System-Aware Reinforcement Learning for Optimized Implicit Imbalance Participation in Belgium

Fabio Pavirani
IDLab, Ghent University – imec
Ghent, Belgium
fabio.pavirani@ugent.be

Seyed Soroush Karimi Madahi
IDLab, Ghent University – imec
Ghent, Belgium
seyedsoroush.karimimadahi@ugent.be

Bert Claessens
Beebop
Belgium
bert@beebop.ai

Chris Develder
IDLab, Ghent University – imec
Ghent, Belgium
chris.develder@ugent.be

***Abstract*—The increasing integration of renewable energy sources into electrical grids has disrupted the balance between production and consumption. To address this challenges, transmission system operators such as the Belgian one have introduced imbalance tariffs that penalize harmful energy deviations. Although the imbalance settlement mechanism allows balance responsible parties to dynamically adjust their energy positions, it also exposes them to significant risks. In fact, Belgian imbalance prices are determined retrospectively at the end of each settlement block, meaning that energy deviations occur under uncertain pricing conditions. Reinforcement learning (RL) offers a promising solution for navigating this uncertainty thanks to its ability to manage stochastic environments and deliver long-term rewards. However, achieving profitable participation in imbalance settlement requires more than just handling price volatility; it also demands a deep understanding of the grid dynamics. This paper examines how enriching an RL agent's observation space with grid-related data can enhance its awareness of system dynamics and improve decision-making. We specifically focus on the agent's performance during unstable periods – i.e., quarters where last-minute deviations in system imbalance heavily influence prices – by introducing a related metric. Using the soft actor-critic algorithm, we control a simulated battery energy storage system participating in the Belgian imbalance settlement, leveraging historical data spanning three years. Our findings indicate that, compared to a system-agnostic RL agent (i.e., an agent that does not have grid-related values in the observation space), the system-aware agents develop more effective policies, particularly during unstable quarters.**

***Index Terms*—battery, deep machine learning, grid balancing, imbalance settlement, reinforcement learning**

I. Introduction

A. Problem description

Following the shift of modern electrical grids towards low-emission power generation assets (e.g., solar and wind farms) [1, 2], significant effort has been required by transmission system operators (TSOs) to keep an operational balance between electrical production and consumption. Among the regulatory mechanisms applied in Western Europe, TSOs designed a special imbalance tariff (or imbalance settlement) targeting balance responsible parties (BRPs) that deviate from their energy schedule hindering the grid balance. In Belgium, the imbalance tariff is applied in 15-minute blocks. During each block, any deviation by BRPs is charged with an imbalance price, which is calculated retrospectively based on the system power balance measured throughout the period. The pricing mechanism is designed to penalize deviations that worsen the system imbalance (SI): higher prices are applied when consumption exceeds generation (a negative SI), and lower prices when generation exceeds consumption (a positive SI). This structure creates an implicit balancing framework that not only helps TSOs maintain the SI between operational bounds, but also offers BRPs with flexible assets – such as batteries – the opportunity to make a profit from active participation. While such participation can contribute to grid stability, it is hard to pursue it due to the intrinsic volatile nature of the system SI. Thus, it is challenging and risky for BRPs to implicitly adjust their schedules profitably. BRPs need to consider many factors, such as the amount of flexibility they can allow to allocate, the amount of risk they are willing to face, and the current system status.

B. Related Works

Multiple methods and approaches have been explored to implicitly participate in the imbalance settlement through an energy storage system. Given the high level of uncertainty covering the mechanism, stochastic optimization has been proposed as a technique for passive balancing. Works such as [3–5] focus on such model-based approaches using techniques more broadly referred to as Model Predictive Control (MPC) [6]. Such methods usually require a prediction of future imbalance prices or system imbalances. Works such as [7–10] attempted to predict the future prices. However, given their intrinsic volatility, producing accurate predictions is particularly challenging. Indeed, accurate predictions require a high amount of data that is not available in real-time to BRPs, and will always face inaccuracies due to unpredictable events [11]. Moreover, MPC techniques typically have a high inference time, hindering the real-time participation of the mechanism. An alternative approach is offered by reinforcement learning (RL) algorithms. RL is well-suited to handle stochastic dynamics and, once trained, typically incurs negligible inference time cost. Moreover, model-free

RL algorithms are purely data-driven and do not require any model of the system. As a result, RL can be a preferable alternative to MPC techniques for leveraging real-time prices, such as the imbalance ones [12]. Different contributions have been proposed to enhance the usage of RL in the imbalance settlement. For instance, in [13] a risk-averse policy has been obtained through a Distributional RL algorithm that leverages the imbalance prices using a cycle-constrained battery. In [14] they obtained more interpretable and safe RL policies through a policy distillation procedure. Although these studies have significantly advanced the algorithmic aspects of RL for imbalance settlement, they have overlooked the impact of incorporating grid-related features on the agent's performance and behavior.

