Intra-day Bidding Strategies for Storage Devices Using Deep Reinforcement Learning

Ioannis Boukas*, Damien Ernst*, Anthony Papavasiliou† and Bertrand Cornélusse*
*Department of Electrical Engineering and Computer Science
University of Liege, Liege, Belgium
Email: {ioannis.boukas, dernst, bertrand.cornelusse }@uliege.be
†Center for Operations Research and Econometrics (CORE)
Universite Catholique de Louvain, Louvain la Neuve, Belgium Email: anthony.papavasiliou@uclouvain.be

***Abstract*—The problem faced by the operator of a storage device participating in a continuous intra-day (CID) market is addressed in this paper. The goal of the storage device operator is the maximization of the cumulative rewards received over the entire trading horizon, while taking into account operational constraints. The energy trading is modeled as a Partially Observable Markov Decision Process. An equivalent state representation and high-level actions are proposed in order to tackle the variable number of the existing orders in the order book. The problem is solved using deep reinforcement learning (RL). Preliminary results indicate that the agent converges to a policy that scores higher total revenues than the "rolling intrinsic".**

I. Introduction

The efficient integration of renewable energy resources (RES) in future power systems as directed by the recent worldwide energy policy drive [1] has given rise to discussions related to the security, sustainability and affordability of the power system("The Energy Trilemma"). In this context, flexible energy sources such as storage devices (e.g. pumped-hydro storage units) able to accommodate the variability of the RES generation have a key role [2]. There is a need for a market place where such systems can valorise their smart planning and their provision of flexibility services to the power system [3]. High accuracy on the generation output of RES can only be achieved closer to the time of physical delivery. In that sense, a real-time energy market would be the most suitable candidate for storage devices.

The participation of storage devices in short-term energy markets has been extensively studied in the literature and often cast as an optimal resource allocation problem. In [4] a multi-stage stochastic programming framework is selected to represent the different sequential trading floors, namely the day-ahead (DA), the intra-day (ID) and the balancing market. The optimal bidding curves of a hydro-reservoir for the Nordic spot market are derived accounting for price uncertainty in [5]. Several measures, such as the value of the stochastic solution are used to estimate the quality of the considered formulation.

In these approaches, the intra-day market is considered as auction-based and is modeled as a single recourse action. For each trading period, the optimal quantity offered is derived according to the realization of various stochastic variables. However, in reality, for most countries trading in intra-day market is a continuous process.

The continuous intra-day (CID) market participation for a thermal generator is considered in [6] and approximate optimal strategies are derived. However, the ability to trade through an order book for multiple future periods and the time overlap between trading and imbalance settlement is not taken into account. In this paper we adopt the continuous market framework as proposed in [7]. We extend the trading agents considered in [7], where each agent is supposed to select the price to buy or sell its energy in a constant range. We propose a novel approach where the agent can learn an optimal trading policy through the interaction with a market simulator.

In this paper, we extend the real-time bidding strategies proposed in [8] for the case of a storage device. The sequential decision making problem of participating in the CID market is formulated as a Partially Observable Markov Decision Process (POMDP). The trading agent is supposed to dynamically select the orders that maximize its benefits through the entire horizon. The dynamics of the storage system as well as the specifications of the ID market are modeled. Due to the high dimensionality and the dynamically evolving size of the order book we motivate an equivalent state representation and the use of high-level actions. The goal of the selected actions is the identification of the opportunity cost of trading. We solve the intra-day trading problem of a storage device using reinforcement-learning techniques, more specifically the Deep Q-Network proposed in [9]. The resulting optimal policy is evaluated using real data from the German ID market [10].

