Jan 17th, 12:00 AM

Managing Intermittent Renewable Generation with Battery Storage using a Deep Reinforcement Learning Strategy

Yuchen Zhou
Karlsruhe Institute for Technology (KIT), Institute of Information Systems and Marketing(IISM), Karlsruhe, Germany, yuchen.zhou@student.kit.edu

Sarah Henni
Karlsruhe Institute for Technology (KIT), Institute of Information Systems and Marketing(IISM), Karlsruhe, Germany, sarah.henni@kit.edu

Philipp Staudt
Karlsruhe Institute for Technology (KIT), Institute of Information Systems and Marketing(IISM), Karlsruhe, Germany, philipp.staudt@kit.edu

Managing Intermittent Renewable Generation with Battery Storage using a Deep Reinforcement Learning Strategy

Yuchen Zhou[1], Sarah Henni[1], Philipp Staudt[1]

[1] Karlsruhe Institute for Technology (KIT), Institute of Information Systems and Marketing, Karlsruhe, Germany
udwyg@student.kit.edu, {sarah.henni,philipp.staudt}@kit.edu

Abstract. Most of Germany's existing wind and solar plants have been losing their subsidies after 20 years of operation since 2020. Without support schemes, the challenges for the renewable operators are the intermittent generation and the fluctuating power prices. Consequently, lower-than-expected revenues and high revenue variability make it more difficult for the renewable operators to be active on power markets. Therefore, the renewable operators have to be profit effective as well as cope with the high variability of their revenue. This paper proposes a deep reinforcement learning (DRL) based model to adjust the renewable operators' short-term energy supply using a battery storage strategy. The simulative empirical evaluation shows that the renewable operators can be profitable on the market and improve their revenue stability using the proposed DRL based battery storage strategy.

Keywords: deep reinforcement learning, battery storage system, renewable generation, maximizing profit, revenue variability

1 Introduction

As sustainable and environmentally friendly sources of electricity, renewable energy generators have the potential to replace power generation from conventional power plants using fossil fuels. In Germany, a significant growth of wind and solar power plant installation has been observed, supported by fixed feed-in tariffs guaranteed for 20 years. Nonetheless, with the reform of the Renewable Energy Act taking effect in 2017, it was made compulsory for large facilities to enter into the so-called "direct marketing": operators must sell their power directly on the wholesale electricity market without the guaranteed fixed price [1]. Furthermore, most of Germany's existing wind and solar stations are losing their subsidies successively after 20 years of operation [2].

Renewable operators without support schemes are directly exposed to market risks. Unlike conventional energy sources, wind and solar power fluctuate with the weather and are non-dispatchable. Consequently, revenues of renewable operators can vary considerably [3]. On the wholesale power market, power prices change relatively

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

quickly throughout the day due to the fluctuating demand and supply patterns and the lack of storage possibilities [4]. Meanwhile, the increasing penetration of renewable generation (RG) pushes more expensive conventional generation down the merit order, and even decreases the price [5]. Very low and even negative wholesale prices can result in lower-than-expected revenues, and the high revenue variability makes it more difficult for renewable operators to be active on long-term forward power markets [6]. Thus, the renewable operators must generate strategies to be profitable on power markets and to dampen the revenue variability.

Battery storage systems (BSSs) provide means to make RG dispatchable and become an enabling technology for RG regarding various services in the power system [7]. By charging the battery at lower prices and discharging it at higher prices, the renewable operators can shift their energy sales from times when demand is low and supply is high to times with better revenue opportunities. Hence, they can increase their profit or avoid large variability in revenues. Recent research has proposed deep reinforcement learning (DRL) based solutions to allow for more short-term flexibility in the market through investor-owned BSS for energy arbitrage (EA) [8–10]. EA in this context means buying electricity at lower prices and selling it later at higher prices. In literature, it has become common to refer to this as arbitrage. However, an important characteristic of arbitrage is that no risk is associated which is not the case for these strategies, since it is uncertain whether the price spreads at the spot market will be sufficient to cover storage cycling costs. We therefore choose to name it energy arbitrage. As an effective data mining technology, DRL algorithms can fully explore and utilize the fluctuation patterns in the historical datasets to generate optimized strategy [8]. However, the joint operation of the renewable generators and BSS causes huge initial investment costs due to the high battery prices. As is indicated in [11], a dynamic sizing of the storage capacity might be more profitable if the storage operator is modeled as an independent market entity and offers storage service to the renewable operator. Additionally, most researchers only consider maximizing the profit of the system owner. Only few researchers have addressed the problem of revenue variability reduction. In line with the requirements, our paper therefore addresses the following research question:

"How can a DRL model be applied to maximize the profit as well as reduce the variability of a renewable operator's revenue using a BSS service in the context of intermittent renewable generation?"