C. *Motivations and Contributions*

Since the imbalance prices are strongly correlated to the state of the system, it is crucial for optimized participation to take into consideration grid-related measurements to make more informed decisions. While the agent can still infer profitable patterns even without directly observing grid-related measurements, feeding it these values increases its system-awareness. This enhanced awareness, in turn, improves its revenues and understanding of the problem. Motivated by this gap, we present an ablation study showing the effects of different grid-related features when fed into the agent's observable space.

Specifically, we expanded the observation space of the agent with *real-time SI*, *historical measurements*, and *embedded temporal information*. We assess how these enhancements affect the agent's behavior and revenue, particularly during periods of significant price volatility. Moreover, we introduce a metric to indicate the system-awareness of the agent – i.e., the ability to take actions that match the agent's expectations of the grid dynamics. Our results show how grid-aware agents obtain superior rewards compared to agnostic ones.

II. Methodology

A. *Reinforcement Learning*

RL is a control technique deployed to solve a problem formulated as a Markovian Decision Process (MDP) [15]. An MPD (in its deterministic formulation) is defined as the tuple $(\mathcal{S},\, \mathcal{A},\, f(s,a),\, \rho(s,a))$, where $\mathcal{S}$ is the *State space*, $\mathcal{A}$ is the *Action space*, $f : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{S}$ is the *Transition function*, and $\rho : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ is the *Reward function*. The general objective of the RL agent is to find and select the optimal sequence of actions that maximizes the rewards obtained. To learn that, the agent normally interacts with the environment (i.e., the MDP) multiple times to gather experience samples. The agent will then use these samples to learn a policy $\pi : \mathcal{S} \longrightarrow \mathcal{A}$ that will define the learned behavior of the RL agent. Two major distinctions can be made about RL algorithms, that are:

- **Model-based algorithms:** an internal model of the environment is used by the agent to obtain its policy,
- **Model-free algorithms:** the agent learns a policy without relying on any direct model formulation.

and:

- **Policy-based algorithms:** a parametric policy is directly learned,
- **Value-based algorithms:** the policy is obtained through the learning of value functions that estimate the value of each state-action pair.

In our experiments, we deployed the Soft Actor-Critic (SAC) algorithm [16, 17]. SAC is a model-free algorithm, allowing us to avoid building an explicit model of the imbalance settlement. Moreover, SAC is primarily considered a policy-based algorithm (although it also uses value-based mechanisms to learn the policy), allowing us to directly learn a neural network (NN)-based policy working with a continuous action space. While some studies argue that discrete action spaces can achieve near-optimal performance [18], continuous action spaces enable finer control over uncertainty, making our analysis more insightful. In addition, SAC has already been proven to be a valid algorithm for the imbalance implicit balancing task [13], thus motivating our choice. Next, we describe the MDP formulation used in our experiment.

B. *Problem Formulation*

For simplicity with the notation, we consider the definition of the MDP's state to be equivalent to the vector observable by the agent, meaning our MDPs are fully observable.

1) State Space: Our experiments involve an ablative analysis to evaluate the impact of different state features. We define a basis state space as:

$$\mathcal{S} \doteq \{s_t = (\mathrm{qh}_t,\, \mathrm{qhm}_t, \mathrm{SoC}_t, \lambda_t, c_t) \;\; ; \; \forall t \in \mathbb{Z}^+\} \;, \qquad (1)$$

where t denotes the timestep index of the state with each timestep spanning 1 minute, $\mathrm{qh} \in \{0,\, 1,\, \ldots,\, 95\}$ denotes the quarter-hour, $\mathrm{qhm} \in \{0,\, 1,\, \ldots,\, 14\}$ denotes the progression of the quarter-hour, $\mathrm{SoC} \in [0, 1]$ denotes the (percentage of) state of charge stored in the battery, λ is the last observable price published by the TSO as an estimation of the final price of the quarter-hour, and $c \in \mathbb{R}^+$ is the amount of cycles[1] consumed by the battery in the last 24 hours. On top of this state definition, we assessed the values of these additions:

