Game Theoretic Issues in Cognitive Radio Systems (Invited Paper)

Jane Wei Huang and Vikram Krishnamurthy
Statistical Signal Processing Research Lab
University of British Columbia, Vancouver, Canada
Email: {janeh, vikramk}@ece.ubc.ca

***Abstract*— The ability to model independent decision makers whose actions potentially affect all other decision makers makes game theory attractive to analyze the performances of wireless communication systems. Recently, there has been growing interest in adopting game theoretic methods to cognitive radio networks for power control, rate adaptation and channel access schemes. This work presents several results in game theory and their applications in cognitive radio systems. First, we compute the Nash equilibrium power allocation and rate adaptation policies in cognitive radio systems using static game and dynamic Markovian game frameworks. We then describe how mechanism design helps to design a truth revealing channel access scheme. Finally, we introduce the correlated equilibrium concept in stochastic games and its application to solve the transmission control problem in a cognitive radio system.**

***Index Terms*— Cognitive Radio, Game Theory, General-Sum Markovian Dynamic Game, Switching Control Game, Nash Equilibrium, Mechanism Design, Correlated Equilibrium**

I. Introduction

Game theory was first introduced by J. V. Neumann and O. Morgenstern in [1] in 1944. It is a discipline aiming at modeling situations in which decision makers have to make specific actions that have mutual, possibly conflicting consequences. Game theory has been widely used as an analysis tool in economic systems [2], [3]. Recently, with the introduction of ad-hoc networks and cognitive radio systems, has game theory been considered as an adequate tool to design wireless self-organized networks [4], [5]. Specifically, game theoretic models have been developed to better understand congestion control, routing, power control, trust management and other issues in wired and wireless communication systems. The goal of this paper is to present recent results in game theory applied to the design and analysis of cognitive radio networks.

A. Why is Game Theory Relevant to Cognitive Radio Systems?

With an increasing demand on data rates and new applications, spectrum crowding and congestion continue to grow. A report released by the Federal Communications Commission (FCC) in 2002 [6] suggests that while some spectrum bands are over-utilized and crowded, many other licensed spectrum bands are under-utilized. This fact motivates the development of new technologies and standards in wireless communication systems that seek to use these under-utilized licensed bands. The idea of cognitive radio systems [7]–[9] is one possible method to achieve more efficient utilization of the available spectrum resources. While the traditional approach to ensure the co-existence of multiple systems is to split the available spectrum into frequency bands and allocate them to different licensed (primary) users, the dynamic spectrum access in cognitive radio systems improves the spectrum utilization by detecting unoccupied spectrum holes and assigning them to unlicensed (secondary) users.

The term cognitive radio was coined in 1999 by J. Mitola III in [10]. Dynamic spectrum access is an important aspect of cognitive radio. It can be achieved in various ways including *underlaying*, *overlaying*, or *interweaving* the signals of secondary users with that of the primary users, while keeping the interferences as low as possible.

In cognitive radio networks with reduced functionality base stations (no central authority) and autonomous cognitive radios, game theory can be naturally applied to achieve the decentralized operation and self configuration features. In a game theoretic setting, cognitive radios can be viewed as selfish rational players each seeking to optimize its own utility. The interest of an individual cognitive radio may conflict with that of the network, in which case game theory can be straightforwardly applied, as it traditionally analyzes situations where player objectives are in conflict.

B. Cognitive Radio Power Control with Static Game Theoretic Approach

In a cognitive radio network, proper power control is of importance to ensure efficient operation of both primary and secondary users. Even without the presence of primary users, power control is still an issue among secondary users since the signal of one user may cause interference to the transmissions of others. Thus, how to develop an efficient power allocation scheme that is able to jointly optimize the performance of multiple cognitive radios in the presence of mutual interference is of interest to such a system.

Static game framework has been used to compute the Nash equilibrium power allocation policies in cognitive

Manuscript received June 29, 2009; revised August 3, 2009; accepted September 2, 2009.

doi:10.4304/jcm.4.10.790-802

radio networks [8], [11], [12]. Different from a dynamic game where players make sequence of decisions, a static game is one in which all players make decisions one time simultaneously, without knowledge of the strategies of other players. The Nash equilibrium of a game is a set of strategies, one for each user, such that no user has the incentive to unilaterally change its action. In a Nash equilibrium, any change in the strategy by a user would lead that user to have less payoff than if it keeps the current strategy. Section II gives an example of a cognitive radio system where each user aims to maximize its information rate subject to the transmission power constraint. A distributed asynchronous iterative water-filling algorithm is used to compute the Nash equilibrium power allocation policy of such a system using static game theoretic approach [12].

C. Switching Control Game: A Special Type of Stochastic Dynamic Game

Most games considered in wireless communication systems to date are static games. Stochastic dynamic game theory is an essential tool for cognitive radio systems as it is able to exploit the correlated channels in the analysis of decentralized behaviors of cognitive radios.

The concept of a stochastic game, first introduced by Lloyd Shapley in early 1950s, is a dynamic game played by one or more players. The elements of a stochastic game include system state set, action sets, transition probabilities and utility functions. It is an extension of the single player Markov decision process (MDP) to include the multiple players whose actions all impact the resulting payoffs and next state. A switching control game [13]–[15] is a special type of stochastic dynamic game where the transition probability in any given state depends on only one player. It is known that the Nash equilibrium for such a game can be computed by solving a sequence of Markov decision processes. Section III shows an example where the rate adapt problem in a Time Division Multiple Access (TDMA) cognitive radio system is formulated as a switching control Markovian game and a value iterative optimization algorithm is proposed to compute the Nash equilibrium for such a game [16].

D. Mechanism Design in Cognitive Radio Systems

An efficient spectrum assignment technology is essential to a cognitive radio system, which allows secondary users to opportunistically utilize the unoccupied spectrum holes based on agreements and constraints. These secondary users have to coordinate with each other in order to maintain the order and result in maximum efficiency. It motivates the development of spectrum access approaches in cognitive radio systems. The opportunistic scheduling in cognitive networks assuming the scheduling is fully aware of primary user transmissions are considered in [17] and [18], while [19]–[21] consider the scenario with only partial primary user activity information is available.

However, all the existing opportunistic scheduling approaches overlook the fact that the secondary users may be owned by different agents and they may work in competitive rather than cooperative manners. These selfish users can become so sophisticated that they lie about their states to optimize their own utilities at the cost of reducing the overall system performance. It requires mechanism design theory in order to prevent this from happening. Mechanism design is the study of designing rules for strategic, autonomous and rational players to achieve predictable global outcome [22] using game theoretic approach. A milestone in mechanism design is the Vickrey-Clark-Groves (VCG) mechanism, which is a generalization of Vickrey's second price auction [23] proposed by Clark [24] and Groves [25]. The particular pricing policy of the VCG mechanism makes reporting true values the dominant strategy for all the players. Section IV is an example where we model each user in a cognitive radio as a selfish player aiming to optimize his own utility and we try to find a mechanism which ensures efficient resource allocation within the network [26].

E. Correlated Equilibrium of a Dynamic Markovian Game

The fundamental solution concept for dynamic Markovian games is Nash equilibrium, however, it suffers from limitations, such as non-uniqueness, loss of efficiency, non-guarantee of existence. In game theory, a correlated equilibrium is a solution concept which is more general than the Nash equilibrium [27], [28]. A correlated equilibrium is defined as follows. Each player in a game chooses his action according to his observation of the value of a signal. A strategy assigns an action to every possible observation a player can make. If no player would deviate from the recommended strategy, the distribution is called a correlated equilibrium. Compared to Nash equilibria, correlated equilibria offer a number of conceptual and computational advantages, including the facts that new and sometimes more "fair" payoffs can be achieved, that correlated equilibria can be computed efficiently for games in standard normal form, and that correlated equilibria are the convergence notion for several natural learning algorithms. Furthermore, it has been argued that the correlated equilibria are the natural equilibrium concept consistent with the Bayesian perspective [28]. Section V is one of such examples where it formulates the user scheduling problem in a cognitive radio network using stochastic dynamical game framework with the goal of obtaining the correlated equilibrium policy [29].

F. Organization of this Paper

This paper is organized as follows: Section II formulates the power allocation problem in a cognitive radio system using static game framework. Section III then introduces a special type of dynamic game: A switching control game and uses it to solve the rate adaptation problem in a TDMA cognitive radio system. Section IV applies mechanism design to obtain a truth revealing

opportunistic scheduling algorithm and Section V computes the correlated equilibrium in a stochastic dynamic game. Finally, Section VI concludes the paper with some open issues on applying game theory in cognitive radio systems.

II. Distributed Transmission Power Control: Asynchronous Iterative Water-filling

An iterative water-filling algorithm (IWFA) is proposed in [11] to obtain the Nash equilibrium for multiuser power control problem in a digital subscriber line system. In the problem formulation, the user power allocation problem in a interference channel system is modeled as a noncooperative game, and the existence and uniqueness of a Nash equilibrium are established for a two-player version of such a game. However, the IWFA suffers from low convergence rate in a system with large number of users. In order to overcome this disadvantage, an improved asynchronous iterative water-filling algorithm (AIWFA) was proposed in [12]. The AIWFA is based on the asynchronous framework as described in [30] which allows all the users to update in a completely asynchronous way. This feature makes AIWFA applicable to all practical cognitive radio systems.

The system model we consider here is a Gaussian frequency-selective interference channel with multiple cognitive radio users and multiple receivers. It is aimed to find a distributed power allocation scheme without the coordination among users. In the system model, we assume there are K secondary users and each user has G subcarriers. $\mathcal{K} = \{1, 2, \ldots, K\}$ is used to denote the set of users. Denoting the power allocation of user k over subcarrier g as $p_k(g)$, the system constraint can be written as follows.

$$\sum_{g=1}^{G} p_k(g) \leq P_k, \tag{1}$$

$$p_k(g) \leq P_k^{\max}(g), \tag{2}$$

where P_k denotes the total transmission power of the kth user and $P_k^{\max}(g)$ denotes the power limit on the gth subcarrier of the kth user. (1) is the total power constraint on each user and (2) is the spectral mask constraint and it is imposed to eliminate the interference from each user over specified spectrum bands.

