# Stochastic Game Formulation for Cognitive Radio Networks

Fangwen Fu and Mihaela van der Schaar
Electrical Engineering Department, University of California, Los Angeles (UCLA),
Los Angeles, California, USA
{fwfu, mihaela}@ee.ucla.edu

***Abstract*— In this paper, we model the various wireless users in a cognitive radio network as a collection of selfish, autonomous agents that strategically interact in order to compete for the dynamically available spectrum opportunities. We propose a stochastic game framework to model how the competition among users for spectrum opportunities evolves over time. At each stage of the dynamic resource allocation, a spectrum moderator auctions the available resources and the users strategically bid for the required resources. Based on the observed resource allocation and corresponding rewards from previous allocations, we propose a best response learning algorithm that can be deployed by wireless users to improve their bidding policy at each stage. The simulation results show that by deploying the proposed best response learning algorithm, the wireless users can significantly improve their own performance in terms of both the packet loss rate and the incurred cost for the used resources.**

***Keywords- Multi-user Resource Management; Interactive Learning, Cognitive Radio Networks, Stochastic game***

## I. INTRODUCTION

One vision for emerging cognitive radio networks assumes that certain portions of the spectrum will be opened up for secondary users (SUs), which can autonomously and opportunistically share the spectrum once primary users (PUs) are not active [3][4][5]. Importantly, in cognitive radio networks, heterogeneous wireless users with different utilities, delay tolerances, traffic characteristics, interference avoidance, knowledge and ability to adapt will need to coexist in the same band. Current solutions do not provide fair or efficient resource management for delay-sensitive applications such as multimedia streaming in the cognitive radio networks.

Thus, to enable the proliferation of multimedia applications over cognitive radio networks, wireless solutions for dynamic spectrum access and resource management will need to consider the system dynamics, as well as the heterogeneity of wireless users. Moreover, SUs will need to possess learning abilities to be able to strategically influence and adapt to the dynamic spectrum division. Using their knowledge, SU can proactively harvest resources based on their dynamic resource requirements as well as optimally adapt their cross-layer transmission strategies to the environment dynamics and time-varying gathered resources. Such dynamic and competitive solutions for spectrum access and protocol design lead to more efficient and fair wireless networks than current solutions, which require SUs to blindly follow predetermined or static protocol rules [1].

In our considered cognitive radio networks, the SUs are modeled as rational and strategic. We model the spectrum management as a stochastic game [6] in which the SUs simultaneously and repeatedly compete for the available network resources. The competition for the dynamic resources is assisted by a central coordinator (similar to that in existing wireless LAN standards such as 802.11e HCF [7]). We refer to this coordinator as the central spectrum moderator (CSM). The role of the CSM is to allocate resources to the SUs based on pre-determined utility maximization rule. In this paper, to explicitly consider the strategic behavior of the autonomous SUs and the informationally-decentralized nature of the competition for wireless resources, we assume that the CSM deploys an auction mechanism for dynamically allocating resources. In order to capture the network dynamics, we allow the CSM to repeatedly auction the available spectrum opportunities based on the PUs' behaviors. Meanwhile, each SU is allowed to strategically adapt its bidding strategy based on information about the available spectrum opportunities, its source and channel characteristics, and the impact of the other SUs bidding actions. Using this stochastic wireless allocation framework, we can develop a learning methodology for SUs to improve their policies for playing the auction game, i.e. the policies for generating the bids to compete for the available resources. The detailed learning algorithm is presented in [2]. Specifically, during the repeated multi-user interaction, the SUs can observe partial historic information of the outcome of the auction game, through which the SUs can estimate the impact on their future rewards and then adopt their best response in order to effectively compete for the channel opportunities.

The paper is organized as follows. In Section II, we introduce a stochastic game formulation for multi-user interaction in the cognitive radio networks. In Section III, we specify the details of the stochastic game in our cognitive radio networks and characterize the best response to play this game. In Section IV, we present the simulation results, followed by the conclusions in Section V.

## II. DYNAMIC MULTI-USER WIRELESS RESOURCE GAME FORMULATION

As mentioned in the introduction, we focus in this paper on developing wireless resource markets for secondary networks (SN). In SN, the SUs can opportunistically utilize the network resources that are vacated by the PUs. For illustration purposes, we assume that the SN consists of $M$ SUs, which are indexed by $i \in \{1,\cdots,M\}$. The SUs compete for the dynamically available transmission opportunities based on their own "private" information and knowledge about other SUs and available resources (and/or PUs' behaviors). In each time slot $\Delta T$, the SUs compete with each other for spectrum access and, given the allocated transmission opportunities, they deploy optimized cross-layer strategies to transmit their delay-sensitive bitstreams.

This work was supported in part by NSF CAREER Award CCF-0541867 and in part by grants from ONR.

