games

MDPI

Article

Boltzmann Distributed Replicator Dynamics: Population Games in a Microgrid Context

Gustavo Chica-Pedraza [1,*], Eduardo Mojica-Nava [2] and Ernesto Cadena-Muñoz [3]

[1] School of Telecommunications Engineering, Universidad Santo Tomás, 110311 Bogotá D.C., Colombia
[2] Department of Electric and Electronic Engineering, Universidad Nacional de Colombia, 111321 Bogotá D.C., Colombia; eamojican@unal.edu.co
[3] Department of Systems and Industrial Engineering, Universidad Nacional de Colombia, 111321 Bogotá D.C., Colombia; ecadenam@unal.edu.co
* Correspondence: gustavochica@usantotomas.edu.co

Abstract: Multi-Agent Systems (MAS) have been used to solve several optimization problems in control systems. MAS allow understanding the interactions between agents and the complexity of the system, thus generating functional models that are closer to reality. However, these approaches assume that information between agents is always available, which means the employment of a full-information model. Some tendencies have been growing in importance to tackle scenarios where information constraints are relevant issues. In this sense, game theory approaches appear as a useful technique that use a strategy concept to analyze the interactions of the agents and achieve the maximization of agent outcomes. In this paper, we propose a distributed control method of learning that allows analyzing the effect of the exploration concept in MAS. The dynamics obtained use Q-learning from reinforcement learning as a way to include the concept of exploration into the classic exploration-less Replicator Dynamics equation. Then, the Boltzmann distribution is used to introduce the Boltzmann-Based Distributed Replicator Dynamics as a tool for controlling agents behaviors. This distributed approach can be used in several engineering applications, where communications constraints between agents are considered. The behavior of the proposed method is analyzed using a smart grid application for validation purposes. Results show that despite the lack of full information of the system, by controlling some parameters of the method, it has similar behavior to the traditional centralized approaches.

Keywords: Boltzmann distribution; economic dispatch problem; Multi-Agent Systems; population dynamics; revision protocol

Citation: Chica-Pedraza, G.; Mojica-Nava, E.; Cadena-Muñoz, E. Boltzmann Distributed Replicator Dynamics: Population Games in a Microgrid Context. *Games* **2021**, *12*, 8. https://doi.org/10.3390/g12010008

Received: 30 October 2020
Accepted: 4 January 2021
Published: 15 January 2021

1. Introduction

The study of large-scale control distributed systems has been the focus of scientists in recent decades. Several models and techniques are used to deal with challenges related to this area such as the high cost in terms of computational requirements, the communication structure, and the estimation of the data needed to perform a task in large-scale systems. Multi-Agent Systems (MAS) is a tool for addressing this kind of problems and is usually employed in the framework of game theory. In this context, the study of interactions of agents have received special attention due to the use of strategies that allow agents maximizing their outcomes. In this regard, the work in [1] provides connections among games, learning, and optimization in networks. Other works are focused on games and learning [2] or studied algorithms for distributed computation in dynamic networks topologies [3]. In [4], authors analyzed applications of power control using both distributed and centralized game theory frameworks. Applications of smart grid control are found in [5]. Other works focused on cases in which coordination and negotiation problems lead to the analysis of agents interactions [6]. Other approaches related to applications in power control based on game theory are found in [7].

The literature of game theory differentiates between three kinds of games. Continuous games where the concept of pure strategy is stated for agents and one agent can have many of it. Matrix games where agents are individually treated and allowed to take only one shot play simultaneously. The last case is related to dynamic games, which assume players can learn somehow about the actions and the states of the environment. This feature implies an agent can learn about the taken decision to adjust its behavior [8]. This kind of games has to respond to the following challenges: the modeling of the environment where the interaction between players takes place, the modeling of players' goals, and finally, how to prioritize players actions and how to estimate the amount of information a player has [9].

Evolutionary game theory (EGT) introduces an approach that belongs to dynamic games related to the study of dynamics of agents that change over time [10]. The evolutionary stable strategy was the concept that made EGT popular due to its relationship with biology and the way it supports the understanding of natural behaviors [11]. For this reason, the use of EGT is traditionally assumed to be positive (descriptive), rather than normative (prescriptive). Nonetheless, when EGT is used in the study of particular empirical phenomena to predict some behavior, sometimes it appears to be more on the normative side [12]. In this regard, the replicator dynamics (RD) approach is based on the understanding of EGT, which has been used in several real life control applications. The combination of revision protocols (which gives a description on how agents select and change strategies), and population games (which specifies interactions among agents) leads to the concept of evolutionary game dynamics [11]. The process of modeling large-scale systems is usually developed using this evolution perspective, since its mathematical background allows describing this process by using differential equations [10].

Applications of EGT can be found in many fields of Engineering, some of them focused on the economic dispatch problem (EDP) of microgrids for electric generators, control of systems of communications access, optimization problems, among others [8]. Some of the advantages of modeling engineering problems using the EGT approach are the easiness to associate a game with a problem of the engineering field, where the objective function and the strategies can be stated as payoff functions, and the relationship between the concepts of optimization and Nash Equilibrium, which is possible under specific conditions that allow satisfying the first order optimization conditions of the Karush–Kuhn–Tucker. Finally, it is worth noting that EGT obtains games solutions using local information. In this sense, when addressing engineering problems, distributed approaches emerge, which is relevant due to the cost of implementation of centralized schemes and its high complexity [8]. Distributed population dynamics has some advantages in comparison with some techniques such as the dual decomposition method, which needs a centralized coordinator [13]. This property minimizes costs related to communication structure. Moreover, compared with distributed algorithms for learning in normal-form-games, distributed population dynamics have no fails in applications that use constraints over all the variables immersed in the decision process [14]. This feature fits properly for dealing with problems related to resource allocation cases such as the smart city design [15]. In this kind of applications, there is an increasing need for integration between the digital technologies and the distributed power generation to make electric networks robust, flexible, reliable and efficient. However, the actual model of a centralized grid must be changed to a distributed model based on microgrids [16]. In this regard, microgrids consider control operations individually, as they separate the economic dispatch, the power generation and the secondary frequency [17]. The economic dispatch action is managed with static optimization concepts [18] or even methods such as the direct search operating off-line [19]. When the analysis includes the generator, loads or power line losses into the distributed model, it is more difficult to analyze the problem. Other approaches are unable to analyze the dynamic conditions such as the time dependence of the economic dispatch [20]. Several approaches have proposed to deal with these challenges. For example, in [21] is proposed a microgrid centralized energy management system with an operation in stand-alone mode that analyzes the static behavior. Other approaches work with a distributed control strategy based on power line signaling used for energy storage systems [22]. Some applications of the MAS technique for distributed economic problems are found in [23], whose contribution was to consider the delays of the communication system. Other authors propose microgrid architectures based on distributed systems such as the microgrid hierarchical control [24].

This paper proposes an approximation to tackle some of the challenges found in the revised literature. Our main objective is to introduce a more realistic control method of learning that allows analyzing the effect of the exploration concept in MAS, i.e., communication between agents. Since the RD has proven to emerge from simple learning models [25], this work uses the Q-learning dynamics from the reinforcement learning theory as a way to include the concept of exploration into the classic exploration-less RD equation. Therefore, by merging these frameworks, it is feasible to find a way to deal with dynamics immersed in the environment where the feedback of each agent depends not only on itself but also on other agents, and where communications constraints between them are part of the system. To make the analysis, the Boltzmann distribution is used to introduce a distributed version of the RD as a tool for controlling the behavior of agents under specific conditions. In this regard, the obtained method uses a temperature parameter and an extra term after the derivation process, which allow adjusting the behavior of the learning agents and to give a connection between the selection-mutation mechanism from EGT and the exploitation-exploration scheme from RL. Although this feature meets the traditional positive condition of EGT techniques (i.e., models what happens when agents interact), its applications to the control area should be considered normative rather than positive, since the tuning factor can be used to meet the desired behavior beforehand. To understand these features, this approach attempts to explain agents decision-making using experimental data to deal with the economic dispatch problem, which is one of the common issues in a smart grid. The results of this solution are compared with the classic centralized approach of RD.

