Reinforcement Learning for Repeated Power Control Game in Cognitive Radio Networks

Pan Zhou, *Student member*, Yusun Chang, *Member* and John A. Copeland, *Fellow*

***Abstract*—Cognitive radio (CR) users are expected to be uncoordinated users that opportunistically seek the spectrum resource from primary users (PUs) in a competitive way. In most existing works, however, CR users are required to share the interference channel information and power strategies to conduct the game with pricing mechanisms that incur the frequent exchange of information. The requirement of significant communication overheads among CR users impedes *fully* distributed solutions for the deployment of CR networks, which is a challenging problem in the research communities. In this paper, a robust distributed power control algorithm is designed with low implementation complexity for CR networks through reinforcement learning, which does not require the interference channel and power strategy information among CR users (and from CR users to PUs). To the best of our knowledge, this research provides the solution for the first time for the incomplete-information power control game in CR networks. During the repeated game, CR users can control their power strategies by observing the interference from the feedback signals of PUs and transmission rates obtained in the previous step. This procedure allows achieving high spectrum efficiency while conforming to the interference constraint of PUs. This constrained repeated stochastic game with learning automaton is proved to be asymptotically equivalence to the traditional game with complete information. The properties of existence, diagonal concavity and uniqueness for the game are studied. A Bush-Mosteller reinforcement learning procedure is designed for the power control algorithm, and the properties of convergence and learning rate of the algorithm are analyzed. The performance of the learning-based power control algorithm is thoroughly investigated with simulation results, which demonstrates the effectiveness of the proposed algorithm in solving variety of practical CR network problems for real-world applications.**

***Index Terms*—Cognitive Radio, Power Control, Repeated Games, Bush Mosteller, Reinforcement Learning, Nash Equilibrium, Stochastic Optimization, Lagrangian Multiplier.**

I. Introduction

IN COGNITIVE Radio (CR) [1] networks, power control deals with the selection of proper transmit power for CR users' transmissions that achieves high spectrum efficiency by enabling CR users to reuse the PUs' spectrum bands under the interference constraints imposed by PUs. In the next generation wireless communications, CR users are expected to be uncoordinated opportunistic users, whereas there are conflicting interests among the CR users [2], [3]. This motivates the use of noncooperative game theory to perform researches on CR networks (see a survey paper [2]). Compared to traditional centralized solutions, the game-theoretical approach has the advantage of distributed implementation for CR networks: each CR user is only interested in their own utility maximization, and it does not need to know other users' payoff (or utility function). However, to perform the power control game, an optimal power strategy for each CR user requires its rivals' (i.e. other CR users in the network) interference and power transmitting strategy. These overheads cost channel resources and the information exchange among CR users by using non-cooperative game theory would scale in an order of $\Theta(N^2)$, which is infeasible in the large network deployment.

In the power control game of CR networks, two fundamentally conflicting objectives exist in CR networks. On the one hand, each CR user wishes to achieve higher Signal-to-Interference plus Noise Ratio (SINR) and transmit at higher rates, which results in better Quality of Service (QoS); on the other hand, this higher SINR is obtained at the expense of increased interference to other CR users and to PUs. This selfish behavior of users would lead to performance degradation [5], which is called the "*tragedy of commons*" in economics. To solve this problem, most existing works introduce "pricing" [4] as an incentive scheme to facilitate more efficient resource utilization for selfish CR users [6], [7], [26], [9], [12], [10]. However, the pricing-based scheme requires a certain degree of cooperation among CR users, and it is difficult to configure and implement pricing schemes in CR networks, since there lacks a centralized coordinator and/or there is no single *creditable* user in CR networks. Moreover, price-based incentive schemes can incur significant overheads in the algorithm design. For example, in [11], a self-incentive pricing scheme is adopted, where each CR user checks the best pricing factors for every power control strategy at the Nash equilibrium. This scheme has high implementation complexity, which might not be suitable in the large network deployment.

As we have noticed, the above works [6], [7], [26], [9], [12], [10] focused on power control in the slow channel-varying or no-fading environments, where frequent channel information exchange is not necessary. Power control for CR networks in fading environments has also studied [17] [14] [13]. These works indicate that the acquisition of adequate Channel State Information (CSI) from PUs, which results in better throughput for CR users in fast time varying channels, is a hard problem. Moreover, due to the opportunistic and heterogeneous nature of CR users, CR users are most likely to be autonomous users. The usual assumption of spontaneous

Manuscript received 15 January 2011; revised 28 July 2011. Parts of this work have been presented at *IEEE Global Telecommunications Conference (IEEE GLOBECOM 2010)* [28].

P. Zhou and J. A. Copeland are with Communications Systems Center (CSC) Lab, School of ECE, Georgia Institute of Technology, Atlanta, GA, 30332, USA (e-mail: {pzhou9, john.copeland}@ece.gatech.edu).

Y. Chang is with Electrical Engineering, Division of Engineering, Southern Polytechnic State University, Marietta, GA, 30060, USA (e-mail: yusun@gatech.edu). He is also affiliated with the Georgia Institute of Technology. He is the corresponding author for this article.

Digital Object Identifier 10.1109/JSAC.2012.120106.

willingness to exchange their private information, e.g., channel and power strategy, is unrealistic for conflict of interest CR users.

Our work is motivated by the drawbacks of classic game-theoretical approaches stated above, and the growing interest from the research community that concerns on fully distributed power control algorithms for CR networks (see related work in Section II): the demand of robust distributed power control algorithms with low communication overheads for opportunistic and competitive CR users. In this paper, we formulate the power control problem as an incomplete-information non-cooperative game in CR networks, where the interference channel and power strategy information among CR users and the interference channel information from CR users to PUs are not available. The power control game as an interaction process is modeled by repeat games [19]. The proposed new framework requires CR users to be "cognitive" enough as in [15] that could decode the link control information from PUs' feedback channels. During every stage of the repeated game, each CR user will only utilize the interference information explored from PUs' control link and the transmission rate (by trying certain power control strategy) obtained in the last stage to decide the power control strategy in this stage with no other information. A reinforcement-learning approach is designed for the repeated power control game. The proposed algorithm is suitable for a wide scope of practical CR networks. The possible network scenarios are: 1) there are no creditable CR users; 2) no infrastructure-supported CR networks; 3) CR users are not willing to exchange their private information; 4) the large network deployment condition, where channel information exchange does not provide scalability; 5) fast time varying environments, where channel estimation among CR users is slow; 6) PUs' activity are unknown, i.e., PUs' on-and-off distributions are unknown (the interference-power constraints might change during this time for mobile PUs).

Similar to the classic game-theoretical approach, the objective of the proposed power control game is as follows: CR users need to compete for maximizing their own average utility in the repeated game; on the other hand, each CR user needs to satisfy the interference power constraints imposed by PUs during the course of the game. The formulated power control problem is a *constrained* repeated game with learning automaton. The solution of this repeated game is to choose the optimal transmit power strategy at each stage that maximizes the average utility and conform the interference power constraints from PUs eventually. As a learning automaton, the CR user will adjust the mixed transmit power strategies according to its own realization of utility and the observed interference level at every time step with no other information. The power control game is proved to be asymptotically equivalent to the classic game with complete information. The properties of existence, diagonal concavity and uniqueness for the matrix game at each individual CR user are studied. One of the best known stochastic models of reinforcement learning, the Bush-Mosteller reinforcement learning procedure [32], is used to design the power control algorithm. The properties of convergence and learning rate of the algorithm are analyzed as well. Finally, the learning-based power control algorithm is implemented to perform simulations.

The remainder of the paper is organized as follows. In section II, we summarize related works and their contributions on power control in CR networks. In Section III, the system model, notations and the behavior of the repeated power control game as learning automation are described. In Section IV, the mixed strategies and Nash equilibrium are defined, and the Learning-based Game problem is formulated. In Section V, the properties of the Nash equilibrium are studied. In Section VI, the learning algorithm for the repeated power control game is proposed. In section VII, the convergence and learning rate are analyzed. Simulation results of the algorithm are investigated in Section VIII. Finally, the conclusion of this study is presented in Section IX.

II. Related Works

Power Control problem is one of the key research issues in spectrum underlay CR networks. Especially, the non-cooperative behaviors among CR users by applying game theory has gained intensive attention in recent years in CR networks, e.g., [26], [3], [17], [6], [7], [9], [10], [11], [12]. Almost all of them assume the CSI among CR users and from CR users to PUs is known[1]. However, the CSI estimation among CR users and/or PUs is a hard problem [13][2] that researchers usually ignored. Especially, in the scenario of autonomous CR networks, obvious cooperation among PUs and CR users are not permitted in CR. The *pricing* scheme in CR networks has been considered in recent literatures [6], [7], [9], [10], [11], [12], [18], [26]. Pricing has the advantage to improve the spectrum utilization. However, it requires cooperative behaviors among CR users and/or PUs, and the communication cost is high.

Recently, the research community shows more and more interests on designing fully distributed power control algorithms for CR networks. This is especially important for competitive and autonomous CR users for future wireless communications in hostile radio-access environments. Two of the key research issues are as follows: a) how to reduce or eliminate the overhead of CSI and power strategy information exchange among CR users that grants a robust distributed network; b) how to effectively estimate and explore the CSI from PUs for CR users' power control. For the research issue b), several works have been proposed while not much on the research issue a). In [13], the authors proposed a novel power control scheme to maximize the capacity of a *single* CR user by exploring the PU's CSI, which is based on the measurement of the average interference-power constraint at the PU-Rx. In [16], the authors considered the scenario on utilizing the PU's ACK/NAK information to maximize the utility of a *single* CR user. However, both [16] and [13] considered only a single SU. In [15], the authors considered the power control of multiple CR users by letting them listen to the PU's feedback channel as an external inference signal for coordination among

[1]Usually, researchers assume the CSI estimation among the CR-Tx and CR-Rx is via classical channel training, estimation, and feedback mechanisms. For CSI estimation from the PU-Tx to the CR-Rx, they assume CR users have the preknowledge on the PR-Tx power level and the channel reciprocity

[2]The PUs' CSI to the CR-Tx can be obtained by applying, 1) eavesdropping the CSI feedback from the PU-Rx to the PU-Tx; 2) the feedback from a cooperative sensing node located in the vicinity of the PU-Rx.

distributed CR transmitters. This distributed power control algorithm can approximate the optimal solution without PU cooperation, central controller/monitor, or inter-SU message passing. However, the proposed algorithm is only effective for the *sum-utility* maximization, where the utility of each CR user only requires the CSI of itself. The utility function in [15] does not take other CR users' interference into account, which avoids the interference CSI estimation among CR users.

The formulated problem and the proposed algorithm in [15] have the drawback of inaccurate throughput characterization, which provides a *suboptimal* solution for CR networks. Moreover, the *sum-utility* maximization is not suitable for competitive and autonomous CR users. The learning-based power control proposed in this paper posses distinctions from previous works in the following aspects: 1) this work provides a solution for the *incomplete-information* power control for *competitive* and *autonomous* CR users for the first time; 2) different from [15], our work considers the interference among multiplier CR users (attacks the research issue b)) that provides an accurate throughput characterization for each CR user; 3) this work discloses the relationship between the incomplete-information power control game and the classic complete-information power control game; 4) the convergence and learning rate upper-bound is provided theoretically; 5) the algorithm is robust and adaptive to dynamic wireless environments and varying PUs' activities.

Learning-based power control game has also been studied in other wireless networks. In [25], a non-cooperative power control algorithm with repeated games was studied for ad hoc networks. The authors provided the important insight that a felicitous intelligent learning behavior with self-incentive dynamics could eventually converge to steady state with a satisfactory system performance. In [26], a distributed discrete power control problem was formulated as an N-person nonzero sum game. The proposed stochastic learning power control algorithm were proved to converge to a stable Nash equilibrium.

In our earlier work [28], a framework for multiple CR users to perform distributed power control through reinforcement learning is presented. In this study, more thorough researches are illustrated on the learning mechanism for wireless environments, the properties and advantage of applying the Bush-Mosteller reinforcement scheme, and the convergence and learning rate analyses for the proposed power control algorithm.

III. Repeated Constrained Power Control Game

A. System Model as learning Automaton

The system model is illustrated in Fig. 1, where there are N CR transceiver pairs playing the repeated power control game with PUs over time. There are M PUs in the network region. Each PU's behavior might be changing with time, where its activity is distributed according to some probability that is not known to CR users. The interference CSI condition among CR users and the CSI from each CR user to each PU are unknown, and they can change in time. We consider a synchronous slotted time structured spectrum access for PUs and CR users as in [15] [16] during the long time repeated

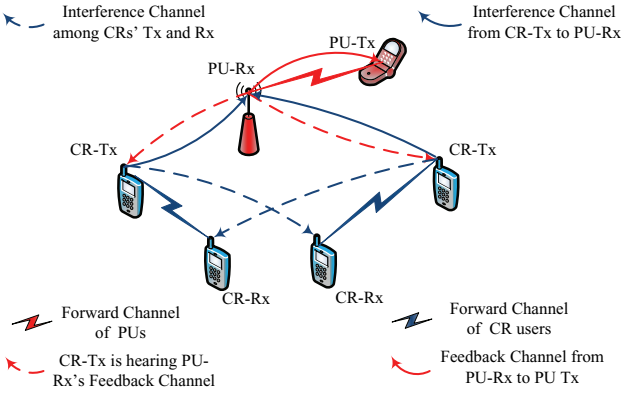


Fig. 1. CR users explore PUs-RX's feedback information; they don't have channel and power strategy information from other CR users and to PUs.

game. At each time slot t, each CR transmitter forwards its data with transmit power strategies from $\mathsf{P}_k^t \in \mathbf{P}_k \triangleq \{\mathsf{P}_k(1), \mathsf{P}_k(2), ..., \mathsf{P}_k(N_k)\}$, which is the action set of the learning automaton of user k; and N_k is the number of total actions of the CR user k. The transmit power space for the CR network is denoted as $\mathbf{P} = \prod_{k=1}^{N} \mathbf{P}_k$. CR users can utilize its feedback channel (Rx-to-Tx) to estimate its own CSI. Let us denote $s_k^t = (b_l^t, G_{kk}^t)$ as the CSI of the CR user k at time t, where $b_l^t = 0$ ($b_l^t = 1$) indicates the inactive (active) of PU l's transmission at time t, and G_{kk}^t denotes the channel gain of the CR user k. Since we assume the blind CSI from other CR users and PUs, the expected payoff u_k^t and the interference function (from the CR user k to the PU l) $\eta_{k,l}^t(b_l^t)$ for each CR user is unknown before the game.

