Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 11741–11747

Optimal Scheduling of Battery Energy Storage Systems Using a Reinforcement Learning-based Approach

Alaa Selim * **Huadong Mo** * **Hemanshu Pota** * **Daoyi Dong** *

* *School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2610 Australia (e-mail: a.selim@adfa.edu.au, huadong.mo@adfa.edu.au, h.pota@adfa.edu.au, d.dong@adfa.edu.au).*

Abstract: This article proposes a novel energy management algorithm that controls the battery energy storage system (BESS) and on-grid supply. It employs the deep-Q-network agent with prioritized experience replay, and its efficacy is validated and verified by comparison to a benchmark method for mixed integer linear programming. The grid and energy storage systems are governed by switching operations initiated by BESS controllers via the automatic transfer switch. The primary objective is to accomplish optimal scheduling of batteries one day in advance to reduce electricity costs while maintaining battery health and primary power supply reliability. The methods proposed in this work provide practicable grid and battery operation patterns that test all conceivable planning scenarios for energy storage operation. Finally, a comparative analysis is performed to evaluate the efficacy of the proposed BESS operation scheduling methods.

Keywords: Energy Storage, Control optimization, State of charge, Deep-Q-network (DQN), Prioritized experience replay, Mixed integer linear programming (MILP).

1. INTRODUCTION

According to the Australian Energy Market Operator (AEMO) reports in Datta et al. (2018), and Aryai and Goldsworthy (2022), several plans have been proposed to provide scenario-based projections to 2051 for solar PV and batteries installations for accommodating the residential community loads. These plans aimed to assist AEMO in managing the commercial battery systems under defined operation modes of batteries such as solar shift mode and tariff optimization mode. Additionally, controlling solar PV and batteries onsite is a non-trivial problem due to many state space variables, which cannot be easily deduced and needs advanced algorithms to allocate the control set-points within the system constraints. Consequently, new technologies for managing the distributed energy resources (DER) with the grid have been developed to deduce the optimized operational set-points for the system operator.

Several techniques of control algorithms were investigated and applied for the optimal control of BESS. Benchmark models for the BESS control problem were conducted based on deterministic approaches and linear programming techniques. In Jeddi et al. (2019), the dynamic programming method showed better results in optimizing battery dispatch using stochastic control strategies considering historic load profiles, weather forecasts, and optimal grid limit. In Kong et al. (2021), a dynamic programming-based control scheme was used to deduce the BESS's optimal charging/discharging decisions over a billing cycle, targeting to minimize the customer's home energy cost subject to a specific tariff structure.

Another research by Vedullapalli et al. (2019) proposed the model predictive control (MPC) as the most emerging technique for the battery scheduling problem since it computes the trajectory of future control inputs to optimize the future behavior of the plant output. Furthermore, MPC was used in parallel with mixed integer linear programming (MILP) to provide operation schedules for BESS, heating, ventilation, and air conditioning (HVAC) for an office building, which significantly contributed to power bills' reduction while achieving comfortable set-points for HVAC devices. Similar to Rosewater et al. (2019), where MPC showed a better performance and was more robust than the state-of-the-art approach, achieving lower bills for energy customers and eliminating the optimistic shortfall. Additionally, MPC can handle uncertainty scenarios for the PV-ESS system, using forecast information for decision-making as in the research by Nair et al. (2021) which obtains a robust scheduled operation pattern for BESS. However, there were few comparison results between model-based, rules-based, and learning-based approaches for the BESS problem in terms of bill savings, scheduled patterns, and battery degradation.

The latest advances in deep reinforcement learning (DRL) agents in Wei et al. (2022), Ren et al. (2018), Mocanu et al. (2018), Nguyen et al. (2020), Zhang et al. (2019), Perera and Kamalaruban (2021), Cao et al. (2020) and Gao et al. (2022) have led to a revolutionary mutation in the system control problems. DRL agents have introduced