The rest of this paper is organized as follows. In Section 2, a brief summary of game theory and Q-learning is introduced, and the replicator dynamics equations are presented from the perspective of EGT. Section 3 describes a distributed neighboring approach for the Boltzmann control method, based on the replicator dynamics behavior. Section 4 discusses some relevant concepts obtained in Section 3 regarding the domains of reinforcement learning and the evolutionary game theory. Section 5 introduces a smart grid application based on an economic dispatch problem, where the distributed approach of the Boltzmann control method is applied. Finally, some conclusions are outlined in Section 6.

2. Theoretical Framework

Game theory presents a set of mathematical equations to analyze the background in decentralized control issues. A game is usually formed by a set of agents (players) with the same population behavior that select a strategy to perform an action. The strategy of an agent allows it to maximize rewards in cases where the action selected was appropriate or may result in punishment if it was not [26]. This behavior is analyzed using theory of learning. In this context, the reinforcement learning (RL) scheme gives the main features to understand the direct relationship between signals, actions, states, and the environment. At each step of their interaction, each player receives a notification of the present environment state and a reinforcement signal, and consequently, it selects a strategy. Each player of the game has the purpose of finding a policy that provides him the best profits after mapping states to actions [27]. Figure 1 shows the RL framework, where an agent at time t takes an action a_t, depending on the information perceived in the environment state S_t. Then, the environment changes to state s_{t+1} and, consequently, the agent perceives a consequence of its taken actions r_{t+1}, e.g., reward or punishment.

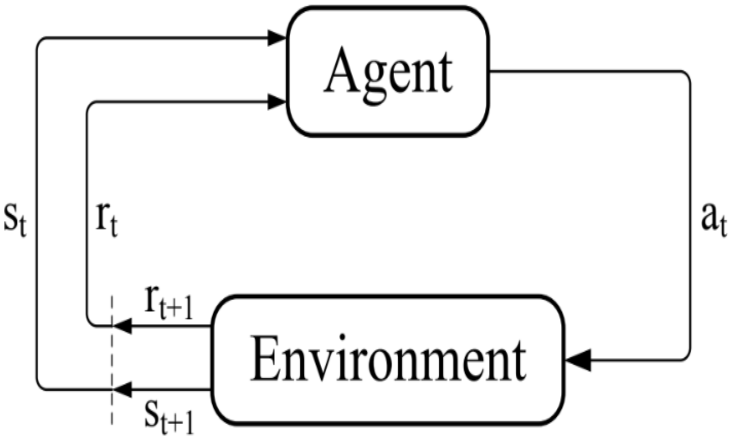


Figure 1. RL framework. Adapted from [25].

Traditional RL methods are characterized by a structure of estimated value functions [27]. A value of a state or a pair state-action is the total amount of reward an agent can achieve, starting from that state. This implies that the optimal value function has to be found to know the policy that best satisfies payoffs. This process can be done with the value iteration algorithm and the Markov decision process [28] in cases where the environment is well-known. In other cases, Q-learning positions itself as an adaptive value iteration method, which needs no specific model of the environment. Equation (1) describes how the Q-learning interaction is done [29]:

$$Q_{t+1}(s,a) \leftarrow (1-\alpha)Q_t(s,a) + \alpha(R + \gamma max_{a'} Q_t(s',a')) \tag{1}$$

The estimation starts in $Q_{t+1}(s,a)$ taking $t+1$ as time value, and reaches the $Q_t(s',a')$, where s' is the most recent value of a player state s after executing action a. $max_{a'}$ receives the highest Q Value from s' by choosing the action that maximizes the Q Value. The term α describes the common step size parameter, R represents the immediate reinforcement, and γ represents a discount factor.

In cases where agents have full access to information of the game and no communication constraints, classical theory of games together with theory of learning are useful tools to model centralized control applications. However, these cases are a close representation of ideal situations, which have some limitations when representing real life conditions, interactions, and rationality of individuals. In this sense, Evolutionary Game Theory (EGT) intends to relax the concept of rationality, by replacing it with biological concepts and behaviors such as natural selection and mutation [30]. The fundamental notions of EGT came from the evolutionary biology field [31,32]. In this scope, the strategies of the players are genetically encoded and called genotypes, which refer to the behavior of each player used to compute its payoff. Each player genotype outcome is conditioned by the number of other types of players in the environment.

EGT makes that the population strategies start to evolve through the construction of a dynamic process, where the Replicator Dynamics equation represent the expected value of this procedure. The context of an evolutionary system usually refers to two concepts: mutation and selection. Mutation gives variety to the population, while selection is related to prioritize some varieties in which each genotype represents a pure strategy $Q_j(n)$. All the replicator dynamics offspring inherits this behavior. Equation (2) shows the general equation of the RD [25]:

$$\frac{dx_i}{dt} = [(Ax)_i - x \cdot Ax]x_i \tag{2}$$

where x_i represents the amount of density of a specific population playing strategy i, and the payoff matrix is represented by A. This matrix has different payoff values that each replicator gets from the other players. The population state (x) is usually described by the vector of probability $x = (x_1, x_2, ..., x_J)$, which shows the different amount of densities of each type of replicator in the population. Therefore, $(Ax)_i$ represents the payoff that the player i obtains in a population with x state. In this case, the average payoff will be represented as $x \cdot Ax$. The growth rate $\frac{dx_i}{dt}$ of the whole population using strategy i equals the difference between the current strategy payoff and the average payoff in the population [31].

Relating EGT and Q-Learning

The authors in [33] considered a way to associate the schemes of Q-learning and RD in the context of a two-players game, in which agents play different strategies. This association is possible considering that players are Q-learners. To model this scenario, two systems of differential equations are stated: one for player Q (columns) and another for player P (rows). In case $A = B^t$, the general Equation (2) for RD is used, in which x_i is replaced by p_i or q_i. The payoff matrix is stated by A or B, and the population state (x) is represented by p or q, based on the behavior of the player to be modeled. Thus, $(Ax)_i$ changes to $(Aq)_i$ or $(Bp)_i$ and represents the payoff that the player i obtains in a population with p or q state. Similarly, the growth rate $\frac{dx_i}{dt}$ changes to $\frac{dp_i}{dt}$ or $\frac{dq_i}{dt}$ for player Q and player P, respectively. The foregoing can be described by the following differential equations system [25]:

$$\frac{dp_i}{dt} = [(Aq)_i - p \cdot Aq]p_i \tag{3}$$

$$\frac{dq_i}{dt} = [(Bp)_i - q \cdot Bp]q_i \tag{4}$$

Equations (3) and (4) represent the RD equations for two populations. The growth rate of each population depends on the other populations, where A and B represent two necessary payoff matrices to calculate the rate of change from these differential equation systems for the two different current players in the problem.

To see the relationship of the RD equations with the Q-learning approach, first, it is necessary to introduce Equation (5), which describes the Boltzmann distribution:

$$x_i(\delta) = \frac{e^{\tau Q_{a_i}(\delta)}}{\sum_{j=1}^{n} e^{\tau Q_{a_j}(\delta)}} \tag{5}$$

where the probability of playing the i strategy at δ step time is represented by $x_i(\delta)$, and τ represents the temperature. The authors in [33] used Equation (5) to derive the Q-Learning continuous time model for a two-players game, where $\frac{dx_i}{dt}$ is taken as $\dot{x}_i$. This result is shown in Equation (6).

$$\frac{\dot{x}_i}{x_i} = \tau \Big[\frac{dQ_{a_i}}{dt} - \sum_{j=1}^{n} \frac{dQ_{a_i}}{dt} x_j \Big] \tag{6}$$

In order to solve the expression $\frac{dQ_{a_i(t)}}{dt}$ in Equation (6), the Equation (1) is used to model the update rule for the first Q-learner player, which yields an expression that describes the difference equation for the Q-function as shown in Equation (7):

$$\Delta Q_{a_i}(\delta) = \alpha \Big[R_{a_i}(\delta + 1) + \gamma \max Q - Q_{a_i}(\delta) \Big] \tag{7}$$