II. Continuous Intra-Day market design

The CID market is a continuous process similar to the stock exchange market as presented in [7]. The need for a CID market is motivated by the reduction of imbalance costs, the optimization of participants' portfolios closer to real-time and the better exploitation of flexibility [11]. Each market product $x \in X$, where X is the set of all available products, corresponds to the physical delivery of energy in a pre-defined time-slot. As presented in Figure 1, every time-slot is defined by its starting point t_d^x and its duration λ. Participants express their willingness to buy or sell energy by posting orders o_i^x, where $i \in N \subseteq \mathbb{N}$ corresponds to the index of each order posted in order book O^x for product x. The trading process for time-slot x opens at $t_o^x = t_d^x - \tau$ and closes at t_c^x. For every time-step

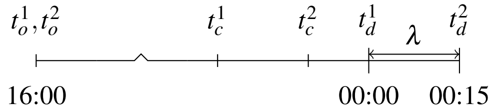


Fig. 1: Trading time-line for products Q-1 and Q-2

TABLE I: Order Book for Q-1 and time-slot 00:00-00:15

i	**Type**	v **[MW]**	p **[€/MWh]**	
4	"Sell"	6.25	36.3	
2	"Sell"	2.35	34.5	⟵ ask
1	"Buy"	3.15	33.8	⟵ bid
3	"Buy"	1.125	29.3	
5	"Buy"	2.5	15.9	

t in the trading horizon $t_o^x < t < t_c^x$, each participant has the possibility to place new orders or adjust existing orders.

For instance, in the German CID market, trading of hourly ($\lambda = 1$ hr) or quarterly ($\lambda = 15$ min) products for day D opens at 3p.m. or 4 p.m. of day $D-1$ respectively. The gate closes 30 *min* before the actual energy delivery. The time-line for trading products Q-1 and Q-2 that correspond to the 15 min time-slots 00:00-00:15 and 00:15-00:30 respectively, is presented in Figure 1.

In practice, the available contracts (buy and sell orders) fall in three categories, i.e. the market order where no limit price is specified (order is matched at the best price), the limit order that contains a price limit and can be matched only at that or at a better price and the market sweep order that is executed immediately (fully or partially) or gets canceled. Limit orders may appear with restrictions related to their execution and their validity. For instance, an order that carries the specification Fill or Kill should either be fully and immediately executed or canceled. An order that is specified as All or Nothing remains in the order-book until it is entirely executed. In this paper, for the sake of simplicity all the orders are assumed to be limit orders without any particular specifications.

After the gate opens, participants submit orders with the predefined specifications. The orders are treated according to the first come first serve (FCFS) rule. Table I contains all the available orders o_i^x defined by their type ("sell" or "buy), the volume level v and the price level for each energy unit p. The difference between the most expensive buy order (bid") and the cheapest sell" order (ask") defines the bid-ask spread of the product. A deal between two counter-parties is struck when the price p_{buy} of a "buy" order and the price p_{sell} of a "sell" order satisfy the condition $p_{buy} \geq p_{sell}$. This condition is tested at the arrival of each new order. The volume of the transaction is defined as the minimum quantity between the "buy" and "sell" order $\min(v_{buy}, v_{sell})$. The residual volume will remain available in the market at the same price.

III. Problem statement

The problem faced by the storage device operator is the selection of the optimal sequence of orders that maximizes its revenues over the entire trading horizon. The sequential decision making problem for ID market participation is formulated as a Partially Observable Markov Decision Process (POMDP) as in [8].

Two modules are used to describe the simulation environment: the "Storage" module models the transition dynamics of the storage device and the "ID Market Simulator" simulates the transition dynamics of the ID market. The state of the trading agent $s_t \in S = \left\{s_t^I, s_t^E\right\}$ is composed of the internal $s_t^I \in S^I$ ("Storage" module) and the external $s_t^E \in S^E$ ("ID Market Simulator") state. The agent can interact with the simulation environment by selecting an action a_t and observing the subsequent state of the environment. The trading agent can decide whether to accept (partially or fully) or not the existing orders o_i^x for each open product in the order book O^x. The action matrix is $a_t \in A = \{0,1\}^{|N|\times|X|}$, where X is the set of available products and $N \subseteq \mathbb{N}$ is the number of unmatched orders for each product. The transition from state s_t to the next state s_{t+1} is described by equation (1), where the arrival of new orders is denoted by the exogenous parameter ω_t sampled from a process as shown in equation (2).

