Distributed Learning in Hierarchical Networks

Hélène Le Cadre
CEA, LIST
91191 Gif-sur-Yvette CEDEX
FRANCE
Email: helene.le-cadre@cea.fr

Jean-Sébastien Bedo
France-Télécom/Orange
6, place d'Alleray, 75015 Paris
FRANCE
Email: jeansebastien.bedo@orange.com

***Abstract*—In this article, we propose distributed learning based approaches to study the evolution of a decentralized hierarchical system, an illustration of which is the smart grid. Smart grid management requires the control of non-renewable energy production and the integration of renewable energies which might be highly unpredictable. Indeed, their production levels rely on uncontrolable factors such as sunshine, wind strength, etc. First, we derive optimal control strategies on the non-renewable energy productions and compare competitive learning algorithms to forecast the energy needs of the end users. Second, we introduce an online learning algorithm based on regret minimization enabling the agents to forecast the production of renewable energies. Additionally, we define organizations of the market promoting collaborative learning which generate higher performance for the whole smart grid than full competition.**

***Index Terms*—Algorithmic Game Theory, Coalition, Distributed Learning, Regret.**

I. Introduction

We describe and test distributed learning algorithms in the context of a hierarchical highly connected network of agents, an illustration of which is the smart grid. A first class of algorithms will be used to control the non-renewable energy production. Then a second class will be proposed to favor the introduction of renewable energy sources. The first class of algorithms belongs to the reinforcement learning category. Tit for tat and fictitious play are well-known in repeated game theory [3], [18]. They require different knowledge levels i.e., the access to the agents' previous decisions under tit for tat and the access to the full history of the agents' past decisions under fictitious play. They form a natural bridge with more sophisticated learning rules used in machine learning such as regret based learning [2], [18]. This latter is a rather original approach in the machine learning community [1], [2]. When properly designed, its performance to learn mixed strategies over various datasets overperform other classical machine learning approaches such as support vector machines, autoregressive processes and artificial neural networks [6]. The originality of this article for the machine learning community relies on the study of these algorithms in a decentralized and hierarchical framework resulting from the double leader-follower structure of the game.

We have chosen to place the model in the context of the smart grid although the results that we derive are quite generic and can be applied to any hierarchical network with a distributed access to the scarce resource and no capacity storage. Such framework is quite common in the revenue management literature [16] and can be applied to many other industries like to model the interconnections between autonomous systems or content providers and access providers in the Internet, to better understand the relations between suppliers and retailers in the retail industry, etc. Going back to the smart grid, we give a broad definition of it. Initially, smart grids are networks enabling a decentralized production of the energy and involving bidirectional energy flows which are controlled by a complex, global and secured communication network. The network is said to be *smart* because it is capable of integrating efficiently each agent's action (producers, providers and consumers) in order to guarantee a sustainable and secure supply at lower cost.

Nowadays, in Europe and in France especially, traditional electrical networks rely on nuclear based energies [4]. The main difficulty is to adapt the production level so as to meet the uncertain demand level. In this article, we build a first model where two learning strategies based on tit for tat and fictitious play are used to adapt the production level to meet the demand level. In this context, the distributed learning algorithms can be used as distributed control algorithms providing boundaries on the agents' biases in the prediction process.

Because of their structure, smart grids offer a substantial opportunity for the integration of renewable energies. Renewable energies like wind power, photovoltaic, geothermal, biomass, small hydroelectric facilities, etc., are highly unpredictable since they depend on uncontrolable exterior factors like wind, level of sunshine, etc. They are cleaner but their production being more difficult to forecast, they are far more difficult to integrate into the electrical network. It requires to develop efficient learning algorithms to predict both the consumers' demand, which can be highly erratic due to their new active role in the grid, and the renewable energy productions.

In practice, the electrical network based on the smart grid model is composed of a multitude of microgrids. Microgrids are modern, small-scale versions of the centralized electrical system. They can be either sellers in case where they have a surplus of power to transfer, or buyers in case where they need to buy additional power to meet their demand. Saad et al. propose a distributed microgrid coalition formation algorithm enabling a decrement of 31% of the average power losses relative to the non-cooperative case, in [14]. However, their result relies on strong assumptions. Indeed, they make the

VALUETOOLS 2012, October 09-12, Cargèse, France

DOI 10.4108/valuetools.2012.250217

hypothesis that the consumers' demand is random since it depends on unpredicable factors such as consumption level, consumption behavior, etc. [8]. Additionally, they make the simplifying assumption that the power surplus which is defined as the difference between the total power and the demand, is distributed according to a known density function.

Actually, the justification of the fitting of a specific parametric density function requires the game designer to learn at least its parameters. In the statistical learning literature, there are three major learning approaches, each of them corresponding to a particular abstract learning task: supervised learning, unsupervised learning and reinforcement learning [2], [17]. Tasks that fall within the paradigm of reinforcement learning are control and online optimization problems, games and other sequential decision making tasks. Learning based on regret minimization as described in [2], belongs to this category. Additionally, we observe in [6] that the performance resulting from learning based on regret minimization tested on real data bases, are clearly superior to the ones obtained using supervised learning approaches. As a result, these points have convinced us to use a learning approach based on regret minimization. The difficulty is then to extend the already existing method to a distributed learning framework and to clearly identify how such a decentralized access to the information will affect the smart grid economic organization.

The existing literature on distributed learning primarly focuses on distributed learning algorithms that are suitable for implementation in large scale engineering systems [7], [12], [17]. The results mainly concentrate on a specific class of games, called games of potential [15]. This class of games is of particular interest since they have inherent properties that can provide guarantees on the convergence and stability of the system. However, there exist some limitations to this framework. The most striking one is that it is frequently impossible to represent the interaction framework of a given system as a potential game [9].

The learning game studied in this paper belongs to the category of *repeated uncoupled games*. Indeed one agent cannot predict the forecasts and so actions of the other agents at a given time period. To take his decision i.e., optimal prices and traded quantities of energy, each agent is aware of the history of forecasts of all the agents and of his utility. Recent work has shown that for finite games with generic payoffs there exist completely uncoupled learning rules i.e., rules where the agents observe only their own prediction history and their utility, that lead to Nash equilibria that are Pareto optimal [12]. Marden et al. exhibit a different class of learning procedures that lead to Pareto optimal vector of actions that do not necessarily coincide with Nash equilibria [9]. A well-known illustration of the practical interest of Pareto optimality can be found in the prisoner dilemma where the Nash equilibrium is inefficient compared to the Pareto optimum that would be obtained if the prisoners had collaborated [10]. However, one serious concern regarding Pareto optimality is that the optimum is generally not unique and deciding which one should be implemented by the system might in some cases, require contracts, communication or bargaining mechanisms to be designed at the beginning of the game. Of course this is not always the case, as proven by [9] and [12]. Under conditions stating that it is not possible to divide the interacting agents into two distinct subsets that do not mutually interact with one another, Marden et al. prove that the game dynamics induce a Markov process over the finite state space which is defined as the set of the triples containing the chosen action, the resulting utility and an additional binary parameter called the agent's mood [9]. Then they focus on characterizing the support of the limiting stationary distribution i.e., the stable states. In particular, they prove that any stable state maximizes the social welfare under an initial assumption on agent interdependency. Compared with our model, Marden et al. study a stochastic game where the agents always control their production. They take as an example the wind turbines which can adapt their power to maximize the whole wind farm's social welfare defined as the sum of the total power produced by each turbine. In our game setting, the production of the renewable energies cannot be controlled since it relies on exogeneous events. Therefore, it requires to introduce decentralized learning approaches based on online optimization [1] and to study the resulting economic interactions using a game theoretic approach.