The time amount spent between two repetitions of the Q-values updates is stated by σ with $0 < \sigma \leq 1$, and $Q_{a_i}(\delta\sigma)$ represents the Q-values at time $k\sigma$. Therefore, if this equation takes an infinitesimal scheme, after taking the limit $\sigma \to 0$, Equation (7) gets to Equation (8).

$$\frac{\dot{x}_i}{x_i} = \tau\alpha\Big[R_{a_i} - \sum_{j=1}^{n} x_j R_{a_j} + \sum_{j=1}^{n} x_j (Q_{a_j} - Q_{a_i})\Big] \tag{8}$$

Since $\frac{x_j}{x_i}$ equals $\frac{e^{\tau}\Delta Q_{a_j}}{e^{\tau}\Delta Q_{a_i}}$, the second part of Equation (8) can be expressed in logarithm terms:

$$\alpha\Big[\tau \sum_j x_j (Q_{a_j} - Q_{a_i})\Big] = \alpha\Big[\sum_{j=1}^{n} x_j \ln\Big(\frac{x_j}{x_i}\Big)\Big] \tag{9}$$

After reorganizing and substituting the last expression in Equation (8), it takes the form of Equation (10).

$$\frac{\dot{x}_i}{x_i} = \alpha\tau\Big[R_{a_i} - \sum_{j=1}^{n} x_j R_{a_j}\Big] + \alpha\Big[\sum_{j=1}^{n} x_j \ln\Big(\frac{x_j}{x_i}\Big)\Big] \tag{10}$$

To introduce the use of the payoff matrices in a two player games, it is possible to write R_{a_i} as $\sum_j a_{ij} q_j$, where the expression for the first player is:

$$\dot{x}_i = x_i \alpha\tau\Big[(Aq)_i - x \cdot Aq\Big] + x_i \alpha\Big[\sum_{j=1}^{n} x_j \ln\Big(\frac{x_j}{x_i}\Big)\Big] \tag{11}$$

Equation (11) is the representation of the derivation of the continuous time model for Q-learning developed by [33], which uses the Boltzmann concept. This expression can be understood as a centralized approach, very similar to the original form of Equation (3), which is the classic RD form to model the behavior of player P, in the context of a two-player game. The main differences are influenced by the Boltzmann model with the introduction of α and τ constants, and the appearance of the second term. Applications of this approach have been developed in the context of 2×2 games, multiple state, and multiple player games [25]. However, its application to solve real life problems is still an open issue.

The next Section introduces our proposal, which can be understood as a distributed control method of learning. This novel approach uses part of the bases proposed in [33]. Our main contribution is the introduction of a Boltzmann distributed perspective, in which the concept of population dynamics from EGT is considered and where players can only play neighboring strategies. This control method considers communications constraints among agents, i.e., a context where agents have incomplete information of the system.

3. The Proposed Boltzmann Distributed Control Method

In this section, the proposed Boltzmann distributed replicator dynamics method is presented. From the perspective of the classic replicator dynamics, we use Equation (11) as the depart point. This expression can be used not only in a two-players game, but in multi-players games. Consequently, Equation (11) can be stated as shown in Equation (12):

$$\dot{x}_i = \alpha x_i \tau\Big[f_i(x) - \overline{f}(x)\Big] + \alpha x_i\Big[\sum_{j=1}^{n} x_j \ln\Big(\frac{x_j}{x_i}\Big)\Big], \tag{12}$$

where a portion of a particular type of population will increase or decrease if its individuals have a higher/lower fitness than the population average. The state vector $x = (x_1, x_2, ..., x_n)^n$ with $0 \leq x_i \leq 1, \forall_i$ and $\sum_{i=1}^{n} x_i = 1$ is used to describe the population and represents the fractions belonging to each of n types. Consider the fitness function $f_i(x)$, where i represents the fitness type. Then, the population fitness average is described by $f(x) = \sum_j x_j f_j(x)$. By applying the last assumptions, this expression changes to

$$\dot{x}_i = \alpha x_i \tau\Big[f_i(x) - \sum_{j=1}^{n} x_j f_j(x)\Big] + \alpha x_i\Big[\sum_{j=1}^{n} x_j \ln\Big(\frac{x_j}{x_i}\Big)\Big]. \tag{13}$$

Considering that the first term of Equation (14) takes the form of the centralized equation for the RD, this work proposes its adaptation to a decentralized form, looking forward to computing the agents local information to deal with communications constraints. Equation (14) shows the decentralized form of this stage as follows:

$$\dot{x}_i = \underbrace{\alpha x_i \tau \Big[f_i(x) - \sum_{j=1}^{n} x_j f_j(x) \Big]}_{\textit{Centralized}} = \underbrace{\alpha x_i \tau \Big[f_i(x) \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} x_j f_j(x) \Big]}_{\textit{Decentralized}} \tag{14}$$

where the expression $\sum_{j=1}^{n} x_j$ is equal to the unit. This is explained because the term x_j within the sum represents the probabilities of playing the jth strategy. Similarly, for the second term in Equation (13), after doing some Algebra and applying logarithms rules, this expression takes the following form:

$$\underbrace{\alpha x_i \Big[\sum_{j=1}^{n} x_j \ln\Big(\frac{x_j}{x_i} \Big) \Big]}_{\textit{centralized}} = \underbrace{-\alpha x_i \Big[\ln x_i - \sum_{j=1}^{n} x_j \ln x_j \Big]}_{\textit{Decentralized}} \tag{15}$$

Finally, by replacing Equations (14) and (15) in Equation (13), it gets to the form of Equation (16), which represents the Decentralized Replicator Dynamics rule in terms of Boltzmann probabilities.

$$\dot{x}_i = \alpha x_i \tau \Big[f_i(x) \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} x_j f_j(x) \Big] - \alpha x_i \Big[\ln x_i - \sum_{j=1}^{n} x_j \ln x_j \Big] \tag{16}$$

This expression is further analyzed in Section 4 from the perspective of EGT and its relationship with the selection-mutation concept. It is also analyzed from the perspective of the exploration-exploitation approaches, and their effect in MAS. Meanwhile, the first parenthesis in Equation (16) refers to changes in the individuals proportion playing the i-th strategy that needs full information of all the payoff functions and the entire population state. Therefore, population dynamics need full-information to evolve. Nonetheless, as one of the main objectives of this work is finding a way to control applications in which agents have no access to the entire information of the system, dependency on full-information must be avoided in some scenarios, where for instance, communication infrastructure limitations, size of systems and privacy issues hinder the availability of this information. Consequently, the analysis of the population structure defines the dynamics that describes the agent behavior. In this sense, the classic assumption for the population structure is that the population has a full and well-mixed structure, i.e., the population structure allows agents to select a specific strategy with the same probability of selecting any other. Figure 2a clarifies this concept by illustrating some agents playing a game. In this case, each agent is represented by an element and the strategies selected are denoted by the element shape, for example paper, stone, or scissors.

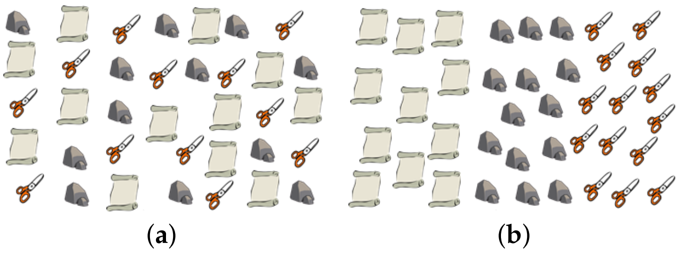


Figure 2. (**a**) Classic population structure approach; (**b**) More realistic population approach.

