The Impact of Reward Shaping in Reinforcement Learning for Agent-based Microgrid Control

Valentin Père, Fabien Baillon, Mathieu Milhé, Jean-Louis Dirion

▶ **To cite this version:**

Valentin Père, Fabien Baillon, Mathieu Milhé, Jean-Louis Dirion. The Impact of Reward Shaping in Reinforcement Learning for Agent-based Microgrid Control. ESCAPE 32 - European Symposium on Computer Aided Porcess Engineering, Jun 2022, Toulouse, France. pp.1459-1464, 10.1016/B978-0-323-95879-0.50244-7 . hal-03754056

HAL Id: hal-03754056

https://imt-mines-albi.hal.science/hal-03754056v1

Submitted on 27 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Impact of Reward Shaping in Reinforcement Learning for Agent-based Microgrid Control

Valentin Père[a], Fabien Baillon[a], Mathieu Milhe[a] and Jean-Louis Dirion[a]

[a]*Université de Toulouse, IMT Mines Albi, UMR CNRS 5302, Centre RAPSODEE, Campus Jarlard, F-81013 Albi Cedex 09, France*
valentin.pere@mines-albi.fr

Abstract

In order to reduce CO2 emissions, electricity networks must increasingly integrate renewable energies. Microgrids are distributed electrical networks with their own generation and load, often supported by an electrical storage system. It can be connected to the external electrical network or isolated. Since electricity consumption, price and renewable production are stochastic phenomena, the control of microgrids must adapt to uncertainties. Data-driven models and in particular reinforcement learning (RL) have become efficient algorithms in high-level microgrid control. RL are agent-based algorithms, which interact with their environment and learn with a numerical reward signal. A certain behavior can implicitly be expected when the reward system is formulated. For example, a reward system that encourages the agent to interact as little as possible with the external network will explicitly increase the autonomy of the microgrid. Implicitly, it can be expected to schedule the battery to maximize the ratio of renewable energy used to the amount producible. Q-learning algorithm has been used due to its performance in discrete action space, which simplified the benchmark complexity. An agent is trained with different reward functions commonly found in the literature related to data-driven microgrid control algorithms. The agent parameters do not vary from one case study to another. Indicators are set up to evaluate the agent behavior. They are based on implicit behavioral criteria in the definition of the reward system such as the ratio of renewable energy used, the amount of energy stored during peak hours, etc. This study enables to find a way to rationalize the choice of a reward system to control in a near-optimal way microgrid while meeting implicit secondary objectives. It could lead to a choice on weighting coefficient in a combination of reward functions.

Keywords: Microgrid, Reinforcement Learning, Control, Reward

1. Introduction

1.1. Electricity storage scheduling with reinforcement learning

To remain reliable despite uncertainties in renewable electricity production and consumption, microgrid control must be efficient. Bidram and Davoudi (2012) has distinguished 3 categories of control. The first 2 categories are frequency and voltage regulation with very low time scale. The third category is what is called high-level control in this study. It concerns power flow long-term planning with higher time granularity (minimum 15 minutes). This high level planning can be done with several methods (Abdelhedi et al. (2018)): rule-based, optimization-based or learning-based methods are efficient. However, with uncertainties of electricity consumption, price and renewable generation, complete physical model based approaches are inappropriate. Thus, optimization-based methods will have difficulties to achieve optimal planning without predicting the future values of the stochastic variables. Data-driven methods are proven to be efficient in this context.

Especially, reinforcement learning (RL) algorithms (Sutton and Barto (1995)) learn policies given an environment and objectives. A RL agent makes decisions in its environment using Markov decision processes, the environment responds and the agent receives a reward signal, indicating if the reached environment state is suitable or not in respect to the objective. With this signal, it learns to value states or actions taken in specific states and thus can build a control policy according to the defined reward system.

High-level control can focus on every microgrid unit, including consumption (demand side management), controllable production (unit commitment) and electricity storage scheduling. This study aims to provide a view of the impact of the choice of reward signal in storage scheduling with respect to implicit indicators.

1.2. Case Study

The system studied is a simulation of a microgrid composed by photovoltaïc (PV) panels for electricity production, a point of consumption, an electrochemical battery for short-term storage system and hydrogen storage for long-term electricity storage. This simulation has data-driven units (electricity consumption and PV production) and analytical models (storage). Both data and short-term storage characteristics are taken from François-Lavet et al. (2016). The microgrid has two operation modes: connected to the main grid and isolated.

