Multi Power-Market Bidding: Stochastic Programming and Reinforcement Learning

Kim K. Miskiw
Karlsruhe Institute of Technology
kim.miskiw@kit.edu

Nick Harder
University of Freiburg
nick.harder@inatech.uni-freiburg.de

Philipp Staudt
Carl von Ossietzky University of Oldenburg
philipp.staudt@uol.de

Abstract

The growing importance of short-term electricity trading over independent subsequent markets in Europe presents market participants with intricate decision challenges. Established solutions based on stochastic programs are often used but suffer from shortcomings such as the curse of dimensionality in multi-stage decision processes. Reinforcement learning is a promising alternative. However, best practices for the comparison of the two approaches and the ex-post evaluation of reinforcement learning are not yet established. In this paper, we offer a comparison of stochastic programs and reinforcement learning and propose measures for a comparative performance evaluation between the two approaches. We demonstrate them on an empirical case study over subsequent market stages of the German market zone within the coupled European power market.

Keywords: Reinforcement Learning, Multi-market Bidding, Stochastic optimization, Sequential Decision Problems

1. Introduction

The share of intermittent renewable generation in the energy system is continuously growing (eurostat, 2023) and, consequently, the uncertainty and volatility on electricity markets rise. In Europe, in light of self-dispatch, markets react by raising the temporal granularity of electricity trading. For example, the auction interval of the German control reserve markets was decreased from weekly to daily (regelleistung.net, 2022). In reaction to the rising uncertainty, short-term markets are increasingly liquid (e.g., the traded volume on the continuous European intraday market (IDM) increased by 21% in 2021 alone (EpexSpot, 2022)).

In light of these developments, the former dominance of the auction-based pay-as-cleared one-shot day-ahead market (DAM) as the central marketplace for dispatch in the coupled European electricity market is diminishing giving way for increasingly short-term sequential dispatching decisions. Consequently, without a central system operator, market participants face a sequential decision problem under uncertainty regarding these decisions (Klæboe et al., 2022). In this environment, portfolio optimization over all markets can potentially increase the overall portfolio value (Löhndorf & Wozabal, 2022; Miskiw et al., 2023).

Solving a sequential decision problem in the presence of uncertainty can be summarized under the term stochastic optimization (Powell, 2021). However, studies that consider multiple revenue streams for example woth a Stochastic Program (SP) suffer from the curse of dimensionality as multiple scenarios over multiple market stages lead to exponentially increasing number of combined scenarios. This results in high run-times and requires detailed and careful modeling of the underlying uncertainty representation (Powell, 2021).

An alternative approach to these sequential decision problems is Reinforcement Learning (RL) (Powell, 2021), which currently receives considerable attention. However, despite the growth of RL applications, some researchers, as well as practitioners remain skeptical about its performance (Perera & Kamalaruban, 2021). Part of this skepticism might be caused by the absence of established practices for evaluating these black-box models even ex-post in the context of energy markets.

Benchmarking RL and deep reinforcement learning (DRL) in particular, pose significant challenges due to their inherent complexities. Unlike traditional solution methods, which often operate in discrete and well-defined environments, DRL involves training neural networks on complex and continuous state and action spaces. The high dimensionality of these spaces and the and non-linearity of the underlying neural networks make it difficult to define standardized benchmarks or formulate performance guarantees and to judge the quality of decisions of such algorithms, even ex-post. Potentially, SPs offer an alternative avenue to assess the performance of RL algorithms. Powell (2021)

states that "the research community has only begun to exploit the[se] links" between stochastic optimization approaches. With this paper, we aim to contribute to the exploration and exploitation of similarities between SP and RL applied to energy markets and propose evaluation concepts and benchmarks. We answer the following two research questions:

1. What are similarities and differences between current RL and SP applications in terms of modeled decision complexity, uncertainty, and performance comparison?
2. What are options and challenges of benchmarking RL to SP in sequential electricity markets?

To this end, we first introduce the two solution methods and formalize them according an established unifying framework for stochastic optimization (Powell, 2021) in Section 3. We then introduce the specific multi-market bidding problem in Section 4 and answer the described research questions along with an empirical study in Sections 5 and 6. We begin by outlining an overview of the related literature.

2. Related Work

One method for handling sequential and uncertain settings is the application of multi-stage SP, as in (Heredia et al., 2018) and (Löhndorf & Wozabal, 2022). A literature review of solution methods using optimization approaches for a stochastic representation of multiple revenue streams is outlined by Finnah (2022). The uncertainty in sequential multi-market bidding problems is sometimes simplified by assuming perfect foresight or evaluating single revenue streams (Klæboe et al., 2022). This reduces the computational burden and complexity compared to SPs with multiple stages (Aasgård, 2022; Klæboe et al., 2022).

