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

Ioannis Boukas IOANNIS.BOUKAS@ULIEGE.BE
Department of Electrical Engineering and Computer Science
University of Liege
Liege, Belgium

Damien Ernst DERNST@ULIEGE.BE
Department of Electrical Engineering and Computer Science
University of Liege
Liege, Belgium

Anthony Papavasiliou ANTHONY.PAPAVASILIOU@UCLOUVAIN.BE
Center for Operations Research and Econometrics (CORE)
Universite Catholique de Louvain
Louvain la Neuve, Belgium

Bertrand Cornélusse BERTRAND.CORNELUSSE@ULIEGE.BE
Department of Electrical Engineering and Computer Science
University of Liege
Liege, Belgium

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".

Keywords: Electricity markets, Storage devices, Intra-day market, Deep-Q Networks

1. Introduction

The efficient integration of renewable energy resources (RES) in future power systems as directed by the recent worldwide energy policy directive (Commission (2017)) 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 (Papalexopoulos et al. (2016)). 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 (Nasrolahpour et al. (2016)). 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.

In this paper, we present a summary of published work in real-time bidding strategies for the case of a storage device, presented in Boukas et al. (2018b). The sequential decision making prob-

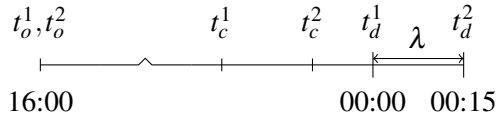


Figure 1: Trading time-line for products Q-1 and Q-2

lem 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 Mnih et al. (2015). The resulting optimal policy is evaluated using real data from the German ID market (EPEXSPOT (2017)).

2. Continuous Intra-Day market design

The CID market is a continuous process similar to the stock exchange market as presented in Ilic et al. (2012). 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 (Scharff and Amelin (2016)). 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 $i \in N_x \subseteq \mathbb{N}$ corresponds to the index of each order posted in order book O^x for product x (Table 1). The trading process for time-slot x opens at t_o^x 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 possibility to place new orders or adjust existing orders.

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 1 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.

3. 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 Boukas et al. (2018a).

Table 1: 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	

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 $a_t \in A = \{0,1\}^{|N_x|\times|X|}$, where X is the set of available products and $N_x \subseteq \mathbb{N}$ is the number of unmatched orders for each product x. 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).

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

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

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 (3).

$$r_t = \sum_{x=1,i=1}^{X,N_x} a_{t,x,i} v_{t,x,i} p_{t,x,i}. \tag{3}$$

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 (4) 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) Sutton and Barto (1998). The discount factor $\gamma \in [0,1]$ is used to adjust the strategy of the agent to be myopic or not.

$$G_t = \sum_{k=0}^{T-t-1} \gamma^t \cdot r_{t+k+1} \tag{4}$$

4. Solution technique

The state-action value function Q following a stationary policy μ is defined by Sutton and Barto (1998) as :

$$Q(s,a) = \mathbb{E}_\mu\left[G_t | s_t = s, a_t = a\right] \tag{5}$$

The optimal solution to the problem defined in equations (1)-(4) corresponds to the identification of the time-variant policy that maximizes the expected returns over the trading horizon T. The optimal policy $\pi^* = [\mu_0^*, \mu_1^*, \mu_2^*, ..., \mu_T^*]$ is approximated with the stationary policy μ given by solving the equations (6)-(7).

$$Q^*(s,a) = \max_{\mu} \mathbb{E}_{\mu}\left[\sum_{k=0}^{T-t-1} \gamma^t \cdot r_{t+k+1} | s_t = s, a_t = a\right] \tag{6}$$

$$\mu^* = \arg\max_{\mu} Q(s,a) \tag{7}$$

The agent is able to learn the state-action value function Q by approximating it using a Deep Q-Network Mnih et al. (2015). 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 of the Q-Network ($Q(s_t, a; \theta)$) are updated using mini-batch samples of quadruples $\left\{(s_t, a_t, r_t, s_{t+1})_j\right\}$ obtained by simulated experience. As proposed in Mnih et al. (2015), at each iteration k, Q-values obtained from taking gradient steps towards the minimization of the temporal difference error δ_j, as shown in equations (8) and (9). The goal is the back-propagation of total rewards early stages in the decision process. Parameters $\alpha, \gamma \in (0,1]$ are selected such that the convergence process is enhanced.

$$\delta_j = r_t + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta_k) - Q(s_t, a; \theta_k), \forall j \in J \tag{8}$$

