TIME-VARYING CONSTRAINT-AWARE REINFORCEMENT LEARNING FOR ENERGY STORAGE CONTROL

Jaeik Jeong, Tai-Yeon Ku & Wan-Ki Park
Energy ICT Research Section, Electronics and Telecommunications Research Institute (ETRI)
Daejeon 34129, Republic of Korea
{jaeik1210,kutai,wkpark}@etri.re.kr

ABSTRACT

Energy storage devices, such as batteries, thermal energy storages, and hydrogen systems, can help mitigate climate change by ensuring a more stable and sustainable power supply. To maximize the effectiveness of such energy storage, determining the appropriate charging and discharging amounts for each time period is crucial. Reinforcement learning is preferred over traditional optimization for the control of energy storage due to its ability to adapt to dynamic and complex environments. However, the continuous nature of charging and discharging levels in energy storage poses limitations for discrete reinforcement learning, and time-varying feasible charge-discharge range based on state of charge (SoC) variability also limits the conventional continuous reinforcement learning. In this paper, we propose a continuous reinforcement learning approach that takes into account the time-varying feasible charge-discharge range. An additional objective function was introduced for learning the feasible action range for each time period, supplementing the objectives of training the actor for policy learning and the critic for value learning. This actively promotes the utilization of energy storage by preventing them from getting stuck in suboptimal states, such as continuous full charging or discharging. This is achieved through the enforcement of the charging and discharging levels into the feasible action range. The experimental results demonstrated that the proposed method further maximized the effectiveness of energy storage by actively enhancing its utilization.

1 INTRODUCTION

Energy storage devices, such as batteries, thermal energy storages, and hydrogen systems, play a pivotal role in mitigating the impact of climate change (Aneke & Wang, 2016; Jacob et al., 2023). These storage technologies are instrumental in capturing and efficiently storing excess energy generated from renewable sources during peak production periods, such as sunny or windy days. By doing so, they enable the strategic release of stored energy during periods of high demand or when renewable energy production is low, thereby optimizing energy distribution and reducing reliance on traditional fossil fuel-based power generation. It can be utilized in energy arbitrage by attempting to charge when surplus energy is generated and energy prices are low or even negative, and conversely, discharging during periods of energy scarcity when prices are high (Bradbury et al., 2014). The integration of energy storage with arbitrage strategies contributes to grid stability, enhances overall energy reliability, and fosters a more sustainable energy ecosystem. Energy storage devices are utilized in various capacities, ranging from small-scale applications for households to large-scale units for the overall grid operation, contributing to mitigating climate change (Ku et al., 2022; Inage, 2019).

To enhance the utility of energy storage devices, determining optimal charge and discharge levels for each time period is crucial. In recent times, reinforcement learning techniques have gained prominence over traditional optimization methods for this purpose (Cao et al., 2020; Jeong et al., 2023). Unlike conventional optimization approaches, reinforcement learning allows for dynamic adaptation and decision-making in response to changing conditions, enabling energy storage systems to continuously learn and improve their performance over time. This shift towards reinforcement learning reflects a recognition of its ability to navigate complex and dynamic environments. It offers

a more adaptive and effective solution to optimize charging and discharging strategies for energy storage devices across diverse temporal patterns. This transition in methodology underscores the importance of leveraging advanced learning algorithms to maximize the operational efficiency of energy storage systems in real-world, time-varying scenarios.

The continuous nature of energy storage device charge and discharge levels poses a challenge when employing discrete reinforcement learning techniques such as Q-learning. These algorithms operate with a predefined set of discrete actions, e.g., fully or partially charging/discharging (Cao et al., 2020; Rezaeimozafar et al., 2024). It limits their suitability for tasks involving continuous variables, and thereby constraining the system's ability to explore and optimize across a continuous range of charge and discharge values. Consequently, the utilization of these methods may lead to suboptimal solutions, as the algorithms cannot fully capture the intricacies of the continuous action space (Lillicrap et al., 2015). Due to this limitation, continuous reinforcement learning approaches are often employed such as proximal policy optimization (PPO) (Schulman et al., 2017), to better address the need for precise decision-making in the continuous spectrum of charge and discharge levels.

