# Optimal Scheduled Control Operation of Battery Energy Storage System using Model-Free Reinforcement Learning

Alaa Selim
School of Engineering and Information Technology,
University of New South Wales, Canberra, Australia.
email: a.selim@adfa.edu.au

2022 IEEE Sustainable Power and Energy Conference (iSPEC) | 978-1-6654-8522-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/iSPEC54162.2022.10033035

***Abstract*—Driven by the tremendous increase in rooftop solar panels and battery installations in Australian states, several studies have been conducted to efficiently manage battery operations with the imported grid power through battery energy storage systems (BESS). Therefore, it is crucial for the BESS to carefully decide the power set-points of the installed batteries to maintain user comfort while operating household appliances. Additionally, BESS should be capable of reducing the electricity bills by optimally managing the battery operation to sustain times of higher tariff prices for the imported grid power. This paper formulates the scheduled operation of the BESS as a Markov decision Process (MDP) that enables the BESS to figure out numerous scenarios and decides the optimal power set-points for both batteries and grid power. A model-free reinforcement learning approach is proposed to manage the batteries' power-sharing and grid operation set-points to solve this MDP problem. This approach utilizes the advantages of the Deep Deterministic Policy Gradient (DDPG) algorithm to decide the shared power set-points every 5 minutes interval for the day ahead operation of the BESS. Finally, the proposed model is trained and validated using historical data of the Australian National Electricity market to offer an optimal scheduled control pattern for the daily BESS operations.**

***Index Terms*—Battery energy storage system, Control optimization, State of charge, Power sharing, Model-free, Reinforcement learning.**

## I. Introduction

CONTROL and management of battery energy storage systems (BESS) is an important area of research, with several advanced techniques being developed to improve their efficiency, longevity, and integration with the power grid. These techniques include adaptive charging [1], power smoothing [2], and smart grid algorithms [3], among others. Continued research in this field is crucial for the development of sustainable and reliable energy systems. Another important aspect of controlling battery energy storage systems is managing their interaction with other components of the power grid. This includes coordinating the charging and discharging of the battery with the operation of grid operators, to maximize their efficiency and minimize their impact on the grid. Researchers have developed algorithms for this purpose, such as "smart grid" algorithms that use real-time data to optimize the operation of the power grid.

Deep Reinforcement Learning (DRL) is currently applied as one of the best smart grid algorithms for grid modernization and solving convex optimization problems [4]. DRL differs from these methods, where it can learn decisions by interacting with an environment and adjusting its actions according to feedback. Similar to closed-loop feedback control, the DRL agent works to tune its actions based on the reward defined, instantaneous state value, and its current action. The agent keeps tuning its actions and saving its experience with the environment in a replay buffer medium or in the form of weight updates for the trained deep networks. One of the key parameters in the training phase of the DRL agent is the selection of state variables to monitor. In BESS problems, parameters like tariff prices, load profiles, and battery capacity can provide useful information for the agent, which can determine the best action to take in the present state. Consequently, we start monitoring them in a real-time manner for better training of the DRL. As training progresses, the DRL agent learns to take actions in an exploratory way for the state variables and environment to reach the global optimal value and avoids the sub-optimal results [5].

DRL agents have introduced many promising solutions for the BESS control problem with dynamic pricing environments as in [6]. Starting with the discrete action control, it can be performed by Deep-Q-learning network (DQN) DRL agents as in [7] to control the charging and discharging of batteries by setting discrete samples for each one. Further, in [8], the DQN method was significantly improved to be reliable for deciding BESS decision variables, including uncertainties and mitigating overestimation of the standard Q-learning approach. In [9], the proposed DRL algorithm is used for controlling ESS to arbitrage in real-time electricity markets under price uncertainty and learns the proper stochastic control policy for BESS control. Similar to [10] and [11], an online energy scheduling DRL controller provides real-time feedback to consumers to encourage more efficient electricity use and the DRL is used to learn the optimal bidding strategy from the dynamic environment of the Australian electricity market In our work, we are trying to focus on using continuous control actions using deterministic policy gradient DRL approach that can execute continuous actions for optimizing BESS

operations and learns to respond to the variable tariff pricing controlled by the Australian distribution companies.

