Learning to Use the Spectrum in Self-Configuring Heterogenous Networks: A Logit Equilibrium Approach

Samir M. Perlaza, Samson Lasaulce, Hamidou Tembine, Merouane Debbah

Learning to Use the Spectrum in Self-Configuring Heterogenous Networks: A Logit Equilibrium Approach

S.M. Perlaza
Alcatel Lucent Chair in
Flexible Radio - SUPELEC.
France
Samir.MedinaPerlaza@supelec.fr

S. Lasaulce
Laboratoire des Signaux et
Systèmes (LSS) - CNRS,
SUPELEC, Univ. Paris Sud.
France
Lasaulce@lss.supelec.fr

H. Tembine
Telecommunications Dept. -
SUPELEC
France
Hamidou.Tembine@supelec.fr

M. Debbah
Alcatel Lucent Chair in
Flexible Radio - SUPELEC.
France
Merouane.Debbah@supelec.fr

ABSTRACT

In this paper, we study the particular scenario where several transmitter-receiver pairs communicate subject to mutual interference due to the usage of the same frequency bands. In particular, we focus on the case of heterogeneous networks, where radio devices have different interests (utility functions), transmit configurations (sets of actions), as well as different signal processing and calculation capabilities. The underlying assumptions of this work are the followings: (i) the network is described by a set of states, for instance, the channel realization vector; (ii) radio devices are interested in their long-term average performance rather than instantaneous performance; (iii) each radio device is able to obtain a measure of its achieved performance at least once after updating its transmission configuration. Considering these conditions, we model the heterogenous network by a stochastic game. Our main contribution consists of a family of behavioral rules that allow radio devices to achieve an epsilon-Nash equilibrium of the corresponding stochastic game, namely a logit equilibrium. A thorough analysis of the convergence properties of these behavioral rules is presented. Finally, our approach is used in the context of a classical parallel interference channel in order to compare with existing results.

1. INTRODUCTION

Self-configuring heterogeneous wireless networks is the term generally used to identify distributed networks where radio devices from different classes exploit common resources (e.g., frequency bands typically). Often, the difference from one device to another relies mainly of the physical layer technology, e.g., Wi-Fi, Bluetooth, Zigbee, etc, and also on the type of applications the devices are designed for. Within this framework, coordination between radio devices is limited due to the unfeasibility of any kind of message exchange, and thus, each radio device must autonomously determine its own optimal transmit configuration. This is one of the reasons why game theory and the concept of Nash equilibrium (NE) [12] has been widely accepted in the analysis of this kind of networks.

The relevance of the NE stems from the fact that once it is achieved, each radio device's transmission configuration is optimal with respect to the transmit configuration of all the other devices. Thus, the NE is clearly a desired solution from each radio device's standpoint and appears to be relevant in wireless networks where neither coordination at all nor cooperation is possible. Nonetheless, achieving NE is not an easy task. As the main constraint, we highlight the fact that radio devices are not able to observe neither the transmit configuration (e.g., the transmit power level or the channel selected) for the other devices nor the instantaneous global state of the network, i.e., channel realizations, energy constrains and quality of service requirements of all the active radio devices. Thus, the lack of information naturally constraints each radio device to determine its optimal transmit configuration at a given instant.

From this perspective, an increasing interest has been observed in the design of behavioral rules to allow radio devices to achieve an NE configuration as a result of a short interaction with its counterparts, similar to a learning process [8]. In this direction, the best response dynamics (BRD) [5] and fictitious play (FP) [2] has been largely used in wireless communications [24, 16, 22, 25, 14] and have been shown to converge to NE in certain network topologies.

The main constraint in BRD, FP and its variants is the fact that each radio device must observe the transmit configurations of all the other devices, the actual game state and possess a closed form expression of the utility function, which is clearly a very demanding condition in practical scenarios. In some network topologies and depending on the performance metric, this condition can be weakened and a simple broadcast message from each receiver might be enough to implement either BRD or FP [24]. However, the amount of signaling might be too high depending on the number of dimensions of the scenario, e.g., the number of frequency bands or transmit antennas.

More elaborated behavioral rules for achieving equilibria are based on reinforcement learning (RL) [3, 21, 29]. In RL, the information required by each radio device is simply an

Gamecomm 2011, May 16, 2011 - Paris, France.

observation of its own achieved performance at least every time it changes its transmit configuration. The principle of RL is as follows. After observing the current value of its utility, each radio device updates a probability distribution over all its feasible transmit configurations (or actions). At each update, the probability of the played action increases or decreases depending on the observed utility. In the wireless communication domain, this idea has been used and has been proved to converge to NE in some particular radio resource allocation scenarios [28, 30]. The main advantages of RL with respect to BRD and FP are numerous (provided it converges to NE). For instance, RL is less demanding in terms of information: only a numerical observation for the achieved utility at each game stage is sufficient to implement the RL rule.

However, aside from all the attractive advantages of RL, it has a critical drawback: each observation of the utility is used to directly update the probability distributions without maintaining an estimate of the performance achieved with each transmit configuration. This fact might lead the network to converge to a stationary state which is not an NE. We say stationary, in the sense that none of the radio devices changes its transmit configuration since it is unable to identify that other transmit configurations might bring a higher performance. Consider for instance, the simple power allocation game described in [17] and general examples in [9]. Motivated by this observation, in this paper, we introduce a kind of behavioral rules which are known in the domain of Markov decision processes as actor-critic algorithms [6, 27, 7]. Here, each radio device simultaneously learns both the time-average performance achieved with each of its transmit configurations and the equilibrium probability distribution. This estimation helps to solve the problem encountered in behavioral rules based on reinforcement learning, where convergence is observed but the final network configuration does not correspond to an NE. In particular, contrary to the RL algorithms described above, whenever these behavioral rules lead to a stationary network configuration, it corresponds to a logit equilibrium (LE), which is indeed, an epsilon-close Nash equilibrium concept.

The paper is organized as follows. In Sec. 2 the problem of spectrum sharing is formulated. In Sec. 3, such a problem is modeled by a stochastic game and the concept of ϵ-equilibrium and state-independent behavioral strategies is introduced. In Sec. 4 a particular ϵ-equilibrium known as logit equilibrium is introduced. In Sec. 5 and Sec. 6 a family of behavioral rules which allow radio devices to achieve a logit equilibrium are presented and a thorough analysis of its convergence in our particular system model is presented. In Sec. 7, we provide some numerical examples in order to evaluate the performance of the proposed behavioral rule. The paper is concluded by Sec. 8

2. SYSTEM MODEL

In the following, we describe a decentralized self-configuring network where transmitters aim to optimize their individual spectral efficiency by setting up their spectrum access scheme, i.e., number of channels to access, allocated power per channel. Nonetheless, as we shall see in Sec. 3, the contributions of this paper can be applied to scenarios where radio devices have different performance metrics and different sets of transmit parameters to set up.