During each time slot, a "state" of network resources can be defined to represent the available transmission opportunities in a SN, which is denoted by $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$, where $\boldsymbol{\mathcal{W}}$ is the set of possible resource states. We can also define "states" for the SUs. For instance, the states may represent their private information, which includes the traffic and channel characteristics. The current state of a SU $i$ is denoted by $s_i \in \mathcal{S}_i$, where $\mathcal{S}_i$ is the set of possible states of SU $i$.

At each time slot, SU $i$ will deploy an action to compete for the network resources. This action is denoted by $a_i \in \mathcal{A}_i$, where $\mathcal{A}_i$ is the set of possible external actions. An example of the actions in wireless networks is the selected transmit power in interference channels or the declared resource request like the TSPEC in 802.11e WLANs.

In this paper, we formulate the multi-user wireless cognitive resource competition as a stochastic game. Formally, the stochastic game is defined as a tuple $(\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{A}}, \boldsymbol{P_s}, P_w, \boldsymbol{\mathcal{R}})$, where $\boldsymbol{\mathcal{I}}$ is the set of agents (SUs), i.e. $\boldsymbol{\mathcal{I}}=\{1,...,M\}$, $\boldsymbol{\mathcal{S}}$ is the set of state profiles of all SUs, i.e. $\boldsymbol{\mathcal{S}}=\mathcal{S}_1 \times \cdots \times \mathcal{S}_M$ with $\mathcal{S}_i$ being the state set of SU $i$, $\boldsymbol{\mathcal{W}}$ is the set of network resource state. $\boldsymbol{\mathcal{A}}$ is the joint action space $\boldsymbol{\mathcal{A}}=\mathcal{A}_1 \times \cdots \times \mathcal{A}_M$, with $\mathcal{A}_i$ being the external action set available for SU $i$ to play the resource sharing game, $\boldsymbol{P_s}$ is a transition probability function defined as a mapping from the current state profile $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, corresponding joint actions $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ and the next state profile $\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}$ to a real number between 0 and 1, i.e. $\boldsymbol{P_s}: \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{S}} \mapsto [0,1]$. $P_{\boldsymbol{w}}$ is a transition probability function defined as a mapping from the current resource state $\boldsymbol{w} \in \boldsymbol{\mathcal{W}}$ and the next state $\boldsymbol{w}' \in \boldsymbol{\mathcal{W}}$ to a real number between 0 and 1, i.e. $P_{\boldsymbol{w}}: \boldsymbol{\mathcal{W}} \times \boldsymbol{\mathcal{W}} \mapsto [0,1]$. $\boldsymbol{\mathcal{R}}$ is a reward vector function defined as a mapping from the available resource $w$, the current state profile $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$ and corresponding joint actions $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ to an $M$-dimensional real vector with each element being the reward to a particular agent, i.e. $\boldsymbol{\mathcal{R}}: \boldsymbol{\mathcal{W}} \times \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{A}} \mapsto \mathbb{R}^M$.

The state transition $P_{\boldsymbol{w}}$ for the network resource state is determined by the PUs, and not by the SUs. In other words, the SUs' actions will not affect the network resource state transition. This structure actually opens an opportunity to allow the PUs to be agents with higher priorities in this stochastic game.

## III. Specification of Stochastic Game for Coginitve Radio Networks

As an illustration example, we consider that the SN can be formed across $N$ channels, each indexed by $j \in \{1,...,N\}$. At each time slot, each channel is assumed to be in one of the following two states: ON (this channel is currently used by the PUs) or OFF (this channel is not used by the PUs and hence is an opportunity for the SUs to use). Within each time slot, the channel is only OFF or ON [8]. At time slot $t \in \mathbb{N}$, the availability of each channel $j$ is denoted by $w_j^t \in \{0,1\}$, with $w_j^t$ being 0 if the channel is in ON state, and being 1 if it is in OFF state. The channel availability profile for the $N$ channels is represented by $\boldsymbol{w}^t = \left[w_1^t,...,w_N^t\right]$, which is the state of the network resource at time slot $t$.

### A. States for each SU

We assume that the SU deploys a delay-sensitive application. The data of the application layer is packetized with an average packet length $\ell$. In this paper, we consider multimedia applications, where the application packets have a hard delay deadline, i.e. the packets will expire in $J$ stages after they are ready for transmission. Then, we can define the state of the buffer as $\boldsymbol{v}_i^t = \left[v_{i1}^t,\cdots,v_{iJ}^t\right]^T$, where $v_j^t (1 \le j \le J)$ is the number of packets waiting for transmission that have a remaining life time of $j$ time slots.

The condition of channel $j$ experienced by SU $i$ is represented by the Signal-to-Noise Ratio (SNR) and it is denoted as $c_{ij}^t$ in dB. The channel condition profile is given by $\boldsymbol{c}_i^t = \left[c_{i1}^t,\cdots,c_{iN}^t\right]$. To model the dynamics experienced by SU $i$ at time $t$ in the cognitive radio network, we define a "state" $s_i^t = (\boldsymbol{v}_i^t, \boldsymbol{c}_i^t) \in \mathcal{S}_i$, which encapsulates the current buffer state as well as the state of each channel.