10.1016/j.ifacol.2023.10.546

many promising solutions for the BESS control problem, as illustrated with examples in the existing literature. Starting with the discrete action control in Yan et al. (2021), it can be performed by Deep-Q-learning network (DQN) agents to control the charging and discharging of batteries by setting discrete actions for each cluster of batteries. Further, in Bui et al. (2019), the DQN method was significantly improved to be more reliable for deciding BESS decision variables, including uncertainties. Additionally, this work has shown the potential of enhancing the Q-learning method for significant state space problems of BESS. In our work, we utilize the modified Q-learning algorithm for the optimal decision of discrete action control. In Xu et al. (2019), the proposed DRL algorithm was used to control the energy storage system (ESS) to arbitrage in real-time electricity markets under price uncertainty and learn the proper stochastic control policy for BESS control. Similar to Mocanu et al. (2018), the online energy scheduling DRL controller provides real-time feedback to consumers to encourage more efficient electricity use. DRL also monitored a highly dimensional database, including information about photovoltaic power generation, electric vehicles, and building appliances. In the DRL agent model, the optimization problem is solved in the form of the Markov Decision Process (MDP). The agent explores and exploits multiple scenarios during the training phase and then develops an optimal policy for the best decision during testing. Moreover, the batch of sampled data for BESS as in Mbuwir et al. (2018) can be introduced to the problem after defining the dimensions of the state and action spaces. DRL has also been extended to deal with optimal battery control under cycle-based degradation.

This research in the literature has inspired us to consider DRL as a good candidate for solving BESS optimization problems and locating global solutions instead of model-based approaches. In addition, the DRL can easily incorporate battery health and renewable uncertainty objectives through the defined reward function. Then, it can determine the globally optimal solution based on the updated policy discovered. Furthermore, the existing MILP methods define numerous constraints and develop relaxation methods to form a complex MILP to better guarantee global solutions instead of sub-optimal ones. DRL has advantages over these methods by only penalizing constraint violations without requiring the solver to construct numerous technical constraints, as in the MILP case.

This paper proposes an approach based on machine learning that regulates BESS and grid supplies for optimal energy management, thereby extending battery life and reducing electricity costs. The main contributions are summarized as follows:

- A DQN with a PER-based DRL approach is used for learning the safe and optimized capacity scheduling policy for the energy storage and grid operation in the context of performing the BESS optimal operation, which deals with dynamic tariff prices for the Australian electrical market, including uncertainties. This algorithm is found to have good reliability for guaranteeing the global optimal solution for the studied problem.
- Relaxed MILP is used as the benchmark method for deducing the optimal switching criteria for BESS units. The proposed learning-based approach is therefore guaranteed to be successful by using the MILP results as a reference for BESS operation.
- The proposed algorithm focuses on integrating BESS controllers and automatic transfer switches for the first time to govern grid and energy storage systems. This setup ensures seamless transitions between different power sources and enhances the reliability of the power supply.
- Using evaluation metrics (i.e., switching count, battery operation time) to validate the learning-based approach for BESS control actions and final electricity costs.

2. PROBLEM FORMULATION

The problem formulation is based on deducing a scheduled operation for BESS and grid supply to obtain a highly optimized performance for the energy dispatch and maintain the life cycle of BESS. The optimization problem is formulated as follows:

$$\min \sum_{t=0}^{T} \gamma_t^g \times P_t^g + \sum_{t=0}^{T} \gamma_t^b \times \left(P_t^b - P_{t-1}^b\right) + \sum_{t=0}^{T} \gamma_t^g \times \gamma_p \times T_t^g \tag{1}$$

subject to

$$\sum_{t=0}^{T} P_t^b + P_t^g = P_t^d + P_t^{unc.}, \tag{2}$$

$$\gamma_t^g \in [0,1], \gamma_t^b \in [0,1], \tag{3}$$

$$P_t^{b,min} < P_t^b < P_t^{b,max}, \tag{4}$$

$$P_t^{g,min} < P_t^g < P_t^{g,max}, \tag{5}$$

$$P_t^{gex,min} < P_t^{gex} < P_t^{gex,max}, \tag{6}$$

$$\gamma_t^g + \gamma_t^b = 1, \tag{7}$$

$$P_t^b + P_t^{pv} + P_t^g \geq P_t^d + P_t^{unc.}, \tag{8}$$

$$P_t^{gex} \leq P_t^{pv} - P_t^d, \tag{9}$$

$$P_t^{b,ch} \leq P_t^{pv} - P_t^d, \tag{10}$$

$$P_t^{b,disch} \leq P_t^d - P_t^{pv}, \tag{11}$$

$$P_t^g \leq P_t^d - P_t^{pv}, \tag{12}$$

$$P_t^{b,ch} + P_t^{b,disch} = P_t^b, \tag{13}$$

$$SOC_{t+1} = SOC_t + \Delta T \left(\alpha_{b,ch} P_t^{b,ch} + \frac{P_t^{b,disch}}{\alpha_{b,disch}}\right), \tag{14}$$

$$SOC_t^{min} < SOC_t < SOC_t^{max}. \tag{15}$$

The objective function shown in (1) aims to minimize the power imported from the grid and reduce the rate of power being dispatched from batteries to keep their remaining energy for the end of the day. To optimize these objectives, discrete decision variables of γ_t^g and γ_t^b are used to control the switching operation between batteries and the grid supply to find the optimal solution of the

