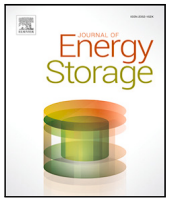
Journal of Energy Storage 73 (2023) 109144

Contents lists available at ScienceDirect

Journal of Energy Storage

journal homepage: www.elsevier.com/locate/est



Research papers

Optimize the operating range for improving the cycle life of battery energy storage systems under uncertainty by managing the depth of discharge

Seon Hyeog Kim [a], Yong-June Shin [b,*]

[a] Digital Convergence Research Laboratory, Electronics and Telecommunications Research Institute, 218, Gajeong-ro, Daejeon, 34129, Yuseong-gu, Republic of Korea

[b] The School of Electrical and Electronic Engineering, Yonsei University, 50, Yonsei-ro, Seoul, 03722, Seodaemun-gu, Republic of Korea

ARTICLE INFO

Keywords:
Battery aging
Battery energy storage system (BESS)
Battery management
Depth of discharge (DOD)
Deep reinforcement learning
Time-of-use

ABSTRACT

Globally, renewable energy penetration is being actively promoted by renewable energy 100% (RE100) policies. BESS operators using time-of-use pricing in the electrical grid need to operate the BESS effectively to maximize revenue while responding to demand fluctuations. Battery energy storage (BESS) is needed to overcome supply and demand uncertainties in the electrical grid due to increased renewable energy resources. BESS operators using time-of-use pricing in the electrical grid need to operate the BESS effectively to maximize revenue while responding to demand fluctuations. However, excessive discharge depth and frequent changes in operating conditions can accelerate battery aging. Deep discharge depth increases BESS energy consumption, which can ensure immediate revenue, but accelerates battery aging and increases battery aging costs. The proposed BESS management system considers time-of-use tariffs, supply deviations, and demand variability to minimize the total cost while preventing battery aging. In this study, we investigated a BESS management strategy based on deep reinforcement learning that considers depth of discharge and state of charge range while reducing the total operating cost. In the proposed BESS management system, the agent takes actions to minimize the total operating cost while avoiding excessive discharge depth and low state of charge. A series of experiments using a real BESS demonstrated that the proposed BESS management system has improved performance compared to the existing methods.

1. Introduction

Renewable energy deployed to achieve carbon neutrality relies on battery energy storage systems to address the instability of electricity supply. BESS can provide a variety of solutions, including load shifting, power quality maintenance, energy arbitrage, and grid stabilization [1]. Previous research has proposed an energy management system (EMS) operation strategy that integrates BESS, PV, and vehicle-to-grid functions to maximize the benefits of BESS [2,3]. Mixed-integer linear programming was implemented to solve various grid scenarios to reduce operating costs and peak hour consumption [4,5]. Model predictive control (MPC) is a modern optimal control strategy that can efficiently handle non-linearity and operational constraints. MPC can provide improved performance and is well suited to EMS problems. In [6,7], MPC was used to maximize the economic benefits of BESS and minimize the BESS performance degradation under different system constraints. However, MPC performance can be affected by load/PV uncertainties [8].

Existing energy management studies using BESSs have focused on reducing electricity costs in time-of-use (TOU) tariffs, while the aging conditions of the BESS has not been seriously considered. In [9], the state-of-charge (SOC) range affected battery aging. A scheduling algorithm considering battery degradation was proposed in [10–13]. Excessive depth of discharge (DOD) can ensure immediate revenue, but BESSs typically do not cycle beyond their maximum rate capacity. Increasing DOD due to excessive charge/discharge for economic gain increases the risk of BESS fire and accelerates battery aging. In [14,15], the state of health (SOH) and end of life (EOL) of a battery is highly dependent on depth of discharge (DOD) conditions. Lithium-ion batteries are typically designed to last longer when charged to a moderate SOC range, such as 20%–80%. In additions, deep discharging can cause internal stress on the battery, which can lead to other issues such as reduced charging capacity and decreased overall performance. The capacity degradation of a battery is accelerated by repeated deep discharges and recharges at high SOC [16].

The aforementioned studies have demonstrated improvements in charge and discharge scheduling, but they are model-based approaches

* Corresponding author.
E-mail addresses: seonh@etri.re.kr (S.H. Kim), yongjune@yonsei.ac.kr (Y.-J. Shin).

https://doi.org/10.1016/j.est.2023.109144
Received 4 May 2023; Received in revised form 17 September 2023; Accepted 29 September 2023
Available online 16 October 2023

that rely heavily on information from system models. To ensure optimal operation even in complex environments, BESS management methods based on reinforcement learning (RL) have been proposed. Model-free approaches that do not require system model information have achieved great success in decision-making applications using RL [17]. Deep neural networks (DNNs) have also overcome the problem caused by the small state space of Q-learning. In [18], deep reinforcement learning (DRL) combining RL and DNNs provided an effective EMS without specific user information. In [19], a method to optimize scheduling in DR based on deep Q-learning (DQN) combining Q-learning and DNNs was proposed. To overcome the high-dimensional DQN problem and avoid being trapped in local optimization, double-deep Q-learning (DDQN) was proposed. The authors in [20] implemented DDQN to learn optimal battery control policies considering price uncertainty and battery degradation. To solve the DQN discrete action problem, deterministic policy gradient (DPG) has been proposed. However, DPG suffers from low sampling efficiency and slow convergence due to large variance of gradient estimation. To overcome these drawbacks, a deep deterministic policy gradient (DDPG) method was proposed. The authors in [13] showed simulation results to derive an optimal BESS control strategy based on DDPG. Recently, soft actor criticism (SAC), a state-of-the-art DRL strategy that accelerates convergence and improves optimization performance, has been used to intelligently optimize EMS. The authors in [21–23] developed a method based on SAC that outperforms other DRL methods in optimizing EMS in complex environments.

Existing DRL-based BESS scheduling methods have demonstrated improved performance through simulation verification. However, they have not simultaneously considered the DOD conditions of the BESS and the degradation cost due to the uncertainty of load/generation. In addition, performance analysis based on actual battery test results has not been addressed. Based on this literature review, this paper proposes a state-of-the-art DRL-based BESS scheduling that can learn optimized control to reduce grid operating costs, including the degradation cost, in a complex environment. We compare the BESS scheduling method using DRL with real battery DOD tests in similar environments to analyze the impact on battery life and operating costs.

The remainder of this paper is organized as follows. Section 2 describes a grid environment model and the battery aging model. The BESS management procedure based on DRL is introduced in Section 3. In Section 4, we apply the proposed methods to various grid scenarios based on real-world grid datasets and actual battery tests, and Section 5 summarizes and concludes the paper.

2. Environment model

2.1. Grid and time-of-use

An electrical grid consists of a primary energy resource, the electricity grid, and renewable resources such as demand-driven loads, BESS, and PV. BESSs are installed to reduce the cost of electricity through arbitrage and to balance the energy imbalance caused by the uncertainty and irregular demand of solar power generation. The power balance constraint that must be satisfied at all times can be formulated as follows:

$$\underbrace{P_t^{DE}}_{\text{Demand}} = \underbrace{P_t^{G}}_{\text{Utility grid}} + \underbrace{\eta_e P_t^{BESS}}_{\text{BESS}} + \underbrace{\eta_p P_t^{PV}}_{\text{PV}}, \tag{1}$$

where P_t^{DE} is the grid demand, P_t^{G} is the electrical power from the utility grid, P_t^{BESS} is the BESS charging/discharging power, η_e is the efficiency of the BESS inverter (η_e depends on the charge/discharge operating conditions), P_t^{PV} is the PV output power, and η_p is the efficiency of the PV inverter

Table 1
Time-of-use tariff.

