Reinforcement-Learning Based Threshold Policies for Continuous Intraday Electricity Market Trading

Gilles Bertrand
CORE, Université catholique de Louvain
Email: gilles.bertrand@uclouvain.be

Anthony Papavasiliou
CORE, Université catholique de Louvain
Email: anthony.papavasiliou@uclouvain.be

***Abstract*—Continuous intraday electricity market has become increasingly important in recent years, due to the increasing integration of renewable resources in power systems. Trading in this market is challenging due to the multistage nature of the problem, its high uncertainty, and the fact that decisions need to be made rapidly in order to lock in profitable trades.**

We cast the problem of trading in continuous intraday markets as a reinforcement learning problem, and tackle the problem using policy function approximation. We specifically parametrize the trading policy using price thresholds, and optimize the choice of these thresholds using the REINFORCE algorithm. We demonstrate the effectiveness of our proposed policy by showing that it outperforms the method, classically used in the industry, rolling intrinsic of 4.2% (out of sample) on the 165 last days of 2015 in the German continuous intraday market.

I. Introduction

The integration of renewable sources in Germany has recently increased from 18.2% in 2010 to 32.2% in 2016 [1]. This proliferation is largely driven by the 2020 climate and energy package [2] which has been adopted by the European Union in 2008, and which targets sourcing 20% of the EU energy consumption from renewable sources by 2020. This increase of renewable production implies that the market requires more flexibility close to real time. Consequently, the Germany continuous intraday market (CIM) which allows agents to correct their trading positions in the actual day of operations where renewable supply conditions are revealed has become increasingly active. Specifically, traded volumes in the German CIM have increased from 1005 TWh in 2010 to 4070 in 2016 [3]. This market is therefore becoming an interesting option for fast-moving assets such as pumped hydro storage to valorize their flexibility.

CIM follows a format which is distinct from that of balancing markets and day-ahead auctions, and have therefore been analyzed separately in the literature. The literature can be classified into the three following categories. (i) The first category of papers focuses on modeling the statistical properties of the CIM [6] and [7]. (ii) The second category optimizes trading strategies by placing strong assumptions on the behavior of the CIM price through parametric models [8] and [9]. (iii) The third approach focuses on developing methods without invoking prior assumptions on either the model or the data which is also the approach that we follow. In [10] the authors propose a heuristic method for covering the position of a wind farm. In [11], the problem of a storage unit is modelled as a partially observable Markov Decision Process and solved ”in sample” using value function approximation. In [12], the problem of a trader not owning any asset and covering its position in the balancing market is presented. The authors model the problem as a one step reinforcement learning problem and use policy function approximation coupled to the REINFORCE algorithm in order to solve it.

The contribution of this paper is threefold: (i) We cast the problem of bidding in the CIM as a reinforcement learning problem. (ii) We employ policy function approximation in order to solve the reinforcement learning problem. The randomized policy that we propose is characterized by parameters which describe the price thresholds above which buy orders[1] should be accepted, and below which sell orders should be accepted. The form of our policy will be justified by the insight that we gain from analyzing the KKT conditions of the problem in deterministic form. In order to optimize the policy parameters, we use the REINFORCE algorithm, as described in [13] and [14]. (iii) We propose some behaviors that should be included in the policy in order to make it successful. We will present the general idea of these behaviors and explain with complete details how to parametrize the policy for one of them.

II. Overview of the German electricity market

The positioning of the CIM in the German electricity market design is presented in figure 1. Short-term market operations commence with the clearing of the day-ahead market on the day before actual operations (D-1), at 12 noon [15]. The intraday auction (IA) is then conducted at 3 pm [16] on D-1. The CIM opens at 3 pm on D-1 for hourly products, and at 4 pm on D-1 for quarterly products. The market closes 30 minutes before delivery [3]. Following delivery, the imbalances of market participants are settled at the imbalance settlement phase.

In this paper we will focus exclusively on developing methods for the CIM. We further restrict our attention to methods that do not result in imbalance, in order to focus on the intertemporal arbitrage opportunities that are offered by pumped hydro assets, and avoid instead risky policies that make profit by speculating on the imbalance price. In the German CIM, buy and sell bids arrive randomly and are ‘grabbed’ by market participants who find the bids favourable. The specific question that we focus on in this paper is which bids should be selected by owners of pumped hydro storage assets. The bids are characterized by the delivery period (hour

[1] Buy/sell order means that somebody wants to buy/sell power from us.

or quarter within an hour), their type (buy or sell), the selling or buying price (in €/MWh) and quantity (in MWh).

