Learning-Based Optimization of the Exchange Price in a Heterogeneous Market

Guénolé Chérot, Roman Le Goff Latimier, Benjamin Cajna, H. Ben Ahmed

Learning-Based Optimization of the Exchange Price in a Heterogeneous Market

Guénolé CHÉROT, Roman LE GOFF LATIMIER, Benjamin CAJNA, Hamid BEN AHMED
SATIE, UMR CNRS 8029, UniR – ENS Rennes
Rennes, France
guenole.cherot@ens-rennes.fr

***Abstract*—In coming years, energy systems are likely to be organized as a heterogeneous system, where independent energy communities coexist with a main market. The price of electricity exchanges between communities and the main system should then reflect not only production costs, but also network constraints, potential distribution congestion and incentives towards local flexibilities. This price would necessarily be local, and setting it optimaly would require an all-knowing operator. The present contribution aims to investigate the potential of reinforcement learning to predict this exchange price. A minimalist case study is introduced to improve the interpretability and generalizability of the results obtained. In particular, learning speeds will be studied in order to discuss the volume of data required to guarantee a given level of performance. The transfer of trained algorithms from one case study to another will also be discussed.**

***Index Terms*—Reinforcement learning, price forecast, energy community, heterogeneous systems, congestions, flexibility**

I. INTRODUCTION

As a result of the energy transition, power flows through electrical systems are likely to increase significantly [1]. Indeed, to achieve the targets for reducing CO2 emissions, the electrification of entire industrial sectors is a major contribution, as long as the electricity used is produced by low-carbon sources. The synergy between electric vehicles and renewable energies is particularly relevant for instance [2]. This development of distributed generation and storage capacities, made up of a large gathering of low-power resources, is profoundly reshaping the structure and organization of the power system: it is fostering the emergence of new players – the prosumers – and new forms of electricity exchange [3]. However, the historically vertical and unidirectional organization of power systems, structured around a wholesale electricity market, is not suited to easily integrating distributed energy resources [4]. In order to preserve the benefits and robustness of interconnection, it is key to avoid the proliferation of communities on the edge of the main network aiming for self sufficiency. We therefore need to come up with another model that would allow them to be willingly integrated, while making the most of the flexibility and reactivity of small scale management. This has led to the emergence of new operating concepts such as local energy markets [5], energy communities [6] and peer-to-peer (P2P) electricity markets [7]. At European Union level, the legislative work leading up to the Clean Energy Package [8] aimed to promote decentralization of the European electricity system by giving consumers an active role and empowering them. It was with this package of measures that the notions of renewable energy communities and citizen energy communities were introduced. In this context, it seems likely that power grids will become heterogeneous systems, where energy communities coexist with the conventional centralized system [9].

In the literature, several works have addressed this topic. Moret and Pinson [10] have formulated a community market where prosumers are allowed to share their energy at the community level, or trade with the outside world via a third-party supervisor responsible for interfacing with the market and the system operator (SO). In [11], Morstyn and McCulloch propose a P2P market platform enabling prosumers to exchange energy with each other and with the wholesale market. Similarly, in [12] the authors introduce an exchange platform that provides an interface between prosumer communities and wholesale markets, and coordinates the community's operational decisions on supply and demand.

However, although these studies propose different organizations for the operation of a power system where energy communities and the conventional system interact, they do not take into account the physical constraints of the power grid [13]. Yet, managing such a heterogeneous system requires the development of new management rules, particularly with regard to the interaction between a centralized market, an energy community and the system operator - transmission or distribution [14], [15]. Indeed, the exchange price between the community and the main grid must take into account not only generation costs, but also transmission constraints such as congestion or voltage limits. On the transmission network level, network management includes reconfigurations of the interconnection topology [16]. On the other hand, at distribution network level, actions include reconfigurations [17], but also the management of local flexible agents [18]. Within such networks, the constraints taken into account relate to both voltage levels [19] and line congestion [20]. In addition, within a distribution network, control is likely to operate as an incentive signal sent to agents rather than as a direct action. Indeed, assets are there owned by numerous private individuals, rather than by the SO or by a small number of producers. Setting a price that induces agents to take the required action is therefore an even thornier issue than actually performing that action - while assuming we've been able to determine it beforehand. Consequently, further exploration of

the relationship between community energy exchanges and the wholesale market is an important topic. Recently, Faria [21] proposed integrating SO into a P2P market either by penalizing congestion-causing exchanges, or by incentivizing players via a flexibility market.

This problem can naturally be solved by a constrained optimization approach, potentially distributed between the network operator and the energy community [22]. However, it would then be necessary to provide for exchanges of information: either the objective functions of the community players, or at least an iterative exchange of the dual variables of the optimization [2]. The operational deployment of such an approach would therefore face regulatory and technical difficulties [23]. It would seem preferable for the system operator to be able to announce to the community the energy price for each time step of a time horizon, in order to respect its physical constraints [24].