A. Problem Formulation

With the above system setup and constraints (1,2), each user aims to maximize its transmission rate in a distributed way. Denote R_k as the maximum achievable rate of the kth user, and it can be expressed as:

$$R_k = \frac{1}{G}\sum_{g=1}^{G} \log\left(1 + \frac{|h_{k,k}(g)|^2 p_k(g)}{\sigma_k^2(g) + \sum_{l=1,l\neq k}^{K} |h_{k,l}(g)|^2 p_l(g)}\right), \tag{3}$$

with $h_{i,j}$ denoting the channel quality between the jth cognitive radio user and ith receiver on the gth subcarrier. $|\sigma_k(g)|^2$ is used to denote the variance of the zero-mean circularly symmetric complex Gaussian noise at the kth receiver over the gth subcarrier. The term $\sum_{t=1,t\neq k}^{K} |h_{k,t}|^2 p_t(g)$ is the total interference caused by all other users to user k.

Using $\mathbf{p}_k = \{p_k(1), p_k(2), \ldots, p_k(G)\}$ to denote the power allocation vector of the kth user and $\mathbf{p}_{-k} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{k-1}, \mathbf{p}_{k+1}, \ldots, \mathbf{p}_K\}$ to denote the power allocation strategies of the remaining $K-1$ remaining users, the power allocation strategy of all the users in the system can be denoted as $\mathbf{p} = \{\mathbf{p}_1, \ldots, \mathbf{p}_K\} = \{\mathbf{p}_k, \mathbf{p}_{-k}\}$. We denote $\mathcal{P}_k$ as the set of transmission policies of user k that satisfy the system constraints (1, 2), it is specified as:

$$\mathcal{P}_k = \left\{\mathbf{p}_k : \sum_{g=1}^{G} p_k(g) \leq P_k,\ p_k(g) \leq P_k^{\max}(g)\right\}. \tag{4}$$

We use $\mathbf{p}^* = \{\mathbf{p}_1^*, \ldots, \mathbf{p}_K^*\}$ to denote the Nash equilibrium power allocation strategy. Given $\mathbf{p}_{-k}^*$, the optimal power allocation strategy of the kth user $\mathbf{p}_k^*$ is the solution to the following optimization problem:

$$\max_{\mathbf{p}_k} R_k(\mathbf{p}_k, \mathbf{p}_{-k}), \quad s.t. \quad \mathbf{p}_k \in \mathcal{P}_k, \ (\forall k \in \mathcal{K}). \tag{5}$$

Here, $R_k(\mathbf{p}_k, \mathbf{p}_{-k})$ (specified in (3)) is the maximum achievable transmission rate of the kth user.

B. Asynchronous Iterative Water-filling Algorithm

Based on the above problem formulation, an asynchronous iterative water-filling algorithm is proposed to obtain the Nash equilibrium policy [12]. We use n to denote the iteration index and $\mathcal{N} = \{0, 1, 2, \ldots\}$ to indicate the iteration index set. Due to the asynchronous feature of the AIWFA, not every user updates its power allocation strategy at each iteration n. We use $\mathcal{N}_k$ to indicate the iteration index set for user k where user k updates its policy $\mathbf{p}_k^n$. Here, $\mathbf{p}_k^n$ is specified as the power allocation policy of the kth user at the nth iteration.

The AIWFA is outlined in Algorithm 1. μ_k is the water-level parameter chosen to satisfy the power constraint of the kth user (1) and $[x]_a^b$ is the Euclidean projection of x onto the interval $[a, b]$. The algorithm can be summarized as follows. In step 1, we initialize the iteration index n and the initial power allocation vector $\mathbf{p}^0$, where $\mathbf{p}^0$ satisfies the system constraints (1,2). If $n \in \mathcal{N}_k$, we update the transmission policy of the kth user at the nth iteration $\mathbf{p}_k^n$ according to the water-filling algorithm, otherwise, the power allocation policy of the kth user remains unchanged. The algorithm terminates when $\mathbf{p}^n$ converges. It is shown in [12] that the convergence of Algorithm 1 is guaranteed if one of the following conditions is satisfied.

$$\begin{aligned} \frac{1}{w_k}\sum_{k\neq l} \max_{n\in\mathcal{D}_k\cap\mathcal{D}_l} \frac{|h_{k,l}|^2}{|h_{k,k}|^2}\frac{P_l}{P_k} w_l &< 1, \forall k \in \mathcal{K}; \\ \frac{1}{w_l}\sum_{k\neq l} \max_{n\in\mathcal{D}_k\cap\mathcal{D}_l} \frac{|h_{k,l}|^2}{|h_{k,k}|^2}\frac{P_l}{P_k} w_k &< 1, \forall l \in \mathcal{K}; \end{aligned} \tag{6}$$

where $\mathbf{w} = \{w_1, w_2, \ldots, w_K\}$ is any positive vector. $\mathcal{D}_k$ denotes the set $\{1, 2, \ldots, G\}$ possibly deprived by the subcarrier indicates the user k would never use as the best response set to any strategies used by the other users, for the given set of transmission power and propagation channels [12].

Algorithm 1 Asynchronous Iterative Water-filling Algorithm

Step 1: Set $n = 0$; Initialize $\mathbf{p}^0$ with $\mathbf{p}_k \in \mathcal{P}_k$ for $\forall k \in \mathcal{K}$.
Step 2: Update the transmission policy of each user:
for $k = 1 : K$ **do**
  if $n \in \mathcal{N}_k$ **then**
    for $g = 1 : G$ **do**
      $p_k^{n+1}(g) = \left[\mu_k - \frac{\sigma_k^2(g)+\sum_{l=1,l\neq k}^{K}|h_{k,l}(g)|^2 p_l(g)^n}{|h_{k,k}(g)|^2 p_l(g)^n}\right]_0^{P_k^{\max}(g)}$
    end for
  else
    $\mathbf{p}_k^{n+1} = \mathbf{p}_k^n$.
  end if
end for
Step 3: If $\mathbf{p}^{n+1} \neq \mathbf{p}^n$, then $n = n + 1$; otherwise, $\mathbf{p}^{n+1}$ is the system Nash equilibrium power allocation policy.

III. Switching Control Markovian Dynamic Game

In this section, we are going to formulate the second user rate adaptation problem in a cognitive radio network as a constrained general-sum switching control Markovian dynamic game. A switching control game [13]–[15] is a special type of game where the transition probability in any given state depends on only one player. It turns out that we can solve such type of game by a finite sequence of Markov decision processes.

The system model considers the secondary user rate adaptation problem in cognitive radio networks where multiple secondary users attempt to access a spectrum hole [16]. We assume a Time Division Multiple Access (TDMA) cognitive radio system model (as specified in the IEEE 802.16 standard [31]) that schedules one user per spectrum hole at each time slot according to a predefined decentralized scheduling policy. Therefore, the interaction among secondary users is characterized as a competition for the spectrum hole and can naturally be formulated as a dynamic game. By modeling transmission channels as correlated Markovian sources, the transmission rate adaptation problem for each user can be formulated as a general-sum switching control Markovian dynamic game with a latency constraint. The transmission policy of such a game takes into account the secondary user channel qualities, as well as the transmission delay of each secondary user.

A. System Description

This subsection introduces the system model (Fig. 1). We consider a TDMA system with K secondary users where only one user can access the channel at each time slot according to a predefined decentralized access rule. The access rule will be described later in this section. The correlated block fading channel of each user is modeled as a Markov chain. The rate control problem of each secondary user can then be formulated as a constrained Markovian dynamic game. More specifically, under the predefined decentralized access rule, the problem presented is a special type of game, namely a switching control Markovian dynamic game.

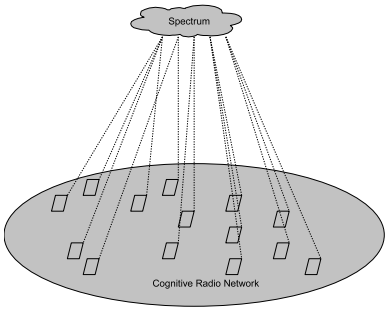

Figure 1. A K user cognitive radio system network where all the users are trying to access to the spectrum hole.

1) System States and TDMA Access Rule: We are going to denote the time slot index as t and $t \in \mathcal{T} = \{0, 1, 2, \dots\}$. The channel quality state of user k at time t is denoted as h_k^t and it is assumed to belong to a finite set $\{0, 1, \dots, Q_h\}$. The channel state can be obtained by quantizing a continuous valued channel model comprising of circularly symmetric complex Gaussian random variables that depend only on the previous time slot. The composition of channel states of all the K users can be written as $\mathbf{h}^t = \{h_1^t, \dots, h_K^t\}$. Assuming that the channel state $\mathbf{h}^t \in \mathcal{H}, t \in \mathcal{K}$ is block fading and each block length equals to one time slot, the channel state can be modeled using a finite states Markov chain model. The transition probability of the channel states from time t to $t+1$ can be denoted as $\mathbb{P}(\mathbf{h}^{t+1}|\mathbf{h}^t)$.

Let b_k^t denote the buffer occupancy state of user k at time t and it belongs to a finite set $b_k^t \in \{0, 1, \dots, L\}$. The composition of the buffer states of all the K users can be denoted as $\mathbf{b}^t = \{b_1^t, \dots, b_K^t\}$ and $\mathbf{b}^t$ is an element of the secondary user buffer state space $\mathcal{B}$.

