Towards Human-Steerable Battery Energy Storage System Optimization: A Novel MDP Framework

Sai Krishna Gottipati
AI Redefined Inc
Montreal, Canada
sai@ai-r.com

Cloderic Mars
AI Redefined Inc
Montreal, Canada
cloderic@ai-r.com

Julien Gabaud
AI Redefined Inc
Montreal, Canada
julien.g@ai-r.com

Vahid Abdollahi
AI Redefined Inc
Montreal, Canada
vahid@ai-r.com

Laila El Moujtahid
AI Redefined Inc
Montreal, Canada
laila@ai-r.com

Matthew E. Taylor
AI Redefined Inc
University of Alberta
Edmonton, Canada
matt@ai-r.com

ABSTRACT

This paper introduces a novel formulation for optimization of the BESS (battery energy storage system) problem, a crucial component for driving renewable energy production to profitability. We review the existing literature on several optimization methods and make a case for the need of this novel MDP formulation that will be a foundation for learning from human demonstrations and feedback, prioritizing different constraints and real-world situations in addition to optimizing for profitability.

KEYWORDS

BESS, Green AI, Reinforcement Learning, Social Good

1 INTRODUCTION

In the pursuit of a sustainable and resilient energy future, the integration of energy storage solutions, alongside renewable energy sources, such as solar and wind, has become paramount. "Storage reduces total carbon dioxide emissions from the electricity system by utilizing overgeneration from zero-marginal emissions sources such as wind and solar to replace generation from the coal and natural gas fleet" [13].

The emergence of Battery Energy Storage Systems (BESS) has played a pivotal role in addressing this challenge, offering a means to store excess energy during periods of high production and releasing it during times of increased demand. Optimizing BESS is a challenge in itself, involving the navigation of the intricate interplay of factors such as avoiding energy waste, meeting demand shortfalls, and maximizing revenue in volatile open energy markets, through long-term energy acquisition contracts such as pre-purchase agreements (PPA[1]) or participating in global grid stability with mechanisms such as frequency control ancillary services (FCAS[2]). Figure 1 describes a model of the system.

A significant complication arises from the unpredictability of external factors crucial to this optimization process, namely weather patterns that influence solar and wind energy production, context-driven variations in energy consumption, and fluctuating prices in energy markets similar to stock markets. Balancing these factors is further complicated by the multiplicity of objectives. For instance, it might be financially prudent to forego fulfilling a PPA in favor of a more lucrative open market option, but this might impact future contract negotiations. The delicate trade-offs extend to decisions such as exceeding recommended charging cycles of a battery system to enhance revenue, which might reduce its lifetime.

Navigating this multifaceted optimization challenge often falls on human operators, who possess a nuanced understanding of the context and are empowered to make critical judgment calls. A robust BESS optimization system, therefore, should not only be sophisticated in its algorithms but also be bidirectional, seamlessly incorporating human operator feedback into its decision-making processes.

In recent years, techniques to take human feedback into account have been developed on top of the reinforcement learning (RL) framework, namely the reinforcement learning from human feedback (RLHF) family of algorithms [8, 18]. In this work, we take a first step towards this direction by demonstration of the applicability of RL based methods to the BESS optimization problem.

We review the existing literature and propose a novel Markov decision process (MDP) formulation for the BESS system. We introduce an open-source environment and establish different baseline algorithms including heurisitc, reinforcement learning, and imitation learning algorithms. We foresee an active community of developers and researchers pushing the frontiers of this novel system to reduce carbon dioxide emissions and make clean energy more profitable.

2 RELATED WORK

With the increasing global adoption of renewable energy sources, driven by their inherent variability influenced by weather conditions, electricity markets are experiencing increased volatility. Consequently, there is an increasing need to develop accurate simulations of power plants that participate in these markets. Such simulations facilitate the exploration of power generation management strategies aimed at achieving a more resilient power generation system, leveraging on-site power storage facilities to better align with market demand profiles. This tool is essential to maximize the profitability of renewable energy producers, thus stimulating further growth in this sector in conjunction with other conventional power generation methods.

