Integrating Human-in-the-Loop for Safe Reinforcement Learning in Optimal Operation of Community Energy Storage Systems

Van-Hai Bui, Guilherme Vieira Hollweg, and Wencong Su
Department of Electrical and Computer Engineering
University of Michigan-Dearborn, MI, USA

Abstract

Reinforcement learning (RL) is becoming a potential solution for solving optimization problems in power and energy systems. However, a major issue with conventional RL is that it does not guarantee the safe operation of critical infrastructures such as microgrids or power systems. Therefore, this paper proposes a safe RL-based optimization framework with a human-in-the-loop approach for the operation of a community energy storage system (CESS) in community microgrid (MG) systems. The proposed framework not only maximizes the CESS's profit but also reduces the amount of load shedding in the MG during emergency situations. To demonstrate the effectiveness of the proposed framework, safe Q-learning is implemented to optimize the operation of the CESS with human input, aiming to avoid all catastrophic actions at critical states.

1 Introduction

A community microgrid (MG) is a small-scale power system, which is a localized group of electricity sources and loads [Cornélusse et al., 2019]. The MG comprises various sources of distributed generation, such as diesel generators (DGs), photovoltaics (PV), wind turbines, energy storage systems (ESSs), and intelligent control mechanisms, providing a reliable and efficient energy supply. Therefore, community MGs can offer various benefits to regional and rural communities, such as improving the reliability of their electricity network and reducing their electricity bills [Syed and Morrison, 2021], [Salehi et al., 2022]. The MG typically connects to a utility grid in normal operation, but it can also disconnect and operate in island mode under emergency conditions, functioning autonomously to supply the community's demand.

Workshop on Artificial Intelligence for Critical Infrastructure (AI4CI 2024) @ IJCAI'24, Jeju Island, South Korea, https://sites.google.com/view/aiforci-ijcai24/, August 4, 2024. Eds: F. Silva, W. Su, R. Glatt, Y. Wang.

Corresponding authors: V.H. Bui (vhbui@umich.edu) and W. Su (wencong@umich.edu)

In order to efficiently operate community MGs, many optimization algorithms have been introduced to minimize operation costs or enhance system reliability. For instance, [Al-Sorour et al., 2022] have presented a peer-to-peer mechanism aimed at reducing net energy trading with the utility. The approach uses a two-day-ahead energy forecast and also allows energy trading among local prosumers using a mixed-integer linear programming (MILP) model. [Balderrama et al., 2019] have developed a two-stage linear programming optimization framework for a community islanded MG in a rural neighborhood. However, such methods often show an inability to adapt dynamically to changing conditions. They require a reconfiguration of the model and objectives as system dynamics change. Therefore, they frequently face many challenges in handling the uncertainty in the MG system caused by the high penetration of renewable distributed generators (RDGs).

To overcome such limitations, a machine learning-based optimization approach, specifically, reinforcement learning (RL), is becoming crucial. RLs offer the ability to continuously learn and adapt to new situations without human reprogramming [Erick et al., 2020]. [Hasan et al., 2022] have developed a control strategy for MGs in which universal droop control, virtual inertia control, and an RL-based control mechanism have been combined for online tuning of the controllers' parameters. [Mbuwir et al., 2020] have proposed a framework combining distributed optimization and RL for congestion management. This approach offers microgrids a potential solution to provide flexibility and avoid congestion problems in distribution grids.

Despite the potential of RL in optimizing the operation of MGs, implementing conventional RL algorithms in critical infrastructure such as power and energy systems and MGs poses significant safety risks. In order to avoid catastrophic failures, integrating safety into the conventional RL-based optimization framework is necessary and is gaining increasing attention in both academia and industry [Qiu et al., 2022], [Gu et al., 2022].

Therefore, in this study, a safe Q-learning-based optimization framework with a human-in-the-loop approach is proposed for the optimal operation of a CESS in a community

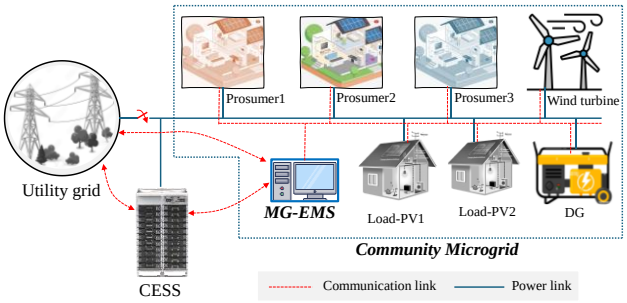

Fig. 1. Community microgrid systems.

