Real-Time Bidding Strategies from Micro-Grids Using Reinforcement Learning

Ioannis BOUKAS
ULiège - Belgium

Damien ERNST
ULiège - Belgium

Bertrand CORNÉLUSSE
ULiège - Belgium

{ioannis.boukas, dernst, bertrand.cornelusse}@uliege.be

ABSTRACT

We address the problem faced by the operator of a microgrid participating in a continuous real-time market. The microgrid consists of distributed generation, flexible loads and a storage device. The goal of the microgrid operator is the maximization of the profits over the entire trading horizon, while taking into account operational constraints. The variability of the Renewable Energy Sources (RES) is considered and the energy trading is modeled as a Markov Decision Process. The problem is solved using reinforcement learning (RL). The resulting optimal real time bidding strategy of a microgrid is discussed.

INTRODUCTION

The concept of smart microgrid is an efficient way to integrate renewable generation and to exploit the available flexibility in a decentralized manner. A microgrid consists of variable distributed generation (e.g. solar, wind), storage systems and flexible demand. The smart control of these components at a local decentralized level results in a more efficient utilization of these resources [1]. Owing to its flexible nature the microgrid can support the operation of the grid by managing its energy production and consumption, and by providing ancillary services. In order to succeed in the integration of microgrids to the existing distribution network, financial incentives should be given to investors. Aside from the more efficient use of resources and the integration of green renewable energy, the main driving factor for investment is the benefit arising by the services provided to the main grid. There is a need for a market floor where the microgrids could reach out in order to valorize their smart planning and the provision of flexibility services to the power system. High accuracy on the generation output of RES can only be achieved close to physical delivery. In that sense, a real-time energy market would be the most suitable candidate for microgrids [2].

Little attention is given in the literature to the dynamic interaction of microgrids with the energy markets. In [3], the participation of the Skagen CHP unit combined with heat storage in the spot and regulation markets illustrates experimentally that such structures are able to provide services to the main grid. They benefit directly from the financial rewards, and indirectly from the efficient operation of the resources. The interaction of microgrids through a wholesale market in an islanded power system is proposed in [4]. The authors propose a multi-agent case study, where each agent constructs bidding curves and participates in a competitive environment while optimizing its own operation. However, the system is not considered to be connected to the main grid and the variety of existing market floors is not taken into account. Short-term energy markets participation has been studied extensively in the case of large-scale units. In [5], the authors address the sequential decision making problem of a hydro-electric plant participating in the Nord pool day-ahead and intra-day market. The intra-day market floor is usually assumed to be auction-based. For each trading period, the optimal quantity to be offered is derived according to the realization of various stochastic variables. However, in practice, trading in the intra-day market is a continuous process, where participants can exchange energy bilaterally, almost until the physical delivery.

In this paper, in accordance to this market structure, we assume a framework similar to [6] where the intra-day participation is formulated as a continuous process and solved explicitly. We model the decision making process of a microgrid operator that is exchanging energy in a real-time market. We extend the trading agents proposed in [2], where the agent is supposed to select the price to buy or sell its energy in a constant range, by proposing a novel approach where the agent can learn an optimal trading policy through interaction with a market simulator. After a brief description of the real-time market framework, we detail the sequential decision making problem faced by the microgrid operator, explain and illustrate our proposed solution method, and conclude.

REAL-TIME MARKET DESIGN

The real-time market is a continuous process similar to the stock exchange market as presented in [2]. Participants can trade energy through bilateral trades. Each market product $x \in X$, where X is the set of all available products, corresponds to the physical delivery in a time-slot. As presented in Figure 1, every time-slot is defined by its starting point t_d^x and its duration δ. 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 t in the trading horizon $t_o^x < t < t_c^x$, each participant has the ability to place new orders or adjust existing orders. There are three types of contract in practice, namely the market order (order matched at the best available price), the limit order