	Time	Electricity price [cent/kWh]
TOU Tariff	Off-Peak	2.35
	Mid-Peak	4.30
	On-Peak	32.6

2.2. Battery energy storage system

BESS scheduling is optimized by considering demand/supply forecasts, TOU and SOC. The inequality constraints include the utility grid's power capacity limits and EMS, as follows:

$$P_{min}^{G} \le P_t^{G} \le P_{max}^{G}, \tag{2}$$

$$P_{min}^{BESS} \le P_t^{BESS} \le P_{max}^{BESS}, \tag{3}$$

Constraint (2) is the utility grid constraint and constraint (3) is the charging/discharging rate limit in BESS scheduling imposed by the EMS. To prevent the battery from being over-charged or over-discharged, the BESS SOC limit is defined as follows:

$$SOC_{min} \le SOC_t \le SOC_{max}, \tag{4}$$

The defined SOC estimation method estimates residual capacity by calculating the BESS charge/discharge power per hour based on a energy capacity. The SOC is utilized to estimate the available energy of the BESS. The SOC of the BESS can be calculated as follows:

$$SOC_{t+1} = \begin{cases} SOC_t + \eta_c \frac{|P_t^{BESS}| \cdot \Delta t}{E_B}, & P_t^{BESS} > 0, \\ SOC_t - \eta_d \frac{|P_t^{BESS}| \cdot \Delta t}{E_B}, & P_t^{BESS} < 0, \end{cases} \tag{5}$$

where η_c represents the charging conversion efficiency and η_d represents the discharging conversion efficiency. E_B is the energy capacity of the BESS, which gradually decreases as the battery ages, so updating information about E_B can improve the accuracy of SOC estimation.

The EMS plays an essential role in optimal operational scheduling using BESSs, as it considers the grid states and TOU. The TOU pricing provides consumers with opportunities to manage their electricity cost by shifting use from on-peak periods to off-peak periods. The TOU C_t^{TOU} is presented in Table 1 [24]. The electricity price during off-peak hours is 2.35 cent/kWh, whereas that during on-peak hours is 32.6 cent/kWh. This TOU pricing can save electricity costs for on-peak loads by utilizing BESS at off-peak to charge energy at a lower cost. Thus, the operating cost C_t^{o} is determined by the utility grid as well as the BESS charging/discharging schedule, and can be defined as follows:

$$C_t^{o} = (P_t^{G} + P_t^{BESS}) \times C_t^{TOU}, \quad s.t. \quad (1) - (4) \tag{6}$$

The objective of the proposed EMS is to optimize the BESS scheduling over a finite period such that the grid operates economically while reducing the aging costs during demand and supply uncertainties. This objective function applies to similar electrical systems; these include EVs such as BESSs.

2.3. Battery aging model

The limited BESS lifespan is a critical factor in grid long term operation planning. Frequent charging/discharging will reduce the BESS lifespan. In general, it is not recommended to discharge a battery entirely, as this dramatically shortens its life. In other words, there is a trade-off between the electricity and BESS aging costs in BESS management. Increasing the BESS running time and cycling can reduce the electrical costs but accelerate aging, which results in higher replacement costs. Without careful management, cyclical use causes

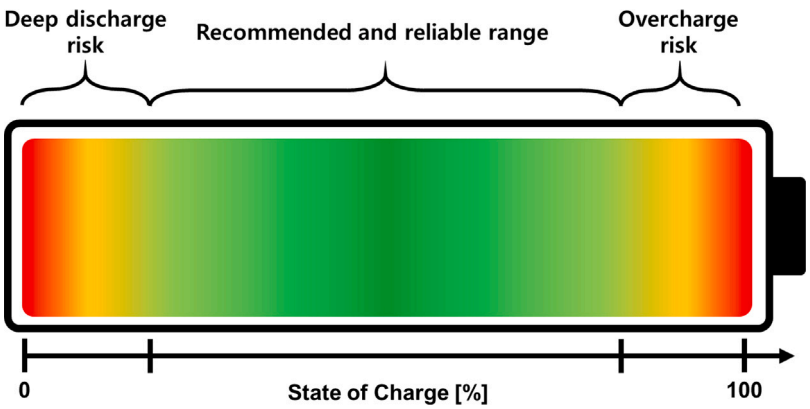


Fig. 1. Battery lifespan impact of SOC operating range.

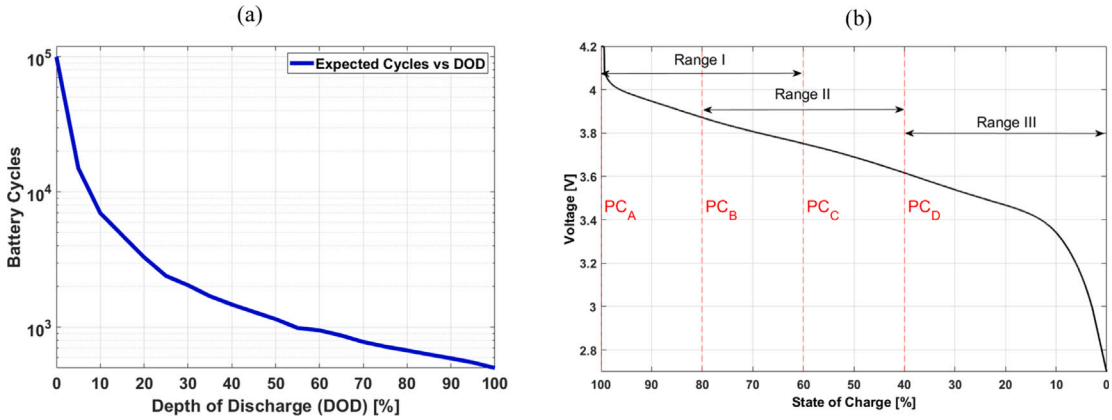


Fig. 2. (a) Cycle life depending on DOD. (b) Partial cycling of the BESS.

the BESS to age rapidly, which results in BESS system replacement costs [25]. In [26], an EMS that considers the ESS battery degradation cost was proposed. However, the aging indices used in previous studies did not facilitate the evaluation of the cyclic aging of daily scheduling. Therefore, a proactive BESS management system is required to optimize economic operation while minimizing aging factors; such a system is described below.

2.3.1. Depth of Discharge (DOD)

A battery's lifetime is highly dependent on the DOD. The DOD indicates the percentage of the battery that has been discharged relative to the battery's overall capacity. Deep discharge reduces the battery's cycle life, as shown in Fig. 1. Also, overcharging can cause unstable conditions. To increase battery cycle life, battery manufacturers recommend operating in the reliable SOC range and charging frequently as battery capacity decreases, rather than charging from a fully discharged SOC or maintaining a high SOC. Therefore, as suggested in this paper, deep discharge should be avoided by utilizing BESS scheduling that considers the DOD. Fig. 2(a) illustrates the relationship between the DOD and the cycle life; the wider the DOD range, the shorter the battery's life cycle. The DOD is calculated as follows:

$$D_k = \max(SOC_t) - \min(SOC_t) \tag{7}$$

where D_k denotes the DOD at the kth cycle and t is the time stamp.

2.3.2. Operating range of BESS

The impact of aging varies depending on the SOC ranges where the battery operation is concentrated, which can be evaluated using a partial cycling (PC) [9]. The PC reflects the BESS degradation conditions based on SOC range. The SOC range X is divided into four ranges: A (100%–80%), B (79%–60%), C (59%–40%), and D (39%–0%), as shown in Fig. 2(b). The power output during the time the battery spends in SOC range X is written as [9]:

$$\rho_x = \frac{\int_{t_0}^{T} SOC_x dt}{\int_{t_0}^{T} SOC_{Total} dt} \times 100 \quad [\%], \tag{8}$$

In Eq. (8), the numerator is the cumulative power output during the time the battery spends in each SOC range x. Weighting functions are then used to calculate the PC value as follows:

$$\rho = (a \times \rho_A + b \times \rho_B + c \times \rho_C + d \times \rho_D), \tag{9}$$

where a, b, c, and d are the linear weighting factors that are determined by the BESS scheduling conditions. Based on the battery manufacturer's data sheets and research [14–16], they recommend operating the BESS in the 20%–80% range. A charging at high SOC range accelerates battery aging as a result of problems such as corrosion and electrolyte stratification [9]. As shown in Fig. 2(b), cycling in a low SOC range (range D, ρ_D) causes more damage to the battery than cycling in other SOC ranges, so d has the highest weight for capacity degradation. In both high and low SOCs (ρ_A), excessive charging can increase DOD and deteriorate the cycle life [27]. Therefore, a is set to be higher than b and c.