In continuous reinforcement learning, however, challenges also arise when determining charge and discharge levels due to the dynamic nature of the state of charge (SoC) over time. The range of feasible charge and discharge actions varies based on the evolving SoC. For example, setting actions like charging to a negative value (e.g., complete charging to -1) and discharging to a positive value (e.g., complete discharging to 1) results in a feasible action range of 0 to 1 when the battery is fully charged, and -1 to 0 when the battery is fully discharged. Nevertheless, current approaches often struggle to effectively address such time-varying action ranges. Currently, when actions fall outside the designated range, a common solution involves charging up to the maximum SoC or discharging up to the minimum SoC (Jeong & Kim, 2021; Kang et al., 2024). However, this approach introduces a potential challenge wherein the SoC may become stuck in a fully charged or fully discharged state during the learning, limiting its ability to explore within the full spectrum of SoC states. Additionally, there are approaches that assign a cost or negative reward proportional to the extent by which actions deviate outside the designated range (Lee & Choi, 2020; Zhang et al., 2023). However, there is a potential for overly conservative learning, as the emphasis leans heavily towards actions that remain within the designated range. Addressing these issues is crucial for adapting to the time-varying action ranges associated with the changing SoC over time.

In this paper, we propose a continuous reinforcement learning approach to address these challenges in energy storage device control. The key innovation lies in augmenting the conventional objective functions of the actor and critic with an additional supervising objective function designed to ensure that the output actions at each time step fall within the feasible action space. In contrast to traditional supervised learning, where the training encourages the output to approach specific values, the introduced supervising objective function focuses on constraining the output within a particular range. This inclusion is pivotal as bringing the output within the feasible action space enables the activation of charging and discharging operations in energy storage devices. This proactive approach helps prevent the energy storage from getting stuck in suboptimal states like complete charge or discharge, facilitating the exploration of more optimal actions. The integration of the supervising objective function enhances the adaptability and efficiency of continuous reinforcement learning in optimizing energy storage operations over changing states. We conducted experiments related to energy arbitrage and found that the addition of the supervising objective function effectively addressed the challenge of the system becoming stuck in suboptimal conditions. This objective function prevents the energy storage device from being trapped in states like complete charge or discharge and promotes continual exploration for optimal energy arbitrage.

2 METHODS

In this section, we present a continuous reinforcement learning model combined with a supervising objective function. Modern continuous reinforcement learning comprises actor and critic components, each with distinct objective functions depending on the specific reinforcement learning algorithm employed. In this paper, we adopt the proximal policy optimization (PPO) algorithm, known for its compatibility with long short-term memory (LSTM) (Schulman et al., 2017), to address energy storage device control problems. When tackling control problems associated with energy storage devices, the majority of reinforcement learning states often involve time-series data such as SoC,

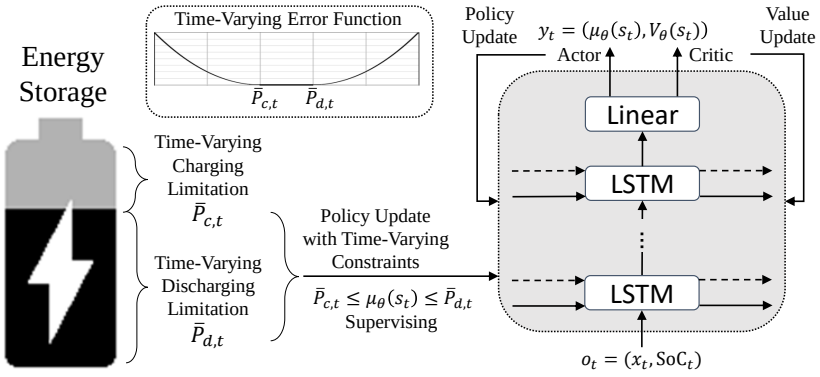


Figure 1. A framework of the proposed method.

energy generation, energy demand, and energy prices. Given this temporal nature, combining PPO with LSTM becomes particularly advantageous. This combination facilitates effective learning and decision-making in scenarios where the state representation is composed of sequential data.

The overall framework of the proposed method is shown in Figure 1. The input to the LSTM at each time step is referred to as an observation because, in the context of time-series data, the value corresponding to each time is a partially observable state. At time t, the LSTM input, denoted as o_t, comprises the SoC at time t, represented as SoC_t, and other variables x_t pertinent to the objectives of energy storage device control. These additional variables may include factors such as energy generation, demand, prices, or other relevant parameters, depending on the specific goals of energy storage control. The past observations construct the state at time t as $s_t = (o_0, o_1, \cdots, o_t)$. In our problem, the action is the charging and discharging amount, and since the action can be known according to changes in the SoC within the state, past actions are not separately included as the state. With the LSTM parameters θ, the output of the LSTM consists of the actor's output, denoted as $\mu_\theta(s_t)$, and the critic's output, denoted as $V_\theta(s_t)$. Here, $\mu_\theta(s_t)$ represents the mean of the Gaussian policy, while $V_\theta(s_t)$ signifies the estimated value. The standard deviation of the Gaussian policy is predetermined based on the desired level of exploration. During the training phase, action a_t is sampled from the Gaussian policy, and during the actual testing phase, $\mu_\theta(s_t)$ serves as the action a_t (Zimmer & Weng, 2019). Based on the s_t and a_t, reward r_t and the next observation o_{t+1} are given from the environment.

