Time-Varying Constraint-Aware Reinforcement Learning for Energy Storage Control

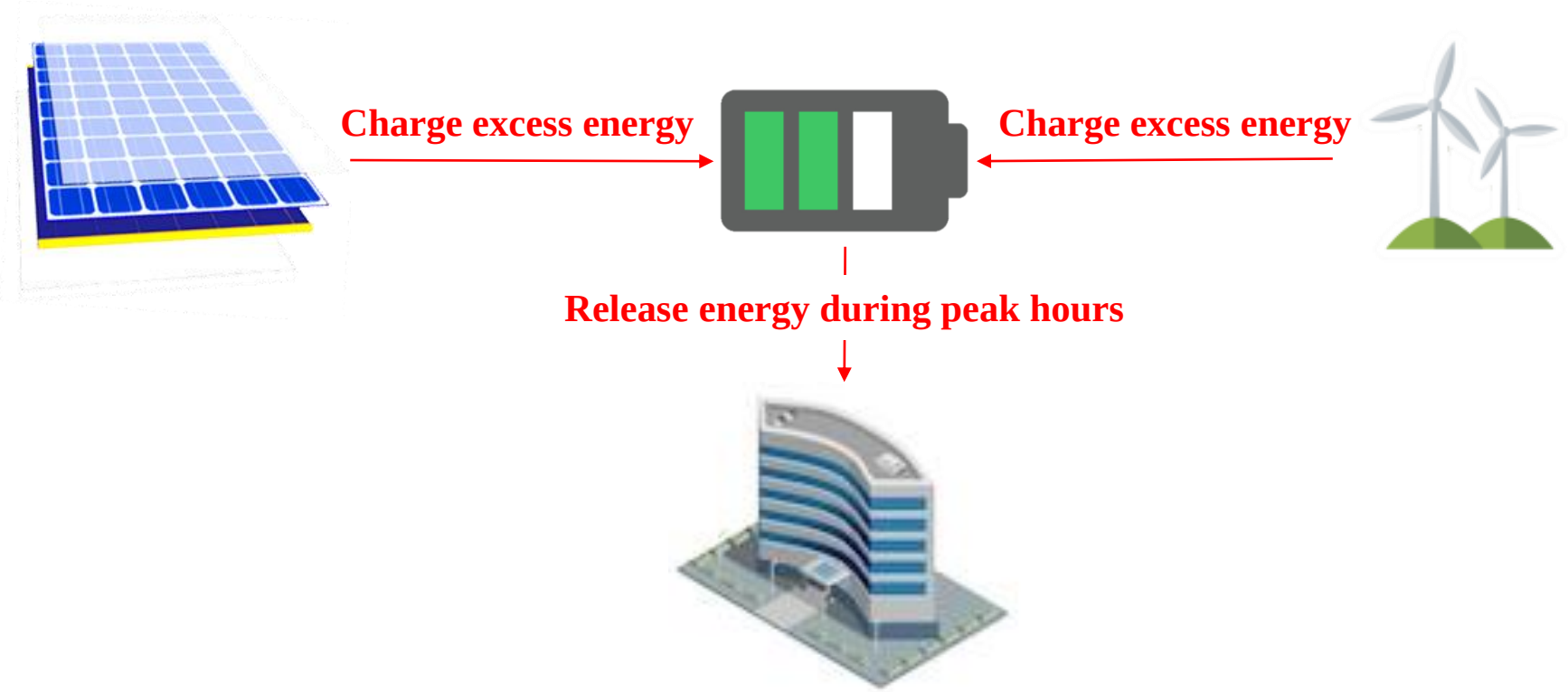
Tackling Climate Change with Machine Learning at ICLR 2024

Jaeik Jeong, Tai-Yeon Ku, Wan-Ki Park
Electronics and Telecommunications Research Institute (ETRI)

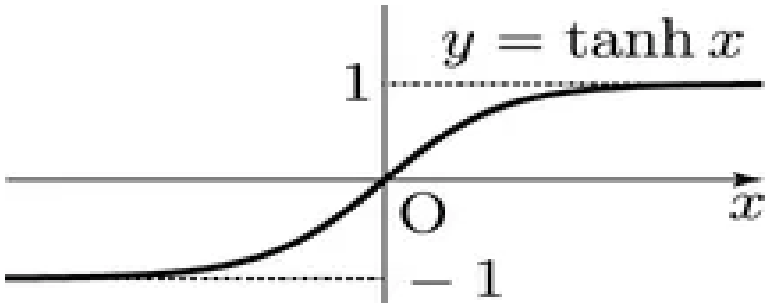
Presenter: Jaeik Jeong
2024. 05. 11.