MG system. The CESS is controlled to maximize its profit in grid-connected mode and to minimize the load shedding amount in the MG during islanded mode. Adding humans in the loop in the training process can ensure that the CESS agent's decisions are continuously overseen and adjusted by experienced operators. Therefore, the CESS agent can guarantee that its operation is always within the predefined safety operation boundary. This approach also ensures not only that the energy supplied is cost-effective but also that it enhances the system's reliability against operational risks. The effectiveness of the proposed model is also validated in the numerical results section using several test scenarios.

2 System Model

In this section, we develop an optimization framework for a CESS integrated with a community MG in both grid-connected and islanded modes. The main operational objective aims to maximize the profit of the CESS and to minimize the operational cost as well as the load shedding amount of the MG system. Additionally, all safety issues are also addressed with safe Q-learning and the human-in-the-loop approach.

2.1 Community Microgrid Systems

A typical community MG is depicted in Fig. 1, which includes multiple prosumers, wind turbines, photovoltaic systems (PVs), and diesel generators (DGs) to supply its demand [Cornélusse et al., 2019], [Trivedi et al., 2022]. The optimal operation of the community MG is carried out by a microgrid energy management system (MG-EMS). The community MG is also connected to a community energy storage system (CESS) and to the utility grid for economical operation and enhanced MG reliability.

Although the operation of the CESS is independent from the community MG in normal operation, the CESS is also controlled to support the MG and maintain power balance during emergency operations, including islanded mode. The CESS is controlled using the Q-learning method, and safety operation is taken into account with human feedback during the training process. The Q value is set by human input for specific state-action pairs, as illustrated in Fig. 2.

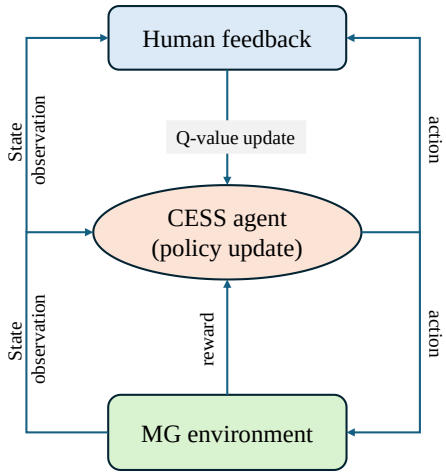

Fig. 2. Safe Q-learning-based operation with human feedback (Q-value adjustment).

2.2 System Operation Strategy

This section presents the system operation strategy, including safe Q-learning-based operation of the CESS and MILP-based operation of the MG. Algorithm 1 presents a detailed safe Q-learning framework with a human-in-the-loop approach. The CESS agent observes the system state and determines an action using the Q-table. During the training process, the agent executes the selected action and receives feedback from the MG after running the MILP model. The adjusted reward is then calculated based on the profit of the CESS and potential load shedding information from the MG-EMS. The agent observes the new state and updates the Q-table policy. The same process is carried out with a large number of episodes for training convergence. The Q-table is used to determine the optimal operation of the CESS.

The detailed interaction among the CESS agent, MG-EMS, and human monitoring is summarized in Fig. 3. State information is observed by both the CESS agent and a human, and then the CESS agent selects an action, informing both the MG-EMS and the human monitor. The MG-EMS performs optimization and estimates the load shedding that should be performed in the MG system, while the human monitor determines the Q value applied for some critical states to avoid catastrophic actions, as shown in Tables 1 and 2. Finally, the Q table is updated using state, action, new state, and adjusted reward information. The optimal output of the CESS can be found with an optimal policy (i.e., Q-table). Human monitoring not only ensures that the agent does not take catastrophic actions at critical states but also ensures that the training process is converging and stable by regularly validating the Q-table.

A detailed mathematical model is presented in the next section for both MG and CESS operation.