$$s_{t+1} = f\left(s_t, a_t, \omega_t\right) \tag{1}$$

$$\omega_t \sim p_{\mathscr{W}}(\cdot) \tag{2}$$

At every time-step t in the trading horizon, the internal state $s_t^I \in S^I = \{s_t^B\}^{|X|}$ contains the variables that describe the transition dynamics of the storage device. In particular, it contains the projection of the state of charge of the storage device s_t^B for every open time-slot. The internal state s_t^I is updated according to function g, based on the action matrix a_t selected by the agent:

$$s_{t+1}^I = g\left(s_t^I, a_t\right). \tag{3}$$

The state of the ID market is represented by the external state $s_t^E \in S^E = \{v, p\}^{|N|\times|X|}$. At each time-step t of the trading horizon the "ID Market Simulator" outputs the set of orders O. As shown in equation (4), the state at $t+1$ depends on the state of the previous time-step t, the orders accepted by the trading agent a_t and the stochastic arrival of new orders ω_t. The orders accepted by the agent are removed from the order book at the next time-step $t+1$.

$$s_{t+1}^E = z\left(s_t^E, a_t, \omega_t\right) \tag{4}$$

The transition dynamics of the whole system (1) are described by (2) and (4) as

$$f\left(s_t, a_t, \omega_t\right) = F\left(g\left(s_t^I, a_t\right), z\left(s_t^E, a_t, \omega_t\right)\right) \tag{5}$$

The instantaneous reward signal $r_t = \rho\left(s_t, a_t, s_{t+1}\right)$ collected after each transition is defined as the trading revenues at time-step t as shown in equation (6).

$$r_t = \sum_{x=1, i=1}^{X,N} a_{t,x,i} v_{t,x,i} p_{t,x,i}. \tag{6}$$

The objective of the trading agent is the maximization of the total received rewards in the end of the trading horizon. Thus, we define in equation (7) the return G_t at each time-step t, as the sum of the discounted rewards received over the rest of the trading horizon (roll-out) [12]. The discount factor $\gamma \in [0,1]$ is used to adjust the strategy of the agent to be myopic or not.

$$G_t = \sum_{k=0}^{T-t-1} \gamma^t \cdot r_{t+k+1} \tag{7}$$

IV. Solution technique

The state-action value function Q following policy π is defined by [12] as :

$$Q(s,a) = \mathbb{E}_\pi \left[G_t | s_t = s, a_t = a\right] \tag{8}$$

The optimal solution to the problem defined in equations (2)-(7) corresponds to the identification of the policy that maximizes the expected returns over the trading horizon T. The optimal policy $\pi^* = [a_0^*, a_1^*, a_2^*, ..., a_T^*]$ is given by solving

$$Q^*(s,a) = \max_\pi \mathbb{E}_\pi \left[\sum_{k=0}^{T-t-1} \gamma^t \cdot r_{t+k+1} | s_t = s, a_t = a\right] \tag{9}$$

$$\pi^* = \arg\max_\pi Q(s,a) \tag{10}$$

The agent is able to learn the state-action value function Q by approximating it using a Deep Q-Network [9]. Through a series of episodic interactions with its environment, the agent can extract an optimal policy without the need for an explicit model of the system. A neural network (NN) is used to approximate the value function due to the large and continuous state space. The parameters θ_k of the Q-Network ($Q(s_t, a; \theta_k)$) are updated using samples of quadruples (s_t, a_t, r_t, s_{t+1}) obtained by simulated experiences. The objective function to be minimized is the temporal difference error δ. The goal is the back-propagation of total rewards early stages in the decision process. As proposed in [9], the final Q-values obtained from training the neural network are the solution to the supervised learning problem presented in equations (11) and (12), with $\alpha, \gamma \in (0,1]$.

$$\delta = r_{t+1} + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_k) - Q(s_t, a; \theta_k) \tag{11}$$

$$\theta_{k+1} = \theta_k + \alpha \delta \nabla_{\theta_k} Q(s_t, a; \theta_k) \tag{12}$$

V. State-space representation

The high-dimensional continuous external state $s_t^E \in S^E = \{v,p\}^{|N| \times |X|}$ defined in section III is used to describe the state of the CID market. Owing to the variable (non-constant) amount of orders $|N|$ in each order book O^x for product $x \in X$, the state-space S^E does not have a constant size. In order to approximate the Q-function using a NN as described in section IV it is necessary to find an approximate representation of the external state with constant size.