- **Quarter-hour embedding[2]:** An oscillatory relation between each quarter-hour of a day and the system imbalance (and imbalance price) is capturable from the historical data [19]. This is observable in Figs. 4 to 6, showing the average yearly SI and imbalance prices measured in each quarter of three different years. To help the agent capture this correlation, we directly fed the agent with the historical average imbalance prices measured in each quarter of the train and validation set: $\mathrm{qh}^*_t \in \mathbb{R}$.
- **System imbalance:** To help the agent understand when the price of the current quarter-hour is potentially unstable (i.e., the final price might differ from the estimated

[1] The battery consumes 1 cycle when it discharges an amount of total energy equals to its energy capacity.

[2] The name embedding is used as this addition follows similar intuitions as a deep embedding layer. Note that the encoded values are not learned by the agents, but rather manually fed based on historical values.

one), we directly fed the agent with the average SI observed so far in the settlement period: $\mathrm{SI}_t \in \mathbb{R}$.

- **Historical information:** SI and imbalance price patterns exhibit temporal dependencies. To allow the agent to leverage this, we fed it with the last $\tau \in \mathbb{N}^+$ observations of average SI and final imbalance price in the previous quarters. In our experiments, we fixed $\tau = 2$.

2) Action Space: The action space describes the (dis)charging actions the battery can perform to create a deviation in the BRP schedule, and therefore an implicit participation in the imbalance settlement. Each action corresponds to the power applied to the battery in the corresponding timestep, with negative values corresponding to discharging actions and positive values to charging actions. We assume a power capacity $P^{\max}$ that sets the maximum power of the battery:

$$a_t \in \mathcal{A} \doteq \left[-P^{\max}, P^{\max}\right] \; ; \; \forall t \in \mathbb{Z}^+. \tag{2}$$

In our experiments, we fixed $P^{\max} \doteq 1\,\mathrm{MW}$

3) Reward Function: The reward function is formulated as the revenue obtained through the imbalance settlement. Formally, the reward in a given timestep $t \in \mathbb{Z}^+$ is defined as:

$$r_t \doteq -\hat{\lambda}_t \frac{a_t}{60} \; ; \tag{3}$$

where $\hat{\lambda}_t$ is the actual imbalance price determined at the end of the quarter.

4) Transition Function: Part of the transition function defines the amount of energy exchanged with the grid given a certain battery state and following a certain action. We assume a battery with a certain energy capacity $E^{\max}$. Formally, denoting E_t as the energy stored in the battery in timestep t, then $E_t \doteq \mathrm{SoC}_t E^{\max} \in [0, E^{\max}]$. The dynamics of the battery are then defined as:

$$\mathrm{SoC}_{t+1} \doteq \mathrm{SoC}_t + \frac{\Delta^{\mathrm{E}}}{E^{\max}} \; ; \; \forall t \in \mathbb{Z}^+, \tag{4}$$

where Δ^{E} is the energy released from (injected into) the battery, as of Eq. (5). In our experiments, we fixed $E^{\max} \doteq 2\,\mathrm{MWh}$. Moreover, we assumed a constant efficiency value $\eta \in [0, 1]$.

$$\Delta^{\mathrm{E}}_t \doteq \begin{cases} \frac{a_t}{60}\eta, & \text{if } a_t \geq 0 \\ \frac{a_t}{60}\left(2 - \eta\right), & \text{if } a_t < 0 \end{cases} \tag{5}$$

In our experiments, we fixed $\eta \doteq 0.9$. Finally, we constrained the battery to use at most 1 discharging cycle per day. This constraint makes the evaluation more realistic as the overall battery usage should be limited to preserve its lifetime performance.

III. Experiments setup

A. Datasets

For our experiments, we used historical data comprising the imbalance prices and system imbalance (SI) in Belgium for the years 2021, 2022, and 2023.[3] This data is publicly available on the Belgian TSOs (Elia) website [20]. We divided the data into three sets:

- **Training set:** Contains all days except those allocated to the test and validation sets. The agent explores the environment and collects experience samples using this set.
- **Validation set:** Comprises the second-to-last five days of each month.
- **Test set:** Consists of the last five days of each month and is used to validate the agent's final performance on unseen data.

The final agent parameters are selected based on the configuration that maximizes rewards in the validation set, and the test set is used for out-of-sample validation. Last, we also compared them to the optimal (most profitable) scenario obtained using a multi-integer linear programming technique assuming perfect knowledge of the future prices.