Algorithm 1: Safe RL with human-in-the-loop approach

1. Initialize MG environment state
2. Load RL agent with initial policy Q table
3. **For loop**: each episode from 1 to N:
4. Reset environment to a starting condition
5. **While loop**: each timestep from 1 to K:
6. Agent observes current state
7. Selects action based on Q table and epsilon greedy
8. Carry out action in the environment
9. Formulate and run MILP
10. *Input:* Current state, action, operational constraints
11. *Output:* Adjusted reward
12. Update agent with new state and adjusted reward
13. Ensure actions meet pre-defined safety constraints *(tables 1 and 2)*
14. Environment transitions to new state based on action
15. Adjust policy based on the learning algorithm
16. **End while loop**
17. **End for loop**
18. Analyze performance of the policy
19. Determine the optimal operation of CESS
20. **End**

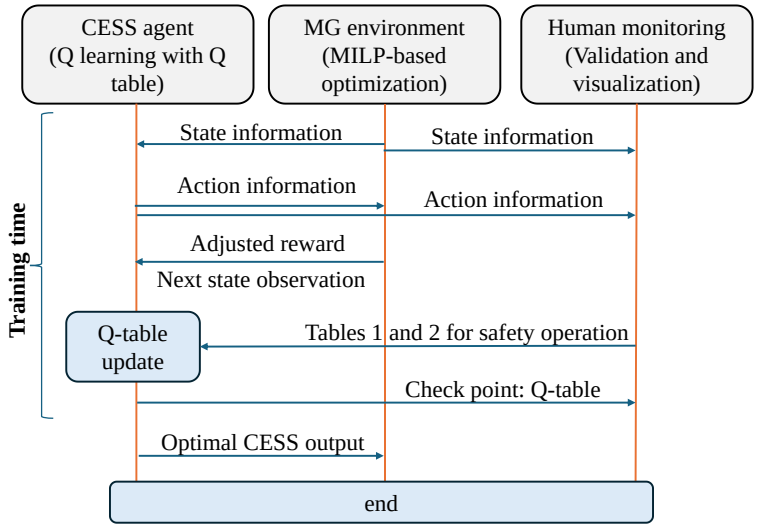

Fig. 3. Interaction among CESS agent, MG-EMS, and human monitoring and feedback.

2.3 Mathematical Model

First, a detailed mathematical model is presented, showing the action selection of the CESS agent and Q-table update in different scenarios. At each given state s_t, the agent selects an action following Equation (1) with the epsilon- greedy algorithm. The actions could be {-1, 0, 1}, corresponding to the {discharging mode, idle mode, charging mode} of the CESS.

$$a_t = \begin{cases} argmax\{Q(s_t, a_i)\}, & if\ \beta \geq \varepsilon \\ rand\{a_i\}, & if\ \beta < \varepsilon \end{cases} \quad (1)$$

$$\forall i \in \{-1,0,1\}$$

where: a_t is a selected action at t, s_t is the observed state at t, $Q(s_t, a_i)$ is Q value at state s_t and action a_i with i in {-1,0,1}, β is a random value between 0 and 1, ε is epsilon value for epsilon greedy algorithm.

The reward of the CESS is calculated using Equation (2) based on the charging/discharging amount.

$$r_t = -PR_{buy,t}.P_{char,t} + PR_{sell,t}.P_{dis,t}\ \forall t \in T \quad (2)$$

where: r_t is the reward value at t, $PR_{buy,t}$ and $PR_{sell,t}$ are buying price and selling price with the utility grid or MG at

t, $P_{char,t}$ and $P_{dis,t}$ are the amount of charging and discharging power at t, respectively.

The final reward for the CESS agent should be the sum of the CESS profit and the potential penalty for load shedding at the MG, as expressed in Equations (3) and (4).

$$r_t^{mg} = -pen_t.P_{shed,t} \qquad \forall t \in T \qquad (3)$$

$$r_t = r_t + r_t^{mg} \qquad \forall t \in T \qquad (4)$$

where: r_t^{mg} is the reward at t considering the response of MG environment. pen_t and $P_{shed,t}$ are penalty and load shedding amount at t, respectively.