RL is used increasingly for different tasks in the energy domain and especially for energy trading problems (Di et al., 2020; Perera & Kamalaruban, 2021; Yang et al., 2020). RL promises to overcome the mentioned curse of dimensionality by developing an optimal bidding, i.e., operation strategy through a forward-looking procedure, rather than requiring exhaustive evaluations of all combinations of possible system states. More importantly, it does not require any a priori quantification of the uncertainty and it can make strategic decisions without full knowledge of the system (Perera & Kamalaruban, 2021). Most of the previous work only considers one market as in (Lehna et al., 2022) but simplified research settings on multi markets exists (Al-Gabalawy, 2021; Anwar et al., 2022; Demir et al., 2023; Di et al., 2020).

To evaluate the performance of developed RL agents, current literature evaluates RL algorithms in various non-standardized ways. Due to the stochasticity of RL performance, Müller-Brockhausen et al. (2022) suggest that RL is slipping into a replicability crisis, which highlights the need for proper comparison and benchmarking. Some RL applications in the energy domain rely on rule-based approaches to benchmark RL applications (Al-Gabalawy, 2021; Demir et al., 2023; Lehna et al., 2022). Heuristics enable a straightforward comparison and provide a lower bound for the solution of a sequential decision problem, but so far, no standardized and broadly accepted approach has emerged. Others use a deterministic optimization assuming perfect foresight as an upper bound (Berlink et al., 2015; Qiu et al., 2016). Such an optimization assumes complete knowledge of the future and cannot replicate the uncertainty associated with multi-market bidding in practice.

SPs might be a more suitable benchmark because they also consider uncertainty. Some studies aim to replicate an SP with an RL algorithm and evaluate their performance in comparison, which we discuss further in the remainder.

3. Solution Methods: Stochastic optimization

Stochastic optimization in itself is a term comprising different research communities that attempt to solve sequential decision problems under uncertainty (Powell, 2021). In the next section, we introduce solution methods of this domain and focus those that have been used to benchmark RL, namely SPs. To bridge the formal differences between the methods, we use the framework introduced by Powell (2021). According to Powell (2021), a sequential decision problem can be formalized as the trajectory in Equation (1). We start with an initial state (S_0), in which a decision (x_0) is made. After a state an information update follows (W_1) that leads to the next state (S_1) and so on.

$$(S_0; x_0; W_1; S_1; x_1; :::; W_3; S_3; x_3) \tag{1}$$

3.1. Multi-Stage Stochastic Programming

SP can be interpreted as a version of deterministic programming, optimizing a target function that includes random variables. These random variables are discretized in possible realizations $!$ in the finite probability space , often referred to as scenarios. Consequently, SPs can quickly become hard to solve as they are roughly speaking j j times larger than

Figure 1. Translation of the information availability in a three-stage sequential decision problem into a scenario tree.

Figure 2. Formulation of the MDP for a decision in a sequential decision problem following Powell, 2021.

their deterministic counterpart. This can result in the curse of dimensionality, which leads to long run-times (Miskiw et al., 2023). To limit their complexity, a lot of effort is invested into the de nition of the respective scenarios (Powell, 2021). The scenario generation is a multi-step process that requires several modeling choices itself. First, the stochastic process is modeled, and then scenarios are generated and reduced (Heitsch & Keles, 2010).

Multiple decision stages require the quanti cation of scenario trees prior to optimization, as shown in Figure 1. The state S_0 comprises all the initial information, meaning of one entire scenario tree with all the paths and their realization probabilities. The exogenously available information that is revealed at time step t $(W_t^{K_t}; t \in T; k_t \in K_t)$ is equivalent to the realization of the different scenario paths along the tree, which consequently reveals updated information and shapes the subsequent scenario paths and leads to state S_t. T is the set of stages at which new information is revealed, and K_t is the set of discrete scenario realizations, namely the uncertainty quanti cation, de ned for this stage.

Within the scenario trees, the SP determines the optimal solution, considering the uncertainties and dynamics of the sequential problem. This results in a transparent and interpretable result.

3.2. Deep Reinforcement Learning

RL algorithms are based on Markov Decision Processes (MDP) (Bellman, 1957). Formally, an MDP is de ned by a tuple $(S; A; P; R;)$, where S is the state space of the system. Based on the current state s, an action a is chosen. Then, the state of the system changes in accordance with the transition probability $P(s' | s; a)$ from the transition probability distribution P. The goal is to nd a policy that maximizes the total reward $R(s; a; s')$, whereby is the weight with which later rewards are discounted. The policy maps states to actions, i.e., $(s) = a$.