ETRI
한국전자통신연구원

- Energy Storage and Climate Change
  - Energy storage and its smart integration are crucial for transitioning to decentralized and renewable energy production
    - Storing excess energy generated from renewable sources during peak production periods
    - Releasing stored energy during periods of high demand

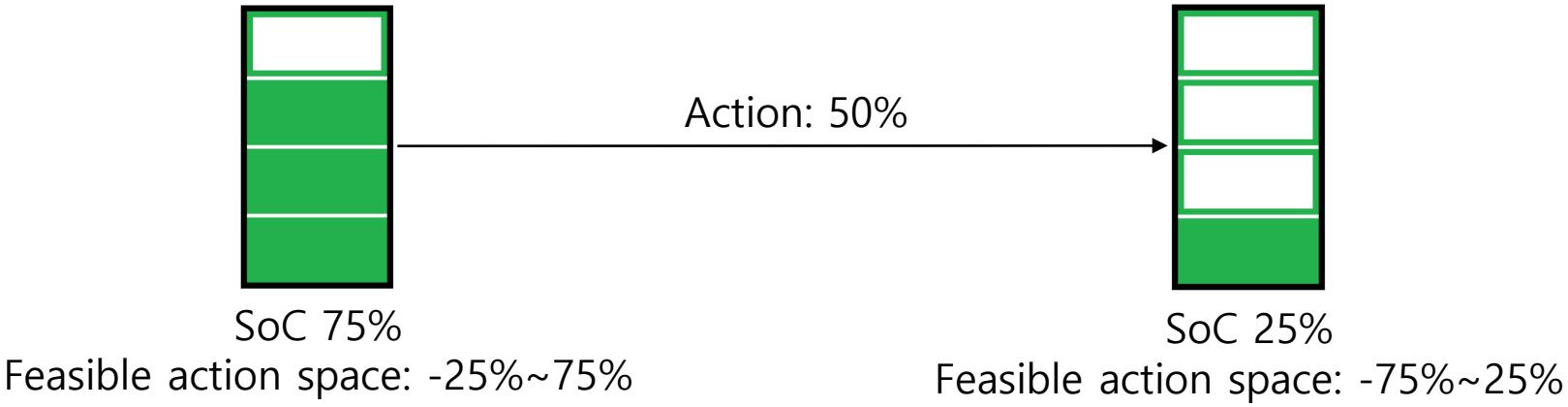


- Energy Storage and Continuous Reinforcement Learning
  - Reinforcement learning is useful in controlling energy storage in the presence of uncertainties such as renewable energy generation, energy demand, and energy prices
  - Control of energy storage can be interpreted as a problem of determining charging and discharging amounts over time
    - Since these amounts are continuous values, continuous reinforcement learning should be employed
  - In continuous reinforcement learning, when there's a need to constrain the range of actions, the set of actions that can be outputted is termed the feasible action space
    - For instance, if the possible range of action 'a' is from 'α' to 'β', then the feasible action space becomes '$\alpha < a < \beta$'.



