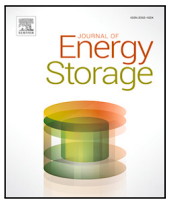
Journal of Energy Storage 104 (2024) 114377

Contents lists available at ScienceDirect

Journal of Energy Storage

journal homepage: www.elsevier.com/locate/est



Research papers

Distributional reinforcement learning-based energy arbitrage strategies in imbalance settlement mechanism

Seyed Soroush Karimi Madahi [a,*], Bert Claessens [b,a], Chris Develder [a]

[a] IDLab, Department of Information Technology, Ghent University – imec, Technologiepark-Zwijnaarde 126, 9052 Ghent, Belgium
[b] BEEBOP, Belgium

ARTICLE INFO

Keywords:
Battery energy storage systems (BESS)
Distributional soft actor–critic (DSAC)
Imbalance settlement mechanism
Reinforcement learning (RL)
Risk-sensitive energy arbitrage

ABSTRACT

Growth in the penetration of renewable energy sources makes supply more uncertain and leads to an increase in the system imbalance. This trend, together with the single imbalance pricing, opens an opportunity for balance responsible parties (BRPs) to perform energy arbitrage in the imbalance settlement mechanism. To this end, we propose a battery control framework based on distributional reinforcement learning. Our proposed control framework takes a risk-sensitive perspective, allowing BRPs to adjust their risk preferences: we aim to optimize a weighted sum of the arbitrage profit and a risk measure (value-at-risk in this study) while constraining the daily number of cycles for the battery. We assess the performance of our proposed control framework using the Belgian imbalance prices of 2022 and compare two state-of-the-art RL methods, deep Q-learning and soft actor–critic (SAC). Results reveal that the distributional soft actor–critic method outperforms other methods. Moreover, we note that our fully risk-averse agent appropriately learns to hedge against the risk related to the unknown imbalance price by (dis)charging the battery only when the agent is more certain about the price.

1. Introduction

Climate change has been a motivation for transitioning toward a decarbonized electricity grid on both the supply and the demand side. The European Commission aims to reach carbon neutrality by 2050 [1]. To achieve this target, the penetration of renewable energy sources (RES) needs to dramatically increase. The International Renewable Energy Agency's report of 2023 states that the total power capacity of RES in the world grew from 1.57 TW in 2013 to 3.37 TW in 2022.[1] However, this trend makes electricity generation more uncertain due to the dependence of RES production on weather conditions. Consequently, the increase in the share of RES leads to an increase in the mismatch between generation and consumption.

Given this potentially increasing mismatch between production and consumption, transmission system operators (TSOs) are facing challenges in maintaining the balance of the grid. Following the liberalization of the European electricity system, the balancing responsibility of TSOs has been outsourced to balance responsible parties (BRPs) [2]. Each unbalanced BRP is penalized by an imbalance price at the end of each imbalance settlement period. According to the electricity balancing guideline (EBGL), published by the European Network of Transmission System Operators for Electricity (ENTSO-E), the main objective of the imbalance settlement mechanism is to make sure that BRPs support the system balance in an efficient way and to stimulate market participants in restoring the system balance [3]. Also, EBGL states that a single imbalance pricing method should be used to calculate the imbalance cost: the settlement price should be the same for both negative and positive imbalances. Such a single imbalance pricing encourages BRPs to deviate from their day-ahead nomination to help the TSO with balancing the grid and to reduce their cost. The wide usage of RES in addition to the single imbalance pricing provides an opportunity for BRPs to reduce their cost using an arbitrage strategy in the imbalance settlement mechanism. For this purpose, recently battery energy storage systems (BESS) have attracted the attention of BRPs due to their fast response time [4], high efficiency [5], and significant decreases in cost of recent battery technology [6].

Energy arbitrage in this imbalance settlement mechanism is challenging because of high uncertainties in imbalance price and near real-time decision-making. Due to these challenges, as well as the recent change in the imbalance pricing methodology, few research works have been conducted on the arbitrage in the imbalance settlement mechanism [2,7,8]. Most of the cited studies have formulated control strategies for BESS using model-based optimization methods, such as stochastic optimization and robust optimization. Despite their promising results, deploying model-based optimization methods for the

* Corresponding author.
E-mail address: seyedsoroush.karimimadahi@ugent.be (S.S. Karimi Madahi).

[1] https://www.irena.org/Publications/2023/Mar/Renewable-capacity-statistics-2023.

https://doi.org/10.1016/j.est.2024.114377
Received 12 April 2024; Received in revised form 11 October 2024; Accepted 22 October 2024
Available online 6 November 2024

arbitrage problem sometimes is not straightforward due to potential non-convexities in the problem. Furthermore, sometimes these optimization methods, especially stochastic optimization, suffer from high computational time during inference because of solving an optimization problem repeatedly. This restricts the applicability of such optimization methods for problems with relatively short decision-making time intervals, such as our minute-based energy arbitrage problem. In addition, these methods fall short in problems where obtaining an accurate model of the system is difficult or the system is partially observable. For example, modeling a real electricity market is challenging due to partially known model parameters and uncertainties [9].

Given the above challenges, few research works have focused on risk management in the arbitrage problem in the imbalance settlement mechanism. Generally, market participants have different risk preferences. For example, BRPs have more conservative arbitrage strategies in the imbalance settlement mechanism because of highly volatile imbalance prices. In other words, BRPs assign higher weights to scenarios with lower revenues and deviate from risk-neutral decision-making. Thus, to provide a more practical solution, a risk-averse perspective needs to be considered in the arbitrage strategy, while to the best of the authors' knowledge, most previous studies have largely ignored risk management. Moreover, a battery's lifetime mainly depends on its charging/discharging operations. Frequently switching between charging and discharging can significantly reduce the battery cycle life and thus decrease the net profit, due to an increased operational cost of the BESS.

In summary, shortcomings and weaknesses in previous studies of arbitrage strategies are that they: (i) do not consider a risk-sensitive perspective; (ii) do not explore model-free alternatives for an arbitrage problem in the imbalance settlement mechanism; and (iii) do not study an arbitrage problem in the imbalance settlement mechanism with a minute-based decision-making time resolution. To address these shortcomings (further elaborated in Section 2), in this paper, we propose a distributional RL-based control framework for a risk-sensitive energy arbitrage strategy in the imbalance settlement mechanism for BESS. The proposed control framework (Section 3) aims to maximize the arbitrage profit as well as a risk measure by constraining the daily number of cycles for the battery. To the best of our knowledge, our work presented here is the first that adopts distributional RL for BESS to perform energy arbitrage in electricity markets while considering the primary operational constraints of a BESS. Thanks to distributional RL, our proposed framework improves over other methods in energy arbitrage in three aspects: (i) outstanding performance (ii) ability to learn risk-sensitive policies (iii) stability in learning. Besides its ability to introduce risk-awareness, our solution particularly adopts a model-free, data-driven approach. This allows it to efficiently deal with nonlinear/ non-convex objective functions and constraints which risk-awareness typically entails. We believe distributional RL methods are proper methods for risk management, since they learn the complete probability distribution of random returns instead of the expected return. The proposed control framework can be tuned according to the risk preference of BRPs from a fully risk-averse perspective to a fully risk-seeking one. In this paper, we start from two state-of-the-art reinforcement leaning (RL) methods, i.e., deep Q-learning (DQN), as a value-based method, and soft actor–critic (SAC), as a policy gradient method. We extend these vanilla DQN and SAC methods with a distributional perspective (i.e., DDQN, DSAC) and a risk-aware component in the loss function (Section 4). The performance of the proposed control framework is evaluated on the Belgian imbalance prices of 2022 (Sections 5 and 6). Overall, our contributions in this paper are that we propose a distributional RL-based control framework

(i) that achieves a risk-sensitive arbitrage strategy with a tunable risk tolerance by optimizing a weighted sum of the arbitrage revenue and a risk measure in the imbalance settlement mechanism;
(ii) for BESS to attain implicit balancing in the imbalance settlement mechanism, while considering a constraint on the daily number of cycles;
(iii) for which we compare the performance of value-based and policy gradient RL methods in a highly uncertain trading market.

Next, in Section 2 we first outline previous studies on the energy arbitrage of BESS in electricity markets and highlight research gaps. Section 3 formalizes the problem formulation of energy arbitrage in the imbalance settlement mechanism. Our adopted RL methods for solving it, and the rationale for their use, are then explained in detail in Section 4. Quantitative results on a case study are presented in Section 5, and finally, Section 6 summarizes our overall conclusions.

2. Background and related work

Energy arbitrage is a technique to achieve financial profits by purchasing energy when the price is cheap and selling it when the price is expensive [10]. This section provides a review of previous works on the energy arbitrage of energy storage systems from various perspectives and highlights research gaps.

2.1. Energy arbitrage for energy storage systems

Target market: Energy arbitrage can be performed within *a single* electricity market to take advantage of varying prices at different hours. For instance, [11] considers the *day-ahead market*: a day-ahead dispatch model is proposed for a liquid air energy storage coupled with an liquified natural gas (LNG) regasification process in day-ahead electricity and LNG gas markets. On the other hand, [12] focuses on the *real-time* market. In particular, first, they obtain the maximum potential profit from the real-time market using a linear optimization program, assuming perfect foresight for future prices. Then, a shrinking-horizon control algorithm is developed for the energy arbitrage strategy of a BESS in the real-time market, by considering forecast errors on the future real-time prices.

Clearly, arbitrage can also consider a combination of *multiple* markets, to benefit from a price difference between two or more electricity markets. For example, [13] provides deterministic model formulations to aggregate multiple arbitrage opportunities for electricity storage by considering all three short-term markets, i.e., day-ahead, intraday, and real-time markets. In [14], the risk management of BESS bidding is studied in both day-ahead and intraday markets. In [15], the planning framework for electric vehicle (EV) aggregators is proposed to participate in the day-ahead market and to react to imbalance prices.

Energy arbitrage algorithm: Some studies have focused on model-based optimization methods to solve energy arbitrage problems. For example [16] uses a bi-level approach for the joint optimization of transmission revenues (using the MW-mile scheme) and day-ahead market participation through a BESS. They use robust optimization to deal with uncertainties. The authors in [17] propose a model predictive control (MPC) framework for designing aging aware arbitrage strategies for BESS in the intraday market. Also stochastic models have been used, e.g., as in [18], which uses it to maximize the energy arbitrage revenue of a BESS under uncertainty in both day-ahead and real-time markets.

On the other hand, several research works have been conducted to obtain optimal arbitrage strategies using model-free RL. Han et al. [19] propose an arbitrage strategy based on Q-learning-based to maximize operating profit in the day-ahead market. Similarly, [20] proposes a deep-RL approach to solve the electricity arbitrage problem in the day-ahead market. As an example for the intra-day market, [21] proposes an RL-based framework for the strategic participation of a BESS.