This paper proposes a learning-based approach that controls BESS and grid supplies for achieving the optimal energy management, extending batteries life-time and reducing electricity bills. The main contributions are summarized as follows:

- BESS control problem is reformulated for considering continuous control actions to achieve a globally optimal solution that minimizes the daily electricity bills and maximizes the battery longevity.
- A deterministic policy gradient-based DRL approach is used for allocating power-sharing set-points for installed batteries and grid supply to provide a scheduled control pattern for BESS daily operation.
- The proposed approach is tested with dynamic tariff prices for the Australian electrical market including uncertainties to show a significant reduction in electricity bills for residential consumers by almost 40 percent.

## II. Problem Formulation

The problem formulation is based on deducing a scheduled operation for BEES and grid supply to obtain a highly optimized performance for the energy dispatch and maintain the economical added-value of BESS. In this section, the optimization problem is formulated as follows:

$$\min_{\alpha_g,\alpha_b} \sum_{t=0}^{T} \alpha_g.T_t^g.P_t^g + \sum_{t=0}^{T} \alpha_b.P_t^b \tag{1}$$

Subjected to

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

$$0 < \alpha_b < 1 \tag{3}$$

$$0 < \alpha_g < 1 \tag{4}$$

$$\alpha_g + \alpha_b \leq 1 \tag{5}$$

$$P_t^b + P_t^g \geq P_t^{di} \tag{6}$$

$$P_t^{b,min} < P_t^b < P_t^{b,max} \tag{7}$$

$$P_t^{g,min} < P_t^g < P_t^{g,max} \tag{8}$$

$$SOC_{t+1} = SOC_t + \Delta T \left(\alpha_b P_t^b\right) \tag{9}$$

The main objective function shown in (1) aims to minimize the grid power imported (i.e., minimizing tariff prices) and also minimize the rate of energy dispatched by battery for a longer life cycle. This can be achieved by controlling $\alpha_g$ and $\alpha_b$, which provided the percentage of power-sharing between the grid and batteries respectively. There is a trade-off relationship here for these two variables, where you can not minimize both locally as they depend on each other. For example, if you are trying to minimize the grid share by setting $\alpha_g$ to 1, this will result in $\alpha_g$ being 0 and thus will violate the energy balance constraints.

State variables for the studied horizon of the system are identified as follows: $P_t^b$ is the power dispatched by the energy storage system within storage limits of $P_t^{\text{b,max}}$ and $P_t^{\text{b,min}}$. $P_t^g$ is the power supplied by the grid within the $P_t^{\text{g,max}}$ and $P_t^{\text{g,min}}$ limits. $T_t^g$ is the instantaneous tariff price, $P_t^d$ is the total power demand for the system,$P_t^{pv}$ is the rooftop solar power and $P_t^{unc.}$ is the power needed to balance uncertainties in generation and demand. The idea of solving this formulation is to find the optimal ratio of power-sharing, which can result in finding the minimum global optimal value for this objective function. Finally, at each time step of this objective function, we experience an entirely different state, that makes the problem behaves as the the framework of the Markov Decision Process (MDP).

### A. System Modeling

The proposed control algorithm is based on the conceptual model shown in Fig. 1. The proposed controller can be model-based or model-free, where it depends on the model plant of the batteries and grid. In our work, we will follow the model-free approach, which makes the model can be generalized for any residential user and not limited to the technical specs of the model-based controller. The controller is assumed to collect data on a real-time basis every 5 minutes through communication modules attached to each studied device. For reading and writing data, it is assumed that the controller is connected to the smart inverter, where it can control power set-points for the battery dispatching and thus use the grid power to complement the total load demand. This assumption elaborates on how we can physically control power sources through the smart BESS unit. Consequently, the agent decides the optimal power-sharing set points between batteries and the grid supply.

## III. Deep Reinforcement Learning-Based Control of BESS

BESS management problem is first cast into MDP. The system experiences different conditions at each time step, represented by the state vector. Actions are taken at each time step based on the updated state from the last time step and with respect to the boundaries of the system input actions. Deep Deterministic Policy Gradient (DDPG) [12] is proposed for solving the MDP problem. This approach is a model-free off-policy algorithm for learning continuous actions. It combines ideas from DPG (Deterministic Policy Gradient) and DQN (Deep Q-Network). It uses experience replay and slow-learning target networks from DQN, and it is based on DPG, which can operate over continuous action spaces.