New packets arrive at the buffer at each time slot and we denote the number of new incoming packets of the kth user at time t as f_k^t, $f_k^t \in \{0, 1, 2, \dots, \infty\}$. The composition of the incoming traffic of all the K users can be denoted as $\mathbf{f}^t = \{f_1^t, \dots, f_K^t\}$, it is an element of the incoming traffic space $\mathcal{F}$. For simplicity, the incoming traffic is assumed to be independent and identically distributed (i.i.d.) in terms of time index t and user index k. The incoming traffic is not a part of the system state but it affects the buffer state evolution.

Use $\mathbf{s}_k^t = [h_k^t, b_k^t]$ to denote the state of user k at time t, the system state at time t can then be denoted as $\mathbf{s}^t = \{\mathbf{s}_1^t, \dots, \mathbf{s}_K^t\}$. The finite system state space is denoted as $\mathcal{S}$, which comprises channel state $\mathcal{H}$ and secondary user buffer state $\mathcal{B}$. That is, $\mathcal{S} = \mathcal{H} \times \mathcal{B}$. Here $\times$ denotes a Cartesian product. Furthermore, $\mathcal{S}_k$ is used to indicate the state space where user k is scheduled for transmission. $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_K$ are disjoint subsets of $\mathcal{S}$ with the property of $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_K$.

The system adopts a TDMA cognitive radio system model (IEEE 802.16 [31]). A decentralized channel access algorithm can be constructed as follows: At the beginning of a time slot, each user k attempts to access the channel after a certain time delay WT_k^t. The time delay of user k

can be specified via an opportunistic scheduling algorithm [32], such as

$$WT_k^t = \frac{\gamma_k}{b_k^t h_k^t}. \quad (7)$$

Here γ_k is a user specified quality of service (QoS) parameter and $\gamma_k \in \{\gamma_p, \gamma_s\}$. If user k is a primary user, $\gamma_k = \gamma_p$, otherwise, $\gamma_k = \gamma_s$. By setting $\gamma_p << \gamma_s$, the network does not allow the transmission of secondary users with the presence of primary users. As soon as a user successfully access a channel, the remaining users detect the channel occupancy and stop their attempt to access. Use k^{*t} to denote the index of the first user which successfully accesses the spectrum hole, in the case where there are multiple users with the same minimum waiting time, k^{*t} is chosen from those users with equal probability.

2) Action and Costs: If the kth user is scheduled for transmission at time slot t, its action a_k^t represents the bits/symbol rate of the transmission. Assuming the system uses an uncoded M-ary quadrature amplitude modulation (QAM), different bits/symbol rates determine the modulation schemes, that is, $M = 2^{a_k^t}$.

Transmission cost: When user k is scheduled for transmission at time instant t, that is, $\mathbf{s}^t \in \mathcal{S}_k$, the cost function of user k depends only on a_k^t, as all the other users are inactive. Let $c_i(\mathbf{s}^t, a_k^t)$ denote the transmission cost of user $i, i \in \mathcal{K}$ at time t. Specifically, $c_i(\mathbf{s}^t, a_k^t)$ is chosen to be the bit error rate (BER) of user i during the transmission. Thus, the costs of all the users in the system can be specified as:

$$\begin{aligned} c_k(\mathbf{s}^t, a_k^t) &\geq 0 \\ c_{i,i\neq k}(\mathbf{s}^t, a_k^t) &= 0. \end{aligned} \quad (8)$$

Holding cost: Each user has an instantaneous QoS constraint denoted as $d_i(\mathbf{s}^t, a_k^t)$, $i = 1, \ldots, K$. If the QoS constraint is chosen to be the delay (latency constraint) then $d_i(\mathbf{s}^t, a_k^t)$ is a function of the buffer state b_i^t. The instantaneous holding costs will be subsequently included in an infinite horizon latency constraint.

B. Transition Probabilities and Switching Control Game Formulation

1) Transition Probabilities: With the above setup, the decentralized transmission control problem in a Markovian block fading channel cognitive radio system can now be formulated as a switching control game. In such a game [14], the transition probabilities depend only on the action of the kth user when $\mathbf{s} \in \mathcal{S}_k$. This feature enables us to solve such a game by a finite sequence of Markov decision processes. According to the property of the switching control game, when the kth user is scheduled for transmission, the transition probability between the current composite state $\mathbf{s}^t = [\mathbf{h}^t, \mathbf{b}^t]$ and the next state $\mathbf{s}^{t+1} = [\mathbf{h}^{t+1}, \mathbf{b}^{t+1}]$ depends only on the action of the kth user a_k^t, which can be specified as

$$\begin{aligned} &\mathbb{P}(\mathbf{s}^{t+1}|\mathbf{s}^t, a_k^t) \\ &= \prod_{i=1}^{K} \mathbb{P}(h_i^{t+1}|h_i^t) \cdot \prod_{i=1, i\neq k}^{K} \mathbb{P}(b_i^{t+1}|b_i^t) \cdot \mathbb{P}(b_k^{t+1}|b_k^t, a_k^t) \end{aligned} \quad (9)$$

The buffer occupancy of user k evolves according to Lindley's equation [33]

$$b_k^{t+1} = \min\left([b_k^t - a_k^t]^+ + f_k^t, L\right). \quad (10)$$

The buffer state of user $i \in \mathcal{K}$, $i \neq k$ evolves according to the following rule:

$$b_i^{t+1} = \min(b_i^t + f_i^t, L).$$

The buffer state transition probability of user k depends on its incoming traffic distribution and its action, which is

$$\mathbb{P}(b_k^{t+1}|b_k^t, a_k^t) = \begin{cases} \mathbb{P}(f_k^t = b_k^{t+1} - [b_k^t - a_k^t]^+) & b_k^{t+1} < L \\ \sum_{x=L-[b_k^t - a_k^t]^+}^{\infty} \mathbb{P}(f_k^t = x) & b_k^t = L \end{cases}.$$

For those users who are not scheduled for transmission, the buffer state transition probabilities only depends on the incoming traffic, which can be written as

$$\mathbb{P}(b_i^{t+1}|b_i^t) = \begin{cases} \mathbb{P}(f_i^t = b_i^{t+1} - b_i^t) & b_i^{t+1} < L \\ \sum_{x=L-b_i^t}^{\infty} \mathbb{P}(f_i^t = x) & b_i^{t+1} = L \end{cases}. \quad (11)$$

2) Switching Controlled Markovian Game Formulation: We use π_i $(i = 1, 2, \ldots, K)$ to denote the transmission policy vector of the ith user. With a slight abuse of notation, $\pi_i(\mathbf{s})$ is used to denote the transmission policy of user i in state $\mathbf{s}$ and is a component of π_i. $\pi_i(\mathbf{s})$ lives in the same space as the action a_i of the ith user. Assume at time instant t user k is scheduled for transmission according to the system access rule which is specified in (7). The infinite horizon expected total discounted cost of the ith $(i = 1, 2, \ldots, K)$ user under transmission policy π_i can be written as:

$$C_i(\pi_i) = \mathbb{E}_{\pi_i}\left[\sum_{t=1}^{\infty} \beta^{t-1} \cdot c_i(\mathbf{s}^t, a_k^t)\right] \quad (12)$$

where $0 \leq \beta < 1$ is the discount factor. The expectation of the above function is taken over the system state $\mathbf{s}^t$ which evolves over time index t. If we denote the holding cost of user i at the tth time slot as $d_i(\mathbf{s}^t, a_k^t)$, the infinite horizon expected total discounted latency constraint can be written as

$$D_i(\pi_i) = \mathbb{E}_{\pi_i}\left[\sum_{t=1}^{\infty} \beta^{t-1} \cdot d_i(\mathbf{s}^t, a_k^t)\right] \leq \widetilde{D}_i, \quad (13)$$

where $\widetilde{D}_i$ is a system parameter depending on the system requirement. Note here that we assume the latency constraint is valid in our problem formulation. $\widetilde{D}_i$ is chosen so that the set of policies that satisfy such a constraint is non-empty. This assumption will be discussed more specifically in Section III-C.

Equations (9,12,13) define a constrained switching control Markovian game. Our goal is to compute a Nash equilibrium policy $\pi_i^*, i \in \mathcal{K}$ (which is not necessarily unique) that minimizes the discounted transmission cost (12) subject to the latency constraint (13). The following result shows that a Markovian switching control game can be solved using a sequence of Markov decision processes (MDPs).

Result: [14, Chapter 3.2]The constrained switching control Markovian game (12,13) can be solved by a finite sequence of MDPs (as described in Algorithm 2). At each step, the algorithm iteratively updates the transmission policy π_i^n of user i given the transmission policies of

the remaining users. The optimization problem of each iteration can be mathematically written as:

$$\pi_i^{*n} = \{\pi_i^n : \min_{\pi_i} C_i^n(\pi_i) \text{ s.t. } D_i^n(\pi_i) \leq \widetilde{D}_i, i \in \mathcal{K}.\} \quad (14)$$

C. Value Iteration Algorithm

We present a value iterative algorithm in this subsection to compute a Nash equilibrium solution to the constrained Makovian dynamic game optimization problem described in (14). A value iterative optimization algorithm was designed to calculate the Nash equilibrium for an unconstrained general-sum dynamic Markovian switching control game [14]. Therefore, we first transfer the problem in (14) to an unconstrained one using Lagrangian dynamic programming and then apply the value iterative algorithm specified in Algorithm 2 to compute the Nash equilibrium solution.