In EGT, each agent has the same probability of receiving a revision opportunity. At the moment of receiving it, agents randomly select one of their neighbors and, according to the revision protocol selected, they may change their own strategy to the one of their neighbors. Since players have been considered full-well-mixed in the population, the probability that any opponent chooses and plays any of the strategies of the structure is the same Figure 2a. In contrast, Figure 2b illustrates a non-full-well-mixed population, i.e., a more realistic approach. In this case, the probability of receiving an opportunity to make a revision is the same for all agents, but the probability that a neighbor chooses and plays a particular strategy is not the same. For example, if one agent receives a revision opportunity when playing a scissors strategy, the probability of selecting an opponent playing a paper strategy is null, because no scissors are near to any paper. However, for this same agent, the probability of selecting an opponent playing a scissors or stone strategy is more balanced and higher than in the paper case.

To obtain a mathematical representation of this behavior, the graph $G = (T, L, M)$ represents the dynamics among agents. The group T is linked with the strategies the agent can select. Set L represents the meeting probability among strategies. It is very likely that L is higher than zero if there is a link between two strategies. In order to contextualize the forecast in the adjacency matrix $M = [a_{ij}], a_{ij} = 1$ means strategy j and i can meet each other, while $a_{ij} = 0$ means strategy j and i cannot, in this way we define the neighborhood of agent i as Ni. In this regard, full-well-mixed and non-full-well-mixed populations may be specified with two types of graphs. Figure 3a shows a representation using a complete graph for the full-well-mixed populations and Figure 3b for the case of non-full-well-mixed populations. The topology of the graph depends on the specific structure of the population. In the specific case of this work, undirected graphs are assumed, i.e., strategies j and i have the same probability as strategies i and j to meet each other.

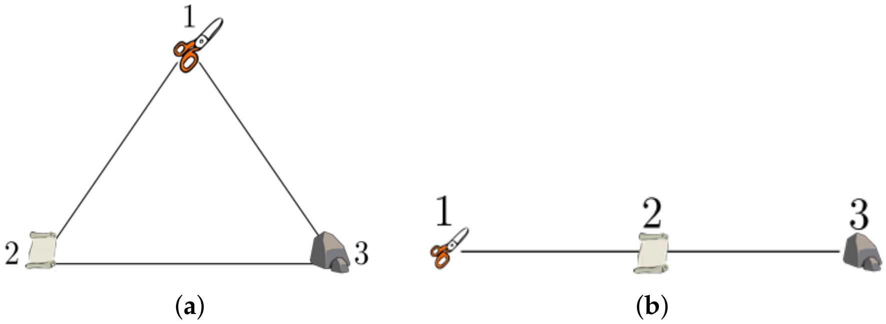


Figure 3. Example of graphs for (**a**) full-well-mixed population and (**b**) non-full-well-mixed population.

Up to this point, in order to control agents interactions is needed to meet the population structure constraint of non-full-information dependency despite the distributed control model stated in Equation (16) as depicted in Figures 2b and 3b. To achieve this condition, the approach of [34] is considered. The authors used the pairwise proportional-imitation protocol to introduce the neighboring concept, as shown in the following expression:

$$p_{ij} = p_j[f_j(p_{Ni}) - f_i(p_{Ni})]_+ \tag{17}$$

where the computation of p_i just needs the knowledge of the portions of the population playing neighboring strategies, which leads us to make the following assumption:

Assumption 1. *Operations using the pairwise proportional-imitation protocol and its update are neighboring-based, i.e., the payoff function as well as the iterations in the sums depend on the neighbors that have effective communication with the i-th player.*

Under this assumption, the resulting distributed replicator dynamics that meet the population structure constraints and allow agents to control the computation of non-full-information are stated in Equation (18):

$$\dot{x}_i = \alpha x_i \tau \Big[f_i(x_{Ni}) \sum_{j \in Ni}^{n} x_j - \sum_{j \in Ni}^{n} x_j f_j(x_{Nj}) \Big] \tag{18}$$

where $f_{i/j}(x_{Ni/j})$ represents the payoff function for player i/j evaluated in the proportion of its effective communication neighbors, and the term $\sum_{j \in Ni}^{n} x_j$ represents a sum only with the effective communication neighbors. Since our assumption of the neighboring concept has been only applied to the first part of Equation (16), the introduction of this concept to the second part of the equation (second parenthesis) takes the following form:

$$-\alpha x_i \Big[\ln x_i - \sum_{k \in Ni}^{n} x_k \ln x_k \Big] \tag{19}$$

where the term k must be understood as the i-th neighbor with an active communication link using strategy j. The last couple of expressions represent the behavior of the i-th player in terms of Boltzmann probabilities for the Distributed Replicator Dynamics or BBDRD. Equation (20) shows the complete form of the control method grouping both approaches, i.e., the distributed and the neighboring concepts.

$$\dot{x}_i = \alpha x_i \tau \underbrace{\Big[f_i(x_{Ni}) \sum_{j \in Ni}^{n} x_j - \sum_{j \in Ni}^{n} x_j f_j(x_{Nj}) \Big]}_{Exploitation} - \alpha x_i \underbrace{\Big[\ln x_i - \sum_{k \in Ni}^{n} x_k \ln x_k \Big]}_{Exploration} \tag{20}$$

As mentioned before, the BBDRD equation also demonstrates the use of the exploitation and exploration concepts from RL, and the selection-mutation perspective from EGT, which are discussed in the next Section. The use of this approach and an illustration of its application to control in a smart grid context is presented in Section 5 of this document.

4. Evolutionary Game Theory Approach

In this Section, the BBDRD shown in Equation (20) is analyzed from the point of view of the RL and also in an evolutionary perspective. This analysis takes relevance to understand the incorporation of the exploration concept in the RD equation and its analogy to the EGT domain.

4.1. Selection-Mutation Perspective

The first part of the dynamics of Equation (20) has the classic structure of the RD, which allows bearing in mind the Q-learner dynamics from the perspective of EGT, since the selection mechanism is implicit in it. Consequently, the second part of the expression represents the mutation behavior for Q-learning, i.e.,

$$x_i \alpha \left(\sum_{k \in Ni}^{n} x_k \ln(x_k) - \ln(x_i) \right) \tag{21}$$

Two entropy values can be recognized from Equation (21): the value of the strategy x_i and the whole probability distribution x. The entropy terms can be described as:

$$E_i = -x_i \ln(x_i) \tag{22}$$

and

$$E_n = -\sum_{k \in Ni}^{n} x_k \ln(x_k) \tag{23}$$

where E_i describes the available information related to strategy i and E_n describes the information of the whole distribution. As a result, the mutation expression can be rewritten as:

$$-\left(\alpha x_i E_n - \alpha E_i\right) \tag{24}$$

The next expression represents the mutation equation derived, taking into account the distinction between old and new states of x_i.

$$\sum_{k \in Ni}^{n} \epsilon_{ik} x_k - x_i \tag{25}$$

The term ϵ_{ik} in Equation (25) describes the mutation rate of players using strategy i that change it to the strategy of the k neighbors with an active communication link, e.g., the strategy j. For values of k higher or equal to 1, ϵ_{ik} is grater than or equal to zero.

From the perspective of EGT, on the one hand, mutation in the context of Q-Learning dynamics is directly related to entropy expressing the strategy state, but this connection is not new, as it has been noticed that entropy is increased by mutation [35]. In [36], this connection is explained from the point of view of the thermodynamics, considering that the mutation has the tendency to increase entropy. On the other hand, the dynamics of the Q-learners shows that the concept of selection is developed according to the RD, where a strategy can be favored by or independent of the resulting payoff, which is closely related to it is opponent behavior. The mutation mechanism is also present. This effect is calculated by comparing the entropy strategy value with the entropy value of the entire population.

4.2. *Exploration-Exploitation Perspective*

Reinforcement learning has the challenge of compensating the exploration and exploitation mechanisms. To obtain the maximum profit, an agent has to take an action, and normally it selects actions that returned a big reward previously. However, to recognize these actions, the agent has to take actions it had not taken previously. By associating exploitation and exploration with mutation and selection mechanisms, a biological interpretation of the RL exploration-exploration concept can be obtained. To make it clearer, the first term in the summation of Equation (20) selects the best strategies all the time, which fits totally on the exploitation concept. Similarly, the importance of introducing the exploration term into the RD equation is its direct relationship with the entropy terms of Equation (21). It is worth noting that the higher the entropy value, the higher the uncertainty level is for selecting one strategy. Thus, the exploration term increases entropy and provides variety at the same time. In this sense, the concepts of exploration and mutation are very close, since both of them provide variety, and give at the same time a feature of heterogeneity to the environment.