We perform a case study under real market prices by simulating a solar park and a wind farm that use a BSS service agent to increase profits and counteract revenue variability.

2 Related Works

To schedule short-term energy supply with BSS, conventional programming methods mainly include mixed integer linear programming (MILP), and dynamic programming (DP) [12, 13]. In [12], the authors propose a storage bidding strategy that includes price and quantity bids based on stochastic price forecasts using a probability density function. The optimization problem is reformulated into a MILP and solved using a

standard LP solver. In [13], a control algorithm based on DP using weather and consumption predictions is proposed for a renewable energy system coupled with an energy storage system, to limit the grid power ramp-rate and to optimize energy trading for the system owner. However, these methods often have large computational costs and rely on very accurate forecasts which are not available for power markets [14]. Moreover, the potentially high dimensionality of the state space makes these methods unsuitable for applications in power system [15].

Considering the random nature of RG and market prices, as well as the time-coupled feature of the battery state of charge (SoC), this problem can be modelled and solved through DRL [10]. By providing the observation of the environment as input for an artificial neural network (ANN), DRL can solve many real-world problems with continuous and high-dimensional data [14]. Recently, many papers studied the application of DRL in energy supply scheduling with BSS, which shows promising results [8-10]. In [8], a data-driven controller using DRL is proposed to increase a wind power producer's revenue given uncertain wind power generation and electricity prices. The simulation results of the case study conducted on a wind farm show that the uncertainties can be effectively handled and high revenues for the power producer can be ensured. In [9], the authors propose a DRL based method to optimize the control policy for battery charging and discharging with the purpose of maximizing the profit considering an accurate battery degradation model. The empirical results based on the historical U.K. wholesale market prices show the effectiveness and the economic advantage compared to a model based on MILP. In [10], a DRL based agent is proposed for investor-owned PV-BSS to maximize the profit by providing stacked services in power systems. The proposed method is tested using real market data.

It is worth mentioning that different measures are used to control the violation of battery charging and discharging constraints in these studies. In [9], actions of the agent are discretized as relative values regarding the maximum charging and discharging power of the current battery capacity, whereas in [10] they are continuous coefficients regarding the energy management unit. The authors also propose a safety control algorithm in [10] to ensure that the operating constraints of the battery are strictly satisfied by always regulating invalid charging and discharging values into the safety range. In [8], actions are concrete amounts, and the penalty fee is calculated in the reward function for violation of constraints. Using absolute values lacks flexibility in practice and applicability in systems with intermittent RG. The accompanying control algorithm is more suitable for stacked services in [10] but could be complex for renewable operators in our case. Therefore, we make an adaptive design by using discrete relative values for the action space and by considering a penalty term in the reward function for violation of the charging and discharging constraints.

3 Methodology

In this section, we first describe the background of reinforcement learning (RL) and deep Q-network (DQN), then present the framework of the proposed DRL based model, and finally introduce the measure we use to estimate the revenue variability.

3.1 RL background and DQN

RL is framing of problems in which an agent learns its optimal behavior in regard to a given objective through numerous trial-and-error interactions with a dynamic environment. Generally, RL problems are described as Markov decision process (MDP), which is modelled as a four-tuple $< S, A, p_a, r_a >$ [17], where:

- S is a set of states, which contains agent's observation from the environment,
- A is a set of actions the agent can take,
- $p_a(s, s') = p_r(s_{t+1} = s' | s_t = s, a_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t + 1$,
- $r_a(s, s')$ is the immediate reward passed from the environment to the agent by taking action a and changing the state s to state s'.

An RL agent interacts with its environment in a sequence of discrete time steps. At each time step t, the agent observes the current state, $s_t \in S$, and on that basis chooses an action, $a_t \in A(s)$, that it communicates to the environment. One step later, the agent receives a numerical reward, $r_{t+1} \in R$, which implies how good or bad that action was. The environment then changes to a new state, S_{t+1}. This process continues to a finite time step T and thus causes a sequence of experiences of the whole episode. The goal of the agent is to maximize the cumulative discounted reward $G = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, where $\gamma \in [0,1]$ is the discount rate to trade off the immediate and long-run rewards.