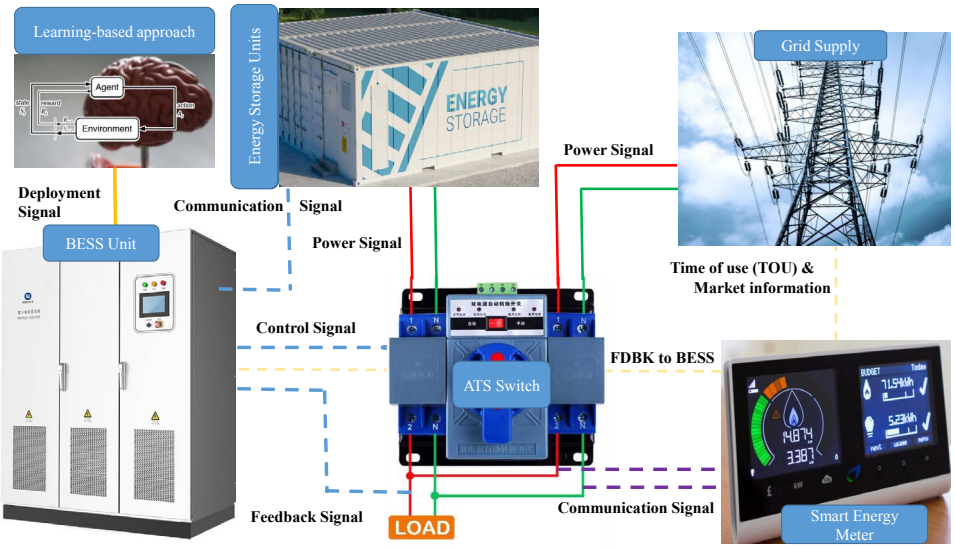

Figure 1. Conceptual model for smart BESS.

objective function while respecting all system constraints listed from (2) to (15). State variables for the studied horizon of the system are identified as follows: P_t^{pv} is the output solar PV measured through a smart inverter at each time step, P_t^b is the power dispatched by the energy storage system within storage limits of the maximum $P_t^{b,max}$ and the minimum $P_t^{b,min}$. P_t^g is the power supplied by the grid within the $P_t^{g,max}$ and $P_t^{g,min}$ limits, P_t^{gex} represents the exported power to the grid within the $P_t^{gex,max}$ and $P_t^{gex,min}$ limits. SOC_t is computed based on the P_t^b and needs to be within limits of SOC_t^{min} and SOC_t^{max}. T_t^g is the instantaneous electricity bill price with a weighted penalization coefficient of γ_p. P_t^d is the total power demand for the system, and $P_t^{unc.}$ is the power needed to balance uncertainties in generation and demand. The objective function represents the switching action pattern for the BESS that is determined by γ_t^g and γ_t^b in a way that only one power source is switched on based on (7) and this switching action is only feasible in cases, where no violations happen to the maximum and minimum boundaries of the energy system. Additionally, this control solution respects the charging and discharging rates in $P_t^{b,ch}$ and $P_t^{b,disch}$ while assuming that charging and discharging efficiency ($\alpha_{b,ch}$, $\alpha_{b,disch}$) are equal to 90 percent. Lastly, for $P_t^{g,ex}$ and $P_t^{b,ch}$ in (9) and (10), it depends if there is a surplus power and the optimal solution is to divide this power between charging batteries and grid export and achieve optimal power sharing. The same idea in (11) and (12), for supplying load with power beside solar PVs in times of lack of generation, we can have a sharing power percentage between power supplied by the grid and dispatched from batteries with respect to the system limits.

To solve this objective function, system parameters are modelled and computed as follows:

- **Energy Storage Charging Computation**: State of charge of battery is computed based on the battery capacity and as shown in the formulation below. These equations are computed at each time step to show the overall status of each battery used in the system. The state of charge is computed based on the amperes dispatched, which is calculated from the energy dispatched for steady state operation.

$$\Delta \text{SOC} = \frac{C}{C_n} = \frac{\Delta Ah}{C_n} = \frac{-\int_{t_0}^{t_0+\tau} I_\text{b}(t)dt}{C_n} 100\% \tag{16}$$

where C is the current battery capacity due to changing and discharging patterns and C_n is the capacity of the battery when it is freshly used. ΔAh is the ampere-hour dispatched at each time step and I_b is the charging and discharging current for each time step.