The article is organized as follows. In Section II, we introduce the economic interplays between the agents, describe the repeated game setting and the optimization program for each agent. Then the double Stackelberg game is solved in Section III in complete and partial information frameworks. In Section IV, we consider the case where producers control their productions while service providers need to make predictions about the microgrid energy needs to optimize their prices and traded quantities of energy. The performance of two learning algorithms: tit for tat and fictitious play are evaluated analytically. In Section V, the integration of renewable energies in the grid requires the development of an efficient online learning algorithm based on regret minimization. We finally prove that collaborative learning through a grand coalition generates higher performance for the whole smart grid than in the case of individual learning under full competition.

II. The model

The smart grid ecosystem is made of three categories of agents:

- K energy producers $e_1, ..., e_K$ which can be associated with nuclear plants, photovoltaic park managers, wind farm administrators, etc. The produced energy can be either non-renewable like in the case of nuclear plants or renewable when it comes from photovoltaic parks, wind farms, etc.
- n service providers $s_1, ..., s_n$ which might buy energy from each of the K producers and route it through their transport network to their clients.
- n microgrids $\mathcal{M}_1, ..., \mathcal{M}_n$, each one of them being committed with a single energy provider. We assume that each end user contracts with only one service provider

and does not churn from one service provider to another during all the period of our study. This assumption holds well if we consider local or regional utility companies for example.

We introduce two individual sequences: $\nu_i^s(t)$, which contains the energy needs issued from the clients of service provider s_i and $\nu_k^e(t)$, which coincides with producer e_k's production, at time period t. The economic relationships between the agents in the grid are pictured in Figure 1. The symbol \$ is used to represent the directed monetary transfers between the involved agents.

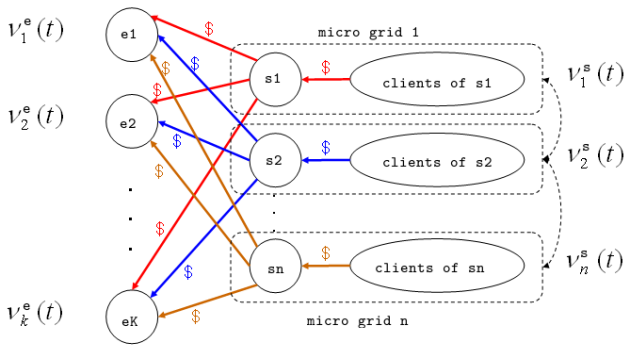


Fig. 1. Economic relations between the agents in the grid.

A. Repeated game setting

We model the interplay between all the agents through a repeated game. At each time period t of the game:

(1) The energy producers e_k communicate their prices $\tilde{p_k}(t) > 0$ for one energy unit (i.e., Wh) to the service providers

(2) The service providers s_i place energy quantity orders to energy producers: the quantity ordered by s_i to e_k is denoted by $q_{ik}(t)$

(3) The service providers s_i communicate their prices $p_i(t) > 0$ for one energy unit to their end users

The end users $\mathcal{M}_i$ need $\nu_i^s(t)$ energy units for the period (which could depend on weather, cooking, etc.)

(4) The end users decide to find alternative sources of energies for $a_i(t)$ energy units. They buy the rest of their needs $\Big(\nu_i^s(t) - a_i(t)\Big)$ to service provider s_i

Each energy producer e_k produces $\nu_k^e(t)$ energy units

The energy producers distribute their production to service providers: e_k delivers $\alpha_{ki}(t)\nu_k^e(t)$ energy units to service provider s_i

($\star$) The energy producers incur penalties if they did not fulfill the energy orders placed by the service providers

($\star$) The service providers incur penalties if they did not fulfill the energy needs of their end users

The penalties are proportional to the difference between the initial energy order and the final energy delivery. More precisely, e_k incurs the penalty $\tilde{\gamma_i}\Big(q_{ik}(t) - \alpha_{ki}(t)\nu_k^e(t)\Big)_+$ with $\tilde{\gamma_i} > 0$, from service provider s_i, and s_i incurs the penalty $\gamma_i\Big(\nu_i^s(t) - a_i(t) - \sum_k \alpha_{ki}(t)\nu_k^e(t)\Big)_+$ with $\gamma_i > 0$, from end users $\mathcal{M}_i$. The penalties are donated to the unbiased regulator who is supposed to control the overall energy distribution system. Various mechanisms of transfer can then be implemented by the unbiased regulator. The penalties can be distributed to shareholders of the energy producers or used for social purposes.

B. Optimization program for each agent

The only decision variable for the end users is the quantity of energy that they decide to get from alternative sources: $a_i(t)$. We assume that the end users have no lever to influence their random energy needs: $\nu_i^s(t)$. The information available to end users is the price of an energy unit from their service provider and their energy needs $\nu_i^s(t)$ for the time period. We assume that finding alternative energy sources rather than buying it to the service provider has some costs for the end users. More precisely, finding $a_i(t)$ energy units through alternatives costs them $\frac{a_i(t)^2}{2}$ per time period. As a result, the total cost of energy for the end users $\mathcal{M}_i$ is:

$$p_i(t)\Big(\nu_i^s(t) - a_i(t)\Big) + \frac{a_i(t)^2}{2} \tag{1}$$

End users $\mathcal{M}_i$ choose $a_i(t)$ in order to minimize their total cost of energy depending on the energy price.

The decision variables for each service provider s_i are the energy unit price $p_i(t)$ and the energy orders $q_{ik}(t)$ for each energy producer e_k. The only information available to service provider s_i when he makes his decision are the energy unit prices $\tilde{p_k}(t)$ of all the energy producers. He has to forecast the energy needs of his customers and the energy production of all the energy producers. Following our description of the interplay between the agents, the utility for service provider s_i at time period t is:

$$\begin{aligned}\pi_i(t) &= p_i(t)\Big(\nu_i^s(t) - a_i(t)\Big) - \sum_{k=1,\dots,K} q_{ik}(t)\tilde{p_k}(t) \\ &- \gamma_i\Big(\nu_i^s(t) - a_i(t) - \sum_{k=1,\dots,K} \alpha_{ki}(t)\nu_k^e(t)\Big)_+ \end{aligned} \tag{2}$$

where for any real x, we set: $x_+ = \max\{x; 0\}$. Service provider s_i chooses his energy unit price and his energy orders toward energy producers so that $\pi_i(t)$ is maximized.

The only decision variable for each energy producer e_k is the energy unit price $\tilde{p_k}(t)$ that he proposes to the service providers. We assume that the energy producer cannot influence directly the energy he produces at each time period. This assumption holds well if, for example, we look at a wind turbine farm without any investment in an additional wind turbine during the study period. The variation of the wind intensity will make vary the energy produced without any lever for the energy producer. When energy producer e_k makes his decision, he has no information because he is the first agent to play in the time period. He has to forecast the energy quantity

that he will produce and the energy orders of all the service providers.