B. Evaluation Metrics

To assess the impact of grid-related inputs in the state space, we compared the results obtained by adding each feature. Each feature addition was evaluated across the three years of historical data, with three different random seeds per year to ensure robustness. For every combination of state space, year, and seed, we performed a hyperparameter tuning cycle. To evaluate the agents, we used three distinct metrics:

- **Daily Reward:** The agent's daily profit on the test set.
- **Cycle consumed:** Assesses the efficiency of battery usage by measuring the number of cycles consumed by the battery. Given two agents with the same daily rewards, the one consuming fewer cycles is preferred as it increases battery longevity.
- **(Price) Mismatched Cost:** Quantifies the agent's ability to make informed decisions based on the system status. Mismatched costs represent the financial difference between the agent's expected revenue (based on the published approximate price, λ_t) and the actual revenue (based on the final imbalance price, $\hat{\lambda}_t$). Formally, the mismatched cost at timestep t is:

$$C_t \doteq \frac{a_t}{60}\left(\hat{\lambda}_t - \lambda_t\right). \tag{6}$$

Lower mismatched costs denote a better understanding of the system conditions, indicating that the agent anticipates price discrepancies and adjusts its actions accordingly. For example, if the agent performs a discharging action ($a_t < 0$) while expecting a higher price ($\lambda_t > \hat{\lambda}_t$), it will receive less revenue than anticipated, resulting in a higher mismatched cost. We acknowledge that, while the mismatched costs metric provides insight into the agent's ability to anticipate deviations in imbalance prices, it does not fully capture the agent's awareness of system dynamics. In particular, a high mismatched cost does not necessarily imply poor decision-making ability. For instance, if the final imbalance price turns out to be lower than the approximated one, the agent may still benefit

[3] Because the three years faced different magnitudes of prices, each year is evaluated independently.

TABLE I: Results averaged for each seed and day of evaluation. The indicated percentages are relative to the baseline agent (System-Agnostic).

Year	**Features**	Daily Rewards	Daily Cycles	Mismatched Costs
2021	MPC Optimal	499 (+66.7%)	0.900 (+8.8%)	-33.0 (-152.1%)
2021	System-Agnostic	299 (0.0%)	0.827 (0.0%)	63.4 (0.0%)
2021	SI	298 (-0.4%)	0.826 (-0.1%)	62.7 (-1.1%)
2021	History	298 (-0.5%)	**0.776 (-6.2%)**	**46.8 (-26.1%)**
2021	Embedding	**308 (+2.7%)**	0.823 (-0.5%)	59.6 (-6.0%)
2021	SI+Embedding	307 (+2.4%)	0.831 (+0.5%)	60.4 (-4.8%)
2021	All Features	294 (-1.9%)	0.787 (-4.8%)	51.6 (-18.5%)
2022	MPC Optimal	871 (+60.6%)	0.916 (+6.3%)	-148.1 (-185%)
2022	System-Agnostic	542 (0.0%)	0.861 (0.0%)	174.1 (0.0%)
2022	SI	565 (+4.3%)	0.862 (+0.1%)	**138.4 (-20.5%)**
2022	History	510 (-6.0%)	**0.807 (-6.3%)**	162.9 (-6.5%)
2022	Embedding	546 (+0.7%)	0.881 (+2.2%)	171.5 (-1.5%)
2022	SI+Embedding	**575 (+6.1%)**	0.901 (+4.6%)	145.6 (-16.4%)
2022	All Features	502 (-7.5%)	0.828 (-3.9%)	133.9 (-23.1%)
2023	MPC Optimal	692 (+75.5%)	0.920 (+6.8%)	-167.4 (-214.7%)
2023	System-Agnostic	394 (0.0%)	0.861 (0.0%)	145.9 (0.0%)
2023	SI	405 (+2.7%)	0.896 (+4.1%)	**102.9 (-29.5%)**
2023	History	387 (-1.8%)	**0.782 (-9.1%)**	118.2 (-19.0%)
2023	Embedding	383 (-2.8%)	0.881 (+2.3%)	137.6 (-5.6%)
2023	SI+Embedding	**410 (+4.0%)**	0.904 (+4.9%)	114.4 (-21.6%)
2023	All Features	350 (-11.2%)	0.833 (-3.3%)	105.8 (-27.5%)

from discharging energy if the final price remains sufficiently high. In such cases, the agent incurs a mismatched cost while still executing an optimal action. Nevertheless, the metric remains a useful proxy for assessing how well the agent accounts for price uncertainty in its decision-making.