In Figure 2a the market depth for the products Q-1 to Q-6 at one time instant is presented. The market depth for each side ("sell" or "buy") at any point in time is defined as the total available volume in each order book O^x. The market depth per price level for "sell" or "buy" orders is computed by accumulating the available volume in ascending or descending price order respectively. In Figure 2a the bid-ask spread can be identified as the distance between the most expensive "buy" and the cheapest "sell" order. The bid-ask spread and the market depth are indicators of the liquidity in a market. In a liquid market there is always a counter-party willing to exchange a product in minimum time while fulfilling several requirements.

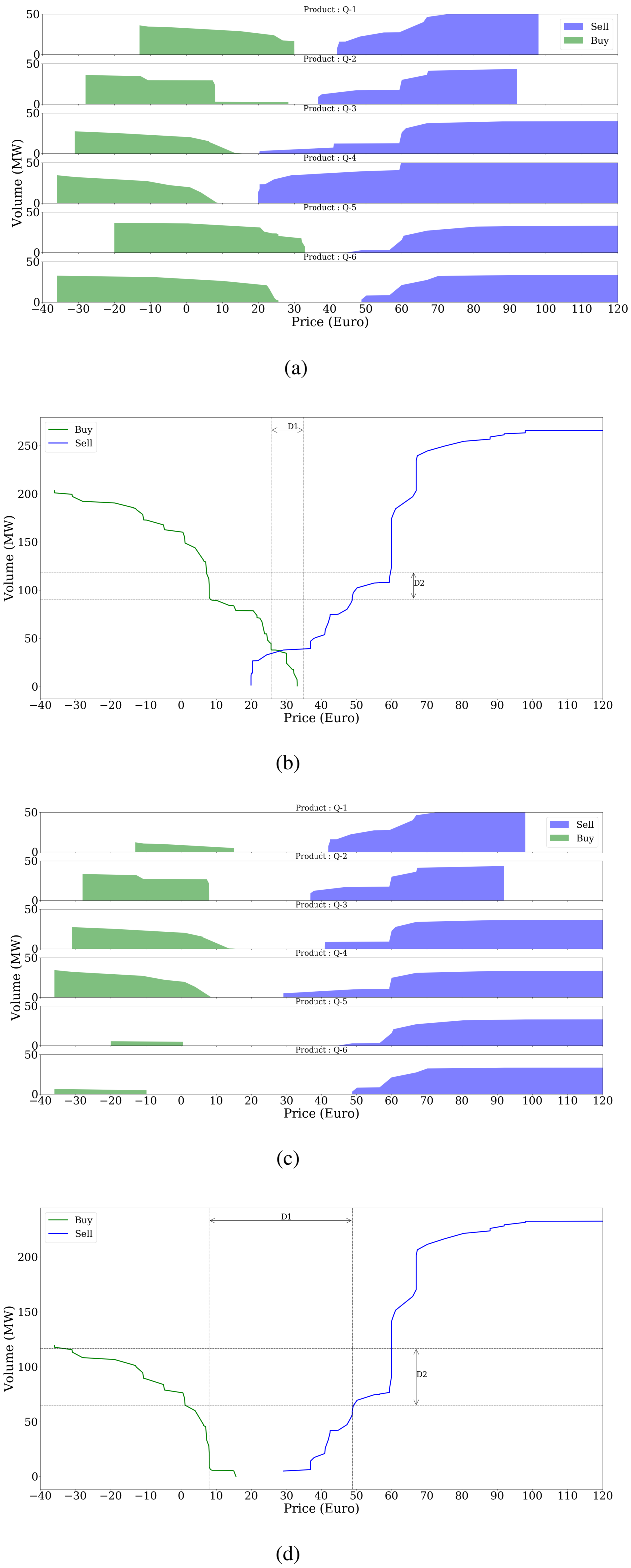
In the case of a storage device the main profit-generating mechanism is the arbitraging between two time-steps. Its functionality allows the charging of electricity in periods of low prices and discharging in periods of high prices. In the framework of the CID market a storage device would buy energy from those willing to sell at a low price and would sell this volume back to those willing to buy at a higher price. For instance, in Figure 2a a storage device would buy volume for product Q-4 and sell volume back for product Q-5.