Algorithm 2 Value Iterative Optimization Algorithm

Step 1:
Set $m = 0$; Initialize l.
Initialize $\{\mathbf{V}_1^0, \mathbf{V}_2^0, \ldots, \mathbf{V}_K^0\}$, $\{\lambda_1^0, \lambda_2^0, \ldots, \lambda_K^0\}$.
Step 2: Inner Loop: Set $n = 0$;
Step 3: Inner Loop: Update Transmission Policies;
for $k = 1 : K$ **do**
for each $s \in \mathcal{S}_k$,
$\pi_k^n(s) = \arg\min_{\pi_k^n(s)} \Big\{ c(s, a_k) + \lambda_k^m \cdot d_k(s, a_k) + \beta \sum_{s'=1}^{|\mathcal{S}|} \mathbb{P}(s'|s, a_k) v_k^n(s') \Big\}$;
$v_k^{n+1}(s) = c(s, \pi_k^n(s)) + \lambda_k^m \cdot d_k(s, \pi_k^n(s)) + \beta \sum_{s'=1}^{|\mathcal{S}|} \mathbb{P}(s'|s, \pi_k^n(s)) v_k^n(s')$;
$v_{i=1:K, i\neq k}^{n+1}(s) = \lambda_i^m \cdot d_i(s, \pi_k^n(s)) + \beta \sum_{s'=1}^{|\mathcal{S}|} \mathbb{P}(s'|s, \pi_k^n(s)) v_i^n(s')$;
end for
Step 4: If $\mathbf{V}_k^{n+1} \leq \mathbf{V}_k^n, k \in \mathcal{K}$, set $n = n + 1$, and return to Step 3; Otherwise, go to Step 5.
Step 5: Update Lagrange Multipliers
for $k = 1 : K$ **do**
$\lambda_k^{m+1} = \lambda_k^m + \frac{1}{l} \Big[D_k(\pi_1^n, \pi_2^n, \ldots, \pi_K^n) - \widetilde{D}_k \Big]$
end for
Step 6: The algorithm stops when $\lambda_k^m, k \in \mathcal{K}$ converge, otherwise, set $m = m + 1$ and return to Step 2.

The algorithm can be summarized as follows. We use $\mathbf{V}_{k=1,2,\ldots,K}^n$ to represent the value vector at nth inner iteration and $\lambda_{k=1,2,\ldots,K}^m$ to represent Lagrange multiplier at mth outer iteration. The algorithm mainly consists of two parts: the outer loop and the inner loop. The outer loop updates the Lagrange multiplier of each user and the inner loop optimize the transmission policy of each user under fixed Lagrange multipliers. The outer loop index and inner loop index are m and n, respectively. It could be seen from Algorithm 2 that the interaction among all the secondary users is through the update of value vectors since $v_{i=1:K, i\neq k}^{n+1}(s)$ is a function of $\pi_k^n(s)$. The inner loop runs at every time slot and the inner loop iteration period equals to the time slot period.

In Step 1, we set the outer loop index m to be 0 and initialize the step size l, the value vector $\mathbf{V}_{k=1,2,\ldots,K}^0$ and Lagrange multipliers $\lambda_{k=1,2,\ldots,K}^0$. Step 3 is the inner loop where at each step we solve kth user controlled game and obtain the new optimal strategy for that user with the strategies of the remaining players fixed. Step 4 updates the Lagrange multipliers based on the discounted delay value of each user given the transmission policies $\{\pi_1^n, \pi_1^n, \ldots, \pi_K^n\}$. $\frac{1}{l}$ is the step size which satisfies the conditions for convergence of the Robbins-Monro algorithm. This the sequence of Lagrange multipliers $\{\lambda_1^m, \ldots, \lambda_K^m\}$ with $m = 0, 1, 2, \ldots$ converges in probability to $\{\lambda_1^*, \ldots, \lambda_K^*\}$ which satisfy the constrained problem defined in (14) [34], [35]. The algorithm terminates when certain accuracy of $\lambda_{k=1,2,\ldots,K}^m$ is obtained, otherwise, go to Step 2.

Since this is a constrained optimization problem, the optimal transmission policy is a randomization of two deterministic polices [33]. Use $\lambda_{k=1,2,\ldots,K}^*$ to represent the Lagrange multipliers obtained with the above algorithm. The randomization policy of each user can be written as:

$$\pi_k^*(s) = q_k \pi_k^*(s, \lambda_{k,1}) + (1 - q_k) \pi_k^*(s, \lambda_{k,2}), \quad (15)$$

where $0 \leq q_k \leq 1$ is the randomization factor and $\pi_k^*(s, \lambda_{k,1})$, $\pi_k^*(s, \lambda_{k,2})$ are the unconstrained optimal policies with Lagrange multipliers $\lambda_{k,1}$ and $\lambda_{k,2}$. Specifically, $\lambda_{k,1} = \lambda_k^* - \Delta$ and $\lambda_{k,2} = \lambda_k^* + \Delta$ for a perturbation parameter Δ. The randomization factor of the kth user q_k is calculated by:

$$q_k = \frac{\widetilde{D}_k - D_k(\lambda_{1,2}, \ldots, \lambda_{K,2})}{D_k(\lambda_{1,1}, \ldots, \lambda_{K,1}) - D_k(\lambda_{1,2}, \ldots, \lambda_{K,2})}. \quad (16)$$

The convergence proof of the inner loop of Algorithm 2 can be referred to [14, Chapter 6.3]. The intuition behind the proof is as follows: The value vector $\mathbf{V}_k^{(n)}$ ($k \in \mathcal{K}$) is nonincreasing on the iteration index n in the value iteration algorithm. There are only a finite number of strategies available for the optimal policy π_k^* for $k \in \mathcal{K}$. It can be concluded that the algorithm converges in a finite number of iterations. This value iterative algorithm obtains a Nash equilibrium solution to the constrained switching control Markovian game with general sum reward and general sum constraint.

Fig. 2 is an example on the performance of Algorithm 2. The system has 2 cognitive radio users, each user has a size 10 buffer, and the channel quality measurements are quantized into two different states, namely $\{1, 2\}$. It could be seen from Fig. 2 that the Nash equilibrium policy of user 1 is a randomized mixture of two pure policies.

IV. Truth Revealing Opportunistic Scheduling

The decentralized channel access algorithm adopt in Section III-A.1 is based on the opportunistic access scheme where each cognitive radio waits for a certain

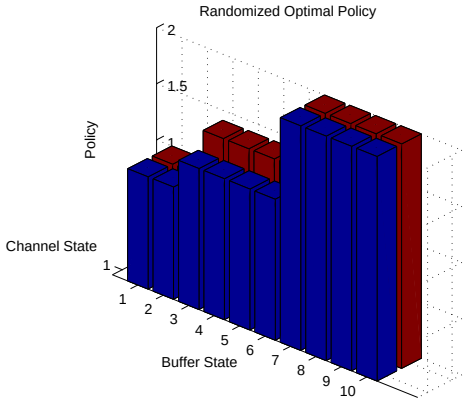

Figure 2. The Nash equilibrium transmission control policy obtained via value iterative optimization algorithm (Algorithm 2). A 2-user system is considered, and each user has a size 10 buffer.

time at the beginning of each time slot and the first user who accesses the channel can use the channel for that time slot. The system is equivalent of having a virtual central scheduler which decides which user is scheduled for transmission at each time slot. However, if different cognitive radios belong to different agents, the selfish cognitive radios may not reveal their true state information to the virtual central scheduler aiming to maximize their own payoffs. Mechanism design theory can be applied to prevent this from happening.

This section considers a multiple user cognitive radio system. We propose a truth revealing system access protocol based on opportunistic scheduling algorithm [26]. The proposed protocol provides better system performance over conventional approaches. By applying the mechanism design theory to the opportunistic scheduling, system users are eliminated from lying and the optimality of the overall system performance is ensured. The pricing mechanism we propose is based on VCG mechanism and it maintains the same desirable economic properties as that of the VCG mechanism.

A. Conventional Opportunistic Accessing Scheme

This subsection describes a conventional opportunistic scheduling algorithm in a K users cognitive radio system. Similar to the previous notations, b_k is used to indicate the buffer state of user k and $\mathbf{b} = \{b_1, b_2, \ldots, b_k\}$. $\hat{b}_k$ is used to represent the buffer state that the kth user reports to the virtual central scheduler. In a truth revealing system, each user reports the true state value and $\hat{b}_k = b_k$, $k \in \mathcal{K}$.

The channel quality of user k is denoted as h_k, specifically, h_k measures the signal to noise radio (SNR). The composition of the channel states of all the K users is denoted as $\mathbf{h} = \{h_1, h_2, \ldots, h_K\}$. Let the symbols per second transmission rate of user k be w_k and bits per symbol rate be m_k (different values of m_k leads to different modulation schemes). Assuming each user uses unit transmission power, the instantaneous throughput is:

$$\rho_k = w_k m_k \big(1 - p_e(h_k, m_k)\big)^{s_k}. \tag{17}$$

Here, s_k denotes the average packet size in bits. $p_e(h_k)$ denotes the BER, which is a function of the SNR and modulation mode of current user k. Assuming the system uses an uncoded M-ary quadrature modulation (QAM), $p_e(h_k)$ can be approximated as [33]:

$$p_e(h_k, m_k) = 0.2 \times \exp\big[\frac{-1.6 h_k}{2^{m_k} - 1}\big]. \tag{18}$$

Applying quantization to the instantaneous throughput ρ_k, we have $\rho_k \in \{0, 1, 2, \ldots, Q_\rho\}$ $(k = 1, 2, \ldots, K)$ with Q_ρ indicating the maximum throughput quantization level. Similarly, $\hat{\rho}_k$ is used to indicate the instantaneous throughput state that user k reports to the central scheduler and $\hat{\rho}_k \in \{0, 1, 2, \ldots, Q_\rho\}$.

For notation convenience, we use $\Theta_k = \{\rho_k, b_k\}$ to represent the true states of the kth user and $\hat{\theta}_k$ to represent the reported states of the kth user. Θ_{-k} denotes the true states of all the remaining $K-1$ users (excluding user k) and $\hat{\Theta}_{-k}$ denotes the corresponding reported values. We use $\Theta = \{\Theta_k, \Theta_{-k}\}$ to denote the true states of all the K users and $\hat{\Theta} = \{\hat{\Theta}_k, \hat{\Theta}_{-k}\}$ to denote the reported states.