As a classic RL algorithm, Q-Learning was developed in 1989 [18]. As suggested by its name, the agent updates the action-value function $Q(s, a)$ recursively:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right] \quad (1)$$

where $\alpha \in [0,1]$ is the learning rate. The updating continues until $Q(s, a)$ converges to the optimal value $Q^*(s, a)$. Thereby a lookup table of the Q-values for each state-action pair is defined, and the agent chooses the action based on the Q-values at each time step of a given state.

Although the standard Q-Learning guarantees convergence, it suffers severely from the so-called curse of dimensionality. To overcome the limitation of tabular Q-Learning, the DQN algorithm was developed [19]. As a combination of deep learning and RL, DQN uses an ANN to approximate the Q-value, which takes the continuous state as input and generates the Q-value for each discrete action. The agent interacts with the environment by choosing the action based on the output of the ANN and stores the past experiences in a large memory. To train the ANN, samples of a fixed size are chosen randomly in each iteration to perform the update of θ_i at iteration i by minimizing the following loss function between the predicted Q-value and the target:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')} \left[(r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i))^2 \right] \quad (2)$$

where θ_i^- are the target network parameters that are only updated every C steps.

3.2 Framework of the Proposed Model

In this section, we detail the proposed DRL based model, the overall framework of which is presented in Figure 1. The key elements of the model are illustrated in the following:

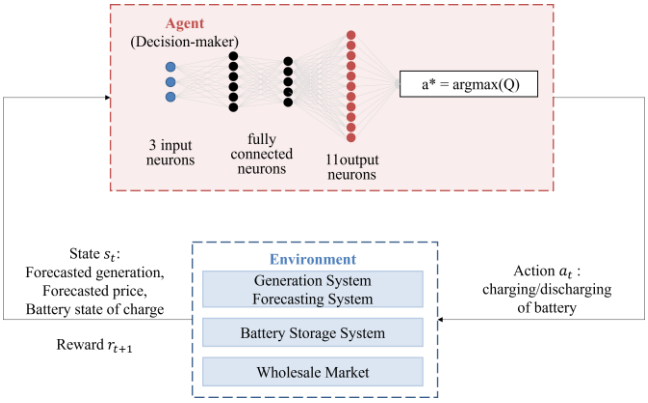

Figure 1. Overall Framework of the Proposed DRL based Model

Agent. The agent learns the optimized policy and decides charging or discharging actions to adjust the energy supply on the market. It uses an ANN taking the state as input to output the Q-values of each state-action pair. The reward from the environment is used to approximate the target Q-value for training the parameters of the ANN.

Environment. The environment is made up of (1) the generation system and the forecasting system of the renewable operator, which provide the forecasts of RG and market prices, (2) the BSS service provider, a separate market entity which charges the renewable operator for the battery charging and discharging and updates the real-time battery SoC, and (3) the wholesale market, which determines the actual power prices.

State Space. The state at time step t is defined as $s_t = (g^{fore}(t), p^{fore}(t), SoC(t-1))$, where g^{fore}, p^{fore} are forecasted RG and price, and $SoC(t-1)) \in [0,1]$ is the SoC at the end of $t-1$. In the actual operation, the SoC must satisfy the capacity constraint $SoC_{min} \leq SoC(t) \leq SoC_{max}$ to ensure that the stored energy in the battery is always within the permitted range. The agent has no impact on g^{fore} and p^{fore}, and is supposed to learn the fluctuation patterns as well as the implicit interrelationship between them if they exist. The SoC is explicitly changeable by the agent.

Action Space. To deal with the intermittent nature of RG, the action is discretized as $a_t \in \{-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$. It specifies the percentage of the maximal allowable charging power $c_{max}(t) = \min\left(g^{fore}(t)\eta^{ch}, e^{up}(t)\right)$, if $a_t <$