Consider a set $\mathcal{K} = \{1, \ldots, K\}$ of transmitters and a set $\mathcal{J} = \{1, \ldots, J\}$ receivers. Each ransmitter sends private information to its respective receiver trough out a set $\mathcal{S} \overset{\triangle}{=} \{1, \ldots, S\}$ of orthogonal channels. Here, the orthogonality is assumed in the frequency domain. All transmitters simultaneously use the same set $\mathcal{S}$ of channels and thus, communications are subject to mutual interference. Let $h_{j,k}^{(s)}(n)$ represent the channel realization between transmitter k and receiver j over channel s at time n. In our analysis, flat fading channels are assumed during the frame period, i.e., the channel realization is assumed time-invariant during the transmission of one frame, however, the channel might vary from frame to frame period. Denote by $\boldsymbol{h}(n) = \left(h_{j,k}^{(s)}(n)\right) \in \mathbb{C}^{J \cdot K \cdot S}$ the vector of channel realizations at interval n and let $\mathcal{H}$ be the finite set of all possible channel realization vectors (in practice, relevant quantities like channel quality indicators in 3G cellular systems are quantized). Let $\boldsymbol{h}^{(i)}$ be the i-th element of the set $\mathcal{H}$, with $i \in \{1, \ldots, |\mathcal{H}|\}$. For each channel use, the vector $\boldsymbol{h}(n)$ is drawn from the set $\mathcal{H}$ following a probability distribution $\boldsymbol{\rho} = (\rho_{\boldsymbol{h}^{(1)}}, \ldots, \rho_{\boldsymbol{h}^{(|\mathcal{H}|)}}) \in \triangle(\mathcal{H})$. That is, $\rho_{\boldsymbol{h}^{(i)}} = \Pr\left(\boldsymbol{h}(n) = \boldsymbol{h}^{(i)}\right)$, with $n \in \mathbb{N}$. The vector of transmitted symbols $\boldsymbol{x}_k(n)$, $\forall k \in \mathcal{K}$, is an S-dimensional random variable with zero mean and covariance matrix $\boldsymbol{P}_k(n) = \mathbb{E}\left(\boldsymbol{x}_k(n)\boldsymbol{x}_k^H(n)\right) = \mathrm{diag}\left(p_{k,1}(n), \ldots, p_{k,S}(n)\right)$. For all $(k,s) \in \mathcal{K} \times \mathcal{S}$, $p_{k,s}(n)$ represents the transmit power allocated by transmitter k over channel s. A power allocation (PA) vector for transmitter $k \in \mathcal{K}$ is any vector

$$\boldsymbol{p}_k(n) = (p_{k,1}(n), \ldots, p_{k,S}(n)) \in \mathcal{A}_k,$$

where, $\forall k \in \mathcal{K}$, $\displaystyle\sum_{s=1}^{S} \boldsymbol{p}_k(n) = p_{k,\max}$ and $p_{k,\max}$ is the maximum transmit power of transmitter k. Following this notation, the power allocation vector $\boldsymbol{p}_k^{(s)}$ represents the s-th element of the set $\mathcal{A}_k$. We denote by $N_k = |\mathcal{A}_k|$ the cardinality of the set $\mathcal{A}_k$. We respectively denote the noise spectral density and the bandwidth of channel $s \in \mathcal{S}$ by N_0 and B_s. The total bandwidth is denoted by $B = \sum_{s=1}^{S} B_s$, independently of the receiver. We denote the individual spectral efficiency of transmitter $k \in \mathcal{K}$ as follows,

$$u_k(\boldsymbol{h}(n), \boldsymbol{p}_k(n), \boldsymbol{p}_{-k}(n)) = \sum_{s \in \mathcal{S}} \frac{B_s}{B} \log_2\left(1 + \gamma_{k,s}(n)\right) \text{ [bps/Hz]}, \tag{1}$$

where $\gamma_{k,s}(n)$ is the signal-to-interference plus noise ratio (SINR) seen by player k over its channel s at time n, i.e.,

$$\gamma_{k,s}(n) = \frac{p_{k,s}(n) g_{j_k,k}^{(s)}(n)}{N_0 B_s + \displaystyle\sum_{i \in \mathcal{K} \setminus \{k\}} p_i^{(s)}(n) g_{j_k,i}^{(s)}(n)}, \tag{2}$$

where, $j_k \in \mathcal{J}$ is the index of the receiver with which transmitter k is associated. Moreover, for all $(j,k) \in \mathcal{J} \times \mathcal{K}$ and $n \in \mathbb{N}$, $g_{j,k}^{(s)}(n) \triangleq \left|h_{j,k}^{(s)}(n)\right|^2$.

In the following of this paper, our interest focuses on designing behavioral rules to determine how radio devices choose their power allocation vectors at each interval n in order to to maximize the time-average of their individual spectral efficiency (1).

3. GAME MODEL

The long-term behavior of the decentralized network described in the previous section can be modeled by a stochastic game. In the following, we formulate such stochastic game and the concept of epsilon-equilibrium. Finally, we describe a particular class of behavioral strategies, to which we restrict the analysis of our game.

3.1 Game Formulation

Consider the stochastic game $\mathcal{G}$ described by the 5-tuple

$$\mathcal{G} = \left(\mathcal{K}, \{\mathcal{A}_k\}_{k\in\mathcal{K}}, \{\bar{u}_k\}_{k\in\mathcal{K}}, \mathcal{H}, \{\rho_{\boldsymbol{h}}\}_{\boldsymbol{h}\in\mathcal{H}}\right). \quad (3)$$

The sets $\mathcal{K} = \{1, \dots, K\}$, $\mathcal{H} = \left\{\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(H)}\right\}$ and $\mathcal{A}_k = \{A_k^{(1)}, \dots, A_k^{(N_k)}\}$, $\forall k \in \mathcal{K}$, represent the set of players, the set of network states and the set of actions of player k, respectively. In this analysis, such sets are assumed finite, nonempty, and time-invariant sets. In the game $\mathcal{G}$, each player represents an *active transmitter* of the network. The set of actions of a given transmitter corresponds to the set of all its feasible *transmission configurations*, for instance, the power allocation vectors, etc. The set $\mathcal{H}$ is the set of all possible channel realizations. The vector $\boldsymbol{\rho} = (\rho_{\boldsymbol{h}^{(1)}}, \dots, \rho_{\boldsymbol{h}^{(|\mathcal{H}|)}}) \in \triangle(\mathcal{H})$ is the probability distribution described in Sec. 2. The function $\bar{u}_k$ is the long-term performance metric of player k, for all $k \in \mathcal{K}$, and will be described later.