To define the sharing coefficients $\alpha_{ki}(t)$, we consider a weighted proportional allocation of resource that allows producers to discriminate energy allocation by providers. This framework is a generalization of the well-known proportional allocation [14] to *weighted energy orders* with penalty coefficients as weights. Such a resource sharing mechanism has already been introduced by Nguyen and Vojnović, in [11]. This means that between two providers booking the same quantity, the one having the highest penalty coefficient will receive the largest part of the producer's available energy. Indeed, the producer wants to minimize his overall penalty and therefore allocates larger parts of his production to providers who appear to him as more threatening than the others. More precisely, in the rest of the article, we will assume that:

$$\alpha_{ki}(t) = \frac{\tilde{\gamma_i} q_{ik}(t)}{\sum_{j=1,...,n} \tilde{\gamma_j} q_{jk}(t)} \tag{3}$$

Then the utility of energy producer e_k at time period t equals:

$$\begin{aligned} \tilde{\pi_k}(t) &= \tilde{p_k}(t) \sum_{i=1,...,n} q_{ik}(t) - \sum_{i=1,...,n} \tilde{\gamma_i} q_{ik}(t) \Big(1 - \tilde{\gamma_i} \nu_k^e(t) \\ &\quad \Big[\sum_{j=1,...,n} \tilde{\gamma_j} q_{jk}(t) \Big]^{-1} \Big)_+ \end{aligned} \tag{4}$$

Energy producer e_k chooses his energy unit price so that $\tilde{\pi_k}(t)$ is maximized.

III. Double Stackelberg game resolution

At first we consider that the double Stackelberg game described in Section II-A is played in a complete information setting on the individual sequences $\nu_i^s(t), \nu_k^e(t),\ \forall i = 1, ..., n,\ \forall k = 1, ..., K$.

Proposition 1: If the system is in energy shortage (Inequality (16)) and if penalties are fair (Inequality (15)), the double Stackelberg game under complete information admits a unique equilibrium :

- for each provider $s_i, \forall i = 1, ..., n,\ a_i(t) = p_i(t)$ is defined by Equation (11) and the quantity orders $q_{ik}(t)$ are defined by Equation (14)
- for each producer $e_k, \forall k = 1, ..., K.$, $\tilde{p_k}(t)$ is defined by Equation (17)

Proof of Proposition 1. The proof detailing the agents' optimal decisions can be found in Appendix and in [5].

We now consider that the individual sequences $\nu_i^s(t), \nu_k^e(t), \forall i = 1, ..., n, \forall k = 1, ..., K$ are not public knowledge. To simplify, we will assume that the energy needs $\nu_i^s(t)$ are all drawn from a common set $\mathcal{X}_s$ and that energy productions $\nu_k^e(t)$ are all drawn from a common set $\mathcal{X}_e$. We assume that $\mathcal{X}_s$ and $\mathcal{X}_e$ are public knowledge and are finite i.e., $|\mathcal{X}_s| < +\infty$ and $|\mathcal{X}_e| < +\infty$.

In the rest of the article, the game settings (energy needs of the microgrids, energy production and penalties) will be chosen so that *we are always in energy shortage* and with fair penalties. It can be ensured easily by choosing the maximum value of $\mathcal{X}_e$ small enough compared to the minimum value of $\mathcal{X}_s$.

According to the analytical expression of the equilibrium derived in the proof of Proposition 1 and in [5], service providers need to forecast the energy productions of each energy producer and the energy needs of the microgrids to optimize their decisions. We will denote by $f_i(X, t)$ the forecast of service provider s_i at time period t for the random variable X. We will also use the following notations:

- $f_i(t) = \Big\{ f_i(\nu_i^s, t), f_i(\nu_1^e, t), ..., f_i(\nu_K^e, t) \Big\}$ to denote the predictions made by service provider s_i about microgrid $\mathcal{M}_i$ instantaneous needs and about the production of each energy producer $e_k,\ k = 1, ..., K$.
- $f(t) = \Big\{ f_1(t), ..., f_n(t) \Big\}$ which contains the forecasts of all the service providers.
- $f_{-i}(y, t) = \Big\{ f_1(t), ..., f_{i-1}(t), y, f_{i+1}(t), ..., f_n(t) \Big\}$ which contains the forecasts of all the service providers except s_i which prediction is set equal to y.
- $\nu(t) = \Big\{ \nu_1^s(t), ..., \nu_n^s(t), \nu_1^e(t), ..., \nu_K^e(t) \Big\}$ which contains the microgrid energy needs and the production of each energy producer $e_k,\ k = 1, ..., K$.

By substitution of the forecasters in the Stackelberg game solution at equilibrium as obtained in Proposition 1 proof and in [5], we infer the optimal decisions for service provider s_i at each time period t:

$$\begin{aligned} p_i(t) &= \frac{f_i(\nu_i^s, t) + \gamma_i}{2} \\ q_{ik}(t) &= \frac{f_i(\nu_k^e, t)}{\tilde{p_k}(t)} \frac{L(i)}{\tilde{\gamma_i}} \frac{n-1}{\delta} \end{aligned}$$

As a result, the utility of service provider s_i at each time period t is:

$$\begin{aligned} \pi_i(t) &= \frac{f_i(\nu_i^s, t) + \gamma_i}{2} \Big(\nu_i^s(t) - \frac{f_i(\nu_i^s, t) + \gamma_i}{2} \Big) - \frac{L(i)}{\tilde{\gamma_i}} \\ &\quad \frac{n-1}{\delta} \sum_{k=1,...,K} f_i(\nu_k^e, t) - \gamma_i \Big(\nu_i^s(t) - \frac{1}{2} \Big(f_i(\nu_i^s, t) \\ &\quad + \gamma_i \Big) - \sum_{k=1,...,K} \frac{f_i(\nu_k^e, t) L(i)}{\sum_{j=1,...,n} f_j(\nu_k^e, t) L(j)} \nu_k^e(t) \Big)_+ \end{aligned} \tag{5}$$

The choice of the agents' forecasting strategies varies depending whether we consider renewable or non-renewable energies. Indeed, in case of non-renewable energies, the producers have the opportunity to control their production and the providers can play on their prices to control their demand level. Therefore, in this case, the distributed learning problem becomes a distributed control problem.

IV. Case of non-renewable energies: a distributed control problem

In this section, we suppose that each producer controls his production. This is typically the case for nuclear plants

or hydraulic centrals which constitute today the majority of energy sources in Europe [4]. Additionally, providers can adjust the demand of their mircrogrid through price incentives.