The environment experienced by each SU is characterized by the packet arrivals from the (multimedia) source (i.e. source/traffic characterization) connected with the transmitter, the spectrum opportunities released by PUs and the channel conditions. Different models can be used by a SU to characterize the environment. However, the accuracy of the deployed models will only affect the performance of the proposed solution, and not the general framework for multi-user stochastic game model presented here.

### B. State transition and stage reward

Since the network resource state is not affected by the actions performed by the SUs, the transition of $\boldsymbol{w}^t$ can be modeled as a finite state Markov chain [8]. The transition probability is denoted by $p_{\boldsymbol{w}}\left(\boldsymbol{w}^{t+1} \mid \boldsymbol{w}^t\right)$. In this example, we assume that the transition probability $p_{\boldsymbol{w}}\left(\boldsymbol{w}^{t+1} \mid \boldsymbol{w}^t\right)$ is known by all the SUs and CSM. However, more complicated models for the network resource state transition can also be involved in our stochastic game framework [9].

When SU $i$ receives the resource allocation $\boldsymbol{z}_i^t$, it can transmit $n_i^t$ packets during time slot $t$, which is denoted as

$$n_i\left(s_i^t, \boldsymbol{z}_i^t\right) = \left\lfloor \frac{\Psi\left(\boldsymbol{c}_i^t, \boldsymbol{z}_i^t\right)\triangle T}{\ell} \right\rfloor, \tag{1}$$

where $\Psi(\cdot)$ is the effective rate function, the form of which depends on the protocols implemented at the SU.

Then, the buffer state can be updated as

$$\begin{bmatrix} v_{i1}^{t+1} \\ \vdots \\ v_{ij}^{t+1} \\ \vdots \\ v_{iJ}^{t+1} \end{bmatrix} = \begin{bmatrix} v_{i2}^t - \max\left(n_i^t - v_{i1}^t, 0\right) \\ \vdots \\ v_{i(j+1)}^t - \max\left(n_i^t - \sum_{m=1}^{j} v_{im}^t, 0\right) \\ \vdots \\ Y_i^t \end{bmatrix}, \tag{2}$$

where $Y_i^t$ is the random variable representing the number of packets arriving at time slot $t$ having life time $J$. The distribution of $Y_i^t$ is denoted by $p_{Y_i^t}(l)$. Hence, the transition probable is given by

$$p\left(\boldsymbol{v}_i^{t+1} \mid \boldsymbol{v}_i^t, \boldsymbol{z}_i^t\right) = \begin{cases} P_{Y_i^t}(l) & \begin{array}{l} if\ \boldsymbol{v}_i^{t+1}\ satisfies\ eq.(2) \\ \&\ Y_i^t = l \end{array} \\ 0 & o.w. \end{cases} \tag{3}$$

The channel condition $\boldsymbol{c}_i^t$ depends on the channel gain and the power level for transmission. The channel gain is generally modeled as a FSMC [10]. In this example, we also consider that the power allocation is constant during the data transmission, and hence, the channel condition $\boldsymbol{c}_i^t$ can be formulated as a FSMC with transition probability $p\left(\boldsymbol{c}_i^{t+1} \mid \boldsymbol{c}_i^t\right)$.

The state transition probability for SU $i$ is given by

$$p_s\left(s_i^{t+1} \mid s_i^t, \boldsymbol{z}_i^t\right) = p\left(\boldsymbol{v}_i^{t+1} \mid \boldsymbol{v}_i^t, \boldsymbol{z}_i^t\right) p\left(\boldsymbol{c}_i^{t+1} \mid \boldsymbol{c}_i^t\right). \quad (4)$$

Here, we assume that the transition of the channel condition is independent of the transition of buffer state. The utility for the delay-sensitive application at time slot $t$ is defined here as

$$u_i\left(s_i^t, \boldsymbol{z}_i^t\right) = \min\left(n_i^t, \sum_{j=1}^{J} v_{ij}^t\right) - \lambda_g \max\left\{v_{i,1}^t - n_i^t, 0\right\}, \quad (5)$$

where $\lambda_g$ is the parameter to trade-off the received and lost packets. More sophisticated utility formulations for multimedia transmission, which consider the explicit impact on the multimedia quality (e.g. PSNR) can be found in [1].

### C. Resource allocation rule

We model the multi-user wireless resource allocation as an auction for spectrum opportunities held by the CSM during each time slot. The SUs calculate the external action $a_i^t$ based on the information about the network resources, and their own private information about the environment they experience.. In this auction game, the external action is the competition bid, i.e. $a_i^t$ is the amount of bid submitted to the CSM. We use the external action and bid interchangeably. Subsequently, each SU submits the bid $a_i^t$ to the CSM. After receiving the bid vectors from the SUs, the CSM computes the channel allocation $\boldsymbol{z}_i^t$ for each SU $i$ based on the submitted bids. To compel the SUs to declare their bids truthfully [11], the CSM also computes the payment $\tau_i^t \in \mathbb{R}_-$ that the SUs have to pay for the use of resources during the current stage of the game. The negative value of the payment means the absolute value that SU $i$ has to pay the CSM for the used resources. The auction result is then transmitted back to the SUs which can deploy their transmission strategies in different layers and send data over the assigned channel. After the data transmission, another auction starts at the next time slot $t+1$. The computation of the allocation $\boldsymbol{z}_i^t$ and payment $\tau_i^t$ is described as follows.