The game $\mathcal{G}$ is played stage by stage *ad infinitum*. At each stage n, every player k adopts an action e.g., a power allocation vector $\boldsymbol{p}_k(n) \in \mathcal{A}_k$. At the end of the stage, player k observes a numerical value $\tilde{u}_k(n)$ of its achieved performance e.g., the individual spectral efficiency $\tilde{u}_k(n) = u_k\left(\boldsymbol{h}(n), \boldsymbol{p}_k(n), \boldsymbol{p}_{-k}(n)\right)$. Note that these observations might be noisy [15]. However, this case is not the focus of the present work.

Note that all the information gathered by player k at stage n is the 2-tuple $(a_k(n), \tilde{u}_k(n)) \in \mathcal{A}_k \times \mathbb{R}$. We denote by $\theta_k(n)$ the available information gathered by player k up to interval n, i.e.,

$$\theta_k(n) = \{(a_k(0), \tilde{u}_k(0)), \dots, (a_k(n-1), \tilde{u}_k(n-1))\}. \quad (4)$$

We refer to $\theta_k(n)$ as the private history of player k at time n. The set of all possible private histories of player k at time n is denoted by $\Theta_k(n)$, and,

$$\Theta_k(n) = \left(\mathcal{A}_k \times \mathbb{R}\right)^n. \quad (5)$$

At each game stage, transmitters select their respective PA vector following a probability distribution

$$\boldsymbol{\pi}_k(n) = \left(\pi_{k,A_k^{(1)}}(n), \dots, \pi_{k,A_k^{(N_k)}}(n)\right) \in \triangle(\mathcal{A}_k)$$

which is built based on its private history $\theta_k(n)$. Here, $\forall n_k \in \{1, \dots, N_k\}$, $\pi_{k,\boldsymbol{p}_k^{(n_k)}}(n)$ represents the probability that player k plays action $\boldsymbol{p}_k^{(n_k)} \in \mathcal{A}_k$ at time n, i.e.,

$$\pi_{k,A_k^{(n_k)}}(n) = \Pr\left(\boldsymbol{p}_k(n) = \boldsymbol{p}_k^{(n_k)}\right). \quad (6)$$

Such probability is built as follows. There exists a function for each time interval n, which we denote by $\sigma_{k,n}$, and is defined as follows,

$$\sigma_{k,n} : \Theta_k(n) \rightarrow \triangle(\mathcal{A}_k). \quad (7)$$

Thus, at each time n player k determines its probability distribution $\boldsymbol{\pi}_k(n)$ based on its private history $\theta_k(n)$. Each player k possesses an infinite sequence of functions,

$$\sigma_k = \{\sigma_{k,n}\}_{n>0}, \quad (8)$$

such that at each game stage n, it is able to generate the corresponding distribution $\boldsymbol{\pi}(n)$. We refer to the sequence of functions σ_k as the behavioral strategy (BS) of player k. In the following, we denote by Σ_k, the set of all possible BS of player k and let $\Sigma = \Sigma_1 \times \ldots \times \Sigma_K$ be the set of all BS profiles.

Now note that, given any behavioral strategy $\boldsymbol{\sigma}$, the initial channel selection profile $\boldsymbol{p}(0) \in \mathcal{A}$ induce a set of sequences of probability distributions $\{\boldsymbol{\pi}_k(n)\}_{n>0}$, for all $k \in \mathcal{K}$. Hence the set of sequences $\{\boldsymbol{\pi}_k(n)\}_{n>0}$ induced by $\boldsymbol{p}(0) \in \mathcal{A}$ together with the initial probability distributions $\boldsymbol{\pi}(0) = (\boldsymbol{\pi}_1(0), \dots, \boldsymbol{\pi}_K(0))$ induce a probability distribution over all the possible sequences of action profiles $\{\boldsymbol{p}(0), \boldsymbol{p}(1), \dots\}$. We denote the expectation with respect to such probability distribution by $\mathbb{E}_{(\boldsymbol{\pi}(0),\boldsymbol{\sigma})}$. Then, the long-term expected performance of player k can be measured by the function, $\bar{u}_k : \Sigma_1 \times \ldots \times \Sigma_K \rightarrow \mathbb{R}_+$, where,

$$\bar{u}_k(\sigma_k, \boldsymbol{\sigma}_{-k} \,|\boldsymbol{\pi}(0)) = \lim_{n\rightarrow\infty} \frac{1}{n} \mathbb{E}_{(\boldsymbol{\pi}(0),\boldsymbol{\sigma})}\left[\sum_{i=0}^{n-1} \tilde{u}_k(i)\right]. \quad (9)$$

The function (9) captures the situation in which the interaction between all transmitters in the network lasts many time intervals and the instantaneous performance is insignificant compared with the performance in all the other time intervals.

In the following, the game $\mathcal{G}$ is analyzed assuming that the aim of each player k is to choose a BS $\sigma_k \in \Sigma_k$ such that it maximizes its performance metric (9) given the BS $\boldsymbol{\sigma}_{-k} \in \Sigma_{-k}$ adopted by all the other players and regardless of the initial action profile $\boldsymbol{a}(0) \in \mathcal{A}$. In particular, we look for a the BS profile $\boldsymbol{\sigma}^* = (\sigma_1^*, \dots, \sigma_K^*) \in \Sigma_1 \times \ldots \Sigma_K$ such that none of the players can obtain a performance improvement by unilaterally using other BS. We provide a more precise concept of this expected solution of the stochastic game $\mathcal{G}$ in the following subsection.