Risk-sensitive energy arbitrage: The risk of bidding in electricity markets stems from the variance between predicted and actual values of the (i) market price, as well as the BRP's portfolio's overall (ii) demand and (iii) supply. However, risk-sensitive arbitrage strategies in different

markets represent distinct strategic bidding behaviors, as explained next. In both day-ahead and imbalance markets, BRPs are exposed to uncertainty in the market price. However, dealing with the risk related to imbalance prices is more challenging: in the *day-ahead market*, there is a strong correlation between day-ahead prices and time periods. Participants typically have sufficiently accurate predictions of peak and off-peak time periods in the day-ahead market. On the other hand, *imbalance* prices are highly uncertain with a weak correlation between imbalance prices and time periods [22]. Consequently, in the imbalance settlement mechanism, the degree of reliance on the predicted imbalance price plays a crucial role in specifying BRPs' risk preference. The risk related to error in the prediction of demand and supply mainly affects BRPs' bidding strategy in the day-ahead market. High prediction error causes an incorrect day-ahead nomination for BRPs, resulting not only in an extra day-ahead cost but also potentially leading to a significant imbalance settlement due to their large deviation from the day-ahead nomination. Articles [23,24] are examples of studies that consider risk management in energy arbitrage problems.

BESS lifetime: The BESS lifetime is a crucial factor in the financial assessment of the energy arbitrage strategy as the operational strategy of BESS significantly influences its lifespan. There are several ways to consider BESS lifetime in arbitrage problems: the authors in [25] prevent frequent cycling by adding a discharge cost to the objective function. In [13], the BESS cycle-life is included as a depreciation cost, which is positive if the cycling rate is greater than the targeted cycling rate, zero otherwise. In [26], the aging cost of BESS is formulated as a function of depth of discharge. Some papers have proposed an RL-based strategy for energy arbitrage, while considering BESS degradation cost [27–30]. Note that in the current work, we constrain the annual number of BESS cycles, reserving considering detailed modeling of BESS degradation costs for future work.

2.2. Research gaps and contributions

Table 1 shows an overview of previous works on energy arbitrage. In terms of target market scenario, we focus on energy arbitrage in the *imbalance* settlement mechanism. The recent change in the imbalance price calculation [3] and an increase in imbalance prices have opened up a new arbitrage opportunity in electricity markets. Fig. 1 demonstrates the rise in Belgian imbalance prices in recent years. However, only few studies have been conducted on energy arbitrage in the imbalance settlement mechanism, due to the high risk involved in this arbitrage. The authors in [2] first implement a new tailored encoder–decoder architecture to generate improved probabilistic forecasts of the future system imbalance. Then, they solve a bi-level robust optimization problem to maximize the revenue from the participation of a BESS in the imbalance settlement. The authors in [7] introduce a novel stochastic model predictive control (MPC) approach to optimize the revenue of BESS in the imbalance settlement mechanism by taking into account battery degradation costs and risk aversion. More specifically, an attention-based recurrent neural network is used to predict the system imbalance and imbalance price. In [8], control strategies are proposed for seasonal thermal energy storage systems to interact with day-ahead and imbalance markets: MPC-based and RL-based controllers are developed for each market interaction to compare the performance of these two controllers in the different electricity markets.

As indicated in Table 1, regarding energy arbitrage algorithm methodology, most previous research works adopt *model-based* optimization methods to solve the arbitrage problem. These methods sometimes require linearization techniques (such as piecewise linear approximation) to transform a nonlinear problem into a linear or mixed-integer convex problem. However, applying these techniques may result in an intractable optimization problem or an inaccurate approximation of the problem. Moreover, these model-based methods need a (probabilistic) forecaster for future imbalance prices to address uncertainty in future prices. In stochastic optimization, such uncertainties can be handled by generating a set of scenarios. Yet, as imbalance prices are highly uncertain, a large number of scenarios are required to correctly reflect the imbalance price distribution, which increases the computational burden to the extent that the problem may become computationally intractable. On the other hand, although robust optimization does not need as many scenarios [31], its obtained solution might be a very conservative strategy and not necessarily the most economical one [32]. Another limitation of the mentioned studies is that only a few of them propose a risk-sensitive arbitrage strategy while considering the lifetime of BESS.

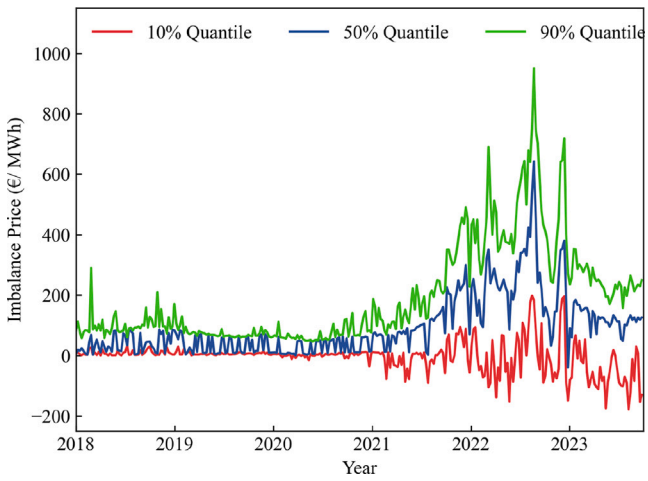


Fig. 1. The evolution of Belgian imbalance prices from 2018 to 2023. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To avoid problems of model-based optimization methods, *RL methods* can be used. RL can learn a (near-)optimal policy for a stochastic nonlinear environment by directly interacting with the environment [33]. In RL, there is no special hypothesis regarding the reward function: it can be linear or nonlinear. In contrast to model-based optimization methods, model-free RL methods do not need prior knowledge or an explicit model of the environment. The agent, by interacting with the environment, captures uncertainties and estimates system dynamics. Another advantage of RL methods is that after training the RL agent, its learned policy can be directly used in a new test setting without requiring solving any optimization problem. Therefore, RL methods are efficient tools for real-time control [34]. The usage of distributional RL elevates our proposed framework above previous RL-based arbitrage methods by providing a framework for risk management and achieving state-of-the-art performance.

3. Problem formulation

In this section, the imbalance settlement mechanism is explained in detail (Section 3.1) and the Markov decision process (MDP) formulation of the energy arbitrage problem in the imbalance settlement mechanism is provided (Sections 3.2 and 3.3).

3.1. Imbalance settlement mechanism

BRPs are responsible for continuously balancing their individual demand and supply. But sometimes BRPs deviate from their traded consumption and generation due to uncertainties in the grid. The total imbalance volume of all BRPs in a single control area is called the total system imbalance [35]. Positive and negative values of the system imbalance indicate the excess and shortage of the generation, respectively. A TSO corrects the system imbalance in real-time by activating reserve capacities offered in the balancing market [36].

Table 1
Comparison with literature.

Paper	Algorithm category		Market					Risk-sensitive	BESS lifetime	Time resolution
	Model-based optimization	RL	DA[a]	ID[b]	RT[c]	BM[d]	ISM[e]			
[13]	✓		✓	✓	✓			✗	✓	15 min
[16]	✓		✓					✓	✗	1 h
[23]	✓		✓	✓		✓		✓	✓	unknown[g]
[17]	✓			✓				✗	✓	15 min
[24]	✓		✓	✓				✓	✗	1 h
[25][h]	✓		✓		✓			✗	✓	5 min
[2]	✓						✓	✓	✗	15 min
[11]	✓		✓					✗	n/a[f]	1 h
[18]	✓		✓		✓			✗	✗	1 h
[12][h]	✓		✓		✓			✗	✗	1 h
[14]	✓		✓	✓				✓	✓	1 h
[7][h]	✓		✓				✓	✓	✓	15 min
[8][h]	✓	✓	✓				✓	✗	n/a	15 min
[20]	✓	✓	✓					✗	✗	1 h
[19]		✓	✓					✗	✗	1 h
[21]		✓		✓				✗	✗	15 min
[26]		✓	✓					✗	✓	1 h
[27]		✓	✓					✗	✓	1 h
[28]		✓	✓					✗	✓	1 h
[29]		✓	✓					✗	✓	1 h
[30]		✓	✓		✓			✗	✓	1 h
[15]		✓	✓				✓	✗	✗	15 min
Ours		✓					✓	✓	✓	1 min

[a] DA: day-ahead market.
[b] ID: intraday market.
[c] RT: real-time market.
[d] BM: balancing market.
[e] ISM: imbalance settlement mechanism.
[f] n/a: not applicable; paper does not consider BESS.
[g] The paper does not explicitly list the timescale used.
[h] These papers studied energy arbitrage within a single market for multiple markets.

A TSO charges BRPs for their imbalance at a price specific to the imbalance settlement period (15 min in most European markets). This mechanism is known as imbalance settlement. The imbalance price is dependent on the reserve volume activated by the TSO [37]. In each imbalance settlement period, the negative imbalance price is equal to the highest activated upward reserve offer (marginal incremental price), and the positive imbalance price is determined by the lowest activated downward reserve offer (marginal decremental price) [38]. Three main imbalance pricing methodologies are used in various countries: (1) dual pricing; (2) two-price settlement; and (3) single pricing [38].

In the dual pricing method, the imbalance price is different for positive and negative imbalances. BRPs penalize for negative and positive imbalances using the marginal incremental price (MIP) and marginal decremental price (MDP), respectively. This pricing method motivates BRPs to keep the balance within their own portfolio without being concerned about the total system imbalance. The main drawback of this method is that there is no incentive for BRPs to deviate from their nomination to restore the grid. For instance, if the total system imbalance is positive and there is a BRP that can reduce this imbalance, then this BRP is not incentivized, but even penalized for deviating from its day-ahead nomination.

In the two-price settlement method, similar to the dual pricing method, different imbalance prices are considered for each imbalance direction. The difference with the dual pricing method is that if the imbalance direction of BRPs is opposite to the total system imbalance direction, the imbalance price is the same as the day-ahead price. Although in this pricing method, BRPs do not face penalties due to their deviation for helping TSO with restoring the grid, the imbalance price is not attractive to create a portfolio imbalance for supporting the grid (typically, day ahead prices are lower than imbalance prices).

In the single pricing method, the imbalance price is the same for both imbalance directions and depends on the total system imbalance. This pricing method provides an opportunity for BRPs to reduce their cost by supporting the grid. For instance, if the total imbalance price is negative and a BRP creates a positive imbalance, the BRP will receive an MIP (imbalance price) which is usually higher than the day-ahead price. In some countries, e.g., Germany, despite using the single pricing method, arbitrage in the imbalance settlement mechanism is prohibited and market players are expected to trade honestly in the markets [39]. Nonetheless, the arbitrage in the imbalance settlement mechanism is a win-win situation for both BRPs and TSOs. On the one hand, BRPs can profit from the arbitrage and indirectly reduce the total system imbalance. On the other hand, this decrease in the total system imbalance results in a lower imbalance price since the TSO does not need to activate more expensive reserve offers.

As mentioned earlier, ENTSO-E aims to harmonize the imbalance settlement mechanism in Europe by implementing the single pricing method for calculating the imbalance price with a 15 min imbalance settlement period. For this reason, the focus of this paper is on the single pricing method. The Belgian imbalance settlement mechanism is a good case study for this research work because since the beginning of 2020, it adopts the single pricing method with a 15 min settlement period [35].