For the PUs' communications, after every primary data transmission, PU-Rxs will feedback an interference term that includes the background noise and the collection of the sum-interference made from CR users. This information is used by PUs to monitor the interference level of the environment for the notification of the violation of interference power constraint messages[3] to CR networks and to conduct their own transmit power control. CR users are assumed to have the capability to explore this feedback information from the control links of PUs as in [13] [15] [16]. The exploration process of this feedback information can have errors and contain not updated values some time[4]. As for the wireless environments, the wireless channel can be modeled as a finite-state Markov channel (FSMC) [20], where fading can be approximated as a discrete-time Markov process with time discretized to a given time interval [21]. Thus, the behavior of each player (or user) in the wireless fading environment can be modeled by a *stochastic variable learning automation* [30] that consists a Markov chain of finite states.

[3]This is essentially the same as in [15] [16]. They assume the PUs measures all the sum-interference from all CR users, and the PU-Rx calculate and feedback the channel outage ACK/NACK to the PU-Tx in fading environments. Our assumption is of the same functionality as in [15] [16], however, it is more feasible for general wireless environments that includes fading and no-fading conditions (See Definition 1 in Section I). The feedback background noise (sum-interference) information is a value of several bits that used for the power control and interference management in the PU-Tx side.

[4]In the following analysis, this means the probability of error exploration of the interference level approaches to 0 almost surely, mathematically.

Let (Ω, H, P) be a probability space, where Ω is the sample space; H is a minimal σ-algebra on subsets of Ω; and P is a probability measure on (Ω, H). The symbol ω denotes an event in the probability space Ω. The stochastic automaton operating in a dynamic wireless environment is a adaptive discrete machine described by the tuple $\{\Xi, \mathbf{P}_k, \{\mathsf{P}_k^t\}, \{s_k^t\}, \{u_k^t\}, \{\eta_k^t(b_k^t)\}, \{p_k^t\}\}$, where Ξ is the automation input bounded set; $\mathbf{P}_k, \{\mathsf{P}_k^t\}, \{s_k^t\}, \{u_k^t\}, \{\eta_{k,l}^t(b_l^t)\}$ are, the action set of transmit power automaton (as we defined above), the sequences of automaton outputs (actions) of transmit power, input channel states, input utilities/payoffs and input interference functions, respectively. Let $p_k^t \triangleq [p_k^t(1), p_k^t(2), ..., p_k^t(N_k)]^T$ denote the mixed strategy of transmit power of user k at time t, where N_k is the number of transmit power strategies of user k. The mixed strategy p_k^t is the conditional probability distribution at time t

$$p_k^t(i) = \Pr\{\omega \in \Omega : \mathsf{P}_k^t = \mathsf{P}_k^t(i)/H_{t-1}\}, \sum\nolimits_{i=1}^{N_k} p_k^t(i) = 1.$$

where
$\mathrm{H}^t = \sigma(\mathsf{P}_k^1, s_k^1, u_k^1, \eta_{k,l}^1(b_l^1), p_k^1; ...; \mathsf{P}_k^t, s_k^t, u_k^t, \eta_{k,l}^t(b_l^t), p_k^t;$
$b_k^t = 0, 1, k = 1, .., N, l = 1, .., M)$ is a σ-algebra that generated by all the histories event over the action sets of CR users. Let $\mathrm{T}^t = \mathrm{T}_k^t$ represent a reinforcement learning (updating) scheme of each CR user which changes the probability vector p_k^t to p_k^{t+1}, that is

$$\mathrm{p}_k^{t+1} = \mathrm{p}_k^t + \gamma_k^t \mathrm{T}^{\mathrm{t}}, \tag{1}$$

where $\mathrm{T}^t = \mathrm{T}^t(p_k^t; \{s_k^t\}^{t=1,...,t}, \{u_k^t\}^{t=1,...,t}, \{\eta_k^t(b_k^t)\}^{t=1,...,t}, \{\mathsf{P}_k^t\}^{t=1,...,t})$, and γ_k^t is a scaler correction factor. The vector $\mathrm{T}_k^t(.) = [\mathrm{T}_{k,1}^t(.), ..., \mathrm{T}_{k,N_k}^t(.)]^T$ satisfies the following conditions that guarantees the valid probability distribution:

$$\sum\nolimits_{i=1}^{N_k} \mathrm{T}_{k,i}^t(.) = 0, \quad p_k^t(i) + \gamma_k^t \mathrm{T}_{k,i}^t(.) \in [0, 1]. \tag{2}$$

for any t and $i = 1, ..., N_k$. The above is the kernel of the the *learning automation* (LA).

B. Repeated Game Behavior

Non-cooperative power control is inherently a repeated process. Each user collects locally the information from the network and then makes the decision of transmit power levels locally. When new information (perhaps dependent on actions in the prior periods) is available, the decision process repeats itself. The learning-based repeated power control game is played in the following way. Consider a noncooperative game in which each player or CR user has only known its own payoff (or utility) realization u_k^t, and $\eta_l^t(b_l^t)$ the realization of interference function (interference term) explored from PUs after the game. At each stage (or time) t, each CR user will simultaneously and independently choose an action according to the probability distribution (mixed strategy) p_k^t, for instance, the transmit power $\mathsf{P}_k(i_k), 1 \le i_k \le N_k, 1 \le k \le N$. Let the achievable data transmission rate r_k^t of CR user k applied to the instant utility:

$$\begin{aligned} r_k^t(\mathbf{P}^t) &= W\log_2(1 + K\mathsf{SINR}(\mathbf{P}^{\mathrm{t}})) \\ &= W\log_2(1 + \frac{K\mathsf{P}_k^t G_{kk}(s_k^t)}{\sum_{j\neq \mathrm{k}}^{j\in N} \mathsf{P}_j^t G_{kj}(s_j^t) + \mathrm{n}_k^t}) \\ &= W\log_2(1 + \frac{K\mathsf{P}_k^t G_{kk}(s_k^t)}{\sum_{j\neq \mathrm{k}}^{j\in N} \mathsf{P}_j^t G_{kj}(s_j^t) + \sigma_k^2 + \sum_{l=1}^{M} \delta_{k,l} b_l^t}) \end{aligned}, \tag{3}$$

where $\mathbf{P}^t = (\mathsf{P}_1^t, \mathsf{P}_2^t, .., \mathsf{P}_N^t) \in \mathbf{P}$. The symbol W denotes the bandwidth of the primary spectrum. K is the SINR gap between a practical data transmission rate and the Shannon capacity limit. The function $G_{kj}(s_j^t)$ represents the channel gain from CR user k to CR user j in channel state s_j^t. The factors $G_{kj}, 1 \le j \le N$, incorporate propagation loss (slow-varying and fast-varying parts), spreading gain, and other normalization constants. Notice that G_{kk} is the path gain on link k. Assuming that not too many close-by nodes transmit at the same time, with reasonable spreading gain, G_{kk} is much larger than other G_{kj}, $(G_{kk} \gg G_{kj})$ $j \neq k$. This assumption is the same as in [27], and we denote it as the *interference non-dominance* condition. The notation n_k^t represents the instant noise of CR user k at time t, in which σ_k^2 is the noise power under the assumption of additive white Gaussian distribution, $\delta_{k,l}$ is the interference from PU l's transmission imposed on CR user k, and b_l^t is the indicator function of the existence of primary transmission.

Since each CR user only knows its own transmit power strategy, the mixed instant utility of CR user k is given as follows

$$\mathrm{u}_k^t(p_k^t) = \sum\nolimits_{i_k=1}^{N_k} \mathrm{r}_k^t(\mathsf{P}_k^t(i_k)) \cdot p_k^t(i_k). \tag{4}$$

Similarly, the instant and the mixed instant interference function of CR user k imposed on PU l is given by

$$\eta_{k,l}^t(\mathsf{P}_k^t(i_k)) = \mathsf{P}_k^t G_{kl}(s_k^t). \tag{5}$$

$$\eta_{k,l}^t(p_k^t) = \sum\nolimits_{i_k=1}^{N_k} b_l^t \cdot \eta_{k,l}^t(\mathsf{P}_k^t(i_k)) \cdot p_k^t(i_k). \tag{6}$$

Then, for each CR user, it changes its mixed strategy according to the accepted reinforcement $\mathrm{p}_k^t \xrightarrow{T_k^t} \mathrm{p}_k^{t+1}$ trying for large t to maximize its individual averaged utility $\mathrm{liminf}_{t\to\infty}\, \mathrm{u}_k^t$ where

$$\Phi_k^t = \frac{1}{t} \sum\nolimits_{\tau=1}^{\tau=t} \mathrm{u}_k^\tau(p_k^\tau), \tag{7}$$

and the sum-interference of all CR users measured by the PU l at time t is

$$\eta_l^\tau(\mathbf{p}^\tau) = \sum\nolimits_{k=1}^{k=N} \eta_{k,l}^\tau(p_k^\tau). \tag{8}$$

Its time average should conform to the interference power constraint C_l. If the *interference-power-constraint-violation* information is reported in its feedback control channel at time t, this information (eavesdropped by CR users) will reinforce the CR users to change their power control strategies in the long run to conform to C_l:

$$\Psi_l^t = \frac{1}{t} \sum\nolimits_{\tau=1}^{\tau=t} \eta_l^\tau(\mathbf{p}^\tau), \mathrm{liminf}_{t\to\infty}\, \Psi_l^t \le C_l. \tag{9}$$

Note that in the above learning game, each individual CR user k only known its own utility and the interference power level that is eavesdropped from PUs' feedback control link after the play in each step. It chooses the transmit power strategy $\mathsf{P}_k^t(i_k)$ with probability $p_k^t(i_k)$ according to the reinforcement learning scheme described in the previous subsection.

In wireless communication researches, the wireless channels or environments are assumed to be stationary. Hence this assumption on a sequence of real functions (i.e., utility and

interference function in this work) is just one special case needed in order for the results in the stochastic analysis to be true. In practice, the wireless environment can experience with non-stationary observation of noise and interference. Let us consider the non-stationary environment, which is characterized by the following two properties:

Assumption 1: The conditional expectations of u_k^t and η_l^t of the environment responses exist, i.e.,

$$\begin{aligned}&\mathrm{E}\{\mathrm{u}_k^t|\mathsf{P}_k^t{=}\mathsf{P}_k(\mathrm{i}_k)\wedge \mathrm{H}_{t-1},\forall \mathrm{k}\in \mathrm{N}\}={r}_k^{i_1,\ldots,i_N}(t)\\&\quad= W\log_2(1+K\mathsf{SINR}(\mathsf{P}^t(i_1),...,\mathsf{P}^t(i_k),...,\mathsf{P}^t(i_N)))\\&\mathrm{E}\{\eta_l^t|\mathsf{P}_k^t{=}\mathsf{P}_k(\mathrm{i}_k)\wedge \mathrm{H}_{t-1},\forall \mathrm{k}\in \mathrm{N}\}={\eta}_l^{i_1,\ldots,i_N}(t)\\&\quad= b_l^t\cdot\left(\textstyle\sum_{\mathrm{k}=1}^{N}\sum_{i_k=1}^{N_k}\mathsf{P}_k^t(\mathrm{i}_k)G_{kl}(s_l^t)\cdot p_k^t(i_k)+\sigma_l^2\right).\end{aligned}\tag{10}$$

and their arithmetic averages tend to a finite limit, with probability one,i.e.,

$$\begin{aligned}&\tfrac{1}{t}\textstyle\sum_{\tau=1}^{\tau=t}{r}_k^{i_1,\ldots,i_N}(\tau)\overset{a.s.}{\underset{t\to\infty}{\to}}{r}_k^{i_1,\ldots,i_N}\\&\tfrac{1}{t}\textstyle\sum_{\tau=1}^{\tau=t}{\eta}_l^{i_1,\ldots,i_N}(\tau)\overset{a.s.}{\underset{t\to\infty}{\to}}{\eta}_l^{i_1,\ldots,i_N}.\end{aligned}\tag{11}$$

The notation ${r}_k^{i_1,\ldots,i_N}(t)$ represents the expected transmission rate of the CR user k given a specific transmit power strategy of $i_1,...,i_N$ for all the N CR users. Similarly, we have the expected interference level ${\eta}_l^{i_1,\ldots,i_N}(t)$ measured by the PU l. The first part of the assumption A1 (10) states some restrictions to the properties of the observation noise. Take the η_l^t in (10) as an example, the condition will be satisfied if at every time t the noise σ_l are bounded in second moments, i.e, $\mathrm{E}\{\sigma_1^2\}<\infty$. This property holds, for example, for Gaussian noises (it is assumed in (3)) and is not true for noise having Cauchy distribution. The second property of assumption A1 (11) represents some sort of stationary in average. It will be satisfied for the utility, interference function and noise, which are stationary in average.

Assumption 2: For any realized action of the transmit power P_k^t at current time t in the repeated power control game, the conditional variance of the utility u_k^t and the interference functions η_l^t of the CR user k are uniformly bounded with probability one, i.e.,

$$\begin{aligned}&E\left\{\left(u_k^t-{r}_k^{i_1,\ldots,i_N}(t)\right)^2/H_{t-1}\wedge\mathsf{P}_k^t{=}\mathsf{P}_k(\mathrm{i}_k)\right\}=\sigma_{t,k}^{+},\\&\quad\max_k\sup_t\sigma_{t,k}^{+}=\sigma_{u,k}^{+}<\infty,\forall k=1,...,N\\&E\left\{\left(\eta_l^t-{\eta}_l^{i_1,\ldots,i_N}(t)\right)^2/H_{t-1}\wedge\mathsf{P}_k^t{=}\mathsf{P}_k(\mathrm{i}_k)\right\}=\sigma_{t,l}^{+},\\&\quad\max_k\sup_t\sigma_{t,l}^{+}=\sigma_{\eta,l}^{+}<\infty,\forall k=1,...,N\end{aligned}\tag{12}$$

The notation $\sigma_{u,k}^{+}$ and $\sigma_{\eta,l}^{+}$ represent the upper bounds of the transmission rate of the CR user k and interference function measured by the PU l. Since the number of transmit power strategies for each CR user is finite and the channel gains of wireless environment (including any fading models) are bounded almost surely [21], the upper bounds $\sigma_{u,k}^{+}$ and $\sigma_{\eta,l}^{+}$ must exist.

Definition 1: A random environment, satisfying conditions (10)$\sim$(12) is said "***asymptotically stationary in the average sense***".