3.2 Nash Equilibrium and ϵ-Equilibrium

In the following, we describe the concept of ϵ-equilibrium and Nash equilibrium in the context of the stochastic game $\mathcal{G}$. First, let us define the ϵ-equilibrium as follows,

Definition 1 (ϵ-equilibrium in the game G). Let $\epsilon > 0$. In the game $\mathcal{G}$, a strategy profile $\boldsymbol{\sigma}^* \in \Sigma_1 \times \ldots \times \Sigma_K$ is an ϵ-equilibrium if it satisfies, for all $k \in \mathcal{K}$ and for all $\sigma_k \in \Sigma_k$,

$$\bar{u}_k(\sigma_k^*, \boldsymbol{\sigma}_{-k}^* \,|\boldsymbol{\pi}(0)) \geqslant \bar{u}_k(\sigma_k, \boldsymbol{\sigma}_{-k}^* \,|\boldsymbol{\pi}(0)) - \epsilon, \quad (10)$$

independently of the initial probability distributions $\boldsymbol{\pi}_k(0) \in \triangle(\mathcal{A}_k)$, $\forall k \in \mathcal{K}$.

An ϵ-equilibrium can be interpreted as a BS profile such that, none of the players can obtain an improvement higher than ϵ by unilaterally changing its own BS. Note also that, by letting $\epsilon = 0$ in Def. 1, the classical definition of Nash equilibrium is obtained.

In the following section, we discuss the feasibility of achieving these equilibrium concepts in the game $\mathcal{G}$.

3.3 Stationary State Independent Behavioral Strategies

Stationary state independent behavioral strategies (SSI-BS) are considered the simplest class of BS in stochastic games [26, 13]. Let the set of stationary state independent behavioral strategy (SSI-BS) profiles be denoted by $\bar{\Sigma}$ and let a given SSI-BS profile be defined as follows,

Definition 2 (Stationary State Independent BS). Consider the game $\mathcal{G}$ and let $\boldsymbol{\sigma} \in \Sigma$ be a behavioral strategy. Then, $\boldsymbol{\sigma}$ is said to be stationary state-independent (SSI) if for all $k \in \mathcal{K}$ and any two private histories $\theta_k(n) \in \boldsymbol{\Theta}_k(n)$ and $\theta_k(m) \in \boldsymbol{\Theta}_k(m)$, with $n \neq m$, it follows that

$$\sigma_{k,n}\left(\theta_k(n)\right) = \sigma_{k,m}\left(\theta_k(m)\right), \quad (11)$$

independently of the states $\boldsymbol{h}(n)$ and $\boldsymbol{h}(m)$.

From Def. 2, it can be implied that for a player k, an SSI-BS does not depend on any of the previous channel selections $\boldsymbol{p}_k(0), \ldots, \boldsymbol{p}_k(n-1)$ and neither on the previous nor current channel realizations $\boldsymbol{h}(0), \ldots, \boldsymbol{h}(n)$. Thus, a SSI-BS $\sigma_k \in \bar{\Sigma}$ can be identified by a vector $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K) \in \triangle(\mathcal{A}_1) \times \ldots \times \triangle(\mathcal{A}_K)$, such that, $\forall k \in \mathcal{K}$ and $\forall (\theta_k(n), \boldsymbol{h}(n)) \in \Theta_k(n) \times \mathcal{H}$, it holds that $\sigma_{k,n}(\theta_k(n)) = \boldsymbol{\pi}_k$. In the sequel, we indifferently use the infinite set of sequences $\sigma_k = \{\boldsymbol{\pi}_k\}_{n>0}$ or the vectors $\boldsymbol{\pi}_k$, with $k \in \mathcal{K}$, to refer to the SSI-BS $\boldsymbol{\sigma}$. Moreover, with a slight abuse of notation, we indifferently write either $\bar{u}_k(\sigma_k, \boldsymbol{\sigma}_{-k})$ or $\bar{u}_k(\boldsymbol{\pi}_k, \boldsymbol{\pi}_{-k})$ to denote the achieved performance of player k given the SSI-BS $\boldsymbol{\sigma}$. In the following, we restrict the analysis of the game $\mathcal{G}$ to the set of SSI-BS. The justifications for this choice are manifold. First, note that none of the players is able to identify the current network state at each stage of the game. Moreover, given the information gathered by player k up to stage $n-1$, i.e., $\theta_k(n-1)$, with $k \in \mathcal{K}$ and $n > 1$, it is not possible to infer any information about the network state $\boldsymbol{h}(n)$. This implies that, player k is unable to calculate an optimal probability distribution $\boldsymbol{\pi}_k(n)$ at each stage n, since the ignorance of the network state at each stage implies the ignorance of a closed form expression of the instantaneous performance metric u_k and the long-term performance metric $\bar{u}_k$. Thus, regardless of the stage n and all the information gathered up to such stage n, all players face the same scenario.

4. LOGIT EQUILIBRIUM

In this section, we introduce the concept of an ϵ-equilibrium known as logit equilibrium. We define the logit equilibrium in the context of the game $\mathcal{G}$ in SSI-BS and we conclude this section claiming its existence and providing some insight on its uniqueness in the game $\mathcal{G}$.

4.1 Logit Equilibrium in SSI-BS

Before we provide a formal definition of the logit equilibrium, we introduce the idea of logit best response.

Definition 3 (Logit Best Response). Consider the game $\mathcal{G}$ and let the vector $\boldsymbol{\pi}_{-k} \in \triangle(\mathcal{A}_1) \times \ldots \times \triangle(\mathcal{A}_{k-1}) \times \triangle(\mathcal{A}_{k+1}) \times \ldots \times \triangle(\mathcal{A}_K)$ represent a given SSI-BS profile, with $k \in \mathcal{K}$. Then, the logit best response of player k, with parameter $\gamma_k > 0$, is the probability distribution $\boldsymbol{\beta}_k^{(\gamma_k)}(\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k})) \in \triangle(\mathcal{A}_k)$ such that, $\boldsymbol{\beta}_k^{(\gamma_k)} : \mathbb{R}^{N_k} \rightarrow \triangle(\mathcal{A}_k)$ is the logit function,

$$\boldsymbol{\beta}_k^{(\gamma_k)}(\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k})) \quad = \quad \left(\beta_{k,A_k^{(1)}}^{(\gamma_k)}\left(\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k})\right), \ldots, \beta_{k,A_k^{(N_k)}}^{(\gamma_k)}\left(\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k})\right)\right)$$

and $\forall n_k \in \{1, \ldots, N_k\}$,

$$\beta_{k,A_k^{(n_k)}}^{(\gamma_k)}\left(\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k})\right) \quad = \quad \frac{\exp\left(\gamma_k \bar{u}_k(\boldsymbol{e}_k^{(n_k)}, \boldsymbol{\pi}_{-k})\right)}{\displaystyle\sum_{m=1}^{N_k} \exp\left(\gamma_k \bar{u}_k(\boldsymbol{e}_k^{(m)}, \boldsymbol{\pi}_{-k})\right)} \tag{12}$$

From Def. 3, it can be implied that at each stage of the game, every power allocation vector of a given transmitter has a non-zero probability of being played, i.e., $\forall k \in \mathcal{K}$ and $\forall n_k \in \{1, \ldots, N_k\}$ and $\forall \gamma_k \in \mathbb{R}_+$, it holds that, $\beta_{k,A_k^{(n_k)}}^{(\gamma_k)}(\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k})) > 0$. More generally, it can be stated that the logit best response in SSI-BS is represented by a probability distribution that assigns high probabilities to the power allocation vectors associated with a high average individual spectral efficiencies and low probability to power allocations associated with low average individual average performance.