To account for immediate rewards in the learning process, a degradation coefficient $c_{d,k}$ is proposed to estimate the reward for every charging or discharging control action. The degradation coefficient can be defined as follows:

$$c_{d,k} = \frac{\rho_k}{D_k}, \tag{10}$$

where $c_{d,k}$ is updated at each episode k based on the last training operation. The degradation level varies depending on the PC value even

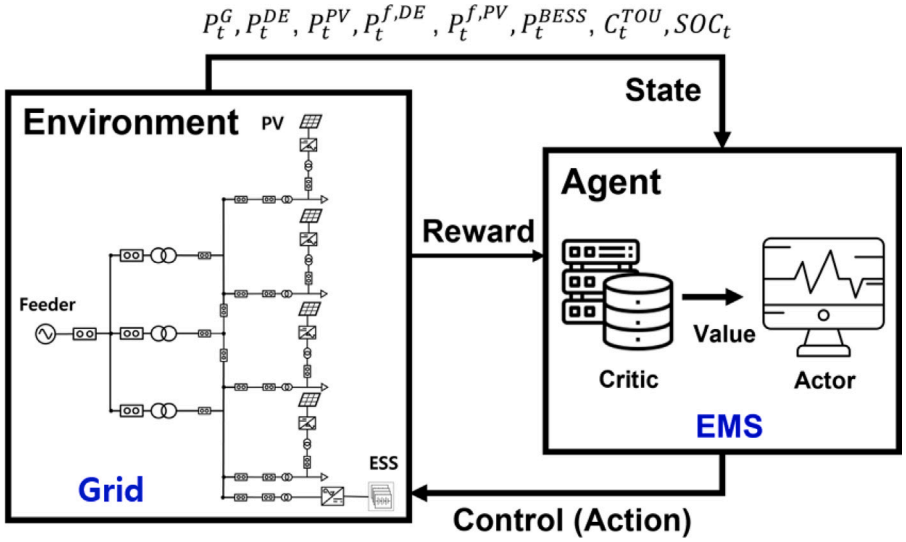


Fig. 3. Energy management system framework based on DRL in grid.

if the DOD is the same. As shown in the example in Fig. 2(b), the DOD of Range I, Range II, and Range III are the same but have different PC values and have different effects on battery aging.

2.3.3. State of Health (SOH)

SOH is a principal parameter that evaluates a battery's lifespan. With the gradual loss of available capacity during aging, the SOH is characterized by the ratio of the battery's remaining available capacity to its initial available capacity, which can be expressed as:

$$SOH(k) = \frac{E_k}{E_0} \times 100[\%], \tag{11}$$

where E_k represents the remaining available capacity at k cycles and E_0 is the initial BESS capacity. In this study, SOH is measured through a complete discharge test to measure the exact capacity degradation of the BESS. However, since such a complete discharge test adversely affects the performance and aging of the battery, SOH is measured every 50 cycles.

3. BESS management using DRL

The BESS-integrated grid considered in this study is installed in a set of buildings located in Seoul, Korea. Fig. 3 shows a schematic diagram of the grid with the BESS and DRL-based EMS system. The environment generates an observation vector s_t from the grid and the battery aging models. The EMS constitutes agents that gradually learn control strategies by leveraging the experience of repetitive interactions with the environment.

In this paper, the agent can observe the uncertainty encapsulated in the data and use a long short-term memory (LSTM) network and DRL techniques to learn the state transitions for the features in the actual data set. The grid state values, forecasting data and TOU are taken directly from the dataset indexed at $t+1$. In contrast, the BESS state values are determined by the control actions taken at time step t. The left part of the workflow (Fig. 4) forecasts the demand and PV using variational mode decomposition (VMD) and LSTM network. LSTM can extract features from the historical data and prevent the vanishing gradient problem. The forecasting method was proposed in our previous study [28]. Using VMD, the demand/PV profile is decomposed into a weekly demand profile and then decomposed into intrinsic mode functions that capture periodic features. Then, LSTM model is trained using intrinsic mode functions from the historical profile. The demand/PV is predicted by integrating the results of analyzing its periodic features. Then, the demand and PV are predicted, concatenated with other states, and fed into the DRL to learn the optimal policy. The deviation between the demand/supply forecasting data and the actual demand can be written as:

$$\Delta P_t = \underbrace{(P_t^{f,PV} - P_t^{PV})}_{\text{PV deviation}} - \underbrace{(P_t^{f,DE} - P_t^{DE})}_{\text{Demand deviation}}, \tag{12}$$

where $P_t^{f,PV}$ is the PV forecast, $P_t^{f,DE}$ is the demand forecast. The deviation of the supply/demand is balanced by the BESS. If the actual PV supply is insufficient due to overestimation, additional BESS discharge is required. If the actual PV supply exceeds the predicted PV supply, BESS charging is instead required.

Actor–critic models are optimized using offline DRL based on the SAC algorithm, as shown in Fig. 4. A critic is trained offline to estimate the demand, PV generation and operating cost. Based on the critic network, an actor is developed to optimize the BESS scheduling, which is updated by the SAC. As the DRL process progresses, the EMS continues to improve the performance and minimize the operational costs. After the offline DRL, the BESS model can directly observe the state using the grid model as well as the BESS degradation model and output a control action to minimize the expected total operating cost. In online applications, the BESS profile generated from the data-driven model based on real-world grid datasets is implemented using an actual battery under similar conditions to observe the battery states according to the charging/discharging pattern, DOD, and PC.

As shown in Fig. 5, the entire BESS system is equipped with eight BESS rack systems (a total of 1 MWh is installed), and each BESS rack system consists of 14 battery modules. Each module also consists of 14 battery packs. In the offline training process, the BESS capacity is set to 1 MWh, which is the same as the actual grid BESS. Since the grid-level BESS in the grid is too large for aging cycles, actual battery testing is implemented in a similar environment at a low level using the same battery pack, which is disassembled from the same model as the actual grid battery module.