The structure of this problem therefore calls for the use of machine learning methods such as neural networks or reinforcement learning. A rich and dynamic literature is currently devoted to these issues: solving the optimal power flow (OPF) [25]; forecasting energy market prices [26]; regulating network congestion [27]. Taking into account the reaction of agents – in addition to evolutions in the system as such – remains a major challenge, however.

The aim of the present contribution is therefore to investigate learning methods for managing flexibilities indirectly within heterogeneous systems. The focus here is on controlling line congestion. Several questions remain open before any operational deployment. For instance: how to guarantee performance, how much data to train [28], how good is the transfer from one community to another [29]? To build on these questions, comparisons between algorirhms on commun ground are necessary. For this reason, we present here a case study designed to be both minimalist and representative. These two qualities are necessary to allow the interpretability of the results and an application as straightforward as possible for the algorithms investigated.

The rest of this article is therefore organized as follows. The case study will be presented in section II. Section III will present on the one hand the modeling of the energy community and how a learning problem can be derived from it. The results obtained will be discussed in section IV.

II. Theoretical and generic case study

Figure 1 presents the case study that is introduce for the purpose of this contribution. An energy community Ψ is connected to the main grid at a single point. It hosts distributed renewable generation, non-flexible consumption but also flexible consumption. Its interconnection with the main grid is a line whose maximum power constraint $\lceil P_l \rceil$ would be exceeded if left unmanaged. The community's exchange price with the external market must therefore be adjusted to ensure that the line's capacity is respected.

Players behaviours within the community are randomly selected from a database. This database contains time series,

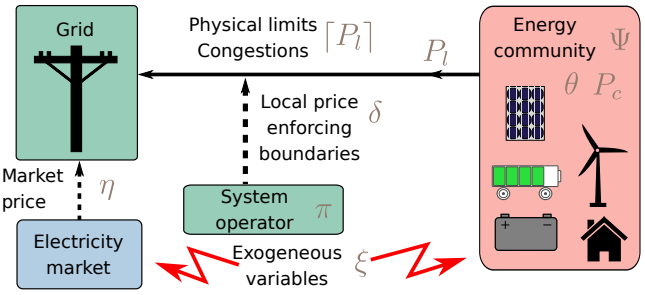

Fig. 1. Case study configuration: an energy community Ψ is connected to the main grid through a line whose maximum capacity $\lceil P_l \rceil$ will be activated. The system operator can reduce exchanges in the line P_l by increasing its cost of use δ. This fosters intra-community exchanges P_c. Prices of the market η and of community players θ vary stochastically, and can be explained by exogenous hidden variables ξ and time t.

with minute time steps, of power exchanged by households, electric vehicles (EVs) and photovoltaic (PV) panels. The number of each type of actor is specified when the community is created. A price θ_n – fixed for the whole simulation and following a normal distribution $\mathcal{N}(\mu, \sigma)$ – is then associated with each actor n. The number of actors, and the values of μ and σ are given in table I.

The community is thus characterized by a set of N_a agents, whose powers evolve over time and whose buying or selling prices $\theta = (\theta_i)_{i \in [\![1; N_a]\!]}$ are fixed. The power variations of each agent are driven by exogenous phenomena ξ – the time of day, the season, the temperature, etc. – which are not specified in the model. This formulation offers two difficulties of interest. On the one hand, the powers of the players in the community are private, so they have to be estimated. On the other hand, the price δ imposed by the SO modifies these powers, making the forecasting task more complex.

In this case study, the community is assumed to be connected to the network at a single point. However, real communities can be made up of geographically distant agents and thus have multiple points of connection with the network [30]. In the context of this contribution, we will consider that such scattered communities can be treated as a sum of communities with a single point of connection, and can therefore be handled separately.

As part of an open and reproducible science approach, all the data used are publicly accessible: [31] for household consumption and wind generation, [32] for electric vehicle consumption and EPEX [1] for energy market prices. The learning algorithms are taken from the implementation proposed by

[1] https://ewoken.github.io/epex-spot-data/

	μ	σ	Number of players
Households	0.25	0.10	40
EV	0.15	0.05	10
PV	0.08	0.02	30

TABLE I
Values selected to create the test case.

Stable-Baselines3 [33] with the default set of parameters. The source codes are open and accessible on GitLab [2].

The simulations were conducted by separating the 365 days in the database into two groups: 90 % of the days were used for training, 10 % for testing. The training data can be viewed several times by the algorithm during the learning process.

