11

Game Theoretic Learning and Pricing for Dynamic Spectrum Access in Cognitive Radio

Michael Maskery[1], Vikram Krishnamurthy[1], and Qing Zhao[2]

[1] University of British Columbia, Canada
{mikem,vikramk}@ece.ubc.ca
[2] University of California at Davis, USA
qzhao@ece.ucdavis.edu

11.1 Introduction

This chapter deals with game theoretic methods for dynamic spectrum access in cognitive radio systems. Cognitive radio systems need to employ dynamic spectrum access methods to efficiently share radio spectrum with other cognitive radios while avoiding interference with legacy systems. Due to the inherent decentralized nature of cognitive radio, dynamic spectrum access strategies need to be decentralized. To address this, we formulate a model in which cognitive radios are players competing for spectrum resources in a game theoretic setting. The players need to access channels in a dynamic and uncertain environment to satisfy demand while respecting system-imposed sharing incentives.

The reader is undoubtedly familiar with the term *Nash equilibrium* in non-cooperative games. In this paper we use a more general equilibrium concept called *correlated equilibrium*. The concept of correlated equilibria in game theory was introduced by Aumann [1,2].[3] Correlated equilibria are easier to characterize and more natural to decentralized adaptive algorithms such as those considered here.

The problem of non-cooperative radio resource allocation is addressed elsewhere in [2–4] from a non-game theoretic perspective, and in [5,6] from a game theoretic one. Of these, [7] is auction-based and does not fit in our framework. Reference [5] is very similar to our approach, even employing similar learning-based ideas, but for a fundamentally different scenario.

Before presenting our main results, including our game theoretic dynamic spectrum access model and adaptive learning algorithm, we begin by reviewing the main ideas in dynamic spectrum access and game theory.

[3] Aumann was awarded the 2005 Nobel Prize in Economics. The Nobel Prize press release in October 2005 reads: “Aumann also introduced a new equilibrium concept, correlated equilibrium, which is weaker than Nash equilibrium, the solution concept developed by John Nash, an Economics Laureate in 1994. Correlated equilibrium can explain why it may be advantageous for negotiating parties to allow an impartial mediator to speak to the parties either jointly or separately, and in some instances give them different information”.

11.1.1 Brief Overview of Dynamic Spectrum Access

The proliferation of a wide range of wireless devices and their applications has resulted in an overly crowded radio spectrum; almost all usable frequencies have already been assigned. This makes one pessimistic about the feasibility of integrating emerging wireless services such as large-scale sensor networks into the existing communication infrastructure.

In contrast to the apparent spectrum scarcity is the pervasiveness of spectrum opportunity. Extensive measurements indicate that, at any given time and location, a large portion of licensed spectrum lies unused. For example, over 62% white space exists in the spectrum under 3 GHz [8]. This paradox between the overly crowded spectrum and the pervasiveness of idle frequency bands in both time and space indicates that spectrum shortage results from the spectrum management policy rather than the physical scarcity of usable frequencies.

The underutilization of spectrum has stimulated a flurry of exciting activities in search for dynamic spectrum access strategies for improved efficiency. Approaches envisioned for dynamic spectrum access fall under three general models: dynamic exclusive use, open sharing and hierarchical access.

The dynamic exclusive use model aims to introduce flexibility to the current command-and-control spectrum regulation policy while maintaining the spectrum licensees' right of exclusive use. Specific approaches include spectrum property rights [9] and dynamic spectrum allotment brought forth by the European DRiVE project [10]. The open sharing model, also referred to as the spectrum commons model [11], draws support from the phenomenal success of wireless services operating in the unlicensed ISM band. It employs open sharing among peer users as the basis for spectrum management. The hierarchical access model can be considered as a hybrid of the above two. The basic idea is to open licensed spectrum to secondary users and limit the interference perceived by primary users (licensees). One approach to spectrum sharing between primary and secondary users is spectrum overlay, which was first envisioned by Mitola [12] under the term "spectrum pooling" and then investigated by the DARPA XG program [13] under the term "opportunistic spectrum access". Another approach is spectrum underlay enabled by the technology of ultra wide band. A more detailed taxonomy of dynamic spectrum access can be found in Chapter 10.

In this chapter, we focus on the overlay approach to dynamic spectrum access. This approach directly targets at idle frequency bands in both time and space by allowing secondary users to identify and exploit instantaneous and local spectrum availability without causing unacceptable interference to primary users.

While conceptually simple, spectrum overlay presents technical challenges across the entire networking protocol stack. Basic components of spectrum overlay include spectrum opportunity identification and spectrum opportunity exploitation. The opportunity identification module is responsible for accurately identifying and intelligently tracking idle frequency bands that are dynamic in both time and space. The opportunity exploitation module takes input from the opportunity identification module and decides whether and how a transmission should take place. The overall design

objective of OSA is to provide sufficient benefit to secondary users while protecting spectrum licensees from interference. We present below a brief overview of major technical issues and recent development in each module. A more detailed survey of technical and regulatory issues in spectrum overlay can be found in [14].