3.1. State

The current state of the information s_t contains the grid model and BESS aging model states. The decision-making process of the Markov decision process (MDP) model for BESS scheduling is proposed in this paper. The MDP signifies that the next state at $t+1$ is only related to the action and state information at time t and is independent of the

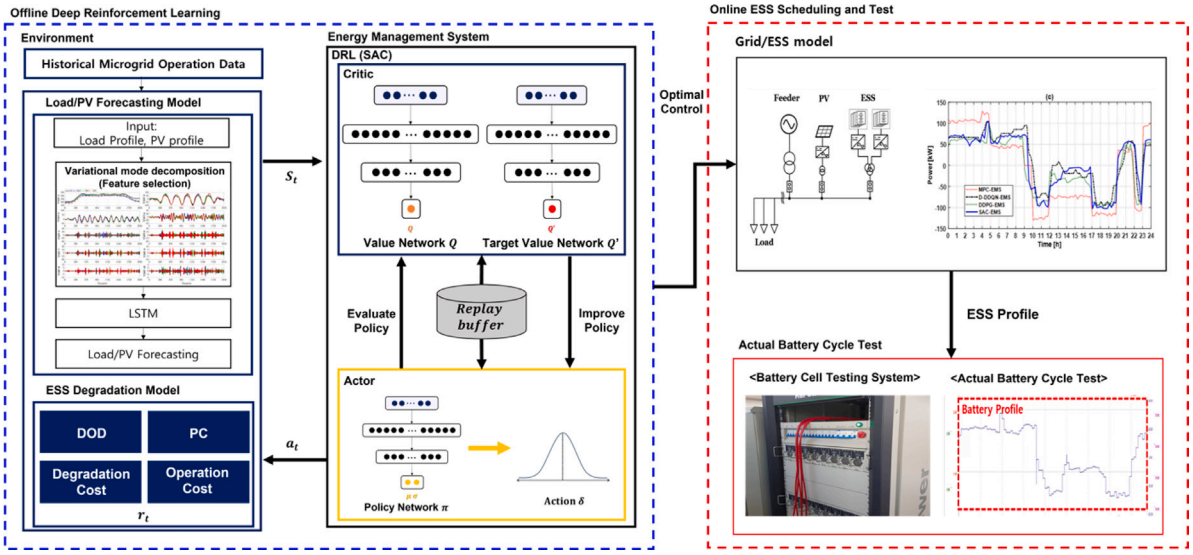


Fig. 4. The workflow of the proposed BESS Scheduling based on SAC.

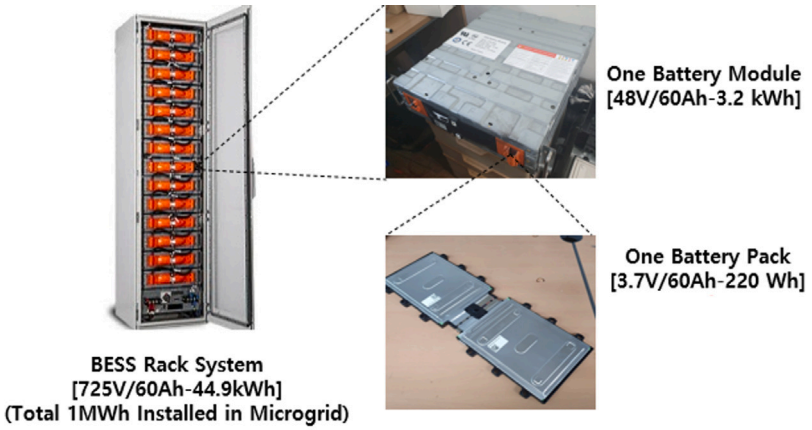


Fig. 5. Energy storage system, battery module and battery pack used in the experiment.

previous state at time $t-1, t-2, t-3, \ldots$. The state $s_t \in S$ at time step t are defined below:

$$s_t = [\underbrace{P_t^G, P_t^{DE}, P_t^{PV}, P_t^{f,DE}, P_t^{f,PV}}_{\text{Microgrid}}, \underbrace{P_t^{BESS}, SOC_t}_{\text{BESS}}, \underbrace{C_t^{TOU}}_{\text{TOU}}], \tag{13}$$

where $P_t^G, P_t^{DE}, P_t^{PV}, P_t^{f,DE}, P_t^{f,PV}$ is from the grid and the forecasting models, and P_t^{BESS} and SOC_t are from the BESS models, C_t^{TOU} is from the TOU pricing.

3.2. Action

The control signal a_t is sent by the EMS to control the BESS power output. Note that the action is chosen by following the strategy π, which will be updated by the SAC algorithm in the direction of a higher reward. The action $a_t : -1 \le a_t \le 1$ is defined as the BESS's normalized power to prevent DRL overestimation and divergence, since the SAC selects an action (charging/discharging or rest) from an action space based on the policy π. The actual power supply can be reconstructed by multiplying P_B (P_B is the maximum power of the BESS). The goal of the proposed algorithm is to find the optimal policy π^* that maximizes the reward (reducing the overall cost).

3.3. Reward and penalty

The reward r_t at time slot t indicates the immediate return, which is obtained when the agent executes the action a_t based on the state s_t. The reward is the key to achieve proper performance in BESS scheduling. In this paper, the goal of BESS scheduling is to maximize the overall electricity cost savings while considering the cost of BESS degradation. Supposing that the experiment start at time slot t in one episode, the cumulative reward is expressed as:

$$R_t = r_t + \gamma \left[r_{t+1} + \gamma r_{t+2} \cdots + \gamma^{T-t-1} r_T \right], \tag{14}$$

where T represents the finite MDP steps and $\gamma \in [0, 1]$ is a discount factor, which is responsible for balancing the current and future return. Thus, given the direction of a policy π, the value function for state s_t can be described as follows:

$$V^{\pi}(s_t) = E\left[R(s_t, t) | s_t = s \right], \tag{15}$$

In addition, the battery degradation penalty and forecasting error penalty functions are proposed to prevent excessive charging/discharging and an increase in DOD/PC, which are defined as follows:

$$\tau_t^D = \begin{cases} \varphi_1 \times c_{d,k} \times C_B, & SOC_{min} < SOC_t \le 0.4, \\ 0 & 0.4 < SOC_t \le 0.6, \\ 0 & 0.6 < SOC_t \le 0.8, \\ \varphi_2 \times c_{d,k} \times C_B, & 0.8 < SOC_t \le SOC_{max}, \end{cases} \tag{16}$$

where φ_1 and φ_2 are the degradation penalty coefficients and C_B is the battery cost per kWh. In this study, C_B is set to a constant value that does not change during daily scheduling [20], and the degradation cost is affected by aging indices such as the DOD and PC. The degradation coefficient $c_{d,k}$ can be determined by the defined aging index, which increase the degradation cost [29].

The proposed BESS scheduling method determines the optimal BESS charging time and charge/discharge rate based on PV and load forecasts. However, deviations in demand and supply will occur due to forecast errors. The proposed DRL and MPC models control the BESS based on predictions, and if the predictions are inaccurate, optimized BESS charging/discharging cannot be achieved, resulting in increased operating costs. Therefore, this study considers a penalty for forecast uncertainty. The forecasting error penalty is defined as follows:

$$\tau_t^F = \begin{cases} \psi_1 \times P_t^{BESS}, & \Delta P > 0, \\ 0 & \Delta P = 0, \\ \psi_2 \times P_t^{BESS}, & \Delta P < 0, \end{cases} \tag{17}$$

where ψ_1 and ψ_2 are the deviation penalty coefficients. This penalty function is imposed for charging the BESS when the actual load is greater than expected or discharging the BESS when the actual load is smaller than expected.

The reward r_t that results from the EMS a_t is set to be equal to the grid's negative overall cost C_t^o. At each time step, the immediate reward can be expressed as follows:

$$r_t(s_t, a_t) = -[C_t^o + \tau_t^D + \tau_t^F], \tag{18}$$

where r_t is the reward of making decision a_t in state s_t.

3.4. Benchmark DRLs for performance evaluation

Before introducing the proposed BESS scheduling results, we briefly introduce some background information on DRL. We compare the performance of the proposed methods using variant DRLs to optimize BESS scheduling.