Some simplifications were made: The maximum power of the battery is not taken into account, the power to be supplied is multiplied by its efficiency and it automatically balances the network provided that its energy capacity is high enough. The RL agent controls the hydrogen storage system. Its maximum power is 1.1kW, its electrolyser efficiency is 0.65 and its fuel cell efficiency is 0.5. No maximum storage capacity is considered. The data are two years of PV production data in Belgium and consumption data respectively. When the net demand for electricity cannot be supplied, the short-term storage is discharged and charged when there is a surplus of energy. The main goal here is to test different reward functions to observe their effects on in behaviors that are implicitly expected from the agent. These behaviors are tracked with indicators.

The RL algorithm used is Deep Q-learning (Mnih et al. (2013)). The agent receives continuous state values that are the electricity consumption, the PV electricity production and the short-term battery state of charge (SOC). All these values are normalized between 0 and 1. The actions that are available for the agent at each timestep are the operating mode of the hydrogen storage system. These actions are discrete, the first action available uses electricity to charge the long-term storage with electrolysis at maximum power if possible. The second action discharges the hydrogen storage with fuel cell at its maximum nominal power (if enough energy is stored). The agent can also choose to do neither. Thus, its action space is composed of three actions.

2. An introduction to Q-learning and deep Q-learning

2.1. Q-learning

The objective of a RL algorithm is to find policies (i.e. probability to take one action from a given state) that maximize the rewards received in an episode (i.e. a series of interaction within the environment). To decide between several actions, the agent values state and action pairs through the rewards following the choice. These pairs are called Q-values and denoted $Q(s,a)$, a stands for action and s for state. The idea behind building an efficient policy is to select the action that maximizes this value from the state in which the agent is located. However, in order to value these pairs, the agent has to explore states and values to sample rewards. A Q-learning agent uses a behavioral policy to sample actions and a learned policy to update the different pairs value. With this algorithm, only the immediate reward perceived by the agent after his action and the following

action (from the next state) that brings to the maximum Q-value are taking into account to update the learned policy.

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha \left[R_{t+1} + \gamma \underset{a}{max} Q(S_{t+1},a) - Q(S_t,A_t) \right] \quad (1)$$

This update rule for Q-value mapping is shown in Equation 1, with $\gamma \in [0;1]$ the discount factor to level out the extent to which future actions are considered in the estimation of a Q-value. Once the agent is trained, an estimate of every Q-values is stored in a table. The agent can then choose every action that maximizes immediate and future reward directly according to this table.

2.2. Deep Q-learning

With Q-learning, the main drawback is from the use of a table, which implicitly requires countable and therefore discrete spaces. Also, wide spaces lead to long computations before the agent is trained.

With Deep Q-learning, the table is replaced by a neural network (NN). In deep learning, the target of a NN has to be stationary. Here, as showed in Figure 1, the target value the NN must predict Q-values, the second part of equation 2 (which the same equation as 1 but factorized in an different way).

$$\begin{aligned} Q(S_t,A_t) &= (1-\alpha) Q(S_t,A_t) \\ &+ \alpha \times (R_t + \gamma \times max_a Q(S_{t+1},a)) \end{aligned} \quad (2)$$

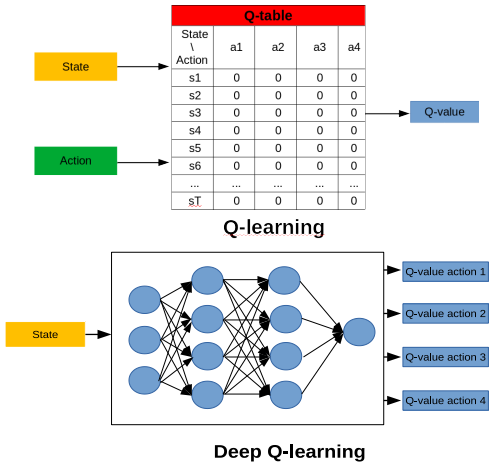


Figure 1: Visual comparison between Q-learning and deep Q-learning

With these Q-values, the problem is that a part of the target, $max_a Q(S_{t+1},a))$ (with S_{t+1} the state at time $t+1$ and a the action that can be taken from state S_{t+1}) depends on the NN that is updated. To solve this problem, an other NN with the same parameters is used to estimate $Q(S_{t+1},A_{t+1})$. These parameters are frozen and actualized slowly. Another reinforcement learning problem is the influence of the output on the next input. This problem is solved with a *ReplayMemory* that injects randomly a sample of state, action, reward, next state and next action into the NN. In this way, the impact of previous prediction is negligible for incoming output and the idea of transition in the system is kept.