3.2. MDP formulation without cycle constraint consideration

The energy arbitrage problem can be formulated as an MDP. An MDP provides a mathematical framework for stochastic sequential decision-making problems and is modeled by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the (discrete) action space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the immediate reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the unknown state transition probability distribution, and $\gamma \in (0, 1]$ is the discount factor [40]. At each time step t, the agent observes the environment state $s_t \in \mathcal{S}$ and takes an action $a_t \in \mathcal{A}$ based on the current state. As a consequence of the taken action, the agent receives a reward value $\mathcal{R}(s_t, a_t)$ and moves to a new state $s_{t+1} \in$

$\mathcal{S}$ with the probability determined by the state transition probability distribution $\mathcal{P}(s_{t+1}|s_t, a_t)$. In the energy arbitrage problem, the agent is a decision maker who decides about the charging/discharging of BESS at each time step. The environment is the external context with which the agent interacts (electricity markets, grid, etc.). We define the MDP formulation of the energy arbitrage problem in the imbalance settlement mechanism without cycle constraints as follows:

(i) *State*: The state at each time step is expressed as

$$s_t = (T_{\text{qh}}, \text{qh}, \text{mo}, \text{SOC}_t, \hat{\pi}_t^{\text{imb}}) \tag{1}$$

where T_{qh} represents the minute of the quarter hour, qh is the quarter hour of the day, mo is the month of the year, SOC_t is the SoC of BESS at time t, and $\hat{\pi}_t^{\text{imb}}$ is the forecasted imbalance price of the current quarter-hour. We used a forecast of the imbalance price because the real imbalance price of the quarter hour is only calculated once the quarter hour is over.

(ii) *Action*: We consider a discrete action space with 3 possible actions, as follows:

$$a_t \in \mathcal{A}, \quad \mathcal{A} = \{-P_{\text{max}}, 0, P_{\text{max}}\} \tag{2}$$

where P_{max} is the maximum (dis-)charging power of the BESS. The action a_t represents a decision on the charging/discharging power at time t. We assume discrete actions, based on lessons learnt from [41]. That study investigated the emergence of extremal switching, or bang–bang behavior, in continuous control RL. They showed competitive performance of RL methods with discrete action space (bang–bang policy) on standard continuous control benchmarks. For this reason, we focus on a discrete action space, leaving further investigation of continuous control in our arbitrage problem as future work.

Further, the agent's actions must respect the maximum and minimum charge and discharge constraints, as well as SoC limits. There are several ways to incorporate such constraints into RL, e.g., adding penalty terms to the reward function [42]. However, the main challenge of adding penalty terms, apart from the difficulty of tuning their weights in the reward function, is that they represent soft penalties. In other words, they only *encourage* meeting constraints, rather than *enforcing them*. To strictly enforce these constraints, we used a function that overrides the action taken by the agent if need be. The function clips the charging power within the feasible range defined by the maximum and minimum limits. Similarly, it blocks the charge action when the SoC exceeds the defined maximum, and the discharge action when the SoC falls below the defined minimum.

(iii) *Reward*: The objective of the agent is to maximize the revenue by buying energy when the imbalance price is low and selling it when the imbalance price is high. Hence, the reward function to be maximized is the negative of the energy cost, defined as follows

$$r_t = -a_t \pi_t^{\text{imb}} \tag{3}$$

where π_t^{imb} is the real imbalance price of the quarter hour in which t lies.

(iv) *State transition function*: In the MDP framework, system dynamics are described by a state transition probability function $\mathcal{P}$. This probability function is unknown in the energy arbitrage problem because of uncertainties in the imbalance price. The agent strives to estimate the state probability distribution through interactions with the environment. However, the state transition for SOC_t is controlled by a_t and can be explicitly formulated as below.

$$\text{SOC}_{\text{t+1}} = \begin{cases} \text{SOC}_{\text{t+1}}^{\text{temp}} & : 0 < \text{SOC}_{\text{t+1}}^{\text{temp}} < 1 \\ 0 & : \text{SOC}_{\text{t+1}}^{\text{temp}} < 0 \\ 1 & : \text{SOC}_{\text{t+1}}^{\text{temp}} > 1 \end{cases} \tag{4}$$

$$\text{SOC}_{\text{t+1}}^{\text{temp}} = \text{SOC}_t + (\max(a_t, 0)\eta_{\text{cha}} + \frac{\min(a_t, 0)}{\eta_{\text{dis}}})\frac{\Delta t}{C_{\text{BESS}}} \tag{5}$$

where C_{BESS} is the maximum capacity of the BESS, and η_{cha} and η_{dis}, denote the charging and discharging efficiency of the BESS, respectively.

3.3. MDP formulation with cycle constraint consideration

Frequent charging/discharging cycles cause an extra cost because they expedite the degradation of BESS. Modeling the aging of BESS is crucial as it indicates a capital loss of BESS investment costs [43]. Due to the dependence of battery lifetime on its operational strategy, the lifetime of a BESS plays an important role in the financial evaluation of the energy arbitrage strategy. Usually, the lifetime of a BESS is determined by the number of complete charge–discharge cycles before its nominal capacity becomes lower than a certain level of its initial rated capacity [44]. Thus, we constrain the daily number of cycles, since it aligns with the designed lifetime and guarantee provided by manufacturers [45]. The MDP formulation with cycle constraint consideration is described next.

(i) *State*: The state is given by

$$s_t = (T_{\text{qh}}, \text{qh}, \text{mo}, \text{SOC}_t, \hat{\pi}_t^{\text{imb}}, n_t^{\text{cyc}}) \tag{6}$$

$$n_t^{\text{cyc}} = \sum_{\text{i=0}}^{\text{t-1}} \frac{|\min(a_i, 0)|\Delta t}{C_{\text{BESS}}} \tag{7}$$

where n_t^{cyc} is the daily consumed number of cycles, calculated using (7).

(ii) *Action*: Similar to the MDP formulation without cycle constraints, the action space is discrete with 3 possible actions. The action is determined as follows

$$a_t = B(u_t, n_t^{\text{cyc}}), \quad u_t \in \mathcal{A} = \{-P_{\text{max}}, 0, P_{\text{max}}\} \tag{8}$$

$$B(u_t, n_t^{\text{cyc}}) = \begin{cases} 0 & : u_t < 0 \wedge n_t^{\text{cyc}} > n_{\text{max}}^{\text{cyc}} \\ u_t & : else \end{cases} \tag{9}$$

where $n_{\text{max}}^{\text{cyc}}$ is the maximum allowed daily number of cycles and $B(.)$ is a backup controller to ensure the daily cycle constraint. The backup controller is used to override the agent action (u_t) when the agent wants to discharge the battery and the daily number of cycles exceeds the maximum allowed value. The introduced backup controller $B(.)$ is part of the explained function in Section 3.2 to enforce the cycle constraint.

(iii) *Reward*: The reward function definition is the same as that of the MDP formulation without cycle constraint.

(iv) *State transition function*: Also the state transition function is the same as that of the MDP formulation without cycle constraint.

4. Reinforcement learning methods

Recently, RL, as a model-free method, has attracted researchers' attention due to its remarkable performance in solving complex sequential decision-making problems such as playing games, robotic control, and autonomous driving. The goal in RL is to learn a policy that maximizes the expected long-term reward. RL methods have been successfully applied to many energy problems such as the smart charging of EVs [46,47], demand response [48], frequency control [49], building control [50], etc. Generally, model-free RL methods can be classified into two categories: value-based methods (e.g., Q-learning, fitted Q-iteration (FQI), DQN, etc.) and policy gradient methods (e.g., actor–critic, deep deterministic policy gradient (DDPG), soft actor–critic (SAC), etc.) [51]. In value-based methods, the Q- (or V-)function is learned (estimated) and the action is chosen based on the learned Q- (or V-)function as to maximize it. On the other hand, policy gradient methods directly learn the policy. In [52], the SAC method has been

proposed as an off-policy actor–critic method. In SAC, the policy is learned by an actor network and the Q-function is estimated by a critic network. The actor aims to maximize the expected reward as well as the entropy of the actor, to encourage the agent to explore the environment more. In this paper, we will use the DQN (as a state-of-the-art method in value-based methods) and SAC (as a state-of-the-art method in policy gradient methods) methods to solve the arbitrage problem formulated as an MDP.

Next, we first highlight main advantages of using RL methods over model-based optimization methods to solve the defined arbitrage problem. Subsequently, we detail the two RL methods adopted in this paper, i.e., DQN and SAC. Finally, we introduce the distributional perspective on RL and the risk-sensitive RL framework.

4.1. Why RL methods?

Model-free RL methods have two main advantages over model-based optimization methods in the defined BESS arbitrage problem:

(1) *Dealing with non-convexity:* in the defined BESS arbitrage problem, there are two parts that may introduce non-convexities: (a) the objective function, and (b) the effect of the agent's actions on the imbalance price. In the risk-neutral case, the *objective function* is defined to maximize the arbitrage revenue (Eq. (3)), which is a linear objective function. It is worth to mention that although in this paper the objective function in the risk-neutral scenario is linear, it can generally be nonlinear or non-convex. On the other hand, in the risk-averse case, the objective is to optimize the weighted sum of the arbitrage profit and a risk measure, which can lead to a nonlinear/non-convex objective function, as some risk metrics are non-convex (e.g., value-at-risk (VaR)). To solve the risk-involved arbitrage problem using model-based optimization techniques, the problem must be formulated either using convex risk metrics (mainly conditional value-at-risk (CVaR)) or applying linear approximation to non-convex risk metrics. For this reason, the choice of risk measure functions in the model-based optimization methods is mainly limited to convex metrics, specifically CVaR. Although the cycle constraint is linear in this paper, a more detailed degradation modeling of the battery can introduce another source of non-linearity to the problem (e.g., BESS degradation cost in [7]). Now, regarding the *effects of the agent's actions on imbalance price*: at each time step, the agent's action can affect the overall system imbalance, which ultimately implies an impact on the imbalance price for the relevant quarter-hour. By considering this impact in the arbitrage problem, the problem becomes a non-convex optimization problem due to a set of non-convex constraints and partially known model parameters in the market model [9]. In this research work, for simplicity, we assume that the agent is of a small scale and does not significantly affect the system imbalance. In other words, our assumption is that the agent is price-taker. Considering a price-maker agent is one of our directions for future work. Similarly, in the current work, for simplicity we ignore the potential effect of other competing agents on the imbalance price, which would introduce another source of nonlinearity to the problem.
In contrast to model-based optimization methods, model-free RL methods directly learn a (near-)optimal policy for a stochastic nonlinear environment. These RL methods do not have any specific hypothesis concerning the reward function, which means they can address sequential non-convex optimization problems without applying linearization techniques. This means that the proposed control framework does not impose any restriction on the definition of nonlinear objective functions or the use of non-convex risk metrics. Also, RL methods do not require a system model (such as an electricity market model) and can implicitly learn this model through interaction with the environment.