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

$$S_t = [P_t^{di}, T_t^g, P_t^d] \tag{10}$$

**Actions Space** $a_t$: the set of actions, the controller can execute at each time step. Also, it determines the continuous operational set-points based on the available power. Formally, we have

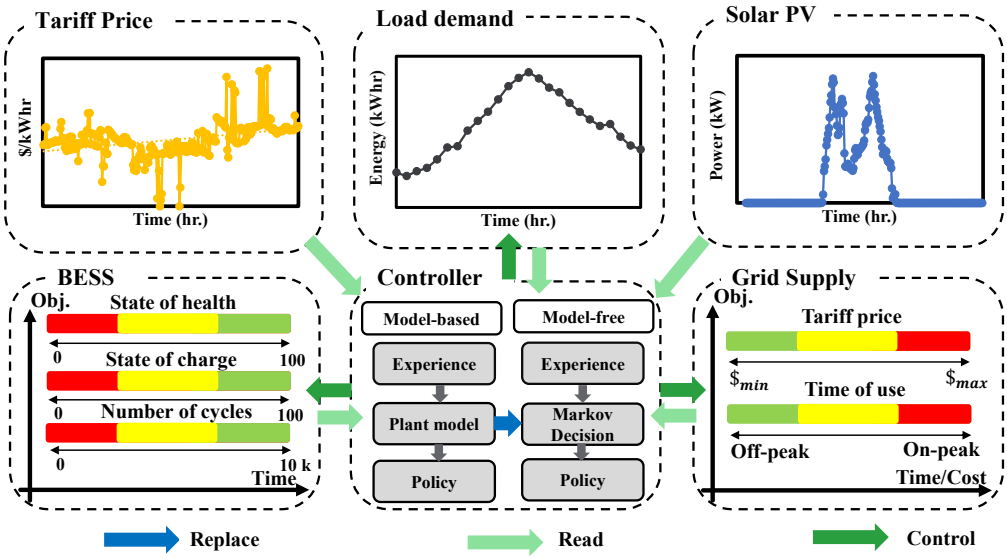


Fig. 1. Conceptual model for smart BESS.

$$a_t = [\alpha_b, \alpha_g] \tag{11}$$

$\alpha_{\mathrm{g}}$and $\alpha_{\mathrm{b}}$ are in scale from 0 to 1 and are used to control sharing between grid and batteries respectively as mentioned in the problem formulation section.

**Reward Function**: The reward function is formulated based on the problem formulation in Section II and then tuned based on the agent solutions to make sure it is not achieving only sub-optimal solutions. .

$$\text{Max} \sum_{t=0}^{T} r\left(s_t, a_t\right) \tag{12}$$

$$\sum_{t=0}^{T} r = -(\eta.\alpha_{\mathrm{g}}.T_t^g.P_t^{di} + \omega.\alpha_{\mathrm{b}}.P_t^{di}) \tag{13}$$

where $\eta$, $\omega$ are the reward coefficients that determine the rewarding and penalization scale. In this problem, we fix these coefficients to be 10 for $\eta$ and $\omega$ This choice was considering these coefficients' impact on the DRL performance for achieving higher reward value with faster convergence [13]. Selecting these coefficients is done by trial and observation for the DRL training results to check if it converges to the optimal solution or gets the problem diverged by picking wrong action values. Additionally, 2nd term in the reward function is found to have more weight compared to the 1st term to guide the DDPG agent in prioritizing a higher power-sharing percentage for batteries.

To enable the DDPG in solving the BESS problem, We need to randomly initialize the state variables, critic-network $Q\left(s, a \mid \theta^Q\right)$ and actor-network $\mu\left(s \mid \theta^\mu\right)$ with weights $\theta^Q$ and $\theta^\mu$. In DDPG, the actor and critic networks are updated based on the policy, which is determined during the training phase to deduce the final target network.$Q'$ and $\mu'$ with weights $\theta^{Q'} \leftarrow \theta^Q$,$\theta^{\mu'} \leftarrow \theta^\mu$. In (14), the initial return $y_t$ of the critic network value is computed based on the initial reward values $r_t$ and state $s_{t+1}$. The critic network then tries to minimize its loss function in (15) based on the $N$ transitions during the training phase, that are stored in the replay buffer. Then we update the actor policy in (16) using the sampled policy gradient method in [12] and update the $\theta^{Q'}$and $\theta^{\mu'}$ as in (17) and (18).