For each CR user, the average utility and interference function from the collection of 2 tensors $U_k=[{r}_k^{i_1,\ldots,i_N}(t)],(1\le k\le N,1\le i_k\le N_k)$ and $Q_l=[{\eta}_l^{i_1,\ldots,i_N}(t)],(1\le l\le M,1\le i_k\le N_k)$ which are *priori unknown*. Thus, we can define the expected utility and expected constraints concerning with all PUs' and/or CR users' CSI and transmit power strategy information as follows.

Definition 2*:* At time t, for the CR user k in the considered game, and for any H_{t-1}-measurable conditional probability distribution $(\mathrm{p}_1^t,...,\mathrm{p}_N^t)$, the *expected utility* is defined as

$$\begin{aligned}U_k(p_1^t,...,p_N^t)&=\mathrm{E}[{r}_k^{i_1,\ldots,i_N}(t)]\\&=\sum_{i_1=1}^{N_1}...\sum_{i_N=1}^{N_N}{r}_k^{i_1,\ldots,i_N}(t)\prod_{s=1}^{N}p_s^t(i_s)\\&=W\log_2(1+\tfrac{K\mathsf{P}_k^t\cdot p_k^tG_{kk}(s_k^t)}{\sum_{j\neq k}^{j\in N}\mathsf{P}_j^t\cdot p_j^tG_{kj}(s_j^t)+\mathrm{n}_k^t})\\&=W\log_2(1+\tfrac{K\sum_{i_k=1}^{N_k}\mathsf{P}_k^t(i_k)p_k^t(i_k)G_{kk}(s_k^t)}{\sum_{j\neq k}^{j\in N}\sum_{i_j=1}^{N_j}\mathsf{P}_j^t(i_j)p_j^t(i_j)G_{kj}(s_j^t)+\mathrm{n}_k^t})\end{aligned},$$

where $\mathsf{P}_j^t\cdot p_j^t=\sum_{i_j=1}^{N_j}\mathsf{P}_j^t(i_j)p_j^t(i_j),1\le j\le N$ is the expected transmit power of the mixed strategy of the CR user j at time t, and the corresponding *expected interference function* is defined as

$$\begin{aligned}Q_l(p_1^t,...,p_N^t)&=\mathrm{E}[{\eta}_l^{i_1,\ldots,i_N}(t)]\\&=\sum_{i_1=1}^{N_1}...\sum_{i_N=1}^{N_N}{\eta}_l^{i_1,\ldots,i_N}(t)\prod_{s=1}^{N}p_s^t(i_s)\\&=b_l^t\left(\sum_{\mathrm{k}=1}^{N}\mathsf{P}_k^t\cdot p_k^tG_{kl}(s_l^t)+\sigma_l^2\right)\\&=b_l^t\left(\sum_{\mathrm{k}=1}^{N}\sum_{i_k=1}^{N_k}\mathsf{P}_k^t(i_k)p_k^t(i_k)G_{kl}(s_l^t)+\sigma_l^2\right)\end{aligned}.$$

We also have the *averaged expected utility* and *interference functions* over time t as follows:

$$\begin{aligned}U_k^t&=\tfrac{1}{t}\textstyle\sum_{\tau=1}^{\tau=t}{U_k}^{\tau}(p_1^{\tau},...,p_N^{\tau}),\\Q_l^t&=\tfrac{1}{t}\textstyle\sum_{\tau=1}^{\tau=t}Q_l^{\tau}(p_1^{\tau},...,p_N^{\tau}).\end{aligned}\tag{13}$$

The following lemma indicates the asymptotic equivalence between (Φ_k^t,Ψ_l^t) and (U_k^t,Q_l^t).

Lemma 1: Under the assumption 1, for any conditional distribution $(\mathrm{p}_1^t,...,\mathrm{p}_N^t)$, it has

$$\Phi_k^t=U_k^t+o_\omega(t^{-1/2}),\Psi_l^t=Q_l^t+o_\omega(t^{-1/2}).$$

Proof: The proof is obtained in the view of Borel-Cantelli lemma [23] and the strong law of large numbers for dependent sequence [24]. $o_\omega(t^{-1/2})$ is a random sequence tending almost surely to zero, and more quickly than $t^{-1/2}$. ■

IV. Nash Equilibrium and Learning Game: Problem Statement

A. Mixed Strategies and Nash Equilibrium

The mixed power control strategy of the CR user k is any sequence of vectors $\mathrm{W}_k=\{p_k^t|\sum_{i=1}^{N_k}p_k^t(i)=1,1\le k\le N\}$ with H_{t-1} measurable components belonging to the simplex $S_{\varepsilon=0}^{N_k}$, i.e.,

$$p_k^t\in S_{\varepsilon=0}^{N_k}=\{p_k^t\in R^{N_k}:p_k^t(i)\ge\varepsilon\ge 0,\sum_{i=1}^{N_k}p_k^t(i)=1\}.\tag{14}$$

Definition 3: The power control strategies $W_1^*,...,W_N^*$ in the noncooperative game are defined as Nash Equilibrium if:

1) The power control strategies are *admissible*, where the interference function Q_l is less than the PU l's interference power constraint C_l, i.e., for all $1 \le l \le M$

$$\max_{p_s^1, s\in N} \limsup_{t\to\infty} Q_l^t = \tilde{Q}(W_1^*, ..., W_N^*) \le C_l, \qquad (15)$$

where C_l are *priori* know values[5];

2) for $\forall k \in N$ and any admissible strategy W_k

$$U(W_1^*, ..., W_N^*) = \min_{p_s^1, s\in N} \liminf_{t\to\infty} U_k^t \ge U(W_1^*, ...W_k..., W_N^*), \qquad (16)$$

where the maximization of (15) and the minimization of (16) are guaranteed over all the initial probability distributions $p_s^1, s \in N$.

Definition 4: The point $(p_1^*, ..., p_N^*)$ is said to be a Nash Equilibrium point of the given N-user CR power control game within the class of mixed strategies W_k, if for each $1 \le k \le N$

$$\begin{aligned} &U_k(p_1^*, ..., p_N^*) = \max_{p_k \in R^k} U_k(p_1^*, ..., p_{k-1}^*, p_k, p_{k+1}^*, ..., p_N^*) \\ &s.t.\ R^k = \bigcap_{l=1}^{M} \{p_k : Q_l(p_1^*, ..., p_{k-1}^*, p_k, p_{k+1}^*, ..., p_N^*) \le C_l\}. \end{aligned} \qquad (17)$$

At the Nash Equilibrium point, no player can increase his payoff by a unilateral change in its strategy.

Definition 5: A mixed strategy is said to be *stationary*, where $\mathrm{T}_k = \{p_k | \sum_{i=1}^{N_k} p_k(i) = 1, 1 \le k \le N\}$ is not changing with time after the learning of the repeated game process. The set of distributions $(p_1^*, ..., p_{\mathrm{N}}^*)$ is called the *stationary* mixed strategy of Nash Equilibrium.

Remark 1: The set of all equilibrium strategies W_k^*, where $1 \le k \le N$, contains the subsets of *admissible stationary strategy* $\{p_k^*\}$ for any $1 \le k \le N, 1 \le l \le M$ and for any $p_k \in S_{\varepsilon=0}^{N_k}$ such that,

$$\begin{aligned} &U_k(p_1^*, ..., p_N^*) \ge U_k(p_1^*, ..., p_{k-1}^*, p_k, p_{k+1}^*, ..., p_N^*). \\ &Q_l(p_1^*, ..., p_{k-1}^*, p_k, p_{k+1}^*, ..., p_N^*) \ge C_l. \end{aligned} \qquad (18)$$

This fact follows immediately from: 1) the continuity property of the utility function U_k and the interference function Q_l 13; 2) the compactness of the simplexes $S_{\varepsilon=0}^{N_k}$; 3) the Nash theorem [29, Th. 1, p.268].

B. Game as Reinforcement Learning

Now we can formulate the repeated power control game problem with a *priori* unknown average utility and constraints. The objective of this power control game is: based on the current information (realized utility and interference), generate the mixed admissible strategy $\{p_k^t\}, \forall k \in N$ to achieve a realizable Nash Equilibrium within the subclass of stationary strategies.

To achieve this objective, we need to emphasize the following fact at first. According to the Nash theorem [29], the set of stationary distribution $(p_1^*, ..., p_N^*)$ that satisfies (18) might contain more than one element. For one of the Nash Equilibrium distribution $(p_1^*, ..., p_N^*)$, the objective is to show if an admissible mixed strategy $W_k, \forall k \in N$ converges to such distribution $(p_1^*, ..., p_N^*)$, then the associated random function of both utility and interference Φ_k^t (7)and Ψ_l^t (9) also converge to the corresponding expected stationary values $U_k(p_1^*, ..., p_N^*)$ and $Q_l(p_1^*, ..., p_N^*)$, respectively. This result is summarized in the following theorem.

Theorem 1: Under the *assumption A1* and *A2*, there exists a strategy $W_k, \forall k \in N$ that is asymptotically stationary realizing a Nash equilibrium:

$$\limsup_{n\to\infty} E\left\{\frac{1}{t}\sum_{\tau=1}^{t}\sum_{k=1}^{N} ||p_k^t - p_k^*||\right\} \overset{a.s.}{=} 0. \qquad (19)$$

Then,

$$\begin{aligned} &\limsup_{n\to\infty} E\left\{\frac{1}{t}\sum_{\tau=1}^{t}\sum_{k=1}^{N} |\Phi_k^t - U_k(p_1^*, ..., p_N^*)|\right\} \overset{a.s.}{=} 0, \\ &\limsup_{n\to\infty} E\left\{\frac{1}{t}\sum_{\tau=1}^{t} |\Psi_l^t - Q_l(p_1^*, ..., p_N^*)|\right\} \overset{a.s.}{=} 0. \end{aligned} \qquad (20)$$

Proof: Let us consider the following sequence,

$$\begin{aligned} S_1^t &= \tfrac{1}{t}\sum_{\tau=1}^{t}\sum_{k=1}^{N} |\Phi_k^t - U_k(p_1^*, ..., p_N^*)| \\ S_{l,2}^t &= \tfrac{1}{t}\sum_{\tau=1}^{t} |\Psi_l^t - Q_l^t(p_1^*, ..., p_N^*)| \end{aligned}. \qquad (21)$$

According to the assumption A1, it follows

$$\begin{aligned} U_k^t(p_1, ..., p_N) &= E\{\Phi_k^t / H_{t-1}\} \\ Q_l^t(p_1, ..., p_N) &= E\{\Psi_l^t / H_{t-1}\} \end{aligned}. \qquad (22)$$

Then, using the above equality (22), we derive the recurrent form of (21).

$$\begin{aligned} S_1^t &= \left(1 - \tfrac{1}{t}\right) S_1^{t-1} + \tfrac{1}{t} |E\{U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1}\}| \\ S_{l,2}^t &= \left(1 - \tfrac{1}{t}\right) S_{l,2}^{t-1} + \tfrac{1}{t} |E\{Q_l^t(p_1^*, ..., p_N^*) - \Psi_l^t / H_{t-1}\}| \end{aligned}.$$

Let us consider S_1^t first. Take the conditional expectation of $(S_1^t)^2$, and in the view of the assumption A2, we derive the following equation,

$$\begin{aligned} &E\left\{(S_1^t)^2 | H_{t-1}\right\} = \left(1 - \tfrac{1}{t}\right)^2 \left(S_1^{t-1}\right)^2 \\ &+ \tfrac{2}{t}\left(1 - \tfrac{1}{t}\right) E\{S_1^t | U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1} | / H_{t-1}\} \ldots (a) \\ &+ \tfrac{1}{t^2} E\left\{|U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1}|^2 / H_{t-1}\right\} \ldots\ldots\ldots\ldots (b) \end{aligned}$$

In part (a) of the above equation, according to *Cauchy-Schwarz inequality*, we have

$$\begin{aligned} &E\{S_1^t | U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1} | / H_{t-1}\} \\ &\le \sqrt{E\left\{(S_1^t)^2 / H_{t-1}\right\}} \sqrt{E\left\{(|U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1}|)^2\right\}}. \end{aligned}$$

Since the transmission rate function $U_k^t(p_1^*, ..., p_N^*)$ are smooth enough, and it can be checked that U_k^t satisfies the following Lipschitz continuity condition.

$$\sqrt{E\left\{(|U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1}|)^2\right\}} \le \beta \sum_{k=1}^{N} \left\|p_k^t - p_k^*\right\|,$$

where β is a constant. According to the assumptions A1 and A2, $E\left\{(S_1^t)^2 / H_{t-1}\right\}$ must be bounded by a constant value. In part (b), in the view of the assumption A2, we have

$$|U_k^t(p_1^*, ..., p_N^*) - \Phi_k^t / H_{t-1}|^2 \le \sigma.$$

[5]Or can also be assumed unknown, in this case, when the CR users violate this constraint, PUs will send the interference-constraints-violation message, and CR users can explore the interference level.

So, accordingly, we have the following inequality for the conditional expectation of $(S_1^t)^2$,

$$E\left\{(S_1^t)^2\,|H_{t-1}\right\}\leq\left(1-\frac{1}{t}\right)(S_1^{t-1})^2+\frac{C_1}{t}\left\|p_k^t-p_k^*\right\|+\frac{C_2}{t^2}.$$

By using the similar procedure, we can obtain the same inequality for the interference power part, $S_{l,2}^t$.

In the view of lemma A. 11 [39], this inequality leads to

$$S_1^t\overset{a.s.}{\underset{t\to\infty}{\to}}0,\qquad S_{l,2}^t\overset{a.s.}{\underset{t\to\infty}{\to}}0. \tag{23}$$

Using lemma A. 13 [39] and considering the mathematical expectation of both sides of the previous inequality, we obtain the desired results (20). ■

Remark 2: This result shows that in the wireless environment, which is "asymptotically stationary in the average", the convergence of stationary Nash equilibrium distribution of $(p_1^*,...,p_N^*)$ always guarantee the convergence of transmission rates and interference functions. The proof of this theorem relies on the application of Robbins-Siegmund theorem (Robbins and Siegmund 1971) [40] in probability theory, a convergence theorem for dependent random sequences.

Therefore, if we can construct an asymptotically (with time) stationary mixed strategy $\{p_k^t\}$ converging to a stationary distribution $(p_1^*,...,\mathrm{p_N^*})$ that realizes a Nash equilibrium, we will achieve the main goal of this learning-based repeated power control game. However, as we know from the game theory, the set of stationary distributions $(p_1^*,...,\mathrm{p_N^*})$ that satisfying (17) might not exist, or unique. To attain our main goal in a rigorous manner, at first, the problem related to the existence and uniqueness of the Nash equilibrium should be solved. In the next section, we deal with these problems.