The actor and critic objective functions in the standard PPO formulation are expressed as follows:

$$L^{PPO}_{\text{actor}}(\theta) = \mathbb{E}_t \left[\min \left(R_t(\theta)\hat{A}_t, \text{clip}(R_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right], \tag{1}$$

$$L^{PPO}_{\text{critic}}(\theta) = \mathbb{E}_t \left[\left(r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t) \right)^2 \right], \tag{2}$$

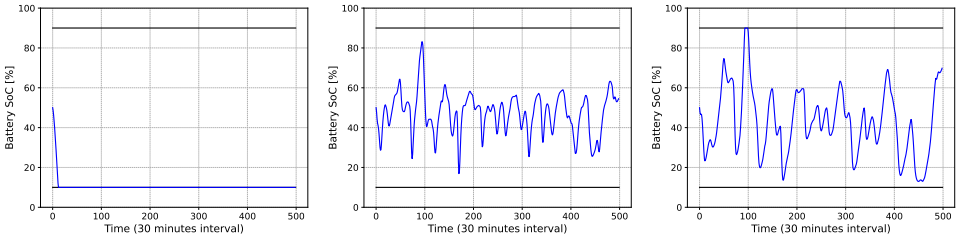
where $R_t(\theta)$ is the ratio of the new and old policies at time t, $\hat{A}_t$ is the estimated advantage function at time t, ϵ is a hyperparameter determining the clipping range, and γ is a discount factor. We are adding a supervising objective function here. Let the charging limitation at time t as $\bar{P}_{c,t}$ and the discharging limitation as $\bar{P}_{d,t}$. These limitations are determined based on the SoC_t. As the SoC varies over time, both the charging and discharging limitations also time-varying. We have defined charging actions as negative and discharging actions as positive, resulting in $\bar{P}_{c,t} \le 0$ and $\bar{P}_{d,t} \ge 0$. The proposed supervising objective function is as follows:

$$L^{PPO}_{\text{supervising}}(\theta) = \min \left(\mu_\theta(s_t) - \bar{P}_{c,t}, 0 \right)^2 + \min \left(\bar{P}_{d,t} - \mu_\theta(s_t), 0 \right)^2. \tag{3}$$

Since $\mu_\theta(s_t)$ serves as the action in the testing phase, we have set the range of $\mu_\theta(s_t)$ to be between $\bar{P}_{c,t}$ and $\bar{P}_{d,t}$. This error function is similar to the mean squared error in supervised learning, with the key distinction that the error is zero within the range of $\bar{P}_{c,t}$ and $\bar{P}_{d,t}$. We finally obtain our main objective, which is minimized at each iteration:

$$L^{PPO}(\theta) = L^{PPO}_{\text{actor}}(\theta) + C_1 L^{PPO}_{\text{critic}}(\theta) + C_2 L^{PPO}_{\text{supervising}}(\theta), \tag{4}$$

where C_1 and C_2 are coefficients of the critic objective function and supervising objective function learning, respectively.



(a) Case 1. (b) Case 2. (c) Case 3 (proposed).

Figure 2. Charging/discharging patterns for 3 cases.

Table 1: Experiment results (30-minutes averaged).

	Metric			
	Charging cost ($)	Disharging revenue ($)	Degradation cost ($)	Total profit ($)
Case 1	-0.000	4.858	-0.080	**4.779**
Case 2	-40.039	53.245	-1.691	**11.514**
Case 3 (proposed)	-37.493	54.085	-1.714	**14.879**

3 RESULTS

In this section, we evaluate the performance of the proposed method by comparing it with two benchmark cases. Case 1 employed a conventional continuous reinforcement learning approach, excluding the equation (3), meaning that the output actions were not restricted to be within the feasible action space. Case 2 incorporated the equation (3) into the reward function. This approach, proposed in (Lee & Choi, 2020), adds negative rewards if the output actions fall outside the feasible action space, rather than explicitly learning the range of output actions. The proposed model is designated as Case 3. We demonstrated the effectiveness of the proposed approach through energy arbitrage experiments based on actual energy price data in 2017 U.K. wholesale market (uk2, 2017). Accordingly, the additional variable x_t becomes the energy price at time t, and the reward r_t is the total profit at time t. We take the first 2000 data points which are sampled every 30 minutes and split the dataset into training set (1000 data points), validation set (500 data points), and test set (500 data points) in chronological order, where the validation set was used for early stopping. We normalize the price data between 0 and 1 by the maximum price \$190.81/MWh. We simulate the proposed method using 100MWh battery with the degradation cost of \$10/MWh. At time slot $t = 0$, we set $\text{SoC}_t = 0.5$, i.e., the half stored energy. We set the minimum and maximum values of the SoC to 0.1 and 0.9, respectively, in order to prevent battery degradation, and the battery charging and discharging model is based on the battery equivalent circuit used in (Cao et al., 2020; Jeong et al., 2023). We used a 2-layer LSTM architecture with 16 neurons and trained the model using the Adam optimizer. All PPO-related parameters were adopted from values commonly used in (Schulman et al., 2017).

