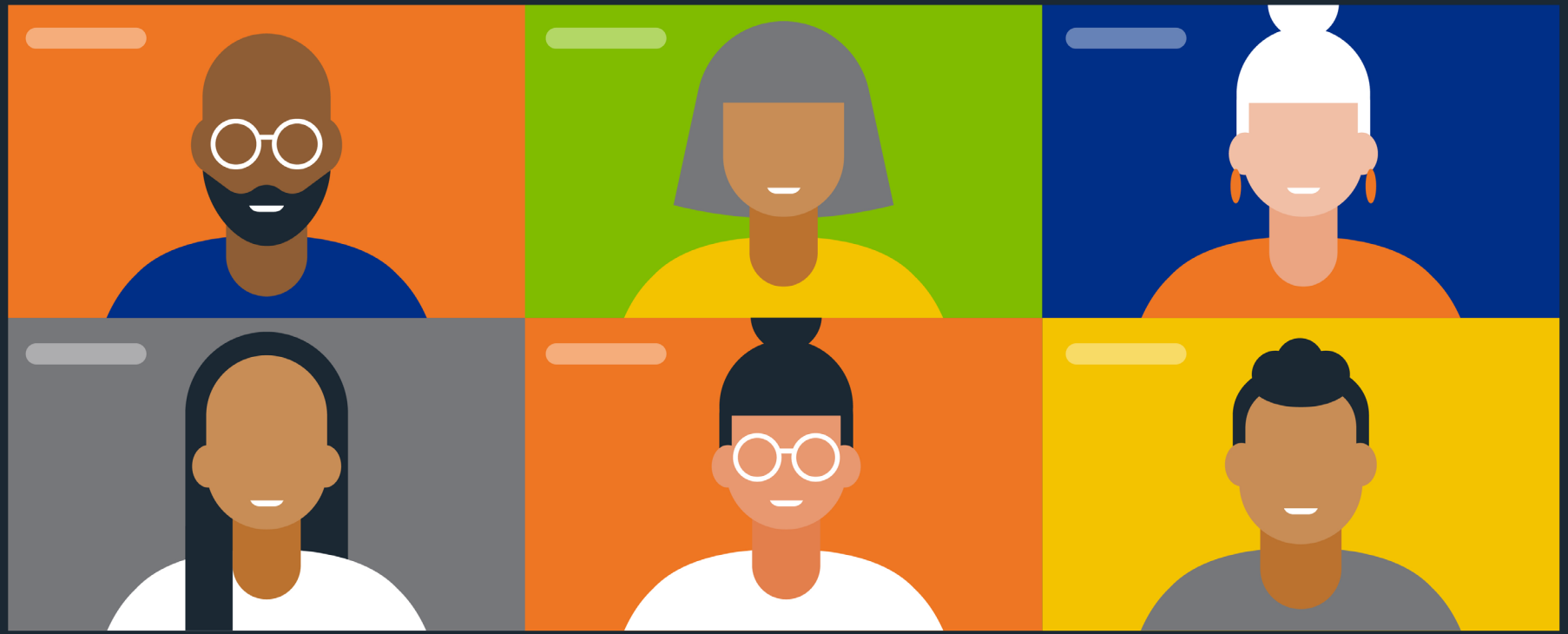


INFORMS ANNUAL MEETING | 2020 VIRTUAL

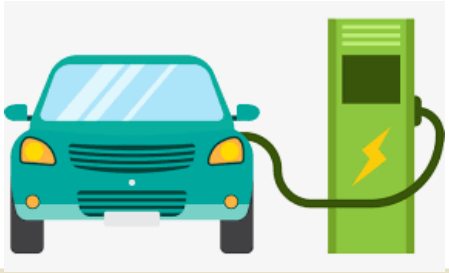
Capacity Scheduling of Battery Storage System for EV Charging and Frequency Regulation: A Proximal Policy Optimization Approach

Speaker: Bin Huang
Department of Electrical and Computer Engineering
Southern Methodist University

1. Overview
2. Safe Control Algorithm
3. Deep Reinforcement Learning Approach
4. Numerical Result
5. Concluding Remarks and Future Work

Background

- The appeal for the low-carbon future spurs the increasing integration of electric vehicles (EVs) to the power grid.
- This trend also brings great challenges to the stability and reliability of the operation of power grid.
- Through providing great flexibility and smoothing power fluctuation, battery storage systems (BSS) are proven to be a qualified solution to the large-scale integration of EV.



Background

- The investors-owned BSSs can be regarded as the independent entity to the power grid, and their ultimate goal is maximizing the revenue.
- Due to the prominent flexibility and fast-response feature, BSS can provide multiple services associated with multiple revenue streams, including peak shaving, reserve, energy arbitrage (EA), frequency regulation (FR), etc.
- By providing the stacked services, the owners of the BSS can exploit the whole capacity of the battery and earn extra profits.
- Fast Frequency Regulation Service, EV Charging, and Energy Arbitrage are considered.