[1]https://www.engie.com/en/news/ppa-power-purchase-agreement-what-is-it
[2]https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem/system-operations/ancillary-services

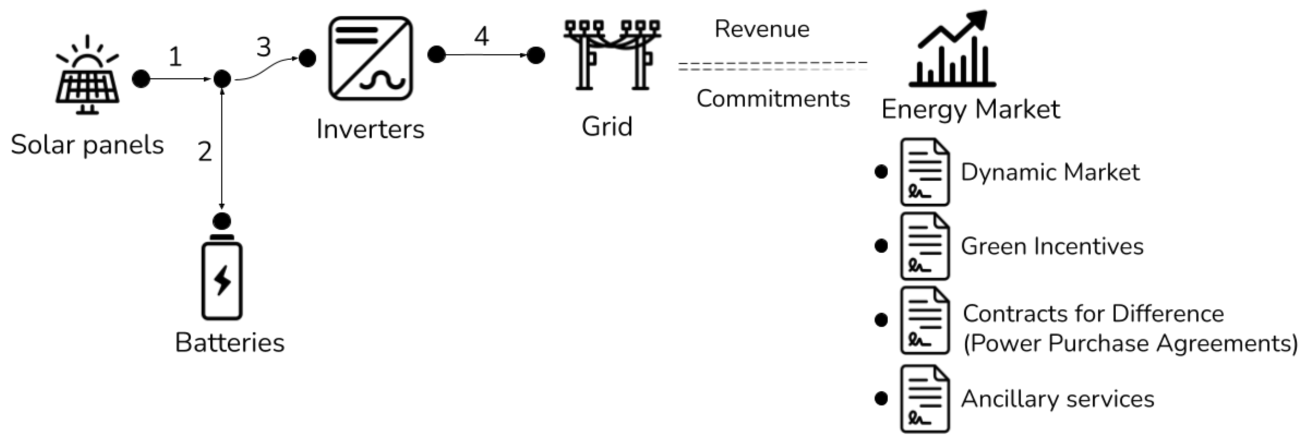

Figure 1: Solar energy production & storage simplified system diagram. Solar panels generate direct current (DC) electricity from sunlight (1), which is combined with the charge / discharge power of the batteries (2), then converted to alternating current (AC) electricity by inverters (3). AC electricity can then be sent to the power grid (4) and sold on the energy market. The introduction of batteries enables the produced energy to be stored to be sold later on, e.g. when the demand, and hence the price, is higher. A plant or portfolio of plants might have power generation commitments under certain power purchase agreements, there might be government incentives for green energy generation, participation in providing ancillary services for grid stabilization, and finally benefit from the dynamic spot price electricity market.

Formulating a mathematical problem to maximize the profit of a power plant entails integrating all pertinent aspects of the energy balance, encompassing electricity production and consumption dynamics, alongside associated supply and sale costs. The objective function typically quantifies the disparity between income and costs, subject to technical constraints governing the system's variables. Conventionally, addressing this optimization problem involves using classical optimization techniques, including linear programming [1], [22], mixed-integer linear programming [20], [5], [28], nonlinear programming [3], [14] and mixed-integer nonlinear programming [23], [26].

Implementing these optimization strategies offers simplicity and rapid execution to identify optimal solutions. Presently, mainstream computational software incorporates proficient solvers capable of efficiently handling mixed-integer linear problems. However, challenges arise when non-linear constraints are introduced into the power plant optimization problem, potentially resulting in non-convex feasible regions and complicating resolution. Some studies adopt various mathematical and heuristic methodologies to tackle these intricacies.