3. Analysis on the impact of reward shaping on implicit indicator

Each and every microgrid system is built with a particular objective. Isolated microgrids follow the objective to be autonomous for example. Some have diesel generator as support electricity production (Kofinas et al. (2018)), and the objective can be to supply users autonomously without this support. Grid-connected microgrids can minimize operation cost or emissions for exemple. Even though objectives and therefore reward systems can be different, whoever made them implicitly expects awaited behaviors. In the case study microgrid presented in 1.2, whether the reward system depends on autonomy or operating cost, the system is expected to buy less energy from the main grid. Of course, in an isolated grid that aims to be autonomous, the equivalent of buying electricity from the main grid to compensate the grid imbalance is poor quality electricity or even blackout periods. Thus, this need of extra energy can be penalized the same way for these very

different systems. Also, systems aims to be energy efficient. Electricity production can not much exceed consumption, or else equilibrium is lost. However, it is important to get as much electricity as possible from the PV panels, whatever the objective. When the system does not need electricity (consumption is supplied, battery is fully loaded and long-term storage is supplied at maximum power), the remain production is extra-energy and lost. Another implicit objective in every case is to minimize excess of electricity.

3.1. Methodology

Every hour, the agent makes a choice in its microgrid. It can charge or discharge the long-term storage, or do nothing. Its action is perceived as an extra electricity demand or production, if the long term storage is charged or discharged respectively. In order to underline the impact of reward shaping, deep Q-learning agents are trained with different reward systems and indicators are identified. First, in an objective of operating cost minimization, different microgrid configurations are tested. The microgrid buys electricity from the main grid at a price of 2€/kWh. As the time granularity of the simulation is one hour, the agent perceives a -2 reward in this case. It occurs when the microgrid net demand (electricity demand action minus electricity production) including additionnal production or consumption from agent action (hydrogen storage charge or discharge perceived as consumption or production) is positive and superior to the electrochemical battery capacity. On the opposite, when this net demand is negative and the extra energy exceeds electrochemical battery capacity, the remaining produced energy is wasted.

Whatever which mode (isolated or grid-connected) is selected, this system is adopted, and the negative reward for buying electricity is applied to penalize isolated microgrid instability (in this case, no electricty is bought, but the simulation is similar). In the grid-connected mode, three configurations are tested, with the objective of reducing operating cost:

Case 1 The system can not sell energy to the main grid. With the operating cost reward system, the only perceived rewards are negative and correspond to the electricity purchased from the main grid.

Case 2 The system can sell energy to the main grid, without power constraint at the common coupling point (the electricity exchange point between the micorgrid and the main grid). In this configuration, agent can receive rewards from selling extra-energy to the main grid. In this specific case, it is impossible to waste energy and PV panels produce 100% of what they should. The electricity sold is four times cheaper than the bought electricity.

Case 3 The system can sell energy to the main grid, with a power constraint at the common coupling point. The agent sells its extra energy and receives positive rewards in this case. However, this amount of sold energy is limited by a power constraint and so is the reward.

In isolated microgrid, as the main objective is to be autonomous, a negative reward for system instability is applied. It is exactly the same system configuration as the first case listed above.

What if the reward system is explicitly giving the agent penalities for wasting excess energy ? To analyse this situation, the first and last configuration are adopted with an extra negative reward equal to the wasted electricity. Again, it has no sense to use this reward system on the second case because no energy waste is allowed. At last, the energy excess negative reward will be applied without the energy bought negative reward to know how the system behave with only this objective.

A comparison of the quantity of excess energy and bought energy from the main grid will be done in the next section.

3.2. Results and analysis

Agents were trained in every case, with three reward systems for each except the second case in which extra energy does not exist. It means seven agent were trained. The convergence of obtained rewards converged at approximately 30 episodes for every agent. Once the training is over, data of their control behavior in the last training episode (1 year) are collected to analyse results.

3.2.1. Excess energy and purchased energy

Surprisingly, there is very little variation in excess energy between the reward systems for each case.
The same pattern is observed with some small variations. Obviously, this amount is smaller when only excess energy governs the agent's reward system. It is larger when reward system only considers operating cost. This has many common features with the purchased energy curves. As shown in Figure 2, more energy is brought from the main grid in winter. The agent tends to buy more energy when its reward system penalize it.

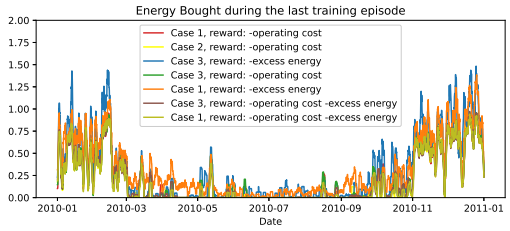
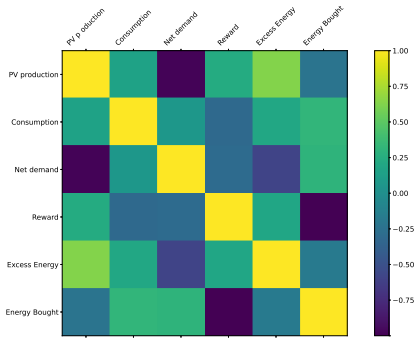


Figure 2: Smooth curve of hourly bought energy in the 3 reward systems

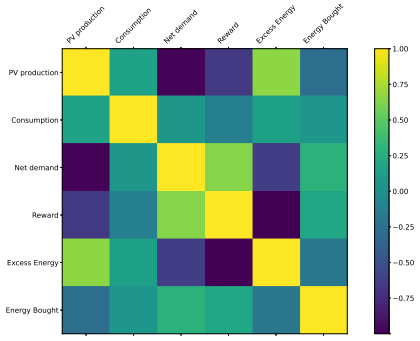
3.2.2. Correlation between variables

Correlation matrices between exogenous variable have been made. It gives important information on how the agent behaves.