The opportunistic access scheme is based on the reported buffer and throughput states. Define A as a *feasible* set, it is a subset of $\mathcal{K} = \{1, 2, \ldots, K\}$, and $A \subseteq \mathcal{K}$. A feasible set is a set which satisfies the system constraint. We specify the system constraint to be the overall transmission power in the system. Specifically, the overall transmission power in the system should be equal or less than the system power limit P. As we specified earlier in this subsection, each user uses an unit transmission power to transmit. Thus, the transmission power constraint on the system is equivalent to the constraint on the total number of users who are transmitting. That is, the number of users who are transmitting simultaneously should be less or equal to the system power limit P. The optimal feasible set A^* to the conventional opportunistic access scheme is a solution to the following optimization problem.

$$A^* = \arg\max \sum_{k \in A} \gamma_k \cdot U_k(\hat{\rho}_k, \hat{b}_k), \tag{19}$$

$$s.t. \quad |A| \leq P. \tag{20}$$

$|A|$ denotes the number of users in set A and $U_k(\hat{\rho}_k, \hat{b}_k)$ denotes the corresponding utility of user k given throughput $\hat{\rho}_k$ and buffer state $\hat{b}_k$. γ_k in (19) is the QoS parameter which depends on the user type. In a cognitive radio system, there are two types of users, primary user and secondary user. Thus, $\gamma_k = \{\gamma_p, \gamma_s\}$. If user k is a primary user, $\gamma_k = \gamma_p$, otherwise, $\gamma_k = \gamma_s$. Furthermore, by setting $\gamma_p \gg \gamma_s$, the network would not allow the transmission of secondary users with the presence of primary user. If there are multiple sets that optimize 19 subject to the constraint 20, A^* is randomly chosen from these set with equal probability. In the decentralized channel access algorithm specified in Section III-A.1, the utility function is specified as $U_k(\hat{\rho}_k, \hat{b}_k) = \hat{h}_k \hat{b}_k$ and the transmission power limit is specified to be $P = 1$ since only one user is allowed for transmission per time slot.

The conventional opportunistic accessing scheme assumes the state information received by the virtual central

scheduler is true, that is, $\hat{\rho}_k = \rho_k$ and $\hat{b}_k = b_k$. However, the conventional opportunistic algorithm may be challenged when the users become so sophisticated and are able to reconfigure themselves to make efficient usage of the local resources (e.g. manage their own reporting data to have the most efficient data transmission). It becomes increasing important to design a mechanism to optimize the overall system performance while ensuring the profit of each user.

B. The Pricing Mechanism

We are going to apply the VCG pricing mechanism to the opportunistic scheduling algorithm, the new mechanism enforces the truth revealing property of each user. The Nash equilibrium of such an algorithm is when each user reports true values.

1) The Pricing Mechanism: Different from the centralized conventional opportunistic scheduling algorithm, the proposed pricing mechanism is a distributed algorithm where each user tries to maximize his own utility function by choosing the reported state values. The buffer and throughput that user k chooses to report to the central scheduler is a solution of the following optimization problem:

$$\begin{aligned}\{\hat{\rho}_k, \hat{b}_k\} &:= \max_{\hat{\Theta}_k} v_k(\Theta_k, \hat{\Theta}_k, \hat{\Theta}_{-k}) \\ &= \max_{\hat{\rho}_k, \hat{b}_k} \alpha^{\gamma_k \rho_k b_k} \times \frac{\prod_{j \in A^*, j \neq k} \alpha^{\gamma_j \hat{\rho}_j \hat{b}_j}}{\prod_{j \in A'} \alpha^{\gamma_j \hat{\rho}_j \hat{b}_j}} I_{k \in A^*} + I_{k \notin A^*},\end{aligned}$$

where the sets A^* and A' are defined in the following ways:

$$\begin{aligned}A^* &:= \arg\max \sum_{j \in A} \gamma_j \hat{\rho}_j \hat{b}_j, \quad s.t. \quad |A| \leq P; \\ A' &:= \arg\max_{\hat{\rho}_k = 0} \sum_{j \in A, j \neq k} \gamma_j \hat{\rho}_j \hat{b}_j, \quad s.t. \quad |A|_{k \notin A} \leq P.\end{aligned}$$

In this optimization problem, α is a fixed constant for the system chosen to be $\alpha > 1$, $I_{\{.\}}$ is the indication function whose value is 1 when the condition is true, otherwise, it is 0.

$v_k(\Theta_k, \hat{\Theta}_k, \hat{\Theta}_{-k})$ is the utility function of user k, which is a function of the true states of kth user, the reported states of kth user and the reported states of all the remaining users. When a user is not scheduled for transmission, his utility function equals to 1, while when a user is scheduled for transmission, his utility function equals to the first part, that is:

$$v_k(\Theta_k, \hat{\Theta}_k, \hat{\Theta}_{-k}) = \begin{cases} \frac{\alpha^{\gamma_k \rho_k b_k} \cdot \prod_{j \in A^*, j \neq k} \alpha^{\gamma_j \hat{\rho}_j \hat{b}_j}}{\prod_{j \in A'} \alpha^{\gamma_j \hat{\rho}_j \hat{b}_j}} & \text{if } k \in A^* \\ 1. & \text{if } k \notin A^* \end{cases}$$

In the above equation, the first part $\alpha^{\gamma_k \rho_k b_k}$ is the gain of user k per unit of subcarrier with throughput state ρ_k and buffer state b_k. The second term $\frac{\prod_{j \in A^*, j \neq k} \alpha^{\gamma_j \hat{\rho}_j \hat{b}_j}}{\prod_{j \in A'} \alpha^{\gamma_j \hat{\rho}_j \hat{b}_j}}$ can be interpret as the number of unit of subcarrier that user k will be allocated which is a function of the state of the remaining users in the system. In other words, the inverse of the second term could be interpreted as the price that user k has to pay to the system if it is scheduled for transmission. Each user select $\{\hat{\rho}_k, \hat{b}_k\}$ to report to the central scheduler aiming to maximize its own utility function.

There is one condition necessary in order to achieve an efficient allocation scheme with selfish agents [36], [37]: if a user k (k=1,2,...,K) reports a false state values $\hat{\Theta}_k \neq \Theta_k$ results in the same value of the utility function as that of if it reports the true value, which is $v_k(\Theta_k, \hat{\Theta}_k, \hat{\Theta}_{-k}) = v_k(\Theta_k, \Theta_k, \hat{\Theta}_{-k}),\ \hat{\Theta}_k \neq \Theta_k$, then the user will choose to report the true values. We name this as the *truth preferred rule*. The interpretation of this rule is that when lying about the states does not bring any benefit to a user, a user would prefer telling the truth.

The pricing mechanism we propose above is based on the VCG mechanism, where we modified the conventional summation form of the utility function into a product form. Such pricing mechanism can be easily related to a practical 802.11 system and interpret the utility function in terms of real physical parameters.

2) Economic Properties of The Pricing Mechanism: The pricing mechanism we proposed above still maintains the same desirable economic properties as that of VCG mechanism, these properties are specified as follows [38], [39]:

1) The mechanism is incentive-compatible in ex-post Nash equilibrium. The best response strategy is to reveal the true state information $\hat{\Theta}_k = \Theta_k$ even after they have complete information about other users Θ_{-k}.
2) The mechanism is individually rational. A selfish agent will join the mechanism rather than choosing not to, because the value of the utility function is non-negative.
3) The mechanism is efficient. Since all the users will truthfully reveal their state information, the opportunistic scheduling algorithm carried out by the central scheduler will maximize the system performance.

The detailed proof of the properties is shown in [26].

Fig. 3 is a numerical example showing the performance of the pricing mechanism we designed. We simulate a 30 users cognitive radio system with each user has 5 buffer states and 10 throughput states. The transmission power constraint on the system is $P = 3$ which is equivalent to that the maximum number of users transmit simultaneously is 3.

In Fig. 3, the x-axis represents the number of iterations and y-axis represents the mean squared error (MSE) of the reported buffer states and throughput states. n is used to denote the iteration index. Defining $\hat{\rho}^n = \{\hat{\rho}_1^n, \ldots, \hat{\rho}_K^n\}$ and $\hat{b}^n = \{\hat{b}_1^n, \ldots, \hat{b}_K^n\}$, the MSE of the reported buffer states and throughput states can be written as:

$$MSE(\hat{\rho}^n) = \frac{1}{K} \times \sum_{k=1}^{K} (\hat{\rho}_k^n - \rho_k^n)^2; \tag{21}$$

$$MSE(\hat{b}^n) = \frac{1}{K} \times \sum_{k=1}^{K} (\hat{b}_k^n - b_k^n)^2. \tag{22}$$

In the figure, the solid curve and the dash curve show the MSE of the reported buffer states and throughput states,

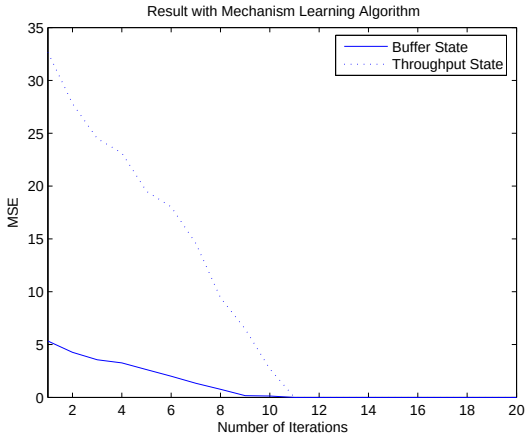

Figure 3. The MSE of the reported buffer states and throughput states after applying the pricing mechanism. The result is of a 30 users system with $L = 5$, $Q_\rho = 10$ and $P = 3$.

respectively. We can see from the figure that the MSE converge to 0 after 11 iterations, at which state, all the users in the system report truthfully and $\hat{\Theta} = \Theta$.