V. Properties of Nash Equilibrium

A. Existence of Nash Equilibrium

To justify the correctness of Lagrange multipliers implementation and existence of nonempty set of admissible transmit power strategies, assume that the given interference constraints satisfy the slater's condition providing the Kuhn-Tucker constraint qualification [31].

Assumption A3: There exists a *feasible* strategy vectors $(\tilde{p}_1^t,...,\tilde{p}_N^t)$ such that $1\leq k\leq N$ and $1\leq l\leq M$, the following inequalities hold:

$$Q_l(\tilde{p}_1^t,...,\tilde{p}_N^t)\leq C_l. \tag{24}$$

The next theorem shows the main result on the existence of Nash Equilibrium of the power control game.

Theorem 2: An Nash equilibrium point (that satisfies (17)) strategy exists for any N-person game satisfying **(A1)** and **(A3)**.

Proof: It follows straight from Th.1 in [33]. Indeed, by **(A1)** the function

$$\begin{aligned}\rho(p,q)&=\sum_{k=1}^{N}U_k(p_1,...,q_k,...,p_N)\\&=\sum_{k=1}^{N}W\log_2(1+\tfrac{K\mathsf{P}_k\cdot q_kG_{kk}(s_k)}{\sum_{j\neq k}^{j\in N}\mathsf{P}_j\cdot p_jG_{kj}(s_j)+n_k^t})\end{aligned} \tag{25}$$

is well defined for any $p=(p_1,...,p_N)$ and $q=(q_1,...,q_N)$ from $R=(R^1\times...\times R^N)$. It is easy to prove $\rho(p,q)$ is continuous in p and q. We can further prove $\rho(p,q)$ is a strictly concave function in q for any fixed p.

Taking the derivative of $\rho(p,q)$ with respect to q_k, we have

$$\nabla_{q_k}\rho(p,q)=\frac{W\left(K\mathsf{P}_kG_{kk}(s_k)\right)}{\sum\limits_{\substack{j\in N\\ j\neq\mathrm{k}}}\mathsf{P}_j\cdot\mathrm{p}_jG_{kj}(s_j)+\mathrm{n}_k+K\mathsf{P}_k\cdot\mathrm{q}_kG_{kk}(s_k)}>0.$$

Taking the derivative of the above equation again, we have the equation at the top of the following page. We obtain the Hessian:

$$\mathbf{H}_k=diag(\nabla_{q_1}^2\rho(p,q),...,\nabla_{q_N}^2\rho(p,q)). \tag{26}$$

Matrix $\mathbf{H}_k$ is obviously *negative* definite: for all vectors v,

$$\mathbf{v}^T\mathbf{H}_k\mathbf{v}=diag(\nabla_{q_1}^2\rho(p,q)v_1^2,...,\nabla_{q_N}^2\rho(p,q)v_N^2). \tag{27}$$

Therefore, $\rho(p,q)$ is a strictly concave function in q for any fixed p within R. Notice that R is convex, compact and nonempty by **(A3)** and (24). Then, the point-to-set mapping $p\in R\to\Gamma p\subset R$ given by

$$\Gamma_p=\left\{q|\rho(p,q)=\max_{z\in R}\rho(p,z)\right\} \tag{28}$$

is upper semi-continuous in R, and, hence, by Kakutani fixed point theorem (see, e.g. [34]), there exists a point $p^*\in R$ such that $p^*\in\Gamma p^*$, that is

$$\rho(p^*,\ p^*)=\max_{z\in R}\rho(p^*,z). \tag{29}$$

This fixed point $p^*\in R$ satisfies (17). Indeed, suppose that (17) is not verified, e.g., for $k=k_0$, there exists a point p_{k_0} such that $p^*(k_0)=(p_1^*,...,p_{k_0},...,p_N^*)\in R$. We have $U_k(p_1^*,...,p_{k_0},....,p_N^*)>U_k(p_1^*,...,p_k^*,....,p_N^*)$ that implies $\rho(p^*,p^*(k_0))>\rho(p^*,p^*)$ which contradicts (29). ■

B. Uniqueness of Nash Equilibria

The condition for uniqueness of equilibria are known as *strict diagonal concavity*. According to Rosen's theorem 2 in [33], we show that if the given matrix game is "strictly diagonal concave", then the corresponding game becomes strictly diagonal concave, and the uniqueness of the equilibria policy follows.

1) Diagonal Concavity Properties: Let us define the function

$$W_r(p,q)=(q-p)\frac{\partial}{\partial p}\rho_r(p,q)+(p-q)\frac{\partial}{\partial q}\rho_r(p,q), \tag{30}$$

where $\rho_r(p,q)=\sum\limits_{k=1}^{N}r_kU_k(p_1,...,q_k....,p_N)$.

Definition 6: A matrix game is *diagonal concave* if there exists positive numbers r_k such that for any $p,q\in R$

$$W_r(p,q)\geq 0; \tag{31}$$

a matrix game is *strict diagonal concave* if there is a positive numbers r_k such that for any $p,q\in R$, the strict inequality holds above. The basic properties of the diagonal concave games can be found in [35].

$$\nabla^2_{q_k}\rho(p,q)=\frac{-W(K\mathsf{P}_k G_{kk}(s_k))^2}{\left(\sum\limits_{j\neq \mathrm{k}}^{j\in N}\mathsf{P}_j\cdot \mathrm{p}_j G_{kj}(s_j)+\mathrm{n}_k+K\mathsf{P}_k\cdot \mathrm{q}_k G_{kk}(s_k)\right)^2}<0.$$

2) Uniqueness of the Power Control Game: As already mentioned above, the equilibrium probability distribution may not be unique. We will prove that the proposed matrix game for CR power control is strict diagonal concave, and admit a unique Nash equilibrium

Lemma 2: The matrix game for CR user power control in the *interference non-dominance* condition is a strict diagonal concave game, i.e.,

$$W_r(p,q)>0. \tag{32}$$

Proof: Let us first expand the function $W_r(p,q)$,

$$\begin{aligned} W_r(p,q) &= (q-p)\tfrac{\partial}{\partial p}\rho_r(p,q)+(p-q)\tfrac{\partial}{\partial q}\rho_r(p,q)\\ &=\tfrac{Wr\cdot(q-p)(KP_kG_{kk}(s_k))}{\sum_{j\neq \mathrm{k}}^{j\in N}P_j\cdot \mathrm{q}_jG_{kj}(s_j)+\mathrm{n}_k+KP_k\cdot \mathrm{p}_kG_{kk}(s_k)}+\\ &\quad\tfrac{Wr\cdot(p-q)(KP_kG_{kk}(s_k))}{\sum_{j\neq \mathrm{k}}^{j\in N}P_j\cdot \mathrm{p}_jG_{kj}(s_j)+\mathrm{n}_k+KP_k\cdot \mathrm{q}_kG_{kk}(s_k)}\\ &=\tfrac{Wr\cdot(p-q)^2(KP_kG_{kk}(s_k))}{\left(\sum_{j\neq \mathrm{k}}^{j\in N}P_j\cdot \mathrm{q}_jG_{kj}(s_j)+\mathrm{n}_k+KP_k\cdot \mathrm{p}_kG_{kk}(s_k)\right)}\\ &\quad\cdot\tfrac{\left(KP_kG_{kk}(s_k)-\sum_{j\neq \mathrm{k}}^{j\in N}P_jG_{kj}(s_j)\right)}{\left(\sum_{j\neq \mathrm{k}}^{j\in N}P_j\cdot \mathrm{p}_jG_{kj}(s_j)+\mathrm{n}_k+KP_k\cdot \mathrm{q}_kG_{kk}(s_k)\right)}, \end{aligned} \tag{33}$$

where $r=(r_1,...,r_k,...,r_N)$. Since we have $G_{kk}(s_k)\gg G_{kj}(s_j)$ in the interference non-dominance condition, and K approaches to 1 in a median SINR value of transmissions, it can guarantee that $KP_kG_{kk}(s_k)-\sum_{j\neq \mathrm{k}}^{j\in N}P_jG_{kj}(s_j)$ in (33) are always positive for the CR users in the most typical weak interference high data rate (SINR) scenarios. Thus, that proves there must exist a r_k such that $W_r(p,q)>0$. ■

Theorem 3: The matrix game for CR user power control admits a unique Nash Equilibrium.

Proof: The proof is straight forward, since the power control game is strict diagonal concave according to lemma 2. It is a direct outcome of Rosen's theorem [33]. ■

Therefore, we can claim that the power control game converges to a unique equilibrium point, during the learning process.

C. Lagrange Multipliers Using Regularization Approach

To get the Nash equilibrium of the CR users' power control strategy in (17), Lagrange multiplier method can be used. The corresponding Lagrange function is given by

$$L_k(p_1,...,p_N;\lambda)=U_k(p_1,...,p_N)-\sum_{l=1}^{M}\lambda_l\left(Q_l(p_1,...,p_N)-C_l\right),$$

where $\lambda_l>0$ and the arguments p_k belong to the $\varepsilon-$simplexes (14). Since the interference power constraint (15) is not strictly convex (multilinearity of interference function) and, as a consequence, any attempt to directly apply the gradient technique for finding its saddle point may fail due to divergence.

One approach for avoiding this problem consists of introducing a regularization term in the corresponding Lagrange function [36]

$$L_k^{\delta}(\mathrm{p}_1,...,\mathrm{p}_N;\lambda)=L_k(\mathrm{p}_1,...,\mathrm{p}_N;\lambda)-\frac{\delta}{2}\left(\|p_k\|^2-\|\lambda_k\|^2\right), \tag{34}$$

where $\lambda_k=(\lambda_{k,1},...,\lambda_{k,M})^T$. These regularized functions are strictly concave with respect to p_k, and strictly convex with respect to r_k. The next theorem describes the dependence of equilibrium-point (or saddle-point) $(p_k^*(\varepsilon_k^t,\delta^t),\lambda_k^*(\varepsilon^t,\delta^t))$ of the regularized function with the regularizing δ_t,ε_k^t for convergence.

Theorem 4: If the sequence $\{\varepsilon_k^t\}$ and $\{\delta_t\}$ are such that

$$\begin{aligned} &\varepsilon_k^t\in(0,1/N_k),\delta^t>0,\lim_{t\to\infty}\delta^t=0,\varepsilon^t=(\varepsilon_1^t,...,\varepsilon_N^t)^T\\ &\lim_{n\to\infty}\varepsilon^t/\delta^t=\upsilon=(\upsilon_1,...,\upsilon_N)^T,\upsilon_k\in[0,\infty),n=1,2,... \end{aligned}$$

the sequence $(p_k^*(\varepsilon_k^t,\delta^t),\lambda_k^*(\varepsilon^t,\delta^t))$ of the equilibrium-points of the corresponding Lagrange function $L_k^{\delta}(\mathrm{p}_1,...,\mathrm{p}_N;\lambda)$ (34) converges to the equilibrium-point $(p_k^{**},\lambda_k^{**})$ of the initial Lagrange function $L_k(\mathrm{p}_1,...,\mathrm{p}_N;\lambda)$, which corresponds to the equilibrium-point (p_k^*,λ_k^*) and has the minimal norm, i,e.,

$$(p_k^*(\varepsilon_k^t,\delta^t),\lambda_k^*(\varepsilon_k^t,\delta^t))\to(p_k^*,\lambda_k^*),t\to\infty. \tag{35}$$

where

$$p_k^{**}=\arg\min_{p^*,\lambda^*}\frac{1}{2}(\|p^*\|^2+\|\lambda^*\|^2). \tag{36}$$

Proof: The proof is similar to [37], we omit it due to space limitation. ■

The following two lemmas are two useful conclusions related to Theorem 4, which will be used in the convergence analysis of the learning algorithm in Section VI.

Lemma 3 For any $p\in P_N$ and for any $\lambda\in R^M$ the following inequality holds:

$$\begin{aligned} &-\tfrac{\delta}{2}\sum_{k=1}^{N}r_k\left(\|p_k^t-p^*(\varepsilon_k^t,\delta^t)\|^2+\|\lambda_k^t-\lambda^*(\varepsilon_k^t,\delta^t)\|^2\right)\\ &\le\sum_{k=1}^{N}r_k[(p_k^t-p^*(\varepsilon_k^t,\delta^t))^T\tfrac{\partial}{\partial p_k^t}L_k^{\delta}(p_1,...,p_N;\lambda_n)\\ &-(\lambda_k^t-\lambda^*(\varepsilon_k^t,\delta^t))^T\tfrac{\partial}{\partial\lambda_n}L_k^{\delta}(p_1,...,p_N;\lambda_n)]\le 0. \end{aligned}$$

where (p_k^*,λ_k^*) is the equilibrium-point of the regularized Lagrange function.