Figure 2 illustrates the charging and discharging patterns for three cases. Case 1 shows a scenario where all the initially stored energy is discharged (sold out) and no further actions are taken. This suggests a failure in learning to manage the costs associated with charging in energy arbitrage, resulting in suboptimal behavior. Case 2 demonstrates reasonable utilization of energy arbitrage, but the proposed Case 3 engages in more active energy arbitrage. Introducing Equation (3) as a negative reward makes the agent conservative towards reaching states of complete charge or discharge, leading to reduced utilization of the energy storage. Table 1 presents the 30-minute average of charging cost, discharging revenue, degradation cost, and total profit for the three cases. It is evident that the proposed Case 3 achieves the highest profit.

4 CONCLUSION

In this paper, we introduce a continuous reinforcement learning approach for energy storage control that considers the dynamically changing feasible charge-discharge range. An additional objective

function has been incorporated to learn the feasible action range for each time period. This helps prevent the energy storage from getting stuck in states of complete charge or discharge. Furthermore, the results indicate that supervising the output actions into the feasible action range is more effective in enhancing energy storage utilization than imposing negative rewards when the output actions deviate from the feasible action range. In future research, we will explore combining offline reinforcement learning or multi-agent reinforcement learning to investigate methods for learning a more optimized policy stably.

ACKNOWLEDGMENTS

This work was supported by the Korea Institute of Energy Technology and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of Korea (No. 2021202090028C).

REFERENCES

The changing price of wholesale UK electricity over more than a decade. https://www.ice.org.uk/knowledge-and-resources/briefing-sheet/the-changing-price-of-wholesale-uk-electricity, 2017. [Online; accessed 30-Jan-2024].

Mathew Aneke and Meihong Wang. Energy storage technologies and real life applications–a state of the art review. *Applied Energy*, 179:350–377, 2016.

Kyle Bradbury, Lincoln Pratson, and Dalia Patiño-Echeverri. Economic viability of energy storage systems based on price arbitrage potential in real-time us electricity markets. *Applied Energy*, 114:512–519, 2014.

Jun Cao, Dan Harrold, Zhong Fan, Thomas Morstyn, David Healey, and Kang Li. Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model. *IEEE Transactions on Smart Grid*, 11(5):4513–4521, 2020.

Shin-ichi Inage. The role of large-scale energy storage under high shares of renewable energy. *Advances in Energy Systems: The Large-scale Renewable Energy Integration Challenge*, pp. 221–243, 2019.

Rhys Jacob, Maximilian Hoffmann, Jann Michael Weinand, Jochen Linßen, Detlef Stolten, and Michael Müller. The future role of thermal energy storage in 100% renewable electricity systems. *Renewable and Sustainable Energy Transition*, 4:100059, 2023.

Jaeik Jeong and Hongseok Kim. DeepComp: Deep reinforcement learning based renewable energy error compensable forecasting. *Applied Energy*, 294:116970, 2021.

Jaeik Jeong, Seung Wan Kim, and Hongseok Kim. Deep reinforcement learning based real-time renewable energy bidding with battery control. *IEEE Transactions on Energy Markets, Policy and Regulation*, 2023.

Hyuna Kang, Seunghoon Jung, Hakpyeong Kim, Jaewon Jeoung, and Taehoon Hong. Reinforcement learning-based optimal scheduling model of battery energy storage system at the building level. *Renewable and Sustainable Energy Reviews*, 190:114054, 2024.

Tai-Yeon Ku, Wan-Ki Park, and Hoon Choi. Energy maestro–transactive energy mechanism. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 6727–6729. IEEE, 2022.

Sangyoon Lee and Dae-Hyun Choi. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. *IEEE Transactions on Industrial Informatics*, 18(1):488–497, 2020.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Mostafa Rezaeimozafar, Maeve Duffy, Rory FD Monaghan, and Enda Barrett. A hybrid heuristic-reinforcement learning-based real-time control model for residential behind-the-meter PV-battery systems. *Applied Energy*, 355:122244, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shulei Zhang, Runda Jia, Hengxin Pan, and Yankai Cao. A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid. *Applied Energy*, 348:121490, 2023.

Matthieu Zimmer and Paul Weng. Exploiting the sign of the advantage function to learn deterministic policies in continuous domains. *arXiv preprint arXiv:1906.04556*, 2019.