V. Correlated Equilibrium in Stochastic Dynamical Game

This section considers the user scheduling problem in a cognitive radio network, with the goal of obtaining the correlated equilibrium policy in a general-sum stochastic dynamical game setting. The correlated equilibrium policy of each user takes into account its transmission rate and transmission delay [29]. The dynamic of the channel quality and buffer state is formulated as a Markov chain and the correlated equilibrium of such a stochastic game is defined based on the Q-functions. The existing non-regret learning algorithm [40], [41] can be easily applied to obtain such correlated equilibrium policies. The correlated equilibrium policy of each user is independent from each other, which enables the decentralized feature of the system. We propose two algorithms, namely, iterative correlated equilibrium algorithm and correlated Q-learning algorithm. Compared to Algorithm 3, Algorithm 4 combines the Q-learning with the non-regret learning which offers a decentralize feature and is more implementable. Distributed learning is an important feature of a cognitive radio system since it does not require central control and each user is able to learn the correlated equilibrium policy independently. Both of the algorithms proposed in this section are novel and have not been applied in cognitive radio networks before.

A. System Model and Transmission Control Utility Function

Different from the system model used in Section III, the system model we use here allows more than one user to access to the spectrum at each time slot.

Similarly, we assume there are K users in the system and t is used as the time slot index. h_k denotes the channel quality of user k and the composition of the channel states is $\mathbf{h} = \{h_1, \ldots, h_K\}$. The channel state over time is formulated as a Markov chain with transition probability denoted as $\mathbb{P}(h^{t+1}|h^t)$. The buffer state of user k is denoted as b_k and $\mathbf{b} = \{b_1, \ldots, b_K\}$. The incoming traffic of user k is f_k and $\mathbf{f} = \{f_1, \ldots, f_K\}$. The system state of user k is $\mathbf{s}_k = \{h_k, b_k\}$ and the composition of the states of all the users is denoted as $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$.

The system is designed to perform effective user scheduling. At each time slot, each user k ($k \in \mathcal{K}$) chooses an optimal action a_k from the action set $\mathcal{A}_k = \{0, 1\}$, where 0 represents *no transmission* and 1 represents *transmission*. The joint action of all the users is denoted as $\mathbf{a} = \{a_1, \ldots, a_K\}$, which is an element of the joint action space, $\mathbf{a} \in \mathcal{A}$. Using standard game theoretic notation, we can write $\mathbf{a} = \{a_k, \mathbf{a}_{-k}\}$ with $\mathbf{a}_{-k}$ standing for the joint actions of other users excluding user k.

The transition probability between the current composite state $\mathbf{s} = [\mathbf{h}, \mathbf{b}]$ and the next state $\mathbf{s}' = [\mathbf{h}', \mathbf{b}']$ is a function of $\mathbf{a}$, which can be expressed as:

$$\mathbb{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) = \prod_{k=1}^{K} \mathbb{P}(h'_k|h_k) \cdot \prod_{k=1}^{K} \mathbb{P}(b'_k|b_k, a_k), \tag{23}$$

where the expression of $\mathbb{P}(b'_k|b_k, a_k)$ is given in (11).

The utility function of user k in state $\mathbf{s}$ given action $\mathbf{a}$ is denoted as $u_k(\mathbf{s}, \mathbf{a})$, and it is specified as:

$$u_k(\mathbf{s}, \mathbf{a}) = \gamma_k \Big[\log\Big(1 + \frac{|h_k|^2 a_k}{\sigma_k^2(n) + \sum_{j=1, j\neq k}^{K} |\tau_{k,j} h_j|^2 a_j}\Big) - \frac{\beta_k b_k}{\frac{1}{K}\sum_{j=1}^{K} b_j}\Big],$$

where γ_k is the QoS parameter and $\gamma_k \in \{\gamma_p, \gamma_s\}$. If user k is a primary user, $\gamma_k = \gamma_p \geq 0$, while if user k is a secondary user, $\gamma_k = \gamma_s \geq 0$. $\gamma_p \gg \gamma_s$ so that the primary users have the priority over secondary users. The first half of the utility function $\log(\ldots)$ represents the maximum achievable transmission rate of user k and it is similar to that in (3), while the second half includes the transmission delay and $\beta_k \geq 0$ denotes the weighting parameter between the transmission rate and the delay. $\tau_{k,j}$ is the channel coefficient determined by the location of users and $\tau_{k,j} \cdot h_j$ measures the channel quality between the jth transmitter and the kth receiver.

B. Correlated Equilibrium in a Markovian Dynamic Game

1) A Review of Correlated Equilibrium in Static Games: To motivate the correlated equilibrium for a dynamic game, we first review the correlated equilibrium for a static game. In a K-user static game setup, each user $k \in \mathcal{K}$ aims to devise a rule for selecting an action a_k from its action set $\mathcal{A}_k$ so as to maximize the expected value of its utility function $u_k(\{a_1, a_2, \ldots, a_K\})$. Since each user is only able to adapt its own action, the optimal action policy depends on the rational consideration of the policies from other users. The adopted solution for such a problem is called an *equilibrium*. Many equilibrium concepts have been developed and the most common one

is the *Nash equilibrium*. In this section, we focus on an important generalization of the Nash equilibrium, known as *correlated equilibrium* [27], [28], which is defined as follows.

Definition 1: Define a joint policy $\boldsymbol{\pi}$ to be a probability distribution on the joint action space $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \mathcal{A}_K$. With the given actions of other users $\mathbf{a}_{-k}$, the policy $\boldsymbol{\pi}$ is a correlated equilibrium, if for every $a_k \in \mathcal{A}_k$, $(k = 1, 2, \ldots, K)$ such that $\pi(\mathbf{a}_{-k}, a_k) > 0$, and any alternative policy $a'_k \in \mathcal{A}_k$, it holds that,

$$\sum_{\mathbf{a}_{-k} \in \mathcal{A}_{-k}} \pi(\mathbf{a}_{-k}, a_k) u_k(\{\mathbf{a}_{-k}, a_k\}) \geq \sum_{\mathbf{a}_{-k} \in \mathcal{A}_{-k}} \pi(\mathbf{a}_{-k}, a_k) u_k(\{\mathbf{a}_{-k}, a'_k\}). \quad (24)$$

One intuitive interpretation of correlated equilibrium is that $\boldsymbol{\pi}$ provides the K users a strategy recommendation from the trusted third-party. The implicit assumption is that the $K-1$ other users follow this recommendation, and user k ask itself whether it is of its best interest to follow the recommendation as well. The equilibrium condition states that there is no derivation rule that could award user k a better expected utility. We also notice that the set of Nash equilibria is obtained by intersecting the polyhedron described above with the additional constraint $\pi(\mathbf{a}_{-k}|a_k) = \pi(\mathbf{a}_{-k}|a'_k)$ for all $a_k, a'_k \in \mathcal{A}_k$ and $\mathbf{a}_{-k} \in \mathcal{A}_{-k}$. Any Nash equilibrium can be represented as a correlated equilibrium when the users can generate their recommendations independently.

One advantage of using correlated equilibrium is that it permits coordination among users, generally through observation of a common signal, which leads to improved performance over a Nash equilibrium [28].

2) The Definition of Correlated Equilibrium in Markovian Games: The stationary policy of the system is only a function of the state instead of the time. We use $\boldsymbol{\pi}_{\mathbf{s}}$ to denote the transmission policy vector of all the users in state $\mathbf{s}$. The policies of all the users over all the states can be denoted as $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_{\mathbf{s}} \in \boldsymbol{\pi}$. Any $\boldsymbol{\pi}_{\mathbf{s}}$ can be decomposed into marginals $(\boldsymbol{\pi}_{\mathbf{s},k}, \boldsymbol{\pi}_{\mathbf{s},-k})$ for any k, where $\boldsymbol{\pi}_{\mathbf{s},k}$ is the marginal probability distribution of the strategy of user k, while $\boldsymbol{\pi}_{\mathbf{s},-k}$ is the marginal distribution of all users but k. Furthermore, each entry of $\boldsymbol{\pi}_{\mathbf{s}}$ is denoted as $\pi_{\mathbf{s}}(\mathbf{a}_{-k}, a_k)$ $(\forall \mathbf{a}_{-k} \in \mathcal{A}_{-k}, \forall a_k \in \mathcal{A}_k)$, which represents the joint probability of taking action $\mathbf{a}_{-k}$ and a_k in state $\mathbf{s}$. The infinite horizon expected total discounted value of user k with initial state $\mathbf{s}^0 = \mathbf{s}$ under transmission policy $\boldsymbol{\pi}$ is:

$$V_k^{\boldsymbol{\pi}}(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty} \beta^t u_k(\mathbf{s}^t, \mathbf{a}^t) \big| \mathbf{s}^0 = \mathbf{s}\right], \quad (25)$$

where $0 < \beta < 1$ is the economic discount factor chosen by the system.

Based on the above definitions, each user k, $k \in \mathcal{K}$ updates its value vector from time t to $t+1$ in the following way,

$$V_k^t(\mathbf{s}) = \max_{\boldsymbol{\pi}_{\mathbf{s},k}} \sum_{\mathbf{a} \in \mathcal{A}} \pi_{\mathbf{s}}(\mathbf{a}) \big[u_k(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathbb{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) V_k^{t-1}(\mathbf{s}')\big]$$

.