The MDP is similar to the sequential decision problem described by Powell (2021) as shown in Figure 2. The state S of the MDP consists of the de ned exogenous information $W_t; t \in T$, which leads us to the resulting new state S_t de ned according to Powell (2021). The set of actions A is equivalent to the set of the decision variables x_t. The reward in the MDP equals the value of the objective function $R(S_t; x_t; S_{t+1})$ from (Powell, 2021). As the domain around RL grew, distinctive ways of deriving the optimal policy developed to handle both discrete and continuous state and/or action spaces. We focus on approaches that can handle continuous state and action spaces, which limits the eligible algorithms mainly to DRL algorithms (Sutton & Barto, 2018). While DRL promises to solve high-dimensional and mixed-integer non-convex problems, they also have drawbacks. As a data-driven black box model, they are limited in their explainability. In addition, the advantage of not having to model the uncertainty explicitly comes at the cost of having to tune DRL algorithms to optimally interpret provided data (Powell, 2021). This lengthy process can lead to over tting on a particular problem domain and introduces stochasticity to the algorithm's performance. This further highlights the need for a comparison of the DRL results with well-established methods and to develop common benchmarks.

4. The Sequential Decision Problem in Electricity Market Trading

In the following, we introduce the sequential electricity market setting in Europe and formulate the sequential decision problem to be solved by RL and SP using Powell's uni ed framework (Powell, 2021).

4.1. European Electricity Markets

In the European electricity market setting employing self-dispatch, power plant operators face an increasingly complex short-term operational optimization problem. In general, they can trade electricity over-the-counter or on power exchanges. The latter span multiple trading options and products as illustrated in a generalized

Figure 3. Simpli ed timeline of trading options on the European exchange balancing and energy markets (following Miskiw et al. (2023)).

Figure 4. Timeline of information availability (above time axis) and decision making (below time axis) of the multi-stage decision problem following Powell (2021) and Miskiw et al. (2023).

and simpli ed way for the European setting in Figure 3. The markets can be separated into those for the provision of balancing capacity and energy (grey shaded background) and energy (green shaded background).

While many speci cations of the general market setting in Europe are standardized across member states, some small differences remain. The speci c lead times named here relate to the German market zone. The rst market to be cleared in a day is the control reserve market (CRM). Several products are traded, which differ by their lead times and product requirements. Next, the DAM is cleared at 12 a.m. Here, the energy delivery for the 24 hours of the next day is traded with an hourly resolution. The third market is the IDM, which can be divided into an initial auction at 3 pm and the subsequent continuous market with a pay-as-bid offer-based mechanism. Quarter-hourly products can be traded on the IDM. The closest auction to real-time is the energy auction for the activation of the control reserve, which takes place 45 minutes before the actual delivery.

4.2. Multi-Market Bidding in the Uni ed Stochastic Optimization Framework

As the bidding decisions on each auction of the described sequence in uence each other (Klæboe et al., 2022), dispatching energy and capacity evolves to the sequential decision problem, and the uncertainties over all market stages must be considered jointly. In the following paragraphs, we model this setting based on (Kraft et al., 2023) and (Miskiw et al., 2023) using the framework provided by Powell (2021) and explain the dynamics of each decision stage.

The available information before each bidding decision in the sequential electricity market setting[1] can be formalized as shown in Figure 4. A market participant aims to maximize her pro t, which comprises the revenue over all markets ($\pi^{CRM;pos}_{W_1}$, $\pi^{CRM;neg}_{W_1}$, $\pi^{DAM}_{W_2}$, $\pi^{IDM}_{W_3}$) in which she takes part minus the variable costs of the dispatch $c^{u}_{W_3;u}$ of her power plants $u \in U$. This is formulated in Equation (2).

$$\max E_{W_1;W_2;W_3|S_0} \Big(\pi^{CRM;pos}_{W_1} + \pi^{CRM;neg}_{W_1} + \pi^{DAM}_{W_2} + \pi^{IDM}_{W_3} - \sum_{u \in U} c^{u}_{W_3;u} \Big) \mid S_0 \tag{2}$$