These mathematical methods attempt to converge towards optimal solutions, a crucial aspect of effectively managing energy resources. In linear programming models, the simplex method [24], [4] emerges as the preferred approach due to its broad applicability, ease of implementation, and computational efficiency. Conversely, for integer programming models, branch-and-bound techniques [17], [21] are predominantly used, allowing an intelligent search for optimal solutions by systematically evaluating feasible integer solutions while considering constraints and bounds. However, due to the complexity and computational demands of these methods, some studies advocate linearizing model equations before solving them.

Heuristic methods, on the other hand, such as Particle Swarm Optimization [23], [11] and Genetic Algorithms [27], offer efficient solutions for energy resource optimization. Particle swarm guides particles in the search space toward optimal solutions with minimal parameter adjustment. Genetic algorithms simulate biological evolution and natural selection, providing flexibility and exploring solution spaces intelligently. Selecting the appropriate parameters of the algorithm is crucial to achieve satisfactory results. However, these methods may not guarantee optimal results.

Deep reinforcement learning (DRL) is an alternative route to pursue to optimize power plant operations in renewable energy markets. Unlike traditional optimization techniques, which rely heavily on predefined models and assumptions, DRL learns directly from interactions with the environment, enabling it to adapt and improve over time. Compared to the classical optimization methods reviewed above, DRL approaches can learn from historical data, are self-adaptable, and learn a good control policy even in a complex environment of optimizing battery energy trading while limiting degradation costs using Deep Q-Learning (DQN) [7]. This inherent flexibility allows DRL to navigate dynamic and uncertain conditions more effectively, potentially unlocking new insights and strategies to optimize power plant operations in renewable energy markets. Double DQN was used in another study [6] to improve the overoptimistic value estimates of DQN by decoupling the selection from the evaluation of an action using a second neural network.

To broaden the action space of the DRL agents from discrete to continuous, necessary for tasks like adjusting battery charge or discharge power, policy gradient techniques are being employed. The deep deterministic policy gradient method [15] initially facilitated the handling of such action spaces and has been further refined by the DRL research community since its inception to achieve better stability and performance. In their study, Harrold et al. [10] employed Rainbow DQN to oversee battery operations in a microgrid, improving energy arbitrage through solar and wind energy utilization, while integrating real-world demand, renewable generation and dynamic energy pricing sourced from wholesale markets, achieving superior performance compared to DDPG and a linear programming model with discrete optimization. Recently, these improved algorithms have been applied to the power management problem. [25] used the soft actor-critic (SAC), twin-delayed deep deterministic policy gradient (TD3), and proximal policy optimization (PPO) to control potentially millions of small-scale assets in private households. Their DRL algorithms outperformed common heuristic algorithms and fell short of the results provided by linear optimization, but by less than a thousandth of the simulation time.

In their study, [19] used battery storage for concurrent energy arbitrage and frequency regulation services, to maximize total revenue while adhering to physical constraints. By tackling the multi-timescale challenge through nested Markov decision process submodels and implementing a co-optimization scheme, their method effectively coordinated these actions. They used the TD3 with an exploration noise decay approach in simulations conducted with real-time electricity prices and regulation signal data, showcasing superior performance compared to DQN.

The studies mentioned above showcased the superiority of DRL in grasping the intricate patterns and uncertainties inherent in power generation and market dynamics, outperforming classical optimization techniques. What sets our work apart is the introduction of a novel MDP formulation designed to tackle the intricacies of energy storage optimization. In doing this, we offer a methodical and rigorous approach to model BESS operations, encapsulating crucial variables such as energy production, consumption, market dynamics, and storage constraints. This formulation not only provides a holistic representation of the optimization challenge, but also facilitates the development of streamlined algorithms for BESS management. We advocate for the widespread adoption of our formulation as the benchmark for future research efforts and industrial applications, as it lays a solid foundation for the promotion of advancements in sustainable energy management practices.

