of 24 hours. The detailed setup description can be found in [21]. We also discretize the battery SOC into 20 status from 0.1 to 0.9, and the initial SOC status is set as 0.1 with $\eta = 0.915$.

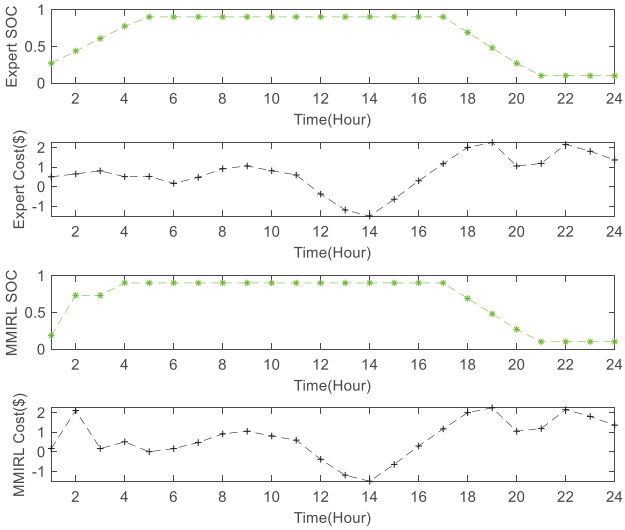

Fig. 2. The comparison results of the DP expert policy and the propose MMIRL algorithm for one of the 5 experiments.

We use 5000 iterations for the Q-learning approach and the DP approach as two expert policies. Similarly, we implement our proposed MMIRL algorithm applied with the Q-learning expert demonstrations and the DP expert demonstrations respectively, and compared with two expert policies and two comparison methods. Because the state in case 2 involved with more parameters and the reward is related to (s, a) pair, these factors lead to the dimension explosion for case 2. In order to speed up the computation, we apply deep Q network to learn the learner's optimal policy with the recovered reward function generated from our MMIRL algorithm. To be fair, 5 experiments are conducted for the proposed MMIRL algorithm for 20000 epochs, and the average results are recorded.

Fig. 2 illustrates the MMIRL algorithm's SOC status and cost in comparison with the DP expert method in one of the

978-1-6654-6441-3/23/$31.00 ©2023 IEEE

TABLE II
SIMULATION RESULTS OF THE MICROGRID FOR CASE 2

Methods	Expert Demos	Total Cost ($)	Cost Accuracy
Q-learning	-	17.95	-
Max Entropy IRL (s)	Q-learning	24.61	73.0%
Imitation Learning	Q-learning	19.94	90.0%
MMIRL (s, a)	**Q-learning**	**18.15**	**98.9%**
Dynamic Programming	-	16.55	-
Max Entropy IRL (s)	DP	29.23	56.6%
Imitation Learning	DP	18.06	91.6%
MMIRL (s, a)	**DP**	**16.63**	**99.5%**

5 experiments. It's clear that the SOC status of DP expert and the MMIRL method is almost the same, except the first five hours. Moreover, when it comes to Table II, our MMIRL algorithm outperforms other comparative methods using both two kinds of expert demonstrations. The MMIRL method using the Q-learning expert demonstrations achieves a $\$18.15$ total cost and 98.9% accuracy. Specially, when the proposed MMIRL algorithm using the DP expert demonstrations, it achieves a lowest $\$16.63$ total cost and 99.5% accuracy. This performance is 8.6% better than the imitation learning method, and 75.8% better than the maximum entropy IRL method using the same DP demonstrations.