11.1.1.1 Spectrum Opportunity Identification

As shown in [15], in a general network setting with spatially varying primary user activity, spectrum opportunity detection needs to be performed jointly by a secondary transmitter and its intended receiver. Specifically, a channel is an opportunity when no primary users in the neighborhood of the secondary transmitter are *receiving* over this channel and no primary users in the neighborhood of the secondary receiver are *transmitting* over this channel. Spectrum opportunity detection thus has both signal processing and networking aspects. The problem can, however, be reduced to a classic signal processing problem: detecting the presence of primary users' signals [15]. Based on the secondary user's knowledge of the signal characteristics of primary users, three traditional signal detection techniques can be employed: matched filter, energy detector (radiometer) and cyclostationary feature detector [16]. A matched filter performs coherent detection. It requires the least number of samples to achieve a given detection power but relies on synchronization and a priori knowledge of primary users' signaling. On the other hand, the non-coherent energy detector requires only basic information of primary users' signal characteristics but suffers from long detection time. Cyclostationary feature detector can improve the performance over an energy detector by exploiting an inherent periodicity in the primary users' signal. Details of this type of detectors can be found in [17]. While classic signal detection techniques exist in the literature, detecting primary transmitters in a dynamic wireless environment with noise uncertainty, shadowing, and fading is a challenging problem that has attracted much research attention [18].

Due to hardware limitation and energy cost associated with spectrum monitoring, a secondary user may not be able to sense all channels in the spectrum simultaneously. In this case, the secondary user needs a sensing strategy for intelligent channel selection to track the time varying spectrum opportunities. The purpose of the sensing strategy is twofold: catch a spectrum opportunity for immediate access and obtain statistical information on spectrum occupancy so that more rewarding sensing decisions can be made in the future. A tradeoff has to be reached between these two often conflicting objectives. Within the framework of partially observable Markov decision processes, optimal opportunity tracking strategies have been studied in [3, 19] and reviewed in Chapter 10.

11.1.1.2 Spectrum Opportunity Exploitation

Once spectrum opportunities are detected, secondary users need to decide whether and how to exploit them. Specific issues include whether to transmit given that opportunity detectors may make mistakes, what modulation and transmission power to use

and how to share opportunities among secondary users to achieve a network-level objective.

The optimal design of spectrum access strategies in the presence of spectrum sensing errors has been addressed in [20,21]. Specifically, the interaction between the spectrum access protocols at the MAC layer and the operating characteristics of the spectrum opportunity detector at the physical layer is quantitatively characterized, and the optimal joint design of opportunity detectors, access strategies and opportunity tracking strategies is obtained. A review of these results is given in Chapter 10.

Modulation and power control in spectrum overlay networks also present unique challenges not encountered in the conventional wired or wireless networks. Since secondary users often need to transmit over non-contiguous frequency bands, orthogonal frequency division multiplexing (OFDM) has been considered as an attractive candidate for modulation in spectrum overlay networks [21–23]. Power control for secondary users needs to take into account the detection range of the opportunity detector, the maximum allowable interference level and the transmission power of primary users [15]. This complex networking issue remains largely open.

Spectrum opportunity sharing among secondary users has been addressed in the context of exploiting locally unused TV broadcast bands (see [1,2,24,25] and references therein). For this type of applications, spectrum opportunities are considered static or slowly varying in time. Real-time opportunity identification is not as critical a component as in applications that exploit temporal spectrum opportunities. It is often assumed that spectrum opportunities at any location over the entire spectrum are known.

In this chapter, we focus on distributed sharing of slowly varying spectrum opportunities among competing secondary users. Differing from the graph coloring approach considered in [1,24], game theory is employed to capture the distributed interaction among selfish secondary users with individual resource demands.

11.1.2 Organization of Chapter

The rest of this chapter is organized as follows. In Sect. 11.2 we introduce the game theoretic equilibrium and learning concepts that are needed to analyze our decentralized spectrum access model. In Sect. 11.3 we present the spectrum access model itself, along with algorithms for estimating channel competition, simultaneous adaptive learning of distributed resource allocation policies and centralized optimization of system-level spectral efficiency. The chapter concludes with a brief summary and discussion.

11.2 Review of Nash and Correlated Equilibrium in Games

Because our dynamic spectrum access model relies on a decentralized decision approach among secondary users, we rely on game theory to provide operational algorithms and performance analysis in this chapter. Thus, in this section, we present

a brief discussion of game theoretic concepts which are to be used, such as Nash and correlated equilibria, as well as an overview of game theoretic learning algorithms by which cognitive radios can adaptively discover how to allocate resources in a competitively optimal fashion.