Substituting the optimal price $\tilde{p}_k(t)$ and traded quantities of energies $q_{ik}(t), \forall i = 1, ..., n$ in producer e_k's utlility as defined through Equation (4), we observe that it is linear increasing in $\nu_k^e(t)$. As a result, producer e_k will maximize his production at any time period, i.e.: $\nu_k^e(t) = \max\{\mathcal{X}_e\}$. The service providers anticipating the behavior of the producers, they will align their forecasts concerning the productions on the value $\max\{\mathcal{X}_e\}$. By differentiation of service provider s_i's utility defined in Equation (5) under energy shortage assumption, we obtain that it is maximized in $f_i(\nu_i^s, t)$ if, and only if, $f_i(\nu_i^s, t) = \nu_i^s(t)$ i.e., the forecast made by s_i about microgrid $\mathcal{M}_i$ energy needs coincides with its true value. The problem of course, is that at the beginning of time period t, provider s_i does not observe $\nu_i^s(t)$. Therefore he needs to develop learning approaches to predict it.

In this section, we apply two learning approaches which are quite classical in the theory of learning in games [3], [18]. While tit for tat requires a rather low level of information, agents basing their forecasts on fictitious play need to keep track of the history of all the past predictions. In spite of this larger memory requirement, fictitious play forms a natural bridge between rather naive type of learning such as tit for tat where the agents' forecasts are based solely on their earlier observation, and more sophisticated rules like regret based learning that will be detailed in the case of renewable energies in Section V.

A. Tit for tat

Tit for tat is a commonly used strategy in repeated game theory. An agent using this strategy begins by cooperating, and then answers to his opponent's previous action. For example, in the repeated prisoner dilemma with infinite or unknown time horizon, the prisoners begin by cooperating, since it enables them to maximize the sum of their utility. They repeat this strategy until one of them defect in which case the other defects too in the next time step [10].

Assuming that providers use tit for tat as learning strategy, they estimate their microgrid energy needs at time period t using the commonly shared energy needs of the microgrid at time period $t-1$. Judging by the form of $\pi_i(t-1)$ as defined in Equation (5), once they have observed the utility value at the end of time period $t-1$, it is quite straigthforward for them to infer $\nu_i^s(t-1)$ since $\pi_i(t-1)$ is linear in $\nu_i^s(t-1)$ and they know all the other terms in the equality. Therefore under tit for tat, for any provider s_i, the forecaster takes the form: $f_i^s(\nu_i^s, t) = \nu_i^s(t-1), \forall i = 1, ..., n$. At time period t, provider s_i forecasts microgrid $\mathcal{M}_i$ energy needs using its previous value.

In the following proposition, we give the analytical expression of provider s_i's loss:

$$l_i\Big(f(t), \nu(t)\Big) = \Big(\pi_i^0(t) - \pi_i(t)\Big)$$

It is the difference between what he would have received if his predictions were correct or unbiased ($\pi_i^0(t)$) and what he really receives in the course of the game.

Proposition 2: Under the assumption that provider s_i's forecasting strategies about microgrid $\mathcal{M}_i$ energy needs are based on tit for tat, his loss at time period t is:

$$\begin{aligned} l_i\Big(f(t), \nu(t)\Big) &= \Big(\nu_i^s(t) - \nu_i^s(t-1)\Big)\Big[\frac{\gamma_i}{2} \\ &- \frac{1}{4}\Big(\nu_i^s(t-1) - \nu_i^s(t)\Big)\Big] \end{aligned}$$

Proof of Proposition 2. We have made the assumption that the penalty coefficients are chosen so that we are always in energy shortage. It corresponds to Case 2 described in Subsection B.2 of Appendix. Practically, this means that the penalty term in Equation (5) is positive. Substituting the tit for tat prediction rule: $f_i(\nu_i^s, t) = \nu_i^s(t-1)$ in Equation (5), we obtain that: $l_i\Big(f(t), \nu(t)\Big) = \frac{\nu_i^s(t)+\gamma_i}{2}\Big(\nu_i^s(t) - \frac{\nu_i^s(t)+\gamma_i}{2}\Big) + \gamma_i\frac{\nu_i^s(t)+\gamma_i}{2} - \Big[\frac{\nu_i^s(t-1)+\gamma_i}{2}\Big(\nu_i^s(t) - \frac{\nu_i^s(t-1)+\gamma_i}{2}\Big) + \gamma_i\frac{\nu_i^s(t-1)+\gamma_i}{2}\Big]$. Simplifying the identical terms of opposite sign, we obtain the stated formula: $l_i\Big(f(t), \nu(t)\Big) = \Big(\nu_i^s(t) - \nu_i^s(t-1)\Big)\Big[\frac{\gamma_i}{2} - \frac{1}{4}\Big(\nu_i^s(t-1) - \nu_i^s(t)\Big)\Big]$. □

B. Fictitious play

Fictitious play is a widely used model of learning [3]. In this process, agents behave as if they think they are facing a stationary, but unknown, distribution of opponents' strategies. This assumption might be too strong in particular if the system in which the agents are learning fails to converge.

Considering fictitious play as the learning rule, producers and providers choose their forecast of microgrid $\mathcal{M}_i$ energy needs as a best response to the empirical distribution of the others' play up to time period $t-1$. As all service providers align their predictions about energy productions on the value $\max\{\mathcal{X}_e\}$ and as the profit of s_i does not depend on the energy needs of other microgrids than $\mathcal{M}_i$, the learning strategy can be determined as solution of the optimization problem:

$$f_i(\nu_i^s, t) = \arg\max_{y \in \mathcal{X}_s} \frac{1}{t-1}\sum_{l=1}^{t-1} \pi_i(l)|_{f_i(\nu_i^s, l) = y} \qquad (6)$$

As in Subsection IV-A, we give in the proposition below the analytical expression of provider s_i's loss.

Proposition 3: Under the assumption that provider s_i's forecasting strategy about microgrid $\mathcal{M}_i$ energy needs is based on fictitious play, his loss at time period t is:

$$\begin{aligned} l_i\Big(f(t), \nu(t)\Big) &= \Big(\frac{\nu_i^s(t)+\gamma_i}{2}\Big)^2 - \Big(\frac{\gamma_i}{2} \\ &+ \frac{1}{4}(\gamma_i\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big)\Big(\nu_i^s(t) + \frac{\gamma_i}{2} \\ &- \frac{1}{4}(\gamma_i + \frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big) \end{aligned}$$

Proof of Proposition 3. Going back to Equation (6) defining the updating rule under fictitious play, we express the

objective function as a function of y:
$\frac{1}{t-1}\sum_{l=1}^{t-1}\pi_i(l)|_{f_i(\nu_i^s,l)=y} = -(\frac{y+\gamma_i}{2})^2 - \frac{L(i)}{\tilde{\gamma}_i}\frac{n-1}{\delta}K\max\{\mathcal{X}_e\} + \frac{\gamma_i}{2}(y+\gamma_i) + \gamma_i\frac{K\max\{\mathcal{X}_e\}L(i)}{\sum_{j=1,...,n}L(j)} + (\frac{y+\gamma_i}{2}-\gamma_i)\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l)$. It is then maximized if, and only if, $y = \frac{1}{2}\Big[\gamma_i + \frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l)\Big]$. By substitution in provider s_i's utility as defined in Equation (5), we obtain provider s_i's loss at time period t:

$$\begin{aligned} l_i\Big(f(t),\nu(t)\Big) &= \Big(\frac{\gamma_i}{2}+\frac{1}{4}(\gamma_i+\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big)^2 - \Big(\frac{\gamma_i}{2} \\ &+ \frac{1}{4}(\gamma_i+\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big)\nu_i^s(t) - \gamma_i\Big(\frac{\gamma_i}{2} \\ &+ \frac{1}{4}(\gamma_i+\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big) \end{aligned}$$