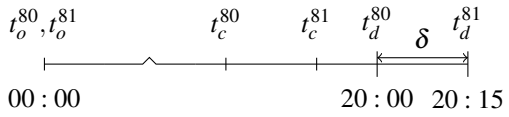


Figure 1: Trading time-line for products Q-80 and Q-81

Table 1: Order Book for Q-80 and time-slot 20:00-20:15

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	

(buy/sell orders executed at a price better than a price limit) and the market sweep order (order that is immediately fully or partially executed or canceled). Limit orders can have different specifications regarding their execution or 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 None remains in the order-book until it is entirely executed. In this paper all the orders are assumed to be limit orders for simplicity. As presented in Table 1, each order o is defined by a type ("sell" or "buy"), an energy volume v and a price per unit p. At every instant, there exists a set of unmatched "sell" and "buy" orders for each product. The difference between the most expensive "buy" order ("bid") and the cheapest "sell" order ("ask") defines the bid-ask spread of the product. A transaction between a "buy" order with price p_{buy} and a "sell" order with price p_{sell} will occur if $p_{buy} \geq p_{sell}$. The validity of this condition is evaluated at the arrival of every new order. The strike price for this transaction is defined as $\min(p_{buy}, p_{sell})$. The volume at which the transaction is executed can be expressed as $\min(v_{buy}, v_{sell})$ and the residual quantity will remain in the market at the same price.

PROBLEM STATEMENT

The considered microgrid is composed of flexible and non-flexible consumption, storage capacity and solar generation. The operator has to choose from the order book the optimal combination of orders that maximizes its revenues over the entire trading horizon.

We formulate the sequential decision making problem for real-time market participation as a Markov Decision Process (MDP). The simulation environment is composed of two modules. The "microgrid" module is used to simulate the transition dynamics of the microgrid components. The "Real-Time Market Simulator" describes the transition dynamics of the real-time market. Together, they form the state $s_t = \{s_t^I, s_t^E\} \in S$ (internal $s_t^I \in S^I$ and external $s_t^E \in S^E$) of the trading agent. The agent has the ability to interact with its environment by taking an action a_t and observing the following state of the environment. The trading agent can either accept or not every unmatched order (o_i^x) for every open time-slot in the order book O^x. The available set of actions 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 every state s_t to the following state s_{t+1} is defined by equation (1), where the stochastic arrival of new orders is denoted by the exogenous parameter ω_t sampled from a process as shown in equation (2).

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

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

At every time-step t in the trading horizon, the internal state $s_t^I \in S^I = \{\phi_t, c_t, s_t^B\}^{|X|}$ contains the variables that describe the transition dynamics of the microgrid. In particular, it contains the solar production ϕ_t, the net consumption c_t and the storage operation state s_t^B for every open time-slot. The function g dictates how the internal state is updated based on the actions taken by the agent a_t:

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

The external state $s_t^E \in S^E = \{v, p, \delta\}^{|N|\times|X|}$ captures the transition dynamics of the real-time market. The "Real-Time Market Simulator" produces the available set of orders O at each trading time-step t based on the stochastic arrival of new orders ω_t and the actions taken by the trading agent a_t, as shown in equation (4). In case the agent has decided to trade energy at time-step t the orders chosen will no longer be available at the next time-step $t+1$.

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

Overall, the transition dynamics of the whole system (1) can be written using (2) and (4) as

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

The instantaneous reward signal $r_t = \rho(s_t, a_t, s_{t+1})$ collected for the transition performed at each time step t, is

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

The goal of the trading agent is to maximize the reward it receives over the entire trading horizon. Thus, at every time-step t, the return G_t is defined as the sum of the discounted rewards the agent will receive over the rest of the trading horizon: $G_t = \sum_{k=0}^{T-t-1} \gamma^t \cdot r_{t+k+1}$ In the case of a finite horizon episode, the discount factor $\gamma \in [0,1]$ is used to simulate whether the agent is myopic or not.

SOLUTION TECHNIQUE

According to [7], the value Q of every state-action pair under policy π is defined as:

$$Q(s,a) = \mathbb{E}_\pi [G_t | s_t = s, a_t = a] \quad (7)$$

To solve this problem we need to find a policy that maximizes the expected discounted rewards over the entire trading horizon T. The optimal policy corresponds to the optimal sequence of actions $\pi^* = [a_0^*, a_1^*, a_2^*, ..., a_T^*]$ and can be inferred by solving

$$Q^*(s,a) = \max_\pi \mathbb{E}_\pi \left[\sum_{k=0}^{T-t-1} \gamma^t \cdot r_{t+k+1} | s_t = s, a_t = a \right] \quad (8)$$

$$\pi^* = \arg\max_\pi Q(s,a) \quad (9)$$

We apply the Q-learning algorithm due to its simplicity and the ease of transitioning from a state value function to an optimal control policy by choosing in every state the action with the highest value. According to this algorithm, the state-action value function is obtained by a series of episodic interactions of the agent with its environment. Convergence is guaranteed by making use of the temporal difference updates of step-size α [7], as shown in equations (10)-(11), with $\alpha, \gamma \in (0,1]$.