Figure 3a shows that rewards are negatively correlated with net demand when operating cost defines reward system. Of course, purchased energy is negatively correlated with rewards. Net demand is the only parameter affected by the agent decision. To increase net demand, the agent has to discharge the long-term storage. In order to do that, the hydrogen storage can not be empty. The agent's game would therefore be to charge and discharge the hydrogen storage at the right time, so that it can discharge when the purchase of electricity is necessary, to alleviate the negative reward. On the contrary, Figure 3b shows that rewards are positively correlated with net demand when excess energy penalizes the agent. The agent's action must therefore charge the hydrogen storage when the PV production exceeds the consumption. He can buy energy without being penalized and emptying the storage (infinite in capacity) does not increase his reward. The graph of stored energy (Figure 4) is interesting, it underlines the fact that long-term



(a) Correlation matrix of exogenous variables in Case 1 with only operating cost considered in reward system



(b) Correlation matrix of exogenous variables in Case 1 with only excess energy considered in reward system

storage is used as short-term storage to to increase the rewards when buying and selling energy define the reward system. What was observed in the correlation matrix is confirmed in the agent behavior, giving penalities only for excess energy tends to make the agent store bigger hydrogen quantity when it is possible. When both excess energy and operating cost are considered, the agent also tends to store hydrogen the way expected in Case 3. Net demand is hightly correlated with PV production and very little correlated with consumption. This is because PV energy production has a range of values that can go very high when non-zero, compared to the consumption which is more constant and low.

When PV panels produce a lot of electricity (between July and September), it may be more attractive to store energy to avoid big excess energy penalities rather negative reward for energy bought. In Case 3, the stored energy can be selled to the main grid in winter. This explains why the agent prefers to store energy in summer in Case 3 rather than in Case 1 (where energy can not be sold) with multi-objective reward system.

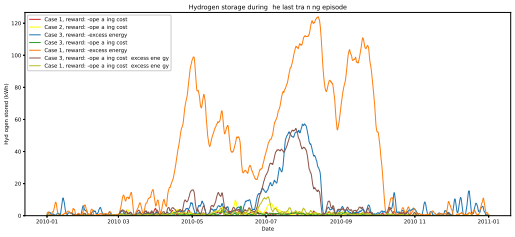


Figure 4: Energy stored in long term storage during the last training episode in every cases

4. Conclusion

The reward functions defines the behavior of a RL based control algorithm in microgrid. Making explicit certain implicitly expected behaviors changed totally the decisions. The multi-objecitive reward system seems more interesting for the microgrid control in this case study. However, the way to modelize electrochemical battery and hydrogen storage was too simplistic. The sizing of the storage systems and PV panels was arbitrary, as was the price of energy both when purchased and when sold to the main grid. In these simulation conditions, the sensibility study of the defined reward system can not be unbiaised. However, it highlighted a behavioral difference of the agent on how to store hydrogen. With a robust sizing of the system and a good storage systems modeling, optimal weights for reward systems can be found throught sensibilization analysis on different criteria.

References

R. Abdelhedi, A. Lahyani, A. C. Ammari, A. Sari, P. Venet, Feb. 2018. Reinforcement learning-based power sharing between batteries and supercapacitors in electric vehicles. In: 2018 IEEE International Conference on Industrial Technology (ICIT). IEEE, Lyon, pp. 2072–2077.

A. Bidram, A. Davoudi, Dec. 2012. Hierarchical Structure of Microgrids Control System. IEEE Transactions on Smart Grid 3 (4), 1963–1976.

V. François-Lavet, D. Taralla, D. Ernst, R. Fonteneau, 2016. Deep Reinforcement Learning Solutions for Energy Microgrids Management. European Workshop on Reinforcement Learning (EWRL 2016), 7.

P. Kofinas, A. Dounis, G. Vouros, Jun. 2018. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. Applied Energy 219, 53–67.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Dec. 2013. Playing Atari with Deep Reinforcement Learning. arXiv:1312.5602 [cs].

R. S. Sutton, A. G. Barto, 1995. Reinforcement Learning: An Introduction, 352.