TABLE I
PARAMETER SETTINGS FOR DQN TRAINING

| Parameters | Value |
|---|---|
| Replay buffer size | 50000 |
| Batch size | 64 |
| discount factor ($\gamma$) | 0.99 |
| tau factor ($\tau$) | 0.005 |
| Learning rate for actor-network | 0.00001 |
| Learning rate for critic network | 0.00002 |
| Number of hidden layers for actor-network | 2 |
| Number of nodes for actor-network | [256,256] |
| Number of hidden layers for critic-network | 2 |
| Number of nodes for critic-network | [256,256] |
| Activation function | ReLU |
| Maximum number of episodes | 3000 |

For action space in DDPG, we can apply noise known as Ornstein-Uhlenbeck process for better exploration of possible MDP scenarios.

$$y_t = r_t + \gamma Q'\left(s_{t+1}, \mu'\left(s_{t+1} \mid \theta^{\mu'}\right) \mid \theta^{Q'}\right) \tag{14}$$

$$L = \frac{1}{N} \sum_t \left(y_t - Q\left(s_t, a_t \mid \theta^Q\right)\right)^2 \tag{15}$$

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q\left(s, a \mid \theta^Q\right)\Big|_{s=s_t, a=\mu(s_t)} \nabla_{\theta_\mu} \mu\left(s \mid \theta^\mu\right)_{s_t} \tag{16}$$

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \tag{17}$$

$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{18}$$

Before deploying the DDPG agent, the OpenAI Gym environment [14] has been developed and registered with the name (Batterycontrol-v1), which defines actions, states, and reward functions. Additionally, this environment defines the system constraints and boundaries we mentioned in Section II.

## IV. Numerical Results

This section provides the simulation results for codes developed using PyTorch library on a laptop computer with 3.6GHZ Intel i7 processor and 32.0 GB RAM. The proposed control methods are used to demonstrate the idea of allocating the power-sharing percentage by continuous actions allocated for the battery system and the grid for the real input model obtained from the Australian retailer of AUSGRID in Fig. 2. For verification and validation, the test is carried out for a 24-hour operation of Solar PV, grid, and batteries. DDPG agent is applied for solving the objective function in (1) to deduce the scheduled pattern of battery and grid operation through allocating the optimal set-points of $\alpha_g$ and $\alpha_b$. The key point in the control optimization algorithm is to achieve the balance for the energy storage that keeps batteries for later use when tariff prices reach higher values. Additionally, we want to find the trade-off relation in the power-sharing percentage between energy storage and grid supply. The environment is coded based on the DRL parameters defined in section III. However, reward function is tuned several times during the exploration phase as the agent diverges in terms of the optimal solution. The trick is that action space limits need to be changed instead of 0 and 1 to 0.5 and 0.7 for batteries 0.3 and 0.5 for the grid. Also, adding penalizing terms for violating these limits. Using this trick, we can force the agent to search within a more bounded search space instead of applying the extreme limits for the objective function. After clearly updating the environment, the training is executed for 13 hours and 36 minutes to come up with the learning curve shown in Fig. 3. This curve indicates the complexity of the solution each episode the DRL tries to solve. However, the reward value is significantly improved for this training time. Due to the time complexity burden, we can use supercomputers in future work to reach a more saturated stage of the learning curve. The power-sharing percentages are finally deduced as shown in Fig. 4 and Fig. 5, which show an intelligent response obtained by the agent. In times of higher tariff prices and on-peak pricing schemes, the higher power-sharing percentage goes to battery to reach 80 percent and grid power share is significantly reduced to almost 20 percent. On the contrary, the grid supply shows a higher share of power supply in times of early morning and late night hours. Finally, we obtain the power set-points for each side as shown in Fig. 6 and Fig. 7 to have a new scheduled control pattern of load power-sharing between storage and the grid. This pattern manages to reduce the electricity bills by almost 40 percent to reach 7132.40 AUD instead of 11141.64 AUD for a daily rate obtained by the conventional BESS unit control.