Control Reserve Market Decision. The rst auction of the day is for capacity in the CRM. The different control reserve products are auctioned in four hourly intervals, namely $ts \in TS$. We model the bidding decision $x^{CRM;pos;bid}_{S_0;lp;ts}$ as a choice of volume bid on price level $lp \in LP$, where the number of elements in the set LP de nes the granularity of the submitted bidding curve. LP can either be an exogenously de ned discrete price level to avoid non-linear relationships or a decision variable $y^{CRM;pos;bid}_{S_0;lp;ts}$ for the price. The acceptance of a bid is modeled with the binary $\hat{}^{CRM;pos;bid}_{W_1;lp;ts}$, which is one if the price level of the bid $y^{CRM;pos;bid}_{S_0;lp;ts}$ is below the realized price in the market (see Constraint (3)). As shown in Constraint (4), the pay-as-bid remuneration in the CRM is accounted for by including the price level $y^{CRM;pos;bid}_{lp;ts}$ in the revenue. It must be noted that the reserve capacity sold to the transmission system operator is a rm commitment.

$$\hat{x}^{CRM;pos;trade}_{W_1;ts;lp} = \hat{}^{CRM;pos}_{W_1;lp;ts} x^{CRM;pos;bid}_{S_0;lp;ts} \quad \forall ts \in TS \tag{3}$$

[1] We focus on the CRM, DAM, and IDM since the participation in the control energy auction is independent of the CRM results and, hence, is an additional income possibility that does not impose restraints on the former stages (Kraft et al., 2023).

$$\hat{x}^{\mathrm{CRM;pos}}_{W_1} = \sum_{ts=1}^{T^S} \sum_{lp=1}^{P} y^{\mathrm{CRM;pos;bid}}_{S_0;lp;ts} \cdot \hat{x}^{\mathrm{CRM;pos;trade}}_{W_1;ts} \quad (4)$$

Day-Ahead Market Decision. In the next step, new information W_1 becomes available, including the results of the CRM auctions and updated forecasts, which leads to state S_1. In state S_1, the bidding decision for the DAM is made. Contrary to the CRM, market participants can take long and short positions in this market since they can be closed in the IDM and do not constitute a firm commitment towards the transmission system operator. Hence, the bidding decision comprises a bid for selling generation $x^{\mathrm{DAM;gen;bid}}_{S_1;lda;h}$ or taking long $x^{\mathrm{DAM;long;bid}}_{S_1;lda;h}$ or short $x^{\mathrm{DAM;short;bid}}_{S_1;lda;h}$ positions, respectively, that pairs a volume with a chosen price levels $y^{\mathrm{DAM;gen;bid}}_{S_1;lda;h}$, $y^{\mathrm{DAM;long;bid}}_{S_1;lda;h}$ and $y^{\mathrm{DAM;short;bid}}_{S_1;lda;h}$. The time resolution of the DAM is hourly, which is denoted by the indices $h \in H$. After this decision, the next exogenous information update W_2 emerges, comprising the DAM clearing and updated forecasts. Based on the realization of the clearing price $\hat{\lambda}^{\mathrm{DAM}}_{W_2;h}$, the generation, long and short bids are accepted or not, which is modeled with the binary $\hat{\alpha}^{\mathrm{DAM;gen}}_{W_2;lda;h}$. From this, we can calculate the actual traded volume from the positions, which is shown for the generation bid $\hat{x}^{\mathrm{DAM;gen;trade}}_{W_2;h}$ in Constraint (5). The traded volume for the long position is calculated analogously, and the revenue from the pay-as-cleared DAM can, consequently, be calculated as in Constraint (6).

$$\hat{x}^{\mathrm{DAM;gen;trade}}_{W_2;h} = \sum_{lda=1}^{LDA} \hat{\alpha}^{\mathrm{DAM}}_{W_2;lda;h} \cdot x^{\mathrm{DAM;gen;bid}}_{S_1;lda;h} \quad (5)$$

$$\forall h \in H$$

$$\hat{r}^{\mathrm{DAM}}_{W_2} = \sum_{h \in H} \hat{\lambda}^{\mathrm{DAM}}_{W_2;h} \cdot \hat{x}^{\mathrm{DAM;gen;trade}}_{W_2;h} + \hat{x}^{\mathrm{DAM;short;trade}}_{W_2;h} \cdot \hat{x}^{\mathrm{DAM;long;trade}}_{W_2;h} \quad (6)$$

Intraday Market Decision. Using the updated information, the IDM stage is performed. The continuous auction of the IDM theoretically needs to be considered with an infinite number of subsequent stages to replicate the continuous trading opportunities. This is often simplified in literature with just one auction using the ID3 price (the weighted average of bid prices 3 hours before delivery) as a uniform price (EpexSpot, 2022; Kraft et al., 2023). This reduces the IDM decision at this state to the same as for the DAM but with a quarter-hourly $qh \in QH$ resolution.