Source: https://cleantechnica.com/2018/10/23/grid-scale-battery-storage-accelerating-in-colorado-australia/

Objective and Challenges

Objective

Develop the optimal scheduling strategies of the investors-owned BSS which performs stacked services.

- Dispatchable capacity is shared between the stacked services dynamically.

Given the complexity of the modelling stacked services, a practical scheduling strategy for BSS should be developed to coordinate the multiple services under strict safe constraints of the power and energy capacity of the battery.

Solution: safe control algorithm for BSS

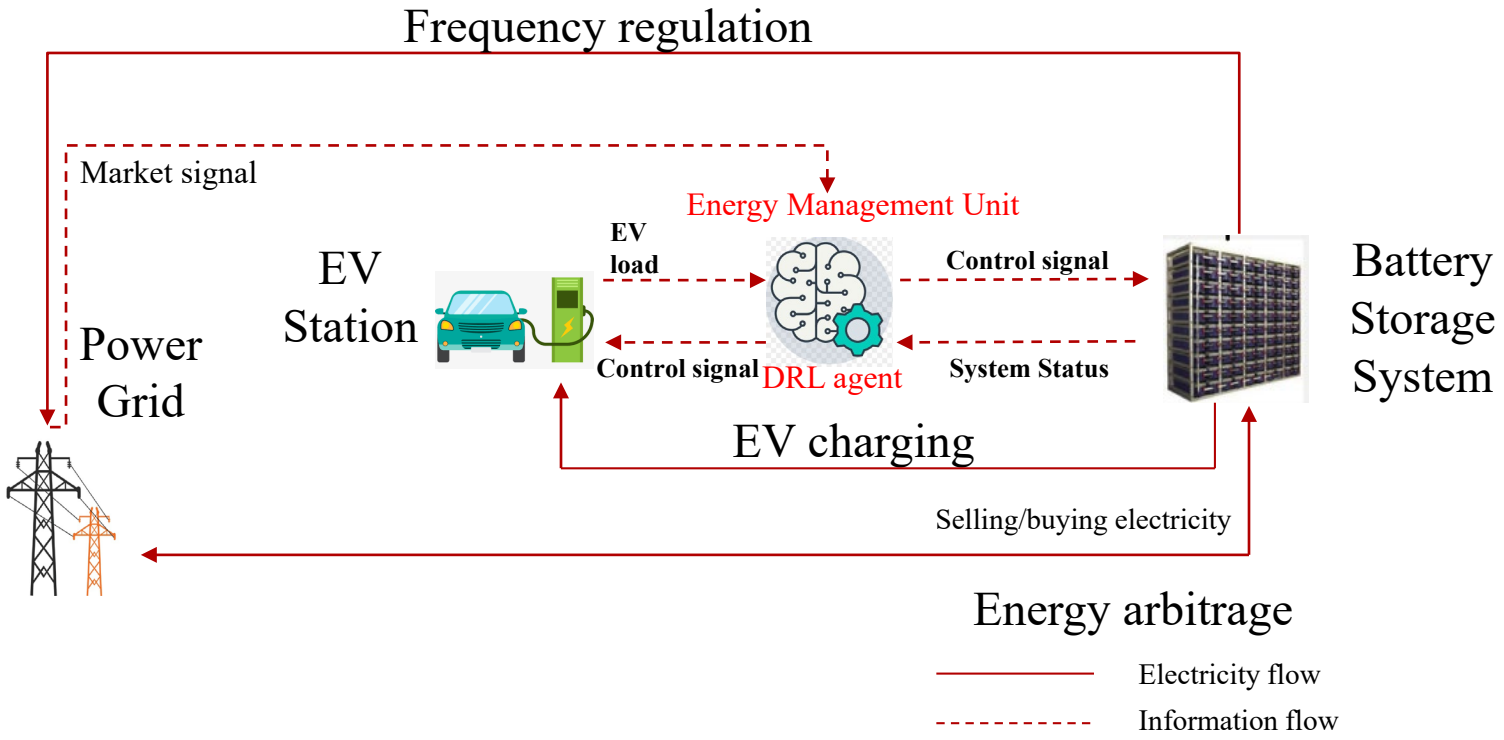
- Ubiquitous uncertainty during the operation.

Electricity market signals, and the power consumption of EV have inherently random features, which result in the battery state of charge (SOC) suffers from severe uncertainties. It is difficult for operators of the BSS to estimate a firm regulation capacity.