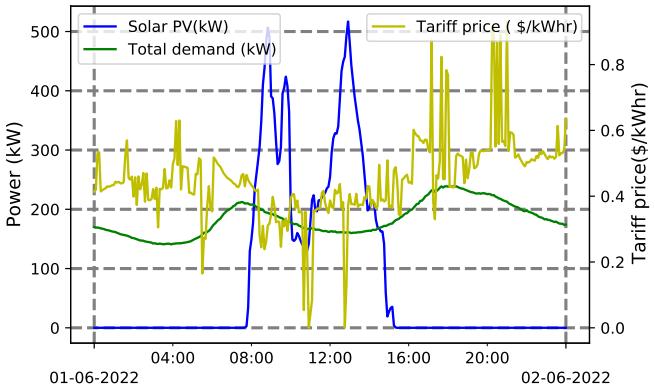


Fig. 2. Input model for the BESS system

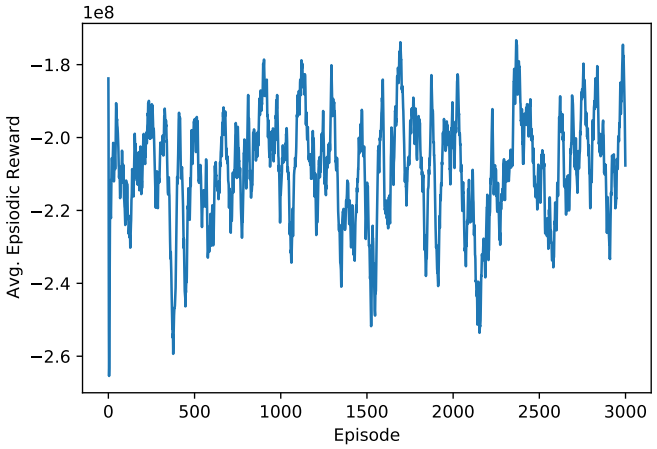


Fig. 3. Learning curve of the DDPG agent

## Conclusion

This paper proposed a model-free reinforcement learning approach to guide the BESS unit in deciding the power-sharing percentages between the available energy storage and the grid supply. The BESS control problem was reformulated considering the control of batteries and grid supply, yielding a complicated convex optimization problem. The proposed DDPG algorithm could learn the optimal control pattern for solving this optimization problem to introduce a scheduled pattern for battery operation, which reduced the electricity bill by almost 40 percent for daily residential customers.