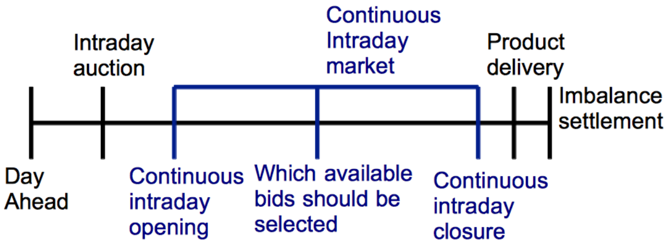

Fig. 1: The sequence of operations in German short-term electricity markets

III. The CIM trading problem as a reinforcement learning problem

We proceed by casting the CIM trading problem as a reinforcement learning problem. To this aim, we define the states variables, the action variables, the transition function and the objective function. We make the following assumptions and simplifications: (i) As the information about the type of the bids (continuous, integer, block) is not disclosed in the German market data set, we consider only continuous bids. This implies that we can accept fractions of bids. (ii) For the sake of simplicity, we only consider hourly products. The extension to quarterly products is straightforward.

A. State variables

Our state can be decomposed into three subsets:

$$S_t = (S^1_{t,d}, S^2_{t,d}, S^3_{t,d}),\ \forall d \in D_t$$

where D_t is the set of delivery hour which can be traded at time step t. These 3 subsets are defined as: (i) The offers available in the CIM at the moment we take the decision $S^1_{t,d}$. (ii) The data to characterize what we have contracted in the past $S^2_{t,d}$. (iii) Exogenous data that we think should influence our decision $S^3_{t,d}$. In this section, we will define the first two sets. We will characterize more precisely the last one in section V where we talk about the behaviors we expect from the policy.

1) Offers available in the CIM: We cannot put all the available offers in the state because it would require to take a decision on all of them which would give an intractable action space. Indeed, in a typical order book, there may exist more than 1000 bids available which would imply more than 2^{1000} possible actions. In order to overcome this problem, we discretize the quantity which can be contracted with $2n+1$ options $-q^n_{td}, -q^{n-1}_{t,d}, \cdots, q^1_{t,d}, 0, q^1_{t,d}, \cdots, q^{n-1}_{t,d}, q^n_{t,d}$. we also define the $2n$ cutoff between the different possibilities $-C_n, -C_{n-1}, \cdots, -C_1, C_1, \cdots, C_{n-1}, C_n$. The cutoffs are defined as $C_i = \frac{q_i + q_{i-1}}{2}, i \in 1 \cdots n$.

Finally, $S^1_{t,d} = (m^1_{t,d}, m^2_{t,d})$ where

- $m^1_{t,d} = (p_{t,d}(-C_n), \cdots, p_{t,d}(C_n))$ with $p_{t,d}(C_i)$, the value of the demand function evaluated at C_i MWh for delivery hour d at time step t;
- $m^2_{t,d} = (\text{rev}_{t,d}(-q_n), \cdots, 0, \cdots, \text{rev}_{t,d}(q_n))$ with $\text{rev}_{t,d}(q_n)$, the revenue of selling q_n MWh for delivery hour d at time step t;

2) Contracted quantity: This set includes the necessary information in order to represent the producer position at time step t. It contains $(v_{t,d}),\ \forall d \in D_t$, where $v_{t,d}$ represents the quantity that would be stored at delivery time d with the trades accepted at time step t or before if the producer does not trade anymore in the future.

B. Action variables

Our action space A_t contains $(a_{t,d}),\ \forall d \in D_t$, where $a_{t,d}$ represents the quantity we sell at time step t. This variable can take values:

$$a_{t,d} \in \{-q_n, \cdots, -q_1, 0, q_1, \cdots q_n\},\ \forall d \in D_t$$

C. Transition function

As we use reinforcement learning, we do not have to model the transition function completely. In our case, it would be really difficult to have a coherent model for the evolution of $S^1_{t,d}$ because it would require jointly modelling the bid arrival process for the different delivery time. On the other side, it is easy to model $S^2_{t,d}$ because its evolution is not stochastic. This model is given by equation 1. The interpretation is that the volume for one delivery time is equal to the volume for the same delivery time at the previous time step at which we subtract the quantity sold at the current time step for every products having an earlier delivery time.