- **System Modeling**: The proposed control algorithm is built based on the conceptual model depicted in Fig. 1. The proposed controller gets feedback from each battery via a feedback communication signal in the integrated smart unit of the battery portal. The instantaneous load demand is obtained by BESS in parallel using the feedback power signal that is transmitted straight to the load. Smart meters constantly send a timestamped communication signal to the BESS to provide an update within 200 milliseconds and monitor grid market prices and peak times. As a result, the agent determines the best combination of switching maneuvers between batteries and grid supply. The controller takes discrete steps through the automatic transfer switch (ATS).

3. DEEP REINFORCEMENT LEARNING-BASED CONTROL OF BESS

The energy management problem is reformulated as a reward function for the DRL. The system experiences different conditions at each time step, represented by a state vector. Actions are taken at each time step based on the updated state from the last time step and concerning the boundaries of the system input actions. Consequently, a reward function is formulated to be fully controlled by the set of actions and states at each time step. The key elements for this controller are shown below.

State Space S_t: the state S_t is used to represent the system status at each time step and is defined as follows:

$$S_t = [P_t^b, P_t^g, SOC_t^k, T_t^g, P_t^d] \tag{17}$$

where SOC is the state of charge of k^{th} battery at time t; the rest of the other state variables are defined in the original problem formulation.

Actions Space a_t: the set of actions consists of these operations that the controller can execute at each time step. Also, it determines the discrete switching set-points of the DERs based on the available power that can be utilized within energy storage limits. Formally, we have

$$a_t = [\gamma_t^g, \gamma_t^b]. \tag{18}$$

These control actions are defined in the problem formulation section. They show the feasible steps that can be applied to the studied horizon and significantly improve the optimization problem.

Reward Function: The reward in this problem is based on the main objective function defined in (1) that aims to maximize the dependency on the batteries through BESS real-time operation while penalizing the higher electricity bill payments and minimizing dependency on the grid network in peak demand times. These objectives are weighted in the reward formulation for their impacts on the system performance during the training and testing. Our target of implementing the DRL agent is to optimize the reward function as defined below,

$$\max \sum_{t=0}^{T} r\left(s_t, a_t\right) \tag{19}$$

$$\sum_{t=0}^{T} r = \eta\gamma_t^g + \omega\gamma_t^b - \psi T_t^g + \delta\ P_t^b - \zeta\ P_t^g \tag{20}$$

where η, ω, ψ, δ, and ζ are the reward coefficients that determine the rewarding and penalization scale. In this problem, we fix these coefficients to be 1000 for η and ω, 1 for ψ, and 10 for δ and ζ. This reward function is formulated to make the trained DRL agent minimizes electricity bills in T_t^g and favors the power supplied by P_t^b than P_t^g. Additionally, this formulation gives equal weights for the switching actions of γ_t^g and γ_t^b to make control actions not biased towards switching either battery or grid side. The choice of reward function's coefficients considers their impact on the DRL performance for achieving higher reward value with faster convergence Katahira (2015). The idea is to keep the same impact for each term on the reward function and not prioritize one over the other. Additionally, these coefficients are found to be fitting the problem while carrying out different training scenarios.

For the DQN to handle the BESS problem, the policy function should be identified by the DQN to find the best reward value by adjusting its hyper-parameters as shown in Table 1 and Katahira (2015). The main target is determining the best policy π for forming the action as $a_t \sim \pi\left(s_t\right)$ with s_t given in (16). To simplify the policy search, we are particularly interested in the parameterized policies offered by $\pi_\theta(\cdot) = \pi(\cdot;\theta)$, with parameter θ optimized. The set of actions, states, and rewards are kept in Q-value format for the self-update of the agent's Q-table