After simplifications and factorizations, we obtain the stated formula: $l_i\Big(f(t),\nu(t)\Big) = \Big(\frac{\nu_i^s(t)+\gamma_i}{2}\Big)^2 - \Big(\frac{\gamma_i}{2} + \frac{1}{4}(\gamma_i\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big)\Big(\nu_i^s(t)+\frac{\gamma_i}{2}-\frac{1}{4}(\gamma_i+\frac{1}{t-1}\sum_{l=1}^{t-1}\nu_i^s(l))\Big)$. □

V. Renewable energies integration in the grid: a distributed learning game

Compared with Section IV and the literature [5], [9], in this section, the energy productions cannot be controlled anymore since it comes from renewable sources exclusively. Service providers optimize their prices and booking quantities at each time period, having no information about the productions and the energy needs of the microgrid at this instant. As a result, the game can be considered as having *partial information* [2]. Its study requires the introduction of performance measures captured here by loss functions, enabling the agents to evaluate the accuracy of their predictions.

We recall the definition of provider s_i's loss which has already been introduced in Subsections IV-A and IV-B:

$$l_i\Big(f(t),\nu(t)\Big) = \Big(\pi_i^0(t) - \pi_i(t)\Big)$$

where $\pi_i^0(t)$ corresponds to provider s_i's utility evaluated in $f_i(\nu_i^s,t) = \nu_i^s(t)$ and $f_i(\nu_k^e,t) = \nu_k^e(t),\ \forall k = 1,...,K$. It means that $\pi_i^0(t)$ contains the utility that provider s_i would have received if his forecasts were perfectly aligned with microgrid $\mathcal{M}_i$ instantaneous needs and with the production of each energy producer.

Having no a priori information about the dynamic evolution of the produced renewable energies and about the energy needs of the microgrids, we assume that everything happens as if the system were in the worst case: Nature and consumers allie together to form a *meta-player* who is supposed to be the most unfavorable to the service providers. It means that the meta-player tries to maximize the sum of the providers' losses. His loss can be expressed as the opposite of the sum of all the providers' losses. Therefore, it takes the form:

$$l\Big(f(t),\nu(t)\Big) = \sum_{i=1,...,n}\Big(\pi_i(t)-\pi_i^0(t)\Big)$$

The agents' external regret over the sequence of time periods $1,...,T$, is expressed as the realized difference between the cumulative loss and the loss of the best prediction i.e., pure strategy (in the sense that this prediction minimizes their cumulative loss).

To be more precise, for service provider s_i, it coincides with the difference between s_i's truly observed cumulative loss and the cumulative loss that would be obtained in case where s_i made the best constant prediction over time interval $[1;T]$. It takes the form:

$$\mathcal{R}_i(T) = \sum_{t=1}^{T} l_i\Big(f(t),\nu(t)\Big) - \min_{y\in\mathcal{X}_s\times\mathcal{X}_e^K}\sum_{t=1}^{T} l_i\Big(f_{-i}(y,t),\nu(t)\Big)$$

Finally, for the meta-player, the regret coincides with the difference between his cumulative loss and the loss of the constant predictions over $[1;T]$ about the unknown sequences minimizing his cumulative loss or equivalently maximizing the sum of the providers' losses over the interval. We have:

$$\mathcal{R}(T) = \sum_{t=1}^{T} l\Big(f(t),\nu(t)\Big) - \min_{z\in\mathcal{X}_s^n\times\mathcal{X}_e^K}\sum_{t=1}^{T} l\Big(f(t),z\Big)$$

The service providers and the meta-player try to determine randomized strategies such that asymptotically their external regrets remain in $o(T)$ where T is the number of time periods which have been played. It means that with probability 1:

$$\lim\sup_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T}\mathcal{R}_i(t) = 0$$

for provider $s_i, \forall i = 1,...,n$ and

$$\lim\sup_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T}\mathcal{R}(t) = 0$$

for the meta-player. Forecasters satisfying these inequalities are said *Hannan consistent* [2].

In the following lemma, we prove that it is possible to construct learning strategies for the service providers which minimize their external regret asymptotically.

Lemma 4: A Hannan consistent learning strategy exists for each service provider s_i.

Proof of Lemma 4. In our case setting, at the end of each time period, service provider s_i knows the energy quantity bought by his customers $\mathcal{M}_i$ and he can infer $\nu_i^s(t)$ from that quantity. s_i also knows the energy which has been delivered by each energy producer e_k to him. He can infer from that the energy which could have been delivered to him, if he had ordered a different quantity $q_{ik}(t)$, all other providers ordering the same energy quantities. As a result, s_i can calculate his loss

for all his possible actions. In [2], it is proved that a Hannan consistent learning strategy always exists when the player can calculate his loss for each possible action at the end of each time period. □

The repetition of the Stackelberg game introduced in Subsection II-A in a context of partial information can be rewritten by introducing randomization in the strategies. We denote by $d_t(f_i) : \mathcal{X}_s \times \mathcal{X}_e^K \to [0;1]$ and $d_t(\nu) : \mathcal{X}_s^n \times \mathcal{X}_e^K \to [0;1]$ the randomized strategies for service provider s_i and for the meta-player respectively at time period t. We then have to cope with a repeated learning game. At each time period t, the repeated game timing introduced in Subsection II-A is updated according to the following rules to incorporate the forecasting tasks of the providers:

(1) All the service providers $s_i,\ i = 1, ..., n$ make their forecasts $f_i(\nu_i^s, t),\ f_i(\nu_k^e, t),\ \forall k = 1, ..., K$ following distributions $d_t(f_i)$ respectively.
(2) Energy producers reveal their energy prices.
(3) Service producers reveal their energy orders $q_{ik}(t)$ and their service prices at the same time.
(4) *The meta-player chooses* $\nu_i^s(t)$ *and* $\nu_k^e(t),\ \forall i = 1, ..., n$ *and* $\forall k = 1, ..., K$ *following the distribution* $d_t(\nu)$.
(5) Each service provider s_i obtains his profit $\pi_i(t)$, the demand of $\mathcal{M}_i$ and the energy quantities offered by each service producer e_k.
Service providers update their forecasting strategies $d_t(f_i)$ and the meta-player updates his forecasting strategy $d_t(\nu)$ depending on the value of the expected utilities.

The step corresponding to the generation of unexpected random events resulting in microgrid energy needs and production variations is now controlled by the meta-player whereas the penalty rules introduced in Subsection II-A remain unchanged.