3.4.1. Double Deep Q Learning (DDQN)

Q-learning uses a critic network and the Q-function, which infers an optimal policy from the state–action pair. The action-value function indicates the extent to which the action taken in each state is effectively denoted by $Q_\pi(s, a)$. The optimal $Q_{\pi^*}^*(s, a)$ is used to represent the maximum accumulative reward of action a_t in state s_t, and the action-value $Q(s_t, a_t)$ is updated using:

$$Q_{\pi^*}^*(s, a) \leftarrow (1 - \theta)Q(s_t, a_t) + \theta \left[r_t + \gamma \max Q(s_{t+1}, a_{t+1})\right], \tag{19}$$

where θ represents the learning rate, which determines the effect of the new reward on the old $Q(s_t)$ value, and γ is the discount factor that balances the immediate and future rewards. However, Q-learning is severely affected by the curse of dimensionality because of its tabular approach to storing the Q-values. To overcome this problem, the value function for the standard Q-learning algorithm is replaced by a DQN with the parameter θ, which is given by the DNN's weights and biases such that $Q_\pi(s, a) \approx Q(s, a, \theta)$. This approximation is subsequently used to define the objective function by the mean-squared error in the Q-function as follows [19,20]:

$$\mathcal{L}(\theta) = \mathbb{E}\left[(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \theta) - Q(s_t, a_t, \theta))^2\right] \tag{20}$$

The DNN accepts a continuous state as an input and outputs an estimate of the Q-function for each discrete action; when acting, the DNN chooses the maximum action-value at each state. However, the max operator leads to overestimations. The DDQN mitigates this problem using two separate networks to decouple the action selection from the target Q value generation. The DDQN uses the following target:

$$y_t^{DDQN} = R_t + \gamma Q\left(s_{t+1}, \arg\min_{a_{t+1}} Q(s_{t+1}, a_{t+1}, \theta), \theta^-\right) \tag{21}$$

The DDQN can more effectively overcome the curse of dimensionality problem by selecting the best actions using the online instead of the target model [30].

3.4.2. Deep Deterministic Policy Gradients (DDPG)

The DDPG is a type of actor–critic-based off-policy method and model-free algorithm based on the DPG, and it can operate in a continuous state and action space. This algorithm uses DNNs to establish two approximation functions from the actor–critic algorithm [13]. The actor network can be described as a policy function $\mu(s|\theta^\mu)$ that deterministically maps states to actions, whereas the critic $Q(s, a)$ is trained using the Bellman equation. While the agent is being trained, the critic and actor network weights are continually updated based on the observed reward at each time step. To facilitate training, these two networks are also created with a copy: a actor target network μ' with parameter $\theta^{\mu'}$ and a critic target network Q' with parameter $\theta^{Q'}$. A loss function L is computed as the mean squared error between the target value and the critic's estimated Q-value, which is written as:

$$L(\theta^Q) = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(s_t, a_t|\theta^Q)^2), \tag{22}$$

where M is the size of the experience mini-batch, and y_i is obtained by applying Eq. (21) as in Q-learning. The critic network's parameters θ^Q are updated by minimizing L across the mini-batch of experiences sampled from the replay buffer. On the other hand, the actor network's parameters are updated according to the gradient of the value function expectation J. The resulting policy gradient $\nabla_{\theta^\mu} J$ is used to update the actor, and is written as:

$$\nabla_{\theta^\mu} J \approx \frac{1}{M}\sum_{i=1}^{M}\left[\nabla_a Q(s_i, a|\theta^Q)|_{a=\mu(s_i|\theta^\mu)} \nabla_{\theta^\mu} \mu(s_i|\theta^\mu)\right] \tag{23}$$

Finally, the target actor and critic networks are updated using a smoothing factor τ to prevent learning instabilities [31]:

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'} \end{aligned} \tag{24}$$

3.4.3. Soft Actor Critic (SAC)

Conventional model-free DRL methods have two limitations: high sampling complexity and weak convergence, which both depend on parameter tuning. To improve the sample efficiency, off-policy algorithms such as the DDPG were proposed, but their performance relies heavily on hyper-parameters. Therefore, the state-of-the-art off-policy DRL algorithm based on maximum-entropy, SAC, is proposed. Similar to the DDPG, the SAC also uses an actor–critic architecture and experience replay buffer that reuses past experiences for an off-policy formulation. Different from DDPG, the primary feature of the SAC is entropy regularization; the algorithm is based on the maximum entropy in the reinforcement learning framework, and its goal is to maximize both the expected rewards and the entropy. This goal is expressed as follows [21–23]:

$$\pi^* = argmax_\pi E_\pi\left[\sum_{t=0}^{\infty}\gamma^t(r_t + \alpha H_t^\pi)\right], \tag{25}$$

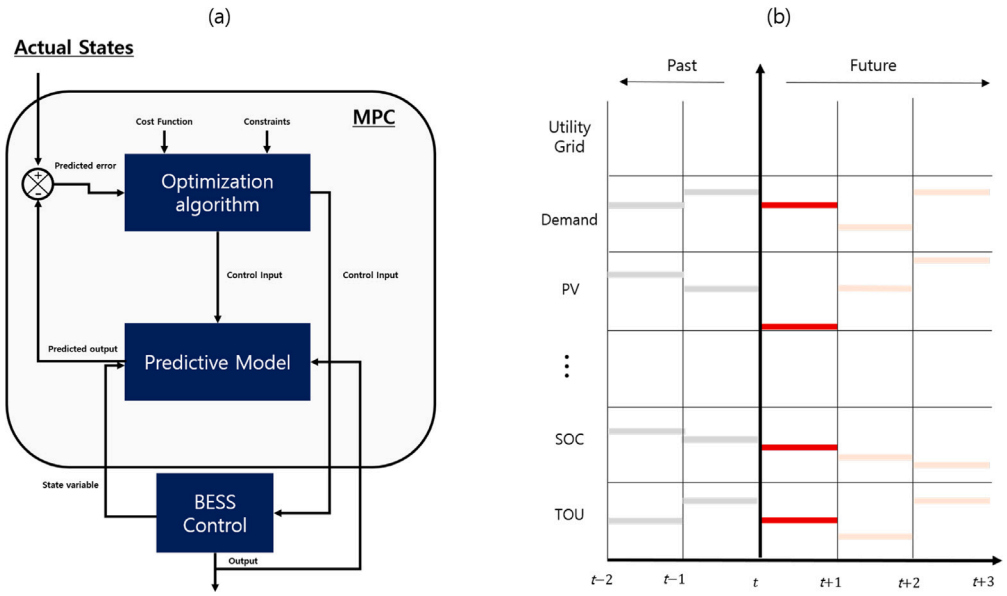


Fig. 6. (a) The MPC-EMS structure for BESS control. (b) The grid/ess model scheme.

where H is the Shannon entropy term that represents the agent's attitude in taking arbitrary actions, and α is a regularization coefficient that indicates the importance of the entropy term for rewards. In general, considering conventional DRL algorithms, α is 0. The maximization of this target function has a close connection with the exploration–exploitation trade-off, and it ensures that the agent is explicitly pushed towards the exploration of new policies and prevented from providing sub-optimal results. As a result, the SAC provides learning robustness and sampling efficiency.

3.4.4. Model Predictive Control (MPC)

MPC is widely used in industry as an effective approach to handling large-scale multivariate constraint control problems. The MPC model is used for performance comparison with the proposed DRL model. The MPC-EMS structure for BESS control and the grid/ess model scheme are shown in Fig. 6. The MPC is to select control actions by iteratively solving an online constrained optimization problem that is designed to minimize a performance index over a finite prediction range based on predictions obtained from a model of the system. The operating/aging cost objective function and constraints are formulated in the same environmental model used for DRL. In the MPC approach, the control inputs for each stage are computed online instead of using precomputed offline. In each sampling period, as shown in Fig. 6(a), the system state is updated, the optimal control problem is solved online, and the controller's time window is stepped back by one step. In this study, the MPC-EMS model is controlled to minimize the operating cost using the same objective function (6) as the DRLs under the same constraints (1)–(4). As shown in Fig. 6(b), the MPC model determined state predictions and actions for each t as follows:

The MPC model determines the control of the BESS by considering the state variables and predicted results. The MPC model plans control commands by considering the state variables at time t, outputs control commands, and receives feedback on the current state variables. It compares and evaluates the output and response of the control commands and updates the cost function to calculate the control input for the next control command.