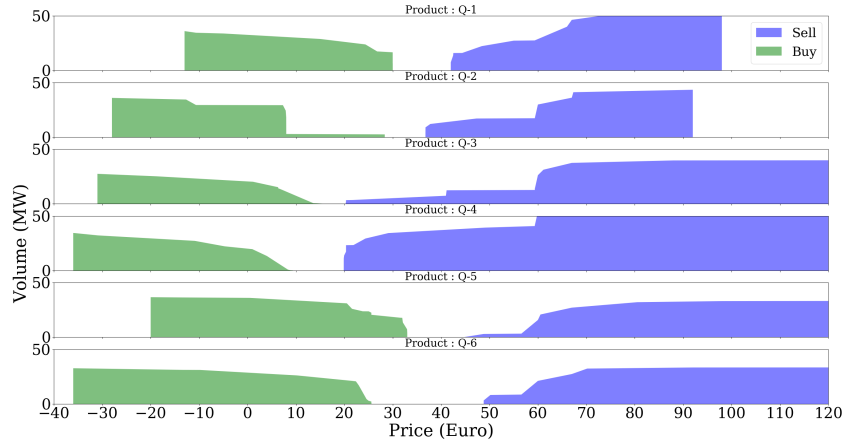
$$\theta_{k+1} = \theta_k - \alpha \sum_j \nabla_{\theta_k} Q(s_t, a; \theta_k)\, \delta_j \tag{9}$$

5. State-space representation

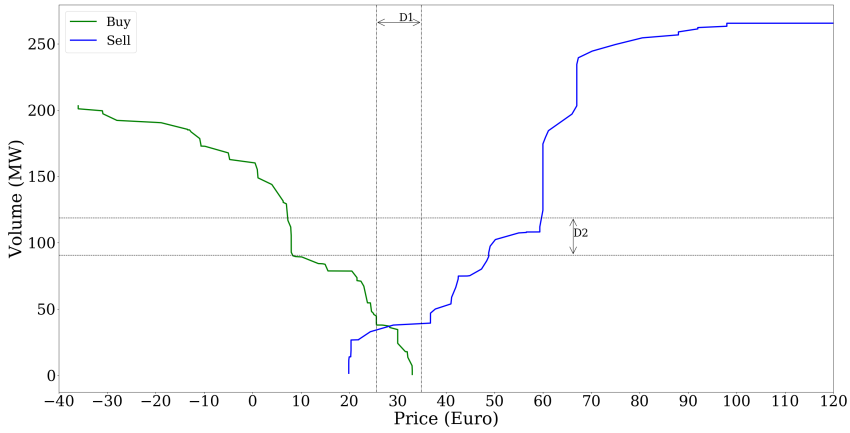
The high-dimensional continuous external state $s_t^E \in S^E = \{v, p\}^{|N_x| \times |X|}$ defined in section 3 is used to describe the state of the CID market. Owing to the variable (non-constant) amount of orders $\mid N_x \mid$ in each order book O^x for product $x \in X$, the state-space S^E does not have a constant size. In order to approximate the Q-function using a NN as described in section 4 it is necessary to find an approximate representation of the external state with constant size.

Owing to the nature of a storage device it is 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 available 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 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. 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 $s_t'^E \in S'^E = \{D1, D2, .., D10\}$ is used to categorize the observed order book based on its profit potential. The state of the trading agent is redefined as $s_t \in S = \left\{s_t^I, s_t'^E\right\}$.



(a)



(b)

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

6. 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 $\mid N_x \mid$ contained in the order book is not constant throughout the trading horizon. Thus, the size of the action space $A = \{0,1\}^{|N_x| \times |X|}$ is not constant. However, in order to ensure the tractability of the problem, a small and discrete action space is necessary Sutton and Barto (1998). 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 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 2. The objective of this strategy formulated in equation (10) is the maximization of the revenues arising

Table 2: Optimizing based on current knowledge.

$$\max_{a_{i,x}} \sum_{x=0}^{X} \sum_{i=0}^{N_x} a_{i,x} v_{i,x} p_{i,x} \tag{10}$$

$$\text{s.t.} \sum_{i=0}^{N_x} a_{i,x} v_{i,x} + Y_x^{ID} + p_x^{DIS} = p_x^{CH} \qquad \forall x \in X \tag{11}$$

$$s_{x+1}^{B} = s_x^{B} + \eta\, p_x^{CH} - \frac{p_x^{DIS}}{\eta} \qquad \forall x \in X \tag{12}$$

$$S^{B,min} \leq s_x^{B} \leq S^{B,max} \qquad \forall x \in X \tag{13}$$

$$0 \leq p_x^{CH} \leq k_x P^{CH,max} \qquad \forall x \in X \tag{14}$$

$$0 \leq p_x^{DIS} \leq (1 - k_x)\, P^{DIS,max} \qquad \forall x \in X \tag{15}$$