Solution: data-driven deep reinforcement learning approach

1. Overview
2. Safe Control Algorithm
3. Deep Reinforcement Learning Approach
4. Numerical Result
5. Concluding Remarks and Future Work

System Overview: Components and Functions

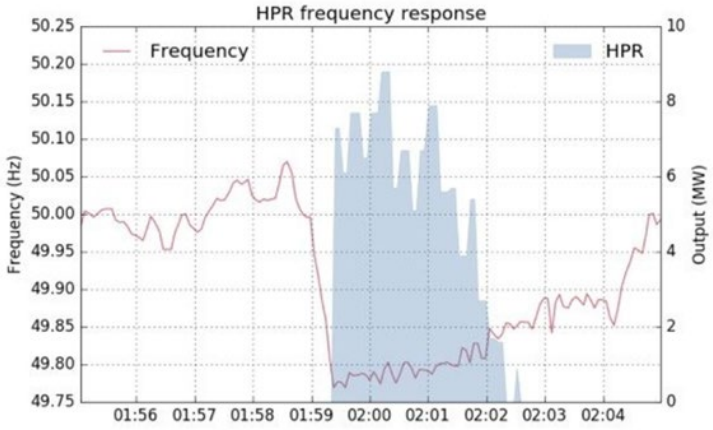


With the advantage of addressing sequential decision-making problem, the DRL agent is applied to perform the energy management task.

Stacked Services

Fast Frequency Regulation Service

- The frequency fluctuation in the power system is mainly caused by the mismatch between the generation and the load.
- Frequency regulation is a tool employed by power grid operators to prevent the system frequency getting too high or too low.



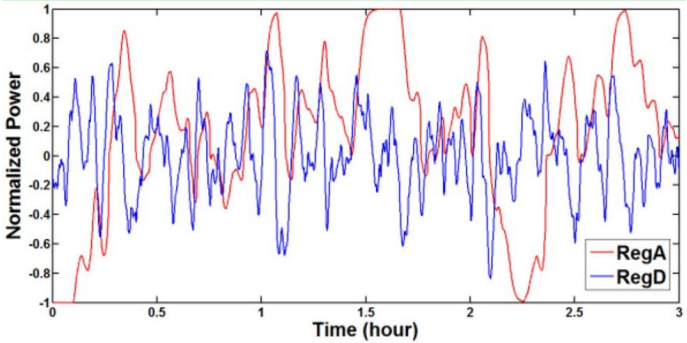
Source: reneweconomy.com

Fast Frequency Regulation Service

PJM market model is adopted.

PJM generates two different types of automated signals that regulation market resources can follow.

It can be observed from the figure that RegD signals fluctuates more frequently.



3-h sample profiles of RegA and RegD.

Reg A: a slower signal that is meant to recover larger and longer fluctuations in system conditions.

Reg D: a fast and dynamic signal that requires resources to respond quickly.

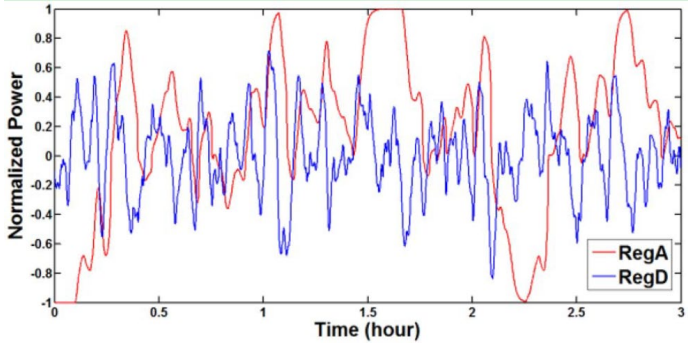
He, Guannan, Qixin Chen, Chongqing Kang, Pierre Pinson, and Qing Xia. "Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life." IEEE Transactions on Smart Grid 7, no. 5 (2015): 2359-2367.

Fast Frequency Regulation Service

BSS is an ideal regulation resource for following the RegD signals due to its fast response feature.

Good match!

RegD's favorable characteristic for BSS is that it requires net zero energy over a 15-min time period (energy neutral), which reduces the amount of obligated reserved energy of BSS.



3-h sample profiles of RegA and RegD.

Reg A: a slower signal that is meant to recover larger, longer fluctuations in system conditions.

Reg D: a fast, dynamic signal that requires resources to respond almost instantaneously.

He, Guannan, Qixin Chen, Chongqing Kang, Pierre Pinson, and Qing Xia. "Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life." IEEE Transactions on Smart Grid 7, no. 5 (2015): 2359-2367.

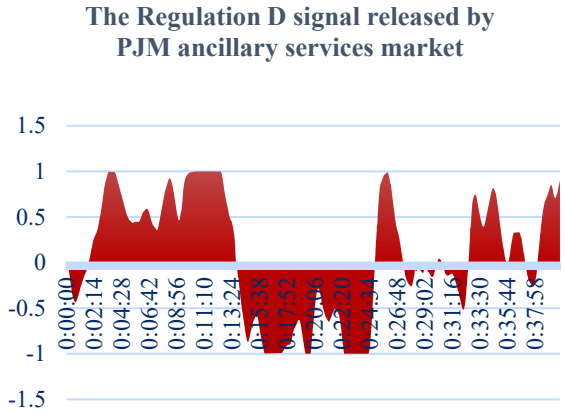
Fast Frequency Regulation Service

time-resolution: 4s converted to 1h

fractional frequency regulation signal (Reg D)

$$q_t = \sum_{j=1}^{J} \left(\frac{\delta_{j,t}^{+}}{\eta_{\text{dis}}} + \delta_{j,t}^{-} \eta_{\text{ch}} \right) \cdot \Delta t \quad (1)$$

q_t represents the cumulative fractional energy consumption in regulation in the hour t. q_t is determined by the actual RegD signal.



q_t is the absolute area enclosed by the RegD curve over 1 hour.