$$\text{v}_{t,d} = \text{v}_{t-1,d} - \sum_{b \in D_t | b \leq d} a_{t,b},\ \forall d \in D_t \tag{1}$$

Previously, we mention that we want to avoid being in imbalance. To this aim, we impose constraints 2 in order to ensure that the reservoir capacity is feasible for each delivery time.

$$0 \leq v_{t,d} \leq V,\ \forall d \in D_t \tag{2}$$

D. Objective

The objective is defined as the sum of the revenue we get for every delivery time of the CIM.

$$R_t = \sum_{d \in D_t} \text{rev}_{t,d}(a_{t,d})$$

IV. Application of policy function approximation

There are two main categories of methods which are used in order to solve model free reinforcement learning. The first one is value function approximation. This method is used on a similar problem in [11]. The limitation is given by the number of data which are needed in order to learn the huge number of parameters of a deep neural network. In this paper, we use policy function approximation as described in [13] and [14]. The idea of this method is to parametrize the policy with respect to a parameter vector θ and to optimize this θ.

More precisely, we optimize our policy $\pi(a|s;\theta)$ with respect to θ where

$$\pi_\theta(a|s) = \mathbb{P}[A_t = a|S_t = s;\theta]$$

The advantage of policy function approximation is that the user can include all his knowledge about the problem directly in the policy parametrization [14]. It also gives policies that are easier to interpret, compared to value function approximation coupled with deep neural network, because it uses far less parameters.

We focus on a policy which is parametrized by buy and sell price thresholds. The threshold policy that we investigate in this paper accepts sell bids if their price is below the buy threshold, and accepts buy bids if their price is above the sell threshold. Our focus on threshold policies is justified by the fact that optimal intertemporal arbitrage in a deterministic setting is indeed achieved by a threshold policy, provided that the bounds of the reservoir are not binding[2]. In the next section, we will present the REINFORCE algorithm which is used in order to optimize θ. Then, we will show how to compute a closed form solution of the quantity needed in this REINFORCE algorithm.

A. REINFORCE algorithm

We employ the REINFORCE algorithm, as defined in [13] and [14], in order to determine the optimal thresholds for our policy. Denote by θ the parameter vector which characterizes a policy function. Then there exists a mapping (typically non-convex) from the parameter θ to the average payoff of the resulting policy. The goal of the algorithm is to optimize this vector θ on the basis of episodes, or epochs, of learning. An episode in the context of our problem is one day of CIM trading on the basis of training data. We consider 24 time steps separated by one hour each with the first one 30 minutes before the delivery of the first product (hour 1). It means that at the first time step, we can trade the 24 products. On the contrary, at the second time step, there is not any available trade for the first delivery time as the market for that delivery time is closed and therefore we can trade the 23 last products. Finally, at the last time step, we can only trade the last product. For every time step, we update θ according to algorithm 3 where g_t is the profit from t to the end of the episode T.

```
• Initialize $\theta$
• for each episode $\{s_1, a_1, r_2, \cdots, s_{T-1}, a_{T-1}, r_T\} \sim \pi(s,a;\theta)$
        for t = 1 : T-1 do
            $\theta = \theta + \alpha\nabla_\theta \log(\pi(s,a;\theta))g_t$     (3)
        end for
  end for
```

It has been proven in [13] and [14] that this algorithm update is following the gradient of the expected profit which gives us the guarantee to converge to a local optimum under standard stochastic approximation conditions for decreasing α.

B. Derivation of the policy and its gradient

The REINFORCE algorithm requires a policy which is differentiable with respect to the parameter vector over which the policy is parametrized. Therefore, we employ a stochastic threshold rather than a deterministic one. This idea has been applied on a simpler problem in [12] for a one step reinforcement learning problem and a single delivery time. More precisely, we define our policy parameter vector, θ, as $\theta = (\mu_X, \exp(\sigma_X), \mu_Y, \exp(\sigma_Y))$, where μ_X and μ_Y are the means of the normal distributions and $\exp(\sigma_X)$ and $\exp(\sigma_Y))$ are the standard deviations that we use for sampling the thresholds X and $\max(X,Y)$. The fact that we select X and $\max(X,Y)$ and not X and Y as thresholds is justified by the the fact that bids that are available in an order book are not matched. Therefore there must exist a bid-ask spread. It is therefore not profitable to simultaneously accept buy and sell bids that correspond to the same delivery time, since this leads to a certain financial loss.