Owing to the nature of a storage device it is then equivalent to represent the individual order books shown in Figure 2a as the aggregated curves presented in Figure 2b. These curves correspond to the aggregated market depth, i.e. the total available volume ("sell" or "buy") per price level for all the "open" products. The intersection of the "sell" and "buy" curves in Figure 2b defines the maximum volume that can be arbitraged by the storage device and serves as an upper bound for the profits at each step in the trading horizon. The market depth for the same products Q-1 to Q-6 at a different time step of the trading horizon is presented in Figure 2c. Figure 2d illustrates that there is no arbitrage opportunity between the products, since the aggregated curves do not intersect.

The need for a low-dimensional state space with constant size and the equivalent representation of the order book with aggregated curves motivate the use of descriptive statistics as presented in Figures 2b and 2d. More precisely we define as $D1$ the signed distance between the 75th percentile of "buy" price and the 25th percentile of "sell" price and as $D2$ the absolute distance between the mean value of "buy" and "sell" volumes. Other measures used are the signed price difference and absolute volume difference between percentiles (25%, 50%, 75%) and the bid-ask spread. The new continuous low-dimensional external state $s_t^{\prime E} \in S^{\prime E} = \{D1, D2, .., D10\}$ is used to categorize the observed order book based on its profit potential. The state of the trading agent is redefined as $s_t \in S = \left\{s_t^I, s_t^{\prime E}\right\}$.

VI. High level actions

At every time-step t in the trading horizon the agent can accept or not each of the available orders in the order book. The total number of orders $|N|$ contained in the order book is not constant throughout the trading horizon. Thus, the size of the action space $A = \{0,1\}^{|N| \times |X|}$ is not constant. However, in order to ensure the tractability of the problem, a small and discrete action space is necessary [12]. Therefore, we define an action space A' composed of two high-level actions. Each of these high-level actions is a mapping into the original action space A. Following the first action, defined as "Idling", no transactions are executed and no adjustment is made to the previously scheduled quantities. Under the second action, defined as "Optimizing based on current knowledge", the agent



(a)

(b)

(c)

(d)

Fig. 2: Market depth per product and the corresponding aggregated curves for profitable (a,b) and non-profitable (c,d) order book.

TABLE II: Optimizing based on current knowledge.

$$\max_{a_{i,x}} \sum_{x=0}^{X} \sum_{i=0}^{N} a_{i,x} v_{i,x} p_{i,x} \tag{13}$$

$$\text{s.t.} \sum_{i=0}^{N} a_{i,x} v_{i,x} + Y_x^{ID} + p_x^{DIS} = p_x^{CH} \qquad \forall x \in X \tag{14}$$

$$s_{x+1}^{B} = s_x^{B} + \eta p_x^{CH} - \frac{p_x^{DIS}}{\eta} \qquad \forall x \in X \tag{15}$$

$$S^{B,min} \leq s_x^{B} \leq S^{B,max} \qquad \forall x \in X \tag{16}$$

$$0 \leq p_x^{CH} \leq k_x P^{CH,max} \qquad \forall x \in X \tag{17}$$

$$0 \leq p_x^{DIS} \leq (1-k_x) P^{DIS,max} \qquad \forall x \in X \tag{18}$$

$$k_x \in \{0,1\} \qquad \forall x \in X \tag{19}$$

$$a_{i,x} \in \{0,1\} \qquad \forall i,x \in N \times X \tag{20}$$

trades based on the observed orders and the state of the storage device at time-step t. The bid acceptance optimization model is presented in Table II. The objective of this strategy formulated in equation (13) is the maximization of the revenues arising from trading, subject to the operational constraints of the storage device. In equation (14) the energy purchased and sold ($\sum_{i=0}^{N} a_{i,x} v_{i,x}$), the past net energy trades (Y_x^{ID}) and the energy discharged by the storage (p_x^{DIS}) must match the energy charged by the storage (p_x^{CH}) for every time-slot x. The energy balance of the storage device, presented in equation (15), is responsible for the time-coupling and the arbitrage between two products (time-slots). The technical limits of the storage level and the charging and discharging process are described in equations (16) to (18). The binary variable k_x restricts the operation of the unit in only one mode, either charging or discharging.