The above optimization problem defines how the strategy of each user $\boldsymbol{\pi}_{\mathbf{s},k}$ is updated at each iteration. Denote $\boldsymbol{\pi}^*$ as the solution to the problem when all the users do the value vector update simultaneously. The value vector of user k converges to $\mathbf{V}_k^{\boldsymbol{\pi}^*}$ under policy $\boldsymbol{\pi}^*$ and $V_k^{\boldsymbol{\pi}^*}(\mathbf{s}) \in \mathbf{V}_k^{\boldsymbol{\pi}^*}$

We are now ready to define the Q-function. Q-function $Q_k^{\boldsymbol{\pi}^*}(\mathbf{s}, \mathbf{a})$ of user k is the total discounted reward of taking action $\mathbf{a}$ in state $\mathbf{s}$ and then following $\boldsymbol{\pi}^*$ thereafter, and it is defined as:

$$Q_k^{\boldsymbol{\pi}^*}(\mathbf{s}, \mathbf{a}) = u_k(\mathbf{s}, \mathbf{a}) + \gamma \cdot \sum_{s' \in \mathcal{S}} \mathbb{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) V_k^{\boldsymbol{\pi}^*}(\mathbf{s}). \quad (26)$$

Therefore, we can have the following relationship between the value vector and Q-function [42].

$$V_k^{\boldsymbol{\pi}^*}(\mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} \pi_{\mathbf{s}}^*(\mathbf{a}) \cdot Q_k^{\boldsymbol{\pi}^*}(\mathbf{s}, \mathbf{a}). \quad (27)$$

The correlated equilibrium in a Markovian dynamic game is then defined as follows.

Definition 2: The stationary policy $\boldsymbol{\pi}^*$ is a correlated equilibrium for the Markovian game described above if $\forall k \in \mathcal{K}$, $\forall \mathbf{s} \in \mathcal{S}$ and $\forall a_k, a'_k \in \mathcal{A}_k$,

$$\sum_{\mathbf{a}_{-k} \in \mathcal{A}_{-k}} \pi_{\mathbf{s}}^*(\mathbf{a}_{-k}, a_k) \cdot Q_k^{\boldsymbol{\pi}^*}(\mathbf{s}, \{\mathbf{a}_{-k}, a_k\}) \geq \sum_{\mathbf{a}_{-k} \in \mathcal{A}_{-k}} \pi_{\mathbf{s}}^*(\mathbf{a}_{-k}, a_k) \cdot Q_k^{\boldsymbol{\pi}^*}(\mathbf{s}, \{\mathbf{a}_{-k}, a'_k\}). \quad (28)$$

C. Distributed Correlated Equilibrium Algorithm

In this section, we first introduce an iterative correlated equilibrium algorithm which uses the non-regret learning algorithm to obtain the correlated equilibrium in the Markovian game. Then, we propose a correlated Q-learning algorithm which combines the non-regret algorithm and Q-learning so as to calculate the correlated equilibrium policy in a distributed way. Both of the algorithms proposed are novel and have not been applied in cognitive radio networks.

First of all, the existence of a correlated equilibrium in any general-sum Markov game can be shown by the following result from [42].

Theorem 3: [42] Every general-sum Markov game has a stationary correlated equilibrium policy.

This result enables the application of the following algorithms.

1) Iterative Correlated Equilibrium Algorithm: Non-regret learning algorithm was first proposed by Hart and Mas-Colell in [40], where they formulate the probabilities the players depart from their current plays to be proportional to the measure of regret for not having used other strategies in the past. The non-regret learning algorithm has been proved to be able to converge to the set of correlated equilibria of the game with probability one. An iterative correlated equilibrium algorithm has been designed to obtain the correlated equilibrium in the Markovian game system model in this section. By using this algorithm, each user in the system does not need to keep the information of the Q-functions of other users and the problem has been solved in a distributed way to some extent.

The iterative correlated equilibrium algorithm is summarized in the table for Algorithm 3. In the algorithm, we

Algorithm 3 Iterative Correlated Equilibrium Algorithm

Outer Loop

Step 1 Initialization:
$n = 0$ and initialize $Q_k^0(\mathbf{s}, \mathbf{a})$ for $\forall k \in \mathcal{K}$, $\forall \mathbf{s} \in \mathcal{X}$ and $\forall \mathbf{a} \in \mathcal{A}$.

Step 2 For $\forall k \in \mathcal{K}$ and $\forall \mathbf{s} \in \mathcal{S}$ do the following:

Inner Loop: Non-regret Learning

Initialize inner loop index $l = 0$, initial regret matrix $\mathbf{R}_k^0$ and action a_k^0.

Action Update: let $i = a_k^l$ and let $a_k^{l+1} = j$ with the probability

$P_k^{l+1}(j) = \frac{\max\{R_k^n(i,j),0\}}{\mu}$ if $j \neq i$;

$P_k^{l+1}(j) = 1 - \frac{\sum_{m \neq i} \max\{R_k^n(i,m),0\}}{\mu}$ if $j = i$.

Regret matrix update:

$H_k^{i,j}(\mathbf{a}^{l+1}) = I_{(\mathbf{a}_k^{l+1}=i)} \cdot \left[Q_k(j, \mathbf{a}_{-k}^{l+1}) - Q_k(i, \mathbf{a}_{-k}^{l+1})\right]$;

$\mathbf{R}_k^{l+1} = \mathbf{R}_k^l + \frac{1}{l+1}\left(\mathbf{H}_k(\mathbf{a}^{l+1}) - \mathbf{R}_k^l\right)$.

Exit inner loop when $\mathbf{P}_k^l$ converges and set $\boldsymbol{\pi}^n = \mathbf{P}^{l+1}$; else set $l = l + 1$ and return back to **Inner Loop**.

Step 3 Update value vectors:
$V_k^n(\mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} \pi_{\mathbf{s}}^n(\mathbf{a}) \cdot Q_k^n(\mathbf{s}, \mathbf{a})$, for $\forall k \in \mathcal{K}$, $\forall \mathbf{s} \in \mathcal{S}$.

Step 4 Update Q-function values:
$Q_k^{n+1}(\mathbf{s}, \mathbf{a}) = u_k(\mathbf{s}, \mathbf{a}) + \gamma \cdot \sum_{\mathbf{s}' \in \mathcal{S}} \mathbb{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) V_k^n(\mathbf{s}')$ for $\forall k \in \mathcal{K}$, $\forall \mathbf{s} \in \mathcal{X}$ and $\mathbf{a} \in \mathcal{A}$.

The iteration terminates when the values of the parameters $\boldsymbol{\pi}^n$ converge; else set $n = n + 1$ and return back to Step 2.

first initialize the outer loop iteration index $n = 0$ and the initial Q-function values in Step 1. Then, we implement the non-regret learning algorithm for each k and $\mathbf{s}$ to learn the new policy. Based on the new policy, we update the value vectors V_k^n and the Q-function values accordingly.

The inner loop of the algorithm is non-regret learning, where l denotes the inner loop iteration. $\mathbf{R}_k^l$ denotes the regret matrix of user k at the lth iteration. Each entry $\mathbf{R}_k^l(i,j)$ indicates the regret value from action i to action j. $\mathbf{P}^l$ denotes the policy of all the users at the lth iteration, while $\mathbf{P}_k^l$ denotes that of user k and $\mathbf{P}^l = \{\mathbf{P}_1^l, \ldots, \mathbf{P}_K^l\}$. Each of $\mathbf{P}_k^l$ entry $P_k^l(j)$ represents the probability of taking action j at iteration l. The new policy of each user is to pick an action according to the action probability $\mathbf{P}^l$. The constant μ is a normalization factor chosen to be $\mu > \sum_{m \neq i} \max\{R_k^n(i,m), 0\}$, ensuring the update of P_k^{l+1} yields valid probabilities. It can be viewed as an inertia parameter, i.e., a higher μ yields lower probability of switching. The term $I_{f(x)}$ is the indicator function which equals to 1 when $f(\mathbf{s})$ is true and 0 otherwise.

The non-regret learning algorithm applied here is based on [41] and its convergence has been proved therein. We observe that in Algorithm 3, each user does not need to know the utility functions and policies of other users which makes the algorithm distributed. Note that in the inner loops, users play the repeated games and the process can be implemented online while the outer loop relies on the knowledge of the probability transition matrix and hence is an offline update. It motivates us to propose the next algorithm, which lifts the requirement of the transition matrix and is implementable online.

2) Correlated Q-learning Algorithm: Q-learning is a type of reinforcement learning technique which learns the Q-function values without the knowledge of the system model. In this section, we are going to propose a correlated Q-learning to calculate the correlated equilibrium policy for the Markovian game. Compared to Algorithm 3, this algorithm does not require the information of the system state transition probability matrix which makes it more practical and implementable.

Algorithm 4 Correlated Q-learning Algorithm

Step 1 Initialize $n = 0$ and $Q_k^0(\mathbf{s}, \mathbf{a})$ for $\forall k \in \mathcal{K}$, $\forall \mathbf{s} \in \mathcal{S}$ and $\forall \mathbf{a} \in \mathcal{A}$.

Step 2 Update $\boldsymbol{\pi}^n$ by non-regret learning.

Step 3 Update the Q-function values:
For $\forall \mathbf{s} \in \mathcal{S}$ and $\forall \mathbf{a} \in \mathcal{A}$:
$Q_k^{n+1}(\mathbf{s}, \mathbf{a}) = (1 - \alpha^n) \cdot Q_k^n(\mathbf{s}, \mathbf{a}) + \alpha^n[u_k(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \sum_{\mathbf{a}' \in \mathcal{A}} \pi_k^n(\mathbf{s}', \mathbf{a}') \cdot Q_k^n(\mathbf{s}', \mathbf{a}')]$
The algorithm terminates when the values of the parameters $\boldsymbol{\pi}^n$ converge; else set $n = n + 1$ and return back to Step 2.

The correlated Q-learning algorithm (Algorithm 4) can be summarized as follows. System initializes the iteration index n and the initial Q-function values in Step 1. Step 2 calculates the correlated equilibrium policy $\boldsymbol{\pi}^n$ with given Q-function values by using the non-regret learning similar to that in Algorithm 3. The system runs the non-regret learning offline. Based on the updated transmission policies, each user updates its Q-function values by using Q-learning. Parameter α^n is the learning rate sequence which is chosen such that $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$. The algorithm terminates when the Q-function converges. The optimal correlated equilibrium and the value function can thus be obtained from the final Q-function.