IV. Results

We first analyze the average results of system-agnostic and system-aware agents. The results are shown in Table I. It is observable that a general improvement is achieved when adding grid-related features compared to the baseline (system-agnostic) agents. Our major observations can be summarized as follows:

- The addition of grid-related features decreases the mismatched costs in every year of evaluation. This is particularly noticeable with the addition of SI or historical features. We attribute this to the additional information enabling agents to better anticipate deviations in final prices when the system balance is uncertain.
- The addition of the current SI measurement generally increases the overall daily reward.
- Adding historical information appears to make agents more cautious. This is reflected by the amount of daily cycles performed, as they are the lowest in each year. That propagates to the amount of daily revenue as well – as more cautious (i.e., less active) participation involves not fully exploiting the battery capacity. In terms of the ratio of revenue over cycled consumption, adding historical information is one of the best options.
- The single addition of the embedded quarter values does not reflect consistent improvements in the daily reward. However, when matched with the addition of SI measurements, the obtained daily rewards are generally the highest.
- Having all the features together significantly decreases the revenue of the agents. We suspect that this is due to the increased space state (historical information alone adds 4 distinct values), that hinders the agent learning procedure.

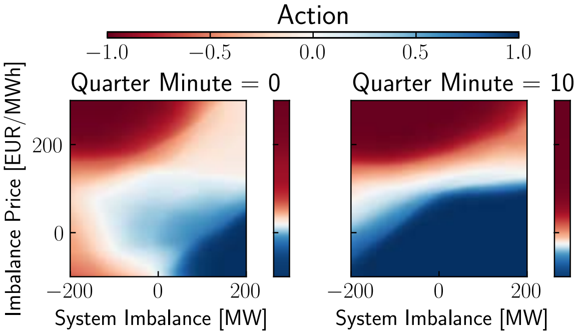

Fig. 1: Actions of a SI-enhanced agent policy when varying the approximated imbalance prices and the SI at the beginning of the quarter (left graph) vs. closer to the end of the quarter (right graph). The policies are compared to the baseline ones, which are independent of the SI values and are indicated as the thinner vertical heat map next to each graph.

We now focus on more specific insights regarding the impact of the added features on the agents' general behavior. We first assess the impact of only adding SI measurement in observation space. In Fig. 1, we visualize an agent policy for 2023 when varying the imbalance prices and the SI values, and we compare it with a grid-agnostic policy (independent of the SI values, as they are not in the baseline observation space). We can observe that for the graph at the left (showing states at the very beginning of the quarter hour), the SI-enabled agent is particularly cautious when the SI magnitude is low, but becomes particularly confident as the magnitude increases, as the sign will hardly change at the end of the quarter hour. This behavior suggests that the agent is now more aware of the system dynamics and their impact on imbalance prices, leading to a better understanding of the fundamental problem it is solving. The uncertainty given by the low-magnitude SI decreases as the quarter hour progresses, as shown in the right-side graph, where the quarter hour is already at an advanced stage and therefore will likely not face any significant changes in the final average SI. This observation gives a valid explanation of why the mismatched costs of the agent when observing the SI are lower than the one of the baseline agent. Indeed, the cases where the final price will significantly differ from the published ones are likely the ones where the sign of the SI is not yet clear i.e., the SI magnitude is low.

In Fig. 2, we visualize the density of actions for the baseline agent and the history-enhanced agent under various scenarios where the published prices differ from the final one. When