III. Modelling as a learning problem

A. Assessing the exchanged power

Figure 2 shows the community merit order at a specific time of the simulation and how it is affected by the evolution of the line usage price. Consumer offers (in orange) are ranked in ascending order of purchase price - the more a consumer is willing to pay, the more likely he is to see his demand satisfied. Production bids are ranked in decreasing order - the cheapest producers will be selected first. Three orders of merit are represented: ⓐ islanded community ⓑ exchanging with the main market without capacity constraints ⓒ including the line fee cost imposed by the SO. Each merit order is associated with an clearing point (price/exchanged power pair) represented by a circle on the figure. In the case ⓐ, no power transits in the interconnection, and around 80 kW is exchanged within the community. In the case ⓑ, the infinite power network offers to buy or sell energy at a price of 0.25 €/kW: all consumers (resp. producers) wishing to buy (resp. sell) at a lower (resp. higher) price will not be selected in the merit order and will not exchange any power. In this scenario, approximately 100 kW will be exported from the community to the grid. In the case ⓒ, the SO charges δ for the use of the line. From the community's point of view, this cost reduces (resp. increases) the exchange price with consumers (resp. producers) on the external network. More power is exchanged in the community, and therefore less in the interconnection: congestion (represented by a rectangle in red) decreases. By further increasing the cost, congestion would decrease until it became non-existent. The power exchanged would be less than or equal to the line capacity, represented in green.

[2]https://gitlab.com/satie.sete/learningoptim_exchprice_heterogenmarket

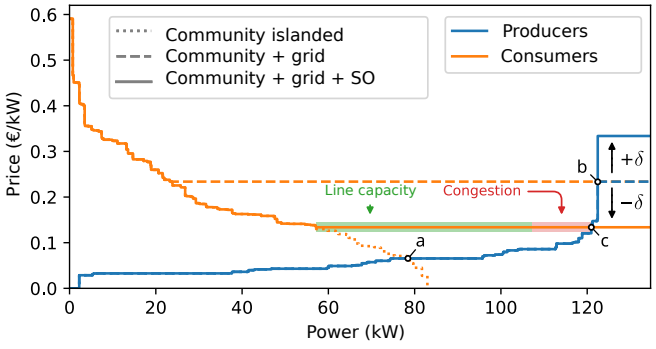

Fig. 2. Impact of the line utilisation fee δ on the merit order within the community. Merit orders of consummers and producers are displayed respectively in orange and in blue. They are shifted according to the combination of considered players: community alone, with grid and with SO. ⓐ, ⓑ and ⓒ represent the various clearing points for this differents configurations. Green and red areas represent respectively the line power limit $\lceil P_l \rceil$ and the power surplus $||\lceil P_l \rceil - |P_l||$ flowing through the line.

The SO's objective is to maintain network integrity while minimizing his impact on power exchanges. We observe that the power transiting the line decreases monotonically as the line fee increases. An all-knowing SO – knowing the preferences of each actor and therefore able to clear the merit order within the community – could therefore easily calculate the optimal cost δ_* for using the line.

In practice, two factors hamper the calculation of δ_*. Firstly, there is a latency between the measurement of network quantities and the sending of a new cost value. The proposed learning algorithms will therefore be compared with a "delayed" DSO, which will apply at time $t + \Delta t$ the optimal strategy $\delta_*^{[t]}$ calculated in t. On the other hand, players' preferences are often unknown. It is therefore impossible to calculate the merit order and the associated δ_*. Only the power transiting the interconnection and the market price of the energy are supposed to be known or measured. In this context, predicting the exchange cost is much more complex, so we'll use a reinforcement learning approach to determine the optimal strategy.

B. Learning setup

The reinforcement learning paradigm is designed to solve sequential decision-making problems, and is therefore ideally suited to this application. The learning scheme is given in figure 3. At each time t, the SO must choose the fee value for using the line $\delta^{[t+1]}$ based on a number of explanatory variables noted $O^{[t]}$ measured at a time step of t: the European market price $\eta^{[t]}$, the previous fee value $\delta^{[t]}$, the power transiting through the interconnection $P_l^{[t]}$, the total power exchanged in the community $P_c^{[t]}$ and time t. The observation is then normalized between -1 and 1 to ensure easier learning. The strategy, also called policy, is $\pi(O^{[t]}) = \delta^{[t]}$. To improve it, the reinforcement agent relies on a single scalar signal called reward R, whose sum must be maximized in expectation. The choice of this function is crucial, as the optimal π_* strategy follows directly from it.