Proof: The right-hand-side inequality is always hold due to the Kuhn-Tucker saddle point theorem [36]; the right-hand-side inequality is obtained according to the property of Kuhn-Tucker saddle point theorem and the Jensen's inequality [36]. The detailed proof is omitted due to space limitation. ■

Lemma 4: Under the same conditions as in Theorem 4, 1) all the possible equilibrium mixed policies can be parameterized by the nonnegative vector parameter $\upsilon\in R^N$ as $(p_1^*(\upsilon),...,p_N^*(\upsilon))$ and the sequences $(p_1^*(\varepsilon_1^t,\delta^t),...,p_N^*(\varepsilon_N^t,\delta^t))$ $\lambda_k^*(\varepsilon^t,\delta^t)$ to the unique equilibrium points $(p_1^*(\upsilon),...,p_N^*(\upsilon))$ and $\lambda_k^*(\upsilon)$, i.e.,

$$\sum_{k=1}^{N}r_k\Big(\left\|p_k(\upsilon)-p^*(\varepsilon_k^t,\delta^t)\right\|^2+\left\|\lambda_k(\upsilon)-\lambda^*(\varepsilon_k^t,\delta^t)\right\|^2\Big)\underset{n\to\infty}{\to}0.$$

2) There exists constants $c_i (i = 1, 2, 3)$ such that

$$\begin{aligned} &\sum_{k=1}^{N} r_k(\left\|p_k^*(\varepsilon_k^{t+1}, \delta^{t+1}) - p_k^*(\varepsilon_k^t, \delta^t)\right\|^2 + \\ &\left\|\lambda_k^*(\varepsilon^{t+1}, \delta^{t+1}) - \lambda_k^*(\varepsilon^t, \delta^t)\right\|^2) \\ &\leq c_1 \left\|\varepsilon^{t+1} - \varepsilon^t\right\| + \\ &c_2 \left\|\delta^{t+1} - \delta^t\right\| + c_3 \left\|\varepsilon^{t+1}/\delta^{t+1} - \varepsilon^t/\delta^t\right\| = \beta_t. \end{aligned} \tag{37}$$

Proof: A similar proof can be found in [37]. The detailed proof is omitted due to space limitation. ■

VI. Reinforcement Learning Algorithm

A. The Complete Information Case

At first, let us review the power control problem in the complete information case, when the expected utility and interference functions[6] are available. In this case, the Lagrange multiplier and the projection of gradient procedure can be applied for the class of diagonal concave games to attain the equilibrium point (see a work [12])

$$\begin{cases} p_k^{t+1} = p_k^t + \gamma_k^t \nabla_{p_k} L_k^{\delta_t}\left(p_1^t, ..., p_N^t; \lambda^t\right) \\ \lambda_{k,l}^{t+1} = \lambda_l^t - \gamma_k^t \nabla_{\lambda_l} L_k^{\delta_t}\left(p_1^t, ..., p_N^t; \lambda^t\right) \\ p_k^{t+1} = \pi_{S_{\varepsilon=0}^{N_k}}\left(p_k^{t+1}\right), \lambda_l^{t+1} = \left[\lambda_l^{t+1}\right]^+ \\ \forall k = 1, .., N, t = 1, 2, ... \end{cases}, \tag{38}$$

where $\pi_{S_{\varepsilon^t=0}^{N_k}}\{\cdot\}$ is the projection operation to the simplex $S_{\varepsilon^t=0}^{N_k}$, λ_l^t is the dual price for η_l^t, and $[\cdot]^+$ is the "take positive part" operator. If the parameters of this procedure satisfy $\gamma_k^t, \delta_t, \varepsilon^t \to 0$ and $\sum_{t=1}^{\infty} \gamma_k^t \delta^t = \infty, \sum_{t=1}^{\infty} \beta^t < \infty$. ($\beta^t$ is defined by (37)), it provides the convergence of p_k^t to the unique equilibrium strategy $(p_1^*(\upsilon), ..., p_N^*(\upsilon))$ in the Lemma 4.

B. The Incomplete Information Case

In case of incomplete information case that concerned in this work, i.e., only the current realizations of utility u_k^t and the function of interference level η_l^t explored from PUs are available, the direct application of the Lagrange multiplier approach is infeasible. The"stochastic approximation" version of (38) can be applied instead of $\nabla_{p_k} L_k^{\delta_t}\left(p_1^t, ..., p_N^t; \lambda^t\right)$ and $\nabla_{\lambda_l} L_k^{\delta_t}\left(p_1^t, ..., p_N^t; \lambda^t\right)$. This new procedure requires the estimation of the mixed power control strategies by using the implementation of the current realizations. This procedure is known as *reinforcement leaning algorithm*, which can be implemented in different ways. In order to present concrete learning procedures, the basic concept of reinforcement schemes in learning automation is presented briefly. [39].

C. Reinforcement Schemes in Learning Automation

It is shown in Section II lemma 1 that the stochastic constrained power control game in (7), (9) is asymptotically equivalent to the problem related to the determination of the equilibrium-point of the regularized function $L_k^{\delta}(p_1, ..., p_N; \lambda)$, using the realizations of the cost function and the constraints. This equivalent problem can be formulated and solved as the behavior of a variable-stochastic automation in multi-teacher environment [38]. Fig. 2 refers to the schematic block diagram for the learning automation to operate in a multi-teacher environment. We note that the normalized procedure processes as a mapping from the teachers' responses $(\zeta_k^t, 1 \leq k \leq N)$ to the learning automation input of CR users' utility (u^t) and (η^t),$1 \leq k \leq N, 1 \leq l \leq M$. The role of the environment is to establish the relation between the actions of the automation and the signals received at its input, which is described as the learning automation in Section I.

Reinforcement schemes are found successful application in the field of learning automata [32]. A reinforcement scheme is similar to the recursive estimation procedure used in adaptive control. The reinforcement scheme generates p^{t+1} from p^t based on incremental changes in the probabilities. Several algorithms for adjusting the probabilities after each sampling period (interaction with the environment) have been proposed [37]. The most commonly used one is a linear updating algorithm have been proposed by Bush and Mosteller [32]. All the reinforcement schemes described in the literature can be considered as being solutions of optimization problems. The following describe the average penalty function for a single user i in the repeated power control game.

$$J = \{\Phi_i(p^t) E_i[1 - \xi_i] - \Psi_i(p^t) E_i[\xi_i]\}. \tag{39}$$

where the functions $\Phi_i(p^t)$ and $\Psi_i(p^t)$ represent the amount of change in the probability vector under the expected reward $(\xi_i = 0)$ and expected penalty $(\xi_i = 1)$. In the case when the complete information on the expected utilities and constraints (i.e., the channel and mixed power strategy information for CR users and/or PUs) is available, then the gradient-like technique can be applied for the class of diagonal concave power control games to attain the equilibrium point. This corresponds to and similar to the traditional power control schemes in CR networks as [11] [12].

To minimize the penalty function J, the reinforcement scheme that sets the gradient of J equal to zero is derived. At the time step t, the following algorithm is obtained:

$$p_i^{t+1} = p_i^t + \frac{\gamma^t}{p_i^t}\left[\frac{\partial \Phi_i\left(p^t\right)}{\partial p_i^t}\left(1 - \xi_i\right) - \frac{\partial \Psi_i\left(p^t\right)}{\partial p_i^t}\xi_i\right]. \tag{40}$$

This derivation is one of the major results of reinforcement schemes. In general, all the existing learning schemes fall into the general framework (40) in the learning-based game-theoretical problem setting up.

In this paper, the widely used Bush-Mosteller reinforcement scheme [32] is adopted, which is a linear reinforcement scheme. And, the function $\Phi(p^t)$ is $p_i^t - \frac{1}{2}(p_i^t)^2$ and function $\Psi(p^t)$ is $(p_i^t)^2$, respectively.

D. Bush-Mosteller-Reinforcement-based Lagrange Multipliers

For the unknown expected utility and constraints of each CR user, the Bush-Mosteller reinforcement scheme and the normalization procedure presented in [37] will be applied hereafter to design a new learning algorithm for the N-user constrained repeated power control game. In fact, we assume

[6]more specifically, the interference function of every CR user k imposed on PU l, $\eta_{k,l}^t$ are known for each CR user in the networks. Thus, the sum-interference η_l^t is known in advance.

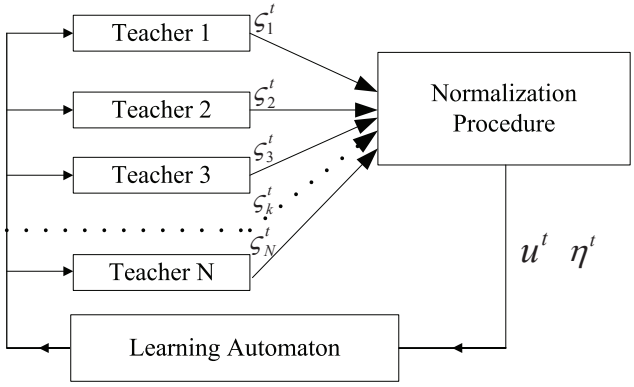


Fig. 2. Learning Automaton of Multi-teacher Environment

that after each stage, the utility as well as the constraints to each CR user are random variables. No information concerning the distribution of the utility and constraints is available. The necessary information is obtained during the course of the game. The learning control is an iterative process involving an adaption at each stage (or time).

Since our formulated power control problem in (7) and (9) is a *stochastic optimization problem* on discrete sets using only the corresponding observations of utility $\mathrm{u}_k^t(p_k^t)$ and interference $\eta_l^t(p_k^t)$ to be optimized. It is evident that we can not directly use these schemes for the optimization purposes. In fact, environment responses (in the probability sense) have to belong to the unit interval $[0,1]$. But the available observations do not obligatory satisfies this condition.

In order to solve this problem, a procedure called "normalization procedure" (as described in Fig.2), which establish the connection between the environment response ζ_k^t and the available observations (u^t) and (η^t), is described in [39] to keep the automation input within the unit segment. For the normalized procedure of the Bush-Mosteller scheme [32], we have the following lemma from [37, Lemma 2, p.268]:

Lemma 5: If, some positive monotonically decreasing $\{\tau_k^t\}$ $(0<\tau_k^t\downarrow 0)\ \forall k\in N$, and positive increasing $\{\lambda^+\}$ $(0<\lambda^+\uparrow\infty)$ sequences, the parameters a_k^t and b_k^t are given by

$$\begin{array}{l}\mathrm{a}_k^t=\tilde{a}_k^t(\sigma_{u,k}^+ +\lambda_t^+\sum_{l=1}^M\sigma_{\eta,k,l}^+ +\delta^t)^{-1}\\ \mathrm{b}_k^t=\tilde{a}_k^t+(N_k-1)(\tau_k^t)^2(1+(N_k-2)\tau_k^t)^{-1}\end{array} \tag{41}$$

where

$$\begin{array}{l}\tilde{a}_k^t=\tau_k^t(1-\tau_k^t)(2(1+(N_k-2)\tau_k^t))^{-1}\\ \tau_k^{t-1}=\varepsilon_k^t.\end{array} \tag{42}$$

then, the automaton input belongs to $(0,1]$.

Now, the "four-step" recursive algorithm for the learning-based power control game is presented.

Algorithm 1 (***Reinforcement Learning-based CR Power Control***)

Step 1: Based on The available data P_k^t=$\mathsf{P}_k(i_k)$, u_k^t, $\eta_{k,l}^t$, $p_k^t(p_k^t(i_k)\geq\varepsilon_k^t>0)$, λ_k^t build the following functions:

$$\tilde{u}_k^t=u_k^t-\delta^t p_k^t(i_k)-\sum\nolimits_{l=1}^M\lambda_l\eta_l^t \tag{43}$$

and normalize (scale) them according to the following procedure:

$$\zeta_k^t=1-(\mathrm{a}_k^t\tilde{u}_k^t+\mathrm{b}_k^t)/p_k^t(i_k) \tag{44}$$

where the sequence $\{a_k^t\}$ and $\{a_k^t\}$ are defined as follows:

$$\begin{array}{l}\mathrm{a}_k^t=\tilde{a}_k^t(\sigma_{u,k}^+ +\lambda_t^+\sum_{l=1}^M\sigma_{\eta,k,l}^+ +\delta^t)^{-1}\\ \mathrm{b}_k^t=\tilde{a}_k^t+(N_k-1)(\tau_k^t)^2(1+(N_k-2)\tau_k^t)^{-1}\\ \tilde{a}_k^t=\tau_k^t(1-\tau_k^t)(2(1+(N_k-2)\tau_k^t))^{-1}\\ \tau_k^{t-1}=\varepsilon_k^t.\end{array} \tag{45}$$

Step 2: Set the initial values of probability distribution among all the transmit power set: $p_k^1(i)$, for $0\leq i\leq N_k, \sum_{i=1}^{N_k}p_k^1(i)=1$. Update the probability distribution p_k^{t+1} and the Lagrange multipliers λ^{t+1} using the following iterative schemes:

$$p_k^{t+1}=p_k^t+\gamma_k^t[e_{N_k}(\mathsf{P}_k^t)-p_k^t+\frac{\zeta_k^t(e^{N_k}-N_k e_{N_k}(\mathsf{P}_k^t))}{N_k-1}] \tag{46}$$

where

$$e_{N_k}(\mathsf{P}_k^t)=(\underbrace{0,...,0,1}_{i_k},0,...,0)^T\in R^{N_k} \tag{47}$$

If $\mathsf{P}_k^t=\mathsf{P}_k(i_k)$ and $e^{N_k}=(1,...,1)^T\in R^{N_k}$. Here, the time-varying correction (adaptation) factors γ_k^t belong to the unit segment. Notice that the conditional mathematical expectation of the normalized environment responses $\tilde{u}_k^t e_{N_k}(\mathsf{P}_k^t)$ is equal to the gradient of the augmented Lagrange function with respect to the probability distributions, i.e.,

$$\mathbf{E}\{\tilde{u}_k^t e_{N_k}(\mathsf{P}_k^t)|H_{t-1}\}:=\frac{\partial}{\partial p_k^t}L_k^\delta(p_1,...,p_N;\lambda_n). \tag{48}$$

The Lagrange multipliers are adjusted according to the following lowing recursion; $\lambda_{k,l}^1>0,\gamma_\lambda^t\geq 0$

$$\begin{array}{l}\lambda_l^{t+1}=[\lambda_l^t-\gamma_\lambda^t\psi_l^t]_0^{\lambda_{t+1}^+}\\ \psi_l^t=\delta^t\lambda_l^t-\eta_l^t+C_l.\end{array} \tag{49}$$

In addition, note that the conditional mathematical expectation of ψ_l^t is equal to the gradient of the augmented (regularized) Lagrange function with respect to the Lagrange multipliers, that is

$$\mathbf{E}\{\psi_l^t|H_{t-1}\}:=\frac{\partial}{\partial\lambda_k^t}L_k^\delta(p_1,...,p_N;\lambda_n). \tag{50}$$

The operator

$$[x]_0^{\lambda_{t+1}^+}=\begin{cases}x, & if\, x\in[0,\lambda_{t+1}^+]\\ \lambda_{t+1}^+, & if\, x>\lambda_{t+1}^+\\ 0, & if\, x<0.\end{cases}$$

Step 3: According to

$$\Pr\{\mathsf{P}_k^{t+1}=\mathsf{P}_k(i)|H_t\}=p_k^{t+1}(i) \tag{51}$$

generate randomly new discrete random variables p_k^{t+1} for each CR user as in the learning stochastic automata implementation, and obtain new observations (realizations) u_k^{t+1} and η_l^{t+1} corresponding to the environment vector-reactions.

Step 4: Return to Step 1.

The positive sequence $\{\varepsilon_k^t\},\{\delta^t\},\{\lambda_t^+\},\{\gamma_k^t\}$, and $\{\gamma_\lambda^t\}$ will be defined next.

This adaptive Learning algorithm is constructed using the Bush-Mosteller reinforcement scheme (46) with the time-varying correction factors γ_k^t, continuous input $\tilde{u}_k^t$, and a normalization procedure which is used to ensure the probability measure. It is easy to verify that $\zeta_k^t \in (0,1)$ for any time t, and $p_k^t \in S_{\varepsilon_n}^{N_k}$.