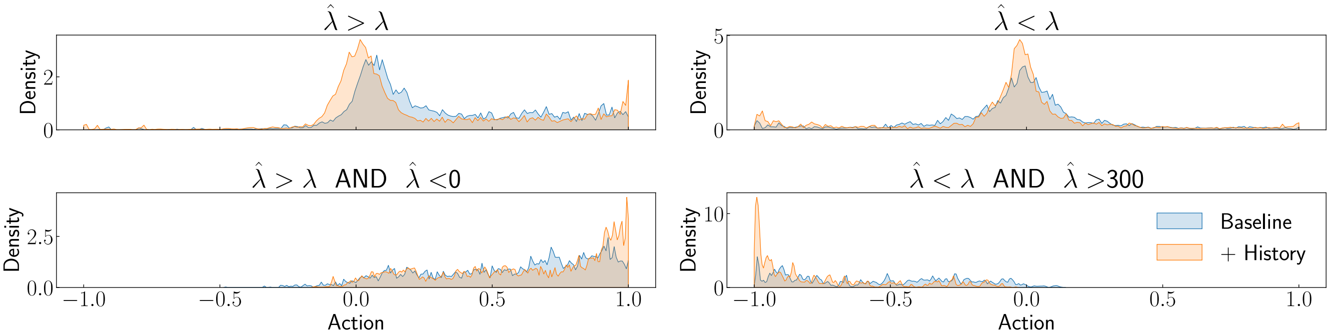

Fig. 2: Density of the agent's action in respective of the mismatched between the final price ($\hat{\lambda}$) and the approximated price (λ). The agent observing historical information better handle errors in the approximated prices compared to the baseline.

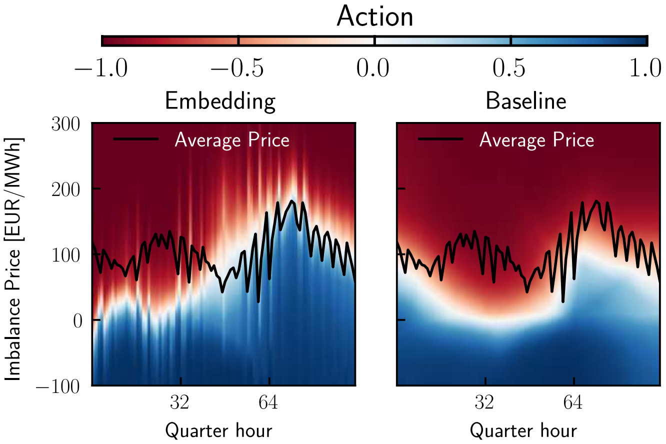

Fig. 3: Actions of an embedding-enhanced agent (left graph) when varying the approximated imbalance prices and the quarter of the day. The policy map is compared with the baseline one (right graph). The line indicates the average imbalance price measured in the evaluated year (2023) for each respective quarter.

the final price exceeds the approximated price ($\hat{\lambda} > \lambda$, as shown in the top left graph), charging actions ($a > 0$) incur a mismatched cost. In this scenario, the history-enhanced agent tends to charge less frequently than the baseline, therefore reducing the mismatched costs. However, when the final prices are sufficiently low – as illustrated in the bottom left graph – it is still optimal to charge the battery. Here, the history-enhanced agent exhibits a higher density of full-power charging actions, indicating that it correctly identifies the revenue-optimal strategy despite an increase in mismatched costs. Analogous observations are noticeable in the graphs on the right, which represent the reverse condition.

Last, Fig. 3 compares the policy actions of the embedded-enriched agent to those of the baseline across different imbalance prices and quarter-hours. The enriched agent more accurately captures the variation between consecutive quarters by following the average price inputs provided as part of its state. Notably, in the latter half of the day, the embedded price signal closely aligns with the agent's threshold for switching between charging and discharging. We also observe a period of higher uncertainty in the embedded agent's actions between quarter 40 and quarter 60. This coincides with increased deviations in the measured prices over the year (see Fig. 6). Although the agent only receives the mean price value for each quarter (without any variance information), it infers the underlying uncertainty from the environmental rewards. Overall, these results confirm that incorporating additional grid-related features into the state space enables the RL agent to better adjust its policy in accordance with the system's expected dynamics

V. Conclusions

In this work, we analyzed the impact of incorporating grid-related features into the observation space of an RL agent performing implicit balancing using a battery storage system. In addition to standard metrics such as overall revenue and battery cycles, we introduced a novel metric – Price Mismatched Costs – which gives an indication of the system-awareness of the agents. Our analysis of historical Belgian data confirms that adding grid-related measurements to the state space yields superior performance across all three metrics. Based on these findings, we propose several promising directions for future research. First, a more detailed investigation into incorporating historical measurements is warranted. Advanced architectures, such as recurrent neural networks, could capture temporal patterns more effectively while keeping the state space compact. Also, our experiments indicate that combining all evaluated features degrade performance – likely due to an excessive increase in state space dimensions. Future studies should explore whether larger or more sophisticated architectures can mitigate this issue. Moreover, an interesting avenue is to integrate a self-learned embedding layer for temporal features directly within the RL network, rather than relying on manually encoded statistical values. Finally, investigating other grid-related features such as balancing signals or diverse temporal indicators (e.g., seasonal variations, holidays) could further enhance the agent's performance.