0, or the percentage of the maximal allowable discharging power $d_{max}(t) = e^{dn}(t)$, if $a_t > 0$. $a_t = 0$ means neither charge nor discharge. $e^{up}(t) = \big(SoC_{max} - SoC(t-1)\big) \cdot U$ and $e^{dn}(t) = (SoC(t-1) - SoC_{min}) \cdot U$ specify the available upward and downward energy to reach the maximal or the minimal energy level of the battery, where U is the nominal battery energy capacity and η^{ch} is the charging efficiency. Additionally, the actual charging and discharging power is also restricted by the maximal charging and discharging rate $R^{ch}/\ R^{dis}$. Therefore, the actual charging power into the battery $e^{in}(t)$ and the discharging power from the battery $e^{out}(t)$ are defined by $e^{in}(t) = \min\big(R^{ch}, |a_t| * c_{max}(t)\big)$, if $a_t < 0$, and $e^{out}(t) = \min\Big(R^{dis}, a_t * d_{max}(t)\Big)$, if $a_t > 0$. Note that at least one of the variables $e^{in}(t)$ and $e^{out}(t)$ is 0 at any time t regarding the choice of a_t to ensure that the battery will not be charged and discharged at the same time. Specifically, if the forecasted generation $g^{fore}(t)$ is 0.8 MWh, the available upward energy of the battery $e^{up}(t)$ is 2 MWh, and the agent chooses action -0.2, then the actual charging power into the battery is $e^{in}(t) = \min(R^{ch}, 0.2 * \min(0.8 \cdot \eta^{ch}, 2))$ MWh. Thereby, the energy supply on the market $e^{supply}(t)$ is adjusted by (3).

$$e^{supply}(t) = g^{fore}(t) - e^{in}(t)/\eta^{ch} + e^{out}(t)\eta^{dis} \tag{3}$$

where η^{dis} is the discharging efficiency, and the transition of the SoC is defined by (4).

$$SoC(t) = SoC(t-1) + \frac{-e^{in}(t)/\eta^{ch} + e^{out}(t)\eta^{dis}}{U} \tag{4}$$

The proposed action space ensures that the charging and discharging power is always within the permitted range. Compared to methods using external controlling systems and models which artificially exclude invalid actions before the decision making, the proposed agent is more self-ruling and learns the policy more autonomously.

Reward. Based on the objective of maximizing the profit as well as reducing the variability of a renewable operator's revenue by adjusting the energy supply at times when there is inverse correlation between generation and prices, we define the immediate reward r_t as (5).

$$r_t = p^{act}(t) \cdot e^{\delta}(t) - C_{battery} \cdot \left|e^{\delta}(t)\right| - P(t) \tag{5}$$

where $p^{act}(t)$ is the actual power price given by the market, $e^{\delta}(t) = -e^{in}(t)/\eta^{ch} + e^{out}(t)\eta^{dis}$ defines the difference of $e^{supply}(t) - g^{fore}(t)$, $C_{battery}$ is the service cost for charging and discharging the battery with 1 MWh energy, and $P(t)$ is the penalty for invalid charging or discharging actions.

The renewable operator's revenue of selling RG on the market without optimization is $p^{act}(t) \cdot g^{fore}(t)$, while the optimized revenue using the DQN strategy is $p^{act}(t) \cdot e^{supply}(t)$. We use $p^{act}(t) \cdot e^{\delta}(t)$ instead of $p^{act}(t) \cdot e^{supply}(t)$ to calculate the profits or losses arising from the charging or discharging actions, since the former doesn't contain $g^{fore}(t)$. Using the fluctuating generation in the reward function can

result in high reward variability and thus have negative influence on the learning performance. $C_{battery} \cdot |e^{\delta}(t)|$ defines the service cost for charging or discharging the battery with $|e^{\delta}(t)|$. In contrast to the papers that only consider the revenue in the reward function, we define the penalty in this paper by (6).

$$P(t) = PF \cdot (pen^{ch}(t) \parallel pen^{dis}(t)) \tag{6}$$

where PF is the penalty factor, which determines the size of the penalty term, and its appropriate value depends on the relative size of the revenue. If PF is set too low, the penalty doesn't have sufficient effect on the learning behavior; if too high, the penalty will prevent accurate learning from the monetary reward. $pen^{ch}(t)$ and $pen^{dis}(t)$ are boolean variables in (7) and (8).

$$pen^{ch}(t) = \begin{cases} 1, if\ a_t < 0\ and\ g^{fore}(t) = 0\ or\ e^{up}(t) = 0 \\ 0, otherwise \end{cases} \tag{7}$$

$$pen^{dis}(t) = \begin{cases} 1, if\ a_t > 0\ and\ e^{dn}(t) = 0 \\ 0, otherwise \end{cases} \tag{8}$$

The reward serves as a numerical signal for the agent to learn the optimal strategy. However, the agent cannot get effective feedback regarding the chosen action by only considering revenue as the reward. In other words, the reward would be zero if the agent decided to charge at times when there is no RG or no storage space in the battery, or to discharge at times when there is no stored energy in the battery, which significantly affects the learning performance. Therefore, we design the penalty to give the agent an explicit negative reward signal at times when the agent chooses invalid actions.