Remark 3: In fact, the proposed algorithm only requires the value of the difference between the current interference level and the interference power constraint in the decision of each CR user's power control strategy (See the second equation in (49)). In the real network implementation, the instant interference level does not need to be obtained every time. CR users can increase their transmit power as they would prefer at the beginning. However, when the interference power constraint is violated, the PUs' would have the feedback information for this violation. Then, CR users will overhear and utilize this information and decrease the power level to meet the interference power constraint finally.

VII. Convergence Analysis and Learning Rate

A. Convergence Analysis

The following theorem acclaims the convergence of the learning-based repeated power control game.

Theorem 5 Suppose that the *Assumption* A1-A3 hold for the learning reinforcement procedure (46)- (49) and the CR power control is diagonal concave. In addition, assume that:

- There exists four nonnegative sequences $\{\varepsilon_k^t\}, \{\delta^t\}, \{\gamma_k^t\}$, and $\{\gamma_\lambda^t\}$ such that $\{\gamma_k^t\} \downarrow 0$, $\delta^t \in (0, \delta^+)$, $\delta^t \downarrow 0$ and $\varepsilon_k^t \in (0, (1/N_k))$, $\varepsilon_k^t \downarrow 0$, $limsup_n(\varepsilon_k^t/\delta^t) < \infty$;
- The updating factor γ_k^t, and γ_λ^t are selected as
$$\gamma_{\lambda,k}^t = \frac{\gamma_k^t a_k^t N_k}{N_k - 1}, 1 \le k \le N. \tag{52}$$
where a_k^t is defined by (45) and γ_λ^t satisfies $\sum_{t=1}^{\infty} \delta^t \sum_{k=1}^{N} \gamma_{\lambda,k}^t = \infty$;
- The following series converges:
$$\sum_{t=1}^{\infty} \left[\varphi^t + \beta_t^2 \left(\delta^t \sum_{k=1}^{N} \gamma_{\lambda,k}^t \right)^{-1} \right] < \infty. \tag{53}$$
where
$\varphi^t = \beta_t^2 + C_\varphi^2 \sum_{k=1}^N (\gamma_{\lambda,k}^t)^2 + 2C_\varphi \beta_t \sum_{k=1}^N \gamma_{\lambda,k}^t$
$+ 2M \sum_{k=1}^N (\gamma_{\lambda,k}^t)^2 \left[(\delta^t \lambda_t^+)^2 + (\Phi_\eta^+)^2 \right]$
$+ 2\sqrt{M} \sum_{k=1}^N (\gamma_{\lambda,k}^t)(\delta^t \lambda_t^+ + \Phi_\eta^+)$.

Then, the mixed strategy of CR users ensure the convergence of the game to the equilibrium point, i.e.,

$$\sum_{k=1}^{N} r_k \left(\left\| p_k^{t+1} - p_k^* \left(\varepsilon_k^{t+1}, \delta^{t+1} \right) \right\|^2 + \left\| \lambda_k^{t+1} - \lambda_k^* \left(\varepsilon_k^{t+1}, \delta^{t+1} \right) \right\|^2 \right) \underset{a.s.}{\to} 0.$$

Proof: See Appendix. ■

Corollary 1: For the class of the algorithm design parameters defined as follows:

$$\begin{aligned} &\gamma_k^t = \gamma_k^0 t^{-\gamma}, \varepsilon_k^t = \varepsilon_k^0 t^{-\varepsilon}, \gamma_k^0, \varepsilon_k^0, \delta^0, \lambda_0^+ > 0 \\ &\delta^t = \delta^0 t^{-\delta}, \lambda_t^+ = \lambda_0^+ + t^\gamma \ln t. \end{aligned} \tag{54}$$

The conditions of Theorem 5 will be verified is

$$\gamma + \varepsilon + \delta + \lambda \le 1 (\gamma > 0, \varepsilon \ge \delta > 0, \lambda \ge 0) \tag{55}$$

and the convergence is ensured if

$$2\gamma > 1. \tag{56}$$

Proof: The proof of the convergence follows directly by substituting (54)-(56) into the conditions of the previous theorem and in view of the fact that

$$\sum_{n=1}^{\infty} t^{-\alpha} \begin{cases} = \infty, if \alpha \le 1 \\ < \infty, if \alpha > 1. \end{cases}$$

■

B. Learning Rate

As known earlier, not only the convergence of the power control game is important but the speed is also essential. For the specific class of the design parameters (54), the next theorem states the convergence rate of the learning game algorithm described above.

Theorem 6: Under the condition of the previous theorems and for the class of design parameters (54), there exists v such that,

$$W_*^t = \sum_{k=1}^{N} r_k \left(\left\| p_k^t - p^*(v) \right\|^2 + \left\| \lambda_k^t - \lambda^*(v) \right\|^2 \right) = o(n^{-v}),$$

where the order v of the adaptation rate satisfies the following constraint:

$$\begin{aligned} &v < v^*(\gamma, \delta, \varepsilon, \lambda) \le v^{**} = \tfrac{1}{3} \\ &v^*(\gamma, \delta, \varepsilon, \lambda) = \min\{2\gamma - 1; \gamma + \delta; \varepsilon - \delta + \gamma; 2\delta\} \end{aligned}$$

and the maximum adaptation rate $v^{**}=v^*(\gamma^*, \delta^*, \varepsilon^*, \lambda^*)$ is reached for

$$\varepsilon = \varepsilon^* = \delta = \delta^* = \frac{1}{6}, \gamma = \gamma^* = \frac{2}{3}, \lambda = \lambda^* = 0.$$

Proof: The expression of $v^*(\gamma, \delta, \varepsilon, \lambda)$ follows from [37] Lemma A.3-2, App. A. The optimal design parameters are the solution of the following constrained optimization problem $v^*(\gamma, \delta, \varepsilon, \lambda) \to max$ over tall the parameters $\varepsilon \ge \delta$ satisfying (54) and (55). The solution of this problem is achieved when all minimal terms within the operator is equal, that is

$$\begin{aligned} v^{**} &= \max_{(\gamma,\delta,\varepsilon,\lambda)} \min\{2\gamma - 1; \gamma + \delta; \varepsilon - \delta + \gamma; 2\delta\} \\ &\le (by\ \gamma \le 1 - \delta - \varepsilon - \lambda) \\ &\max_{1-\delta-\varepsilon-\lambda} \min\{1 - 2\varepsilon - 2\lambda; 1 - \varepsilon - \lambda; 1 - 2\delta - \lambda; 2\delta\} \\ &= \max_{1-\delta-\varepsilon-\lambda} \min\{1 - 2\varepsilon - 2\lambda; 2\delta\} \\ &= (\text{by} \varepsilon = 1/2 - 2\delta - \lambda) \\ &= \max_{\delta \le \varepsilon} 2\delta = 1/3, \text{under } \delta = 1/6 \text{ and } \lambda = 0. \end{aligned}$$

■

VIII. Simulation Results

In this section, simulation results are presented for the repeated power control game in CR networks. We consider there are multiple CR Tx-Rx pairs and one PU Tx-Rx pair sharing a spectrum band with the bandwidth of $1MHZ$. The configuration of CR users and the PU is shown in Fig.3 with their current locations. The simulation parameters are: the distance of the CR Tx-Rx is $250m$; The action sets of transmit power is $\mathsf{P}_k = \{0.02, 0.2, 1\}$*Watt*, the noise of the

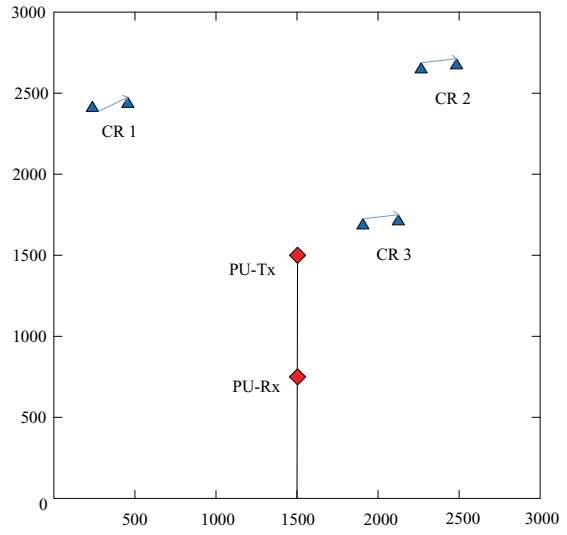


Fig. 3. User location in the network.

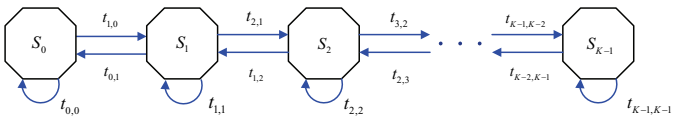


Fig. 4. K-state Markov transitions modeling a Rayleigh fading Channel

measurement is according to a zero-mean Gaussian noise, where its power is $\sigma_k^2 = -118.45dBm$. The channel gain is a log-normal shadowing path loss model, $g_k = (1/d_k)^4$. For the fading, we use the k-state Markov channel (FSMC) to model the Rayleigh fading channel [22], where $k = 8$ and the maximum Doppler frequency is $f_m = 10HZ$. The Channel State transition diagram is shown in Fig. 4, and the state transition probability is listed in Table 1.

This network set-up guarantees the *interference non-dominance* condition, where the interference from other CR users and the background noise is less than the transmit power of each CR user. We consider the PU is always active, where $b_l = 1$. The interference power constraint can change and be set to different values. The default interference power constraint in the PU-Rx is $-115.37dBm$, which is a light interference constraint. As a typical illustration, a random initial probability distribution of the mixed strategies are generated for the CR user 1, the CR user 2 and the CR user 3: $p_1^1 = \{0.59, 0.28, 0.13\}$, $p_2^1 = \{0.74, 0.21, 0.05\}$, and $p_2^1 = \{0.84, 0.07, 0.09\}$, respectively. For the utility function in (3), we set $K = 1$, where the transmission rate is the information-theoretical capacity. The time duration of the learning process for the game is considered in the CR network with the period of $t = 1000$ steps. The parameters of the learning procedure were as follows:

$$\varepsilon = \varepsilon^* = \delta = \delta^* = 16, \gamma = \gamma^* = 2/3, \varepsilon^0 = 0.25$$
$$\lambda = \lambda^* = 0, \lambda_0^+ = \delta^0 = 0.1, \gamma^0 = 0.15,$$

The proposed learning-based power control algorithm is implemented based on all initial values provided above. Then, all the sequences will be decreased according to the learning algorithm until they all converge.

TABLE I
ANALYTICAL VALUES OF THE TRANSITION PROBABILITIES FOR EIGHT-STATE MARKOV CHANNEL WITH THE MAXIMUM DOPPLER FREQUECNCY $f_m = 10HZ$

	$t_{K,K-1}$	$t_{K,K}$	$t_{K,K+1}$
$k=0$	-	0.999359	0.000641
$k=1$	0.000641	0.998552	0.000807
$k=2$	0.000807	0.998334	0.000859
$k=3$	0.000859	0.998306	0.000835
$k=4$	0.000835	0.998420	0.000745
$k=5$	0.000745	0.998665	0.000590
$k=6$	0.000590	0.999048	0.000361
$k=7$	0.000361	0.999639	-

Fig. 5, Fig. 6 and Fig. 7 show the convergence process of the probability distribution of the mixed strategy of each CR users. The first observation from our simulation results is that, whenever we generate a random initial probability distributions of the mixed strategies, the equilibrium state of the transmit power level calculated by the mixed strategy of each user is independent with these initial values. This result confirms the theory of convergence in Section VII. Secondly, when comparing the equilibrium state of the mixed strategy of CR users 1, 2, and 3, we can observe that as the CR Tx-Rx becomes close to the other CR Tx-Rxs, the transmit power level (a dot product of mixed strategy probability and the transmit power set) will decrease in the equilibrium state. However, we do not observe much more interference penalty on CR users' transmit power, when CR users are closer to the PU-Rx. An illustration on this point is the CR user 3. Even though the CR user 3 can hear the PU-Rx better than the CR user 2 and CR user 1, i.e., the interference is larger, the transmit power of the CR user 3 is just a little bit higher than the CR user 2. The inside rationale is as follows: as we know, each CR user has blind CSI information from its Tx to the PU-Rx, and CR users can only explore the sum-interference measured by PU-Rx from its feedback control link, the interference penalty price λ_k^t for each CR user in (49) is the same for each CR user. Correspondingly, the expected-version of subgradient updating process of the interference penalty in (50) that is plugged into (38) is the same for each user. However, in (43), we can observe the utility function of each CR user is different due to the inter-CR-user interference level, the CR user 1 suffers the least interference, the CR user 3 suffers less, and the CR user 2 suffers the most heavy interference. This affects the expected value of subgradient of transmit power in (48), and finally the transmit power updating process in (38). The result can be clearly viewed from (38): the higher the expected subgradient, the higher the equilibrium transmit power. This result is similar to the result in the classic game-theoretical approach without the "pricing" scheme[7]. The simulation results shown in Fig. 5, Fig. 6 and Fig. 7 conform the theoretical analysis.

Fig. 8 shows the learning process of averaged transmission rate for the CR users under the light interference constraints condition of the PU. The results are compared with the theoretical results of the power control game with complete information case: each CR user knows all the CSI and transmit

[7]There is no incentive of the interference power penalty for each individual CR user from PUs based on the interference level of each individual CR user.

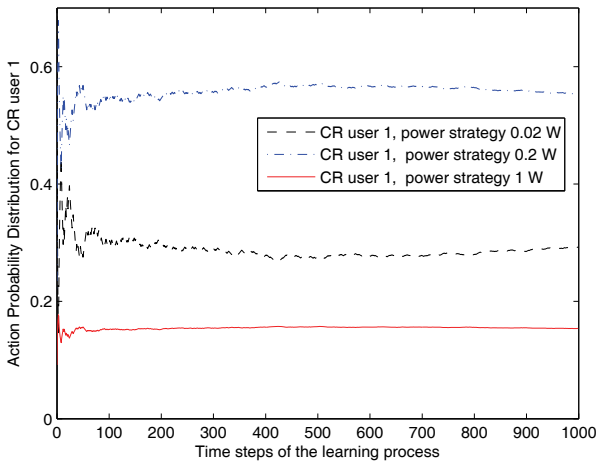


Fig. 5. learning process of the mixed strategy.