Acknowledgemnt

This research was partly funded by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme and the Horizon Europe project BlueBird (https://bluebird-project.eu/ - grant agreement no. 101192452).

References

[1] "World energy transitions outlook 2022: 1.5°c pathway," International Renewable Energy Agency (IRENA), 2022.

[2] "Renewable capacity statistics 2023," International Renewable Energy Agency (IRENA), 2023.

[3] R. Smets, K. Bruninx, J. Bottieau, J.-F. Toubeau, and E. Delarue, "Strategic implicit balancing with energy storage systems via stochastic model predictive control," *IEEE Transactions on Energy Markets, Policy and Regulation*, vol. 1, no. 4, pp. 373–385, 2023.

[4] B. Vatandoust, B. B. Zad, F. Vallée, J.-F. Toubeau, and 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.

[5] S. de Weerd, M. Gibescu, M. de Leeuw, J. van Haperen, and B. Siebenga, "Modelling the economic feasibility of distributed flexibility assets in the dutch electricity markets," in *2022 18th International Conference on the European Energy Market (EEM)*. IEEE, 2022, pp. 1–8.

[6] J. Drgoňa, J. Arroyo, I. C. Figueroa, D. Blum, K. Arendt, D. Kim, E. P. Ollé, J. Oravec, M. Wetter, D. L. Vrabie *et al.*, "All you need to know about model predictive control for buildings," *Annual Reviews in Control*, vol. 50, pp. 190–232, 2020.

[7] V. N. Ganesh and D. Bunn, "Forecasting imbalance price densities with statistical methods and neural networks," *IEEE Transactions on Energy Markets, Policy and Regulation*, 2023.

[8] J. Bottieau, L. Hubert, Z. De Grève, F. Vallée, and J.-F. Toubeau, "Very-short-term probabilistic forecasting for a risk-aware participation in the single price imbalance settlement," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1218–1230, 2019.

[9] M. Narajewski, "Probabilistic forecasting of german electricity imbalance prices. energies 15: 4976," 2022.

[10] J. Dumas, I. Boukas, M. M. de Villena, S. Mathieu, and B. Cornélusse, "Probabilistic forecasting of imbalance prices in the belgian context," in *2019 16th International Conference on the European Energy Market (EEM)*. IEEE, 2019, pp. 1–7.

[11] F. Pavirani, J. Van Gompel, S. S. K. Madahi, B. Claessens, and C. Develder, "Predicting and publishing accurate imbalance prices using monte carlo tree search," *arXiv preprint arXiv:2411.04011*, 2024.

[12] J. Lago, G. Suryanarayana, E. Sogancioglu, and B. De Schutter, "Optimal control strategies for seasonal thermal energy storage systems with market interaction," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 5, pp. 1891–1906, 2020.

[13] S. S. K. Madahi, B. Claessens, and C. Develder, "Distributional reinforcement learning-based energy arbitrage strategies in imbalance settlement mechanism," *Journal of Energy Storage*, vol. 104, p. 114377, 2024.

[14] S. S. K. Madahi, G. Gokhale, M.-S. Verwee, B. Claessens, and C. Develder, "Control policy correction framework for reinforcement learning-based energy arbitrage strategies," *arXiv preprint arXiv:2404.18821*, 2024.

[15] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[16] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[17] T. Haarnoja, A. Zhou, P. Abbeel, and 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.

[18] T. Seyde, I. Gilitschenski, W. Schwarting, B. Stellato, M. Riedmiller, M. Wulfmeier, and D. Rus, "Is bang-bang control all you need? solving continuous control with bernoulli policies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 209–27 221, 2021.

[19] J. Van Gompel, B. Claessens, and C. Develder, "Probabilistic forecasting of power system imbalance using neural network-based ensembles," *arXiv preprint arXiv:2404.14836*, 2024.

[20] Elia, "Open data," https://www.elia.be/en/grid-data/open-data.

Appendix

An oscillatory relation occurs between the SI (imbalance prices) and the quarter of the day. The graphs in Figs. 4 to 6 show this by plotting the average price and SI in each quarter of each year considered in our evaluation. An indication of the variability of the values is also shown through 3 quantile intervals (Q40-Q60, Q25-Q75, and Q10-Q90). In our experiments, we directly exploited this relation by feeding the mapped values (average prices for each quarter) to the agent.