Finally, note also that conversely to the case of the best response in the case of Nash equilibrium [12], the logit best response of player k is unique for all the SSI-BS profiles the other players might adopt.

Using Def. 3, we define the logit equilibrium as follows,

Definition 4 (Logit Equilibrium in SSI-BS). Consider the game $\mathcal{G}$ and let the vector $\boldsymbol{\pi}^* = (\boldsymbol{\pi}_1^*, \ldots, \boldsymbol{\pi}_K^*) \in \triangle(\mathcal{A}_1) \times \ldots \times \triangle(\mathcal{A}_K)$ represent a stationary state-independent behavioral strategy (SSI-BS). Then, $\boldsymbol{\pi}^*$ is a logit equilibrium SSI-BS profile with parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ if for all $k \in \mathcal{K}$, it holds that,

$$\boldsymbol{\pi}_k^* \quad = \quad \boldsymbol{\beta}_k^{(\gamma_k)}\left(\bar{u}_k\left(\boldsymbol{e}_1^{(N_k)}, \boldsymbol{\pi}_{-k}^*\right), \ldots, \bar{u}_k\left(\boldsymbol{e}_{N_k}^{(N_k)}, \boldsymbol{\pi}_{-k}^*\right)\right) \tag{13}$$

At the logit equilibrium, since all actions are played with non-zero probability, at some given game stages the actions taken by player k do not maximize the instantaneous performance u_k, which negatively impacts the long-term performance $\bar{u}_k$. In [15], it has been shown that the maximum loss of performance player k might experience is not higher than $\frac{1}{\gamma_k}\ln(N_k)$, which confirms that the logit equilibrium falls in the class of ϵ-equilibrium described in Def. 1.

4.2 Existence and Uniqueness of the LE

The main result regarding the existence of an LE in the game $\mathcal{G}$ is the following.

Theorem 5 (Existence of the LE). The stochastic game $\mathcal{G}$ has at least one logit equilibrium in the set of stationary state-independent behavioral strategies.

The proof of Theorem 5 is presented in [15]

The uniqueness of the LE in the game $\mathcal{G}$ in SSI-BS is strongly related to the parameters γ_k, with $k \in \mathcal{K}$. For instance, when $\forall k \in \mathcal{K}$, $\gamma_k \rightarrow 0$, there exits a unique LE in SSI-BS and corresponds to the vectors $\boldsymbol{\pi}_k = \frac{1}{N_k}(1, \ldots, 1) \in \triangle(\mathcal{A}_k)$. This LE (uniform probability distributions over the sets $\mathcal{A}_1, \ldots, \mathcal{A}_K$) is unique, independently of the number of NE the game $\mathcal{G}$ might possess. Conversely, when $\forall k \in \mathcal{K}$, $\gamma_k \rightarrow \infty$, the set of LE becomes identical to the set of NE and thus, the game $\mathcal{G}$ exhibits as many LE as NE might exist in $\mathcal{G}$ in SSI-BS.

The number of NE (and thus, the number of LE) of the game $\mathcal{G}$ is often not unique. In fact, it has been already shown that even the simplest cases exhibit several NE. For instance, when $\mathcal{K} = \{1, 2\}$ and $\mathcal{H}$ and $\mathcal{J}$ are unitary, the number of NE can be 1 or 2 in pure strategies [16]. When, $\mathcal{J} = \mathcal{K} = \{1, 2\}$ and $\mathcal{H}$ is unitary, the number of NE can be 1, 2 or 3 in pure strategies [19].

5. LEARNING LOGIT EQUILIBRIA

In this section, we design behavioral strategy profiles $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_K) \in \Sigma$ such that given the information gathered by player k at each stage n, i.e., given the sets $\{\theta_k(n)\}_{n>0}$ for all $k \in \mathcal{K}$, it is able to generate infinite sequences $\{\boldsymbol{\pi}_k(n)\}_{n>0}$, such that, $\lim_{n\rightarrow\infty} ||\boldsymbol{\pi}_k(n) - \boldsymbol{\pi}_k^*|| = 0$, where

$$\boldsymbol{\pi}^* = (\boldsymbol{\pi}_1^*, \ldots, \boldsymbol{\pi}_K^*) \in \triangle(\mathcal{A}_1) \times \ldots \times \triangle(\mathcal{A}_K)$$

is a logit equilibrium in SSI-BS of the game $\mathcal{G}$ (Def. 4).

An important remark is that achieving a logit equilibrium implies that at a given game stage, radio devices are able to build logit best responses in order to face the other players behavioral strategies. Nonetheless, building such a probability requires each player to know the vector $\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k}(n))$.

Thus, radio devices must estimate their corresponding vector $\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k}(n))$ at each game stage n based on their current history $\theta_k(n)$ in order to generate their logit best response.
Let the N_k-dimensional vector

$$\hat{\boldsymbol{u}}_k(n) = \left(\hat{u}_{k,A_k^{(1)}}(n), \ldots, \hat{u}_{k,A_k^{(N_k)}}(n)\right) \tag{14}$$

be the estimation, at the game stage n, that player k possesses of the vector $\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k}(n))$. In the following, we present a result initially introduced in [17, 15], which allows radio devices to simultaneously estimate the vector $\bar{\boldsymbol{u}}_k(\cdot, \boldsymbol{\pi}_{-k}(n))$ and determine the probability distribution $\boldsymbol{\pi}_k(n)$, with which it chooses the power allocation vector $\boldsymbol{p}_k(n)$.