$$k_x \in \{0,1\} \qquad \forall x \in X \tag{16}$$

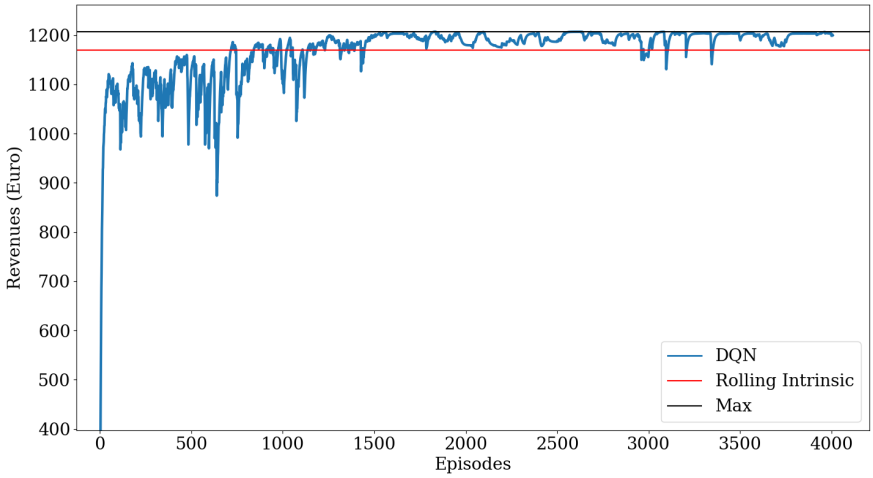
$$a_{i,x} \in \{0,1\} \qquad \forall i,x \in N_x \times X \tag{17}$$

from trading, subject to the operational constraints of the storage device. In equation (11) the energy purchased and sold ($\sum_{i=0}^{N_x} a_{i,x} v_{i,x}$), 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 (12), 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 (13) to (15). 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 2, 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 (7). 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 Lohndorf and Wozabal (2015). 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 (Boukas et al. (2018b)).

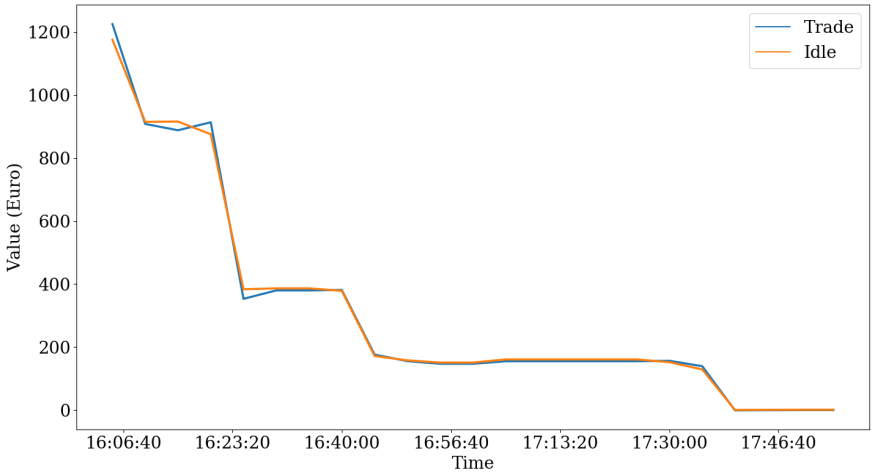
7. Case study

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 Boukas et al. (2018a), 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"Lohndorf and Wozabal (2015). According to this policy, we apply the "Optimizing based on current knowledge" at every time step of the horizon.



(a)



(b)

Figure 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 Boukas et al. (2018a). 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 (5). 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.

8. 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.

References

Ioannis Boukas, Damien Ernst, and Bertrand Cornélusse. Real-time bidding strategies from micro-grids using reinforcement learning. In *CIRED Workshop 2018 (To appear)*, pages 1–4, June 2018a.

Ioannis Boukas, Damien Ernst, Anthony Papavasiliou, and Bertrand Cornélusse. Intra-day bidding strategies for storage devices using deep reinforcement learning. In *International Conference on the European Energy Market, Łódź 27-29 June 2018*, page 6, 2018b.

European Commission. 2030 energy strategy, 2017. URL `https://ec.europa.eu/energy/en/topics/energy-strategy-and-energy-union/2030-energy-strategy`.

EPEXSPOT. Market data intraday continuous, 2017. URL `http://www.epexspot.com/en/market-data/intradaycontinuous`.

D. Ilic, P. G. Da 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)*, pages 1–6, June 2012. doi: 10.1109/DEST.2012.6227918.

N Lohndorf and David Wozabal. Optimal gas storage valuation and futures trading under a high-dimensional price process. Technical report, Technical report, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran,

Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

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)*, pages 1–7, June 2016. doi: 10.1109/PSCC.2016.7541016.

Alex Papalexopoulos, Rod Frowd, Chuck Hansen, Eamonn Lannoye, and Aidan Tuohy. Impact of the transmission grid on the operational system flexibility. *Power Systems Computation Conference (PSCC)*, 2016.

Richard Scharff and Mikael Amelin. Trading behaviour on the continuous intraday market elbas. *Energy Policy*, 88:544 – 557, 2016. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2015.10.045. URL http://www.sciencedirect.com/science/article/pii/S0301421515301713.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.