Controlling specific situations such as the heterogeneity and the communication constraints of a system is challenging when real life applications are addressed. In this process, compensating the exploration-exploitation mechanisms is not an easy task, as it commonly requires a fine adjustment of the immersed parameters in the process of learning. This adjustment is necessary to control the agents actions at the moment of a decision-making process. This problem is addressed with the BBDRD control method presented in Equation (20) as demonstrated in the following Section.

5. Smart Grids Application

In this Section, the Boltzmann-based distributed replicator dynamics is used in a smart grid problem. In this case, we try to maximize the global utility of the power generators or to minimize the total cost of the power generation, in order to satisfy the restrictions on generation capacity and power balance at the same time [37]. This work uses the maximization utility functions to clarify this, while the approaches to EDP have used the static optimization algorithms [18] or methods of offline direct-search [13,19]. The distribution systems management demands works with dynamic environments that in turn need alternative control techniques and tools [8]. A first solution to this problem was made from the perspective of a changing resource allocation approach, as introduced in [38]. The authors modified the population dynamics concept to work in the context of microgrids, with the main purpose of controlling the devices that constitute it. This can be achieved if the loads, generators and other devices share information and cooperate with other elements of the grid. For the network, these elements are controllable devices.

This work presents now the study case, which consists in finding the place to execute the dispatch algorithm at the microgrid. Then, the distributed population dynamics concept is presented by using the BBDRD control approach.

5.1. Case Study

The authors in [39] proposed two control levels. In the lower levels a controllable voltage source of distributed generators (DGs) is connected to loads with an inverter. The output voltage frequency and magnitude are controlled by a droop-gain controller [17]. Figure 4 shows the general microgrid distribution composed of seven DGs.

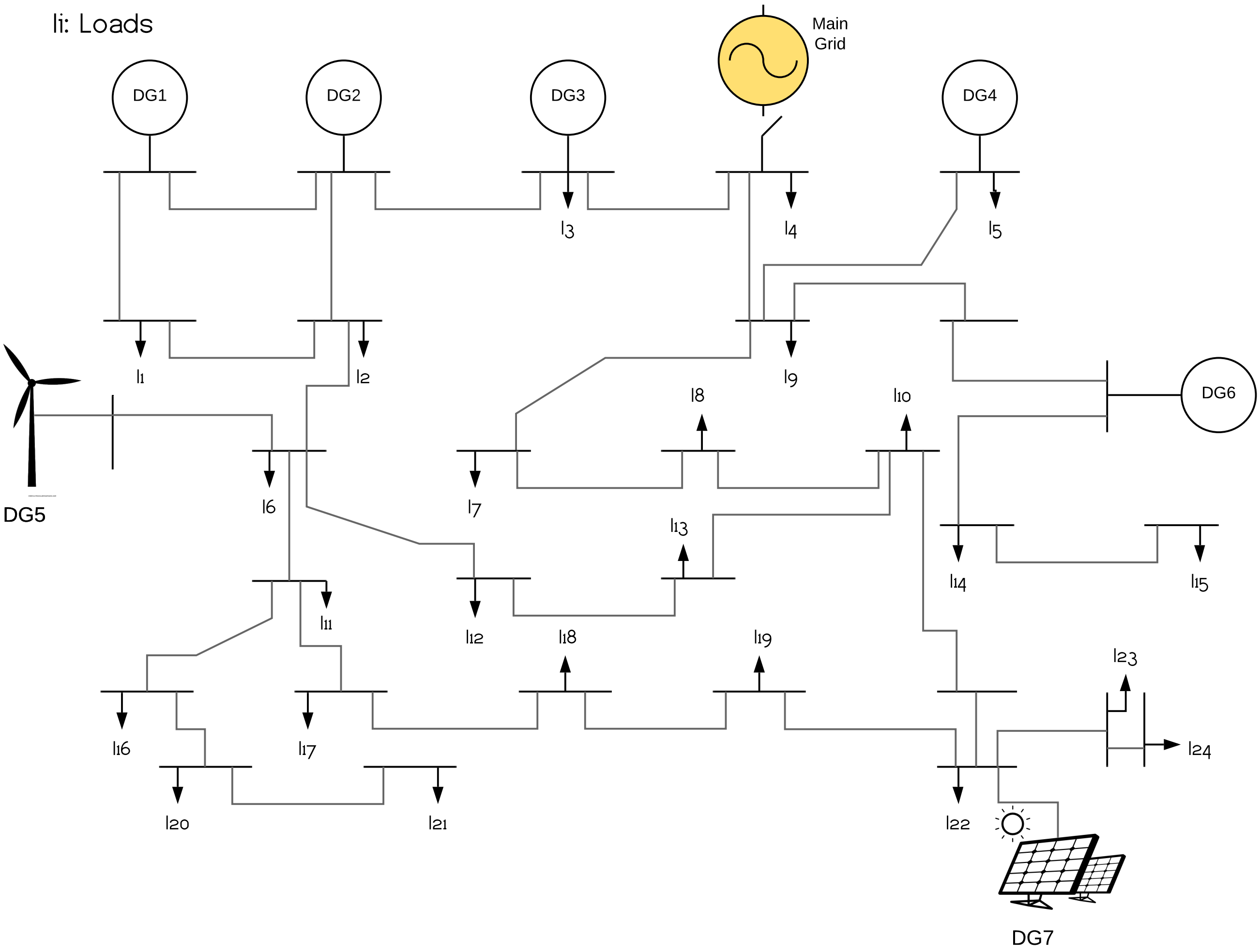


Figure 4. Microgrid distribution. Adapted from [8].

At the highest level, a strategy able to dynamically dispatch set points of power is implemented. The economic constraints, such as the load demands and the power production costs, are sent from the lower level of control and toward the central controller of the microgrid; then, a classic RD is executed. The controller receives dynamical values for load demands and costs, which implies the possibility of including renewable energy resources.

In this case, the dispatch is executed, i.e., the highest level of control. The formulation of the EDP is:

$$\begin{aligned} \max \qquad & J(\varphi) = \sum_{i=i}^{n} J_i(\varphi_i), \\ subject\ \ to \ \ \sum_{i=1}^{n} \varphi_i \ = & \sum_{i=1}^{n} \psi_i \ = \varphi_D \end{aligned} \tag{26}$$

where $0 \ \leq \ \varphi_i \ \leq \varphi_{\max i}, \forall i \in \mathbb{Z}$, n is the total number of DGs, φ_i represents the power set-point for the ith DG, ψ_i represents the loads, φ_D is the total load required by the grid, $\varphi_{\max}$ sets the maximum capacity of generation for the ith DG, and $J_i(\varphi_i)$ is the utility function of each DG. The criterion of the economic dispatch sets the utility function [37], which defines the operation of all generation units using the same marginal utilities as stated in Equation (27)

$$\frac{dJ_1}{d\varphi_1} \ = \frac{dJ_2}{d\varphi_2} \ = ... \ = \frac{dJ_n}{d\varphi_n} \ = \delta, \tag{27}$$

For some $\delta \ > 0$, such that $\sum_{i=1}^{n} \ \varphi_i = \varphi_D$. Based on the EDP criterion of Equation (27), the EDP stated in Equation (26) can have a solution using utility functions with quadratic form for each DG [38].

5.2. Population Games Approach

From population games point of view, this work uses RD to manage the EDP. Figure 5 shows the topology proposed to simulate communication constrains between DGs.