3.3 Evaluation Measure of Revenue Variability

The Value at Risk (VaR) and the Conditional Value at Risk (CVaR) are measures to evaluate the potential loss of investment portfolios [16]. For a given confidence level β, and the probability distribution function of losses over a certain time horizon $F(x)$, the $\beta - VaR = F_x^{-1}(1-\beta)$ defines the lowest value of the β largest losses, whereas the $\beta - CVaR = \int_{-\infty}^{VaR} F_x^{-1}(1-\beta)dx$ defines the conditional expectation of losses beyond β. In other words, the CVaR is the average value of the β largest losses and is therefore higher but more robust than the VaR, which only represents a single value. In this paper, we use the CVaR as a measure for estimating the revenue variability rather than simply using the variance, since it only measures the negative deviations from the mean revenue rather than also punishing the positive deviations.

4 Case Study

The performance of the proposed DRL based model is validated by simulating a virtual solar park with a 1 MW installed capacity using empirical market prices. To compare the results of different generation technologies, a virtual wind farm with a 1 MW

installed capacity is simulated. We assume that the BSS is provided as a service, which is not exclusively used for the described use case but can be accessed temporarily by the solar park and the wind farm operator.

4.1 Data and Model Implementation

We use the hourly price from the German day-ahead wholesale electricity market as actual price [20] and generate price forecasts by adding Gaussian noise to the actual value. For the predicted generation, we use the data generated from [21-23]. We use data from 2018 and 2019 for the training and testing procedure, respectively. We use a BSS with 3 MWh nominal energy capacity. Table 1 shows other technical parameters of the battery.

Table 1. Technical parameters of the battery

Parameter	Value	Parameter	Value
SoC_{max}	0.9	SoC_{min}	0.1
R^{ch}/R^{dis}	0.45	η^{ch}/η^{dis}	0.9

The proposed DRL model is developed using Keras-RL. Table 2 details the training parameters including the deep neural network model architecture. In this paper, $C_{battery}$ is set to 10 €/MWh. In other words, the cyclic cost for charging and discharging the battery with 1 MWh is 20 €. We discuss this in section 5. For the value of PF, we tested 0, 30, and 50 based on our exemplary case. With $PF = 0$, the agent failed to learn how to avoid invalid actions, and with $PF = 50$, the agent only gets slight revenue as he acts too conservatively to avoid invalid actions. We finally set PF to be 30, thus the penalty has sufficient effect on the learning behavior and the agent can achieve higher revenue.

Table 2. Summary of DRL model parameters

Item	Value	Parameter	Value
No. of hidden layers	2	α	0.005
No. of nodes in each layer	32	γ	0.99
Activation function	ReLU	C	500
Optimizer	Adam	T	168
		ε_min	0.2

4.2 Model Performance

During the training procedure, we generate 5 different random seeds for both operators. For each seed we train the DQN agent for 5000 episodes. The convergence process of the mean episode return for both operators is shown in Figure 2. The episode return refers to the sum of rewards over one episode, and the mean episode return in the Y-Axis refers to the simple moving average of the episode return over each 100 episodes. The mean and the standard deviation of the mean episode return over the 5 seeds are

illustrated through the solid lines and the shaded areas, respectively. It can be observed that the value for the solar park converges to -570 after 3590 episodes, and the value for the wind farm converges to -130 after 3500 episodes. Both values converge to a negative number, since the exploration rate ε_min is set to 0.2. There is a 20% probability that the agent randomly chooses non-optimal or invalid actions after the convergence, which are either not profitable for the renewable operator and cause negative profit or invalid charging and discharging choices and cause penalty.

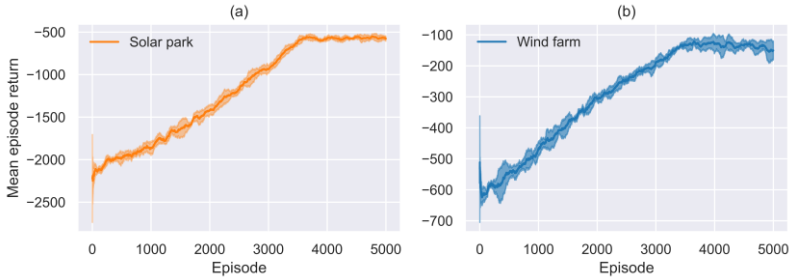

Figure 2. The mean episode return in training ensures convergence for both (a) the solar park and (b) the wind farm.