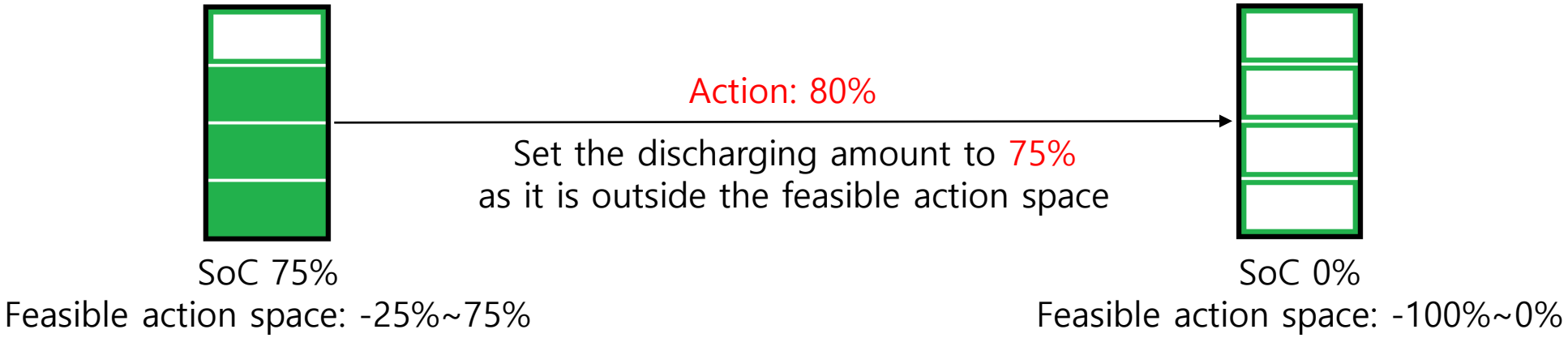
Tanh is useful when restricting the range of actions.
It only resolves the fixed feasible action space problem.

- Energy Storage Systems and Continuous Reinforcement Learning
  - However, in energy storage control, one must consider that the feasible action space continuously changes as the range of charging/discharging amounts varies over each time step
    - If the energy storage system is currently at 75% state-of-charge (SoC), the charging and discharging amounts should be determined within the range of -25% to 75%
      (charging would be represented as negative and discharging as positive)
    - As charging and discharging occur over time, SoC changes, leading to variations in the range of charging and discharging amounts (i.e., the feasible action space continuously changes)



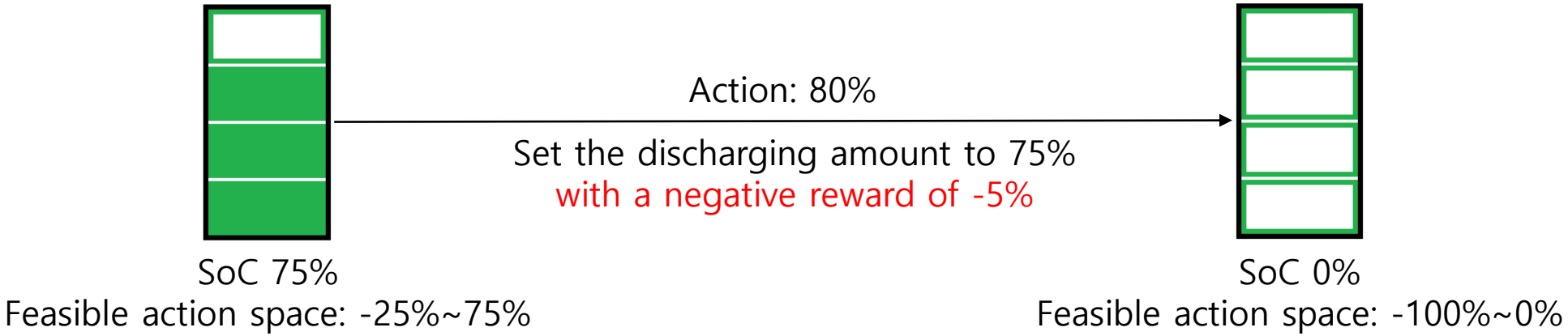
Time-Varying Feasible Action Space

- Traditional Solution for Changing Feasible Action Space
  - Enforcing actions to stay within the designated range
    - If the SoC is at 75% (the feasible action space ranges from -25% to 75%), the learning model adjusts the outputted action to either -25% if it falls below that threshold or 75% if it exceeds it
    - While this method easily resolves the feasible action space issue, it introduces instability in learning