(2) *Tractability and computational complexity:* Model-based optimization methods mostly handle uncertainties using stochastic optimization or robust optimization. Nevertheless, since imbalance prices are highly uncertain, numerous scenarios are required to correctly reflect the imbalance price distribution, which increases the computational burden to the extent that the problem may become computationally intractable. On the other hand, although robust optimization may not need as many scenarios, its obtained solution can tend to be extremely cautious.
That computational time at inference is a notable concern in the considered arbitrage problem in the imbalance market, given the relatively short decision-making time interval (1 min in our case). In this setting, we believe RL methods are suitable for real-time control, given their faster inference compared to model-based optimization methods. In RL methods, during inference, the trained model is directly used to take an action at each time step without the need for repeatedly solving an optimization (as seen in model-based optimization methods).

4.2. DQN

Classical tabular RL methods, e.g., Q-learning, suffer from an issue known as the curse of dimensionality. Since these methods can only be applied to problems with discrete state space, they cannot be used for problems with high-dimensional or continuous state space. In addition, these methods usually need handcrafted state representations [51]. To overcome these limitations, the DQN method uses a deep neural network as a function approximator to estimate the Q-value function parametrized by θ. The Q-value function $Q_\theta(s_t, a_t)$ is learned by minimizing the following loss function:

$$L_Q(\theta) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1})\sim D}\left[\left(r_t + \gamma \max_a Q_{\theta'}(s_{t+1}, a) - Q_\theta(s_t, a_t)\right)^2\right]. \tag{10}$$

The first benefit of DQN is its stability in learning. In [53], two techniques are used to stabilize the learning process. First, the target Q-function $Q_{\theta'}(s_t, a_t)$ is used to calculate next state–action values in Eq. (10). Parameters of the target Q-function (θ') are periodically updated with the most recent θ. Second, agent past transitions are stored in an experience replay buffer (D) and for training Q_θ, mini-batches of experiences are sampled from D. Another benefit of the DQN method is that this method is an off-policy method. The key advantage of off-policy methods is their capacity to learn from historical data since using the current experiences as the training set can easily overfit the policy because the training samples are not independent [54]. In an off-policy setting, a policy learned by the agent is different from a behavior policy used for collecting historical data. Using past transitions for training can significantly improve sample efficiency.

4.3. SAC

Value-based methods have some limitations. The application of these methods is limited to problems with a discrete and low-dimensional action space. Also, these methods learn a deterministic policy, which means for a given state, an action taken by the agent is always the same. Thus, keeping a balance between exploration and exploitation in value-based methods is challenging. Policy gradient methods solve these limitations by learning a policy network that outputs the probability of taking actions in each state. From the existing policy gradient methods, we use SAC because of its superior sample efficiency and stability. In this off-policy method, the policy is learned by an actor network π_ϕ and the Q-function is approximated by a critic network Q_θ. The objective of the actor is to maximize the expected reward as well as maximize the entropy of the actor to encourage the agent to further explore the environment. The loss function of the actor network (J_π) is given by

$$J_\pi(\phi) = \mathbb{E}_{s\sim D, a\sim\pi_\phi}[\alpha \ln \pi_\phi(a|s) - Q_\theta(s, a)] \tag{11}$$

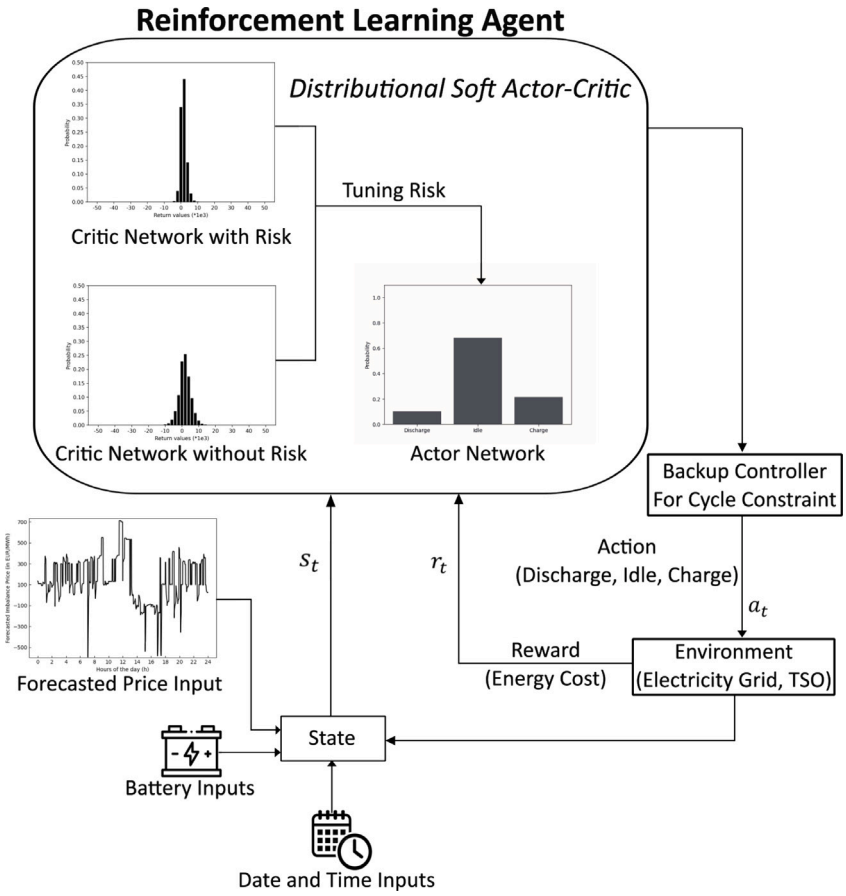


Fig. 2. The overview of the proposed control framework.

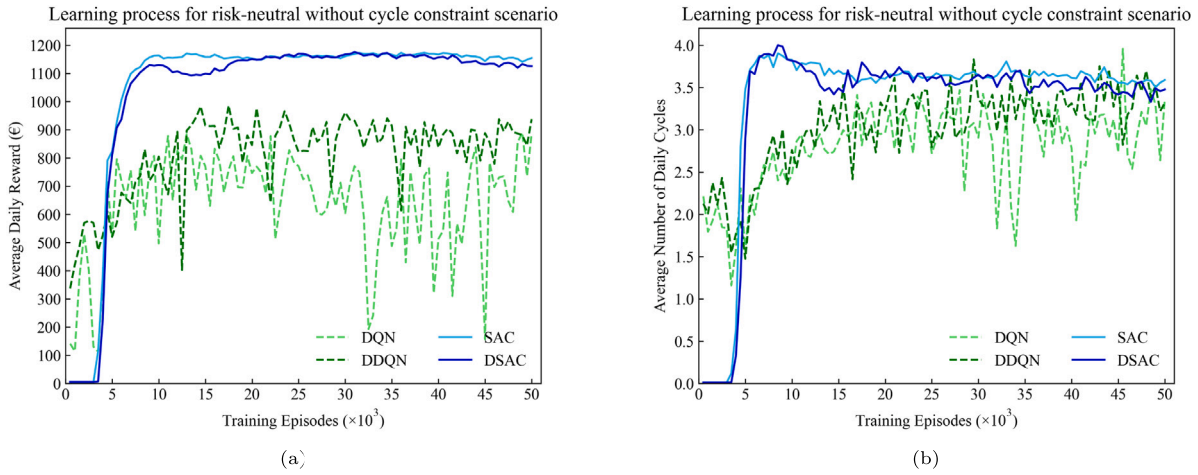


Fig. 3. The learning process of the four RL methods for the risk-neutral without cycle constraint scenario. (a) The average daily profit of the RL methods. (b) The average daily number of cycles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The critic network estimates the soft Q-value function. The loss function of the critic network (L_Q) is formulated as follows:

$$L_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim D}[(y_t - Q_\theta(s_t,a_t))^2] \tag{12}$$

$$y_t = r_t + \gamma \mathbb{E}_{a_{t+1}\sim\pi_\phi}[Q_{\theta'}(s_{t+1},a_{t+1}) - \alpha \ln \pi_\phi(a_{t+1}|s_{t+1})] \tag{13}$$

$$\theta' = \tau\theta + (1-\tau)\theta' \tag{14}$$

In Eq. (13), y_t is an estimated soft-Q-value that is calculated by a modified Bellman equation (the so-called soft Bellman equation). Similar to the DQN method, the target Q-function is used to calculate y_t. After each update of Q_θ, the parameters of $Q_{\theta'}$ are updated according to Eq. (14) with $\tau \ll 1$ to slowly track the learned network [55].

4.4. Distributional RL

A distributional perspective on RL was first introduced in [56]. In distributional RL methods, the probability distribution over returns is estimated rather than a point estimate of the mean. Distributional RL methods offer several advantages, including more stable learning [56],

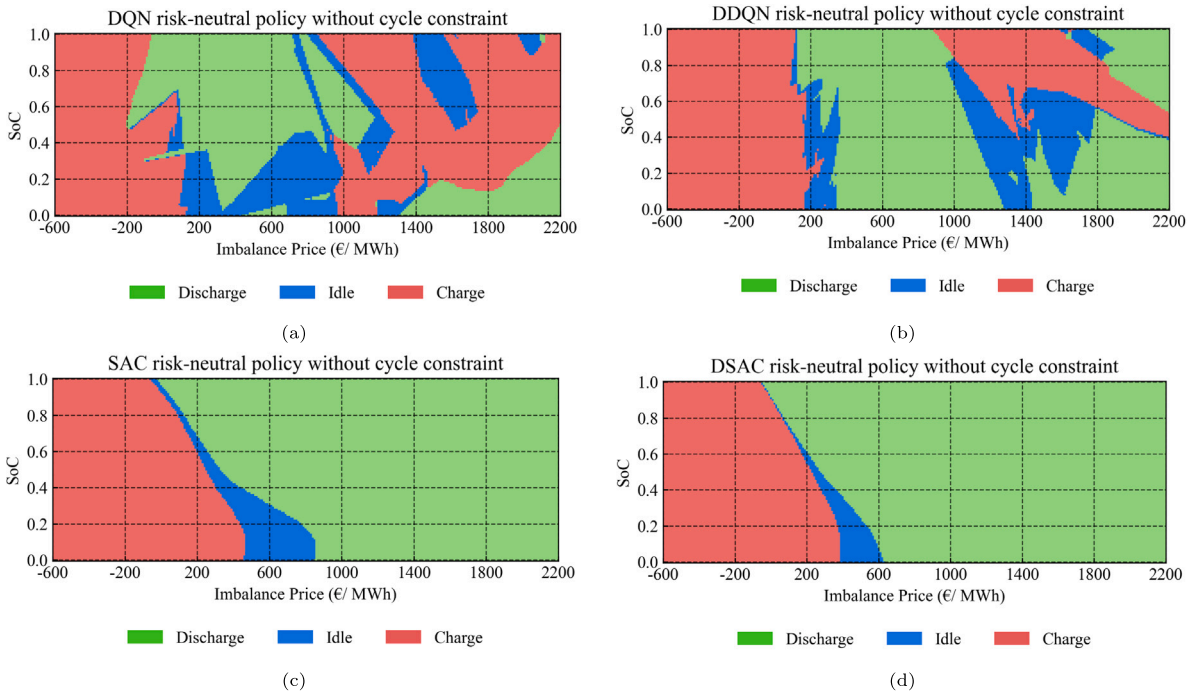


Fig. 4. The projection of the learned policy in the risk-neutral without cycle constraint scenario for (a) DQN, (b) DDQN, (c) SAC, and (d) DSAC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mitigating Q-value overestimation [57], and providing a framework for risk-sensitive learning [58]. In the vanilla DQN method, the core idea is to estimate the Q-value function Q_θ. Going beyond the vanilla DQN method, the distributional DQN (DDQN) method learns the probability distribution of returns ($\mathcal{Z}_\theta$) using the distributional Bellman equation as follows [56]:

$$L_{\mathcal{Z}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim D}[D_{\mathrm{KL}}(\mathcal{T}\mathcal{Z}_{\theta'}(s_t,a_t) \parallel \mathcal{Z}_\theta(s_t,a_t))] \tag{15}$$

$$\mathcal{T}Z(s_t,a_t) \overset{D}{=} r_t + \gamma \max_a \mathbb{E}_{Z\sim\mathcal{Z}_{\theta'}}[Z(s_{\mathrm{r}+1},a)] \tag{16}$$

where $\mathcal{Z}$ is the distribution of returns, $A \overset{D}{=} B$ denotes that two random variables A and B have an equal probability distribution, and $\mathcal{T}\mathcal{Z}_\theta$ indicates the probability distribution of $\mathcal{T}Z$. The distribution of returns can be modeled as a categorical distribution as below.