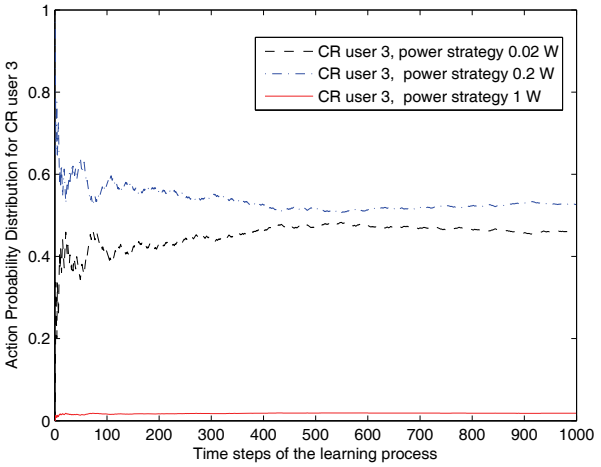


Fig. 7. learning process of the mixed strategy.

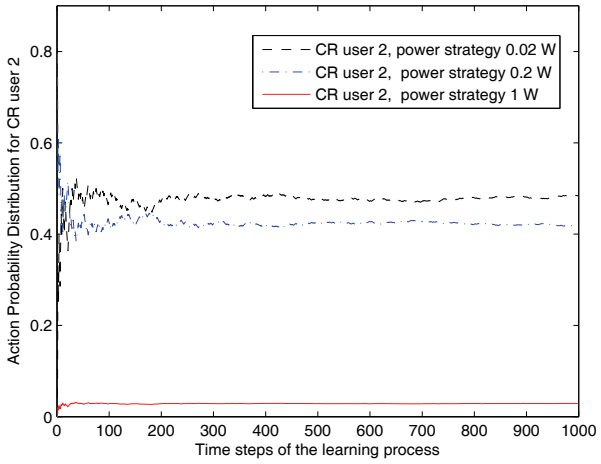


Fig. 6. learning process of the mixed strategy.

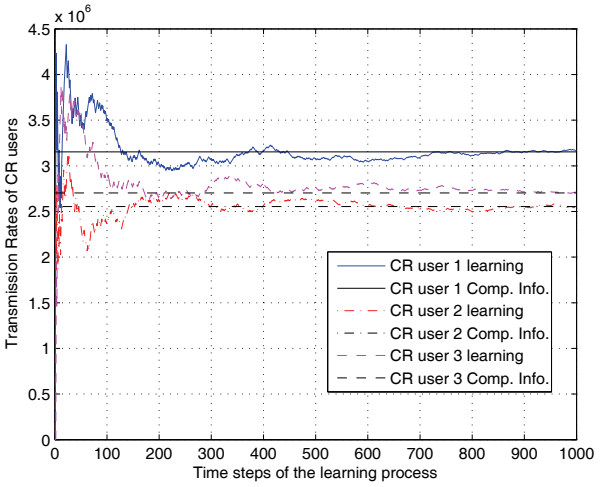


Fig. 8. learning process of the transmission rates.

power strategy in the network, and then the sum-interference of all CR users η_l^t can be calculated by each CR user in the power control process in equ.(38). This scenario is equivalent to the classic power control game case in CR networks without the pricing schemes from PUs as shown in [11], [12], [17]. It can be observed that the averaged transmission rate of each CR user in the learning process will converge and approach to the equilibrium point in the complete information game case, and this simulation results confirm the conclusion of the Lemma 1. In addition, the time interval (or iterations) required to achieve the equilibriums is longer than the classic one. This is due to the expenses of no prior knowledge of CSI and transmitting power strategy of other users in the network. However, in this scenario, CR users only need to explore several bits information of the interference-level from the feedback link of the PU. Then, they use the transmission rates, which are tried in each step with low implementation complexity. This method is suitable and adaptable for highly dynamic changing radio-access environments.

Fig.9 shows the convergence process of the interference power of each CR user in the light interference constraint condition. In this case, the CR users transmit in higher power in the equilibrium condition than the CR users in the heavy interference constraint condition (-117.47dBm) in Fig. 10. An interesting observation is that, when the interference constraint is low (as shown in Fig. 10), the convergence process of interference power of CR users changes much more rapidly in the initial steps than the case when the interference constraint is high (as shown in Fig.9). Especially, this phenomenon is clearly seen for the dominant interference CR users, e.g., the CR user 3 in Fig. 9 and Fig. 10. Moreover, at the initial stages of the repeated power control game, the interference power constraint of the PU is violated, and it is gradually being met during the learning process. This is consistent with the theoretical predictions in (29) (32) (33): the learning algorithm will enforce the transmit power strategy largely to reduce the interference caused by each CR user until it conforms to the interference power constraint.

The last simulation aims to verify the learning rate of the proposed algorithm in large scale networks. To set up a typical simulation scenario, the *interference non-dominance* condition should be guaranteed to provide a unique equilibrium point

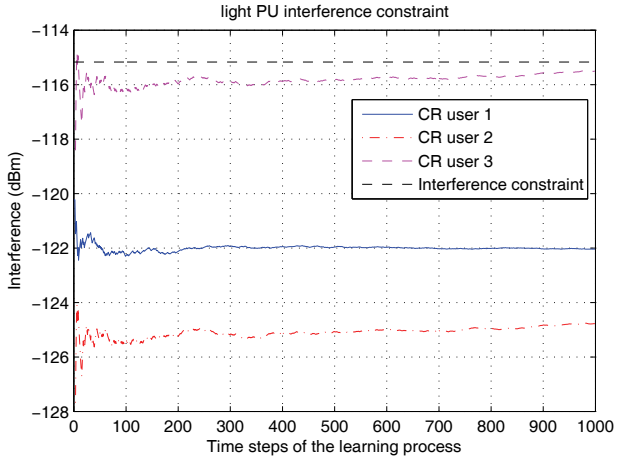


Fig. 9. learning process of the interference level.

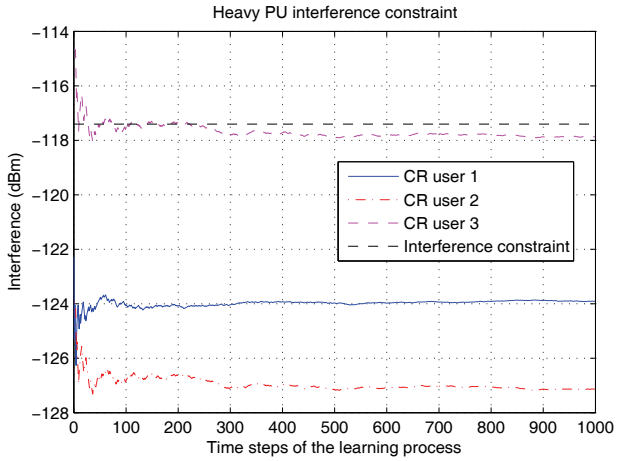


Fig. 10. learning process of the interference level.

for each CR user; However, the network size and game strategy space should be keep unchanged to guarantee a fair comparison, while only the number of CR users is varying. Therefore, the proposed network scenario is determined as follows: consider the same network size as shown in Fig. 3. The PU and its interference power constraint C_l is unchanged. For the CR network, CR transceiver pairs' topology location is kept unchanged. Using the 3-pair CR network as a model guarantees the interference non-dominance condition. By keeping the model CR network size scaling down, we can introduce more model CR networks that results in more CR transceiver pairs into the network. This forms the obtained CR network in the simulation. Thus, the obtained CR network scales in an order of $3N$. Fig. 11 show simulation results of the number of convergence steps *vs* the number of CR users. The convergence condition is declared, where the deviation of probabilities of power control strategy is within 0.005. Since theorem 6 presents an upper bound for all the network topologies and our simulation scenario belongs to one of these specific scenarios, this simulation provides an intuitive view to let researchers perceive how the learning rate performs in large networks. As we can see from the figures, when the convergence step decreases, the number of users scales

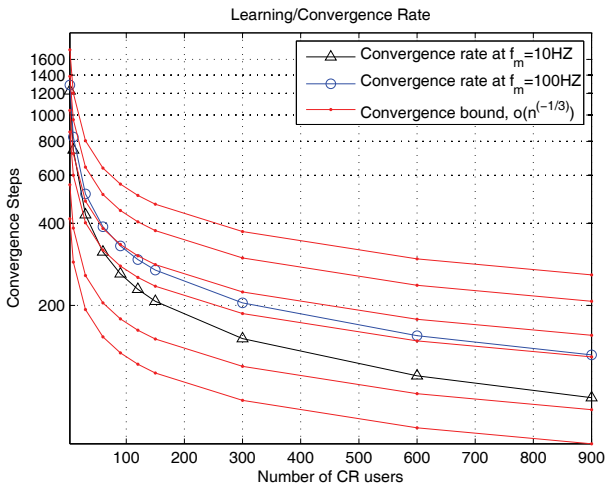


Fig. 11. learning process of the interference level.

up. Moreover, the learning rate is bounded by $o(n^{-1/3})$ in large network deployments. We also compare the results of the convergence rate when the maximum Doppler frequency is $f_m = 10HZ$ and $f_m = 100HZ$. It is observed that the convergence rate in the condition of $f_m = 10HZ$ is faster than $f_m = 100HZ$, which confirms our basic intuition.

IX. Conclusion

In this paper, a robust power control algorithm with low implementation complexity is designed for *competitive* and *autonomous* CR networks for the first time. The general feature of asymptotically stationary in the average sense property is modeled in wireless environments. This formulated problem is an incomplete-information repeated game with learning automata. The interesting properties of the asymptotically equivalence of the classic complete-information case are studied with the proposed incomplete-information power control game as well.

Unlike traditional approaches, the game-theoretical problem is formulated as a strictly diagonal concave game, which proves the uniqueness of the power-control game in a nice mathematical structure and interpretation. To solve this mixed-strategy repeated power-control game, Bush-Mosteller reinforcement is proposed. This reinforcement procedure that uses the Lagrange multipliers and an appropriate regularization is the optimal response within the given constraints for CR users. Using the stochastic optimization techniques, the *convergence* of the game to the Nash equilibrium is analyzed with the *rate of the learning* to be $o(n^{-1/3})$. This proposed algorithm is an effective solution in real-world CR networks where cooperation among CR users and/or PUs is unavailable.

X. Appendix

A. Proof of Theorem 5

Proof: The proof of this theorem is based on the technique of Lyapunov approach and Martingale's theory. As we know, the fulfilling of the assumption **A3** implies that the behavior domain of the game is not empty, that is

$$\bigcap_{k=1,..,N}\bigcap_{l=1,..,M}\left\{p_1^t,...,p_N^t : Q_l\left(p_1^t,...,p_N^t\right)\geq C_l\right\}\bigcap_{k=1,..,N} S_{\varepsilon=0}^{N_k}\neq 0.$$

Under the assumption **A1** and **A2** of this theorem, there exists an time lost t_0 such that for any $t \geq t_0$ the proposed power control game admits an equilibrium point defined by $p_k^*(\varepsilon_k^t, \delta^t)$ and $\lambda_k^*(\varepsilon_k^t, \delta^t)$. Consider the following Lyapunov function:

$$W^t = \sum_{k=1}^{N} r_k \left(\left\| p_k^t - p_k^*(\varepsilon_k^t, \delta^t) \right\|^2 + \left\| \lambda_k^t - \lambda_k^*(\varepsilon_k^t, \delta^t) \right\|^2 \right).$$

Considering the reinforcement algorithm, and by adding and subtracting $p_k^*(\varepsilon_k^t, \delta^t)$ and $\lambda_k^*(\varepsilon_k^t, \delta^t)$ to the first and the second term of the right side of the previous equality, we have

$$W^{t+1} \leq \sum_{k=1}^{N} r_k \left(\begin{array}{l} \left\| (p_k^t - p_k^*(\varepsilon_k^t, \delta^t) + \gamma_k^t A_k^t \right. \\ \left. - \left[p_k^t(\varepsilon_k^{t+1}, \delta^{t+1}) - p_k^*(\varepsilon_k^{t+1}, \delta^{t+1}) \right] \right\|^2 \\ + \left\| \lambda_k^t - \lambda_k^*(\varepsilon_k^t, \delta^t) - \gamma_k^t \psi_k^t \right. \\ \left. - \left[\lambda_k^t(\varepsilon_k^{t+1}, \delta^{t+1}) - \lambda_k^*(\varepsilon_k^{t+1}, \delta^{t+1}) \right] \right\|^2 \end{array} \right)$$

where $\psi^t := (\psi_1^t, ..., \psi_M^t)^T$ and

$$A_k^t = e_{N_k}(\mathsf{P}_k^t) - p_k^t + \frac{\zeta_k^t (e^{N_k} - N_k e_{N_k}(\mathsf{P}_k^t))}{N_k - 1}. \quad (57)$$

Since used the following property of the upper projection operation: $\left| \lambda - [y]_0^{\lambda^+} \right| \leq |\lambda - y|$, which is valid for any $y \in R$, and $\lambda \in [0, \lambda^+]$. Observe that the following bounds hold $\|A_k^t\| \leq C_A = const < \infty$ and $\|\varepsilon_k^t\| \leq \sqrt{M}\left(\delta^t \lambda_t^+ +\right) \sigma_\eta^+, \sigma_\eta^+ = \max_l \sigma_{\eta,l}^+$, which leads to

$$\begin{array}{l} W^{t+1}\left(p_1^{t+1}, ..., p_N^{t+1}; \lambda^{t+1}\right) \leq W^t(p_1^t, ..., p_N^t; \lambda^t) + \\ 2\sqrt{2N}\sqrt{W^t(p_1^t, ..., p_N^t; \lambda^t)}\beta^t + \varphi^t + s^t \end{array} \quad (58)$$

where

$$s^t = 2\sum_{k=1}^{N} \left(\begin{array}{l} \gamma_k^t (p_k^t - p_k^*(\varepsilon_k^t, \delta^t))^T A_k^t \\ -2\gamma_k^t \left(\lambda_{k,\gamma}^t - \lambda_k^*\left(\varepsilon_k^{t+1}, \delta^{t+1}\right) \right)^T \psi_k^t \end{array} \right). \quad (59)$$

Based on (48) and (50), we can the following results:

$$\begin{array}{l} (p_k^t - p_k^*(\varepsilon_k^t, \delta^t))^T E\{A_k^t | H_{t-1}\} \\ = -(p_k^t - p_k^*(\varepsilon_k^t, \delta^t))^T E\left\{ \zeta_k^t e_{N_k}(\mathsf{P}_k^t) | H_{t-1} \right\} \frac{N_k}{N_k - 1} \\ = \frac{a_k^t N_k}{N_k - 1} (p_k^t - p_k^*(\varepsilon_k^t, \delta^t))^T \frac{\partial}{\partial p_k^t} L_k^{\delta^t}(p_1^t, ..., p_N^t; \lambda_k^t) \end{array} \quad (60)$$

and

$$\begin{array}{l} (\lambda_k^t - \lambda_k^*(\varepsilon^t, \delta^t))^T E\{\psi_k^t | H_{t-1}\} \\ = (\lambda_k^t - \lambda_k^*(\varepsilon^t, \delta^t))^T \frac{\partial}{\partial \lambda_k^t} L_k^{\delta^t}(p_1^t, ..., p_N^t; \lambda_k^t). \end{array} \quad (61)$$