After each SU submits the bid vector, the CSM performs two computations: (i) channel allocation and (ii) payment computation. During the first phase, the CSM allocates the resources to SUs based on its adopted fairness rule, e.g. maximizing the total "social welfare":

$$\boldsymbol{z}^{t,opt} = \arg\max_{\boldsymbol{z}^t} \sum_{i=1}^{M} \tilde{h}_i\left(a_i^t, \boldsymbol{z}_i^t, w\right), \quad (6)$$

where $\tilde{h}_i(\cdot)$ is the utility function of SU $i$ seen by the CSM. Note that this utility can be represented by either the effective rate or time on the network allocated to each user, or it can be determined in the utility domain, by considering the resulting utility-rate functions of the deployed multimedia coders [1].

We will consider in this paper, for illustration, a second price auction mechanism [12] for determining the tax that needs to be paid by SU $i$ based on the above optimal channel assignment $\boldsymbol{z}^{t,opt}$. This tax equals:

$$\tau_i^t = \sum_{j=1, j\neq i}^{M} \tilde{h}_j\left(a_j^t, \boldsymbol{z}_j^{t,opt}, w\right) - \max_{\boldsymbol{z}_{-i}^t} \sum_{j=1, j\neq i}^{M} \tilde{h}_j\left(a_i^t, \boldsymbol{z}_i^t, w\right). \quad (7)$$

For simplicity, we can denote the output of the resource allocation game as $\boldsymbol{r}^t = \left(\boldsymbol{z}^t, \boldsymbol{\tau}^t\right) = \Omega\left(\boldsymbol{a}^t, w^t\right)$.

### D. Selecting the Policy for Playing the Resource Management Game

In the cognitive radio network, we assume that the stochastic game is played by all SUs for an infinite number of stages. This assumption is reasonable for applications having a long duration, such as video streaming, videoconferencing etc. In our network setting, we define a history of the stochastic game up to time $t$ as $\boldsymbol{h}^t = \{\boldsymbol{s}^0, w^0, \boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{z}^0, \boldsymbol{\tau}^0, ..., \boldsymbol{s}^{t-1}, w^{t-1}, \boldsymbol{a}^{t-1}, \boldsymbol{b}^{t-1}, \boldsymbol{z}^{t-1}, \boldsymbol{\tau}^{t-1}, \boldsymbol{s}^t\} \in \boldsymbol{\mathcal{H}}^t$, which summarizes all previous states and the actions taken by the SUs as well as the outcomes at each stage of the auction game and $\boldsymbol{\mathcal{H}}^t$ is the set of all possible history up to time $t$. However, during the stochastic game, each SU $i$ cannot observe the entire history, but rather part of the history $\boldsymbol{h}^t$. The observation of SU $i$ is denoted as $\boldsymbol{o}_i^t \in \boldsymbol{\mathcal{O}}_i^t$ and $\boldsymbol{o}_i^t \subset \boldsymbol{h}^t$. Note that the current state $\boldsymbol{s}_i^t$ can be always observed, i.e. $\boldsymbol{s}_i^t \in \boldsymbol{o}_i^t$. Then, a bidding policy $\pi_i^t : \boldsymbol{\mathcal{O}}_i^t \mapsto \boldsymbol{\mathcal{A}}_i$ for SU $i$ at the time $t$ is defined as a mapping from the observations up to the time $t$ into the specific action, i.e. $a_i^t = \pi_i^t(\boldsymbol{o}_i^t)$. Furthermore, a policy profile $\boldsymbol{\pi}_i$ for SU $i$ aggregates the bidding policies about how to play the game over the entire course of the stochastic game, i.e. $\boldsymbol{\pi}_i = (\pi_i^0, ..., \pi_i^t, ...)$. The policy profile for all the SUs at time slot $t$ is denoted as $\boldsymbol{\pi}^t = \left(\pi_1^t, ..., \pi_M^t\right) = \left(\pi_i^t, \boldsymbol{\pi}_{-i}^t\right)$.