Table 1. Parameter settings for DQN training

Parameters	Value
Replay buffer size	100000
Batch size	64
discount factor (Γ)	0.99
Learning rate	0.0005
Coefficient of target network's soft update	0.001
Number of hidden layers	2
Number of nodes	[64,64]
Activation function	ReLU
Maximum number of episodes	15000

and stored in the replay buffer for better performance of the DQN. The following equations

$$\max_\theta \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \Gamma^t r_t\left(s_t, a_t\right)\right] \tag{21}$$

$$Q\left(s_t, a_t\right) := \mathbb{E}_{\pi_\theta}\left[\sum_{\tau=t}^{T} \Gamma^{(\tau-t)} r_\tau\left(s_\tau, a_\tau\right) \mid s_t, a_t\right] \tag{22}$$

$$Q^*\left(s_t, a_t\right) = r_t + \Gamma\mathbb{E}_{s_{t+1}}\left[\max_{a_{t+1}} Q^*\left(s_{t+1}, a_{t+1}\right)\right] \tag{23}$$

show the targeted Q-value $Q(s_t, a_t)$ for the optimal policy and how it can be achieved for the given actions and states. The parameters in (22) represent the transition states of MDP, where the agent memorizes it during the training phase to come up with the final weights of the neural network of the DQN using mini-batches obtained from Q-table. Here comes the role of PER, which plays a significant role in improving the DQN efficiency and stability as in Zeng et al. (2022). Its idea is to use the previous transition states from the memory when computing Q-values and loss functions of the neural network, resulting in significant improvement in the training phase with higher reward values and faster convergence of the learning curve, as shown in Fig. 2, even when compared with Double DQN.

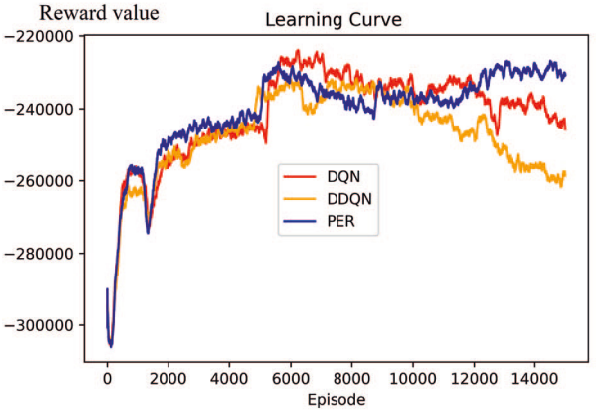

Figure 2. Learning curves of the DQN, DDQN and DQN-PER

Furthermore, PER property is also aimed to minimize the difference between evaluated Q value $Q\left(s_t, a_t \mid \theta\right)$ and target Q value $Q^*\left(s_t, a_t\right)$, which is represented by temporal-difference (TD) error $e_{\mathrm{TD}}\left(\theta\right)$ to improve the updated Q-table results;

$$e_{TD}(\theta) = Q\left(s_t, a_t \mid \theta^i\right) - Q^*\left(s_t, a_t\right). \quad (24)$$

Before deploying the DQN agent, the OpenAI Gym environment has been developed and registered with the name (Batterycontrol-v0), which defines actions, states, and if-conditions for Algorithm 1. Additionally, this environment defines the system constraints and boundaries we mentioned in Section 2. The reward function tends to reach the maximum value after almost 4000 episodes. Before convergence, it experiences the exploration and exploitation phases (i.e., the agent tries possible actions that lead to the highest reward values.) until it finds the optimal policy for this problem.

Algorithm 1 BESS Control Algorithm

Input: SOC_t^K , P_t^d , T_t^g
Output: P_t^b, P_t^g, γ_t^g, γ_t^b
Data: Real-time data-set of solar PVs production, residential load demand and tariff prices on 5 minutes basis; Setting maximum tariff price $T_t^{g,max}$ to 0.5 (AUD/ kWhr)
for t $\epsilon[0,T]$ do **Observe** P_t^d , P_t^{pv}, $P_t^{unc.}$ to compute P_t^{diff} of energy balance (i.e., positive if demand exceeds generation)
Calculate SOC_t^K, P_t^{diff}, T_t^g
Perform BESS discrete Switching
if $P_t^{b,min} < P_t^{diff} < P_t^{b,max}$ and $T_t^g > T_t^{g,max}$ **then**
  Activate **battery mode** by setting γ_t^b to 1
  Allocate P_t^b **set-point**
end
if $T_t^g < T_t^{g,max}$ and $SOC_t^K < 0.2$ **then**
  Activate **grid mode** (γ_t^g to 1)
  Allocate P_t^g **set-point**
end
else
  Activate **optimal charging mode**
  Utilize surplus P_t^{diff} for charging batteries
  if $SOC_t^K > 0.8$, **then** export surplus P_t^{diff} to the grid.
end