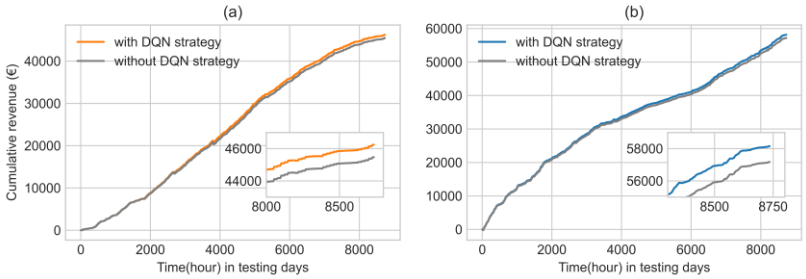

Figure 3. The cumulative revenue in testing days with DQN strategy is higher for both (a) the solar park and (b) the wind farm.

The cumulative revenue over the testing period of 2019 for both operators are shown in Figure 3. We compare the profits that the renewable operator generates when using a BSS service (*with DQN strategy*) with the profits that he generated while directly selling all renewable generation on the market (*without DQN strategy*). It can be clearly observed that both operators achieve higher revenue with the proposed model than selling the RG without the DQN strategy. For the solar park, the renewable operator can get a yearly revenue of about 46 k € with the DQN strategy after subtracting the cost of the BSS, achieving an improvement of 772 € compared to the revenue of about 45 k € without the optimization of the DQN strategy. For the wind farm, the renewable operator earns a yearly revenue of 58 k € with the proposed DQN strategy after subtracting the cost of the BSS, achieving an improvement of 981 € compared to the revenue of 57 k € without the DQN strategy.

4.3 Comparison of Agent Behavior

To assess the usage of the BSS for different operators, we analyze the charging and discharging amount in the testing period. During the whole year of 2019, the total charging amount is 48 MWh, and discharging amount is 43 MWh for the solar park. As for the wind farm, the proposed strategy results in a total charging amount of 52 MWh, and a total discharging amount of 45 MWh. The difference between the charging and discharging amount is caused by the 10% charging and discharging loss. Figure 4 shows the distribution of the total discharging amount in each month, which indicates the difference in agent behavior for two operators. It can be observed that in the summer the agent discharges more for the solar park, and in the winter this amount is higher for the wind farm.

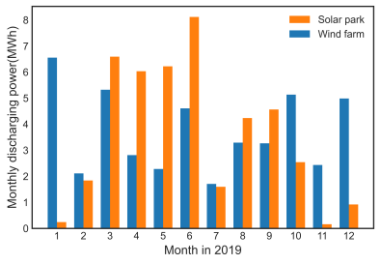

Figure 4. Comparison of the total discharging amount in each month in 2019

To perform a more granular comparison and analyze the difference of the agent behavior described above, we visualize the actions chosen by the agent for both operators under different summer and winter RG and price data over one week. The charging and discharging results are shown in Figure 5. The actual power prices are scaled into the interval of [0,1.5] for better visualization.

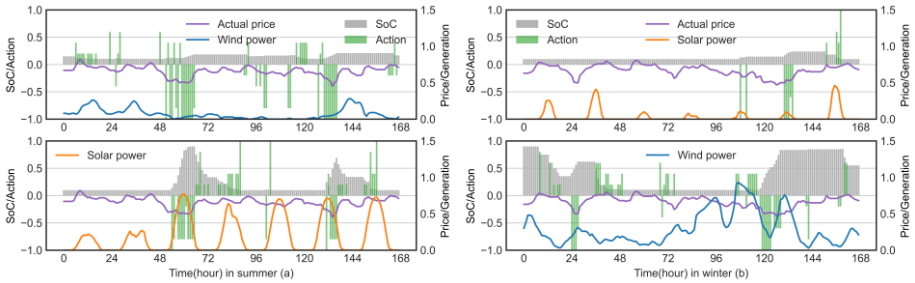

Figure 5. The charging/discharging results over one week for the solar park and the wind farm in (a) summer and (b) winter. (Grey bar: SOC; Green bar: charging (-) / discharging (+) actions; The curves with the right axis represents price and generation)

It can be observed that the agent behaves differently for the two operators. For the solar park, the agent is more active in summer due to the significant inverse correlation between generation and prices. Significant lower-than-average prices occur, for

instance, in hours 48-72 and 132-144 during the summer week when solar power generation is high, while relatively consistent trends of the generation and prices movement can be observed in hours 0-96 in the winter week. For the wind farm, the agent is active in both summer and winter. However, we observe a larger fluctuation of the wind power generation during the winter. Significant inverse correlation between generation and prices can be observed in winter as well, e.g., in hours 0-24 and 96-168 during the winter.