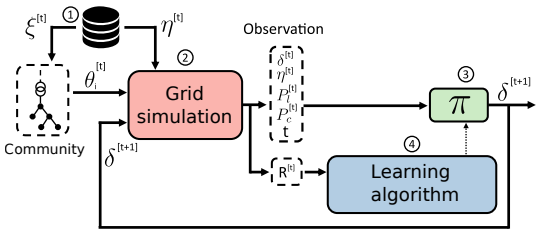

Fig. 3. Reinforcement learning of the fee value for using the grid interconnection. ① The exogenous variables at time t – noted $\xi^{[t]}$– from a database are transmitted to the energy community, which in turn gives the $\theta_i^{[t]}$ prices of each player. ② These prices $\theta_i^{[t]}$, together with the price of the network $\theta^{[t]}$ and the cost of using the line $\theta^{[t]}$ are used to compute the merit order and deduce the observation vector $O^{[t]}$. ③ The observation is transmitted to the SO so that it can deduce the next price $\delta^{[t+1]}$. ④ At the same time, a learning algorithm improves the controller performance with a reward signal that penalizes constraint violations.

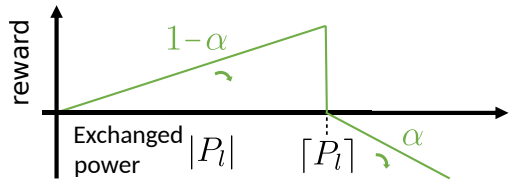

Fig. 4. Shape of the reward function. As α increases, the power exchanged is less rewarded and congestion more penalized.

Since the objective of the SO is to maintain network integrity while minimizing its impact on power exchanges, we have chosen the form illsutrated figure 4 and described by (1) where $\lceil \cdot \rceil$ is the operator $max(\cdot)$. As long as there is no congestion, the reward is proportional to the exchanges on the line. Congestion is penalized by a negative reward, proportional to the amplitude of the congestion. Both terms are weighted by $\alpha \in [0, 1]$, whose influence will be discussed in section IV-C.

$$R = \begin{cases} (1-\alpha) \cdot |P_l| & \text{if } |P_l| < \lceil P_l \rceil \\ -\alpha \cdot \left(|P_l| - \lceil P_l \rceil \right) & \text{otherwise} \end{cases} \quad (1)$$

The recent development of reinforcement learning has led to the emergence of numerous algorithms. We will evaluate two of the most successful. The PPO algorithm – *Proximal Policy Optimization* – [34] is an algorithm based on direct policy optimization. The main idea consists in clipping the gradient when improving the policy. This ensures that the new policy is close to the old one, thus stabilizing learning. The SAC algorithm – *Soft Actor Critic* – [35] is based on learning the state-action function Q_π, an estimator of policy performance. Unlike PPO, this algorithm stores past experience in a memory, making it more effective for problems requiring a limited number of interactions with the environment.

IV. Results and Discussion

A. Time series behaviour

The evolution of the variables of interest over time is shown in figure 5. Without SO (in violet) P_l exceeds the maximum authorized value $\lceil P_l \rceil$ between 9 h and 15 h. This results in a negative reward, which penalizes congestion. On the other hand, the optimal strategy (in green) always respects the constraint by increasing the cost of using the line. Finally, the SAC agent (in orange) respects the constraint 90 % of the time, even if there are a few overruns with a maximum amplitude of $2 \cdot \lceil P_l \rceil$. The cost is well predicted during congestion phases, and overestimated otherwise. This behavior is known from the SAC algorithm: the actions (here δ) are sampled from a Gaussian model initially centered in zero. Learning aims to modify this distribution, but extreme values are difficult to reach. This drawback could be improved by translating the action space so as to center the action around $\delta = 0$.

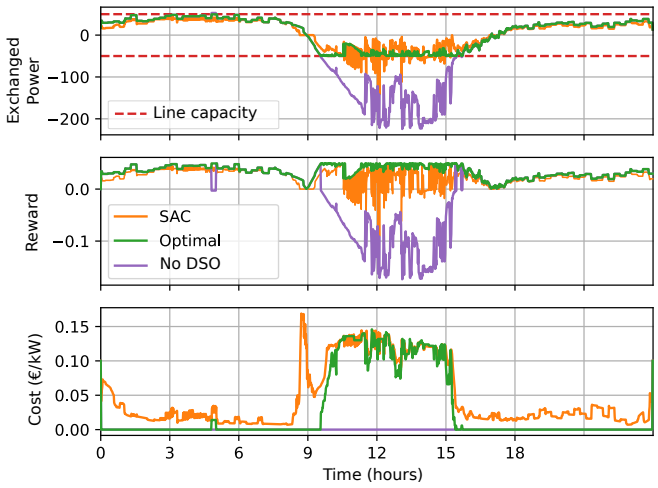

Fig. 5. Example of fluctuations of the exchanged power P_l, of the reward R and of the line fee δ along time according to three control agents: optimal all-knowing control, trained SAC agent and without any SO ($\delta = 0$).