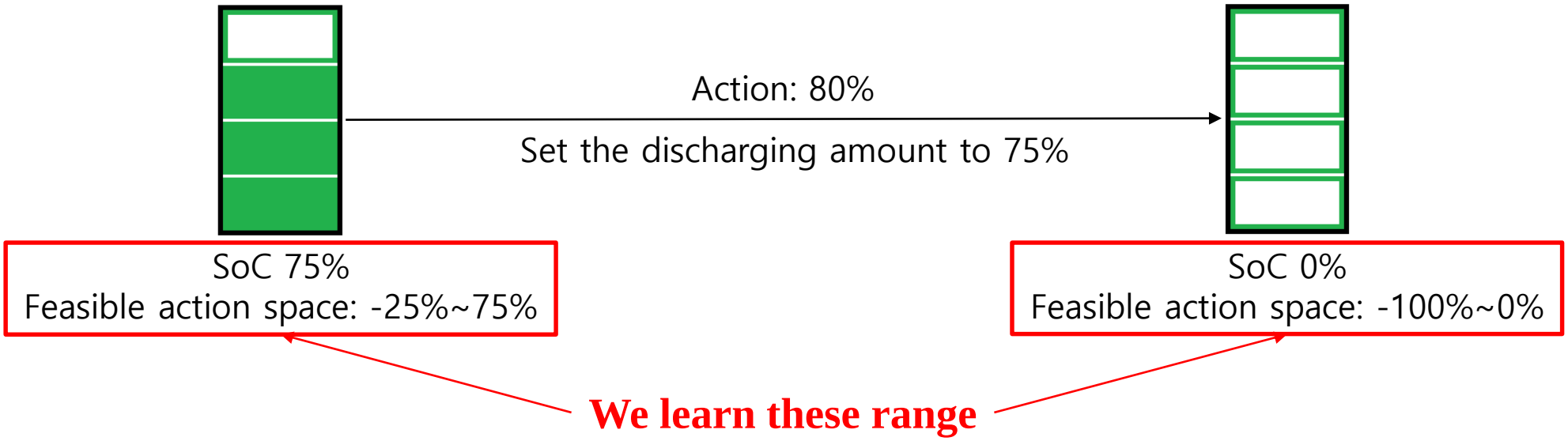
Learning Instability
(SoC may become stuck in a fully charged/discharged state during the learning)

- Traditional Solution for Changing Feasible Action Space
  - Assigning a cost or negative reward proportional to the extent by which actions deviate outside the designated range
    - If the SoC is at 75% (the feasible action space ranges from -25% to 75%) and action is 80%, the discharging amount is 75%, and there is a negative reward of -5%
    - However, there is a potential for overly conservative learning



Conservative Learning
(Emphasis leans heavily toward actions that remain within the designated range)

- Proposed Method
  - There is an additional supervising objective function designed to ensure that the output actions at each time step fall within the feasible action space
    - Prevent the energy storage from getting stuck in suboptimal states (fully charged/discharged)
    - Enabling the activation of charging/discharging operations in energy storage (facilitating the exploration of more optimal actions)