V. CONCLUSIONS

In this paper, a computationally efficient IRL approach called the MMIRL method is proposed to extract the reward function of the microgrid energy scheduling problem from the expert demonstrations. The Q-learning method and DP method were used for generating the expert demonstrations, and the maximum entropy IRL method and the imitation learning method were introduced for comparisons. Two case studies of an energy arbitrage problem and a microgrid system with DERs were conducted to validate the effectiveness of the proposed MMIRL approach. There is a reality that the reward function of microgrid optimization is related to the (s, a) pair. Therefore, the main contribution of our work is the realization of recovering the reward function of the (s, a) pair for microgrid energy scheduling for the first time. Besides, our approach can learn the dispatch policy through the recovered reward function without the need for the computation of state visitation frequency, which is more efficient than conventional IRL methods. Our experiments show that the proposed MMIRL algorithm can achieve more than 99% accuracy in the aspects of total reward or total cost compared with the expert policies, and outperform other existing methods in both cases.

ACKNOWLEDGMENT

This work is partially supported by National Science Foundation under grants # 2047064 and 1949921.

REFERENCES

[1] S. Ishaq, I. Khan, S. Rahman, T. Hussain, A. Iqbal, and R. M. Elavarasan, "A review on recent developments in control and optimization of micro grids," *Energy Reports*, vol. 8, pp. 4085–4103, 2022.

[2] N. Bazmohammadi, A. Anvari-Moghaddam, A. Tahsiri, A. Madary, J. C. Vasquez, and J. M. Guerrero, "Stochastic predictive energy management of multi-microgrid systems," *Applied sciences*, vol. 10, no. 14, p. 4833, 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] H. Tang and S. Wang, "A model-based predictive dispatch strategy for unlocking and optimizing the building energy flexibilities of multiple resources in electricity markets of multiple services," *Applied Energy*, vol. 305, p. 117889, 2022.

[5] K. Zhou, K. Zhou, and S. Yang, "Reinforcement learning-based scheduling strategy for energy storage in microgrid," *Journal of Energy Storage*, vol. 51, p. 104379, 2022.

[6] 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)*. IEEE, 2021, pp. 1–5.

[7] Y. Lin, D. Duan, X. Hong, X. Han, X. Cheng, L. Yang, and S. Cui, "Transfer learning on the feature extractions of sky images for solar power production," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2019, pp. 1–5.

[8] 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, 2022.

[9] S. Li, P. Zhao, C. Gu, J. Li, S. Cheng, and M. Xu, "Online battery protective energy management for energy-transportation nexus," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 8203–8212, 2022.

[10] Y. Li, R. Wang, and Z. Yang, "Optimal scheduling of isolated microgrids using automated reinforcement learning-based multi-period forecasting," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 1, pp. 159–169, 2021.

[11] 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.

[12] Q. Tang, H. Guo, and Q. Chen, "Multi-market bidding behavior analysis of energy storage system based on inverse reinforcement learning," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4819–4831, 2022.

[13] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML, 2000, p. 663–670.

[14] H. Guo, Q. Chen, Q. Xia, and C. Kang, "Deep inverse reinforcement learning for objective function identification in bidding models," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5684–5696, 2021.

[15] Y. Lin, Z. Ni, and X. Zhong, "Multi-virtual-agent reinforcement learning for a stochastic predator-prey grid environment," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[16] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *Aaai*, vol. 8, 2008, pp. 1433–1438.

[17] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *arXiv preprint arXiv:1507.04888*, 2015.

[18] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. JMLR.org, 2016, p. 49–58.

[19] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.

[20] S. Adams, T. Cody, and P. A. Beling, "A survey of inverse reinforcement learning," *Artificial Intelligence Review*, pp. 1–40, 2022.

[21] A. Das, Z. Ni, and D. Wu, "An efficient distributed reinforcement learning for enhanced multi-microgrid management," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–6.

[22] A. Das, Z. Ni, T. M. Hansen, and X. Zhong, "Energy storage system operation: Case studies in deterministic and stochastic environments," in *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2016, pp. 1–5.

[23] H. Shuai, J. Fang, X. Ai, J. Wen, and H. He, "Optimal real-time operation strategy for microgrid: An adp-based stochastic nonlinear optimization approach," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 931–942, 2018.

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