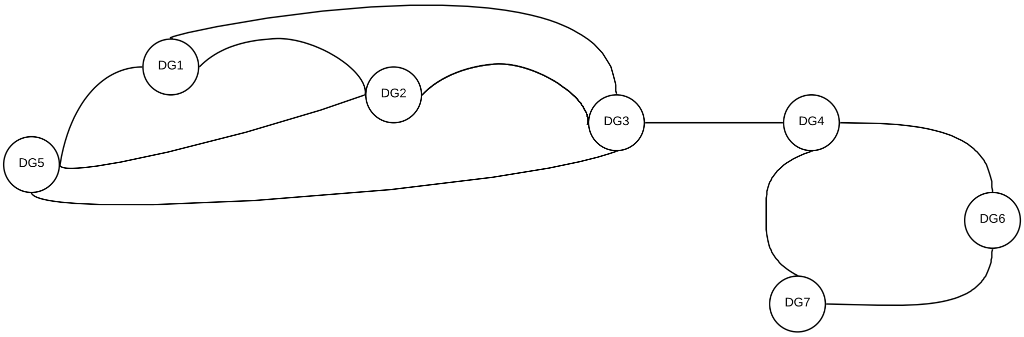


Figure 5. Communication network topology using a graph representation.

For the distributed case, n is the total number of DGs in the system. Let the ith strategy be defined as the act of selecting a DG in the network. In this sense, let φ_i be the amount of power assigned to each DG, which is related to the number of individuals that select the ith strategy in S. The term φ_D is defined as the sum of each power set-point, i.e., $\sum_{i=1}^{n} \ \varphi_i = \varphi_D$ to get a suitable steady state performance. Similarly, to achieve the power balance, the use of $\bar{f} \ = (1/\varphi_D) \ \sum_{i=1}^{n} \ \varphi_i f_i$ allows the invariance of set $\Delta = [\varphi \ \epsilon^t \geq 0 : \sum_{i \ \epsilon \ S} \varphi i = m]$ [40]. This expression guarantees that if $\varphi(0) \ \epsilon \ \Delta$, then $\varphi(t) \ \epsilon \ \Delta, \forall t \ \geq \ 0$, which means that the control strategy should define set-points to ensure the appropriate balance between the generated and the demanded power by the DGs. This behavior allows an appropriate regulation of frequency.

When analyzing the inclusion of technical and economic criteria in the control strategy, the cost and the capacity factors of power generation are very important to condition the final power dispatched to each DG. RD appears to fit appropriately, as its stationary state is reached when the average outcome is equal to all the outcome functions. This feature links

RD and EDP, since it is equal to the economic dispatch criterion of Equation (27) when selecting the outcome function as:

$$f_i(\varphi_i) \ = \frac{dJ_i}{d\varphi_i}\ , \forall i = 1, 2, ..., n, \tag{28}$$

The economic dispatch approach of Equation (27) ensures an optimal solution of the system if constrains are satisfied. The optimization issues may be tackled by marginal utilities in the outcome functions. This is possible because the outcome functions are equal to $\bar{f}$. The outcome selected can be modeled as a function that increases/decreases depending on the distance of the power from/to the set-point desired. In this dynamic, RD allots resources to those DGs that depend on the average outcome. This behavior can be represented by the following function [41]:

$$f(\varpi) \ = r\varpi\left(1 - \frac{\varpi}{k}\right) \tag{29}$$

where k represents the carrying capacity such that the independent variable $\varpi \ \epsilon \ (0, K)$. Here, parameters such as the carrying capacity and the cost factor of generation are used by the outcome functions. Consequently, each DG has an outcome function that can be stated as:

$$f_i(\varphi_i) \ = \frac{dJ_i}{d\varphi_i} \ = \frac{2}{ci}\left(1 - \frac{\varphi i}{\varphi_{max}}\right), \forall i = 1, 2, ..., n, \tag{30}$$

The use of marginal utilities for the outcome functions allows the population game to become a potential game [42]. In Equation (30), the outcome functions lead to functions of quadratic utility for each DG in the optimal EDP [38]. This outcome function has been used in other works such as [38,39,43].

$$J_i(\varphi_i) \ = \frac{1}{ci}\left(2\varphi_i - \frac{\varphi i^2}{\varphi_{maxi}}\right), \quad \forall i = 1, 2, ..., n, \tag{31}$$

5.3. Boltzmann-Based Distributed Replicator Dynamics Approach

The previously described behavior implies the use of a centralized controller and the availability of a high bandwidth infrastructure for communication purposes between nodes. However, we analyze the dynamics immersed in the system, where these classic approaches, e.g., centralized perspectives, could fall short. Therefore, distributed control approximations receives significant attention in the context of a smart grid, because they have proved to be more flexible and adaptable to network variations.

The classic RD approaches assume that the agents have the complete information available to calculate $\vec{f}$. However, sometimes this assumption is not possible due to restrictions of the communication network. For that reason, if we want to analyze the outcomes of the availability of information for the EDP, a Boltzmann-based distributed version of the RD is proposed, as it assumes that the agents have incomplete available information of the system. In this scenario, RD considers the neighbor concept to share local information with other agents. To solve the optimization problem, the following assumptions are needed:

Assumption 2. *The generation units communicate each other using a connected and undirected graph.*

Assumption 3. *The Nash Equilibrium φ^x belongs to the boundaries of a triangle to n-dimensions known as simplex Δ.*

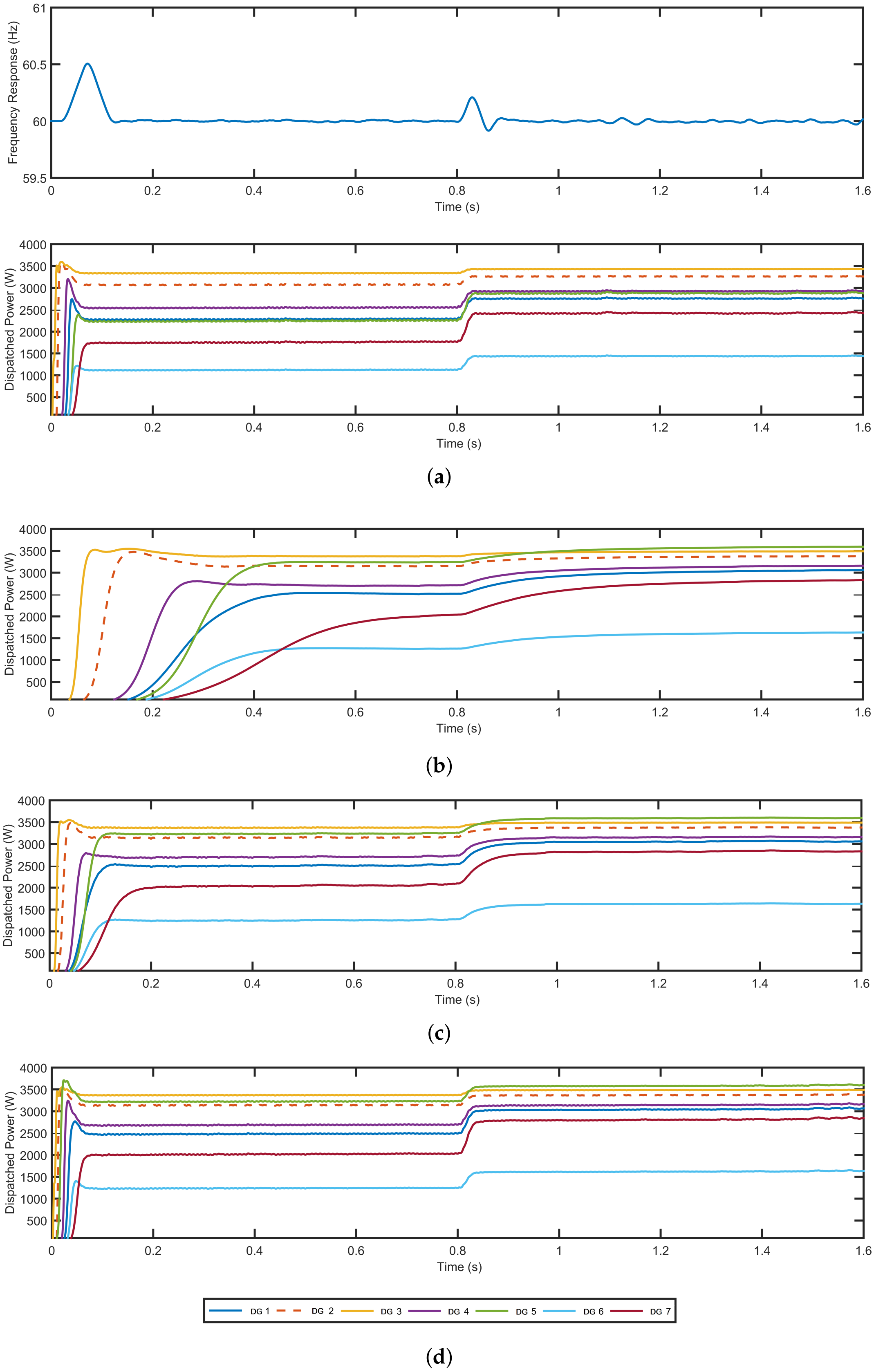
The first assumption means that communication between all generators is guaranteed. Assumption 2 is close in meaning to viability, i.e., if the generators capacity supports the demand of power, then, it is possible to find a solution for the optimization problem, which also means that the power generation capacity belongs to the simplex. If we analyze the concave feature of the utility functions, the problem has a unique solution and the condition of balance is satisfied at steady-state. It is necessary to include the communication graph between the generators to relax the information constraints for the average outcome. To secure the power balance, the results obtained in [44] must be considered, and the invariance of the constraint set Δ is proven. Therefore, the BBDRD can be employed to meet the constraints of the network, given by the graph topology.