Realization Decision. After the IDM clearing (W_3), state S_3 is reached, in which the dispatch is realized. This aggregates all former stages, as the market commitments are known and it must be ensured that they are fulfilled while considering the technical constraints. The dispatch of the power plants $u \in U$ of a market participant can be formulated as in Constraint (7) and the CRM results in Constraints (8). To account for the different temporal resolutions we use $qh(h)$, which denotes the mapping of the quarter hours contained in the respective hour h (e.g., for $h = 1$ follows $qh(1) = 1; 2; 3; 4$).

$$\sum_{u \in U} x^{\mathrm{dispatch;U}}_{S_3;qh;u} = \hat{x}^{\mathrm{DAM;trade;gen}}_{W_2;h} + \hat{x}^{\mathrm{IDM;trade;gen}}_{W_3;qh} \quad \forall h; u; qh(h) \quad (7)$$

$$\sum_{u \in U} x^{\mathrm{reserve;pos;U}}_{S_3;qh;u} = \hat{x}^{\mathrm{CRM;pos;trade}}_{W_1;ts} \quad \forall ts; u; qh(ts) \quad (8)$$

It must be ensured that the speculative positions (long and short) are closed before the actual delivery since high imbalance costs need to be paid otherwise (see Constraint (9)).

$$\hat{x}^{\mathrm{DAM;short;trade}}_{W_2;h} + \hat{x}^{\mathrm{IDM;short;trade}}_{W_3;qh} = \hat{x}^{\mathrm{DAM;long;trade}}_{W_2;h} + \hat{x}^{\mathrm{IDM;long;trade}}_{W_3;qh} \quad \forall h; qh(h) \quad (9)$$

This dispatch must be in line with the capacity constraints of the power plant portfolio and enable the provision of the accepted reserve (Constraints (10) to (14)). In the interest of brevity, we abstain from including other technical constraints, such as ramping. A more complete formulation can be found in (Kraft et al., 2023).

$$x^{\mathrm{reserve;pos;U}}_{S_3;qh;u} + x^{\mathrm{dispatch;U}}_{S_3;qh;u} \le P^{\max}_u \quad \forall qh; u \quad (10)$$

$$x^{\mathrm{dispatch;U}}_{S_3;qh;u} - x^{\mathrm{reserve;neg;U}}_{S_3;qh;u} \ge P^{\min}_u \quad \forall qh; u. \quad (11)$$

$$x^{\text{reserve;neg;U}}_{S_3;qh;u} \quad \text{BIGM} \quad x^{\text{reserve;neg;U}}_{S_3;qh;u} \quad \forall qh;u \tag{12}$$

$$x^{\text{dispatch;U}}_{S_3;qh;u} \quad P^{\min}_{u} \quad x^{\text{reserve;pos;U}}_{S_3;qh;u} \quad \forall qh;u \tag{13}$$

$$x^{\text{reserve;pos;U}}_{S_3;qh;u} \quad \text{BIGM} \quad x^{\text{reserve;pos;U}}_{S_3;qh;u} \quad \forall qh;u \tag{14}$$

5. Qualitative Comparison: Reinforcement Learning & Stochastic Programs

Based on this introduction, we now compare RL and SP applied to multi-market bidding problems currently present in the literature along several axes, namely the modeled decision complexity, uncertainty, and the performance comparison.

Decision Complexity. Modelers try to avoid non-linearities as SPs potentially suffer from high computational times. Market prices, for instance, are often discretized with different prede ned price levels (Kraft et al., 2023; löhndorf & Wozabal, 2022). Models may still include a broad variety of continuous decision variables. For example, Kraft et al. (2023) model three markets, where on each market, a step-wise bidding curve with a minimum of ve different price levels can be submitted for generation, as well as short and long positions. In the interest of comparability, the RL action space should at least be able to represent the same decision complexity as corresponding SPs. However, existing RL studies often employ simpli ed decision representations, such as tabular approaches, which do not enable continuous decision variables (Cao et al., 2020a; Perera & Kamalaruban, 2021). Some studies utilize function approximation methods (Ye et al., 2020), but the action spaces of the bidding strategies are simpli ed. For example, certain models focus on selecting risk factors that indirectly determine the quantity and price of bids (Al-Gabalawy, 2021). Others concentrate solely on choosing the dispatched volume and disregard the uncertainty associated with the bidding price, which de nes whether a bid is successful (Anwar et al., 2022; Cao et al., 2020b; Ochoa et al., 2022). Consequently, scanning the literature reveals that RL approaches often do not yet replicate the complex market environments modeled in state-of-the-art SPs.