Additionally, we can observe that the charging and discharging timing are consistent with the fluctuation pattern of generation and prices for both operators as well as in both seasons, which shows that the well-trained agent can fully explore and exploit the interrelationship between RG and market prices to generate optimized strategies.

4.4 Evaluation of Revenue Variability

In this paper, we use the 90%-CVaR of negative deviations in daily revenue obtained with and without the optimization of the proposed DQN strategy to evaluate the revenue variability. Figure 6 shows the comparison for both operators in 2019.

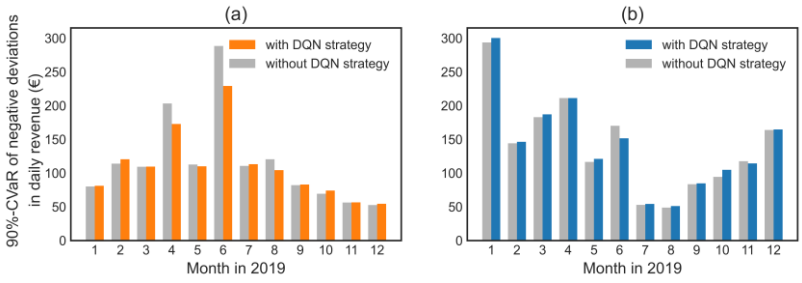

Figure 6. The 90%-CVaR with DQN strategy for (a) the solar park in summer is effectively decreased, for (b) the wind farm is barely reduced except for June.

Without the optimization of the proposed DQN strategy, higher CVaR values are observed for January, March, April, June and December for the wind farm, and in April and June for the solar park. This is consistent with the analysis in 4.3 that the inverse correlation between RG and prices are higher in summer for the solar park and occurs in both summer and winter for the wind farm. A closer look at the values shows that the proposed DQN strategy can significantly reduce the CVaR in April and June for the solar park. However, the CVaR for the wind farm is barely reduced through the DQN strategy except for June. The results indicate that the revenue variability in the summer months can be effectively reduced by coping with the inverse correlation between generation and prices, while the proposed DQN strategy only have limited effect on reduction of revenue variability for the wind farm in the winter.

5 Discussion

Looking back at the simulation results in 4.2 and 4.4, the proposed strategy can reduce the revenue variability for the solar park but only has limited impact for the wind farm. This difference might be explained by different causes of the revenue variability for two generation technologies. Apparently, the large inverse correlation between solar power generation and market prices causes a high CVaR over the summer for the solar park, while the considerable fluctuation of wind power generation has a stronger impact on the revenue variability of the wind farm. The proposed model aims to reduce the negative deviation in daily revenue by adjusting the energy supply at times when there is inverse correlation between generation and prices, and thus has the desired effect for the solar power generator but a limited effect in the wind power operator.

Additionally, the cyclic costs of 20€/MWh are set quite low in view of today's storage costs, which could be potentially uneconomical for the service provider. However, according to [24], costs for battery storage are expected to fall to 100 - 200 €/kWh depending on the technology until 2030. Assuming a cyclic lifetime of 10.000 [25], cyclic costs of 20€/MWh could become feasible. Meanwhile, we assume that the market price spread in the future will increase due to rising shares of intermittent RG, and a higher cyclic cost could be more profitable for both parties. Future works could also go into more detail regarding the opportunity costs of the BSS and whether it would always be available for the service requests of the renewable operator.

6 Conclusion and Future Works

In this paper, we apply DRL to maximize the profit of operators of intermittent renewable generation capacity as well as to reduce the variability of their revenue using a BSS service. We model the BSS service provider as a separate market entity, which charges the renewable operator for using the BSS. In the proposed DRL model, we define the action space as discrete relative values regarding to the maximal available charging and discharging power and design the penalty in the reward function to cope with the intermittent RG. We use CVaR in daily revenue to estimate the revenue variability. The evaluation results using empirical market prices show that it is economically viable to use a BSS service for a simulated wind farm and solar park. The proposed model can effectively improve the revenue stability for the solar park by coping with the inverse correlation of generation and market prices but has limited impact on reducing the negative deviation in daily revenue for the wind farm. Our findings also indicate that the negative deviations in daily revenue for the solar park and the wind farm are caused by different mechanisms. Future works should focus on the impact of the fluctuations of wind power generation and go into more detail on the opportunity costs of the BSS.