4. NUMERICAL RESULTS

This section provides the simulation results for codes developed using the PyTorch library on a laptop computer with a 3.6GHZ Intel i7 processor and 32.0 GB RAM. The proposed control methods are used to demonstrate the idea of allocating the power-sharing percentage by switching actions conducted for the battery system and the grid for the actual input model in Fig. 3. For verification and validation; the test is carried out for a 24-hour operation of Solar PV, grid, and batteries. MILP optimization is applied for minimizing the objective function in (1) to deduce the scheduled pattern of battery and grid operation, as shown in Fig. 4, Fig. 5, and Fig. 6. This optimization is subject to the constraints mentioned in Section 2. Batteries are controlled to dispatch their available power at initial time steps, and then the grid is in service for the remaining night hours due to the lower tariff price. At times of solar energy production, the surplus is utilized in both approaches to charge the batteries and export the surplus power to the grid. Thus, when peak demand exceeds the available solar generation, batteries are dispatched across the loads, and the grid remains disconnected (i.e., minimizing electricity bills as much as the BESS can). Lastly, when batteries start reaching their lower boundaries of SOC, the grid is controlled in steps to supply the remaining load until the remaining energy of the battery vanishes. The critical point in the control optimization algorithm is to achieve the balance for the energy storage that keeps batteries for later use when tariff prices reach higher values. The problem with the MILP algorithm is that the planning of battery operation could have been more efficiently utilized as the battery's remaining energy vanished at a peak time of electricity use, which is our major challenge to handle (time of higher tariff price avoidance). The battery operation should be efficiently managed to plan for the last hours of the day and how to reduce imported grid electricity as much as we can. On the other side, DRL introduces a new pattern of switching actions between batteries and the grid and achieves the trade-off operation between available power sources, as shown in Fig. 7 and Fig. 8, to reduce the electricity bills by almost 51 percent compared to the non-controlled approaches and by 33 percent compared to the linear optimization approach. The algorithm shows smart decisive actions that keep batteries' energy reserve until night at 8:00 p.m. (highest tariff price), when the grid supply is switched off, and battery storage covers that time. These observations have led to the comparative analysis of each proposed system, highlighting some critical metrics shown in Table 2. As for time complexity, MILP usually solves the problem in the range of 0.1 seconds; however, this timing will exponentially increase for higher data dimensions. In contrast, the DRL testing time is much lower than the MILP, while the offline training takes almost 3 hours. Thus, the DRL can be utilized for online testing and can implement requested actions in milliseconds, making it a good fit for real-time applications.

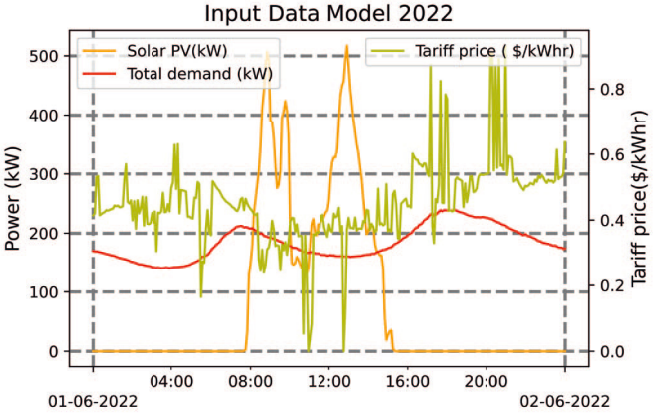

Figure 3. Input model for the BESS system

CONCLUSION

In order to introduce an ideal schedule for the system's operation that significantly decreases electricity costs, this paper proposed a novel energy management algorithm for feasibly managing batteries and grid supply. A complex mixed-integer program was produced by formulating the