Theorem 6 (LE Behavioral Strategy). Consider the game $\mathcal{G}$ and assume that for all $k \in \mathcal{K}$ and for all $n_k \in \{1, \ldots, N_k\}$, it holds that for all $n \in \mathbb{N}$,

$$\begin{cases} \hat{u}_{k,\boldsymbol{p}_k^{(n_k)}}(n) &= \hat{u}_{k,\boldsymbol{p}_k^{(n_k)}}(n-1) + \\ & \alpha_k(n) \frac{\mathbb{1}_{\left\{\boldsymbol{p}_k(n-1) = \boldsymbol{p}_k^{(n_k)}\right\}}}{\pi_{k,\boldsymbol{p}_k^{(n_k)}}(n)} \left(\tilde{u}_k(n-1) - \hat{u}_{k,\boldsymbol{p}_k^{(n_k)}}(n-1)\right), \\ \pi_{k,\boldsymbol{p}_k^{(n_k)}}(n) &= \pi_{k,\boldsymbol{p}_k^{(n_k)}}(n-1) + \\ & \lambda_k(n) \left(\boldsymbol{\beta}^{(\gamma_k)}_{k,\boldsymbol{p}_k^{(n_k)}}(\hat{\boldsymbol{u}}_k(n)) - \pi_{k,\boldsymbol{p}_k^{(n_k)}}(n-1)\right), \end{cases} \tag{15}$$

where, $\boldsymbol{p}_k(0) \in \mathcal{A}_k$, $\hat{\boldsymbol{u}}_k(0) \in \mathbb{R}^{N_k}$ and $\boldsymbol{\pi}_k(0) \in \triangle(\mathcal{A}_k)$ are arbitrary initializations. Consider also the following assumptions and for all $(j,k) \in \mathcal{K}^2$, the learning rates α_k and λ_j satisfy that

$$(B0) \quad \lim_{T\to\infty} \sum_{n=0}^{T} \alpha_k(n) = +\infty \quad \text{and} \quad \lim_{T\to\infty} \sum_{n=0}^{T} \alpha_k(n)^2 < +\infty,$$

$$(B1) \quad \lim_{T\to\infty} \sum_{n=0}^{T} \lambda_k(n) = +\infty \quad \text{and} \quad \lim_{T\to\infty} \sum_{n=0}^{T} \lambda_k(n)^2 < +\infty,$$

and,

$$(B2) \quad \lim_{n\to\infty} \frac{\lambda_j(n)}{\alpha_k(n)} = 0.$$

Then, if the set of coupled stochastic approximation algorithms (15) converge, it holds that,

$$\lim_{n\to\infty} \boldsymbol{\pi}_k(n) = \boldsymbol{\pi}_k^*, \tag{16}$$

$$\lim_{n\to\infty} \hat{u}_{k,\boldsymbol{p}_k^{(n_k)}}(n) = \bar{u}_k(\boldsymbol{e}_{n_k}^{(N_k)}, \boldsymbol{\pi}^*_{-k}), \tag{17}$$

where $\boldsymbol{\pi}_k^* \in \triangle(\mathcal{A}_k)$ satisfies that,

$$\boldsymbol{\pi}_k^* = \boldsymbol{\beta}_k^{(\gamma_k)}\left(\bar{u}_k\left(\boldsymbol{e}_1^{(N_k)}, \boldsymbol{\pi}^*_{-k}\right), \ldots, \bar{u}_k\left(\boldsymbol{e}_{N_k}^{(N_k)}, \boldsymbol{\pi}^*_{-k}\right)\right). \tag{18}$$

The proof of Theorem 6 is presented in the most general case of the game $\mathcal{G}$ in [15], based on previous results on stochastic approximations [9, 1]. The behavioral rule in Theorem 6 has been proved to convergence in several classes of games, e.g., potential games [11]. However, in general, the game $\mathcal{G}$ does not fall in any of those classes and thus, the convergence must be proved.

6. CONVERGENCE ANALYSIS

In the following, we consider some particular cases of the scenario described above and we present some results on the convergence of the proposed algorithms.

6.1 Parallel Multiple Access Channels ($J = 1$)

Note that when $J = 1$, the scenario described before reduces to a parallel multiple access channel [4], and thus, the corresponding game has a potential function ϕ, defined as follows,

$$\phi(\boldsymbol{\pi}) = \sum_{\boldsymbol{h}\in\mathcal{H}} \sum_{s\in\mathcal{S}} \log_2\left(\sigma_s^2 + \sum_{i=1}^{K} p_{i,s} \left|h_{1,i}^{(s)}(n)\right|^2\right) \rho_{\boldsymbol{h}} \prod_{j=1}^{K} \pi_{j,\boldsymbol{p}_j}. \tag{19}$$

Then, from Theorem 4.6.1 in [15], the following holds,

Proposition 7. *When $J = 1$, the algorithm in* (15)*, with $\lambda_1 = \ldots = \lambda_K$, converges to a logit equilibrium of the resulting game $\mathcal{G}$.*

In particular, the BRD has been used to achieve equilibria in this context [24] and proved to converge to NE. Other algorithms such as FP and its variants are also shown to converge [18]. Nonetheless, the advantages of the algorithm in (15) over the BRD (and FP) stem from the fact that there is no synchronization required for the radio devices to coordinate in sequential or simultaneous updates of their configuration [20]. Here, the only requirement for each device is to observe a numerical value of its SINR each time it updates its configuration, independently of all the other devices' strategy update timing. Moreover, contrary to the case of simultaneous BRD, the convergence is always ensured [18].

6.2 Parallel Interference Channels ($J = 2$)

When $J = 2$ receivers, the resulting network topology describes in the simplest case, i.e., when $\forall j \in \mathcal{J}$, $|\mathcal{K}_j| = 1$, the parallel interference channel [4]. In the most general case, i.e., when $\forall j \in \mathcal{J}$, $|\mathcal{K}_j| > 1$, it describes a 2-cell multicarrier cellular channel (in uplink), e.g., an OFDM cellular network system. In both cases, the resulting game is no longer a potential game and thus, convergence of the BRD (and FP) are not always observed [25, 23]. Nonetheless, using our algorithm, convergence is always achieved, as we show in the following proposition,

Proposition 8. *When $J = 2$, $\lambda_i = \lambda_1$, $\forall i \in \mathcal{K}_1$, $\lambda_j = \lambda_2$, $\forall j \in \mathcal{K}_2$ and*

$$\lim_{n\to\infty} \frac{\lambda_1(n)}{\lambda_2(n)} = 0 \tag{20}$$

the algorithm in (15) *converges to a logit equilibrium of the resulting game $\mathcal{G}$.*

The proof of proposition 8 follows from Theorem 4.6.2 in [15].