Our policy is demonstrated in figure 2 for $n = 2$, $q_1 = 10$ MWh and $q_2 = 20$ MWh. C_1 and C_2 are therefore respectively equal to 5 and 15MWh. As indicated in the figure, the probability of accepting 0 MWh of buy bids (denoted as $\pi(s,0;\theta)$) is equal to the upper purple surface. The probability of accepting 10 MWh is the surface indicated in red. Finally, the probability of accepting 20 MWh is the surface indicated in the lower purple area.

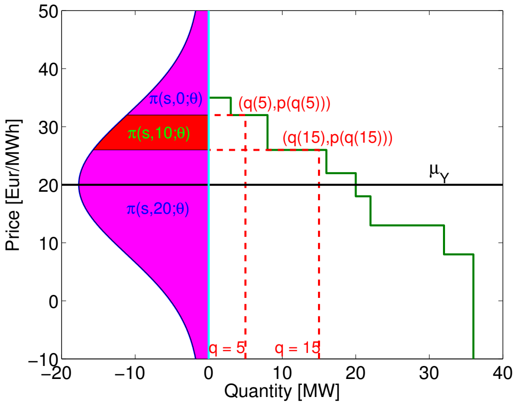

Fig. 2: Threshold for hydro problem. The bell curve indicates the probability density function of the sell threshold. The two purple segments and the red segment of the bell curve indicate the probability of each of the three actions. The green decreasing function corresponds to the buy bids that are available in the order book for a given trading hour.

In order to implement the REINFORCE algorithm, we need to express the probability of each action as a function of the parameters that characterize our policy. These probabilities can be expressed in closed form. We show an example of this closed form in equations 4 and 5.

[2] We demonstrate this in an online appendix which is available at: https://sites.google.com/site/gillesbertrandresearch/publications/app-gm2019

$$\begin{aligned}\pi(s, -q_n; \theta) &= \Pr(p(-C_n) \leq X) \\ &= 1 - F_X(p(-C_n)) \qquad (4)\\ \pi(s, q_n; \theta) &= \Pr(\max(X, Y) \leq p(C_n)) \\ &= F_X(p(C_n))F_Y(p(C_n)) \qquad (5)\end{aligned}$$

The REINFORCE algorithm also requires the derivatives of these probabilities with respect to the parameters that characterize the policy. One of these derivatives is given in equations 6. The others can be computed similarly.

$$\begin{aligned}\frac{\partial \pi(s, -q_n; \theta)}{\partial \mu_X} &= \frac{\partial(1 - F_X(p(-C_n)))}{\partial \mu_X} \\ &= f_X(p(-C_n)) \qquad (6)\end{aligned}$$

V. Policy definition

One of the interests of policy function approximation is to include the user's knowledge in the parametrization of the policy. In this section, we will first show that the threshold can be expressed as any differentiable function of the state. Then, we will present the behaviors that we have added in our policy. After that, we will describe precisely how we parametrize the policy in order to include one of this behaviour.

A. Policy parametrization

Let $f\ R^n \to R^4$ be a differentiable function s.t. $\theta = f(\alpha)$. We can compute the derivative with respect to α by using the chain rule.

$$\begin{aligned}\frac{\partial \pi(s; \theta)}{\partial \alpha} &= \frac{\partial \pi(s; \theta)}{\partial \theta}\frac{\partial \theta}{\partial \alpha} \\ &= \frac{\partial \pi(s; \theta)}{\partial \theta}\frac{\partial f}{\partial \alpha}\end{aligned}$$

Now, we can describe more precisely the subset $S^3_{t,d}$, it contains the exogenous factors which appears in the policy parametrization: $f(\alpha)$.

B. Expected behavior of the policy

The different behaviors we expect from our policy are: (i) Ensure that the energy stored stays in the reservoir limits. (ii) Adapt with respect to the particularity of the trading day. (iii) Adjust with respect to the delivery time. (iv) Adapt with respect to the information received during the day. (v) Adjust with respect to the remaining time before the market closure.

C. Threshold adaptation with the trading day particularity

In this section, we explain how we manage to have a threshold which adapts to the trading day. This is necessary because it is not possible to obtain good results without adapting to the particularity of each day as illustrated on the left graph of figure 3. From this graph, it is clear that it is not possible to set a single threshold which would give good results for the two days because their average level is completely different. In order to account for the difference between the different days, our idea is to include the IA price in the policy parametrization. In order to illustrate that the IA price convey a lot of information about the CIM price[3], we show on the right graph of figure 3 an histogram of the difference between the CIM and the IA price. It can be observed that the difference is centered at 0 and is relatively small.