In both Algorithm 3 and Algorithm 4, at each inner loop iteration, they require the knowledge of the Q-function they calculate at each outer loop. Based on this information, each user performs one table lookup in its Q-function to calculate the Q-utility given the current readings of the state $\mathbf{s}$ and the actions $\mathbf{a}^l$. Two additions and two multiplications update the regret value; and one random number, one multiplication and one comparison suffice to calculate the next action. In the outer loop, an update is done for each state $\mathbf{s}$ and action profile $\mathbf{a}$. The computational complexity for the inner loop is small and hence suitable for implementation. However, the complexity of outer loop depends on the number of states and the size of action profiles. The size of the action set is exponential in the number of users. However, it is possible to regard the actions of other users as one virtual action so as to reduce the dimension of the set of action profiles. This dimension reduction technique enables a dramatic reduction in computational complexity, making an efficient process the update of the Q-functions in the algorithms.

VI. Open Issues and Conclusions

This paper adopt game theoretic approach to solve various problems in cognitive radio systems, such as power allocation, rate adaptation and accessing control. The main game theoretic concepts we used include Nash

equilibria in static games and switching control games, correlated equilibria in stochastic Markovian games and mechanism design.

Despite the recent popularity of game theory in wireless communications and cognitive radio networks, the potential for further research is vast. Here are some open issues related to this paper:

1) The dimension of the system state in a stochastic game is exponential to the number of users of the system, thus, the convergence speed of Algorithm 2,3,4 suffers from the "curse of dimensionality" when the system has large number of users. How to efficiently reduce the state space is yet an issue to solve.
2) Section III only focuses on a special type of game, namely switching control Markovian game. The study of a more general type of Markovian game is still a very interesting area. For example, prove the existence of the Nash equilibrium in a general Markovian game under certain conditions or propose efficient algorithms to obtain a Nash equilibrium policy in a general Markovian game.
3) The pricing mechanism used in Section IV is based on the well-known VCG mechanism. The development of other pricing mechanisms or reputation based mechanism can be one direction of the future work.
4) More analytical results of the correlated equilibrium in a general-sum stochastic game (e.g., the structural result on the correlated equilibrium policy) can be explored.

REFERENCES

[1] J. V. Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[2] R. Gibbons, *Game Theory for Applied Economists*. Princeton University Press, 1992.

[3] S. Hart and A. Neyman, *Game and Economic Theory*. Ann Arbor: The University of Michigan Press, 1995.

[4] A. B. MacKenzie and L. A. DaSilva, *Game Theory for Wireless Engineers*. San Rafael, California: Morgan and Claypool Publishers, 2006.

[5] S. G. Glisic, *Advanced Wireless Communications: 4G Cognitive and Cooperative Broadband Technology*. Wiley, 2007.

[6] F. C. Commussion, *Spectrum Policy Task Force*. Rep. ET Docket no. 02-135, November, 2007.

[7] J. Mitola III, "Cognitive radio for flexible mobile multimedia communications," in *Proceedings of IEEE International Workshop on Mobile Multimedia Communications*, 1999, pp. 3–10.

[8] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, February 2005.

[9] V. K. Bhargava and E. Hossain, *Cognitive Wireless Communication Networks*. New York: Springer-Verlag, 2007.

[10] J. Mitola III and G. Q. Maguire Jr., "Cognitive radio: Making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, August 1999.

[11] W. Yu, G. Ginis, and J. M. Cioffi, "Distributed multiuser power control for digital subscriber lines," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 5, pp. 1105–1115, June 2002.

[12] G. Scutari, D. Palomar, and S. Barbarossa, "Asynchronous iterative water-filling for Gaussian frequency-selective interference channels," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 2868–2878, July 2008.

[13] S. R. Mohan and T. E. S. Raghavan, "An algorithm for discounted switching control stochastic games ," *OR Spektrum*, vol. 55, no. 10, pp. 5069–5083, October 2007.

[14] J. A. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York: Springer-Verlag, 1997.

[15] O. J. Vrieze, S. H. Tijs, T. E. S. Raghavan, and J. A. Filar, "A finite algorithm for the switching control stochastic game," *OR Spektrum*, vol. 5, no. 1, pp. 15–24, March 1983.

[16] J. W. Huang and V. Krishnamurthy, "Transmission control in cognitive radio systems with latency constraints as a switching control dynamic game," in *Proceedings of IEEE CDC*, December 2008, pp. 3823–3828.

[17] C. Peng, H. Zheng, and B. Y. Zhao, "Utilization and fairness in spectrum assignment for opportunistic spectrum access," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 555–576, August 2006.

[18] W. Wang, X. Liu, and H. Xiao, "Exploring opportunistic spectrum availability in wireless communication networks," in *Proceedings of IEEE VTC*, September 2005.

[19] Y. Chen, Q. Zhao, and A. Swami, "Joint design and seperation principle for opportunistic spectrum access," in *Proceedings of IEEE ACSSC*, October/November 2006, pp. 696–700.

[20] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589–600, April 2007.

[21] R. Urgaonkar and M. J. Neely, "Opportunistic scheduling with reliability guarantees in cognitive radio networks," in *Proceedings of IEEE INFOCOM*, April 2008, pp. 1301–1309.

[22] A. MasColell, M. Whinston, and J. R. Green, *Microeconomic Theory*. Oxford: Oxford University Press, 1995.

[23] W. Vickrey, "Counterspeculation auctions and competitive sealed tenders," *Journal of Finance*, vol. 16, no. 1, pp. 8–37, 1961.

[24] E. H. Clarke, "Multipart pricing of public goods," *Public Choice*, vol. 2, pp. 19–33, 1971.

[25] T. Groves, "Incentives in team," *Econometrica*, vol. 41, no. 4, pp. 617–631, 1973.

[26] J. W. Huang and V. Krishnamurthy, "Truth revealing opportunistic scheduling in cognitive radio systems," in *The 10th IEEE International Workshop in Signal Processing Advances in Wireless Communications*, June 2009.

[27] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, pp. 67–96, March 1974.

[28] ——, "Correlated equilibrium as an expression of Bayesian rationality," *Econometrica*, vol. 55, no. 1, pp. 1–18, 1987.

[29] J. W. Huang, Q. Zhu, V. Krishnamurthy, and T. Basar, "Distributed Correlated Q-learning for Dynamic Transmission Control of Sensor Networks *Submitted to IEEE Globecom* ," 2009.

[30] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1989.

[31] *IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE std 802.16-2004, 2004.

[32] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming users in HSDPA multimedia systems," *IEEE Transactions on Multimedia Systems*, vol. 8, no. 4, pp. 844–855, August 2006.
[33] D. Djonin and V. Krishnamurthy, "MIMO transmission control in fading channels - a constrained Markov decision process formulation with monotone randomized policies ," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 5069–5083, October 2007.
[34] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. Wiley-Interscience, 2003.
[35] V. Krishnamurthy and G. G. Yin, "Recursive algorithms for estimation of hidden Markov models and autoregressive models with Markov regime," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 458–476, February 2002.
[36] R. Mirrlees, "An exploration in the theory of optimum income taxation," *Review of Economic Studies*, vol. 38, pp. 175–208, 1971.
[37] P. Dasgupta and E. Maskin, "Efficient auctions," *Quarterly Journal of Economics*, vol. 115, pp. 341–388, 2000.
[38] R. K. Dash, D. C. Parkes, and N. R. Jennings, "Computational mechanism design: A call to arms," *IEEE Intelligent Systems*, vol. 18, no. 6, pp. 40–47, November/December 2003.
[39] R. K. Dash, A. Rogers, N. R. Jennings, S. Reece, and S. Roberts, "Constrained bandwidth allocation in multi-sensor information fusion: a mechanism design approach," in *Proceedings of IEEE Information Fusion Conference*, July 2005, pp. 1185–1192.
[40] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, September 2000.
[41] V. Krishnamurthy, M. Maskery, and G. Yin, "Decentralized adaptive filtering algorithm for sensor activation in an unattended ground sensor network," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 6086–6101, December 2008.
[42] A. Greenwald and M. Zinkevich, "A direct proof of the existence of correlated equilibrium policies in general-sum Markov games," *Brown CS: Tech Report CS-05-07*, July 2005.

Jane Wei Huang (S'05) received her bachelor's degree from Zhejiang University, China in 2005, and M.Phil. from Hong Kong University of Science and Technology, Hong Kong in 2007. She is currently a Ph.D. student in the University of British Columbia.

Her research interests include game theory, Markov decision process, cognitive radio, and sensor networks.

Vikram Krishnamurthy (S'90-M'91-SM'99-F'05) was born in 1966. He received his bachelor's degree from the University of Auckland, New Zealand in 1988, and Ph.D. from the Australian National University, Canberra, in 1992. He is currently a professor and holds the Canada Research Chair at the Department of Electrical Engineering, University of British Columbia, Vancouver, Canada. Prior to 2002, he was a chaired professor at the Department of Electrical and Electronic Engineering, University of Melbourne, Australia, where he also served as deputy head of department. His current research interests include computational game theory, stochastic dynamical systems for modeling of biological ion channels and stochastic optimization and scheduling.

Dr. Krishnamurthy has served as associate editor for several journals including *IEEE Transactions Automatic Control*, *IEEE Transactions on Signal Processing*, *IEEE Transactions Aerospace and Electronic Systems*, *IEEE Transactions Nanobioscience*, and *Systems and Control Letters*. From 2009-2010 he serves as distinguished lecturer for the IEEE signal processing society. From 2010, he serves as editor in chief of *IEEE Journal of Selected Topics in Signal Processing*.