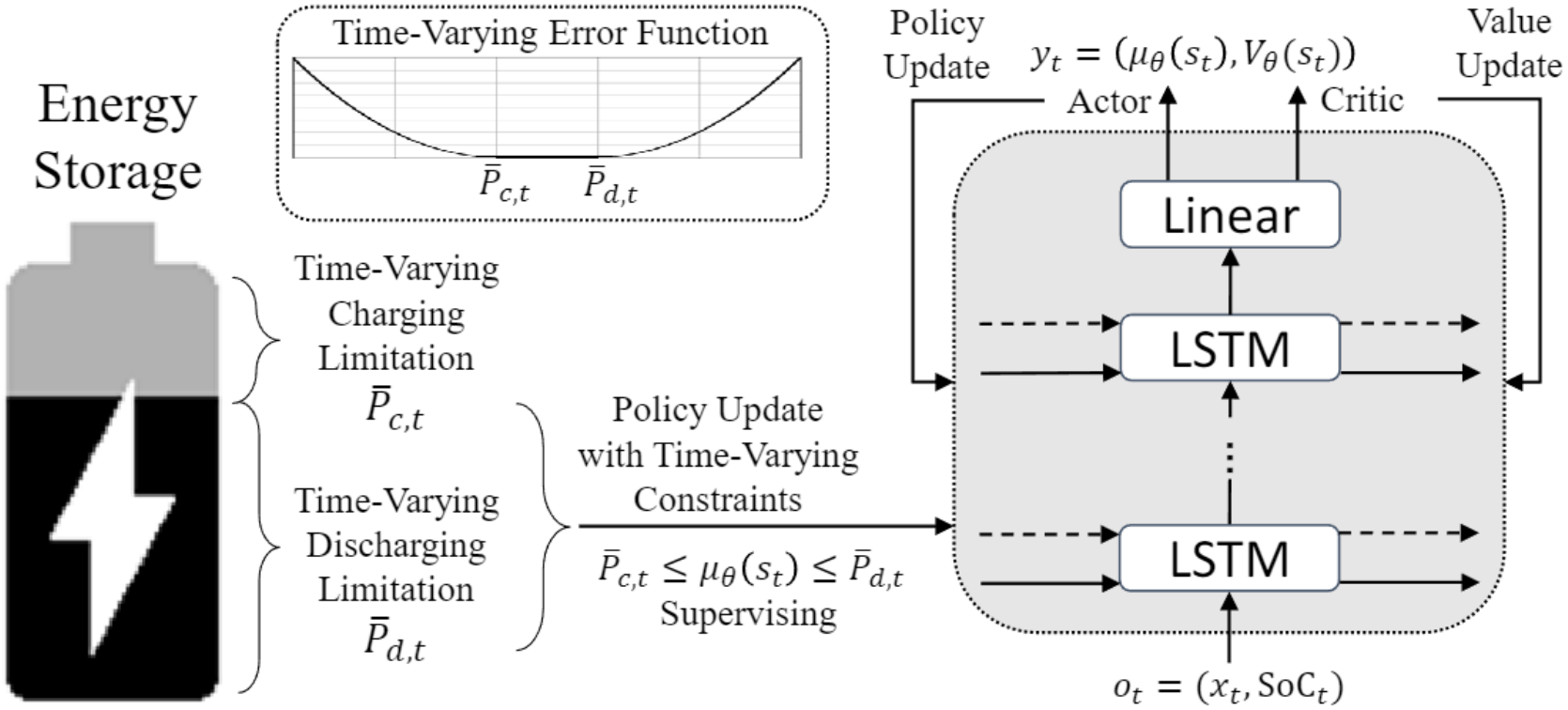
Methods

- Reinforcement Learning Settings
  - In energy storage control problems, the majority of states often involve time-series data
    - Such as SoC, energy generation, energy demand, and energy prices
  - We use a long short-term memory (LSTM) model to address the time-series data
  - We adopt the proximal policy optimization (PPO) algorithm,
    known for its compatibility with LSTM

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

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

- Actor and Critic
  - With the parameters θ and the state s_t, the output consists of the actor's output $\mu_\theta(s_t)$, and the critic's output $V_\theta(s_t)$.
  - During the training phase, action a_t is sampled from the Gaussian policy with mean $\mu_\theta(s_t)$, and during the actual testing phase, $\mu_\theta(s_t)$ serves as the action