5.4. Simulation Results

The proposed BBDRD control model is tested in a study case of a smart grid, with seven distributed generators belonging to a low voltage network. The total power demand of the system is $\varphi_D = 9$ kW; DG 3 is the cheapest and DG 7 is the most expensive. The behavior of DGs 1, 4, 5, and 6 has no substantial cost differences, while DG 2 is the cheapest among them. The system works using a frequency of 60 Hz, where the nominal capacity of all generators is 3.6 kW, except for DG 6 that uses 2 KW.

To have a comparison parameter, we first made the simulation of the classic centralized case, which considers the availability of full-information. Results of this stage are depicted in Figure 6a. In this case, an unanticipated increase in the load of 3 KW and different costs for each generator are observed. The frequency behavior has no significant variations, except in t = 0.8, where the load increment produces a variation of approximately 0.2 Hz, but stabilizes its behavior right after it. Figure 6a also shows the amount of power dispatched to each DG. At the beginning, generator DG 7 dispatches a minimum power in periods of low demand due to its expensive behavior. In contrast, DG 3 goes close to its maximum capacity and stays around this value without being affected by load variations. If the demand increases, DG 7 increases its capacity too, in order to compensate the demand. In the case of DG 6, it almost reaches all its capacity right after the load variation. DGs 1, 4, 5, and 6 have similar behavior because they also have comparable specifications. Finally, DG 2 goes towards its maximum performance due to its cheap behavior.

Once we had the results of the classic centralized case, we proceeded to use the BBDRD control method to compare its behavior. Results of this stage can be observed in Figure 6b–d for different values of τ. The topology shown in Figure 5 was used for simulating the communication constraints in the graphs. As the τ value increases, the system replicates the behavior of the centralized approach. The major differences are observed with low values of τ (Figure 6b), where DGs take more time to reach their working level. In contrast, when using high values of τ (Figure 6d), DGs reach their working levels faster. Regarding communication constraints, DG 5 has a different behavior compared with the centralized approach, i.e., after the increase in demand, it dispatches more power due to the communication constraints of the topology and the influence of the exploration concept of the second term in Equation (20).



(**a**)

(**b**)

(**c**)

(**d**)

Figure 6. Results for a microgrid test system. (**a**) Frequency response and active power response of DGs for the classic RD. The evaluation of the behavior of the Boltzmann-based distributed replicator dynamics is shown for different values of τ as follows: (**b**) $\tau = 0.5$ (**c**) $\tau = 2$, (**d**) $\tau = 5$.

6. Conclusions

The Boltzmann-based distributed replicator dynamics presented in Equation (20) can be described as a distributed control method of learning, which incorporates the exploration scheme from reinforcement learning into the replicator dynamics classic equation. In this regard, the exploration is close in meaning to the mutation concept of EGT, and introduces a way to measure variety in the system using the entropy approach. This distributed control method also uses the scheme of the Boltzmann distribution aiming to introduce the τ parameter for controlling purposes. The selection of a suitable temperature function involves a methodological search and can now be consistently addressed to meet an expected convergence distribution. In this sense, the dynamics of the method have shown properties of the time dependency of the system, which allows using a structured framework in the design of parameters. In terms of stability, the derivation process in Section 3 remains stable if multiple agents are allowed. This fact is evidenced with the incorporation of the population concept into the control method. In this sense, the contribution of the neighboring approach gives the key to avoid centralized schemes and forces agents to take into account only the available information of other players before executing an action. The method was evaluated in a smart grid problem and let us initialize parameters previously. Despite the evolutionary interpretation, this fact states the control method applications in the normative side, rather than in the positive side, which is the common use of EGT.

Problems in Engineering are the representation of real situations where the complexity of the systems can be analyzed using MAS, thanks to the study of the interactions between the agents. EGT has some useful tools to manage these interactions to obtain the control agents behavior. In this work, we show the benefits of using a distributed approach of the EGT for controlling a real life smart grid application. This paper shows the BBDRD performance in experiments where communications constraints are involved, thus becoming a useful tool for the implementation of control strategies in a more realistic Engineering with distributed schemes. This feature takes special relevance due to the possibility to tackle complex systems using local information, allowing considering communications constraints without using a centralized coordinator and avoiding high costs of implementation, which is the case of classic approaches, such as the dual decomposition method. In this regard, the distributed control concept used to deal with the EDP can be extended to other problems in the smart grid context, such as the lack of consistency of renewable generation, physical restrictions of power-flow, and inclusion of power losses.

Results showed that despite the limited information, as the τ value increases and reaches the value of $\tau = 5$, the system replicates the behavior of the centralized approach. In contrast, for $\tau = 0.5$ the performance of the method was far away from results obtained using ideal communication conditions (centralized scheme). Having the capacity to tune behavior of the method beforehand seems to be successful with systems modeled using graphs, where communications constraints between agents arise. This also naturally generalizes to any number of agents or populations. Results also showed that performance of the Boltzmann-based distributed method for solving the economic dispatch problem in a smart grid is convenient from the economic perspective. This feature is explained because the specificities of the DGs were consequent regarding their cost of operation and use of their power capacity. This feature is shown because DG 7 dispatches a minimum power in periods of low demand due to its expensive behavior. In contrast, DGs 2 and 3 are close to their maximum capacity and stay around this value due to their cheap behavior. If the demand increases, DG 7 increases its capacity too to compensate for the demand. The behavior of the other DG also has the expected performance. Ranges of the dispatched power were similar in DGs 1, 4, 5, and 6, which is explained by their similar cost features. When comparing results between centralized and decentralized approaches, it can be noted that both have almost the same performance, except in DG 5, which reaches its maximum right after the demand increased. This feature can be explained from the point of view of the communications constraints in the BBDRD control method.

Author Contributions: Conceptualization, methodology, investigation and formal analysis, G.C.-P. and E.M.-N.; software, validation and writing—original draft preparation, G.C.-P. and E.C.-M. All authors participated in the discussion and proofreading work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Colciencias, grant Doctorado Nacional number 727.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, G.C.P., upon reasonable request.