The Q-table is updated for each state and action pair using the Bellman's equation [Bui et al., 2019], [Sutton et al., 2018], as shown in Equation (5).

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \begin{pmatrix} r_t + \gamma.\max\big(Q(s_{t+1}, a_i)\big) \\ -Q(s_t, a_t) \end{pmatrix} \quad \forall t \in T \qquad (5)$$

where: $\gamma \in [0,1]$ is the discount factor for the updated Q-table.

The operational constraints for the CESS are shown in Equations (6) to (9). The amounts of charging and discharging power are bounded, as expressed in Equations (6) and (7), respectively. These boundaries are calculated at each interval based on the current SOC of the CESS, the capacity, and the efficiency. The current SOC of the CESS is updated using Equation (8) based on the actual amounts of charging and discharging power. Finally, the SOC of the CESS must always remain between the predefined minimum and maximum setpoints of the SOC, as shown in Equation (9).

$$0 \le P_{dis,t} \le SOC_{t-1}.P_{\max}^{CESS}.\eta_{CESS} \qquad \forall t \in T \qquad (6)$$

$$0 \le P_{char,t} \le (1 - SOC_{t-1}).\frac{P_{\max}^{CESS}}{\eta_{CESS}} \quad \forall t \in T \qquad (7)$$

$$SOC\_t = SOC_{t-1} + P_{char,t}.\eta_{CESS} - \frac{P_{dis,t}}{\eta_{CESS}} \quad \forall t \in T \qquad (8)$$

$$SOC_{min} \le SOC_t \le SOC_{max} \; \forall t \in T \qquad (9)$$

where: SOC_t is the state of charge of the CESS at t, $P_{\max}^{CESS}$ is capacity of the CESS, η_{CESS} is the operation efficiency of the CESS, SOC_{min} and SOC_{max} are the minimum and maximum SOC of the CESS, respectively.

The MILP-based mathematical model for the operation of the MG is presented in Equations (10)–(13), including an objective function and three major operational constraints. The objective function aims to minimize the operation cost of the MG in both grid-connected and islanded modes. The operation cost consists of the generation cost of the DG, the trading cost with the utility, and the load shedding penalty in islanded mode, as given in Equation (10). The MG operation status can be determined by 0 or 1, as in Equation (11), depending on its operation mode.

$$\min\left\{\sum_t^{N-t}\Big(C_{dg}.P_{dg,t} + (1-u_t).pen_t.P_{shed,t} + u_t\big(PR_{buy,t}.P_{buy,t} - PR_{sell,t}.P_{sell,t}\big)\Big)\right\} \qquad (10)$$

$$u_t = \begin{cases} 0, & if\ islanded\ mode \\ 1, & if\ grid-connected\ mode \end{cases} \quad \forall t \in T \qquad (11)$$

where: C_{dg} and $P_{dg,t}$ are the generation cost and generation amount, respectively, of the DG at t. u_t is the operation mode of the microgrid system.

The operational boundary of the DG is given in Equation (12).

$$P_{dg}^{min} \le P_{dg,t} \le P_{dg}^{max} \; \forall t \in T \qquad (12)$$

where: P_{dg}^{min} and P_{dg}^{max} are the minimum and maximum setpoints of DG.

Power balance is always maintained in the MG system, as expressed in Equation (13). The power supply typically comes from various sources, including the DG, RDGs, the CESS, and the utility grid; however, load shedding sometimes must still be performed to maintain power balance throughout the system during islanded mode.

$$P_{dg,t} + P_{RDG,t} - P_{char,t} + P_{dis,t} + u_t.\big(P_{buy,t} - P_{sell,t}\big) = P_{load,t} - (1-u_t).P_{shed,t} \quad \forall t \in T \qquad (13)$$

where: $P_{RDG,t}$ is the output of the RDG at t, $P_{load,t}$ is the load amount in the MG system at t.

2.4 Training Safe Q-Learning with HITL

This section presents different rules for the safe operation of the CESS in both grid-connected and islanded modes. In grid-connected mode, two constraints should be considered that form the safety operation boundary of the CESS. The rule is outlined in Table 1, considering the minimum and maximum amounts of SOC of the CESS. If the SOC falls to 10%, charging action is prohibited; if the SOC rises to 90%, discharging action is prohibited, and therefore the Q-value of such state-action pairs is set to -*Infinity*. The CESS agent will always avoid that action in real-time operation.