$$Z(s_t,a_t) = \left\{ z_i \middle| z_i = V_{\min} + \frac{V_{\max}-V_{\min}}{N-1} i, 0 \le i < N \right\} \tag{17}$$

In Eq. (17), $V_{\min}$ and $V_{\max}$ are the maximum and minimum values of random returns, respectively, and N is the number of bins. In distributional SAC (DSAC), the critic network learns the probability distribution of soft returns. The loss function of the critic network in DSAC is similar to Eq. (15), but the calculation of $\mathcal{T}Z(s_t,a_t)$ differs as follows:

$$\mathcal{T}Z(s_t,a_t) \overset{D}{=} r_t + \gamma \mathbb{E}_{a_{\mathrm{t}+1}\sim\pi_\phi, Z\sim\mathcal{Z}_{\theta'}}\left[Z(s_{\mathrm{t}+1},a_{\mathrm{t}+1}) - \alpha \ln \pi_\phi(a_{\mathrm{t}+1}|s_{\mathrm{t}+1})\right] \tag{18}$$

Since the expectation of $Z(s_t,a_t)$ over $\mathcal{Z}_\theta$ is equal to $Q(s_t,a_t)$, the loss function of the actor network is modified as below.

$$J_\pi(\phi) = \mathbb{E}_{s\sim D, a\sim\pi_\phi}[\alpha \ln \pi_\phi(a|s) - \mathbb{E}_{Z\sim\mathcal{Z}_\theta}[Z(s,a)]] \tag{19}$$

4.5. Risk-sensitive RL

By approximating the probability distribution of returns, distributional RL presents a possibility for learning a risk-averse policy. In a risk-neutral RL framework, the agent in each state takes an action that aims to maximize the expected return (Q-value). On the other hand, in the risk-sensitive RL framework, the agent takes an action with the lowest associated risk. The main risk in the arbitrage problem is related to forecasted imbalance prices. The greater the inaccuracy in predicted prices, the higher the associated risk of taking the wrong action.

Risk measures can be used to assess the level of risk associated with a distribution of returns [59]. The loss function of the actor network in the risk-sensitive DSAC can be formulated as follows:

$$J_\pi(\phi) = \mathbb{E}_{s\sim D, a\sim\pi_\phi}\left[\alpha \ln \pi_\phi(a|s) - \mathbb{E}_{Z\sim\mathcal{Z}_\theta}[Z(s,a)] - \beta\Psi[Z(s,a)]\right], \tag{20}$$

where $\Psi[.]$ represents a risk measure function and β is a parameter that controls the trade-off between the expectation value and risk. $\beta = 0$ represents the risk-neutral attitude of the agent. As β increases, the agent becomes more risk-averse. In this paper, value-at-risk (VaR) is applied as the risk measure function:

$$\mathrm{VaR}_\rho(Z) = \inf\{z|\mathrm{CDF}_Z(z) \ge \rho\}, \tag{21}$$

where $\rho \in (0,1]$ is a confidence level. We will set $\rho = 0.1$ in this paper.

5. Simulation results

We will evaluate the performance of the proposed control framework, is explained in Sections 3 and 4, for the energy arbitrage problem.

5.1. Experimental setup

Fig. 2 shows the overview of the proposed control framework, which we test on the Belgian imbalance in 2022 extracted from Elia's website.[2] As mentioned in Section 3.1, Elia publishes two imbalance

[2] https://www.elia.be/en/grid-data/data-download-page?csrt=783739960382611489.

Table 2
Method hyperparameters.

Shared		DQN		SAC	
Parameter	Value	Parameter	Value	Parameter	Value
Discount factor γ	0.9995	Learning rate	5×10^{-4}	Actor learning rate	2×10^{-5}
Soft update factor τ	0.1			Critic learning rate	1×10^{-4}
Experience buffer size	1×10^{6}			Initial α	1
Mini-batch size	16 384			α learning rate	3×10^{-4}
Network hidden layer size	[256,128]				
$V_{\max}$	5000				
$V_{\min}$	−5000				
N	11				

prices: 15-min-based and 1-min-based prices. The reference price for the imbalance settlement of BRPs is the 15-min-based price which is the real imbalance price calculated at the end of the quarter-hour period. The 1-min-based prices, on the other hand, are calculated based on non-validated data, based on the instantaneous system imbalance and prices of cumulative activated regulation volumes on a minute basis. These 1-min-based prices are published to provide additional information to BRPs.[3] We use these non-validated prices as forecasted imbalance prices of the corresponding quarter-hour period. Since the granularity of the forecasted imbalance prices is one minute, the RL agent takes an action every minute. In this work, the day-ahead schedule for the battery is set to zero which means that the battery does not trade in the day-ahead market. However, future work will extend our proposed control framework for arbitrage in both the day-ahead market and imbalance settlement. To train and validate the proposed control framework, the imbalance price dataset is split as follows: the first 20 days of each month are considered as the training set, the 21st day to the 25th day of each month are considered as the validation set, and the remaining days of each month are used as the test set. The considered BESS has a power rating of 1 MW and a maximum capacity of 2 MWh with a round-trip efficiency of 0.9 for both charging and discharging. Since the maximum allowed annual number of cycles for the BESS is 400, the maximum daily number of cycles is set to 1.1. The RL methods are trained with 50 000 episodes and each episode constitutes a full day. The hyperparameters used for the methods are listed in Table 2. The proposed control framework is implemented in Python using the PyTorch package.

We design experiments to answer the following questions:

- Q1: What is the learned arbitrage strategy when there is no limit on the daily number of cycles?
- Q2: How does a daily number of cycles affect the learned arbitrage strategy?
- Q3: What is the effect of the risk-averse perspective on the learned arbitrage strategy?

5.2. Arbitrage strategy without cycle constraint (Q1)

The learning process of the RL methods for the risk-neutral scenario, without considering the cycle constraint, is illustrated in Fig. 3. The performance of the trained RL methods on the test set is indicated in Table 3. Results show that the distributional RL methods outperform the standard RL methods. The reason behind this is that estimating the probability distribution of returns, rather than the expectation of returns, can provide a more stable training target. Also, the distributional RL methods can mitigate instability in the Bellman optimality operator by learning probability distribution of returns. The DDQN method increases the average daily profit by 17% compared to the DQN method. DSAC improves the proportional reward (defined as the ratio of average daily profit to average daily number of cycles) by 2.1% compared to SAC. The comparison between the performance of the distributional and vanilla DQN, and SAC, indicates that the distributional perspective can enhance DQN results to a greater extent. The reason is that the SAC method mitigates instability in the Bellman optimality operator by using an actor network instead of the max operator in the Bellman equation. Therefore, the improvement in the DSAC results is mainly due to stable training target for the critic network. However, the distributional perspective can boost the performance of the vanilla DQN by both providing stable training targets and mitigating instability in the Bellman optimality operator. Results also highlight the superiority of SAC over DQN. This is because SAC can mitigate Q-value overestimations in DQN by replacing the max operator (Eq. (10)) with the expectation operator (Eq. (13)) in the Bellman equation.

To analyze and study the learned policy of the four RL methods, the policy heatmaps are illustrated in Fig. 4. Since SoC and forecasted imbalance price are the two most determinative features for the agent, we show the learned policy with respect to these two input features, which are also informative to interpret the policy. Fig. 4 shows that the SAC and DSAC methods can learn a more meaningful and smooth policy compared to the DQN and DDQN methods. For DQN and DDQN, the Q-value function overestimates the value of rarely seen states and out-of-distribution (OOD) actions in these rare states due to the max operator and the reliance of the estimated Q-values on inputs from the same distribution as its training set. This overestimation results in policies that choose OOD actions. According to Fig. 5, the forecasted imbalance price rarely goes beyond 850 €/ MWh (the probability is 1%). It means that the DQN and DDQN methods overestimate Q-values for this area and take OOD actions. Figs. 4 and 5 reveal some correlation between the learned policy by DSAC and the price distribution. The agent always charges the BESS when the price is within the lower 7% quantile (lower than −60 €/ MWh), regardless of the SoC level. The agent never takes the charging action for the 25% highest prices (prices higher than 380 €/ MWh), even if the BESS is empty. The BESS is always discharged when the price lies in the upper 5% quantile (higher than 640 €/ MWh). For the median price (roughly 220 €/ MWh), the BESS is discharged if the SoC is higher than 60%, does nothing when the SoC is between 60% and 50%, and is charged if the SoC is lower than 50%. Generally, the agent learns a milder slope boundary for the discharge action. If the BESS with a low SoC level is discharged, the agent needs to quickly recharge the BESS to make sure it can still make money. This quick recharging increases the risk of charging at a higher price. Therefore, by decreasing the SoC, the area of idle action becomes larger.

5.3. Arbitrage strategy with cycle constraint (Q2)

Fig. 6 shows the learning process of the RL methods for the risk-neutral scenario when the limitation is applied to the daily number of cycles. Similar to the previous scenario, the DSAC method surpasses other methods by converging to a higher reward with a fewer number of cycles. According to Table 3, although the average daily profit of the DSAC method is less than that of the SAC method, the DSAC method earns this profit by consuming fewer number of cycles. In other words,

[3] https://www.elia.be/-/media/project/elia/elia-site/grid-data/balancing/20190827_end-user-documentation-elia1-minute-publications.pdf.

Table 3
Evaluation of RL methods on the test set in the risk-neutral scenarios.