At each time t:

- MPC model get new state to update the estimate of the current state
- Solve the optimization problem within the constraints
- Apply only the first optimal action and discard the remaining samples

Compared to DRL, MPC trains with a pre-defined model, so the training time is much faster, but external interference or model uncertainty can have a large impact on performance.

4. Case study

In this section, the proposed DRL-based BESS scheduling was implemented and compared with several common solutions, including the DDQN, DDPG and MPC. This section also includes a performance evaluation using simulated scenario based on a real-world grid datasets and actual battery cycle tests. For the purposes of this study, a cycle is defined as 24 h of BESS scheduling starting at 0:00 AM. The dataset time interval used is 5 min, which is the measurement interval of a smart meter, so the control interval in the simulation is also set to 5 min. The composition ratio of the dataset for training data and validation data is 7:3, respectively. The proposed models estimate the total operating costs, including the battery degradation costs, and implement an optimized BESS scheduling for actual batteries to compare the SOH according to the DOD. In addition, to evaluate the generalization of a well-trained agent, three scenarios were introduced in Table 2.

Battery packs with similar initial capacities were used as scheduling cycle test samples. The average full capacity of the battery pack used was 53.81 Ah, which was reduced by 10.31% when compared to the nominal capacity (60 Ah) due to actual grid operation history.

4.1. Training process

The hyperparameters used in this experiments were well-tuned by hyperparameter tuning techniques from previous reports of similar studies. The Adam optimizer was used to learn the DNN weights. The discount factor was set to 0.95, batch size was 64 and the learning rate was set to 0.001 [20,31]. The DDQN action spaces were discretized to 10 kW intervals from −150 kW to 150 kW. The examined DRL methods were implemented using MATLAB and Python.

The training results of the networks of DRL models are shown in Fig. 7. The DRL agents were trained by 7000 episodes, and the cost curve obtained using each DRL method is shown in Fig. 7. Fig. 7(a) represents the performance evolution of the network during the training process of the proposed DRLs. The DDQN, DDPG and SAC are

Table 2
Operating conditions in different scenarios.

Scenario	BESS Initial SOC [%]	Average load consumption [kW]	Average forecasting deviation [%]
SC #1	50	396.23	7.21
SC #2	60	275.41	5.56
SC #3	40	335.87	3.89

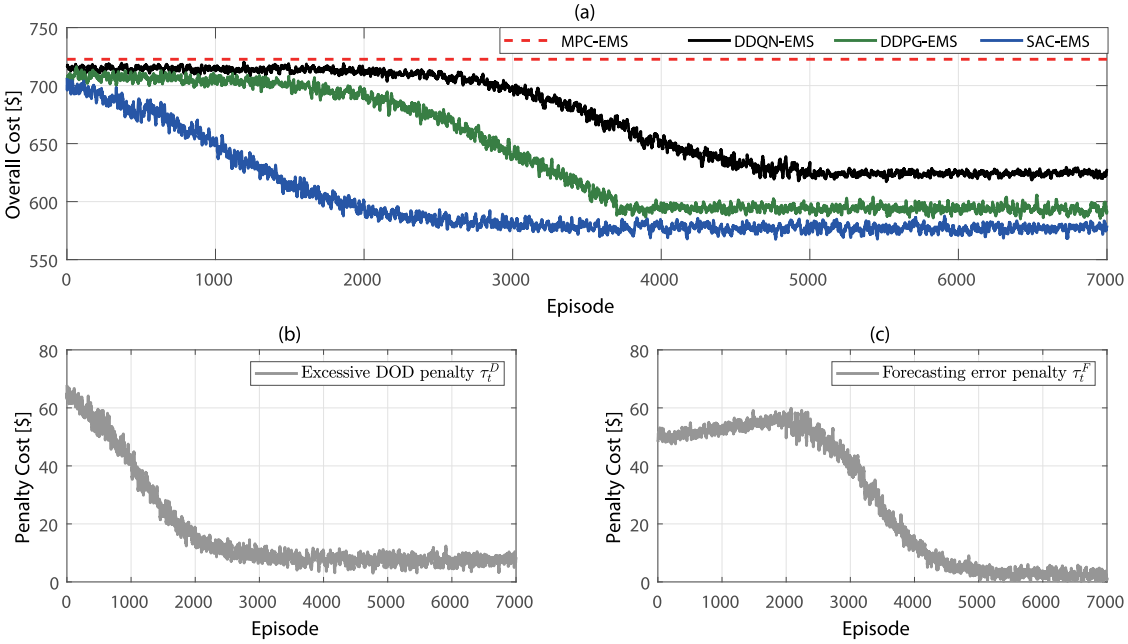


Fig. 7. (a) Daily total operating cost evaluated for the examined DRL methods. (b) Excessive DOD penalty. (c) Forecasting error penalty.

off-policy DRL algorithms, and they use random sampling from inside the replay buffer for training. The DDQN had difficulty reaching the optimized policy until the replay buffers were filled. The DDQN performance did not improve after 5000 episodes and converged. The DDPG demonstrated superior performance when compared to DDQN, but it has a slower learning rate when compared to the SAC. The SAC agent improves its policy continuously during the first 5000 episodes and then stabilizes around a null reward. In Fig. 7(b) and Fig. 7(c), the change in penalty values is presented. The excessive DOD penalty was consistently reduced and converged over 3000 episodes. On the other hand, the forecasting error penalty increased to about 2000 episodes, but they trained rapidly afterward and converged after 5000 episodes Since deviation in the demand/PV forecasting exists, the forecasting error penalty does not converge to zero. Similarly, the battery degradation penalty did not converge to zero because the scheduling SOC range was outside the 40%–80% range. This result occurred because it is advantageous to charge/discharge the BESS even if it undergoes penalties. The experiment demonstrates that the SAC can be learned directly in an environment due to efficient learning.

4.2. Results and discussion

4.2.1. Experimental results

Fig. 8 shows an example of BESS scheduling. In Fig. 8(a), P_t^{DE} represents the actual demand, P_t^{PV} represents the actual PV generation, $P_t^{f,DE}$ represents the predicted demand data, and $P_t^{f,PV}$ represents the predicted PV supply. Fig. 8(b) shows the TOU pricing. Fig. 8(c) represents the experiment results using the proposed scheduling DRL methods (the red dotted line represents the BESS scheduling using the MPC based on TOU, the black dotted line represents the BESS scheduling using the DDQN, the green dotted line represents the BESS scheduling using the DDPG, and the blue solid line represents the BESS scheduling using the SAC). BESS scheduling methods are primarily dependent on TOU, but additional scheduling is performed according to the deviation in supply/demand.

As shown in Fig. 9, the SOC ranges according to the proposed DRL methods are represented and are implemented in scenario 1. Since the MPC-EMS method does not consider the SOC range, the DOD is significantly increased due to excessive charging/discharging. As shown in Table 3, the deep discharge time (DDT), another cause of accelerated battery aging, is defined as the time in which the SOC is less than 40% [9]. The MPC-EMS method uses existing methods based on TOUs without considering the BESS aging conditions. Thus, the MPC-EMS is maximally charged in the lowest TOU range and continuously discharges at peak loads with a DDT of 4.5 h, resulting in a high DOD as shown in Fig. 9. In particular, because of the low initial SOC conditions in scenario 3, there are DDT intervals for all scheduling methods presented in the comparison groups (MPC, DDQN, and DDPG). However, there is no DDT in the SAC-EMS scheduling, which maintained a stable DOD. Similarly, when comparing the overall operating cost, the MPC-EMS method has a higher operating cost than the three methods using DRL (DDQN, DDPG, and SAC). In the most heavily loaded Scenario 1, the MPC-EMS method has a 27% higher operating cost ($ 722.70) compared to the best performing SAC-EMS ($ 567.35). Similarly, MPC-EMS has a 25% higher operating cost ($ 687.32) than SAC-EMS ($ 545.82) in Scenario 2 and 23% higher operating cost ($ 669.91) than SAC-EMS ($ 541.62) in Scenario 3. Although the MPC-EMS method also forecasts the future state of PV, load, etc. and optimizes through iterative calculations reflecting the state of the BESS, the performance is subject to the accuracy of the defined model. Since the DRL has strengths in solving uncertainties and nonlinearities in the environment, it outperforms the MPC-EMS model in conditions with uncertainties such as aging costs of batteries and deviations between demand and load presented in this study.