B. Learning rates

Figure 6 shows the evolution of the average total reward (2) as a function of training time.

$$\overline{R_{tot}} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=0}^{T_{fin}} R_{\pi_i}^{[t]} \right) \quad (2)$$

where N is the number of algorithms trained, T_{fin} is the final time of a simulation and R_π is the reward obtained by the agent following the π policy. Since learning is stochastic, it is necessary to evaluate each algorithm several times - 100 training sessions per algorithm here. Five strategies are evaluated: on the one hand, SAC and PPO, whose performance evolves during training, and on the other hand, the optimal, optimal with delay and SO-free strategies (see section III-A), which constitute a reference.

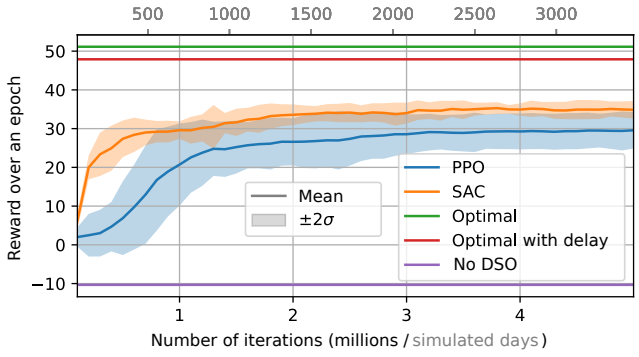

Fig. 6. Evolution of the reward mean value during training according to SAC and PPO algorithms. Three deterministic strategies are displayed as baselines: no control ($\delta = 0$), all knowing optimal control and non anticipative optimal control. The number of iterations is indicated in million (bottom of the figure) and in days (top of the figure). The computation of standard deviations σ among the set of trained agents allows to display in transparency intervals at $\pm 2\sigma$.

Figure 6 shows the amount of data required for training, and compares the convergence speeds of the algorithms. It does not give any direct indication of the algorithms' ability to meet the constraints, as the reward R is an aggregate metric (see section IV-C).

Let's analyze the performance of each strategy. One of the worst possible strategies is to do nothing $\delta = 0$, which leads to an average reward of -10, whereas the optimal and optimal with delay strategies obtain 50.5 and 49 respectively. Throughout the learning process, the SAC algorithm achieves an average reward strictly greater than PPO. Its asymptotic mean reward is also better by about $5\,\%$. SAC is also faster to converge: two million time steps are required, corresponding to 2 to 4 simulated years. These results are well known in the literature: SAC's memory enables it to learn using less data, but at a higher computational cost.

C. Impact of reward parameters

Figure 7 describes the influence of α on control performance, varying between 0 and 1 (see figure 4). Five training runs are performed for each algorithm and each value of α. The average performance is plotted. On the x-axis, the power exchanged must be maximized; on the y-axis, the frequency of congestion must be minimized. The optimum operating point is therefore in the bottom right-hand corner.

The optimal policy (x) dominates all other solutions and leads to no congestion. The optimal policy with delay (+) leads to the exchange of more power – $1\,$kW on average – at the expense of the apparition of some congestions. Finally, the policy without SO leads to the highest exchanges, as well as to a capacity violation in $48\,\%$ of the time.

Let's now take a closer look at the learning algorithms and their sensitivity to the parameter α. For $\alpha = 1$, the power exchanged is not rewarded, so the maximum reward is $\lceil R \rceil = 0$. In this case, the optimal strategy is to impose $\delta = \lceil \delta \rceil$. This is what we see in the bottom left-hand corner of figure 7: the average power exchanged is almost zero, and no congestion is observed.

For $\alpha = 0$, congestion is not penalized, although not rewarded. As a consequence the optimal strategy is not to impose $\delta = 0$, as this would frequently lead to congestion: $P_l > \lceil P_l \rceil$. The maximum reward is reached when $P_l = \lceil P_l \rceil$. The proposed reward signal – as defined in (1) – leads to trade power maximization while minimizing constraints for any α different from 0. For $\alpha = 0.16$, the PPO algorithm exchanges on average $29\,$kW which leads to congestion in $14\,\%$ of the time. For $\alpha = 0.32$, power increases slightly ($32\,$kW) and congestions are more numerous ($15\,\%$ of the time). Generally speaking, for both SAC and PPO, an increase in α goes hand in hand with an increase in exchanged power and congestion. The choice of its value is therefore essential, as it allows the SO to set the learning parameters according to its risk aversion. Finally, we note that SAC dominates some of the solutions given by PPO. This confirms the superiority of SAC for this application case.

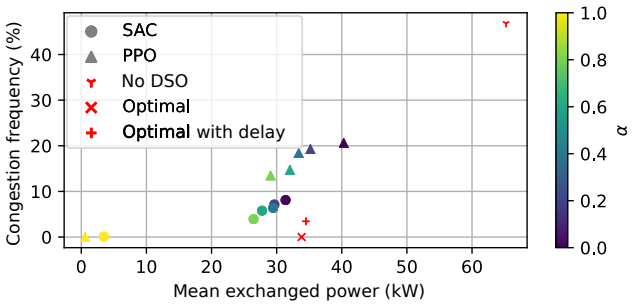

Fig. 7. Pareto front obtained from varying α between 0 and 1. The color of the dots indicates the value of the α parameters. Each circle (resp. triangle) represents the average of five training runs of the SAC (resp. PPO) algorithm. The performance of three untrained strategies (optimal, optimal with delay and unmanaged) is given for comparison.