Methods	Without cycle constraint			With cycle constraint		
	Profit (€/per day)	Cycles (per day)	Proportional profit (€/per cycle)	Profit (€/per day)	Cycles (per day)	Proportional profit (€/per cycle)
DQN	749.9	3.2	235.6	338.0	0.9	399.1
DDQN	877.5	3.2	275.9	397.2	1	405.9
SAC	1147.6	3.7	307.6	504.9	1.1	472.7
DSAC	1148.5	3.6	314.1	486.4	0.9	541.7

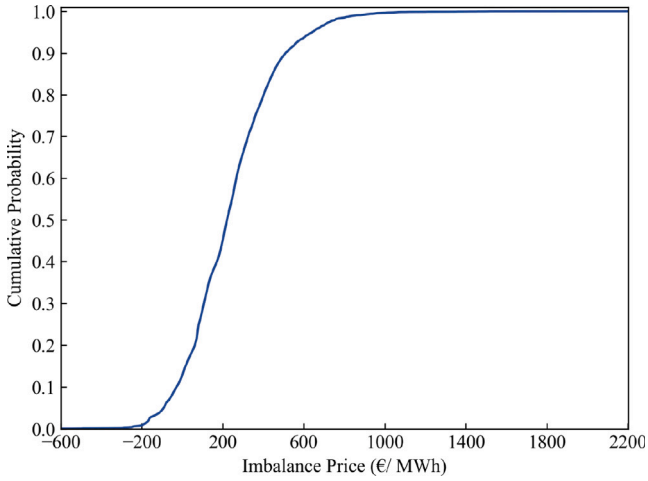


Fig. 5. The cumulative distribution of the imbalance price in 2022.

the DSAC method achieves a 14.6% improvement in the proportional reward per cycle compared to the SAC method. Furthermore, the SAC and DSAC methods converge faster than the DQN and DDQN methods due to their efficient exploration. Since in DQN and DDQN the learned policy is deterministic, the ϵ-greedy exploration technique needs to be used. On the other hand, the SAC and DSAC methods learn a stochastic policy and use the learned probabilities for exploration. Thus, instead of always considering a fixed exploration probability of ϵ for all states, the probability of exploration depends on the current state. For a given state, when the probability of one action is close to 1, the agent almost always exploits and hardly explores. Conversely, when probabilities of all actions are close to each other, the agent most of the time explores to find the best action for that state. Consequently, the SAC and DSAC methods are more data efficient than the DQN and DDQN methods.

The learned policy of DSAC when considering the cycle constraint is illustrated in Fig. 7. Note that the displayed policy is a projection of the learned policy, as the learned policy depends on more than two features and thus is more complicated than the figures shown. The logic behind the learned policy with and without the cycle constraint consideration, which is charging at cheap prices and discharging at expensive prices, is nearly identical. The main difference between these learned policies is in the size of the idle action area. Adding the cycle constraint makes the agent more conservative and increases the idle action area. Moreover, by limiting the number of cycles, the agent recharges the BESS less frequently due to reduced discharging. As a result, in this scenario, the agent recharges the BESS at cheaper prices compared to the previous scenario. To show the performance of the learned DSAC agents in a real-life case, the learned agents are tested using data from March 31, 2022. As Fig. 8 shows, there is one major peak in the imbalance price from 11:00 to 13:15 and one major valley from 13:30 to 17:00 on this day. Both agents properly respond to these prices: the agent without the cycle constraint reacts to roughly all fluctuations in the imbalance price, even small ones (such as the price fluctuation between 4:30 and 6:00, or between 20 and 21:30). However, another agent mostly focuses on more significant fluctuations to limit the number of charging cycles.

Table 4
Evaluation of DSAC method on the test set in the risk-sensitive scenario ($\beta = 3$).

Risk aversion	Profit (€/day)	Cycles (per day)	Proportional profit (€/cycle)	VaR value
$\beta = 0$	1148.5	3.6	314.1	−71
$\beta = 0.3$	796.7	2	399	−48.5
$\beta = 1$	593.9	1.25	474.6	−32.5
$\beta = 3$	518.9	1	518.9	−24.7

5.4. Arbitrage strategy with risk management (Q3)

In risk-averse scenarios, the DQN and SAC methods cannot be directly used as the output of their Q/critic network is a single value, not a full distribution of returns. For this reason, we adopt distributional methods for risk management. Since we already established that distributional DQN exhibits lower performance than distributional SAC in risk-neutral scenarios, we focus on distributional SAC in this section. To answer Q3, we train the DSAC agent without the cycle constraint consideration for varying β values. Results in Table 4 show that the risk-averse agent with $\beta = 3$ experiences a 54.8% reduction in the average daily profit compared to the risk-neutral agent, but given that it avoids risky behavior, we note a higher profit per cycle. Fig. 9 illustrates the difference between the learned critic network for the fully risk-averse and risk-neutral agents. The learned critic network for the fully risk-averse agent is narrower due to applying the risk measure function (VaR) instead of the expectation. Also the VaR values align with this observation: VaR values for the risk-neutral and fully risk-averse critic networks are equal to −589.2€ and −240.5€, respectively. The probability distribution of the hourly profit for test data is shown in Fig. 10. Based on Fig. 10, the risk-averse agent successfully hedges against the uncertainty in the imbalance price and mitigates the tail of the hourly profit distribution.[4] The VaR value of each distribution is provided in Table 4.

Fig. 11 shows the learned risk-averse policy when $\beta = 3$. Compared to Fig. 4, we note that the idle area gets significantly larger: the agent does not discharge the battery when the SoC is low. In this way, the agent makes sure that the battery has always enough energy to inject into the grid when the price is high. Moreover, there is an observable change in the charge threshold that can be justified by Fig. 12. The charge threshold for the risk-neutral agent ranges between 0 and 400 €/ MWh. However, Fig. 12 indicates that within this range, the actual price is significantly uncertain and the chance of charging battery at a price larger than the forecasted value is high. To mitigate this risk, the risk-averse agent learns a lower charge threshold. The risk-averse agent charges the battery at cheaper prices to minimize the risk of charging at a high price resulting from inaccurate price predictions.

[4] Note that both the left- and right-tails are reduced, although from the risk perspective especially the lower (negative) return values should be avoided.

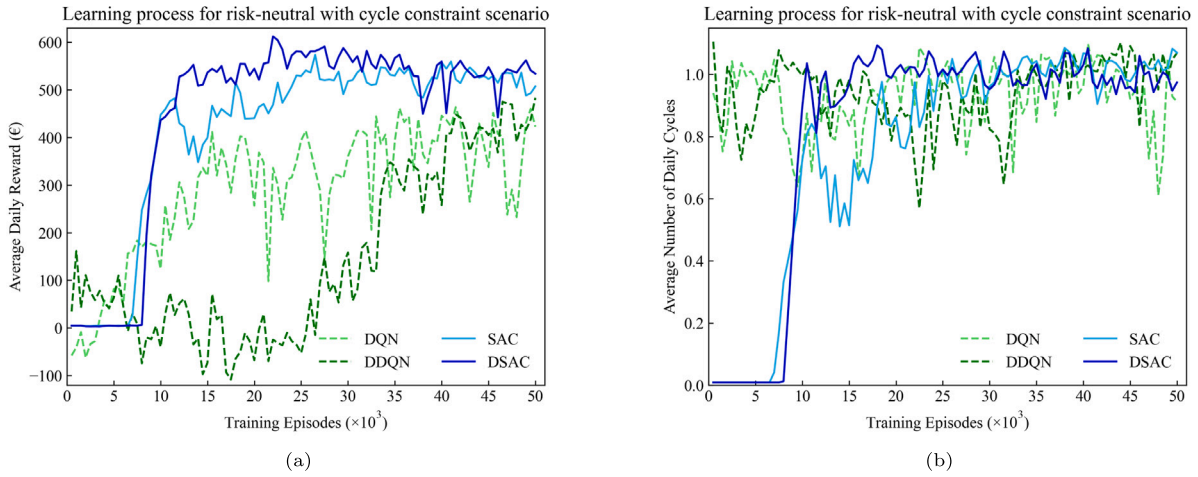


Fig. 6. The learning process of the four RL methods for the risk-neutral with cycle constraint scenario, in terms of (a) the average daily profit, and (b) the average daily number of cycles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

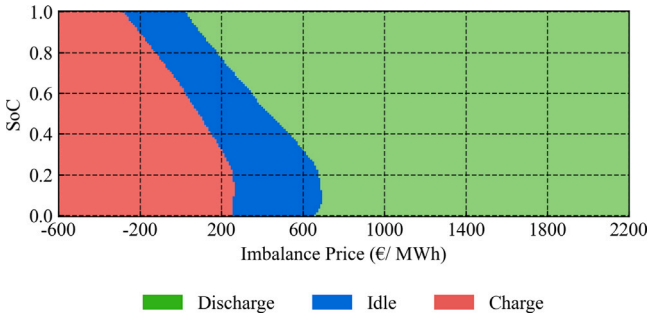


Fig. 7. The projection of the learned policy in the risk-neutral with cycle constraint scenario for DSAC. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

In this paper, a distributional RL-based control framework for BESS was proposed to obtain energy arbitrage strategies in the imbalance settlement mechanism. In the proposed control framework, in addition to considering a constraint on the daily number of cycles, the degree of risk taking in the learned arbitrage strategy can be adjusted based on the risk preference of BRPs. To evaluate the performance of the proposed control framework, two state-of-the-art RL methods, i.e., DQN and SAC, and their distributional variants have been implemented. The results for the Belgian imbalance price in 2022 showed that the DSAC method outperforms other methods (i.e., both the non-distributional baselines as well as DDQN) in all experiments. DSAC improves the average daily profit in the experiment without cycle constraint by 53.1% and in the experiment with cycle constraint by 43.9%, respectively, compared to the (worst performing) DQN method. The dominance of SAC over DQN in terms of data efficiency and mitigating Q-value overestimation, stem from replacing the max operator in the Bellman equation with the expectation operator. Moreover, the distributional methods exhibit better performance than the standard RL methods because they estimate the full probability distribution of returns rather than the expectation of returns, and they resolve instability in the Bellman optimality operator.

In a first experiment, without considering cycle constraints, we noted that the DSAC agent learned a smooth and rational policy: it learned to charge the battery when the price is very cheap (within the lower 7% quantile), discharge when the price is very expensive (within the upper 5% quantile), and take the action based on the SoC for prices in between. In a second experiment, including the cycle constraints, the cycle-aware arbitrage strategy expectedly showed a larger 'idle' action area compared to the case without cycle constraints, effectively leading to a lower number of cycles used. The trained cycle-aware agent tended to respond only to major peaks and valleys in the imbalance price due to the limited number of cycles, while the cycle-unaware agent reacted to almost all fluctuations in the imbalance price. Our study of risk-sensitive agents showed that the risk-averse arbitrage strategies make the distribution of hourly profit narrower and mitigate the tail of the distribution. Indeed, the risk-averse agent charges the battery at lower prices to mitigate the risk associated with inaccurate price forecasts and avoid incurring higher charging costs.