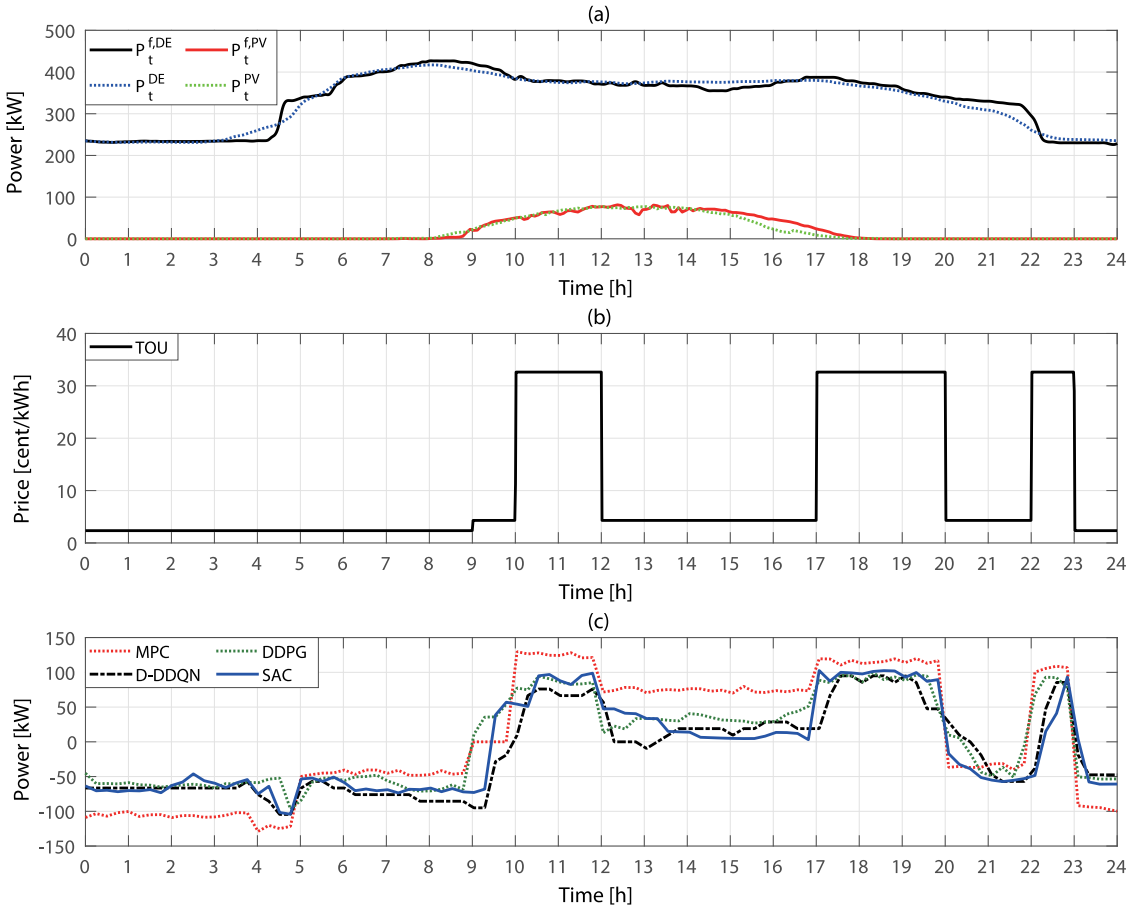


Fig. 8. The BESS scheduling is examined using the proposed DRL methods. (a) Comparison of actual and forecast values. (b) Time-of-Use. (c) BESS Scheduling results.

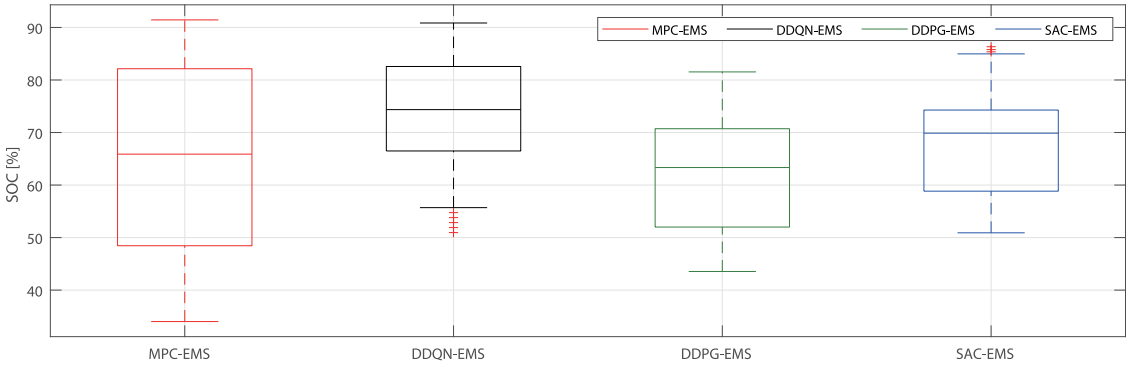


Fig. 9. Average SOC range according to the proposed DRL methods.

Table 3
Daily total operating cost and depth of discharge in the different scenarios.

	MPC			DDQN			DDPG			SAC		
Scenario	SC #1	SC #2	SC #3	SC #1	SC #2	SC #3	SC #1	SC #2	SC #3	SC #1	SC #2	SC #3
Total operating cost [$]	722.70	687.32	669.91	626.83	599.82	588.72	592.25	556.72	549.81	567.35	545.82	541.62
Depth of discharge [%]	57.43	54.21	55.72	39.90	38.73	38.88	37.99	38.24	37.95	35.43	36.72	37.44
Deep discharge time [h]	4.5	2.25	6.75	0	0	2.5	0	0	0.75	0	0	0

4.2.2. Actual battery aging tests

Fig. 10 shows the test results for the application of optimized BESS charging/discharging scheduling to actual batteries and shows the effect of DOD on battery health. After each BESS scheduling, a full discharge was carried out to verify the remaining capacity every 50 cycles. This full discharge shows that the SOH and capacity fade. After 350 cycles, the MPC-EMS capacity loss is about 11.51%. In comparison, the DDQN-EMS and DDPG-EMS capacity losses are 9.63% and 9.15%, respectively. In particular, the SAC-EMS scheduling exhibited the smallest capacity loss, 5.97%, in battery aging tests. The MPC-EMS method demonstrates faster capacity reduction due to the higher DOD while the DRL methods showed slower capacity reduction since they maintained

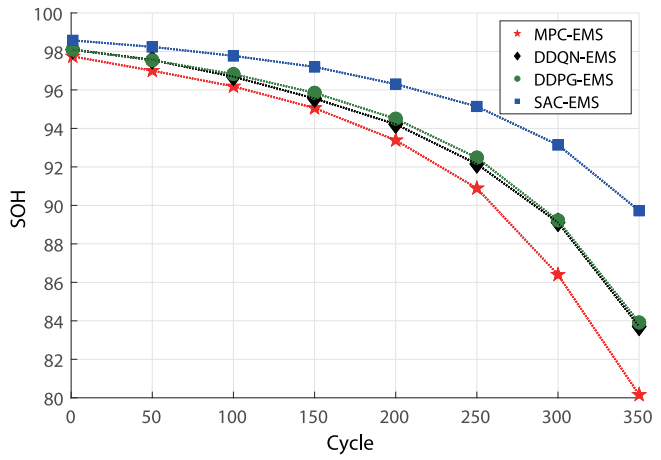


Fig. 10. The SOH changes of battery packs using the proposed BESS scheduling methods.

a stable SOC and avoided the DDT. These experimental results indicate that if BESS maintains a high DOD with a low SOC range, it can reduce the battery lifetime and increase the degradation costs.