Acknowledgments: We express our gratitude to Colciencias for the founding of this Project. We thank Universidad Nacional de Colombia and Universidad Santo Tomás for allowing us to use laboratories, hardware and software for the development of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bacci, G.; Lasaulce, S.; Saad, W.; Sanguinetti, L. Game theory for networks: A tutorial on game-theoretic tools for emerging signal processing applications. *IEEE Signal Process. Mag.* **2015**, *33*, 94–119. [CrossRef]
2. Mu, C.; Wang, K. Approximate-optimal control algorithm for constrained zero-sum differential games through event-triggering mechanism. *Nonlinear Dyn.* **2019**, *95*, 2639–2657. [CrossRef]
3. Zhu, M.; Frazzoli, E. Distributed robust adaptive equilibrium computation for generalized convex games. *Automatica* **2016**, *63*, 82–91. [CrossRef]
4. Najeh, S.; Bouallegue, A. Distributed vs centralized game theory-based mode selection and power control for D2D communications. *Phys. Commun.* **2020**, *38*, 100962. [CrossRef]
5. Tang, R.; Wang, S.; Li, H. Game theory based interactive demand side management responding to dynamic pricing in price-based demand response of smart grids. *Appl. Energy* **2019**, *250*, 118–130. [CrossRef]
6. Främling, K. Decision theory meets explainable ai. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 57–74.
7. Navon, A.; Ben Yosef, G.; Machlev, R.; Shapira, S.; Roy Chowdhury, N.; Belikov, J.; Orda, A.; Levron, Y. Applications of Game Theory to Design and Operation of Modern Power Systems: A Comprehensive Review. *Energies* **2020**, *13*, 3982. [CrossRef]
8. Quijano, N.; Ocampo-Martinez, C.; Barreiro-Gomez, J.; Obando, G.; Pantoja, A.; Mojica-Nava, E. The role of population games and evolutionary dynamics in distributed control systems: The advantages of evolutionary game theory. *IEEE Control. Syst. Mag.* **2017**, *37*, 70–97.
9. Lanctot, M.; Lockhart, E.; Lespiau, J.B.; Zambaldi, V.; Upadhyay, S.; Pérolat, J.; Srinivasan, S.; Timbers, F.; Tuyls, K.; Omidshafiei, S.; et al. OpenSpiel: A framework for reinforcement learning in games. *arXiv* **2019**, arXiv:1908.09453.
10. Sandholm, W.H. *Population Games and Evolutionary Dynamics*; MIT Press: Cambridge, MA, USA, 2010.
11. Hindersin, L.; Wu, B.; Traulsen, A.; García, J. Computation and simulation of evolutionary Game Dynamics in Finite populations. *Sci. Rep.* **2019**, *9*, 1–21. [CrossRef]
12. Grüne-Yanoff, T.; Lehtinen, A. Philosophy of game theory. In *Handbook of the Philosophy of Economics*; Mäki, U., Ed.; Elsevier: Oxford, UK, 2012; pp. 531–576.
13. Palomar, D.P.; Chiang, M. A tutorial on decomposition methods for network utility maximization. *IEEE J. Sel. Areas Commun.* **2006**, *24*, 1439–1451. [CrossRef]
14. Marden, J.R. State based potential games. *Automatica* **2012**, *48*, 3075–3088. [CrossRef]
15. Zhao, L.; Wang, J.; Liu, J.; Kato, N. Optimal edge resource allocation in IoT-based smart cities. *IEEE Netw.* **2019**, *33*, 30–35. [CrossRef]
16. Cagnano, A.; De Tuglie, E.; Mancarella, P. Microgrids: Overview and guidelines for practical implementations and operation. *Appl. Energy* **2020**, *258*, 114039. [CrossRef]
17. Lopes, J.P.; Moreira, C.; Madureira, A. Defining control strategies for microgrids islanded operation. *IEEE Trans. Power Syst.* **2006**, *21*, 916–924. [CrossRef]
18. Ibaraki, T.; Katoh, N. *Resource Allocation Problems: Algorithmic Approaches*; MIT Press: Cambridge, MA, USA, 1988.
19. Ahn, S.J.; Moon, S.I. Economic scheduling of distributed generators in a microgrid considering various constraints. In Proceedings of the 2009 IEEE Power & Energy Society General Meeting, Calgary, AB, Canada, 26–30 July 2009; pp. 1–6.
20. Strbac, G. Demand side management: Benefits and challenges. *Energy Policy* **2008**, *36*, 4419–4426. [CrossRef]
21. Olivares, D.E.; Cañizares, C.A.; Kazerani, M. A centralized optimal energy management system for microgrids. In Proceedings of the 2011 IEEE Power and Energy Society General Meeting, Detroit, MI, USA, 24–28 July 2011; pp. 1–6.
22. Quintana-Barcia, P.; Dragicevic, T.; Garcia, J.; Ribas, J.; Guerrero, J.M. A distributed control strategy for islanded single-phase microgrids with hybrid energy storage systems based on power line signaling. *Energies* **2019**, *12*, 85. [CrossRef]
23. Huang, B.; Liu, L.; Zhang, H.; Li, Y.; Sun, Q. Distributed optimal economic dispatch for microgrids considering communication delays. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 1634–1642. [CrossRef]

24. Vasquez, J.C.; Guerrero, J.M.; Miret, J.; Castilla, M.; De Vicuna, L.G. Hierarchical control of intelligent microgrids. *IEEE Ind. Electron. Mag.* **2010**, *4*, 23–29. [CrossRef]
25. Bloembergen, D.; Tuyls, K.; Hennes, D.; Kaisers, M. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res.* **2015**, *53*, 659–697. [CrossRef]
26. Peters, H. *Game Theory: A Multi-Leveled Approach*; Springer: Berlin/Heidelberg, Germany, 2015.
27. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
28. Ertel, W. Reinforcement Learning. In *Introduction to Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 289–311.
29. Da Silva, F.L.; Costa, A.H.R. A survey on transfer learning for multiagent reinforcement learning systems. *J. Artif. Intell. Res.* **2019**, *64*, 645–703. [CrossRef]
30. Başar, T.; Zaccour, G. *Handbook of Dynamic Game Theory*; Springer: Berlin/Heidelberg, Germany, 2018.
31. Weibull, J.W. *Evolutionary Game Theory*; MIT Press: Cambridge, MA, USA, 1997.
32. Newton, J. Evolutionary game theory: A renaissance. *Games* **2018**, *9*, 31. [CrossRef]
33. Tuyls, K.; Verbeeck, K.; Lenaerts, T. A selection-mutation model for q-learning in multi-agent systems. In Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia, 14–18 July 2003; pp. 693–700.
34. Barreiro-Gomez, J.; Obando, G.; Quijano, N. Distributed population dynamics: Optimization and control applications. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *47*, 304–314. [CrossRef]
35. Eiben, A.E.; Smith, J.E. *Introduction to Evolutionary Computing*; Springer: Berlin/Heidelberg, Germany, 2015.
36. Stauffer, D. *Life, Love and Death: Models of Biological Reproduction and Aging*; Institute for Theoretical Physics: Köln, Euroland, 1999.
37. Aj, W.; Wollenberg, B. *Power Generation, Operation and Control*; John Wiley & Sons: New York, NY, USA, 1996; p. 592.
38. Pantoja, A.; Quijano, N. A population dynamics approach for the dispatch of distributed generators. *IEEE Trans. Ind. Electron.* **2011**, *58*, 4559–4567. [CrossRef]
39. Mojica-Nava, E.; Macana, C.A.; Quijano, N. Dynamic population games for optimal dispatch on hierarchical microgrid control. *IEEE Trans. Syst. Man Cybern. Syst.* **2013**, *44*, 306–317. [CrossRef]
40. Hofbauer, J.; Sigmund, K. *Evolutionary Games and Population Dynamics*; Cambridge University Press: Cambridge, MA, USA, 1998.
41. Britton, N.F. *Essential Mathematical Biology*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
42. Young, H.P.; Zamir, S. *Handbook of Game Theory with Economic Applications*; Technical Report; Elsevier: Amsterdam, The Netherlands, 2015.
43. Mojica-Nava, E.; Barreto, C.; Quijano, N. Population games methods for distributed control of microgrids. *IEEE Trans. Smart Grid* **2015**, *6*, 2586–2595. [CrossRef]
44. Pantoja, A.; Quijano, N.; Passino, K.M. Dispatch of distributed generators under local-information constraints. In Proceedings of the 2014 American Control Conference, Portland, OR, USA, 4–6 June 2014; pp. 2682–2687.