A. Regret based learning

We consider two types of updates for the forecasting randomized strategies $d_t(X)$ at each time period based on the exponential forecaster for signed games: one based on the external regret and the other based on the internal regret [2]. We will use the generic notation $\mathcal{X}$ to refer either to $\mathcal{X}_e$ or to $\mathcal{X}_s$. For a given forecast X, we derive the payoffs $H_X(x,t)$ for each value $x \in \mathcal{X}$ of the forecast at each time period t by going back to the utilities of the agents and by keeping only the terms depending on forecast X. We assume that this is a signed game because the range of values of payoff function $H_X(.)$ might include a neighborhood of 0. We let:

$$\begin{aligned}\mathcal{V}_t &= \sum_{s=1}^{t} Var\Big(H_X(X_s, s))\Big) \\ &= \sum_{s=1}^{t} \mathbb{E}\Big[\Big(H_X(X_s, s)) - \mathbb{E}[H_X(X_s, s)]\Big)^2\Big]\end{aligned}$$

be the sum of the variances associated with the random variable $H_X(X_t, t)$ which is the payoff for forecaster X at time period t assuming that the forecast at time period t has been set to X_t, under the mixed strategy X which is defined over space $\mathcal{X}$. Using the exponential forecaster for signed games with external regret means that the randomized strategy is updated according to the algorithm described below.

External Regret Learning Algorithm

Initialization. For $t = 0$, we set: $w_0(x) = \frac{1}{|\mathcal{X}|},\ \forall x \in \mathcal{X}$.
Step 1 *to* T. The updating rules are the following:

$$\begin{aligned}d_{t+1}(x) &= \frac{w_{t+1}(x)}{\sum_{x\in\mathcal{X}} w_{t+1}(x)},\ \forall x \in \mathcal{X} \\ w_{t+1}(x) &= \exp\Big(\eta_{t+1} \sum_{s=1}^{t} H_X(x,s)\Big) \\ &= d_t(x)^{\frac{\eta_{t+1}}{\eta_t}} \exp\Big(\eta_{t+1} H_X(x,t)\Big),\ \forall x \in \mathcal{X} \\ \eta_{t+1} &= \min\Big\{\frac{1}{2\max\{|H_X(.)|\}}; \sqrt{\frac{2(\sqrt{2}-1)}{e-2}}\sqrt{\frac{\ln|\mathcal{X}|}{\mathcal{V}_t}}\Big\} \\ \mathcal{V}_t &= \mathcal{V}_{t-1} + Var\Big(H_X(X_t, t)\Big)\end{aligned}$$

For the internal regret, it is similar but with $d_t(.) = \sum_{i\neq j} d_t^{i\to j}(.)\Delta_{(i,j)}(t)$ where $d_t^{i\to j}(.)$ is the modified forecasting strategy obtained when the forecaster predicts j each time he would have predicted i and $\Delta_{(i,j)}(t) = \frac{\omega_{(i,j)}(t)}{\sum_{k\neq l}\omega_{(k,l)}(t)}$ with: $\omega_{(i,j)}(t) = \exp\Big(\eta_t \sum_{s=1}^{t-1}\sum_{x\in\mathcal{X}} d_s(x) H_X(x,s)\Big)$.

We see that we need to compute the maximum of the absolute value of the payoff function $|H_X(.)|$ for all forecasts X to run a simulation of the game. This maximum is reached for $x = \min\{\mathcal{X}\}$ or $x = \max\{\mathcal{X}\}$ for all payoff functions except for $H_{f_i(\nu_i^s)}(.)$ because their differentiate with respect to x is never equal to 0. For $H_{f_i(\nu_i^s)}(.)$, the differentiate equals 0 if, and only if, $f_i(\nu_i^s, t) = \nu_i^s(t)$, so the maximum of $|H_X(.)|$ is reached either for $x = \min\{\mathcal{X}\}$ or $x = \max\{\mathcal{X}\}$ or $x = \nu_i^s(t)$.

B. Collaborative learning is better

In this subsection, we want to know how distributed learning by service providers introducing (in)voluntary biases in their predictions, affects the smart grid global performance. Agents will most probably exchange information concerning their forecasts. Some agents might appear as more credible than others and coalitions might emerge. Collaboration will then take place within coalitions. In cooperative game theory literature, a coalition is a group of agents who have incentives to collaborate by sharing resource access, information, etc., in the hope to increase their revenue, knowledge, social welfare (in case of altruism), etc., compared to the case where they behave non-cooperatively [10], [15]. Adapted to our hierarchical learning context, we define coalitions of agents as follows:

- A coalition of agents is a group of agents who share their information and align their predictions to a common value.
- Agents who belong to the same coalition are said to collaborate.

We prove in [5] that under external regret minimization, the smallest upper-bound on the agents' average loss over time interval $[1;T]$ is reached when the service providers integrate a grand coalition. As a result, this economic organization will remain stable provided the agents have incentives to consider the optimization of the whole system performance.

The objective of the next part of this subsection is to test on a toy network made of 2 producers and 3 providers that collaborative learning through a grand coalition provides better guarantees on the smart grid global performance than full competition.

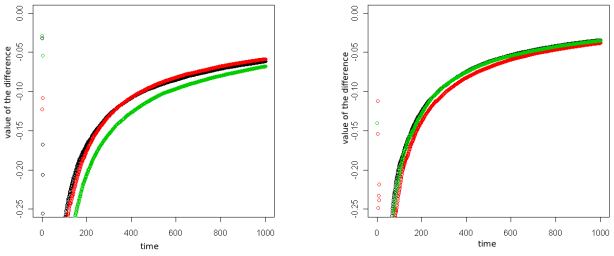
For our numerical illustration, we have chosen $n = 3$ and $K = 2$. We have also used $\gamma_1 = \gamma_2 = \gamma_3 = 0.9$ and $\tilde{\gamma_1} = 0.5$, $\tilde{\gamma_2} = 0.4$, $\tilde{\gamma_3} = 0.6$ and $\mathcal{X}_e = [1;2]$ for the producers, $\mathcal{X}_s = [5;8]$ for the providers which ensure that we are always in energy shortage (cf. Case 2 described in Subsection B.2 of Proposition 1 proof) i.e., that penalties are imposed to the providers.

In the following pictures, we compare the cumulative regret of each agent to the cumulative regret of the same agent who would have forecasted the best value at each time period in terms of payoffs. More precisely, we display:

$$\frac{1}{t}\sum_{s=1}^{t}\sum_{X\in F}\Big(H_X(X_s,s)-\max_x(H_X(x,s))\Big)$$

where F is the generic set of forecasts made by the service provider or the meta-player or the considered coalition.

We start by comparing the cumulative internal and external regrets in the case of full competition between service providers in Figures 2(a) and 2(b).

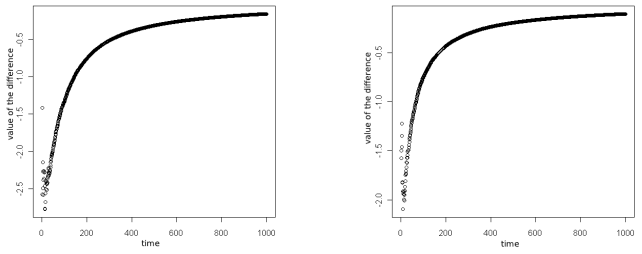


(a) (b)

Fig. 2. Difference between the best achievable cumulative regret and the one obtained with the internal regret minimization algorithm in (a) and with the external regret minimization algorithm in (b) under full competition.