3 ENVIRONMENT

Our BESS model consists of one renewable energy source and one battery. The objective is to decide how much power should be sold to the grid and to charge or discharge or leave the battery idle based on the spot price and the LGC price. The system is represented in Figure 1. We model the MDP as follows: Each time step corresponds to an interval of 5 minutes. Each episode lasts for one day, that is, 288 time steps. The observation at each time step includes the power generation and price values for the last hour, the current and the next hour, the current state of the battery, the maximum allowable charge and discharge of the battery, and the number of time steps remaining until the horizon. The action is a scalar value that tells the total amount of power sold to the grid; this is represented as the edge (4) in Figure 1. If the action is less than the power generated, the additional power generated is used to charge the battery; this is represented as edge (2) in Figure 1. If this value is higher than the generated power, the remaining power is obtained by discharging the battery; this is represented as the edge (2) in Figure 1. However, if at any point during training the action indicates that the battery should be charged or discharged beyond its capacity, we classify it as an illegal action, penalize it heavily, and terminate the episode. During the evaluation phase, we emulate the real-world settings by ignoring the illegal actions, as the inverter is disconnected when the battery is full in the real world (digital twin).

3.1 Generate data of multiple levels of difficulty

The environment is equipped with the capability to generate data (generated power and price) of varying levels of difficulty. This can be used to train agents through curriculum learning. After inspecting the real data of the generated power and the corresponding prices in 10 different provinces in Australia, we generated synthetic data, close to these real data.

At the foundational level, generated power is modeled by a cosine wave, and price is modeled by a sine wave (as they are inversely correlated — if the generated power is higher, the demand would decrease, and hence the price is lower. On the other hand, if the generated power is less, the demand will increase and the price will be higher). When the value of the cosine wave is negative, it is clipped to zero, indicating zero power production during the night. For the corresponding duration, when the generated power is zero, the price is fixed to its maximum value. The generated power values are multiplied by a constant to match the distribution with the real power generation values of a power plant. Similarly, the prices are multiplied by a constant to make them similar to the actual prices.

Furthermore, a random noise sampled (at each time step) from a uniform distribution (with a fixed amplitude) can be added to every time step of the generated power and price values. Increasing the amplitude will make the task harder.

The real power generation curves are characterized by unusual spikes at arbitrary times. To mimic this behavior, we added the provision of adding spikes, i.e., a random noise sampled from a uniform distribution (but with much higher amplitude than the amplitude for noise at each time step) after every few time steps, again characterized by time period and additional position noise.

Finally, we also experimented with real power generation data. Two sources have been used to get real data for this environment. We chose to simulate a typical solar farm in Victoria state in Australia with a 4.7 MW inverter, 8 MW solar panels, and an 11 MWh battery. The price data was obtained from the Australian Energy Market Operator (AEMO). The simulation has been performed with the pvlib library [2] and solar data from the CAMS solar radiation services with pvlib iotools[12].

3.2 Observation and Reward

In addition to the observation and reward described in Section 3, the environment is equipped with several other options. A “mini observation” only includes the price and power generated at the current time step, the current state of the battery, and the number of time steps remaining until the horizon. All of these values are normalized by the corresponding normalization constants. An "observation with noisy forecasts" allows adding additional noise to the forecast values for both prices and power generation.

Agents can be trained and evaluated with different types of rewards. "just revenue" computes the revenue generated at each time step by multiplying the price at the current time step with the total power to the market grid. "Scaled revenue" scales the raw revenue by dividing with an appropriate normalization constant. "Scaled revenue and penalty" adds a penalty of -1 for illegal actions in addition to the scaled revenue. "Survival" gives a reward of -1 for illegal actions and +1 for legal actions.

In our experiments, mini observation along with survival reward were used as quick sanity checks to confirm that the agent can train. We then used standard observation with scaled reward and penalty for training. Both the scaled reward and penalty, and just revenue, were plotted for evaluation episodes.