References

1. Gesetz für den Ausbau erneuerbarer Energien (Erneuerbare-Energien-Gesetz - EEG 2017), https://www.gesetze-im-internet.de/eeg_2014/EEG_2017.pdf (Accessed: 04.08.2021)
2. Gesetz für den Ausbau erneuerbarer Energien, https://www.gesetze-im-internet.de/eeg_2014/ (Accessed: 04.08.2021)
3. Morales, J.M., Conejo, A.J., Perez-Ruiz, J.: Short-Term Trading for a Wind Power Producer. IEEE Trans. Power Syst. 25, 554–564 (2010)
4. Paraschiv, F., Erni, D., Pietsch, R.: The impact of renewable energies on EEX day-ahead electricity prices. Energy Policy 73, 196–210 (2014)
5. Sensfuß, F., Ragwitz, M., Genoese, M.: The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany. Energy Policy 36, 3086–3094 (2008)
6. Gatzert, N., Kosub, T.: Risks and risk management of renewable energy projects: The case of onshore and offshore wind parks. Renewable and Sustainable Energy Reviews 60, 982–998 (2016)
7. Smith, S.C., Sen, P.K., Kroposki, B.: Advancement of energy storage devices and applications in electrical power system. In: 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, pp. 1-8. IEEE (2008)
8. Yang, J.J., Yang, M., Wang, M.X., Du, P.J., Yu, Y.X.: A deep reinforcement learning method for managing wind farm uncertainties through energy storage system control and external reserve purchasing. International Journal of Electrical Power & Energy Systems 119, 105928 (2020)
9. Cao, J., Harrold, D., Fan, Z., Morstyn, T., Healey, D., Li, K.: Deep Reinforcement Learning-Based Energy Storage Arbitrage With Accurate Lithium-Ion Battery Degradation Model. IEEE Trans. Smart Grid 11, 4513–4521 (2020)
10. Huang, B., Wang, J.: Deep Reinforcement Learning-based Capacity Scheduling for PV-Battery Storage System. IEEE Trans. Smart Grid 12, 2272–2283 (2020)
11. Pinson, P., Papaefthymiou, G., Klockl, B., Verboomen, J.: Dynamic sizing of energy storage for hedging wind power forecast uncertainty. In: 2009 IEEE Power & Energy Society General Meeting, pp. 1–8. IEEE (2009)
12. D. Krishnamurthy, C. Uckun, Z. Zhou, P. R. Thimmapuram, A. Botterud: Energy Storage Arbitrage Under Day-Ahead and Real-Time Price Uncertainty. IEEE Transactions on Power Systems 33, 84-93 (2018)
13. Zéphyr, L., Anderson, C.L.: Stochastic dynamic programming approach to managing power system uncertainty with distributed storage. Comput Manag Sci 15, 87–110 (2018)
14. Zhang, Z., Zhang, D., Qiu, R.C.: Deep reinforcement learning for power system: An overview. CSEE JPES 6 (2019)
15. Wang, H., Zhang, B.: Energy Storage Arbitrage in Real-Time Markets via Reinforcement Learning. In: 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE (2018)
16. Rockafellar, R.T., Uryasev, S.: Optimization of Conditional Value-at-Risk. Journal of risk 2, 21–42 (2000)

17. Sutton, R.S., Barto, A.G.: Reinforcement learning. An introduction. The MIT Press, Cambridge Massachusetts (2018)
18. Watkins, C.J., Dayan, P.: Q-learning. Machine learning 8(3-4), 279–292 (1992)
19. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015)
20. SMARD | Marktdaten, https://www.smard.de/home (Accessed: 04.08.2021)
21. Renewables.ninja, https://www.renewables.ninja/ (Accessed: 04.08.2021)
22. Pfenninger S, Staffell I.: Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. Energy 114, 1251-1265 (2016)
23. Staffell I, Pfenninger S.: Using bias-corrected reanalysis to simulate current and future wind power output. Energy 114, 1224-1239 (2016)
24. Energiewende im Kontext von Atom- und Kohleausstieg. Perspektiven im Strommarkt bis 2040, https://www.solarwirtschaft.de/wp-content/uploads/2020/08/EUPD_Energiewende_Studie_Update_2020_webversion.pdf (Accessed: 12.11.2021)
25. Electricity storage and renewables: costs and markets to 2030, https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2017/Oct/IRENA_Electricity_Storage_Costs_2017.pdf (Accessed: 12.11.2021)