TMSG - Training Multiagent Strategies for the Grid

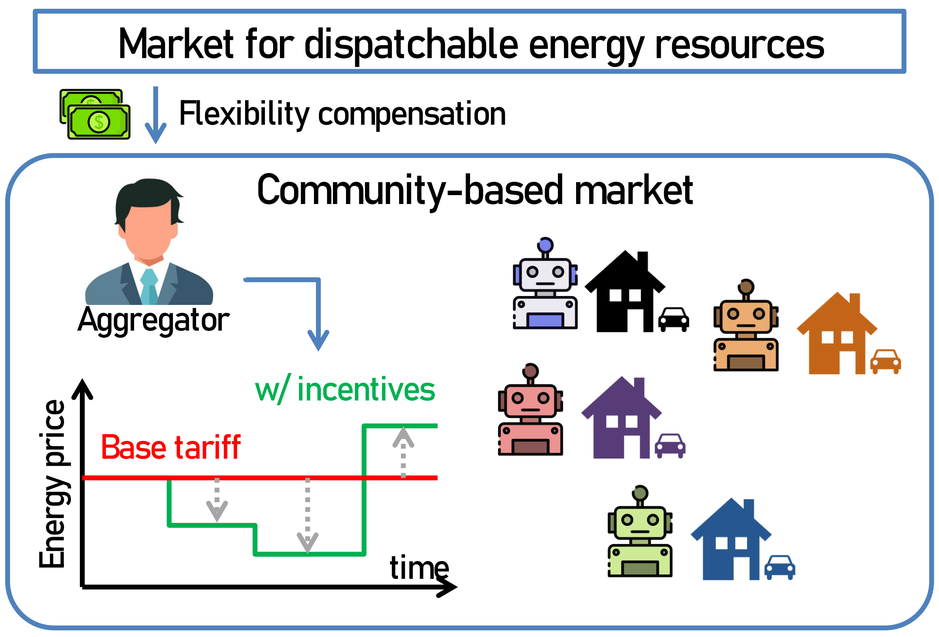
Federico Rosato *(federico.rosato@supsi.ch)*

ISAAC (Institute for Sustainability Applied to the Built Environment), DACD, SUPSI

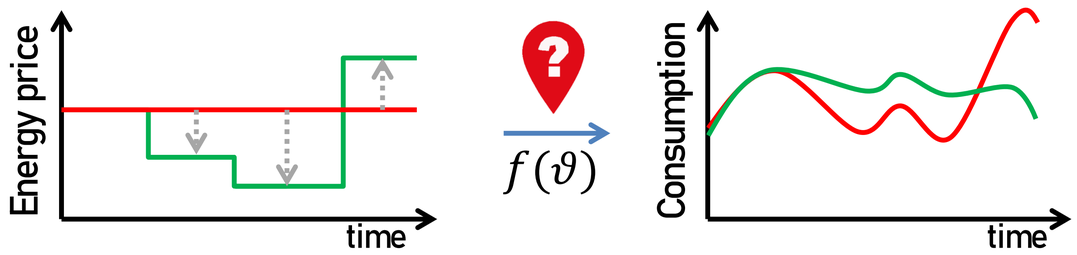
Electrical distribution systems are more and more populated by non-programmable generation, distributed storage, steerable appliances, and electric vehicles, which challenge the traditional distribution grid. On the other hand, smart grid technology offers a technological substrate that can be leveraged to tackle these challenges. The dispersed nature of the appliances renders this background ideal for applying implicit, incentive-driven Demand Response through automated agents. Designing the behavior of these agents, though, is very difficult, since they must strategically react to a stochastic environment whose state evolution also depends on the actions of all the other agents, both seeking economic advantage and user comfort and collectively avoiding the violation of grid constraints. In the TMSG project, we seek to formalize the environment as a Markov Game, a common setting for applying Reinforcement Learning techniques, and then to introduce in this field methods borrowed from recent AI and game theory literature that have seen little application in the Smart Grid setting. The strategic behavior obtained will be analyzed in search of insight into the quantification of flexibility. In the final phase, the project will also involve power flow simulations, including the trained algorithmic agents.

Introduction

In this work, the general setup used is that of an Energy Community with a managed, community-based market. The community acts as a Balancing Services Provider represented by an Aggregator that bids up/down flexibility on a market for dispatchable energy resources for a compensation. The Aggregator then decides a tariff for the day, realizing an implicit steering incentive.