Data source: Byrne, R.H., Concepcion, R.J. and Silva-Monroy, C.A., 2016, July. Estimating potential revenue from electrical energy storage in PJM. In 2016 IEEE Power and Energy Society General Meeting (PESGM)

Fast Frequency Regulation Service

The revenue model of PJM is characterized by the introduction of mileage and the two-part payment based on the capacity bidding of the resources and the performance of the resources.

Bidding power capacity of regulation

Capacity Clearing Price

$$B_{\mathrm{f},t} = P_{\mathrm{f},t} \cdot \varphi \cdot (\lambda_t F_{\mathrm{p},t} + F_{\mathrm{c},t}) \quad (2)$$

Performance Clearing Price

$$\lambda_t = \frac{RegD\ \text{Mileage}}{RegA\ \text{Mileage}}$$

is the ratio between the RegD mileage (absolute summation of movement) to that of RegA. Because RegD's mileage is approximately three times that of RegA, the mileage ratio of RegD is approximately three times larger as well. Thus, qualified resources following RegD would earn three times the performance revenue.

φ performance score of the resources

Performance score reflects how well the resource is following the regulation signal.

Accuracy Delay Precision

He, Guannan, Qixin Chen, Chongqing Kang, Pierre Pinson, and Qing Xia. "Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life." IEEE Transactions on Smart Grid 7, no. 5 (2015): 2359-2367.

EV Charging

EV is modeled here as a random load. It can be partially or fully supplied by the BSS, depending on the decision of the BSS

From the angle of BSS, it can obtain the revenue by selling electricity to EV station.

The revenue from selling power to EV station is represented as:

$$B_t^{\mathrm{ev}} = P_t^{\mathrm{ev,s}} \cdot \Delta h \cdot F_t^{\mathrm{lmp}} \tag{1}$$

where B_t^{ev} is the remunerate of selling solar power; Δh is the time duration and is set to be 1 hour in this paper; F_t^{lmp} is the locational marginal pricing (LMP) at the energy market.

Priority among Stacked Services

Unlike conventional constrained optimization technique, the existing DRL frameworks enable the agent to explore the environment to develop the optimal policy by choosing the action freely during the learning process.

To enforce the constraints, conventional DRL algorithms are dependent on designing the heuristic reward function to guide the learning process of the agent.

The proposed safe control algorithm dispatches the power capacity according to the priority of the stacked services in sequence, thereby bypassing the design of the heuristic reward function.

Priority among Stacked Services

- The control algorithm is based on the following priorities: the power capacity deployed for FFRS is determined first, then EV charging, and finally EA. The rationale is given as follows.

Assume that $q_t < 0$ and $P_t^{\mathrm{EA}} > 0$ at hour t, the services which lead to the rising of the SOC include FFRS, EV charging, and EA. **Through performing the FFRS, the BSS can raise the SOC and get paid as well. In contrast, to raise the SOC, performing EA requires the BSS purchases electricity from the energy market.** As an intermediate, performing EV charging can raise the SOC at no expense. In order to maximize the benefits, EMU should give priority to FFRS, i.e., determine P_t^{f} first. Afterward, based on the remaining upward space of power capacity, EMU determines $P_t^{\mathrm{ev,s}}$. Lastly, P_t^{EA} is designed based on the available upward space of power capacity.

Similar argument can be derived when $q_t > 0$ and $P_t^{\mathrm{EA}} < 0$.

Safe control algorithm

Herein the α, β, ξ are the EV charging, EA, and FFRS ratio coefficient, respectively, which represent the control policy of the EMU. The bound for α and ξ is $[0, 1]$, whereas the bound for β is $[-1, 1]$, where the negative interval represents performing EA by selling electricity.