The reward for SU $i$ at the time slot $t$ is $R_i^t(s_i^t, \boldsymbol{r}_i^t) = u_i\left(s_i^t, \boldsymbol{z}_i^t\right) + \tau_i^t$. Since the resource allocation also depends on other SUs' states and external actions, the reward is further expressed by $R_i^t(s_i^t, \Omega\left(a_i^t, \boldsymbol{a}_{-i}^t, \boldsymbol{w}^t\right))$. We define the *best response* $\beta_i$ for SU $i$ to other SUs' policies $\boldsymbol{\pi}_{-i}^t$ as

$$\beta_i(\boldsymbol{\pi}_{-i}^t) = \arg\max_{\pi_i} Q_i^t((\pi_i^t, \boldsymbol{\pi}_{-i}^t) \mid \boldsymbol{s}^t, \boldsymbol{w}^t) \quad (8)$$

where $Q_i^t(\pi_i^t, \boldsymbol{\pi}_{-i}^t \mid \boldsymbol{s}^t, \boldsymbol{w}^t)$ is the total discounted sum of rewards which is defiend as

$$Q_i^t((\pi_i^t, \boldsymbol{\pi}_{-i}^t) \mid \boldsymbol{s}^t, \boldsymbol{w}^t) = \sum_{k=t}^{\infty} (\alpha_i)^{k-t} R_i^k\left(s_i^k, \Omega\left(a_i^k, \boldsymbol{a}_{-i}^k, \boldsymbol{w}^k\right)\right) \quad (9)$$

The factor $\alpha_i\ (0 \leq \alpha_i < 1)$ is the discounted factor determined by a specific application (for instance, for video streaming applications, this factor can be set based on the tolerable delay). The total discounted sum of rewards in Eq. (9) consists of two parts: (i) the current stage reward and (ii) the expected future reward discounted by $\alpha_i$. Note that SU $i$ cannot independently determine the above value without explicitly knowing the policies and states of other SUs. The SU maximizes the total discounted sum of future rewards in order to select the bidding policy, which explicitly considers the impact of the current bid vector on the expected future rewards.The central issue in the stochastic game is how the best response policies can be determined by the SUs. This will be discussed in Section III.E.

### E. Characterizing the best response policy

Recall that during each time slot, the CSM announces an auction based on the available resources and then SUs bid for the resources. To enable the successful deployment of this resource auction mechanism, we can prove, similarly to our prior work in [1], that SUs have no incentive to misrepresent their information, i.e. they adhere to the "truth telling" policy. We assume that at each time slot $t$, SU $i$ has preference $\mathcal{U}_{ij}^t$ over the channel $j$, which capture the benefit derived when using that channel. The preference $\mathcal{U}_{ij}^t$ is interpreted as the benefit obtained by SU $i$ when using channel $j$, compared to

the benefit when this channel is not used. Note that this benefit also includes the expected future rewards. The optimal bid $a_{ij}^{t,opt}$ that SU $i$ can take on the channel $j$ at time $t$ is the bid maximizing the net benefit $\mathcal{U}_{ij}^t + \tau_i^t$. In the auction discussed in Section III.C, the optimal bid that SU $i$ can make is $a_{ij}^{t,opt} = \mathcal{U}_{ij}^t$, i.e. the optimal bid for SU $i$ is to announces its true preference to the CSM [1]. The proof is omitted here due to space limitations, since it is similar to that in [1]. The payment made by SU $i$ is computed by the CSM based on the inconvenience incurred by other SUs due to SU $i$ during that time slot [1].

Next, we define the preference $\mathcal{U}_{ij}^t$ in the context of the stochastic game model. Using the channel $j$ when it is available, SU $i$ obtains the immediate gain $u(s_i^t, \boldsymbol{e}_j)$ by transmitting the packets in its buffer, where $\boldsymbol{e}_j$ indicates that channel $j$ is allocated to SU $i$ during the current time slot. SU $i$ then moves into next state $\boldsymbol{s}_i^{t+1}$ from which it may obtain the future reward $Q_i^t(\boldsymbol{\pi}^{t+1} \mid \boldsymbol{s}^{t+1}, \boldsymbol{w}^{t+1})$. On the other hand, if no channel is assigned to SU $i$, it receives the immediate gain $u(s_i^t, \mathbf{0})$ and then moves into the next state $s_i^{t+1}$ from which it may obtain the future reward $Q_i^t(\boldsymbol{\pi}^{t+1} \mid \boldsymbol{s}^{t+1}, \boldsymbol{w}^{t+1})$. We define a feasible set of channel assignments to SU $i$'s opponents, given SU $i$'s channel allocation $\boldsymbol{z}_i^t$, as $\boldsymbol{\mathcal{Z}}_{-i}^t(\boldsymbol{z}_i^t)$, with $\boldsymbol{\mathcal{Z}}_{-i}^t(\boldsymbol{z}_i^t) = \{Z_{-i}^t \mid \sum_{k=1,k\neq i}^{M} z_{kj}^t = y_j^t - \boldsymbol{z}_i^t, \forall j, \sum_{j=1}^{N} z_{kj}^t \leq 1, \forall k \neq i, z_{kj}^t \in \{0,1\}\}$.