$$\delta = r_{t+1} + \gamma max_a Q(s_{t+1}, a) - Q(s_t, a) \quad (10)$$

$$Q(s_t, a) \longleftarrow Q(s_t, a) + \alpha\delta \quad (11)$$

Due to the continuous nature of the state space, we approximate the Q^* function with a Neural Network (NN). The parameters θ_k of the Q-Network, denoted by $Q(s_t, a; \theta_k)$, are updated by sampling mini-batches of quadruples (s_t, a_t, r_t, s_{t+1}) of simulated experiences from a memory buffer. The final Q-values are obtained by solving the supervised learning problem presented in equations (12)-(13) as proposed in [8].

$$\delta = r_{t+1} + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_k) - Q(s_t, a; \theta_k) \quad (12)$$

$$\theta_{k+1} = \theta_k + \alpha\delta\nabla_{\theta_k} Q(s_t, a; \theta_k) \quad (13)$$

HIGH LEVEL ACTIONS

The actions available for an intra-day trading agent are the acceptance or not of each order available in the order book at every time step. However, due to the continuously changing nature of the size of the available orders (N), the action space ($A = \{0,1\}^{|N|\times|X|}$) becomes dynamically changing and very large. In order to make the problem tractable, it is necessary to have a discrete and static action space [7]. Therefore, we adopt a small and discrete action space A' that contains two high level strategies that we can map into the original action space A. The first is "Stay idle". Under this strategy the trader chooses not to make any adjustment to the scheduled quantities and thus no transaction takes place. The second is "Optimize based on current knowledge". Under this strategy the trading agent performs transactions based on knowledge regarding the state of the microgrid and the orders that are currently available. The objective of the optimization model presented in Table 2 is the maximization of profit under the operational constraints of the microgrid. In equation (15), the energy balance between the energy purchased and sold ($\sum_{i=0}^{N} a_{i,x} v_{i,x}$), the past net energy trades (Y_x^{ID}), the solar production (ϕ_x) and the energy discharged by the storage (p_x^{DIS}) must match the consumption (c_x) and the energy charged by the storage (p_x^{CH}) for every time-slot x. The state of charge of the storage device is updated for each time-slot according to equation (16). The technical limits of each component are defined in the set of equations (17)-(20). Due to the flexible nature of the demand, the consumption can be adjusted at every time-step between varying limits indicated in (20). The simultaneous charging and discharging of the storage device is prohibited using the binary variable k. At every time step the learning agent will choose between two high level strategies in $a'_t \in A' = [0,1]$. In case $a'_t = 1$, the agent solves the bid acceptance optimization problem presented in Table 2, that will result in the selection of the optimum orders that corresponds to obtaining $a_t \in A$. In the case of idling, the matrix a_t is a zero matrix. In this way, a policy is drawn according to equation (9).

Table 2: Optimize based on current knowledge.

$$\max_{a_{i,x}} \sum_{x=0}^{X} \sum_{i=0}^{N} a_{i,x} v_{i,x} p_{i,x} \quad (14)$$

$$\text{s.t.} \sum_{i=0}^{N} a_{i,x} v_{i,x} + Y_x^{ID} + \phi_x + p_x^{DIS} = c_x + p_x^{CH} \quad \forall x \in X \quad (15)$$

$$s_{x+1}^B = s_x^B + \eta p_x^{CH} - \frac{p_x^{DIS}}{\eta} \quad \forall x \in X \quad (16)$$

$$S^{B,min} \leq s_x^B \leq S^{B,max} \quad \forall x \in X \quad (17)$$

$$0 \leq p_x^{CH} \leq k P^{CH,max} \quad \forall x \in X \quad (18)$$

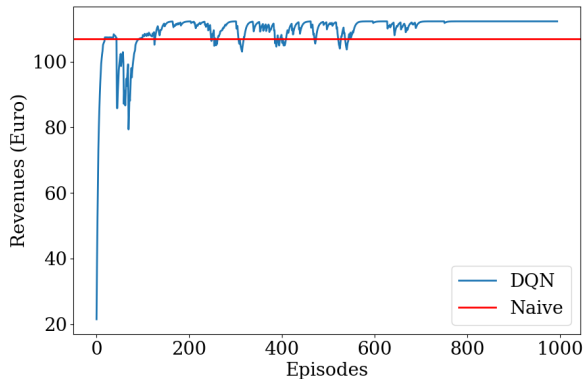
$$0 \leq p_x^{DIS} \leq (1-k) P^{DIS,max} \quad \forall x \in X \quad (19)$$