The core research question therefore pertains the quantification of the flexilibity of the community, modeled as a group of rational self-interested agents parametrized by a vector ϑ including quantitative penalties for user discomfort, coordination penalties, etc. The agents seek to maximize the benefit offered to the user, including absorbing energy from the grid when the price is low, while also avoiding excessive coordination that would disrupt the grid. Knowing the flexibility also allows the Aggregator to balance his bids on the market.



Markov Game modeling

The framework above is modeled as a Markov Game (S, A, P, R, γ). The state space S is composed of three values common to all agents (time of day, weather and forecast), which we call public state, and the values related to flexibilities for each individual agent (SOC of a stationary battery, SOC of an EV, boolean presence of the SOC at the charging point, building thermal inertia). The agents can decide whether to charge the batteries and/or to fill the thermal inertia, giving a space A of 4 possible actions.

The rewards (penalties) experienced by the automatic agents are the dissatisfaction with the thermal condition of the building on the side of the user, the dissatisfaction with the SOC of the EV should the user request it, the price of electricity and a quadratic term for excessive coordination. The state transition probabilities from one state to the next are computed according to a set of both deterministic and probabilistic rules; the resulting matrix P is sparse.

Solution Methods

With the Markov Game model in place, we turn to the search for solution methods. A first method is based on Replicator Dynamics, and consists of the following ODE system, that prescribes the player to evolve a random strategy by incrementing the frequency of an action at a state proportionally to how much value would be gained by performing that pure action all else being equal.

$$\frac{dx_{s,a}^{h}}{dt} = [v_a^x(s,h) - V^x(s,h)]x_{s,a}^{h}$$

$$v_a^x(s,h) = \hat{R}(s,\chi) + \gamma \sum_{s' \in S} \hat{P}(s'|s,\chi) V^x(s',h)$$

This method, albeit offering good theoretical guarantees, is resource-hungry and does not scale well with the number of players and finer discretization of the state space. To tackle this problem, I studied more carefully the peculiar, favorable structural properties of the game being played, identifying essentially two aspects:

- A solution to the «base» Markov Game would require knowledge of the whole state of the game by each agent; in reality, the agents cannot access the private state of the others (imperfect information), so the agent can only reply with a single mixed strategy to the «information set» identified by his measures.
- The agents have the same fundamental nature, avaliable actions and interests. It is meaningful to treat them, at least initially, as completely symmetric (same flexibilities and stochastic characteristics).

I devised a two-step method partly inspired by jam-fold solution methods to tournament poker modeled as a Markov Game. The method is iterative and resembles Policy Iteration, and at its core consists in finding a common strategy for all the symmetric agents by solving the following convex optimization problem at each public state condition $S_{p,h}$:

$$\max_x \sum_{s \in S_{p,h}} \left(\sum_{a \in A} Q(s,a) x_{s,a} - c \sum_{s' \in S_{p,h}} \mathbb{P}(s'|s) \left(\sum_{a \in A} x_{s',a} \xi(a) \right)^2 \right)$$

$$s.t. \quad x \in \mathbb{R}_+^{|S_{p,h}| \times |A|}$$
$$\sum_{a \in A} x_{s,a} = 1 \ \ \forall s \in S_{p,h}$$

This is done at each iteration, and the value of the states is calculated with the appropriate Bellman Equation according to the policy found at the previous iteration. This method, currently under testing, has proven to yield sensible solutions in a reasonable time empirically, albeit convergence is difficult to study due to the intrinsic deep nonstationariety of the multiagent learning problem. Notice the last term of the iteration, which represents the quadratic penalty for eccessive coordination and is the term that «couples» agents together inducing mixed strategies; without it, the method would reduce to standard Policy Iteration. $\xi(a)$ is a function mapping action a to its quantitative effect on coordination, and c is a weight parameter. $\mathbb{P}(s'|s)$ is the probability of an agent being in state s' from the perspective of an agent which is in state s (which is not economical to compute, and therefore is methodologically studied extensively in the complete work).

Remainder of the project

In the next phases of the project, the methods exposed will be expanded, analyzed theoretically and used in analysis campaigns with varying parameters, in order to find sensible strategies and gain insight. The stochastic transitions will be fitted to real data. The trained agents will be co-simulated with the grid power flow to verify grid impact.

The research presented here is supported by the Swiss Federal Office of Energy SFOE as part of SOUR project TMSG.