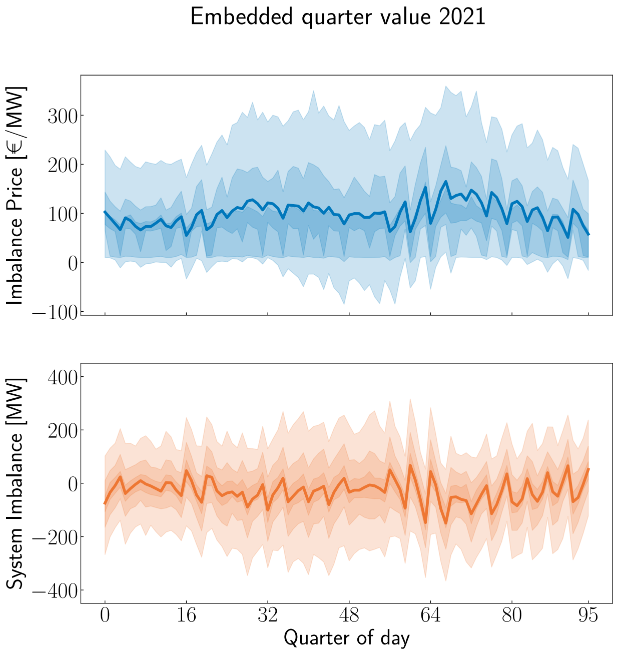

Fig. 4: Average SI and imbalance prices measured in each quarter of 2021. The shaded areas indicate three different quantile intervals.

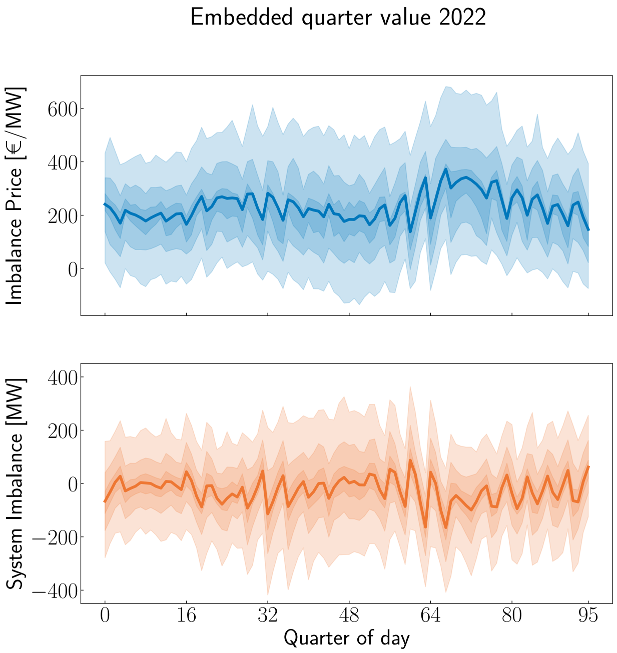

Fig. 5: Average SI and imbalance prices measured in each quarter of 2022. The shaded areas indicate three different quantile intervals.

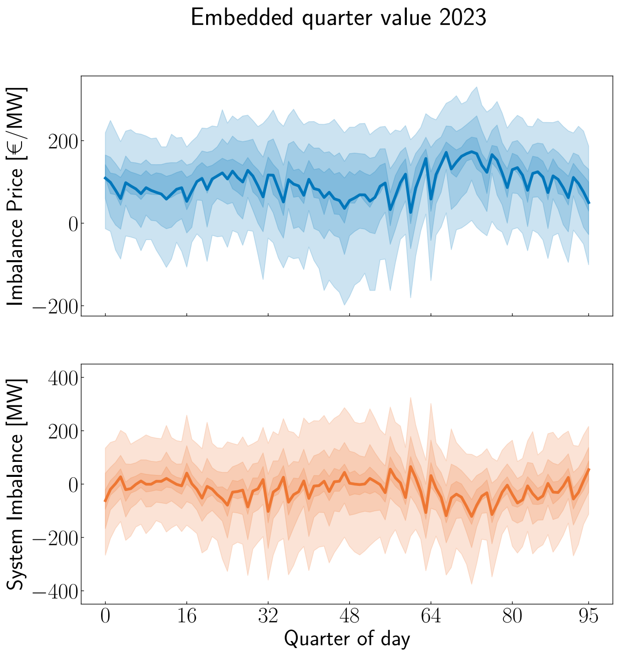

Fig. 6: Average SI and imbalance prices measured in each quarter of 2023. The shaded areas indicate three different quantile intervals.