$$C_x^{min} \leq c_x \leq C_x^{max} \quad \forall x \in X \quad (20)$$

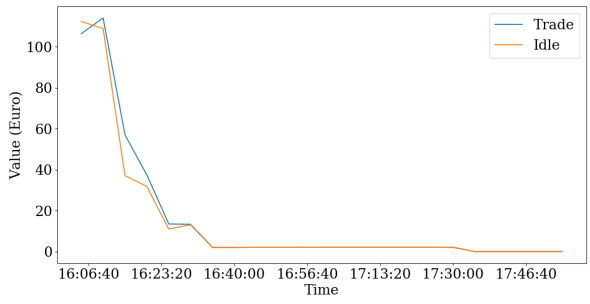
$$k_x \in \{0,1\} \quad \forall x \in X \quad (21)$$

$$a_{i,x} \in \{0,1\} \quad \forall i \in N \quad \forall x \in X \quad (22)$$

The goal of this approach is the identification of the opportunity cost attached to the decision of the agent to trade. Following the naive policy the agent will choose to trade at every time step of the trading horizon. Under this decision, the agent selects a combination of orders that optimizes its



(a) Learning progress



(b) Q-Values evolution

Figure 2: Learning results.

operation and profits. Instead, if the agent decides to wait there might be a better combination of orders appearing in the order-book in 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 it encounters.

CASE STUDY

We apply the proposed methodology for the microgrid using the following parameters: $S^{B,max} = 10kWh$, $P^{CH,max} = P^{DIS,max} = 5kW$, $\eta = 100\%$, $X = \{1,2,3\}$, $\Delta t = 5$ min, $\gamma = 1$, and $\alpha = 0.0005$. The consumption and solar production profiles are considered as in [8]. The real-time market simulator is assumed to model the process of the orders arrival. The appearance of new orders is modeled by a Poisson distribution. The price and the quantity of each order are assumed to follow a Gaussian distribution. For the sake of simplicity we consider three available products. The trading horizon is assumed to be two hours and the agent can take an action every five minutes. The learned policy is tested against the naive policy. Preliminary results indicate that the trading agent is able to converge to a policy. As shown in Figure 2a, under the learned policy the agent receives a cumulative reward 5% higher than that of the naive policy. After 600 episodes the exploration is terminated and convergence is achieved. The evolution of the Q-values for each action over the trading horizon is presented in Figure 2b. The cumulative rewards of the episode are successfully back-propagated to the first trading step. It can be observed that the value of idling/waiting at the first time step results in higher total profits. It is important to note that both actions have the same value in several time steps, because both of them lead to the same (zero) reward. Finally, it can be observed that most of the trading happens early in the episode. This might occur either due to the lack of liquidity in the market or because of the operational constraints of the microgrid.

CONCLUSION

We investigate the participation of a microgrid in a real-time market. The sequential decision making process is formulated as an MDP and solved with RL, and Deep-Q Networks in particular. We propose a novel approach with a small set of discrete high level actions to handle the huge and changing nature of the action space, with the goal of identifying the opportunity cost faced by the trading agent. Our methodology is applied to a test case and preliminary results are promising. The RL agent is indeed able to converge to a better policy than the naive policy in terms of cumulative rewards.

REFERENCES

[1] S. J. Chatzivasiliadis, N. D. Hatziargyriou, and A. L. Dimeas, "Development of an agent based intelligent control system for microgrids," in *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, 2008, pp. 1–6.

[2] D. Ilic, P. Silva, S. Karnouskos, and M. Griesemer, "An energy market for trading electricity in smart grid neighbourhoods," pp. 1–6, Jun. 2012.

[3] H. Lund, A. N. Andersen, P. Østergaard, B. Mathiesen, and D. Connolly, "From electricity smart grids to smart energy systems – a market operation based approach and understanding," vol. 42, 96–102, Jun. 2012.

[4] T. Logenthiran, D. Srinivasan, and A. M. Khambadkone, "Multi-agent system for energy resource scheduling of integrated microgrids in a distributed system," *Electric Power Systems Research*, vol. 81, no. 1, pp. 138–148, 2011.

[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.

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

[8] V. François-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management," in *European Workshop on Reinforcement Learning*, 2016.