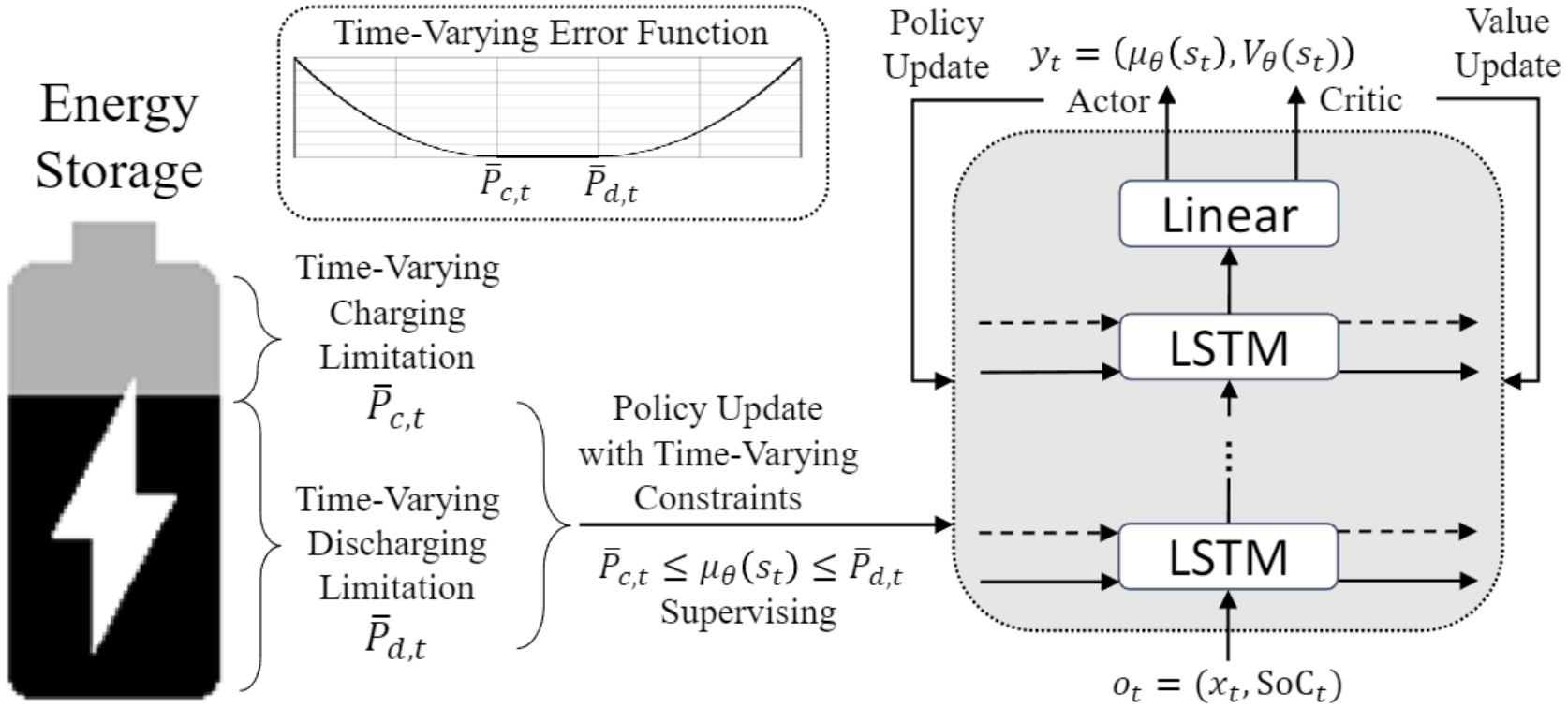


- Proposed Supervising Objective Function
  - An additional objective function was introduced for learning the feasible action range