The preference over the current state can be then computed as

$$\mathcal{U}_{ij}^t(\boldsymbol{s}^t, \boldsymbol{w}^t) = \left[ u_i^t(s_i^t, \boldsymbol{e}_j) + \alpha_i \cdot \sum_{\substack{\boldsymbol{s}^{t+1}\in\boldsymbol{\mathcal{S}} \\ \boldsymbol{w}^{t+1}\in\boldsymbol{\mathcal{W}}}} \left[ p_s(s_i^{t+1} \mid s_i^t, \boldsymbol{e}_j)\, p_w(\boldsymbol{w}^{t+1} \mid \boldsymbol{w}^t) \cdot \sum_{Z_{-i}^t \in \boldsymbol{\mathcal{Z}}_{-i}^t(\boldsymbol{e}_j)} \prod_{k=1}^{M} p_s(s_k^{t+1} \mid s_k^t, \boldsymbol{z}_k^t) Q_i^{t+1}(\boldsymbol{\pi} \mid \boldsymbol{s}^{t+1}, \boldsymbol{w}^{t+1}) \right] \right] - \left[ u_i^t(s_i^t, \mathbf{0}) + \alpha_i \cdot \sum_{\substack{\boldsymbol{s}^{t+1}\in\boldsymbol{\mathcal{S}} \\ \boldsymbol{w}^{t+1}\in\boldsymbol{\mathcal{W}}}} \left[ p_s(s_i^{t+1} \mid s_i^t, \mathbf{0})\, p_w(\boldsymbol{w}^{t+1} \mid \boldsymbol{w}^t) \cdot \sum_{Z_{-i}^t \in \boldsymbol{\mathcal{Z}}_{-i}^t \cdot \mathbf{0}} \prod_{k=1}^{M} p_s(s_k^{t+1} \mid s_k^t, \boldsymbol{z}_k^t) Q_i^{t+1}(\boldsymbol{\pi} \mid \boldsymbol{s}^{t+1}, \boldsymbol{w}^{t+1}) \right] \right] \quad (10)$$

From this equation, it is clear that the true value $\mathcal{U}_{ij}^t$ depends on its own current state $\boldsymbol{s}_i^t$, but also the other SUs' states $\boldsymbol{s}_{-i}^t$, the channel allocations $\boldsymbol{\mathcal{Z}}_{-i}^t(\boldsymbol{e}_j)$ to the other users when channel $j$ is assigned to SU $i$, $\boldsymbol{\mathcal{Z}}_{-i}^t(\mathbf{0})$ when SU $i$ is not assigned to any channel, and the state transition models $p_s(s_k^{t+1} \mid s_k^t, \boldsymbol{z}_k^t), \forall k$. However, the other SUs' states, the channel allocations and the state transition models of other SUs are not known to SU $i$, and it is thus impossible for each SU to determine its preference $\mathcal{U}_{ij}^t(\boldsymbol{s}^t, \boldsymbol{w}^t)$.

Without knowing the other SUs' states and state transition models, SU $i$ cannot derive its optimal bidding strategy $a_{ij}^{t,opt} = \mathcal{U}_{ij}^t(\boldsymbol{s}^t, \boldsymbol{w}^t)$. However, if SU $i$ chooses the bid vector by only maximizing the immediate reward $u_i(s_i^t, \boldsymbol{z}_i^t) + \tau_i^t$, i.e. the total discounted sum of reward degenerates in $Q_i^t(\boldsymbol{s}^t, \boldsymbol{w}^t, \boldsymbol{\pi}) = u_i(s_i^t, \boldsymbol{z}_i^t) + \tau_i^t$ by setting $\alpha_i = 0$. Then, the preference over channel $j$ becomes $\mathcal{U}_{ij}^t(\boldsymbol{s}^t, \boldsymbol{y}^t) = u_i(s_i^t, \boldsymbol{e}_j) - u_i(s_i^t, \mathbf{0})$. Since now $\mathcal{U}_{ij}^t$ only depends on the state $s_i^t$, SU $i$ can compute both the optimal bid vector as well as the optimal bidding policy. We refer to this optimal bidding policy as the "myopic" policy, since it only takes the immediate reward into consideration and ignores the future impact. The myopic policy is referred to as $\pi_i^{myopic}$. To solve the difficult problem of optimal bidding policy selection when $\alpha_i \neq 0$, an SU needs to forecast the impact of its current bidding actions on the expected future rewards discounted by $\alpha_i$. The forecast can be performed using learning from its past experiences.