Similarly, the CESS operational boundaries are also considered in islanded mode. Additionally, two more constraints are considered to assist the MG in maintaining power balance within the system while simultaneously reducing the penalty of the load shedding amount. If there is a power

shortage in the MG, the Q-values for the charging and idle modes are set at *-Infinity* to avoid taking such actions during the real-time operation of the CESS. Conversely, if there is surplus power, the Q-values for the discharging and idle modes are set at *-Infinity*, as given in Table 2.

Con-straints	Q values		
	Discharging mode	Charging mode	Idle mode
$SOC \leq 10\%$	-Inf	-	-
$SOC \geq 90\%$	-	-Inf	-

Table 1: Human safety monitoring in grid-connected mode

Human monitoring, as detailed in both Tables 1 and 2, plays a crucial role in the safe operation of the MG and CESS. It not only ensures that the CESS always operates within the allowed operational boundary but also enhances system reliability by reducing the amount of load shedding.

Constraints	Q values		
	Discharging mode	Charging mode	Idle mode
$SOC \leq 10\%$	-Inf	-	-
$SOC \geq 90\%$	-	-Inf	-
Shortage power	-	-Inf	-Inf
Surplus power	-Inf	-	-Inf

Table 2: Human safety monitoring in islanded mode

3 Numerical Results

This section evaluates the proposed optimization framework for the operation of a community MG and a CESS. The training and validation process of the CESS is presented in detail, and the operation of the MG is also scheduled considering the optimal output of the CESS.

3.1 Input Data

The test system includes a DG, wind turbines, PVs, a CESS, and loads. The system can be connected to or disconnected from the utility grid. The detailed parameters of the DG and CESS, including minimum and maximum operation points, capacity, operational costs, etc., are tabulated in Table 3.

Figure 4 presents the output power of the RDGs, including both PVs and wind turbines, and the total amount of load demand in the MG system. Figure 5 depicts the market price signal, showing the buying/selling price at each interval. The generation cost of the DG is fixed at $0.50/kWh. This input data is used to train the CESS agent and is also taken as input information for the MILP-based optimization model for the community MG system.

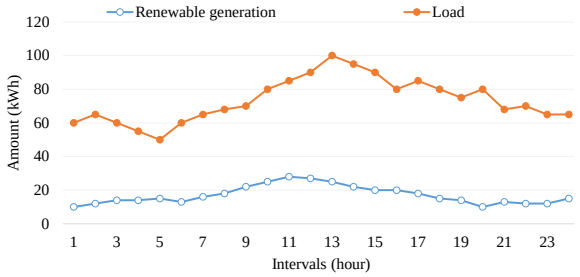

Figure 4. Output power of RDG and load demand.

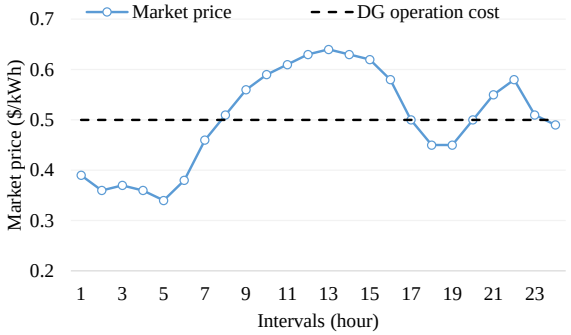

Figure 5. Market price signal and generation cost of DG.

DG	Values	CESS	Values
Min. setpoint (kWh)	0	Min. SOC (%)	10
Max. setpoint (kWh)	100	Max. SOC (%)	90
Generation cost ($/kWh)	0.5	Capacity (kWh)	100
		Init. SOC (%)	50

Table 3: Parameters of DG and CESS

3.2 Optimal Operation of Test MG Systems

The operation of the CESS is determined using an optimal policy Q-table. The Q-table is updated during the training

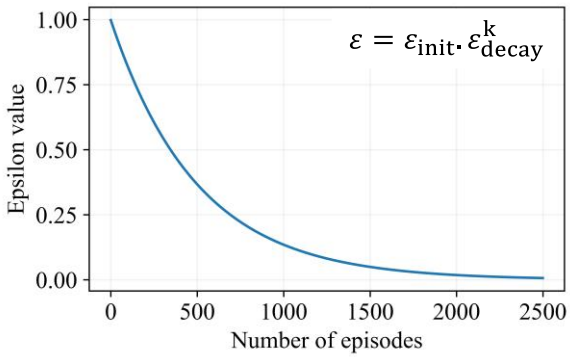

Figure 6. The value of epsilon during the training process.