$\mathrm{SOC}_t^{up} = \overline{SOC} - \mathrm{SOC}_t$, $\mathrm{SOC}_t^{dn} = \mathrm{SOC}_t - \underline{SOC}$
Performing frequency regulation;
if $q_t \leqslant 0$ **then**
  $\xi_{max} = \frac{\mathrm{SOC}_t^{up} U}{-P_{max}^{up} q_t}$
  $\xi_t = \mathrm{clip}(\xi_t, 0, \xi_{max})$
  if $\xi_t < \frac{P^{f,min}}{P_{max}^{up}}$ **then**
    $P_t^f = 0$
  else
    $P_t^f = \xi_t P_{max}^{up}$
  end if
  $\mathrm{SOC}_t^{up} \leftarrow \mathrm{SOC}_t^{up} + \frac{P_t^f q_t}{U}$, $P_{max}^{up} \leftarrow P_{max}^{up,0} - P_t^f$
else
  $\xi_{max} = \frac{\mathrm{SOC}_t^{dn} U}{-P_{max}^{dn} q_t}$
  $\xi_t = \mathrm{clip}(\xi, 0, \xi_{max})$
  if $\xi_t < \frac{P^{f,min}}{-P_{max}^{dn}}$ **then**
    $P_t^f = 0$
  else
    $P_t^f = -\xi_t P_{max}^{dn}$
  end if
  $\mathrm{SOC}_t^{dn} \leftarrow \mathrm{SOC}_t^{dn} - \frac{P_t^f q_t}{U}$, $P_{max}^{dn} \leftarrow P_{max}^{dn,0} + P_t^f$
end if
$B_t^f = P_t^f \cdot \varphi_t \cdot (\lambda_t F_t^{PCP} + F_t^{CCP})$

U Energy capacity
SOC State of Charge

- Derive the threshold ξ_{max} to prevent the over-charging and over-discharging.
- Dispatch P_t^f based on P_{max}^{up} or P_{max}^{dn}
- Update the upward/downward space of SOC; Update the maximum upward/downward available power capacity.

Safe control algorithm

The control logic for allocating power for charging EV is similar to that of frequency regulation. The main difference is that EV load will only lead to the discharging behavior of battery. The electricity flow here is only single direction.

- Derive the threshold $\alpha_{\max}$ to prevent the over-discharging
- Dispatch charging power to EV based on the EV load and α, which is a ratio coefficient between 0-1 determined by EMU.
- Update the downward space of SOC; Update the maximum downward available power capacity

Safe control algorithm

Performing energy arbitrage;

if $\beta_t \geqslant 0$ **then**

$$\beta_{max} = \frac{SOC_t^{up} U}{P_{max}^{up} \Delta h \cdot \eta_{ch}}$$

$$\beta_t = clip(\beta_t, 0, \beta_{max}),\ P_t^{EA} = P_{max}^{up} \cdot \beta_t \quad \text{Purchase electricity}$$

$$P_{max}^{up} = P_{max}^{up} - P_t^{EA}$$

else

$$\beta_{max} = \frac{SOC_t^{dn} U}{-P_{max}^{dn} \cdot \Delta h / \eta_{dis}}$$

$$\beta_t = clip(\beta, 0, \beta_{max}),\ P_t^{EA} = -P_{max}^{dn} \cdot \beta_t \quad \text{Sell electricity}$$

$$P_{max}^{dn} = P_{max}^{dn} - P_t^{EA}$$

end if

$$B_t^{EA} = -P_t^{EA} \cdot \Delta h \cdot F_t^{lmp} \quad \text{revenue}$$

$$Cost_t = c \cdot [|P_{max}^{dn,0} - P_{max}^{dn}| + (P_{max}^{up,0} - P_{max}^{up})] \quad (1) \quad \text{cost}$$

$$SOC_{t+1} = SOC_t - \frac{P_t^f q_t}{U} + \frac{P_t^{pv,e} \Delta h \cdot \eta_{ch}}{U} + \frac{[sgn(\beta)]^+ P_t^{EA} \cdot \Delta h \cdot \eta_{ch}}{U} + \frac{[sgn(-\beta)]^+ P_t^{EA} \cdot \Delta h}{\eta_{dis} U} \quad (2)$$

where sgn is the sign function and $[\ \ ]^+$ is the rectified linear unit (ReLU) function. The adoption of these two function serves as the logic expression: when $\beta_t > 0$, the fifth term of (2) become zero; when $\beta_t < 0$, the fourth term of (2) become zero.

1. Overview
2. Safe Control Algorithm
3. Deep Reinforcement Learning Approach
4. Numerical Result
5. Concluding Remarks and Future Work

Markov Decision Process

Considering the random nature of the PV generation and market signals, and the time-coupled feature of the SOC, the capacity scheduling for PV-BSS is essentially a discrete-time stochastic control process, which can be described as a Markov decision process (MDP).