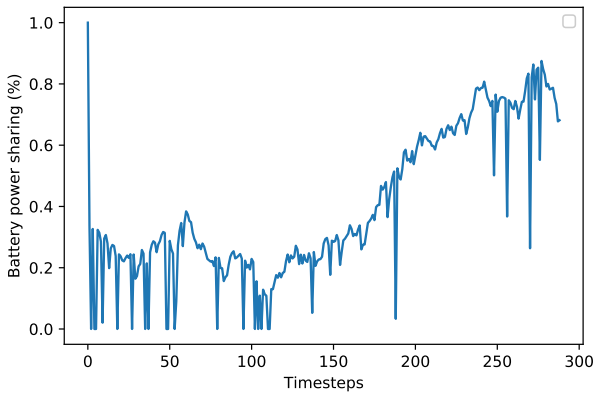


Fig. 4. DDPG optimization results-battery power-sharing setting

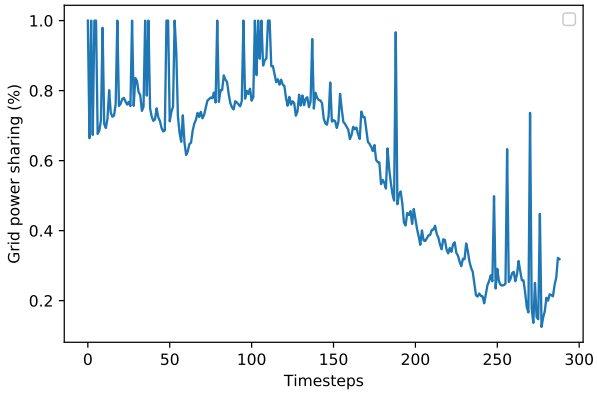


Fig. 5. DDPG optimization results-grid power-sharing settings

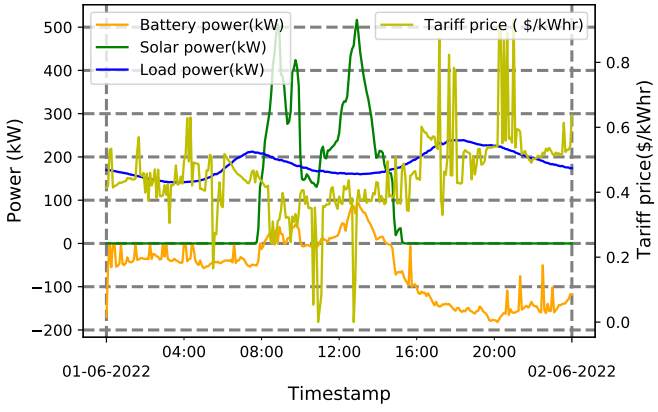


Fig. 6. DDPG scheduled control pattern for battery

In future work, we will apply the proposed algorithm for different pricing schemes in the international market to judge the robustness of the proposed approach. Also, we will use supercomputers for the training phase of the DRL agent for faster and better results of the final computed reward function.

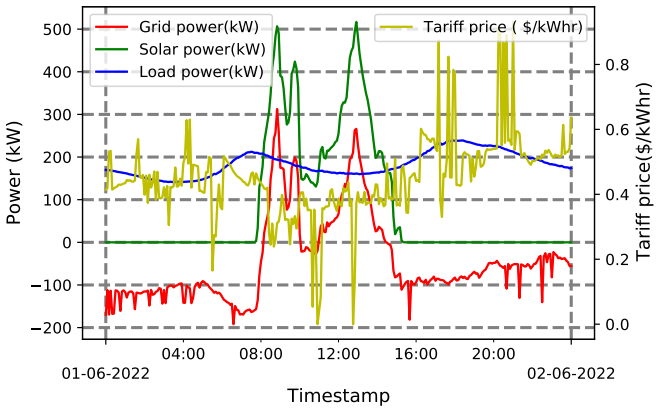


Fig. 7. DDPG scheduled control pattern for grid supply

## REFERENCES


[1] R. N. Anderson, A. Boulanger, W. B. Powell, and W. Scott, "Adaptive stochastic control for the smart grid," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 1098–1115, 2011.

[2] S. Khemakhem, M. Rekik, and L. Krichen, "A flexible control strategy of plug-in electric vehicles operating in seven modes for smoothing load power curves in smart grid," *Energy*, vol. 118, pp. 197–208, 2017.

[3] S. Caron and G. Kesidis, "Incentive-based energy consumption scheduling algorithms for the smart grid," in *2010 First IEEE International Conference on Smart Grid Communications*, pp. 391–396, IEEE, 2010.

[4] J. S. Vardakas, N. Zorba, and C. V. Verikoukis, "A survey on demand response programs in smart grids: Pricing methods and optimization algorithms," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 152–178, 2014.

[5] R. Subramanya, S. Sierla, and V. Vyatkin, "Exploiting battery storages with reinforcement learning: a review for energy professionals," *IEEE Access*, 2022.

[6] E. Brock, L. Bruckstein, P. Connor, S. Nguyen, R. Kerestes, and M. Abdelhakim, "An application of reinforcement learning to residential energy storage under real-time pricing," in *2021 IEEE PES Innovative Smart Grid Technologies-Asia (ISGT Asia)*, pp. 1–5, IEEE, 2021.

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

[8] V.-H. Bui, A. Hussain, and H.-M. Kim, "Double deep $q$-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 457–469, 2019.

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

[10] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, 2018.

[11] M. Anwar, C. Wang, F. de Nijs, and H. Wang, "Proximal policy optimization based reinforcement learning for joint bidding in energy and frequency regulation markets," in *2022 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, 2022.

[12] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[13] K. Katahira, "The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior," *Journal of Mathematical Psychology*, vol. 66, pp. 59–69, 2015.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.