In the view of (59)-(61) and *Assumption* **A2** of this theorem, from the *diagonal concavity property* in **Lemma 3**, it follows:

$$\begin{array}{l} E\{s^t | H_{t-1}\} \\ \overset{a.s.}{=} 2\sum_{k=1}^{N} r_k \left(\begin{array}{l} \frac{a_k^t N_k}{N_k - 1} (p_k^t - p_k^*(\varepsilon_k^t, \delta^t))^T \frac{\partial}{\partial p_k^t} L_k^{\delta^t} \\ (p_1^t, ..., p_N^t; \lambda_k^t) - 2\gamma_\lambda^t \left(\lambda_k^t - \lambda_k^*\left(\varepsilon_k^{t+1}, \delta^{t+1}\right) \right)^T \\ \frac{\partial}{\partial \lambda_k^t} L_k^{\delta^t}(p_1^t, ..., p_N^t; \lambda_k^t) \end{array} \right) \\ \leq -\gamma_\lambda^t \delta^t \sum_{k=1}^{N} r_k \left(\left\| p_k^t - p_k^*(\varepsilon_k^t, \delta^t) \right\|^2 + \left\| \lambda_k^t - \lambda_k^*(\varepsilon_k^t, \delta^t) \right\|^2 \right) \\ = -\gamma_\lambda^t \delta^t W^t. \end{array}$$

Substituting this expression into (58), we obtain

$$W^{t+1} \leq \left(1 - \gamma_\lambda^t \delta^t\right) W^t + 2\sqrt{2N}\sqrt{W^t}\beta^t + \varphi^t. \quad (62)$$

In the above result, we use the inequality $z^r \leq (1-r)z_0^r + (r z_0^{r-1})t$, which is valid for any z and $z_0 > 0$ and $r \in [0, 1]$. let $r = 1/2$,it implies

$$2\sqrt{2N}\sqrt{W^t}\beta^t \leq 2\sqrt{2N}\beta^t\sqrt{z_0} + \sqrt{2N}\beta^t (z_0)^{-1/2} W^t.$$

By setting $\sqrt{2N}\beta^t (z_0)^{-1/2} = \gamma_\lambda^t \delta^t / 2$. it follows:

$$2\sqrt{2N}\sqrt{W^t}\beta^t \leq 4N\left(\beta^t\right)^2 \left(\gamma_\lambda^t \delta^t\right)^{-1} + \gamma_\lambda^t \delta^t W^t / 2.$$

Substituting this estimation above into (62), we have the following upper bound estimation:

$$W^{t+1} \leq \left(1 - \gamma_\lambda^t \delta^t / 2\right) W^t + 4N\left(\beta^t\right)^2 \left(\gamma_\lambda^t \delta^t\right)^{-1} + \varphi^t. \quad (63)$$

Therefore, $\{W^t | H_t\}$ is a quasimartingale. By *Assumption* **A3** and the Robbins-Siegmund theorem (Robbins and Siegmund 1971) [40] in probability theory, the statement of this theorem follows directly from (63). ■

REFERENCES

[1] Akyildiz, I. F., Lee, W.Y., Vuran, M.C. and Mohanty, S., "NeXt Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey," *Computer Networks Journal, (Elsevier)*, vol. 50, pp.2127-2159, September 2006.

[2] B. Wang, Y. Wu and K.J. R. Liu, "Game theory for cognitive radio networks: An overview," *Computer Networks Journal, (Elsevier)*, vol. 54, iss.14, pp. 2537-2561, Oct. 2010.

[3] M. Maskery, V. Krishnamurthy, and Q. Zhao, "Decentralized Dynamic Spectrum Access for Cognitive Radios: Cooperative Design of a Non-cooperative Game, "in *IEEE Trans. Commun.*, vol. 57, No. 2, pp. 459-469, February, 2009.

[4] J., Shu and P., Varaiya, "Pricing Network Services, "In Proceedings of *IEEE Infocom 2003*, San Francisco, CA, pp. 1221-1230, April, 2003.

[5] L. Qiu, Y. Yang, Y. Zhang, and S. Shenker, "On Selfish Routing in Internet-Like Environments," In Proceedings of *ACM SIGCOMM*, Karlsruhe, Germany, pp. 12-25, August 2003.

[6] A. Al Daoud, T. Alpcan, S. Agarwal, and M. Alanyali, "A Stackelberg game for pricing uplink power in wide-band cognitive radio networks," in *Proc. 47th IEEE Conf. Decision Control (CDC 2008)*, pp. 1422-1427, Dec. 2008.

[7] A. Al-Daoud, M. Alanyali, and D. Starobinski, "Secondary pricing of spectrum in cellular CDMA networks," in *Proc. IEEE DySPAN*, pp. 535-542, Nov. 2007.

[8] Y. Xing, R. Chandramouli, and C. M. Cordeiro, "Price dynamics in a competitive agile secondary spectrum access market," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 613-621, Apr. 2007.

[9] D. Niyato and E. Hossain, "Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of nash equilibrium, and collusion," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 192-202, Jan. 2008.

[10] H. Yu, L.Gao, Z. Li, X. Wang and E. Hossain, "Pricing for uplink power control in cognitive radio networks," *IEEE Trans. Veh. Technol.*, Vol. 59, Iss. 4, pp. 1769 - 1778, 2010.

[11] P. Zhou, W. Yuan, W. Liu and W. Cheng, "Joint power and rate control in cognitive radio networks: a game-theoretical approach," *IEEE International Conference on Communications (ICC 2008)*, pp. 3296-3301, 19-23 May 2008.

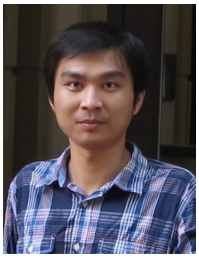
[12] Y. Wu and D. H. K. Tsang, "Distributed multi-channel power allocation algorithm for spectrum sharing cognitive radio networks with QoS guarantee," *The 28th IEEE Conference on Computer Communications (INFOCOM)*, 2009.

[13] R. Zhang, "Optimal power control over fading cognitive radio channel by exploiting primary user CSI," *IEEE Global Telecommunications Conference (GLOBECOM 2008)*, pp. 1-5, Nov. 30-Dec. 4, 2008.

[14] X. Kang, R. Zhang, Y. C. Liang and H.K. Garg, "On Outage Capacity of Secondary Users in Fading Cognitive Radio Networks with Primary User's Outage Constraint," *IEEE Global Telecommunications Conference (GLOBECOM 2009)*, pp. 1-5, Nov. 30-Dec. 4, 2009.

[15] S. Huang, X. Liu, and Z. Ding,"Decentralized Cognitive Radio Control based on Inference from Primary Link Control Information," to appear in *IEEE J. Sel. Areas Commun.*, Special Issue on Advances in Cognitive Radio Networking and Communications.

[16] F. E. Lapiccirella, S. Huang, X. Liu, and Z. Ding, "Feedback-based access and power control for distributed multiuser cognitive networks," *Information Theory and Application (ITA) workshop*, UCSD, 2009.
[17] P. Zhou, Y. Chang and J. Copeland, "Asynchronous Power Control Game with Channel Outage Constraints for Cognitive Radio Networks," to appeaer in *IEEE International Conference on Communications (IEEE ICC 2011)*, Tokyo, Japan, 2011.
[18] J. Huang, R. Berry, and M. L. Honig, "Auction-based spectrum sharing," *ACM Mobile Netw. Appl. J.*, vol. 11, no. 3, pp. 405-418, Jun. 2006.
[19] D. Fudenberg, D. K. Levine, The Theory of Learning in Games. Cambridge, MA: MIT Press, 1998.
[20] A.J., Goldsmith and P.P., Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inf. Theory,* vol. 42, iss.3, pp. 868-886, 1996.
[21] A.J., Goldsmith, *Wireless Communication*, Cambridge University Press, 2005.
[22] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *IEEE Trans. Veh. Technol.*,Vol. 44, iss. 1, pp. 163-171.
[23] A.N., Shiryaev, *Probability*, second edition, Springer Press, 2005.
[24] P., Hall and C.C. Heyde, *Martingale Limmit Theory and Its Applications*, Academic Press, 1980.
[25] C. Long, Q. Zhang, B. Li, H. Yang, and X. Guan, "Non-Cooperative Power Control for Wireless Ad Hoc Networks with Repeated Games", in *IEEE J. Sel. Areas Commun., special issue on Non-Cooperative Behavior in Networking*, Aug. 2007.
[26] Y. Xing and R. Chandramouli, "Stochastic Learning Solution for Distributed Discrete Power Control Game in Wireless Data Networks", *IEEE/ACM Trans. Netw.*, vol: 16, iss.4, 2008, pp: 932-944.
[27] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control," *IEEE J. Sel. Areas Commun.*, Vol.23, pp. 104-116, Jan., 2005.
[28] P. Zhou, Y. Chang and J. Copeland, "Learning through reinforcement for repeated power control game in cognitive radio networks," *IEEE Global Telecommunications Conference (IEEE GLOBECOM 2010)*, pp. 1157-1162, Dec. 5-Dec. 10, Florida, Dec. 2010.
[29] J. Nash, "Equilibrium points in n-person games," Proc. Nat. Acad. USA, vol. 36, pp. 48C49, 1950.
[30] A. S. Poznyak, K. Najim, and E. Gomez, *Self-Learning Control for Finite Markov Chains*. New York: Marcel Dekker, 2000
[31] K. J. Arrow, L. Hurwicz, and H. Uzawa, "Constraint qualifications in maximization problems," *Nav. Res. Logist. Q.*, vol. 8, pp. 175-191, 1961.
[32] Bush, R. and Mosteller, F., "*Stochastic Models of Learning*, " John Wiliey & Son, New York, 1995.
[33] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave N-persons games," *Econometrica*, vol. 33, pp. 520-534, 1965.
[34] J. P. Aubin, *Mathematical Methods of Game and Economic Theory*, Amsterdam, The Netherlands: North Holland, 1979.
[35] N. N. Vorobev, *Foundations of Game Theory: Noncooperative Games*, Basel: Birkháuser, 1994.
[36] Whittle P., *Optimization under Constriants, Wiley-Interscience*, New York, 1971.
[37] A. S. Poznyak and K. Najim, *Learning Automata and Stochastic Optimization*, New York: Springer-Verlag, 1997.
[38] Baba N., *New Topics in Learning Automata Theory and Applications*, Springer-Verlag, Berlin, 1984.
[39] K. Najim and A. S. Poznyak, *Learning Automata: Theory and Applications. New York: Pergamon*, Springer-Verlag, 1994.
[40] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermar- tingales and some applications," *In Optimizing Methods in Statitsics. (JS Rustagi, ed.)*, pp. 233-257, 1971.



Pan Zhou obtained his Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology in 2011 as a member of the Communications Systems Center (CSC). He received his B.S. degree in the *Advanced Class* of Huazhong University of Science and Technology (HUST), and a M.S. degree in Electrical Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2006 and 2008, respectively. He held *honorary degree* and *merit research award* of HUST in his master study. He also received a M.S. degree in Electrical and Computer Engineering from Georgia Institute of Technology in 2010. He was once a student internship at Mobile Communications and Networking Research, NEC Laboratories America, Inc., in summer 2010. His research interest is majorally in wireless communications and networking problems, such as, cognitive radio, wireless scheduling, and theoretical research of network science. The style of simple and mathematics related researches with impressive results are his favorite.

Yusun Chang is an Assistant Professor at Southern Polytechnic State University in Atlanta, and also an Adjunct Assistant Professor leading the Vehicular Networking Group in the Communications Systems Center (CSC) at the Georgia Institute of Technology. He received B.S. and M.S. in Avionics from the Korea Aerospace University in 1993 and 1995, respectively. He instructed air force officers in the Air Force Artillery School for three years in Korea. He received M.S. in Electrical Engineering at Columbia University in 2002 and Ph.D. at the Georgia Institute of Technology in 2007. His research interests include the modeling and the optimization of wireless mobile networks and the implementation of wireless communication systems. He proposes a new paradigm, Sustainable Mobile Networking that improves sustainable environments by using intelligent mobile communication technologies to reduce traffic congestion and air pollution and enhance transportation mobility and safety. He is also the author of the book, "Robust Mobile Networking."

John A. Copeland Prof. John A. Copeland holds the John H. Weitnauer, Jr., Chair in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, and is a Georgia Research Alliance Eminent Scholar. He is the Director of the Communications Systems Center (CSC). This center is doing research on digital communication networks, including wireless sensor networks and WiFi and WiMAX networks, with emphasis on providing security and Quality of Service. Prior to joining Georgia Tech in 1993, Dr. Copeland was Vice President, Technology at Hayes Microcomputer Products (1985-1993), and Vice President, Engineering Technology at Sangamo Weston, Inc. (1982-1985) and served at Bell Labs (1965-1982).

He began his career at Bell Labs conducting research on semi-conductor microwave and millimeter-wave devices. Later, he supervised a group that developed magnetic bubble computer memories. In 1974, he led a team that designed CMOS integrated circuits, including Bell Labs' first microprocessor, the BELLMAC-8. His last contributions at Bell Labs were in the area of lightwave communications and optical logic. At Sangamo Weston he was responsible for R&D groups at ten divisions. At Hayes was responsible for the development of modems with data compression and error control, and for Hayes' representation on CCITT and ANSI standards committees. In 2000 he invented the StealthWatch system for network security monitoring, and founded LANcope, Inc. which today has deployed StealthWatch on over 100 corporate, government, and defense networks.

Dr. Copeland received B.S., M.S. and Ph.D. degrees in physics from the Georgia Institute of Technology. He has been awarded 49 patents and has published over 50 technical articles. In 1970 he was awarded IEEE's Morris N. Liebmann Award for his work on gallium arsenide microwave devices. He is a Fellow of the IEEE and has served that organization as the Editor of the IEEE Transactions on Electron Devices. He served on the Board of Trustees for the Georgia Tech Research Corporation (1983-1993), and Director of the Georgia Center for Advanced Telecommunications Technology (1993-1996).