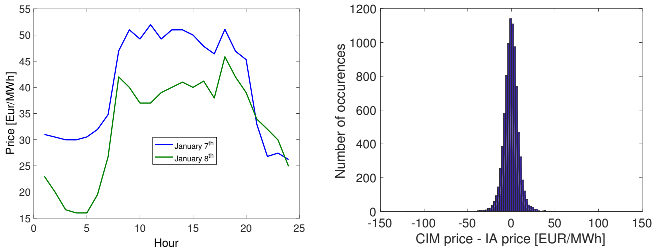

Fig. 3: CIM price for two different days (left). Histogram of price difference between the CIM price and the IA price for year 2015 (right).

After observing that the IA price was a good feature to include in the policy parametrization, θ is changed as in equations 7, 8, 9 and 10 where:

- p_{min} is the minimum of the IA curve.
- p_{max} is the maximum of the IA curve.
- $\alpha_1^{s/b}$ are the weights that will be optimized using the REINFORCE algorithm.

$$\begin{aligned}\mu_X &\leftarrow p_{\text{min}} + \alpha_1^s(p_{\text{max}} - p_{\text{min}}) \qquad (7)\\ \sigma_X &\leftarrow \sigma_X \qquad (8)\\ \mu_Y &\leftarrow p_{\text{max}} - \alpha_1^b(p_{\text{max}} - p_{\text{min}}) \qquad (9)\\ \sigma_Y &\leftarrow \sigma_Y \qquad (10)\end{aligned}$$

The idea behind this parametrization is that p_{min} is a natural candidate in order to initialize the buy threshold but it is a really aggressive value because there are chances that the market will never reach that value (it was the smallest one in the IA). Therefore, we add a security margin represented by α_1^s which tells us of which percentage should we move from p_{min} to p_{max}. This security margin α_1^s will be learned through experience by the REINFORCE algorithm.

VI. Case study: trading in the German CIM

In this section, we present results from the implementation of the proposed policy on the German CIM. The data has been obtained from the European Power Exchange (EPEX). For the purpose of this case study, we consider a pumped storage hydro with a maximum storage capacity of 200 MWh. We use as training set the 200 first days of 2015 and as a test set the remaining 165 last days of 2015. We will start this section by presenting our benchmark method, rolling intrinsic (RI) which is widely used in the industry. Then, we will present the evolution of our method during the learning phase. Finally, we will compare the results of our method with RI.

[3] When we refer to CIM price, we mean the center of the bid-ask spread at a certain moment.

A. Rolling intrinsic method

RI is a classical benchmark as explained in [17]. The idea of this method is to accept any trade which gives a positive profit if the contracted quantity remains in the reservoir bounds. The optimization model of RI is presented in the online appendix.

B. Learning process

In figure 4, we show the evolution of the profit against the iteration. At the beginning, the profit increases quickly, then it stabilizes. This is the classical behavior for reinforcement learning problem.

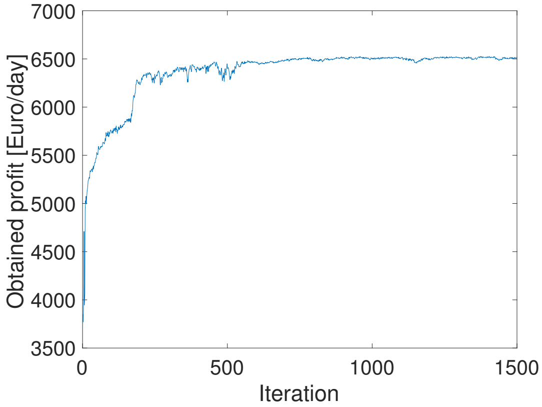

Fig. 4: Profit evolution against the iteration.

C. Profit comparison

In this section, we compare our results with RI on the last 165 days of 2015. As our policy is randomized, the result for each day is the average of 100 runs. In figure 5, we show the profit difference per day between our method and RI. On average, there are 96.16 days at which our profit is better. The average profit improvement compared to RI is 205.84 Euros per day which corresponds to an improvement of 4.2%.