The service providers are in black for s_1, green for s_2 and red for s_3. We can see that in all cases, the difference between regrets converge toward 0 which means that the cumulative payoff obtained at the end of the game following the exponential forecaster strategy is close to the best possible cumulative payoff. This is in coherence with the theoretical result for the internal regret but is better than what we could expect for the external regret which means that we are in a game setting which performs well for regret based learning. We also remark that the algorithm converges faster for the external regret compared to the internal regret.

We compare these graphs with the graphs obtained when service providers integrate a grand coalition in Figures 3(a) and 3(b).



(a) (b)

Fig. 3. Difference between the best achievable cumulative regret and the one obtained with the internal regret minimization algorithm in (a) and with the external regret minimization algorithm in (b) for the grand coalition of service providers.

Again, we observe that the differences between the best achievable regrets and those obtained converge toward 0. The rate of convergence under cooperative learning seems higher than in the non-cooperative case. In addition, we observe that after 400 time periods the sum of differences between regrets under collaborative learning is close to -0.2 whereas the sum of differences between regrets is close to -0.26 in the full competition case. This is in coherence with the theory which says that collaborative learning is better.

VI. Conclusion

In this article, we focus on the identification of distributed learning algorithms enabling the agents to forecast hidden individual sequences in a decentralized hierarchical network which typical illustration is the smart grid. The upper-level agents are called the producers. They provide energy to the lower-level agents. The lower-level agents are called the providers. They provide energy access to the end users. The problem at hand is quite different depending on whether the produced energy is renewable or not. In case where it is non-renewable, the producers control their productions so as to maximize their utility. The providers can also adapt their prices, so as to force the end users to consume as much as they want. However, we prove that their utility is maximized if, and only if, their prediction coincides with the true value of the end users' energy needs which is not observed at the beginning of the time period. As a result, providers need to perform learning to optimize their utility. The performance of two well-known learning algorithms (tit for tat and fictitious play) are derived analytically. When renewable energies are introduced in the smart grid, the energy productions become highly uncertain adding more forecasts to perform. This requires to develop more sophisticated learning algorithms. We propose an online learning algorithm based on regret minimization and analyze

its performance. Furthermore, we show through simulation that collaborative learning generates higher performance for the whole smart grid than full competition.

A possible extension of the work initiated in this paper might be to determine which economic organizations remain stable in case of unexpected attacks i.e., to design a *resilient* system.

APPENDIX

Proof of Proposition 1

The game setting described in Subsection II-A implies that in the relationship producers-providers, producers appear as leaders whereas providers are followers. Identically, in the relationship providers-consumers, providers appear as leaders whereas consumers are mere followers. Under such a setting, the game is called a Stackelberg game and as usual, it should be solved using backward induction [10].

We make the assumption that each energy producer receives at least one energy order from a service provider guaranteeing that the Stackelberg game admits non trivial solutions.

A. Optimization of the end users' decision

To minimize their total cost of energy defined by Equation (1), end users $\mathcal{M}_i$ have to choose $a_i(t)$ so that the differentiate of the total cost of energy equals 0 which means:

$$a_i(t) = p_i(t) \tag{7}$$

We will assume that $p_i(t) < \nu_i^s(t)$ to ensure that this value for $a_i(t)$ does not exceed the energy needs of microgrid $\mathcal{M}_i$. Then the optimal $a_i(t)$ is defined by Equation (7).

B. Optimization of the service providers' decisions

To find his optimal price and energy orders, service provider s_i has to replace $a_i(t)$ by its optimal value in $\pi_i(t)$ defined in Equation (2), and to differentiate the result in $p_i(t)$ and in $q_{ik}(t)$. This differentiation raises two cases.

B.1 Case 1: the energy production fulfills the energy demand of the end users

It is the case when:

$$\nu_i^s(t) - p_i(t) \leq \sum_{k=1,\ldots,K} \alpha_{ki}(t)\nu_k^e(t) \tag{8}$$

Then differentiating the service provider's utility in $q_{ik}(t)$ leads to:

$$\frac{\partial \pi_i(t)}{\partial q_{ik}(t)} = -\tilde{p_k}(t)$$

which means that s_i will try to minimize all his energy orders to maximize his utility. As a result, s_i will tend to break the inequality defining Case 1 in Inequality (8) because $\alpha_{ki}(t)$ will tend toward zero. As a result the optimal decision for s_i will always fall in Case 2 described below or on the frontier between Case 1 and Case 2. The frontier between these two cases is defined by the equation:

$$\nu_i^s(t) - p_i(t) = \sum_{k=1,\ldots,K} \alpha_{ki}(t)\nu_k^e(t) \tag{9}$$

B.2 Case 2: the energy production does not fulfill the energy demand of the end users

It is the case when $\nu_i^s(t) - p_i(t) \geq \sum_{k=1,\ldots,K} \alpha_{ki}(t)\nu_k^e(t)$. Then differentiating s_i's utility gives us:

$$\begin{aligned} \frac{\partial \pi_i(t)}{\partial p_i(t)} &= \nu_i^s(t) + \gamma_i - 2p_i(t) \\ \frac{\partial \pi_i(t)}{\partial q_{ik}(t)} &= -\tilde{p_k}(t) + \gamma_i \nu_k^e(t) \frac{\partial \alpha_{ki}(t)}{\partial q_{ik}(t)} \end{aligned} \tag{10}$$

By using the definition of $\alpha_{ki}(t)$ given in Equation (3), we obtain:

$$\frac{\partial \alpha_{ki}(t)}{\partial q_{ik}(t)} = \tilde{\gamma_i} \frac{C_k(t) - \tilde{\gamma_i} q_{ik}(t)}{C_k(t)^2}$$

where we have let: $C_k(t) = \sum_{j=1,\ldots,n} \tilde{\gamma_j} q_{jk}(t)$. Then going back to System of equations (10), we conclude that the differentiates equal 0 when:

$$p_i(t) = \frac{\nu_i^s(t) + \gamma_i}{2} \tag{11}$$

$$\tilde{p_k}(t) C_k(t)^2 = \gamma_i \nu_k^e(t) \tilde{\gamma_i} \Big(C_k(t) - \tilde{\gamma_i} q_{ik}(t) \Big) \tag{12}$$

On one side, we obtain directly the price for which the differentiate of $\pi_i(t)$ equals 0 through Equation (11). On the other side, Equation (12) can be rewritten as follows:

$$\tilde{\gamma_i} q_{ik}(t) = C_k(t) - \frac{\tilde{p_k}(t) C_k(t)^2}{\nu_k^e(t) \gamma_i \tilde{\gamma_i}} \tag{13}$$

If s_i anticipates that the other service providers will make the same optimization program, replicating Equation (13) for the n service providers and summing them all, results in the following equality:

$$C_k(t) = nC_k(t) - \frac{\tilde{p_k}(t) C_k(t)^2}{\nu_k^e(t)} \sum_{j=1,\ldots,n} \frac{1}{\gamma_j \tilde{\gamma_j}}$$

by definition of $C_k(t)$.