Uncertainty Quanti cation. For SPs, the uncertainty has to be quanti ed explicitly and is provided with exogenous input parameters a priori. The performance and results depend on this underlying representation of uncertainty (Ochoa et al., 2022). Therefore, much effort is dedicated to the right scenario de nition (Russo et al., 2022). In contrast, RL methods are data-driven and do not need to quantify the uncertainty explicitly. The underlying uncertainty is accounted for by modeling the partial observability of the state. This is generally a challenge in the RL domain (Stapelberg & Malan, 2020). Consequently, many multi-market bidding applications assume full observability, where information such as uncertain market prices is included with perfect foresight in the state space (Anwar et al., 2022; Ochoa et al., 2022). The uncertainty in renewable generation is more often accounted for (Lehna et al., 2022). If uncertainty is considered, the state space often incorporates exogenous demand, renewable generation or price forecasts, akin to the uncertainty quanti cation methods used in SPs (Cao et al., 2020b; Demir et al., 2023; Lehna et al., 2022). In summary, the RL approaches have the theoretical advantage of not needing an uncertainty quanti cation. However, current approaches mostly neglect uncertainty in general or only consider a small subset of uncertainties in comparison to SP models.

Performance. Due to the discussed inherent differences in simulation setups, comparing the results of these approaches is challenging. Yet, some studies in the literature attempt to do so. Mohammadi and Hesamzadeh (2022) apply an SP and tabular Q-learning to the same problem. Both methods rely on discrete state spaces. In (Mohammadi & Hesamzadeh, 2022), tabular Q-Learning outperforms the respective SP approach for prosumer management in the IDM. This most likely originates from the uncertainty quanti cation of the SP, but only little information is provided on the multi-step scenario generation process applied. Ochoa et al. (2022) demonstrate that in their study, the SP approach is superior with regard to the daily market income, but the DRL agent achieves less penalized imbalances. Yet, the DRL agent is trained and evaluated on historical data, while the SP only produces results based on the discrete scenario trees. In both comparisons, the RL agent can be expected to outperform the SP. This is because the uncertainty quanti cation in the scenario trees condenses information and underrepresents some of the variance present in the data. The RL agent can leverage this remaining variance to increase its pro t. This plays into the strengths of the RL agent, as it captures patterns that are not re ected in the uncertainty quanti cation of the SP. Both comparisons discuss only to a limited extent how the results from the scenario trees of the SP are linked to historical realizations. (Ochoa et al., 2022) assess their scenario trees with the energy score and Mohammadi and Hesamzadeh, 2022 analyse the effect of the number of scenarios on the generated pro t.

6. Quantitative Comparison: Reinforcement Learning & Stochastic Programs

We evaluate both approaches using an empirical study to tackle the named qualitative differences and to propose comparison measures and benchmarks. The performance evaluation is demonstrated using two different comparisons between SP and RL: Intrinsic and empirical. The process for these comparisons and results in our study are shown in Figure 5.

The intrinsic evaluation is based on the results of the methods relative to their theoretical optimum under perfect foresight. The RL approach is evaluated against the perfect foresight optimization of the decisions on the empirical data. On the contrary, the SP result is evaluated against the Wait-and-see solution within the scenario trees, which allows deciding after one of the uncertain scenario is realized (Conejo et al., 2010). The latter can be averaged over the simulation period by using the probability of each scenario tree, which also results in the average expected daily pro t.

For the empirical comparison, we need to bridge the gap between the results on the SP's scenario trees (expected pro t per scenario tree) and the results of the RL approach on empirical data (over the empirical period). One way of doing so is mapping the empirical days to the scenario trees. Then, the decision made by the SP in the scenario tree can be evaluated based on the actual empirical realization. In the SP context, such a comparison is also called out-of-sample (Conejo et al., 2010), which differs from the meaning of out-of-sample in the context of machine learning.