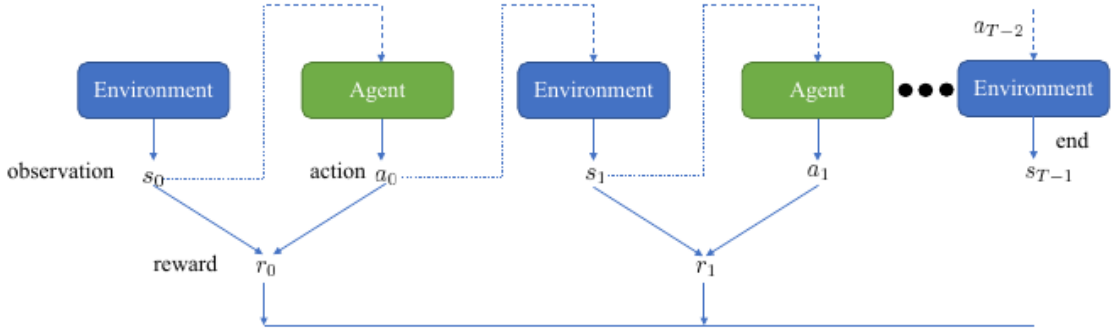


Figure 2: A Markov Decision Process and Trajectory.

Markov Decision Process

The state vector is represented as:

$$\mathbf{S} = [SOC_t^{\mathrm{up}}, SOC_t^{\mathrm{dn}}, F_{t-2}^{\mathrm{lmp}}, F_{t-1}^{\mathrm{lmp}}, F_t^{\mathrm{lmp}}, q_t, Bonus_{t-2}, Bonus_{t-1}, Bonus_t, P_{t-2}^{\mathrm{ev}}, P_{t-1}^{\mathrm{ev}}, P_t^{\mathrm{ev}}] \tag{1}$$

where $Bonus_t = \varphi_t \cdot \left(\lambda_t F_t^{\mathrm{PCP}} + F_t^{\mathrm{CCP}}\right)$.

The action vector is represented as $\mathbf{a} = [\alpha_t, \beta_t, \xi_t]$ with the continuous action space.

The reward function is defined as:

$$r_t = B_t^{\mathrm{EA}} + B_t^{\mathrm{ev}} + B_t^{\mathrm{f}} - Cost_t$$

where r_t also represents the net profit of the PV-BSS at hour t.

Proximal Policy Optimization

- Proximal policy optimization (PPO) is the variant of Trust Region Policy Optimization (TRPO) and Advantage Actor-Critic (A2C).
- PPO can guarantee the safe exploration of the agent while avoiding solving the complicated second-order optimization problem in TRPO.
- Compared with A2C, PPO is much more sample-efficient.
- Compared with deep Q networks(DQN), PPO is more appropriate in dealing with the continuous and multi-dimensional action space.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Proximal Policy Optimization

The **discounted value function** $V^{\pi,\gamma}$ is an indicator to the value of the state through evaluating the impact of state upon the return under policy π:

$$V^{\pi,\gamma}(s_t) := \mathbb{E}_{s_{t+1:\infty}}\left[\sum_{l=0}^{\infty}\gamma^l r_{t+l}\right]$$

Action-value function takes the value of action into account:

$$Q^{\pi,\gamma}(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty} \\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty}\gamma^l r_{t+l}\right]$$

The difference between $Q^{\pi,\gamma}(s_t, a_t)$ and $V^{\pi,\gamma}(s_t)$ is denoted as the **advantage function**: $A^{\pi,\gamma}(s_t, a_t) = Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t)$. The intuition toward $A^{\pi,\gamma}(s_t, a_t)$ is that it measures how much an action is better than others on average, which represents the relative advantage of the action.

$\pi_\theta(a|s)$ is the **policy** of the DRL agent with parameters θ, which represents the actor network in PPO.

$V_\phi(s_t)$ is the **value function** of the DRL agent with parameters ϕ, which represents the critic network in PPO.

Proximal Policy Optimization

$$\hat{J}^{\mathrm{PPO}} = \max_{\theta} \mathop{\mathbb{E}}_{s,a\sim\pi_{\theta_{\mathrm{old}}}} \left[\mathbb{L}\left(s, a, \theta_{\mathrm{old}}, \theta\right)\right]$$ Surrogate clipped objective function

$$\mathbb{L}\left(s, a, \theta_{\mathrm{old}}, \theta\right) = \sum_{(s_t, a_t)} \min\left(\rho_t A^{\pi_{\theta_{\mathrm{old}}}}\left(s_t, a_t\right), \mathrm{clip}\left(\rho_t, 1-\epsilon, 1+\epsilon\right) A^{\pi_{\theta_{\mathrm{old}}}}\left(s_t, a_t\right)\right) \quad (1)$$

$$\rho_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)} \quad (2)$$ Measure how far the current policy is from the previous policy

If advantage is positive: Suppose the advantage for that state-action pair is positive, in which case its contribution to the objective reduces to

$$L\left(s, a, \theta_{\mathrm{old}}, \theta\right) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\mathrm{old}}}(a|s)}, (1+\epsilon)\right) A^{\pi_{\theta_{\mathrm{old}}}}\left(s, a\right) \quad (3)$$

the advantage is positive ⇒ $\pi_\theta(a|s)$ ↑

if $\pi_\theta(a|s) > (1+\epsilon)\pi_{\theta_{\mathrm{old}}}(a|s)$ ⇒ $L\left(s, a, \theta_{\mathrm{old}}, \theta\right) = (1+\epsilon) A^{\pi_{\theta_{\mathrm{old}}}}\left(s, a\right)$ (4)

if $\pi_\theta(a|s) \leq (1+\epsilon)\pi_{\theta_{\mathrm{old}}}(a|s)$ ⇒ $L\left(s, a, \theta_{\mathrm{old}}, \theta\right) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)} A^{\pi_{\theta_{\mathrm{old}}}}\left(s, a\right)$

Stochastic Diagonal Gaussian Policy

Assume that the action is a random variable, the output of the policy network π_θ are assumed to be the expected value of the actions.