3.3 Design principles

All implementations (algorithms) have the same input and output formats. For example, even though the output range of the TD3 algorithm (bounded by [-max-action, +max-action]) is different from that of the heuristic or no-battery baseline (which has an actual range of [0, max-action + sum-action]), we rescale the output of the heuristic and no-battery baseline to ensure that the environment can handle the actions coming from all the implementations in a similar way. Similarly, a no-battery baseline only requires the power generated at the current time step, but we still send the complete observation as its input (the same as the input to RL algorithms).

4 IMPLEMENTATIONS

4.1 No battery baseline

Our first baseline is a scenario with no battery — all the power generated at every time step is sold to the grid.

4.2 Heuristic Algorithm

We introduce a simple heuristic algorithm in which decisions are made based on price forecasts. If the average forecast price for the next hour is less than the current price, it indicates that the price is decreasing. Therefore, we decide to sell all the power generated in the current time step and completely discharge the battery. However, if the average forecast price for the next one hour is greater than the current price, it indicates that the price is increasing. Therefore, we completely charge the battery and only the additional power generated is sold.

4.3 Linear Programming

As a third baseline, we approached the optimization problem presented in Section3 with linear programming using the PuLP library. The performance (measured as cumulative revenue) is slightly worse and the run time is orders of magnitude slower compared to our RL solution.

4.4 RL Algorithms

Twin Delayed DDPG (TD3): TD3 [9] is an off-policy algorithm for continuous control tasks. It builds on DDPG [16] and uses the following techniques to make the RL algorithm more stable: (1) Clipped Double Q-learning which maintains two Q-estimators and updates the loss functions using the smaller Q-value to avoid overestimation bias (2) delayed policy and target network updates compared to Q-function update (3) target policy smoothing as regularization by introducing noise to the target action value.

4.5 Imitation Learning Algorithms and RLHF

As an example of learning from human demonstrations, we train a policy network to imitate the behavior of an RL agent using behavior cloning. Furthermore, we developed a framework to continuously solicit human feedback and preferences for the actions and train a reward model that is used to further fine-tune the policy model. This is very similar to how the large language models (LLMs) are updated in some of the latest work.

5 EXPERIMENTS

As shown in the figure, we compared the performance of our RL algorithm with other baselines on the real data from March 2023 to March 2024. In all the cases with varying noise on the forecast prices, the RL algorithm outperformed the other baselines.

6 CONCLUSION

In this work, we reviewed the current state of the art of BESS optimization techniques and highlighted the need for better frameworks to take human demonstrations and feedback into account and proposed a novel MDP framework and benchmarked heuristic, linear programming, and reinforcement learning algorithms. We demonstrated that RL algorithms generate more revenue compared to other approaches. We advocate for the widespread adoption of our formulation as the benchmark for future research endeavors and industrial applications, as it lays a solid foundation for the promotion of advancements in sustainable energy management practices. More experimental results, data generation methods, real data sources, and other information can be found on our project page https://ai-r.com/research/bessrl

REFERENCES

[1] Arman Alahyari, Mehdi Ehsan, and MirSaeed Mousavizadeh. 2019. A hybrid storage-wind virtual power plant (VPP) participation in the electricity markets: A self-scheduling optimization considering price, renewable generation, and electric vehicles uncertainties. *Journal of Energy Storage* 25 (2019), 100812.

[2] Kevin S. Anderson, Clifford W. Hansen, William F. Holmgren, Adam R. Jensen, Mark A. Mikofski, and Anton Driesse. 2023. pvlib python: 2023 project update. *Journal of Open Source Software* 8, 92 (2023), 5994. https://doi.org/10.21105/joss.05994

[3] Sadra Babaei, Chaoyue Zhao, and Lei Fan. 2019. A data-driven model of virtual power plants in day-ahead unit commitment. *IEEE Transactions on Power Systems* 34, 6 (2019), 5125–5135.

[4] Arijit Bagchi, Lalit Goel, and Peng Wang. 2019. An optimal virtual power plant planning strategy from a composite system cost/worth perspective. In *2019 IEEE Milan PowerTech*. IEEE, 1–6.