### F. Learning for playing the game

A key question is what needs to be learned within a wireless stochastic game in order to improve the policy of an SU. Recall that the optimal bidding policy for SU $i$ is to generate a bid vector that represents its preferences $\mathcal{U}_{ij}^t, \forall j$ for using different channels. From III.E, we can see that SU $i$ needs to learn: (i) the state space of other SUs, i.e. $\boldsymbol{\mathcal{S}}_{-i}$; (ii) the current state of other SUs, i.e. $\boldsymbol{s}_{-i}^t$; (iii) the transition probability of other SUs, i.e. $\prod_{k\neq i} p_s(\boldsymbol{s}_k^{t+1} \mid \boldsymbol{s}_k^t, \boldsymbol{z}_k^t)$; (iv) the resource allocation $\boldsymbol{\mathcal{Z}}_{-i}^t(\boldsymbol{e}_j), \forall j$ and $\boldsymbol{\mathcal{Z}}_{-i}^t(\mathbf{0})$; and (v) the discounted sum of rewards $Q_i^{t+1}(\boldsymbol{\pi}^t \mid (\boldsymbol{s}_i^{t+1}, \boldsymbol{s}_{-i}^{t+1}), \boldsymbol{w}^t)$. However, SU $i$ can only observes the information $\boldsymbol{o}_i^t = \{\boldsymbol{s}_i^0, \boldsymbol{w}^0, \boldsymbol{a}_i^0, \boldsymbol{b}_i^0, \boldsymbol{z}_i^0, \boldsymbol{\tau}_i^0, ..., \boldsymbol{s}_i^{t-1}\boldsymbol{w}^{t-1}, \boldsymbol{a}_i^{t-1}, \boldsymbol{b}_i^{t-1}, \boldsymbol{z}_i^{t-1}, \boldsymbol{\tau}_i^{t-1}, \boldsymbol{s}_i^t\}$ from which SU $i$ cannot accurately infer the other SUs' state space and transition probability. Moreover, capturing the exact information about other SUs requires heavy computational and storage complexity.

Instead, we allow SU $i$ to classify the space $\boldsymbol{\mathcal{S}}_{-i}$ into $H_i$ classes each of which is represented by a representative state $\tilde{\boldsymbol{s}}_{-i,h}, h \in \{1, ..., H_i\}$. By dividing the state space $\boldsymbol{\mathcal{S}}_{-i}$, the transition probability $\prod_{k\neq i} p_s(\boldsymbol{s}_k^{t+1} \mid \boldsymbol{s}_k^t, \boldsymbol{z}_k^t)$ is approximated by $p_s(\tilde{s}_{-i}^{t+1} \mid \tilde{s}_{-i}^t, \boldsymbol{z}_i^t)$, where $\tilde{s}_{-i}^t$ and $\tilde{s}_{-i}^{t+1}$ are the representative states of the classes that $\boldsymbol{s}_{-i}^t$ and $\boldsymbol{s}_{-i}^{t+1}$ belong to. This approximation is performed by aggregating all other SUs' states into one representative state and assuming that the transition depends on the resource allocation $\boldsymbol{z}_i^t$. Note that the classification on the state space $\boldsymbol{\mathcal{S}}_{-i}$ and approximation of the transition probability and discounted sum of rewards affects the learning performance. Hence, a user should tradeoff an increased learning complexity for an increased learning performance. The transition probability $p_s(\tilde{s}_{-i}^{t+1} \mid \tilde{s}_{-i}^t, \boldsymbol{z}_i^t)$ can be approximated using occurrence frequency and the average rewards $V_i^{t+1}((\boldsymbol{s}_i^{t+1}, \tilde{s}_{-i}^{t+1}))$ can be learned using the algorithm similar to the Q-learning [13]. The detailed learning algorithm is presented in [2].

## IV. Simulation Results

In this section, we aim at quantifying the performance of our proposed stochastic interaction and learning framework. We assume that the SUs compete for the available spectrum opportunities in order to transmit delay-sensitive multimedia data.

In this simulation, we consider five SUs competing for the available channel opportunities in the WLAN-like cognitive radio network. The packet arrivals of all the five SUs are modeled using a Poisson process with the same average arrival rate of 2Mbps. The number of channels is 3 and the channel condition of all the five SUs on each channel takes only three values ($K = 3$), which are 18dB, 23dB and 26dB. The transition probabilities are $p_{ij}^{0\to1} = p_{ij}^{0\to2} = 0.4, p_{ij}^{0\to3} = 0.2, p_{1j}^{l\to1} = p_{1j}^{l\to2} = 0.4, p_{1j}^{l\to3} = 0.2, \forall i, j, l$. The parameters of the model of the availability of the channels to the SUs are $p_j^{NF} = 0.7, p_j^{FN} = 0.3$. The length of the time slot $\Delta T$ is also $10^{-2}$ s. Similar parameters are used for the five SUs in order to clearly illustrate the performance differences obtained based on the different strategies.

In this simulation, we consider only two scenarios. In scenario (1), all SUs deploy a myopic bidding strategy $\pi_i^{myopic}, i = 1,2,...,5$, while in scenario (2), SU 5 deploys the multi-user learning-based bidding strategy $\pi_5^{\mathcal{L}_i}$ with the disc and the other SUs deploy the myopic bidding strategy $\pi_i^{myopic}, i = 1,...,4$. The packet loss rate and cost per time slot incurred by the SUs are presented in Table 1. The accumulated packet loss and cost of SU 5 for the five scenarios are plotted in Figure 1(a) and (b), respectively. The average tax and cost is again computed within a time window of $T = 1000$ slots.