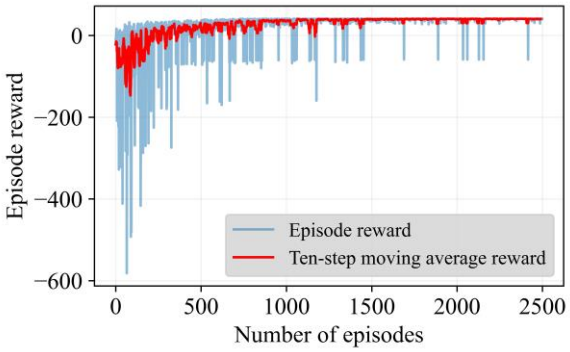

Figure 7. The episode reward of CESS during the training process.

process, integrating human monitoring feedback. The Q-values of critical state-action pairs that cause safety violations are set directly to *-Infinity*. Table 4 shows a few examples of an optimal Q-table with human feedback. This helps to avoid any catastrophic actions during the real-time operation of the CESS.

SOC (%)	Q value		
	Discharging mode	Idle mode	Charging mode
50	39.541542	39.241182	38.940822
60	42.845107	42.804465	42.542311
60	42.847183	41.631251	41.413499
60	42.752587	42.551671	42.632087
70	45.656434	46.516087	42.136382
...	...	...	...
90	0	0	-999999
10	-999999	0	1.008926
10	5.493535	0	0
10	-999999	0	2.322714
10	-999999	0	3.176883

Table 4: Q-table with human feedback (highlighted values)

The training process of the CESS agent is visualized in Figures 6 and 7. Figure 6 shows the value of epsilon during the training process, which represents the probability of selecting a random action. It can be observed that the value of epsilon gradually decreases from 1 to 0. This means that in the early episodes, the agent has a high chance of selecting a random action to explore the environment. When the agent has sufficient knowledge about the environment, the value of epsilon approaches 0, and the agent will choose actions to maximize its cumulative reward. Figure 7 shows the episode reward of the CESS agent during the training process. During the first few episodes, the agent lacks environmental information and does not select optimal actions; therefore, the reward is very low. However, it improves significantly

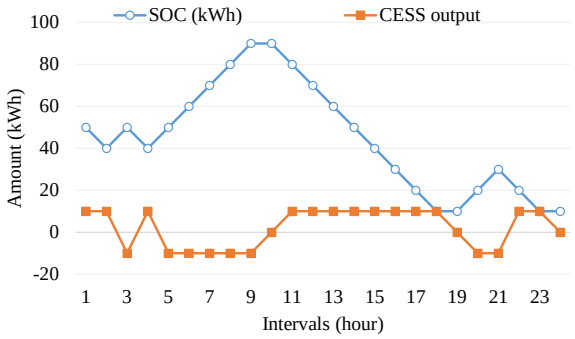

Figure 8. the SOC and optimal output of CESS.

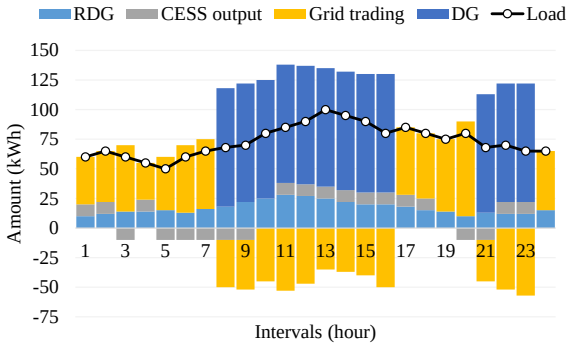

Figure 9. The schedule for operation MG system.

after the training process. The episode reward has converged and reached maximum value. This will also provide an optimal policy Q-table for the optimal operation of the CESS.