6.1. Case Study

To address the differences in decision complexity, we use the same study setup for both methods. We model the positive CRM and DAM for a market participant with a 100 MW P_u^{max} gas- red power plant. Hence, we consider the trajectory in Equation (15).

$$(S_0; [x^{CRM;pos;bid}_{S_0;lp;ts}; y^{CRM;pos;bid}_{S_0;lp;ts}]; W_1; S_1; [x^{DAM;gen;bid}_{S_1;h}; y^{DAM;gen;bid}_{S_1;h}]; W_2; S_2) \quad (15)$$

Since we do not model the IDM stage, long or short bids are not considered because they could not be closed. The technical unit u has constant variable costs of 40 €/MWh. For this asset class, ramping and minimum run constraints are neglectable. The operation from July 2019 to March 2020 in the German market zone is simulated. For the initial information S_0 and the update W_1, we use empirical CRM (regelleistung.net, 2022) and DAM prices as well as publicly available RES forecasts, its updates and the forecasted residual load pro les (Entso-e, 2022).

[2] Please note that for this case $x^{DAM;gen;trade}_{S_1;h}$ and $y^{DAM;gen;trade}_{S_1;h}$ can be set implicitly since it is always optimal to offer the entire not traded capacity into the DAM at marginal costs.

6.2. Stochastic Program

In line with the mathematical formulation of Section 4 and the solution approach outlined in Section 3, a SP for the case study setup is formulated. As pointed out in the qualitative comparison, the quanti cation of uncertainty is crucial for the SP's performance. To do this properly, we use a previously published, stand-alone SP approach. Consequently, the information updates are discretized with constructed scenarios $W_1^{K_1}$ and $W_2^{K_2}$, which are adapted from (Russo et al., 2022) and (Kraft et al., 2023). In their approach, the prices are simulated with one additive time series model and one model for the stochastic residuals with a mean-reversion process and jump regimes. These simulations are clustered into different groups for a set of 18 type days. The derived clusters distinguish between three residual load levels, the season, and whether it is a weekend. For the scenario tree constructions, the simulated prices per cluster are reduced in order to de ne $|K_1| = 10$ scenarios for the CRM prices and $|K_2| = 5$ for DAM prices.

6.3. Deep Reinforcement Learning Model

In the implementation of the learning agent, we employ the Proximal Policy optimization (PPO) algorithm (Schulman et al., 2017). Built upon the foundations of actor-critic methods, PPO addresses the challenge of optimizing policies for sequential decision-making tasks. The implementation is based on the Stable Baselines3 framework by Raf n et al. (2021). The learning agent is trained on the same data set used to construct the uncertainty quanti cation for the SP.

At the rst state S_0, the agent receives the average CRM and DAM market clearing prices from the same time on the preceding day $y^{CRM;pos}_{S_0;ts-6}; y^{DAM}_{S_1;h-24}$ as well as the forecasted solar generation and residual load for the subsequent four hours $P^{solar}_{ts}; P^{load}_{ts}$ and information on the present hour of the day and the season sin and cos values I^{date}_{ts}. The actions spaces of the RL agent in S_0 are $x^{CRM;pos;bid}_{S_0;lp;ts} \in [0; P_u^{max}]$ and $y^{CRM;pos;bid}_{S_0;lp;ts} \in [0; 70]$.

As we aim to maximise the generated revenue, the agent's reward is de ned in Equation (2). To further

Figure 5. Concept graph for obtaining the intrinsic and empirical comparison between SP and RL study results.

guide the agent, we subtract an additional regret term cm_{ts} weighted with . The regret term equals the theoretical optimum the agent could have achieved based on the market prices that were realised.

6.4. Intrinsic Evaluation

The described intrinsic comparison leads to the results for the average daily pro t in Figure 5. It lies in the nature of this comparison, that we evaluate the methods on different grounds. It is rather an evaluation of how well the algorithms perform in regard to their own optimum. The bar of the stochastic program (lower, orange bar) shows the average pro t in comparison to the wait-and-see solution where the realization in the scenario trees can be anticipated. We can see that the SP comes close to its perfect foresight optimum. The RL approach (upper, blue bar) reaches a similar percentage of the perfect foresight optimum on the historical data. Yet, the average daily pro t of the SP is in general higher than with the RL approach, which originates from the scenarios. One reason for this might be that, on average, the pro t potential is greater in the quanti ed scenario trees. Alternatively, it might be caused by the probabilities used to calculate the average pro t for the scenario trees, especially for days when revenue is relatively high. This evaluation does not necessarily allow to compare the two approaches vis-a-vis. However, it is a useful indicator of whether the models are well-parametrized and can serve as benchmarks for each other using different indicators. One such indicator is presented in the following.