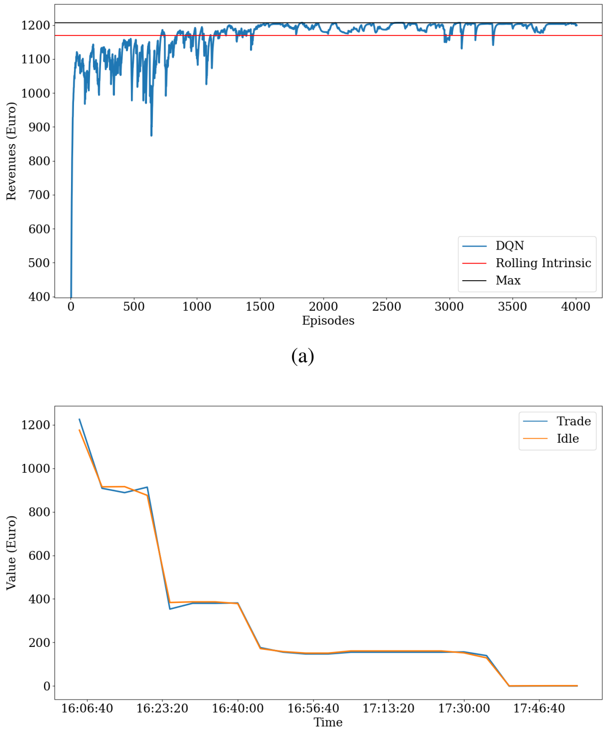
At every time step t the agent can select between two high level actions ($a_t' \in A' = \{0,1\}$). In case $a_t' = 1$, the solution to the bid acceptance optimization problem presented in Table II, is the matrix a_t. In the case of "Idling" ($a_t' = 0$), the matrix a_t is a zero matrix. The optimal policy is drawn according to equation (10).

The approach that we propose in this paper thereby allows us to quantify the value that is associated to the decision of the agent to wait at certain occasions. We compare this approach to an alternative, which we refer to as the "rolling intrinsic" policy, according to which the agent will trade at every time step of the trading horizon based on the current information [13]. In this alternative approach, the agent selects a combination of orders that optimizes its operation and profits. Instead, if the agent decides to wait, there might be a better combination of orders appearing in the order book of the next time step. Thus, by exploiting the experience gained through the interaction with its environment, the agent is able to learn the value of trading or waiting at every different state that it may encounter [8].

VII. Case Study

The proposed methodology is applied for a pumped-hydro energy storage unit using the following parameters: $S^{B,max} = 500MWh$, $P^{CH,max} = P^{DIS,max} = 500$ MW, $\eta = 100\%$, $X = \{Q-1, Q-2, Q-3, .., Q-12\}$, $\Delta t = 5$ min, $\gamma = 1$, and

$\alpha = 0.0005$. In this paper we extend the results presented in [8], by considering 12 quarterly products available for trading. The trading horizon is assumed to be equal to two hours, and the agent can decide on an action every five minutes. Real data from the German CID market are used in order to simulate the arrival of new orders. The neural network that is used in this analysis is a feed-forward multilayer perceptron with four hidden layers and 512 nodes per layer. We compare the obtained policy with the "rolling intrinsic"[13]. According to this policy, we apply the "Optimizing based on current knowledge" at every time step of the horizon.



(a)

(b)

Fig. 3: The evolution of the learning process (a) and the Q-values per action (b).

Preliminary results demonstrate that the agent is able to converge to a policy on a much larger problem than the one investigated in [8]. Figure 3a illustrates that after 3500 episodes the agent has been able to learn a policy that corresponds to higher total revenues than that obtained by the "rolling intrinsic". It is important to note that the agent converges to the policy that results in the maximum observed total revenues.