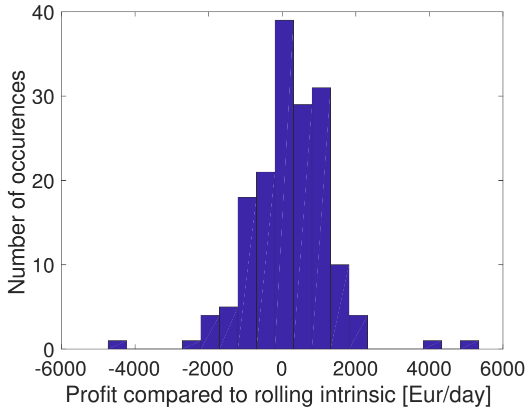

Fig. 5: Extra profit obtained compared to rolling intrinsic for the different days

VII. Conclusions and perspectives

In this paper we tackle the problem of CIM trading for pumped storage hydro resources. We model the problem using reinforcement learning. We focus on policies that are parametrized on price thresholds, and we optimize the resulting policy using the REINFORCE algorithm. We test our threshold policies on the German CIM and compare this with rolling intrinsic method which is commonly used in the industry. We demonstrate that our method performs significantly better than rolling intrinsic method. Moreover our policy function approximation methods can take decisions by only comparing the price of the bid with a threshold which verifies our requirement of short time to take decisions.

Acknowledgment

This research was funded by the ENGIE Chair on Energy Economics and Energy Risk Management, by an Electrabel grant on “Modeling the Value of Flexibility at Sub-Hourly Operating Time Scales”, and by the Belgian National Science Foundation (FSR-FNRS) through a FRIA grant. The authors also want to thank Gauthier de Maere and Guillaume Erbs for their helpfull comments during the development of this work.

References

[1] Eurostat [Online]. Available: http://ec.europa.eu/eurostat/statistics-explained/images/d/de/Table_3-Share_of_electricity_from_renewable_sources_in_gross_electricity_consumption_2004-2016.png

[2] EU [Online]. Available: https://ec.europa.eu/clima/policies/strategies/2020_en

[3] EPEXSPOT [Online]. Available: https://www.epexspot.com/fr/donnees_de_marche/intradaycontinuous

[4] T. K. Boomsma. ”*Bidding in sequential electricity markets: The Nordic case*” European Journal of Operational Research, 2014.

[5] S. Braun. ”*Hydropower Storage Optimization Considering Spot and Intraday Auction Market*” Energy Procedia vol. 87, pp. 36-44, 2016.

[6] R. Kiesel and F. Paraschiv, ” *Econometric analysis of 15-minute intraday electricity prices*” Energy Economics, Volume 64, 77-90, May 2017.

[7] R. Kiesel. ”*Modeling market order arrivals on the intraday power market for deliveries in Germany with Hawkes processes with parametric kernels*”, Energy Finance Christmas Workshop, 2017.

[8] R. Aid et al., ”*An optimal trading problem in intraday electricity markets*”, arXiv:1501.04575, 2015.

[9] S. Braun and R. Hoffmann. ”*Intraday Optimization of Pumped Hydro Power Plants in the German Electricity Market*” Energy Procedia vol. 87, pp. 45-52, 2016.

[10] A. Skajaa et al., ”*Intraday Trading of Wind Energy*” IEEE Transactions on Power Systems Volume 30, Issue 6, Nov. 2015.

[11] I. Boukas et al., ”*Intra-day Bidding Strategies for Storage Devices Using Deep Reinforcement Learning*” 15th International Conference on the European Energy Market, Dresden, Germany, June 27-29, 2018.

[12] G. Bertrand and A. Papavasiliou, ”*An Analysis of Threshold Policies for Trading in Continuous Intraday Electricity Markets*” 15th International Conference on the European Energy Market, Dresden, Germany, June 27-29, 2018.

[13] R. J. Williams, ”*Simple statistical gradient-following algorithms for connectionist reinforcement learning*”, Machine Learning, vol. 8, no. 23, 1992.

[14] R. Sutton and A. Barto, ”*Reinforcement Learning: An Introduction*” Nov 2017.

[15] EPEXSPOT [Online]. Available: https://www.epexspot.com/en/product-info/auction/germany-austria

[16] EPEXSPOT [Online]. Available: https://www.epexspot.com/en/product-info/intradayauction/germany

[17] N Lohndorf and D. Wozabal, *Optimal gas storage valuation and futures trading under a high-dimensional price process*, Technical report, Tech. Rep., 2015.