[5] Rémi Bourbon, Sandra Ulrich Ngueveu, Xavier Roboam, Bruno Sareni, Christophe Turpin, and David Hernández-Torres. 2019. Energy management optimization of a smart wind power plant comparing heuristic and linear programming methods. *Mathematics and Computers in Simulation* 158 (2019), 418–431.

[6] Van-Hai Bui, Akhtar Hussain, and Hak-Man Kim. 2019. Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties. *IEEE Transactions on Smart Grid* 11, 1 (2019), 457–469.

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

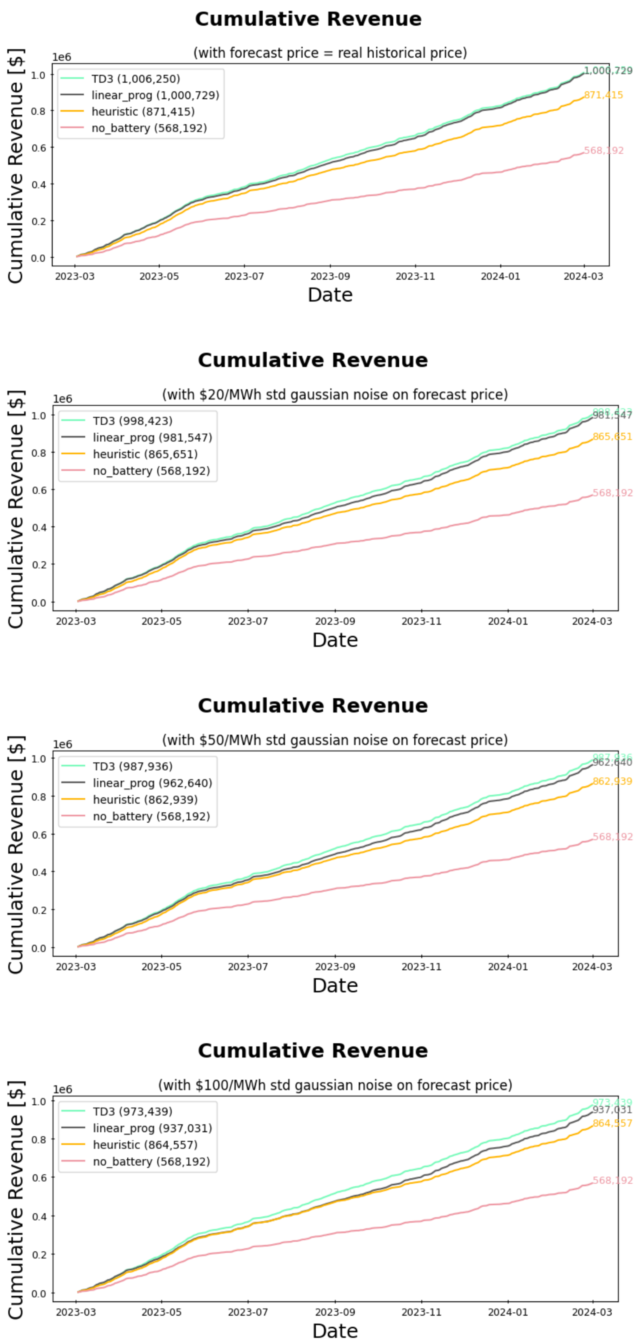

Figure 2: Cumulative revenues measured over 1 year with varying noise on price forecasts. TD3 outperformed all other existing approaches in all the cases

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

[9] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1582–1591. http://proceedings.mlr.press/v80/fujimoto18a.html

[10] Daniel JB Harrold, Jun Cao, and Zhong Fan. 2022. Data-driven battery operation for energy arbitrage using rainbow deep reinforcement learning. *Energy* 238 (2022), 121958.