When the PPO agent attempts to determine the action based on the observation, it depends on the sampling the actions from the multivariate normal distribution:

$$a = \mu_\theta(s) + \sigma_\theta(s) \odot \mathbf{x} \tag{1}$$

where $\mathbf{x}$ is the samples vector of a standard multivariate normal distribution: $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$; μ and σ are the mean and standard deviation of the action vector;

As for the stochastic diagonal Gaussian policies, $\pi_\theta(a_t|s_t)$ can be derived as:

$$\pi_\theta(a_t|s_t) = \frac{1}{(2\pi)^{k/2} \prod_i^k \sigma_i} \exp(-\sum_{i=1}^{k} \frac{(x-\mu_i)^2}{2\sigma_i^2}) \tag{2}$$

Training Scheme

1. Initialization of the actor network and critic network.
2. Let agent interact with the environment and collect the set of trajectories samples.
3. Advantage estimation based on the current value function.
4. Update the parameters of the policy network

$$\hat{J}^{\mathrm{PPO}} = \max_{\theta} \mathop{\mathbb{E}}_{s,a\sim\pi_{\theta_{\mathrm{old}}}} \left[\mathbb{L}\left(s, a, \theta_{\mathrm{old}}, \theta\right)\right]$$

$$\mathbb{L}\left(s, a, \theta_{\mathrm{old}}, \theta\right) = \sum_{(s_t, a_t)} \min\left(\rho_t A^{\pi_{\theta_{\mathrm{old}}}}\left(s_t, a_t\right), \mathrm{clip}\left(\rho_t, 1-\epsilon, 1+\epsilon\right) A^{\pi_{\theta_{\mathrm{old}}}}\left(s_t, a_t\right)\right)$$

5. Update the value function network

$$\phi_{new} = \arg\min_{\phi} \frac{1}{T} \sum_{t=0}^{T-1} \left(V_{\phi}\left(s_t\right) - \hat{R}_t\right)^2$$

The steps 2, 3, 4, and 5 will repeat N epoch.

1. Overview
2. Safe Control Algorithm
3. Deep Reinforcement Learning Approach
4. Numerical Result
5. Concluding Remarks and Future Work

Test Parameters

- Case studies are conducted based on the real-world data from PJM energy and regulation market in 2018.
- Both PPO and A2C use multilayer perceptron (MLP) with two hidden layers as the policy network and value function network, respectively.
- The scheduling cycle is one week (168 h) in the case studies. Thus, the trajectory length is 168. The market data in 2018 are split into training and testing set: the first nine months are training set, and the rest three months are testing set. For each epoch, 12 trajectories are collected to update the PPO and A2C agents.

Test Parameters

Parameters	Value
S_0	0.5
ϕ	0.95
η_{dis} / η_{ch}	0.9/0.9
$\underline{S}$ / $\overline{S}$	0.1/0.9
U	30 MWh
c	0.5$/MW
P^{dn}_{max} / P^{up}_{max}	-10MW/10MW
Δh / Δt	1h / 4s

Table 1: Parameters of the PV-BSS

Parameters	Value
ϵ	0.2
γ	0.91
λ	0.97
lr_π / lr_V	5.7e-4/1.2e-7
N_π / N_V	80/80
KL^{max}	0.015
$\log \sigma_\theta$	-0.6

Table 2: Hyperparameters of the PPO

Performance of PPO

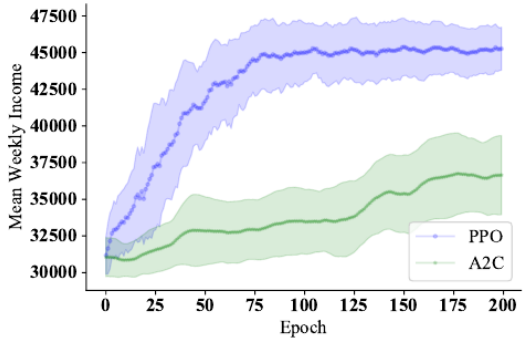


Figure 3: Training process of PPO and A2C.

Learning performance index: average weekly income

In terms of profitability, PPO outperforms A2C and random policy by 50.6% and 23.5%, respectively.