Note that the form of the reward function is not discussed here, only the α parameter evolves. We followed the guidelines below to create the reward function. i) In the congestion-free zone, the reward must increase with power. If it were decreasing or constant, the algorithm would be rewarded for limiting the flow in the line, which goes against the role of the SO. ii) In the congested zone, the reward must decrease with power, as large-scale congestion is the most dangerous for the network. This being the case, instead of a linear relationship between R and P_l, we could have chosen a polynomial, exponential or other relationship. This would have modified the algorithm's performance. However it is likely that the general conclusions of this article would not be altered.

D. Transfer learning

As seen in figure 6, learning requires a large number of interactions with the environment. During this phase, the agent makes sub-optimal decisions that could have serious consequences for the network. Sharing a pre-trained algorithm could greatly accelerate learning time and reduce the risks associated with exploration.

The literature on transfer learning [29] offers dozens of approaches, whose performance varies widely depending on the problem at hand. Some methods are specific to reinforcement learning [36], [37]. They mainly rely on the existence of one or more expert policies π_X – resulting from prior training – to train the transfer policy π_T. The performance of π_T and π_0 – policy without transfer – is characterized by four metrics: reward gain at initialization, time gain to reach a certain threshold, asymptotic reward gain and regret expressed as (3). Since π_* is the optimal policy, it is always positive.

$$\mathcal{R}(T, \pi_T, \pi_*) = \sum_{t=0}^{T} R_{\pi_*}^{[t]} - R_{\pi_T}^{[t]} \tag{3}$$

The most intuitive method, called policy transfer, consists in using an expert policy π_X to initialize the policy by transfer π_T: $\pi_T^{[0]} = \pi_X$. This approach generally improves performance at initialization, but makes little difference to the learning

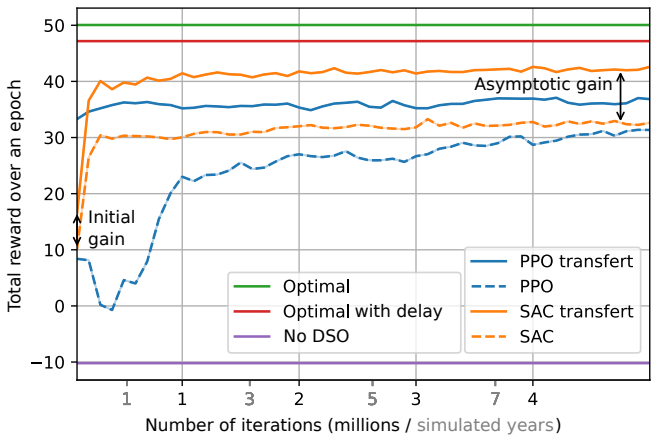

Fig. 8. Performance of transfer learning compared to simple training. The performance of three strategies requiring no training (optimal, optimal with delay and unmanaged) is given for comparison. Initial and asymptotic gains are shown for the SAC algorithm.

curve and asymptotic performance. In addition, it does not take advantage of the possible presence of several expert policies.

Learning by demonstration consists in using a group of expert policies and learning by selecting those with the smallest forecast errors. This approach will be explored in a future contribution.

Figure 8 shows the performance of the policy transfer. Each of the N expert policies $\pi_{X,i}$ has been trained on a different energy community Ψ_i. As described in section II, they differ in the agents of which they are composed. They are then used to initialize an agent to be trained in the Ψ_T community: $\pi_{T,i}^{[0]} = \pi_{X,i}$. Six agents per transfer are trained in this way. Their performance will be compared to that of six π_0 agents trained from scratch on the Ψ_T community. To simplify reading, only the best agent in each category is shown.

For PPO, the initial gain is 25, which is very encouraging, as it means that π_T is making close-to-optimum decisions right from the start of training. The risk of congestion is therefore limited during the first interactions with the new environment. The regret is $\mathcal{R} = 7.3 \cdot 10^7$. The asymptotic performance gain is less than 5 and might have been zero if π_0 had been trained longer. For SAC, the initial gain is small, but learning is faster. The regret is $\mathcal{R} = 4.9 \cdot 10^7$. The asymptotic gain is about 10. This can be explained by the fact that π_T explores its environment more efficiently than π_0. It is therefore less sensitive to local minima.