7. NUMERICAL ANALYSIS

In this section, we provide a numerical analysis of the performance achieved by radio devices following the behavioral rule proposed in this article. First, we focus on the achieved performance in limited time. Here, our interest focuses in determining the sum spectral efficiency when the learning period is limited. Note that the theoretical analysis requires infinite time for convergence, which is not practically appealing. Second, we focus on the impact of the number of choices each radio devices might possess at a given time. Here, we verify the counter intuitive result which states that increasing the set of choices each radio device possesses during the whole game realization might lead to worse global performance.

7.1 Convergence in Finite Time

In the following, we say that a learning algorithm converges to either a LE, if there exists a number $n \in \mathbb{N}_1$, given a sufficiently small number $\epsilon > 0$, such that,

$$\forall k \in \mathcal{K},\ ||\boldsymbol{\pi}_k(n) - \boldsymbol{\pi}_k^*|| < \epsilon, \tag{21}$$

where $\boldsymbol{\pi}^* = (\boldsymbol{\pi}_1^*, \ldots, \boldsymbol{\pi}_K^*)$ is a LE strategy profile, respectively. We measure the convergence time, as the number of iterations required to observe convergence in the sense of (21), assuming that all convergence updates occur periodically at a constant frequency. We refer as achieved performance to the time-average utility observed by the radio devices from the first strategy update up to convergence. As another note, we point out the fact that there exist other ways to measure convergence other than Euclidian distance as in (21). For instance, using the Kullback-Leibler divergence as in [10]. In the following, we evaluate numerically the convergence time and the achieved performance by radio devices in the particular cases described above.

Consider the decentralized wireless network described in Sec. 2, when there exist only two transmitters, i.e., $\mathcal{K} = \{1, 2\}$, two receivers, i.e., $\mathcal{J} = \{1, 2\}$ and two channels $\mathcal{S} = \{1, 2\}$. We limit the set of power allocation vectors, to the channel selection case, i.e.,

$$\begin{aligned} \mathcal{A}_k \quad = \quad & \Big\{ \boldsymbol{p}_k^{(s)} = p_{k,\max}\, \boldsymbol{e}_s : \forall s \in \mathcal{S},\ \boldsymbol{e}_s = (e_{s,1}, \ldots, e_{s,S}), \\ & \forall r \in \mathcal{S} \setminus s,\ e_{s,r} = 0,\ \text{and } e_{s,s} = 1 \Big\}. \end{aligned} \tag{22}$$

In order to facilitate a fair comparison of the behavioral rule in Theorem 6 with existing results [20], we consider the set $\mathcal{H}$ is unitary. This implies that the stochastic game reduces to play the same game repeatedly *ad infinitum*. At each stage, the corresponding one-shot game $\mathcal{G}(\boldsymbol{h})$ might have either one NE in pure strategies or two NE in pure strategies plus one NE in mixed strategies [19]. Here, we generate 10000 channel realizations. For each channel realization, we calculate the sum of individual spectral efficiencies at the NE. We treat separately the case of unique and multiple equilibria. Similarly, for each channel realization, we determine the sum of individual spectral efficiencies achieved by both transmitters when their learning time is limited to a fixed number of time intervals (game repetitions). In Fig. 1, we plot the average sum of individual spectral efficiencies as a function of the number of times the game is let to be repeated. On the left, we consider the case of a unique equilibrium and on the right, the case of multiple equilibria is considered.

In Fig. 1, it is clearly seen that the longer the radio devices are left to learn, the better their achieved performance. Interestingly, after certain number of iterations, radio devices are able to achieve an average sum spectral efficiency which is better than the worst NE, i.e., the NE with the lowest sum spectral efficiency. Another important remark, is the impact of the parameter γ_k, for all $k \in \mathcal{K}$. In Fig. 1, it is also shown that the smaller γ_k the more interest player k has in playing uniformly all its own actions. Conversely, when the parameter γ_k is large, the corresponding radio devices is tempted to use more often the best configuration and thus, it achieves a better performance.

7.2 Impact of the Number of Choices

In this subsection, we increase the number of available channels and we let each transmitter to use either a unique channel or any subset of adjacent channels. Thus, if we consider S channels, the cardinality of the sets $\mathcal{A}_k$ is $\frac{S}{2}(1 + S)$, for all $k \in \mathcal{K}$. Here, we generate 10000 channel realizations and we let radio devices to learn through 1000 game repetitions. In Fig. 2, it is shown that increasing the number of available channels leads to a loss of spectral efficiency. This observation is due to the fact that letting each radio device to use an additional channel implies increasing the number of possible power allocation vectors. This implies that the radio device has to try more power allocation vectors in order to estimate the individual spectral efficiency it obtains with each of them. However, such power allocation vectors might not be those bringing the highest individual spectral efficiency, and thus, using them reduces the average individual spectral efficiency.

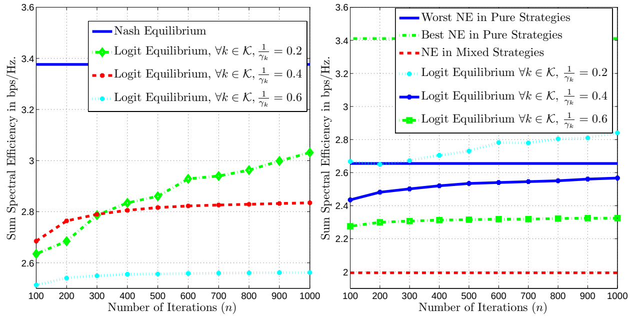

Figure 1: Achieved sum spectral efficiency in the two-transmitter two-receiver two-channel game, when transmitters are limited to channel selection. Here $\alpha_1(n) = \alpha_2(n) = \frac{1}{n^{(\frac{3}{4})}}$, $\alpha_2(n) = \frac{1}{n^{(\frac{2}{3})}}$ and $\lambda_1(n) = \frac{1}{n}$. Moreover, $SNR = \frac{p_{k,\max}}{\sigma^2} = 10$ dBs. (left) Case of unique NE. (right) Case of multiple NE.