Concerning open research questions and future work, we first note that in this paper, the day-ahead schedule for the battery was set to zero. Follow-up work will generalize the proposed control framework by taking into account energy arbitrage between the day-ahead market and the imbalance settlement mechanism. Studying the effect of considering a continuous action space instead of a discrete one forms another next step to take.

Also, in this paper, we clipped actions post-hoc to satisfy constraints. However, since this post-hoc correction step is not considered during the learning process, it might impact the final performance of the proposed framework. A possible direction to address constraints explicitly is to use differentiable implicit layers that enforce them. By incorporating a differentiable layer into the agent network, the agent network can be trained in an end-to-end fashion to satisfy constraints.

As we mentioned earlier, our current work can be extended by modeling the agent as a price-maker one. For this purpose, the problem needs to be formulated as a bi-level optimization problem, where the upper level optimizes the arbitrage profit along with the risk measure, while the lower level is related to the balancing market clearing problem. Considering a multi-agent system to study the effect of other competing agents on the imbalance price is an additional direction for future work.

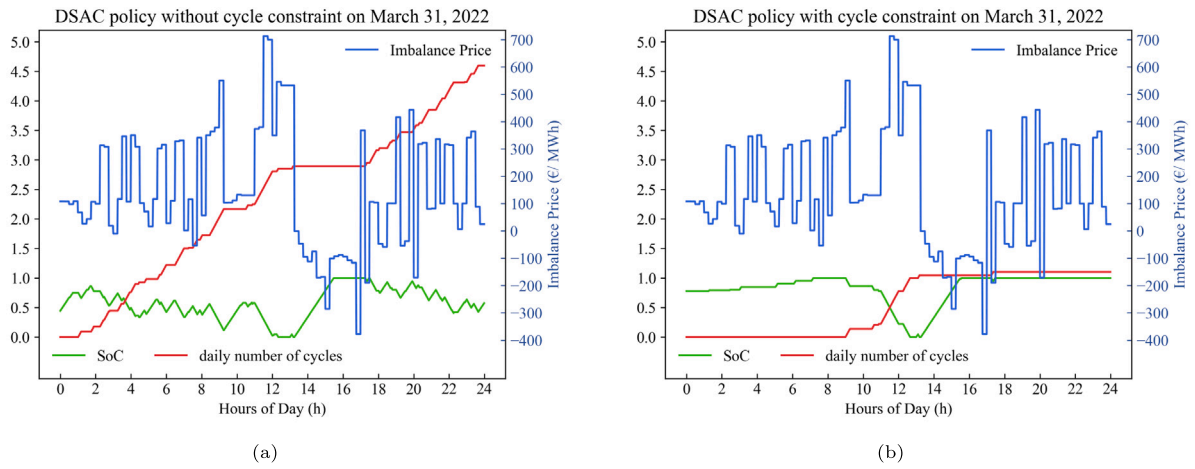


Fig. 8. The performance of the trained agent by the DSAC method on March 31, 2022 (a) without and (b) with considering cycle constraint. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

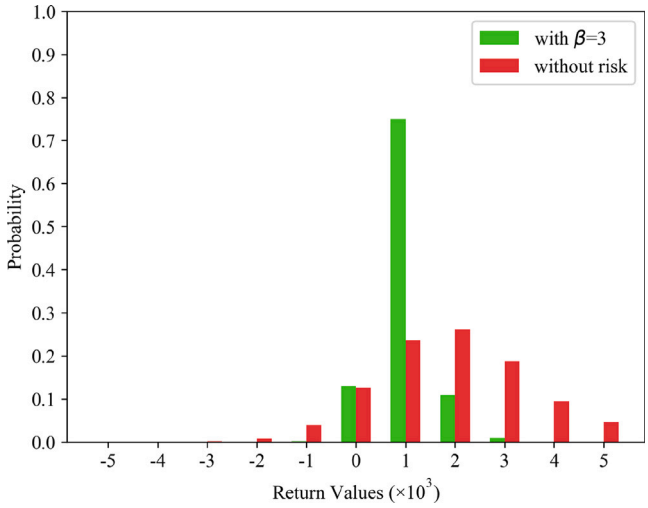


Fig. 9. The learned critic network for the risk-neutral (β = 0) and risk-averse (β = 3) agents. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

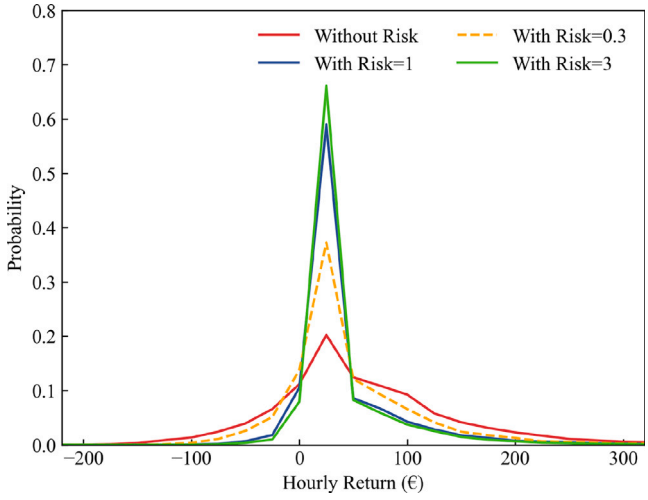


Fig. 10. The probability distribution of hourly profit with and without the risk. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

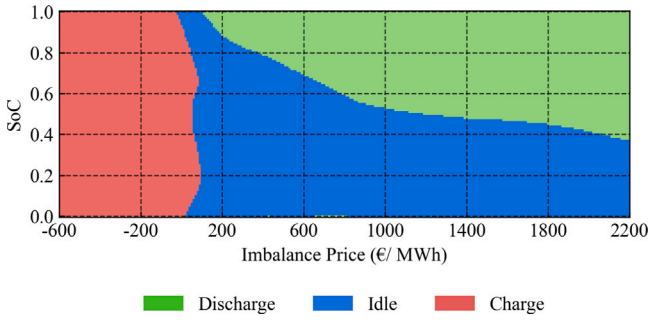


Fig. 11. The projection of the learned policy using DSAC for the risk-averse agent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

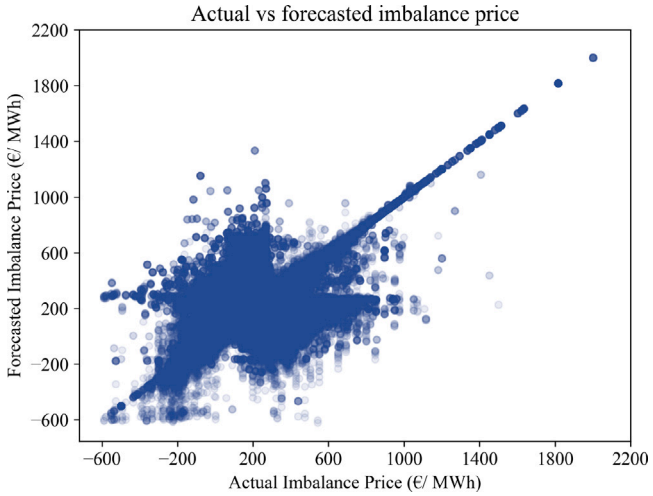


Fig. 12. Actual imbalance price vs. forecasted imbalance price.

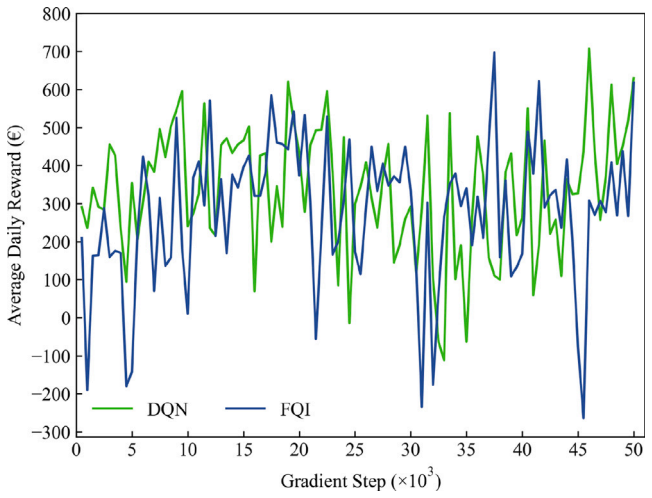


Fig. A.13. The learning process of the DQN and FQI methods for the small experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRediT authorship contribution statement

Seyed Soroush Karimi Madahi: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Bert Claessens:** Writing – review & editing, Supervision, Conceptualization. **Chris Develder:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 3.5 in order to improve the language and check the grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has received funding from the Horizon 2020 Project RENergetic (grant no. 957845) and the Energy Transition Fund (FOD Economy) via the FlexMyHeat project.

Appendix. Comparing DQN with FQI

The FQI method [60] is another widely used value-based method. In [8], FQI is used to obtain a 15-min-based arbitrage strategy in the imbalance settlement mechanism. In this section, a small experiment is carried out to compare the performance of the DQN and FQI methods. In this experiment, the methods are trained on the first nine days of February and evaluated on February 10, 2022. The architecture of the neural network used in the FQI method is the same as that of the DQN method. The experience replay buffer size, number of iterations, and number of episodes are 16 384, 400, and 500, respectively. In accordance with Fig. A.13, both methods perform almost similarly. However, the run time of the FQI method is roughly 5 times greater than that of the DQN method and even gets worse by increasing the experience replay size and the number of episodes. The reason for the longer run time for FQI is its number of iterations: in each episode, the Q network is trained for the mentioned number of iterations. Thus, the FQI method is inappropriate for obtaining the arbitrage strategy.

Data availability

The data that has been used is confidential.