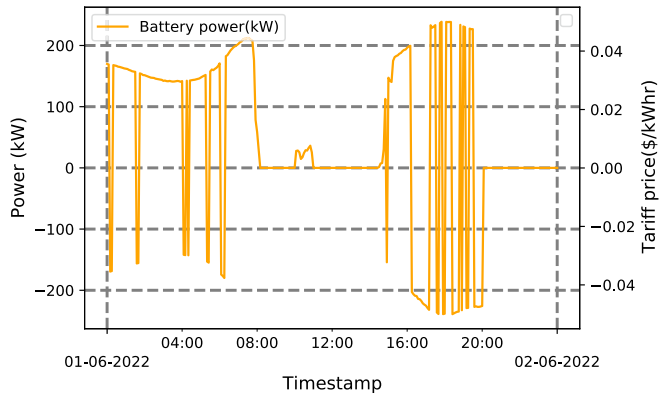

Figure 4. MILP optimization results-battery settings

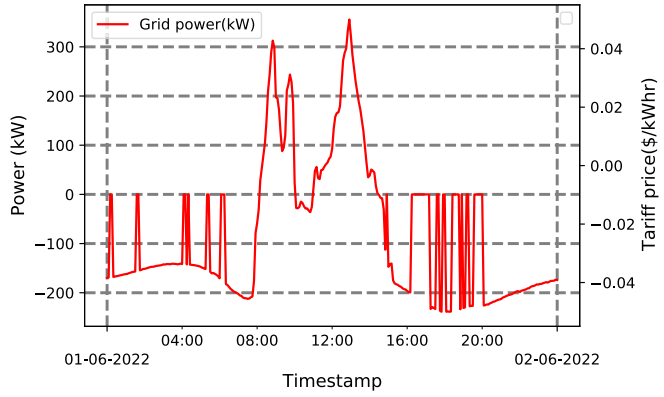

Figure 5. MILP optimization results-grid settings

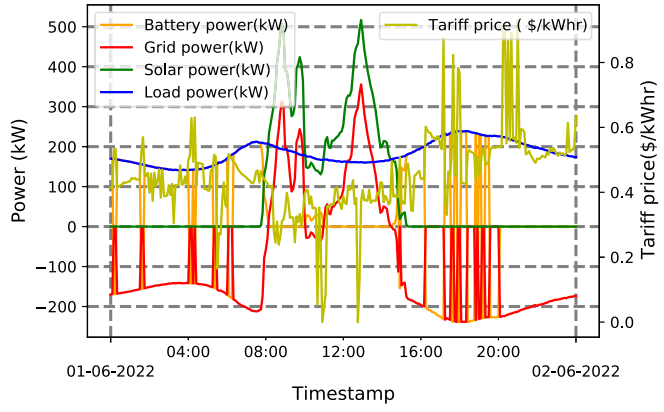

Figure 6. MILP optimization results-system overview

Table 2. Comparative analysis based on proposed key metrics

Metrics	No Control	MILP	DRL
Electricity Price ($)	16205.14	11372.25	7850.7
Switching actions no.	-	43	158
Battery in-service (min)	-	215	790
Operation violations	-	0	0

energy management system problem with the control of batteries and grid supply in mind. The suggested method outperformed linear optimization for lowering electricity costs and delivered the best decisive actions while in charging/discharging mode. It also introduces a

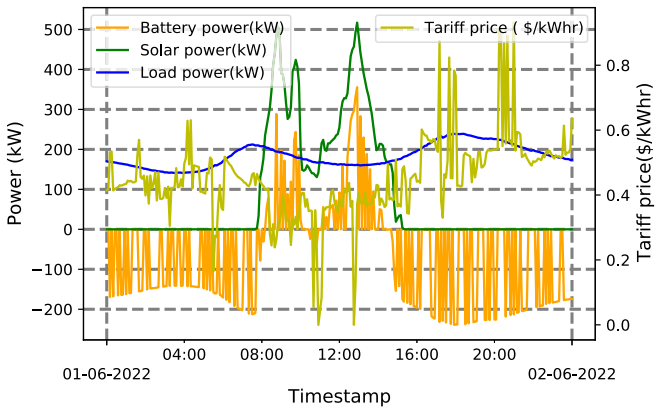

Figure 7. DRL optimization results-battery side

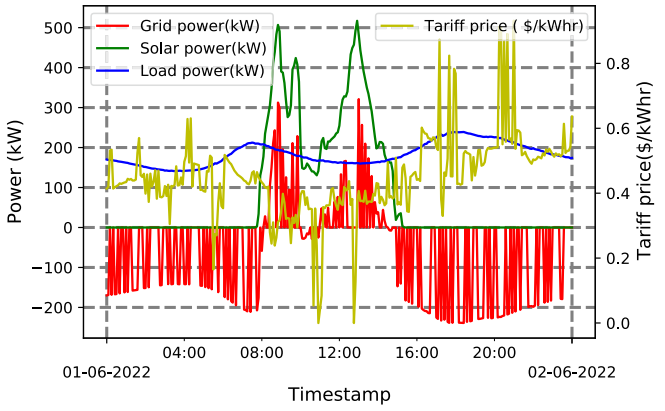

Figure 8. DRL optimization results-grid side

new framework for an ATS-controlled scheduled operation between the grid and battery. In subsequent work, we will train the DRL agent to allocate continuous operation set-points for the energy storage and the grid supply and implement this algorithm on real-time simulators that interface with real batteries for a better study of the electrochemical behavior of energy storage systems.