5. Conclusion

This study proposes the development of the BESS scheduling method to address the grid energy management problem. Data-driven DRL optimization methods have been proposed because it is difficult to have a perfect physical/predictive model in actual BESS operation; the BESS method considers the battery's SOC range to reduce the operation/degradation cost and extend the battery's lifetime. This proposed method leverages the performance of the state-of-the-art SAC DRL in combination with the battery aging model, which is designed using the battery aging index.

The proposed methods are implemented in an actual battery test and contribute to real-time scheduling implementation. The proposed method's performance was evaluated by performing various case study to verify its adaptability in various situations; additionally, the aging cycle test shows that BESS management considering SOC/DOD conditions can extend the battery's lifetime. Furthermore, optimization for DOD could become even more important when long-term operation of the BESS is considered. The proposed approach is expected to be more economical because long-term operation must also include long-term maintenance/replacement costs due to battery aging. Therefore, it is necessary to operate the BESS in an optimized DOD range to avoid increasing costs and capacity loss due to aging.

CRediT authorship contribution statement

Seon Hyeog Kim: Methodology, Writing – original draft, Validation, Investigation, Experiments, Visualization, Reviewing and writing. **Yong-June Shin:** Resource, Funding acquisition, Conceptualization, Supervision, Reviewing & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Seon Hyeog Kim reports financial support was provided by Yonsei University. Seon Hyeog kim reports a relationship with Korea Institute of Energy Technology Evaluation and Planning that includes: funding grants. Seon Hyeog Kim has patent pending to Seon Hyeog Kim.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (No. KETEP-20202020800290 & 20202020900290) and by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning, (No. NRF-2020R1A2B5B03001692). The authors would like to thank LG electronics for the real-world grid datasets and battery energy storage system.

References

[1] M. Wang, et al., The value of vehicle-to-grid in a decarbonizing California grid, J. Power Sources 513 (2021) 230472.
[2] L. Luo, Optimal scheduling of a renewable based microgrid considering photovoltaic system and battery energy storage under uncertainty, J. Energy Storage 28 (2020) 101306.
[3] J. Wu, et al., Energy management strategy for grid-tied microgrids considering the energy storage efficiency, IEEE Trans. Ind. Electron. 65 (12) (2018) 9539–9549.
[4] H.A.U. Muqeet, A. Ahmad, Optimal scheduling for campus prosumer microgrid considering price based demand response, IEEE Access 8 (2020) 71378–71394.
[5] Y. Li, et al., Optimal scheduling of an isolated microgrid with battery storage considering load and renewable generation uncertainties, IEEE Trans. Ind. Electron. 66 (2) (2019) 1565–1575.
[6] F. Garcia-Torres, et al., Optimal economic schedule for a network of microgrids with hybrid energy storage system using distributed model predictive control, IEEE Trans. Ind. Electron. 66 (3) (2019) 1919–1929.
[7] F. Garcia-Torres, C. Bordons, Optimal economical schedule of hydrogen-based microgrids with hybrid storage using model predictive control, IEEE Trans. Ind. Electron. 62 (8) (2015) 5195–5207.
[8] U. Raveendrannair, et al., An analysis of multi objective energy scheduling in PV-BESS system under prediction uncertainty, IEEE Trans. Energy Convers. 36 (3) (2021) 2276–2286.
[9] L. Liu, et al., Managing battery aging for high energy availability in green datacenters, IEEE Trans. Parallel Distrib. Syst. 28 (12) (2017) 3521–3536.
[10] M.A. Ortega-Vazquez, Optimal scheduling of electric vehicle charging and vehicle-to-grid services at household level including battery degradation and price uncertainty, IET Gener. Transm. Distrib. 8 (6) (2014) 1007–1016.
[11] C. Zhou, et al., Modeling of the cost of EV battery wear due to V2G application in power systems, IEEE Trans. Energy Convers. 26 (4) (2011) 1041–1050.
[12] B. Xu, et al., Factoring the cycle aging cost of batteries participating in electricity markets, IEEE Trans. Power Syst. 33 (2) (2018) 2248–2259.
[13] Yan, et al., Deep reinforcement learning-based optimal data-driven control of battery energy storage for power system frequency support, IET Gener. Transm. Distrib. 14 (25) (2020) 6071–6078.
[14] S.-J. Park, et al., Depth of discharge characteristics and control strategy to optimize electric vehicle battery life, J. Energy Storage 59 (2023) 106477.
[15] R.D. Deshpande, et al., Physics inspired model for estimating 'cycles to failure' as a function of depth of discharge for lithium ion batteries, J. Energy Storage 33 (2021) 101932.
[16] M. Eskandari, et al., Battery energy storage systems (BESSs) and the economy-dynamics of microgrids: Review, analysis, and classification for standardization of BESSs applications, J. Energy Storage 55 (Part B) (2022) 105627.
[17] S. Lee, D.H. Choi, Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, Sensors 19 (18) (2019) 3937.
[18] Y. Du, F. Li, Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning, IEEE Trans. Smart Grid 11 (2) (2020) 1066–1076.
[19] Z. Wan, et al., Model-free real-time EV charging scheduling based on deep reinforcement learning, IEEE Trans. Smart Grid 10 (5) (2019) 5246–5257.
[20] J. Cao, et al., Deep reinforcement learning based energy storage arbitrage with accurate lithium-ion battery degradation model, IEEE Trans. Smart Grid 11 (5) (2020) 4513–4521.
[21] B. Zhang, et al., Soft actor-critic–based multi-objective optimized energy conversion and management strategy for integrated energy systems with renewable energy, Energy Convers. Manag. 243 (2021) 114381.
[22] S. Wang, et al., Deep reinforcement scheduling of energy storage systems for real-time voltage regulation in unbalanced LV networks with high PV penetration, IEEE Trans. Sustain. Energy 12 (4) (2021) 2342–2352.
[23] J. Wu, et al., Battery thermal- and health-constrained energy management for hybrid electric bus based on soft actor-critic DRL algorithm, IEEE Trans. Ind. Electron. 17 (6) (2021) 3751–3761.

[24] Korea electric power corporation (KEPCO), 2020, [Online]. Available: https://home.kepco.co.kr.
[25] T.A. Lehtola, A. Zahedi, Electric vehicle battery cell cycle aging in vehicle to grid operations: a review, IEEE Trans. Emerg. Sel. Topics Power Electron. 9 (1) (2021) 423–437.
[26] C. Ju, et al., A two-layer energy management system for microgrids with hybrid energy storage considering degradation costs, IEEE Trans. Smart Grid 9 (6) (2018) 6047–6057.
[27] E. Wikner, T. Thiringer, Extending battery lifetime by avoiding high SOC, Appl. Sci. 8 (2018) 1825–1840.
[28] S. Kim, et al., Deep learning based on multi-decomposition for short-term load forecasting, Energies 11 (12) (2018) 3433.
[29] Z. Wang, et al., Dueling network architectures for deep reinforcement learning, in: Proc. Int. Conf. Learning Representations, 2016.
[30] V. Bui, et al., Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties, IEEE Trans. Smart Grid 11 (1) (2020) 457–469.
[31] Y. Ye, et al., Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning, IEEE Trans. Smart Grid 11 (4) (2020) 3068–3082.