[11] Daniel Hropko, Ján Ivaneckỳ, and Ján Turček. 2012. Optimal dispatch of renewable energy sources included in virtual power plant using accelerated particle swarm optimization. In *2012 ELEKTRO*. IEEE, 196–200.

[12] Adam R. Jensen, Kevin S. Anderson, William F. Holmgren, Mark A. Mikofski, Clifford W. Hansen, Leland J. Boeman, and Roel Loonen. 2023. pvlib iotools—Open-source Python functions for seamless access to solar irradiance data. *Solar Energy* 266 (2023), 112092. https://doi.org/10.1016/j.solener.2023.112092

[13] Jennie Jorgenson, A Will Frazier, Paul Denholm, and Nate Blair. 2022. *Storage futures study: Grid operational impacts of widespread storage deployment.* Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

[14] Rakkyung Ko, Daeyoung Kang, and Sung-Kwan Joo. 2019. Mixed integer quadratic programming based scheduling methods for day-ahead bidding and intra-day operation of virtual power plant. *Energies* 12, 8 (2019), 1410.

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

[16] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1509.02971

[17] Zuoyu Liu, Weimin Zheng, Feng Qi, Lei Wang, Bo Zou, Fushuan Wen, and You Xue. 2018. Optimal dispatch of a virtual power plant considering demand response and carbon trading. *Energies* 11, 6 (2018), 1488.

[18] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 2285–2294.

[19] Yushen Miao, Tianyi Chen, Shengrong Bu, Hao Liang, and Zhu Han. 2021. Co-optimizing battery storage for energy arbitrage and frequency regulation in real-time markets using deep reinforcement learning. *Energies* 14, 24 (2021), 8365.

[20] Hrvoje Pandžić, Igor Kuzle, and Tomislav Capuder. 2013. Virtual power plant mid-term dispatch optimization. *Applied energy* 101 (2013), 134–141.

[21] Samaneh Pazouki and Mahmoud-Reza Haghifam. 2016. Optimal planning and scheduling of energy hub in presence of wind, storage and demand response under uncertainty. *International Journal of Electrical Power & Energy Systems* 80 (2016), 219–239.

[22] Mette Kirschmeyer Petersen, Lars Henrik Hansen, J Bendtsen, Kristian Edlund, and Jakob Stoustrup. 2013. Market integration of virtual power plants. In *52nd IEEE conference on decision and control*. IEEE, 2319–2325.

[23] Jing Qiu, Ke Meng, Yu Zheng, and Zhao Yang Dong. 2017. Optimal scheduling of distributed energy resources as a virtual power plant in a transactive energy framework. *IET Generation, Transmission & Distribution* 11, 13 (2017), 3417–3427.

[24] Morteza Rahimiyan and Luis Baringo. 2019. Real-time energy management of a smart virtual power plant. *IET Generation, Transmission & Distribution* 13, 11 (2019), 2015–2023.

[25] Jan Martin Specht and Reinhard Madlener. 2023. Deep reinforcement learning for the optimized operation of large amounts of distributed renewable energy assets. *Energy and AI* 11 (2023), 100215.

[26] Zhongfu Tan, Guan Wang, Liwei Ju, Qingkun Tan, and Wenhai Yang. 2017. Application of CVaR risk aversion approach in the dynamical scheduling optimization model for virtual power plant connected with wind-photovoltaic-energy storage system with uncertainties and demand response. *Energy* 124 (2017), 198–213.

[27] Jean-François Toubeau, Zacharie De Grève, and François Vallée. 2017. Medium-term multimarket optimization for virtual power plants: A stochastic-based decision environment. *IEEE Transactions on Power Systems* 33, 2 (2017), 1399–1410.

[28] Chixin Xiao, Danny Sutanto, Kashem M Muttaqi, and Minjie Zhang. 2020. Multi-period data driven control strategy for real-time management of energy storages in virtual power plants integrated with power grid. *International Journal of Electrical Power & Energy Systems* 118 (2020), 105747.