REFERENCES

Aryai, V. and Goldsworthy, M. (2022). Controlling electricity storage to balance electricity costs and greenhouse gas emissions in buildings. *Energy Informatics*, 5(1), 1–23.

Bui, V.H., Hussain, A., and Kim, H.M. (2019). Double deep q-learning-based distributed operation of battery energy storage system considering uncertainties. *IEEE Trans. Smart Grid*, 11(1), 457–469.

Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., and Blaabjerg, F. (2020). Reinforcement learning and its applications in modern power and energy systems: A review. *Journal of modern power systems and clean energy*, 8(6), 1029–1042.

Datta, U., Kalam, A., and Shi, J. (2018). Battery energy storage system to stabilize transient voltage and frequency and enhance power export capability. *IEEE Trans. Power Systems*, 34(3), 1845–1857.

Gao, Y., Matsunami, Y., Miyata, S., and Akashi, Y. (2022). Operational optimization for off-grid renewable

building energy system using deep reinforcement learning. *Applied Energy*, 325, 119783.

Jeddi, B., Mishra, Y., and Ledwich, G. (2019). Differential dynamic programming based home energy management scheduler. *IEEE Trans. Sustainable Energy*, 11(3), 1427–1437.

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology*, 66, 59–69.

Kong, W., Luo, F., Jia, Y., Dong, Z.Y., and Liu, J. (2021). Benefits of home energy storage utilization: An australian case study of demand charge practices in residential sector. *IEEE Trans. Smart Grid*, 12(4), 3086–3096.

Mbuwir, B.V., Kaffash, M., and Deconinck, G. (2018). Battery scheduling in a residential multi-carrier energy system using reinforcement learning. In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 1–6. IEEE.

Mocanu, E., Mocanu, D.C., Nguyen, P.H., Liotta, A., Webber, M.E., Gibescu, M., and Slootweg, J.G. (2018). On-line building energy optimization using deep reinforcement learning. *IEEE Trans. Smart Grid*, 10(4), 3698–3708.

Nair, U.R., Sandelic, M., Sangwongwanich, A., Dragičević, T., Costa-Castelló, R., and Blaabjerg, F. (2021). An analysis of multi objective energy scheduling in pv-bess system under prediction uncertainty. *IEEE Trans. Energy Conversion*, 36(3), 2276–2286.

Nguyen, T.T., Nguyen, N.D., and Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 50(9), 3826–3839.

Perera, A. and Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137, 110618.

Ren, Z., Dong, D., Li, H., and Chen, C. (2018). Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6), 2216–2226.

Rosewater, D., Baldick, R., and Santoso, S. (2019). Risk-averse model predictive control design for battery energy storage systems. *IEEE Trans. Smart Grid*, 11(3), 2014–2022.

Vedullapalli, D.T., Hadidi, R., and Schroeder, B. (2019). Combined hvac and battery scheduling for demand response in a building. *IEEE Trans. Industry Applications*, 55(6), 7008–7014.

Wei, Q., Ma, H., Chen, C., and Dong, D. (2022). Deep reinforcement learning with quantum-inspired experience replay. *IEEE Transactions on Cybernetics*, 52(9), 9326–9338.

Xu, H., Li, X., Zhang, X., and Zhang, J. (2019). Arbitrage of energy storage in electricity markets with deep reinforcement learning. *arXiv preprint arXiv:1904.12232*.

Yan, L., Liu, W., Jiang, W., Li, Y., Li, R., and Hu, S. (2021). Deep reinforcement learning based optimization of battery charging and discharging management for data center. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.

Zeng, L., Yao, W., Shuai, H., Zhou, Y., Ai, X., and Wen, J. (2022). Resilience assessment for power systems under sequential attacks using double DQN with improved prioritized experience replay. *IEEE Systems Journal*.

Zhang, Z., Zhang, D., and Qiu, R.C. (2019). Deep reinforcement learning for power system applications: An overview. *CSEE Journal of Power and Energy Systems*, 6(1), 213–225.