Then as $C_k(t)$ is not zero because each producer e_k receives at least one order of energy, by dividing the previous equation by $C_k(t)$ and reordering we obtain:

$$C_k(t) = \frac{\nu_k^e(t)}{\tilde{p_k}(t)} \frac{n-1}{\delta}$$

where $\delta = \sum_{j=1,\ldots,n} \frac{1}{\gamma_j \tilde{\gamma_j}}$. By replacing $C_k(t)$ in Equation (13), we obtain the energy orders for which the differentiates of $\pi_i(t)$ equals 0:

$$q_{ik}(t) = \frac{\nu_k^e(t)}{\tilde{p_k}(t)} \frac{n-1}{\delta \tilde{\gamma_i}} L(i) \tag{14}$$

where we have introduced the notation $L(i) = 1 - \frac{n-1}{\delta \gamma_i \tilde{\gamma_i}}$ to simplify future calculations.

Presently, we have to check that the price and energy orders for which the differentiates of $\pi_i(t)$ equal 0 satisfy the conditions of Case 2.

First, it is easy to check that the price is positive through Equation (11). However, the energy orders defined in Equation (14) are non-negative if, and only if, $1 \geq \frac{n-1}{\delta\gamma_i\tilde{\gamma_i}}$ which is equivalent to:

$$\gamma_i\tilde{\gamma_i} \geq \frac{n-1}{\delta} \tag{15}$$

This inequality means that the penalties related to s_i are close to the penalties related to the other service providers. Indeed, if all penalties are equal to γ, then $\delta = \frac{n}{\gamma^2}$ and Inequality (15) is true for all service providers. On the contrary, if all penalties are equal to γ except for s_1 which has a penalty of $\frac{\gamma}{n-1}$, then $\delta = \frac{(n-1)n}{\gamma^2}$ and Inequality (15) becomes $n \geq (n-1)^2$ which is false as soon as $n > 2$.

Second, by replacing the energy orders defined by Equation (14) in Equation (3), we obtain $\alpha_{ki}(t) = \frac{L(i)}{\sum_{j=1,\dots,n} L(j)} = L(i)$ meaning that the total energy delivered to the customers of s_i is $\sum_{k=1,\dots,K} \alpha_{ki}(t)\nu_k^e(t) = L(i)\sum_{k=1,\dots,K} \nu_k^e(t)$. As a result, the price and energy orders for which the differentiates of $\pi_i(t)$ equal 0 verify the inequality defining Case 2 if, and only if:

$$\nu_i^s(t) \geq \gamma_i + 2L(i) \sum_{k=1,\dots,K} \nu_k^e(t) \tag{16}$$

This inequality states that the total production of energy by energy producers should not be too large compared to the energy needs of customers.

If Inequalities (15) and (16) are true, the optimum for s_i is reached for $p_i(t)$ defined by Equation (11) and $q_{ik}(t)$ defined by Equation (14). If one of these inequalities is not true, then the optimum for s_i is reached on the frontier defined by Equation (9).

C. **Optimization of the energy producers' decision**

After substituting $q_{ik}(t)$ and $C_k(t)$ by the expressions found in the previous section in energy producer e_k's utility as defined in Equation (4), we obtain:

$$\tilde{\pi_k}(t) = \nu_k^e(t)\frac{n-1}{\delta}\Big(\sum_{i=1,\dots,n} \frac{L(i)}{\tilde{\gamma_i}} - \sum_{i=1,\dots,n} \Big(\frac{L(i)}{\tilde{p_k}(t)} (1 - \frac{\tilde{p_k}(t)\tilde{\gamma_i}\delta}{n-1})_+\Big)\Big)$$

The only part of this equation depending on $\tilde{p}_k(t)$ has always a negative impact on the profit of the energy producer under the assumption of fair penalties. Indeed, in that case, as raised in the previous section, we have: $L(i) \geq 0$ for all service providers s_i. As a result, to maximize his profit, the energy producer has to choose $\tilde{p_k}(t)$ such that the part depending on $\tilde{p}_k(t)$ in the above equation equals 0. It implies that the term $1 - \frac{\tilde{p_k}(t)\tilde{\gamma_i}\delta}{n-1}$ is inferior to 0 for all $i = 1, ..., n$. It is equivalent to: $\tilde{p_k}(t) \geq \frac{n-1}{\delta\tilde{\gamma_i}}$. Consequently, the optimal price for the energy producer with fair penalties is defined by:

$$\tilde{p_k}(t) = \frac{n-1}{\delta \min_{i=1,\dots,n}\{\tilde{\gamma_i}\}} \tag{17}$$

In theory, the price could be higher than this value and it would change nothing for the utility of the energy producer. But the energy producer has an incentive to be moderate on his price to avoid competition from other energy producers. □

REFERENCES

[1] Bubeck S., Online Optimization, Lecture notes, Princeton university, Department of Operations Research and Financial Engineering, 2012

[2] Cesa-Bianchi N., Lugosi G., Prediction, Learning, And Games, Cambridge university press, 2006

[3] Fudenberg D., Levine D. K., The Theory of Learning in Games, the MIT Press, 1998

[4] de Ladoucette P., Chevalier J.-M., The electricity of the future: a mondial challenge, Economica, 2010

[5] Le Cadre H., Bedo J.-S., Collaborative Learning is Better, working paper of the Alternative Energy and Nuclear Power Commission, 2012

[6] Le Cadre H., Potarusov R., Auliac C., Energy Demand Prediction: A Partial Information Game Approach, in proc. EEVC 2011

[7] Li N., Marden J. R., Decoupling coupled constraints through utility design, Discussion paper, Department of ECEE, university of Colorado, Boulder, 2011

[8] Li H., Zhang W., QoS routing in smart grids, in proc. IEEE global communication conference, 2010

[9] Marden J. R., Young H. P., Pao L. Y., Achieving Pareto Optimality Through Distributed Learning, Oxford Economics discussion paper n^{o}557, 2011

[10] Myerson R., Game Theory: An Analysis of Conflict, Harvard university press, 2006

[11] Nguyen T., Vojnović M., Weighted Proportional Allocation, in proc. of ACM Sigmetrics 2011

[12] Pradelski B. R., Young H. P., Learning efficient Nash equilibria in distributed systems, discussion paper, Department of Economics, university of Oxford, 2010

[13] Saad W., Han Z., Debbah M., Hjorrungnes A., Başar T., Coalitional Game Theory for Communication Networks: A Tutorial, IEEE Signal Processing Magaeinr, Special Issue on Game Theory, vol.26, pp.77–97, 2009

[14] Saad W., Han Z., Poor V. H., Coalitional Game Theory for Cooperative Micro-Grid Distribution Networks, 2-nd IEEE International Workshop on Smart Grid Communications, 2011

[15] Shapley L. S., Stochastic games, in proc. of the National Academy of Sciences of the United States of America, vol.39, pp.1095–1100, 1953

[16] Talluri K. T., van Ryzin G. J., The Theory and Practice of Revenue Management, International Series in Operations Research & Management Science, vol.68, 2005

[17] Young H. P., Learning by trial and error, Games and Economic Behavior, vol.65, pp.626–643, 2009

[18] Young P. Y., Strategic Learning and its limits, the Arne Ryde Memorial Lecture Series, Oxford university press, 2004