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

  - We finally obtain our main objective

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



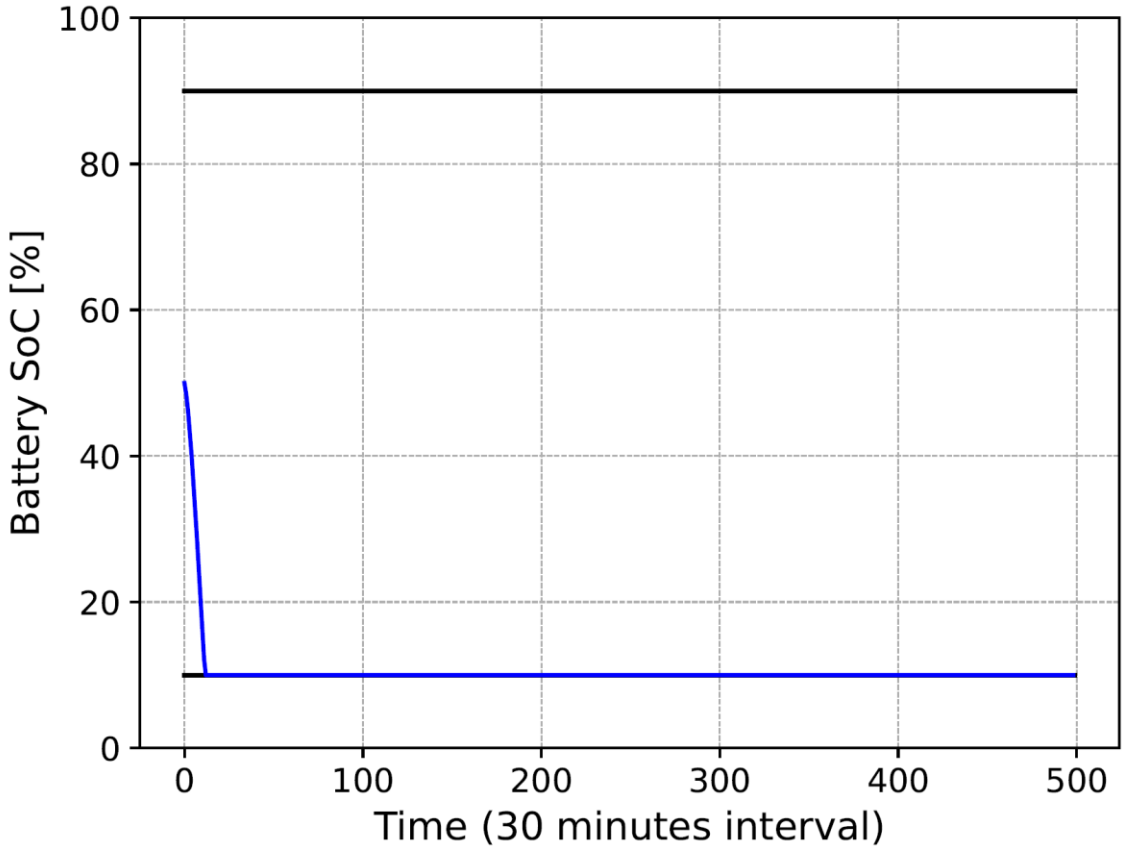
Results

- Experiment Settings
  - We demonstrated the effectiveness of the proposed approach through energy arbitrage experiments based on actual energy price data [1].
  - 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.
  - We set initial SoC=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.
  - 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 a degradation cost of \$10/MWh.

[1] 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.

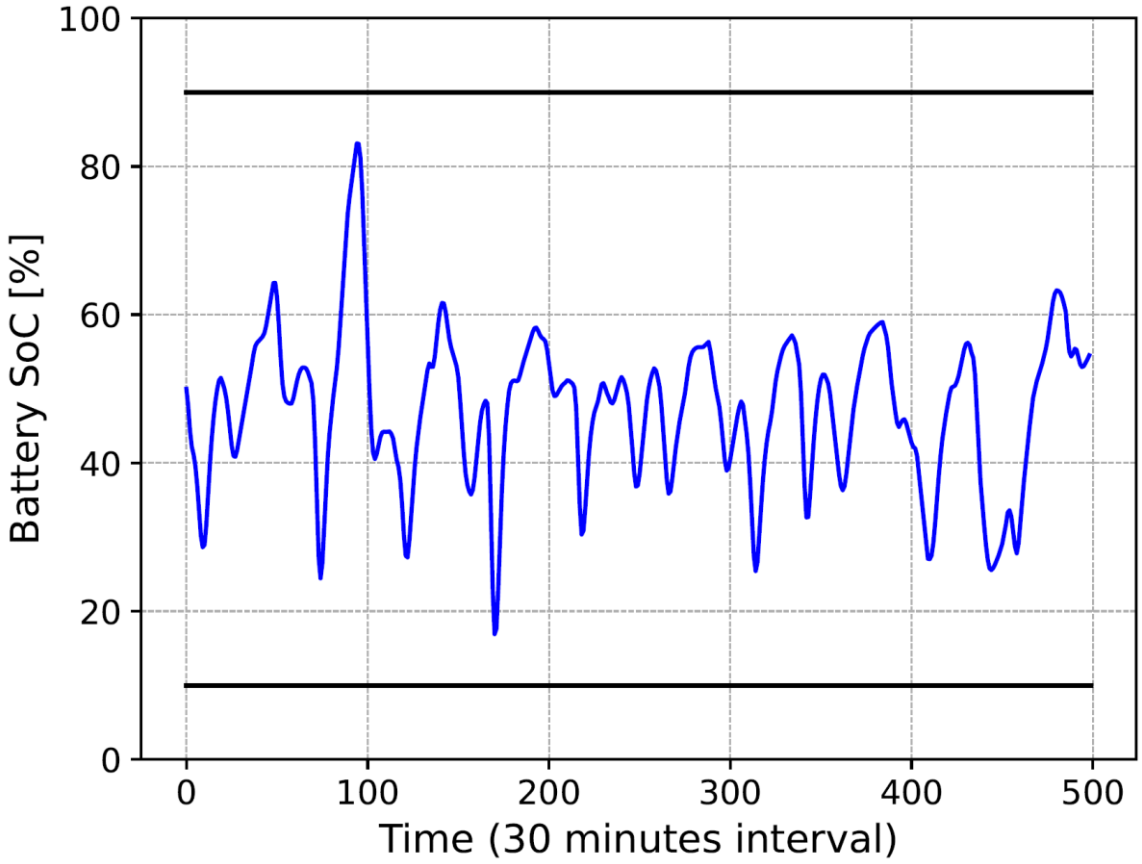
Results

- Case 1
  - Case 1 employed a conventional continuous approach, excluding the supervising equation.
  - It shows a scenario where all initial energy is discharged (sold out) with no further actions.
  - This suggests a failure in learning to manage the costs associated with charging in energy arbitrage, resulting in suboptimal behavior.