6.5. Empirical Evaluation

For this comparison, the SP is applied to the empirical data of the simulation horizon. We do so by mapping an empirical day to the respective scenario tree. The trees are distinguished based on season, weekend or weekdays and residual load level. The former two can be matched easily, while an empirical day is assigned to the residual load level by nding the level where the average absolute deviation between the empirical day and scenario tree residual load is the lowest[3]. Then, the optimal decision according to the SP is executed on the empirical day. The cumulative pro t of this comparison for a weekday in the transition season is shown in Figure 6. As the scenario tree realizations deviate from the actual realizations, the cumulative pro t is lower. The scenario trees condense information and underrepresent some of the variance present in the data. For example, the scenario trees do not contain outliers. Analysing the decisions in detail, it becomes clear that the SP approach more often bids relatively low prices in the CRM market, since it underestimates the historical DAM pro t opportunities. The RL agent, on the other hand, mostly bids on the CRM market during periods with high prices and otherwise participates in the DAM. Yet, the results and comparisons show why an SP, as it is considered here, suffers from disadvantages when compared to RL based on empirical data.

[3] This is not necessarily optimal and, we could assign a day to multiple type days and weigh the results. However, we propose this rst approach here.

Figure 6. Empirical comparison of cumulative profit made by the solution approach applied to empirical realizations of a weekday in the transition season.

7. Conclusion

Calculating optimal strategies to bid sequentially on multiple electricity markets poses a challenging task for algorithmic solutions due to the dimensionality of the state and action spaces and inherent uncertainty. This makes the bidding process a sequential decision problem that can be addressed using stochastic optimization. As these classical approaches face the curse of dimensionality and are challenging to implement in online settings, reinforcement learning has gained popularity as an alternative solution.

This paper aims to add to the discussion on the comparison between stochastic programs and reinforcement learning. We attempt to bridge the gap between the methods by formalising the multi-market bidding problem as a sequential decision under uncertainty for both methods based on (Powell, 2021). We explore the similarities and differences in current applications of reinforcement learning and stochastic programs and consequently answer the stated research questions. Our qualitative analysis reveals that while reinforcement learning is theoretically able to reflect the continuous action and state spaces as well as partial observability induced by uncertainty, the applications do not yet reach the complexity of state-of-the-art stochastic programs in this domain. While the literature suggests that reinforcement learning outperforms the respective stochastic programs when applied to similar cases, some challenges remain. We discuss two options to benchmark reinforcement learning with stochastic optimizations in an empirical study. The study uses data from the control reserve and day-ahead market of the German market zone within the European power market and conducts an intrinsic and empirical comparison. The intrinsic comparison shows that both methods perform well relative to their theoretical optimums. This is a prerequisite for meaningful benchmarking. We then compare both methods on empirical data. This reveals the expected outperformance of the stochastic program by the reinforcement learning approach as the latter can exploit more variance in the data. Yet, more work needs to be done to correctly map the results from the scenario trees to empirical applications, as we propose and evaluate one naive approach. Our results contribute to the discussion of using and benchmarking DRL in electricity market modeling. However, both the evaluation of deep reinforcement learning as well as its benchmarking with stochastic programs are still nascent fields with a need for further research and consensus building within the community.

References

Aasgard, E. K. (2022). The value of coordinated hydropower bidding in the nordic day-ahead and balancing market. Energy Systems, 13(1), 53–77. https://doi.org/10.1007/s12667-020-00388-7

Al-Gabalawy, M. (2021). Reinforcement learning for the optimization of electric vehicle virtual power plants. International Transactions on Electrical Energy Systems, 31(8), e12951. https://doi.org/10.1002/2050-7038.12951

Anwar, M., Wang, C., de Nijs, F., & Wang, H. (2022). Proximal policy optimization based reinforcement learning for joint bidding in energy and frequency regulation markets. 2022 IEEE Power & Energy Society General Meeting (PESGM), 1–5. https://doi.org/10.1109/PESGM48719.2022.9917082

Bellman, R. (1957). A markovian decision process. Journal of Mathematics and Mechanics, 6(5), 679–684. Retrieved June 12, 2023, from http://www.jstor.org/stable/24900506

Berlink, H., Kagan, N., & Reali Costa, A. H. (2015). Intelligent decision-making for smart home energy management. Journal of Intelligent & Robotic Systems, 80(S1), 331–354. https://doi.org/10.1007/s10846-014-0169-8

Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., & Blaabjerg, F. (2020a). Reinforcement learning and its applications in modern power and energy systems: A review. Journal of Modern Power Systems and Clean Energy, 8(6), 1029–1042. https://doi.org/10.35833/MPCE.2020.000552

Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., & Blaabjerg, F. (2020b). Reinforcement learning and its applications in modern power and energy systems: A review. Journal of Modern Power Systems and Clean Energy, 8. https://doi.org/10.35833/MPCE.2020.000552