The optimal operation of the CESS is presented in Figure 8, using the optimal Q-table after the training process. The CESS is charged to the maximum value during the low-price intervals and then is discharged to the minimum value during high-price intervals to maximize its profit. The MG-EMS receives the actual output power of the CESS and carries out optimization to schedule the operation of the entire MG system, including the DG and trading with the grid. It can be seen from Figure 9 that the power balance is always maintained throughout the time-scheduling horizon.

4 Conclusions

This study has proposed a safe RL-based optimization model with a human-in-the-loop approach for the operation of a CESS. The model not only ensures that the CESS maximizes its profit but also guarantees safe operation. The actual output of the CESS is shared with the MG-EMS to determine the optimal operation of the entire MG system. The numerical results show that the CESS plays a crucial role in the operation of the community MG system, especially in islanded mode. The CESS is controlled to minimize the penalty of load shedding within the MG system.

Acknowledgements

The author's work was supported by the University of Michigan-Dearborn's Office of Research "Research Initiation &

Development."

References

[Cornélusse et al., 2019] B. Cornélusse, I. Savelli, S. Paoletti, A. Giannitrapani, and A. Vicino. A community microgrid architecture with an internal local market. *Applied Energy*, 242, pp.547-560, 2019.

[Syed and Morrison, 2021] M.M. Syed and G.M. Morrison. A rapid review on community connected microgrids. *Sustainability*, 13(12), p.6753, 2021.

[Salehi et al., 2022] N. Salehi, H. Martínez-García, G. Velasco-Quesada, and J.M. Guerrero. A comprehensive review of control strategies and optimization methods for individual and community microgrids. *IEEE Access*, 10, pp.15935-15955, 2022.

[Al-Sorour et al., 2022] A. Al-Sorour, M. Fazeli, M. Monfared, A. Fahmy, J.R. Searle, and R.P. Lewis. Enhancing PV self-consumption within an energy community using MILP-based P2P trading. *IEEE Access*, 10, pp.93760-93772, 2022.

[Balderrama et al., 2019] S. Balderrama, F. Lombardi, F. Riva, W. Canedo, E. Colombo, and S. Quoilin. A two-stage linear programming optimization framework for isolated hybrid microgrids in a rural context: The case study of the "El Espino" community. *Energy*, 188, p.116073, 2019.

[Erick et al., 2020] A.O. Erick and K.A. Folly. Reinforcement learning approaches to power management in grid-tied microgrids: A review. In *2020 Clemson University Power Systems Conference (PSC)*, pp. 1-6, IEEE, March 2020.

[Hasan et al., 2022] M.M. Hasan, I. Zaman, M. He, and M. Giesselmann. Reinforcement Learning-Based Control for Resilient Community Microgrid Applications. *Journal of Power and Energy Engineering*, 10(09), 2022.

[Mbuwir et al., 2020] B.V. Mbuwir, F. Spiessens, and G. Deconinck. Distributed optimization for scheduling energy flows in community microgrids. *Electric Power Systems Research*, 187, p.106479, 2020.

[Qiu et al., 2022] D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac. Safe reinforcement learning for real-time automatic control in a smart energy-hub. *Applied Energy*, 309, p.118403, 2022.

[Gu et al., 2022] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, Y. Yang, and A. Knoll. A review of safe reinforcement learning: Methods, theory, and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[Trivedi et al., 2022] R. Trivedi, S. Patra, Y. Sidqi, B. Bowler, F. Zimmermann, G. Deconinck, A. Papaemmanouil, and S. Khadem. Community-based microgrids: Literature review and pathways to decarbonise the local electricity network. *Energies*, 15(3), p.918, 2022.

[Bui et al., 2019] V.H. Bui, A. Hussain, and H.M. Kim. Q-learning-based operation strategy for community battery energy storage system (CBESS) in microgrid system. *Energies*, 12(9), p.1789, 2019.

[Sutton and Barto, 2018] R.S. Sutton and A.G. Barto. Reinforcement learning: An introduction. *MIT Press*, 2018.