References

[1] Council of the European Union European Parliament, Regulation (EU) 2021/1119 of the European parliament and of the council of 30 june 2021 establishing the framework for achieving climate neutrality and amending regulations (EC) no 401/2009 and (EU) 2018/1999 ('European climate law'), 2021.
[2] J. Bottieau, L. Hubert, Z.D. Grève, F. Vallée, J.F. Toubeau, Very-short-term probabilistic forecasting for a risk-aware participation in the single price imbalance settlement, IEEE Trans. Power Syst. 35 (2020) 1218–1230.
[3] European Network of Transmission System Operators for Electricity, Explanatory document to all tsos' proposal to further specify and harmonise imbalance settlement in accordance with article 52(2) of commission regulation (EU) 2017/2195 of 23 november 2017, establishing a guideline on electricity balancing, 2018.
[4] Y. Yang, Y. Ye, Z. Cheng, G. Ruan, Q. Lu, X. Wang, H. Zhong, Life cycle economic viability analysis of battery storage in electricity market, J. Energy Storage 70 (2023) 107800.
[5] A. Krupp, R. Beckmann, P. Draheim, E. Meschede, E. Ferg, F. Schuldt, C. Agert, Operating strategy optimization considering battery aging for a sector coupling system providing frequency containment reserve, J. Energy Storage 68 (2023) 107787.
[6] K. Fida, K. Imran, K.K. Mehmood, P. Bano, A. Abusorrah, A.K. Janjua, Optimal battery energy storage system deployment from perspectives of private investors and system operators for enhancing power system reliability, J. Energy Storage 69 (2023) 107882.
[7] R. Smets, K. Bruninx, J. Bottieau, J.-F. Toubeau, E. Delarue, Strategic implicit balancing with energy storage systems via stochastic model predictive control, IEEE Trans. Energy Mark. Policy Regul. (2023) 1–14.
[8] J. Lago, G. Suryanarayana, E. Sogancioglu, B.D. Schutter, Optimal control strategies for seasonal thermal energy storage systems with market interaction, IEEE Trans. Control Syst. Technol. 29 (2021) 1891–1906.
[9] M. Dolanyi, K. Bruninx, J.-F. Toubeau, E. Delarue, Capturing electricity market dynamics in strategic market participation using neural network constrained optimization, IEEE Trans. Power Syst. (2023).
[10] B. Ellis, C. White, L. Swan, Degradation of lithium-ion batteries that are simultaneously servicing energy arbitrage and frequency regulation markets, J. Energy Storage 66 (2023) 107409.
[11] H. Khaloie, F. Vallée, Day-ahead dispatch of liquid air energy storage coupled with lng regasification in electricity and lng markets, IEEE Trans. Power Syst. (2023).
[12] S. Vejdan, S. Grijalva, The value of real-time energy arbitrage with energy storage systems, in: IEEE Power and Energy Society General Meeting 2018-August, IEEE Computer Society, 2018.
[13] T. Brijs, F. Geth, C.D. Jonghe, R. Belmans, Quantifying electricity storage arbitrage opportunities in short-term electricity markets in the CWE region, J. Energy Storage 25 (2019) 100899.
[14] H. Khaloie, J. Faraji, F. Vallée, C.S. Lai, J.-F. Toubeau, L.L. Lai, Risk-aware battery bidding with a novel benchmark selection under second-order stochastic dominance, IEEE Trans. Ind. Appl. (2023).
[15] F. Ruelens, B.J. Claessens, R. Belmans, G. Deconinck, Sequential decision-making strategy for a demand response aggregator in a two-settlement electricity market, in: 2016 European Control Conference, ECC, IEEE, 2016, pp. 1229–1235.
[16] M.R. Ansari, M. Yaghtin, M. Kazemi, A bi-level approach for participation of hybrid transmission operating companies in the day-ahead market, considering energy storage systems, J. Energy Storage 61 (2023) 106765.
[17] N. Collath, M. Cornejo, V. Engwerth, H. Hesse, A. Jossen, Increasing the lifetime profitability of battery energy storage systems through aging aware operation, Appl. Energy 348 (2023) 121531.
[18] D. Krishnamurthy, C. Uckun, Z. Zhou, P.R. Thimmapuram, A. Botterud, Energy storage arbitrage under day-ahead and real-time price uncertainty, IEEE Trans. Power Syst. 33 (2017) 84–93.
[19] G. Han, S. Lee, J. Lee, K. Lee, J. Bae, Deep-learning-and reinforcement-learning-based profitable strategy of a grid-level energy storage system for the smart grid, J. Energy Storage 41 (2021) 102868.
[20] G. Xu, J. Shi, J. Wu, C. Lu, C. Wu, D. Wang, Z. Han, An optimal solutions-guided deep reinforcement learning approach for online energy storage control, Appl. Energy 361 (2024) 122915.
[21] I. Boukas, D. Ernst, T. Théate, A. Bolland, A. Huynen, M. Buchwald, C. Wynants, B. Cornélusse, A deep reinforcement learning framework for continuous intraday market bidding, Mach. Learn. 110 (2021) 2335–2387.
[22] Y. Bu, P. Li, H. Yu, H. Ji, G. Song, J. Xu, J. Li, J. Zhao, Risk-managed operation of community integrated energy systems in day-ahead and real-time markets based on portfolio theory, Sustain. Energy Grids Netw. 36 (2023) 101243.

[23] H. Khaloie, A. Anvari-Moghaddam, J. Contreras, J.-F. Toubeau, P. Siano, F. Vallée, Offering and bidding for a wind producer paired with battery and CAES units considering battery degradation, Int. J. Electr. Power Energy Syst. 136 (2022) 107685.
[24] A. Akbari-Dibavar, K. Zare, S. Nojavan, A hybrid stochastic-robust optimization approach for energy storage arbitrage in day-ahead and real-time markets, Sustainable Cities Soc. 49 (2019) 101600.
[25] N. Zheng, X. Liu, B. Xu, Y. Shi, Energy storage price arbitrage via opportunity value function prediction, in: 2023 IEEE Power & Energy Society General Meeting, PESGM, IEEE, 2023, pp. 1–5.
[26] Y. Dong, Z. Dong, T. Zhao, Z. Ding, A strategic day-ahead bidding strategy and operation for battery energy storage system by reinforcement learning, Electr. Power Syst. Res. 196 (2021) 107229.
[27] A. Das, D. Wu, Optimal coordination of distributed energy resources using deep deterministic policy gradient, in: 2022 IEEE Electrical Energy Storage Application and Technologies Conference, EESAT, IEEE, 2022, pp. 1–5.
[28] J. Park, T. Kwon, M.K. Sim, Optimal energy storage system control using a Markovian degradation model—Reinforcement learning approach, J. Energy Storage 71 (2023) 107964.
[29] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, K. Li, Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model, IEEE Trans. Smart Grid 11 (5) (2020) 4513–4521.
[30] T.-H. Wang, Y.-W.P. Hong, Learning-based energy management policy with battery depth-of-discharge considerations, in: 2015 IEEE Global Conference on Signal and Information Processing, GlobalSIP, IEEE, 2015, pp. 992–996.
[31] J. Engels, B. Claessens, G. Deconinck, Combined stochastic optimization of frequency control and self-consumption with a battery, IEEE Trans. Smart Grid 10 (2) (2017) 1971–1981.
[32] X. Zhao, Y. Yang, M. Qin, Q. Xu, Day-ahead dispatch of novel battery charging and swapping station based on distributionally robust optimization, J. Energy Storage 63 (2023) 107080.
[33] L. Buşoniu, T. de Bruin, D. Tolić, J. Kober, I. Palunko, Reinforcement learning for control: Performance, stability, and deep approximators, Annu. Rev. Control 46 (2018) 8–28.
[34] D. Qiu, Y. Wang, W. Hua, G. Strbac, Reinforcement learning for electric vehicle applications in power systems: A critical review, Renew. Sustain. Energy Rev. 173 (2023) 113052.
[35] J. Baetens, J. Laveyne, G. Van Eetvelde, L. Vandevelde, Imbalance pricing methodology in Belgium: Implications for industrial consumers, in: 2020 17th International Conference on the European Energy Market, EEM, IEEE, 2020.
[36] J. Lago, K. Poplavskaya, G. Suryanarayana, B.D. Schutter, A market framework for grid balancing support through imbalances trading, Renew. Sustain. Energy Rev. 137 (2021) 110467.
[37] B. Vatandoust, B.B. Zad, F. Vallée, J.-F. Toubeau, K. Bruninx, Integrated forecasting and scheduling of implicit demand response in balancing markets using inverse optimization, in: 2023 19th International Conference on the European Energy Market, EEM, IEEE, 2023, pp. 1–6.
[38] I.G. Marneris, A.V. Ntomaris, P.N. Biskas, C.G. Baslis, D.I. Chatzigiannis, C.S. Demoulias, K.O. Oureilidis, A.G. Bakirtzis, Optimal participation of RES aggregators in energy and ancillary services markets, IEEE Trans. Ind. Appl. 59 (2022) 232–243.
[39] T. Matsumoto, D. Bunn, Y. Yamada, Mitigation of the inefficiency in imbalance settlement designs using day-ahead prices, IEEE Trans. Power Syst. 37 (2021) 3333–3345.
[40] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.
[41] T. Seyde, I. Gilitschenski, W. Schwarting, B. Stellato, M. Riedmiller, M. Wulfmeier, D. Rus, Is bang-bang control all you need? solving continuous control with bernoulli policies, Adv. Neural Inf. Process. Syst. 34 (2021) 27209–27221.
[42] W. Saunders, G. Sastry, A. Stuhlmueller, O. Evans, Trial without error: Towards safe reinforcement learning via human intervention, 2017, arXiv preprint arXiv:1707.05173.
[43] J. Engels, B. Claessens, G. Deconinck, Techno-economic analysis and optimal control of battery storage for frequency control services, applied to the German market, Appl. Energy 242 (2019) 1036–1049.
[44] C. Zhou, K. Qian, M. Allan, W. Zhou, Modeling of the cost of EV battery wear due to V2G application in power systems, IEEE Trans. Energy Convers. 26 (2011) 1041–1050.
[45] Y. Hu, M. Armada, M.J. Sánchez, Potential utilization of battery energy storage systems (BESS) in the major European electricity markets, Appl. Energy 322 (2022) 119512.
[46] N. Sadeghianpourhamami, J. Deleu, C. Develder, Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning, IEEE Trans. Smart Grid 11 (2019) 203–214.
[47] S. Sultanuddin, R. Vibin, A.R. Kumar, N.R. Behera, M.J. Pasha, K. Baseer, Development of improved reinforcement learning smart charging strategy for electric vehicle fleet, J. Energy Storage 64 (2023) 106987.
[48] Z. Li, Q. Meng, X. Yan, Y. Lei, X. Wu, J. Liu, L. Wang, et al., Reinforcement learning-based demand response strategy for thermal energy storage air-conditioning system considering room temperature and humidity setpoints, J. Energy Storage 72 (2023) 108742.
[49] A.H. Yakout, H.M. Hasanien, R.A. Turky, A.E. Abu-Elanien, Improved reinforcement learning strategy of energy storage units for frequency control of hybrid power systems, J. Energy Storage 72 (2023) 108248.
[50] G. Gokhale, B. Claessens, C. Develder, PhysQ: a physics informed reinforcement learning framework for building control, 2022, arXiv preprint arXiv:2211.11830.
[51] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, F. Blaabjerg, Reinforcement learning and its applications in modern power and energy systems: A review, J. Mod. Power Syst. Clean Energy 8 (2020) 1029–1042.
[52] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: International Conference on Machine Learning, PMLR, 2018, pp. 1861–1870.
[53] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.
[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[55] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, 2015.
[56] M.G. Bellemare, W. Dabney, R. Munos, A distributional perspective on reinforcement learning, in: International Conference on Machine Learning, PMLR, 2017, pp. 449–458.
[57] J. Duan, Y. Guan, S.E. Li, Y. Ren, Q. Sun, B. Cheng, Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors, IEEE Trans. Neural Netw. Learn. Syst. 33 (2021) 6584–6598.
[58] T. Théate, D. Ernst, Risk-sensitive policy with distributional reinforcement learning, Algorithms 16 (2023) 325.
[59] X. Ma, L. Xia, Z. Zhou, J. Yang, Q. Zhao, Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning, 2020, arXiv preprint arXiv:2004.14547.
[60] M. Riedmiller, Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method, in: Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16, Springer, 2005, pp. 317–328.