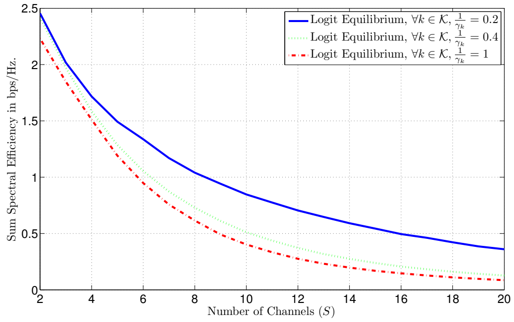

Figure 2: Achieved sum spectral efficiency in the two-transmitter two-receiver N-channel game, when transmitters are limited to use only one channel or any combination of adjacent channels. Here $\alpha_1(n) = \alpha_2(n) = \frac{1}{n^{(\frac{3}{4})}}$, $\alpha_2(n) = \frac{1}{n^{(\frac{2}{3})}}$ and $\lambda_1(n) = \frac{1}{n}$. Moreover, $SNR = \frac{p_{k,\max}}{\sigma^2} = 10$ dBs.

8. CONCLUSIONS

This paper presents a framework to learn distributed strategies in communication networks where the transmitters may not know much about the structure of the game, have different action sets, have different utility functions, and only know the realizations of their utility and nothing else (in particular the utility function is not assumed to be known). The interesting case where transmitters learn at different speeds is addressed, which is also a possible feature in heterogeneous networks. In this context, we have shown that the proposed (behavioral) strategies are independent of the closed-form expression of the performance metric and the set of transmit configurations of each radio device. Here, the required conditions for our results to be valid are: Each device must possess a finite set of feasible actions and it must be able to observe its achieved instantaneous performance at least once after each update of its transmit configuration. When the proposed behavioral rule is used and convergence is observed, then the system is said to achieve a Logit equilibrium. A convergence analysis was carried out for a particular scenario considering as performance metric the individual spectral efficiency.

9. REFERENCES

[1] V. Borkar. Stochastic approximation with two timescales. *Systems Control Lett.*, 29:291–294, 1997.

[2] G. W. Brown. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376, 1951.

[3] M. F. Bush R. Stochastic models of learning. *Wiley Sons, New York.*, 1955.

[4] T. M. Cover and J. A. Thomas. Elements of information theory. *Wiley-Interscience*, 1991.

[5] D. Fudenberg and J. Tirole. Game theory. *MIT Press*, 1991.

[6] V. R. Konda and V. Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM J. Control Optim.*, 38(1):94–123, 1999.

[7] V. R. Konda, John, and J. N. Tsitsiklis. Actor-critic algorithms. In *SIAM Journal on Control and Optimization*, pages 1008–1014. MIT Press, 2001.

[8] S. Lasaulce and H. Tembine. *Game Theory and Learning in Wireless Networks: Fundamentals and Applications*. Elsevier Academic Press, Oct. 2011.

[9] S. D. Leslie and E. J. Collins. Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Ann. Appl. Probab.*, 13(4):1231–1251, 2003.

[10] P. Mertikopoulos, E. V. Belmega, A. Moustakas, and S. Lasaulce. Distributed learning policies for power allocation in multiple access channels. *IEEE Journal on Selected Areas in Communications*, 30(1), Jan. 2012.

[11] D. Monderer. Potential games. *Games and Economic Behavior*, 14:124–143, 1996.

[12] J. F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, 1950.

[13] A. Neyman and S. Sorin. *Stochastic Games and Applications*. NATO Science Series, 1999.

[14] J.-S. Pang, G. Scutari, F. Facchinei, and C. Wang. Distributed power allocation with rate constraints in Gaussian parallel interference channels. *IEEE Trans. on Info. Theory*, 54(8):3471–3489, Aug. 2008.

[15] S. M. Perlaza. *Game Theoretic Approaches to Spectrum Sharing in Decentralized Self-Configuring Networks*. PhD thesis, Télécom ParisTech, Jul. 2011.

[16] S. M. Perlaza, E. V. Belmega, S. Lasaulce, and M. Debbah. On the base station selection and base station sharing in self-configuring networks. *3rd ICST/ACM International Workshop on Game Theory in Communication Networks*, Oct. 2009.

[17] S. M. Perlaza, H. Tembine, and S. Lasaulce. How can ignorant but patient cognitive terminals learn their strategy and utility? In *the 11th IEEE Intl. Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2010)*, Marrakech, Morocco, June 2010.

[18] S. M. Perlaza, H. Tembine, S. Lasaulce, and V. Q. Florez. On the fictitious play and channel selection games. In *IEEE Latin-American Conference on Communications (LATINCOM)*, pages 1–5, Bogota, Colombia, Sept. 2010.

[19] L. Rose, S. M. Perlaza, and M. Debbah. On the Nash equilibria in decentralized parallel interference channels. In *IEEE Workshop on Game Theory and Resource Allocation for 4G*, Kyoto, Japan, Jun. 2011.

[20] L. Rose, S. M. Perlaza, S. Lasaulce, and M. Debbah. Learning equilibria with partial information in wireless networks. *IEEE Communications Magazine, special issue Game Theory in Wireless Communications*, Sep. 2011.

[21] P. Sastry, V. Phansalkar, and M. Thathachar. Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information. *IEEE Transactions on Systems, Man and Cybernetics*, 24(5):769–777, May 1994.

[22] G. Scutari, S. Barbarossa, and D. Palomar. Potential games: A framework for vector power control problems with coupled constraints. *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006.

[23] G. Scutari, D. Palomar, and S. Barbarossa. Simultaneous iterative water-filling for Gaussian frequency-selective interference channels. In *IEEE International Symposium on Information Theory*, pages 600–604, July 2006.

[24] G. Scutari, D. Palomar, and S. Barbarossa. Optimal linear precoding strategies for wideband non-cooperative systems based on game theory – part II: Algorithms. *IEEE Trans. on Signal Processing*, 56(3):1250–1267, mar. 2008.

[25] G. Scutari, D. Palomar, and S. Barbarossa. The MIMO iterative waterfilling algorithm. *IEEE Transactions on Signal Processing*, 57(5):1917–1935, May 2009.

[26] E. Solan. Stochastic games. *Encyclopedia of Database Systems, Springer*, 2009.

[27] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 1057–1063, 2000.

[28] Y. Xing and R. Chandramouli. Stochastic learning solution for distributed discrete power control game in wireless data networks. *IEEE/ACM Trans. Networking*, 16(4):932–944, 2008.

[29] H. P. Young. Strategic learning and its limits (arne ryde memorial lectures sereis). *Oxford University Press, USA*, 2004.

[30] W. Zhong, Y. Xu, M. Tao, and Y. Cai. Game theoretic multimode precoding strategy selection for mimo multiple access channels. *IEEE Signal Processing Letters*, 17(6):563 – 566, jun. 2010.