V. Conclusion and Future Works

The present contribution evaluated different learning techniques to predict the optimal price between an energy community and the main market, taking into account the capacity of the line connecting the community to the main grid. The *Soft Actor Critic* method was found to be more efficient than the *Proximal Policy Optimization* method, both in terms of learning speed and asymptotic performance. We show that learning of the line fee value can be achieved in reasonable times – of the order of one to two simulated years – and that a simple mechanism such as policy transfer can significantly speed up the convergence time while minimizing regret. This confirms the value of sharing information between communities. What is more, the proposed reward function can be modified by the DSO according to its risk aversion. These results have been achieved on the basis of a deliberately minimalist case study that has been developed in order to emphasize the interpretability of the results and lead to potentially generalizable rules. As part of a reproducible, open-science approach, the code developed and the data used are publicly accessible on a GitLab repository [3].

The perspectives of this study are as follows. First, learning must be robustified to identify configurations that can lead to excessive congestion, and setting line fees accordingly. This can be achieved by changing the form of the reward function, by integrating a so-called "pessimistic" supervisor to prevent exploration of risky states, or by improving transfer learning, notably through demonstration learning. Secondly, fee prediction needs to be generalized to all the lines in the network. The formalism of the optimal power flow informs us of the existence of nodal prices, enabling the network to be perfectly controlled and the optimal operating point to be reached. The prediction of these prices would be a generalization of the present contribution, allowing us to take into account voltage constraints, which are the predominant constraints within distribution networks.

References

[1] H. R. Galiveeti, A. K. Goswami, and N. B. Dev Choudhury, "Impact of plug-in electric vehicles and distributed generation on reliability of distribution systems," *Eng. Sci. Technol. an Int. J.*, vol. 21, no. 1, pp. 50–59, feb 2018.

[2] R. L. G. Latimier, G. Chérot, and H. B. Ahmed, "Online learning for distributed optimal control of an electric vehicle fleet," *Electric Power Systems Research*, vol. 212, p. 108330, 2022.

[3] C. Inês, P. L. Guilherme, M.-G. Esther, G. Swantje, H. Stephen, and H. Lars, "Regulatory challenges and opportunities for collective renewable energy prosumers in the eu," *Energy policy*, vol. 138, p. 111212, 2020.

[4] X. Jin, Q. Wu, and H. Jia, "Local flexibility markets: Literature review on concepts, models and clearing methods," *Appl. Energy*, vol. 261, p. 114387, mar 2020.

[5] F. Teotia and R. Bhakar, "Local energy markets: Concept, design and operation," *2016 Natl. Power Syst. Conf. NPSC 2016*, feb 2017.

[6] S. Moroni, V. Alberti, V. Antoniucci, and A. Bisello, "Energy communities in the transition to a low-carbon future: A taxonomical approach and some policy dilemmas," *Journal of Environmental Management*, vol. 236, pp. 45–53, 4 2019.

[7] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin, "Peer-to-peer and community-based markets: A comprehensive review," pp. 367–378, apr 2019.

[8] EU Parliament, "Directive (EU) 2019/944 on Common Rules for the Internal Market for Electricity and Amending Directive 2012/27/EU," Tech. Rep. L 158, 2019.

[9] S. Kerscher and P. Arboleya, "The key role of aggregators in the energy transition under the latest European regulatory framework," *Int. J. Electr. Power Energy Syst.*, vol. 134, p. 107361, jan 2022.

[10] F. Moret and P. Pinson, "Energy collectives: A community and fairness based approach to future electricity markets," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3994–4004, 2018.

[3] https://gitlab.com/satie.sete/learningoptim_exchprice_heterogenmarket

[11] T. Morstyn and M. D. McCulloch, "Multiclass energy management for peer-to-peer energy trading driven by prosumer preferences," *IEEE Transactions on Power Systems*, vol. 34, pp. 4005–4014, 9 2019.

[12] J. M. Zepter, A. Lüth, P. C. Del Granado, and R. Egging, "Prosumer integration in wholesale electricity markets: Synergies of peer-to-peer trade and residential storage," *Energy and Buildings*, vol. 184, pp. 163–176, 2019.

[13] G. Cui, Q.-S. Jia, and X. Guan, "Energy management of networked microgrids with real-time pricing by reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 570–580, 2023.

[14] I. Bouloumpasis, D. Steen, and L. A. Tuan, "Congestion Management using Local Flexibility Markets: Recent Development and Challenges," in *Proc. 2019 IEEE PES Innov. Smart Grid Technol. Eur. ISGT-Europe 2019*. Institute of Electrical and Electronics Engineers Inc., sep 2019.

[15] B. Cajna, R. Le Goff Latimier, and H. Ben Ahmed, "Hybrid Market Design Considering Heterogeneous System Operator Preferences," in *Proc. Innov. Smart Grid Technol. Eur. Conf. ISGT 2024*, Dubrovnik, Croatia, 2024.