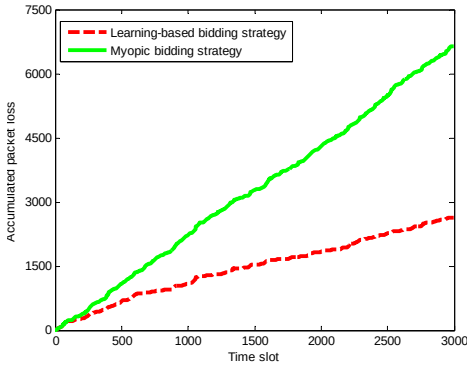
From Table 1, we note that SU 5 significantly reduces the packet loss rate by 14.6% and average cost by 16.1% by adopting the best response learning-based bidding strategy. Figure 1(a) and (b) further verify the improvement of the performance for SU 5. However, other SUs' performances are decreased, as they need now to compete against a learning SU (i.e. SU 5), which is able to make better bids for the available resources.

## V. CONCLUSION

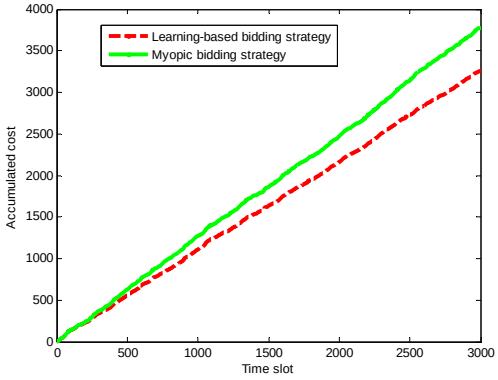
In this paper we model the cognitive radio resource allocation problem as a "stochastic game" played among strategic SUs. At each stage of the game, the CSM deploys a generalized second price auction mechanism to allocate the available spectrum resource. The SUs are allowed to simultaneously and independently make bid decision on that resource by considering their current states, experienced environment as well as the estimated future reward. To improve the bid decision at each stage, we propose a best response learning algorithm to predict the possible future reward at each state. The simulation results show that our proposed learning algorithm can significantly improve the SUs' performance. Our future work will focus on analyzing the performance of cognitive radio networks where multiple SUs are deploying various learning strategies and protocols.

Table 1. Performance of SU 1~5 in the five SUs network

| | SU 1 | | SU 2 | |
|---|---|---|---|---|
| | Packet Loss Rate (%) | Average cost | Packet Loss Rate (%) | Average cost |
| 1 | 21.14 | 1.2002 | 19.99 | 1.1666 |
| 2 | 25.03 | 1.2992 | 24.20 | 1.2993 |
| | SU3 | | SU 4 | |
| 1 | 22.05 | 1.2123 | 21.37 | 1.1949 |
| 2 | 25.72 | 1.3338 | 26.02 | 1.3568 |
| | SU 5 | | | |
| 1 | 24.17 | 1.3101 | | |
| 2 | 9.56 | 1.0988 | | |




(a)




(b)

Figure 1. The accumulated packet loss and cost of SU 5 in the two scenarios, (a) accumulated packet loss over the time slot; (b) accumulated cost over the time slot

## REFERENCES


[1] F. Fu, and M. van der Schaar, "Non-collaborative resource management for wireless multimedia applications using mechanism design," IEEE Transaction on Multimedia, vol. 9, no. 4, pp. 851-868, Jun. 2007.

[2] F. Fu, and M. van der Schaar, "Learning to Compete for Resources in Wireless Stochastic Games," IEEE Transactions on Vehicular Technology, to appear.

[3] Federal Communications Commission, "Spectrum Policy Task Force," Rep. ET Docket No. 02-135, Nov. 2002.

[4] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," IEEE J. Sel. Areas Commun., vol. 23, no. 2, Feb. 2005.

[5] F. A. Ian, W.Y. Lee, M.C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless network: a survey," Computer Networks, vol 50, no. 13, Sept. 2006.

[6] D. Fudenberg, and D. K. Levine, "The theory of learning in games," Cambridge, MA: MIT Press, 1999.

[7] "IEEE 802.11e/D5.0, wireless medium access control (MAC) and physical layer (PHY) specifications: Medium access control (MAC) enhancements for Quality of Service (QoS), draft supplement," June 2003.

[8] S. Shankar, C.T. Chou, K. Challapali, and S. Mangold, "Spectrum agile radio: capacity and QoS implications of dynamic spectrum assignment," Global Telecommunications Conference, Nov. 2005.

[9] L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. Artificial Intelligence, Volume 101, pp. 99-134, 1998.

[10] Q. Zhang, and S.A. Kassam, "Finite-state Markov model for Rayleigh fading channels," IEEE Transaction on Communications, vol. 47, no. 11, Nov. 1999.

[11] M. Jackson, "Mechanism theory," In the Encyclopedia of Life Support Systems, 2003.

[12] P. Klemperer, "Auction theory: A guide to the literature," J. Economics Surveys, vol. 13, no. 3, pp. 227-286, Jul. 1999.

[13] C. Watkins, and P. Dayan, "Q-learning." Technical Note, Machine Learning, vol. 8, 279-292, 1992.