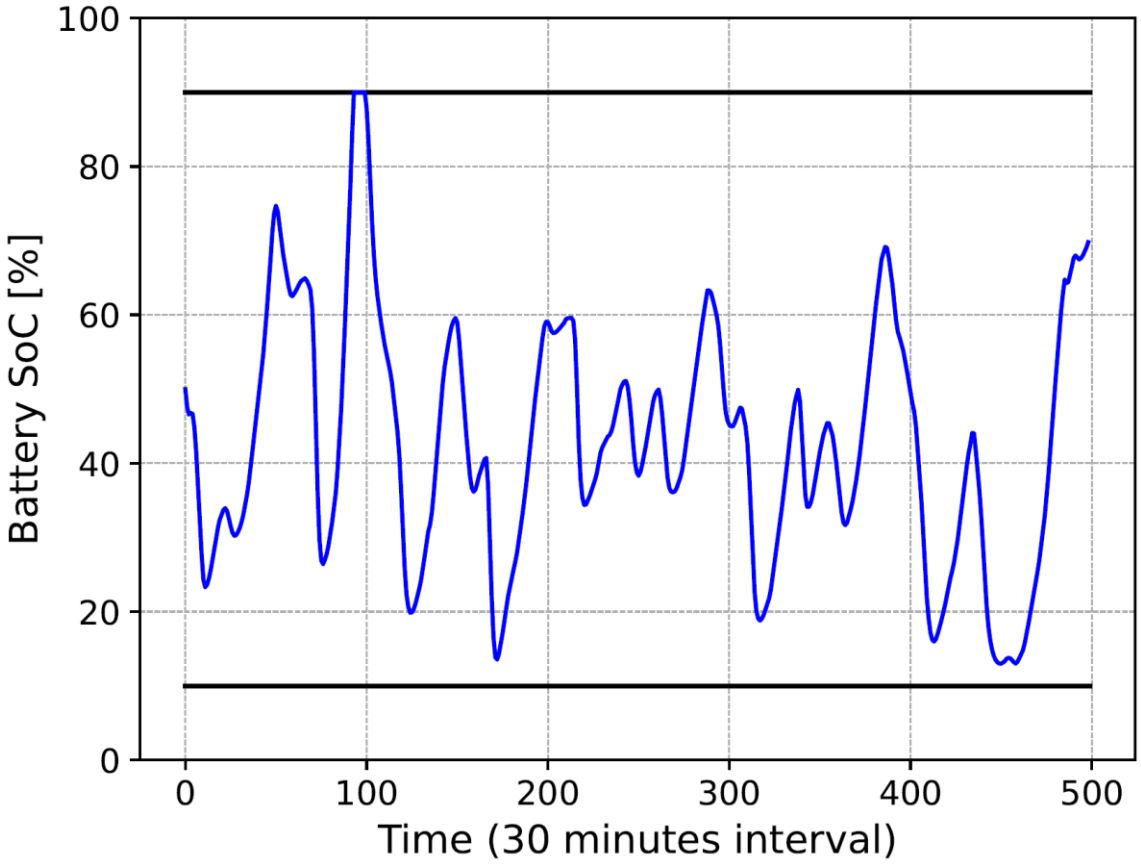
Results

- Case 2
  - Case 2 incorporated the supervising equation into the reward function.
  - This approach adds negative rewards if the output actions fall outside the feasible action space, rather than explicitly learning the range of output actions.
  - It demonstrates reasonable utilization of energy arbitrage.



Results

- Case 3
  - Case 3 designates the proposed model.
  - It engages in more active energy arbitrage.
  - Introducing supervising equation as a negative reward makes the agent conservative towards reaching complete charge/discharge, leading to reduced utilization of the energy storage.



Results

- Total Profit
  - The table below 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.

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**

Conclusion

- Conclusion
  - 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.