[16] A. Marot, B. Donnot, C. Romero, B. Donon, M. Lerousseau, L. Veyrin-Forrer, and I. Guyon, "Learning to run a power network challenge for training topology controllers," *Electric Power Systems Research*, vol. 189, p. 106635, 2020.

[17] S. H. Oh, Y. T. Yoon, and S. W. Kim, "Online reconfiguration scheme of self-sufficient distribution network based on a reinforcement learning approach," *Applied energy*, vol. 280, p. 115900, 2020.

[18] J. Wang, W. Xu, Y. Gu, W. Song, and T. C. Green, "Multi-agent reinforcement learning for active voltage control on power distribution networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3271–3284, 2021.

[19] A. Petrusev, M. A. Putratama, R. Rigo-Mariani, V. Debusschere, P. Reignier, and N. Hadjsaid, "Reinforcement learning for robust voltage control in distribution grids under uncertainties," *Sustain. Energy, Grids Networks*, vol. 33, p. 100959, mar 2023.

[20] O. G. M. Khan, A. Youssef, M. Salama, and E. F. El-Saadany, "Management of congestion in distribution networks utilizing demand side management and reinforcement learning," *IEEE Systems Journal*, vol. 17, no. 3, pp. 4452–4463, 2023.

[21] A. S. Faria, T. Soares, T. Orlandini, C. Oliveira, T. Sousa, P. Pinson, and M. Matos, "P2P market coordination methodologies with distribution grid management," *Sustain. Energy, Grids Networks*, p. 101075, may 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352467723000838

[22] A. Kargarian, J. Mohammadi, J. Guo, S. Chakrabarti, M. Barati, G. Hug, S. Kar, and R. Baldick, "Toward Distributed/Decentralized DC Optimal Power Flow Implementation in Future Electric Power Systems," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 2574–2594, 2018. [Online]. Available: http://www.ieee.org/publications{_}standards/publications/rights/index.html

[23] R. Le Goff Latimier, H. Ben Ahmed, and B. Multon, "Distributed optimisation with restricted exchanges of information: charging of an electric vehicle fleet," 2018.

[24] A. Heydari, M. Majidi Nezhad, E. Pirshayan, D. Astiaso Garcia, F. Keynia, and L. De Santoli, "Short-term electricity price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm," *Appl. Energy*, vol. 277, p. 115503, nov 2020.

[25] M. Chatzos, T. W. Mak, and P. V. Hentenryck, "Spatial Network Decomposition for Fast and Scalable AC-OPF Learning," *IEEE Trans. Power Syst.*, vol. 37, no. 4, pp. 2601–2612, jul 2022.

[26] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron, "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark," *Appl. Energy*, vol. 293, p. 116983, jul 2021.

[27] R. Henry and D. Ernst, "Gym-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems," *Energy AI*, vol. 5, p. 100092, sep 2021.

[28] C. Shi, R. Wan, R. Song, W. Lu, and L. Leng, "Does the markov decision process fit the data: Testing for the markov property in sequential decision making," in *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814. International Machine Learning Society (IMLS), feb 2020, pp. 8766–8776. [Online]. Available: https://arxiv.org/abs/2002.01751v1https://github.com/RunzheStat/TestMDP

[29] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, jan 2021.

[30] V. Z. Gjorgievski, S. Cundeva, and G. E. Georghiou, "Social arrangements, technical designs and impacts of energy communities: A review," *Renew. Energy*, vol. 169, pp. 1138–1156, may 2021.

[31] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J. Albrecht, and Others, "Smart*: An open data set and tools for enabling research in sustainable homes," *SustKDD, August*, vol. 111, no. 112, p. 108, 2012.

[32] Å. L. Sørensen, K. B. Lindberg, I. Sartori, and I. Andresen, "Residential electric vehicle charging datasets from apartment buildings," *Data in Brief*, vol. 36, p. 107105, 2021.

[33] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann, "Stable Baselines3," \url{https://github.com/DLR-RM/stable-baselines3}, 2019.

[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv*, vol. 1707.06347, jul 2017. [Online]. Available: https://arxiv.org/abs/1707.06347v2http://arxiv.org/abs/1707.06347

[35] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 5. International Machine Learning Society (IMLS), jan 2018, pp. 2976–2989. [Online]. Available: https://arxiv.org/abs/1801.01290v2

[36] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Res.*, vol. 10, pp. 1633–1685, dec 2009. [Online]. Available: https://dl.acm.org/doi/10.5555/1577069.1755839

[37] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer Learning in Deep Reinforcement Learning: A Survey," *arXiv Prepr.*, vol. abs/2009.0, sep 2020. [Online]. Available: https://arxiv.org/abs/2009.07888v5http://arxiv.org/abs/2009.07888