PPO uses only about 36*12*168 = 72576 experience samples to achieve the optimal performance of A2C, while A2C needs 175*12*168 = 352800 experience samples.

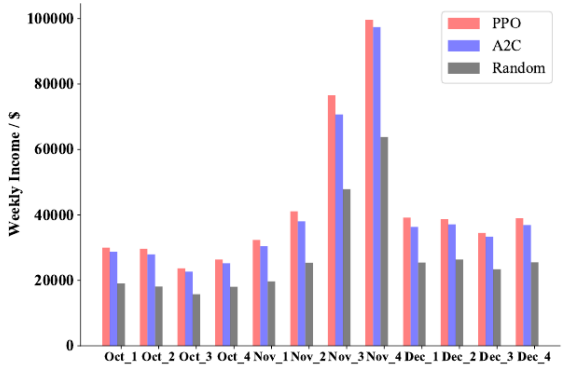


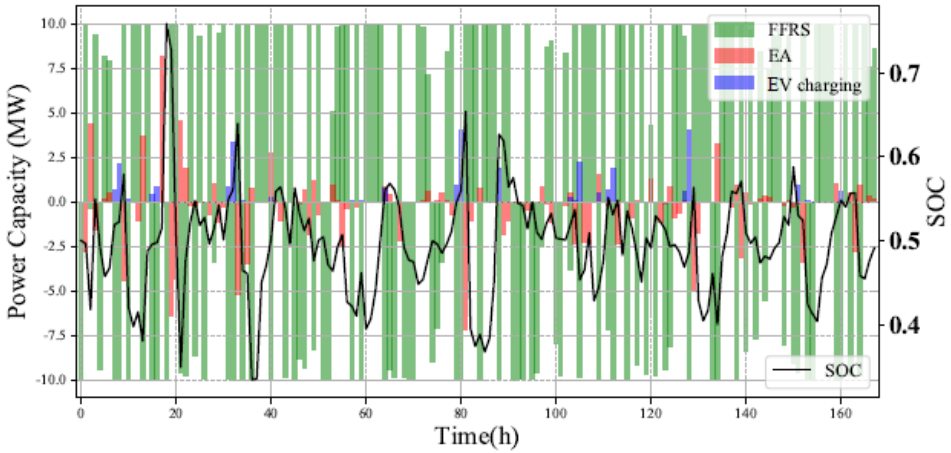
Figure 4: The revenue on the testing data.

The performance upon the testing dataset can demonstrate the generalization of the DRL agent

PPO agent is dominant over the A2C agent and random agent.

PPO agent: $510,582
A2C agent: $484,779
Random agent: $328,462

Capacity Scheduling Result



- SOC curve of the battery is within the safe range over the whole scheduling cycle.
- The power capacity deployed to the stacked service is within the safe constraints.
- The power capacity of BSS deployed for performing FFRS is dominant over all other services most of the time.
- The trained PPO agent consumes only 78 ms to make the scheduling decision for a week.

Capacity Scheduling Result

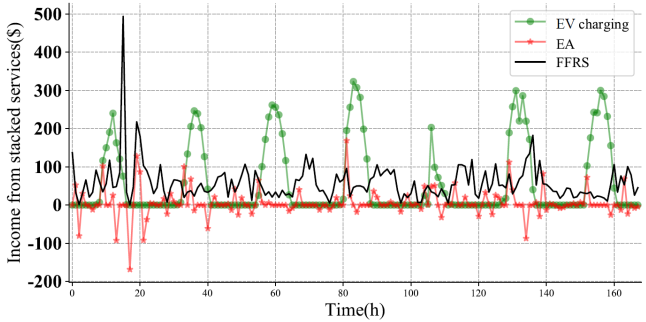


Figure 8: Power capacity deployed for EA and LMP price in the first week in January.

Net profit: $21036
selling power to EV station: $9610.83
EA: $834.30
FFRS: $9729.48
deprecation cost: -$861.43

Most of the time, the PPO agent choose to purchase electricity at a relatively low price and sell at a relatively high price.

1. Overview
2. Safe Control Algorithm
3. Deep Reinforcement Learning Approach
4. Numerical Result
5. Concluding Remarks and Future Work

Concluding Remarks and Future Work

- This research proposes a pragmatic solution to the capacity scheduling of BSS, which performs the stacked services.
- A safe control algorithm of BSS is proposed to ensure the safe operation of the PV-BSS.
- The PPO-based DRL agent is developed to cooperate with the control algorithm to improve the profitability of PBSS.
- Case studies based on the real data of PJM energy and regulation market are conducted. The results demonstrate that the PPO agent with the proposed safe control algorithm is capable of generating the safe scheduling schemes while maximizing the net profit of BSS.
- Comparative results demonstrate the superior performance of PPO agent to A2C agent in terms of optimization results and sample efficiency.
- A more detailed and practical battery degradation model should be taken into consideration because the excessive usage of the battery caused by providing the stacked services will impair the lifespan of the battery.