The evolution of the Q-values for each action over the trading horizon is presented in Figure 3b. In effect, the Q-values obtained correspond to the expected value of the returns of each state-action pair as indicated in equation (8). For instance, the cumulative rewards of the episode are successfully back-propagated to the first trading step and the values for both actions decrease as the episode advances. There are time-steps in the episode where idling instead of trading results in higher total revenues and consequently has a larger value. It is also important to note that for several time-steps both actions take similar values because both lead to the same (zero) instantaneous reward. We can identify several points that the values slightly increase as the episode progresses. This occurs due to the approximation error of the Q-function and highlights the significance of a more adequate state representation. Finally, the optimal policy is obtained by following the sequence of actions that has the maximum Q-value and results in the highest cumulative rewards.

VIII. Conclusion

The participation of a storage device in the CID market is investigated. In this novel approach, the sequential decision making problem is modeled as a POMDP and solved using Deep-Q networks. Due to the variable size of the order book, a new state representation and the use of high-level actions were motivated. The main goal is the identification of the opportunity cost faced by the trading agent between trading and idling. The proposed methodology is applied to real ID data from the German market. Preliminary results demonstrate the ability of the agent to learn an optimal policy that results in higher revenues than the "rolling intrinsic".

In future work, the proposed methodology will be used to train the agent with a larger dataset and to validate its performance on unseen data. Moreover, a better representation of the state will be able to minimize the numerical errors.

IX. Acknowledgments

This research is supported by an ENGIE Electrabel grant on "Modeling the Value of Flexibility on Sub-Hourly Operating Time Scales". Bertrand Cornélusse would like to thank Nethys for its financial support.

References

[1] E. Commission. (2017). 2030 energy strategy, [Online]. Available: https://ec.europa.eu/energy/en/topics/energy-strategy-and-energy-union/2030-energy-strategy.

[2] A. Papalexopoulos, R. Frowd, C. Hansen, E. Lannoye, and A. Tuohy, "Impact of the transmission grid on the operational system flexibility," *Power Systems Computation Conference (PSCC)*, 2016.

[3] E. Nasrolahpour, H. Zareipour, W. D. Rosehart, and S. J. Kazempour, "Bidding strategy for an energy storage facility," in *2016 Power Systems Computation Conference (PSCC)*, 2016, pp. 1–7.

[4] T. K. Boomsma, N. Juul, and S.-E. Fleten, "Bidding in sequential electricity markets: The nordic case," *European Journal of Operational Research*, vol. 238, no. 3, pp. 797 –809, 2014, ISSN: 0377-2217. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221714003695.

[5] S. E. Fleten and T. K. Kristoffersen, "Stochastic programming for optimizing bidding strategies of a nordic hydropower producer," *European Journal of Operational Research*, vol. 181, no. 2, pp. 916 –928, 2007, ISSN: 0377-2217.

[6] R. Aïd, P. Gruet, and H. Pham, "An optimal trading problem in intraday electricity markets," *Mathematics and Financial Economics*, vol. 10, no. 1, pp. 49–85, 2016.

[7] D. Ilic, P. G. D. Silva, S. Karnouskos, and M. Griesemer, "An energy market for trading electricity in smart grid neighbourhoods," in *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 2012, pp. 1–6.

[8] I. Boukas, D. Ernst, and B. Cornélusse, "Real-time bidding strategies from micro-grids using reinforcement learning," in *CIRED Workshop 2018*, 2018, pp. 1–4.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, ISSN: 00280836. [Online]. Available: http://dx.doi.org/10.1038/nature14236.

[10] EPEXSPOT. (2017). Market data intraday continuous, [Online]. Available: http://www.epexspot.com/en/market-data/intradaycontinuous.

[11] R. Scharff and M. Amelin, "Trading behaviour on the continuous intraday market elbas," *Energy Policy*, vol. 88, pp. 544 –557, 2016, ISSN: 0301-4215. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301421515301713.

[12] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st. Cambridge, MA, USA: MIT Press, 1998.

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