11.2.1 Equilibrium Definitions

For a game with L players, the problem of each player $l = 1, 2, \ldots, L$ is to devise a rule for selecting their own action $\boldsymbol{X}^l$ from a set $\boldsymbol{S}^l$ (with size S^l), in order to maximize the expected value of a given utility function $u^l(\boldsymbol{X}^1, \boldsymbol{X}^2, \ldots, \boldsymbol{X}^L)$. Since each player only controls one of L variables, the problem requires careful consideration of the actions of other players, which are unknown in advance.

The central concept in non-cooperative game theory is an equilibrium, which identifies stable operating points of the system under certain conditions, such as common knowledge of rationality. The most common such equilibrium is due to Nash [24], defined as follows:

Definition 11.1. *For each player l, who takes random action $\boldsymbol{X}^l$, define a strategy π^l to be a probability distribution on $\boldsymbol{S}^l$, so that $\pi^l(x^l) = Pr(\boldsymbol{X}^l = x^l)$ for all $x^l \in \boldsymbol{S}^l$. Label the joint (random) action of all players by $\boldsymbol{X}$, and define the strategy profile π to be the product of all individual strategies, so that $\pi(x) = Pr(\boldsymbol{X} = x) = \prod_{k=1}^{L} \pi^k(x^k)$. ($\boldsymbol{X}$ resides on the space $\mathbb{S} = \boldsymbol{S}^1 \times \boldsymbol{S}^2 \times \ldots \times \boldsymbol{S}^L$.) We may write any strategy profile π as (π^l, π^{-l}) for any l, where π^{-l} is the strategy profile of all players but l. The expected utility to l resulting from π is*

$$u^l(\pi) = \sum_{x \in \mathbb{S}} u^l(x)\pi(x). \tag{11.1}$$

Now, π is a Nash equilibrium if each π^l is an optimal response to the collection π^{-l} of strategies of other players. That is,

$$u^l(\pi^l, \pi^{-l}) \geq u^l(\sigma^l, \pi^{-l}) \tag{11.2}$$

for all $l = 1, 2, \ldots, L$ and all possible alternative strategies σ^l.

The notation (σ^l, π^{-l}) means that l uses strategy σ^l instead of π^l.

In this chapter, we find it useful to focus on an important generalization of the Nash equilibrium, which was proposed in [1,2] and is known as the *correlated equilibrium*. This is defined as follows:

Definition 11.2. *Define a joint strategy π to be a probability distribution on the product space $\mathbb{S} = \boldsymbol{S}^1 \times \boldsymbol{S}^2 \times \ldots \times \boldsymbol{S}^L$. That is, $\pi(x) = Pr(\boldsymbol{X} = x)$ for joint actions $\boldsymbol{X}, x \in \mathbb{S}$. (The expected utility to l resulting from π is again as in (11.1).) We may decompose any strategy π into marginals (π^l, π^{-l}) for any l, where π^l is the marginal action distribution (strategy) of l, and π^{-l} is the marginal strategy of all players but l. Now, π is a correlated equilibrium if*

$$u^l(\pi^l, \pi^{-l}) \geq u^l(\sigma^l, \pi^{-l}), \tag{11.3}$$

for all $l = 1, 2, \ldots, L$ *and all possible alternative marginal strategies* σ^l *that are a function of* π^l.

In the correlated equilibrium, strategy π provides each player l with an action "recommendation" a^l. Based on this, and knowing π, a player could calculate an a posteriori probability distribution for the actions of other players, and hence an expected utility for each action. The equilibrium condition states that there is no deviation rule (represented by a function σ^l of π^l) that would award l a better expected utility. Combining (11.1) and (11.3), we obtain the equivalent condition:

$$\sum_{x^{-l} \in \boldsymbol{S}^{-l}} \pi(j, x^{-l})[u^l(k, x^{-l}) - u^l(j, x^{-l})] \leq 0 \tag{11.4}$$

for all $l = 1, 2, \ldots, L$, and $j, k \in \boldsymbol{S}^l$. That is, for any recommendation j to l, there is no profitable deviation k. The correlated equilibria comprise a convex set, given by:

$$\mathcal{CE} = \{\pi \in \Delta(\mathbb{S}) : \pi \text{satisfies (11.4)} \; \forall \, l, j, k\}. \tag{11.5}$$

The correlated equilibrium concept permits coordination between players, and can lead to improved performance over a Nash equilibrium [1]. If a correlated equilibrium distribution $\pi(s)$ can be written as a product of independent marginals $\pi(s) = \prod_{k=1}^{L} \pi^k(s^k)$, then it also satisfies the definition of a Nash equilibrium. The set (11.5) is also structurally simpler than the set of Nash equilibria; it is a convex set, whereas the Nash equilibria are isolated points at the extrema of this set [25]. Since the set of correlated equilibria is convex, fairness between players can also be addressed in this domain. Finally, decentralized, online adaptive procedures (see below) naturally converge to (11.5), whereas the same is not true for Nash equilibria (the so-called law of conservation of coordination [26]).

11.2.2 Adaptive Learning of Equilibria

A particularly interesting application of game theory is its usefulness in developing adaptive procedures in multiagent environments. Such procedures enable components of a system to learn a satisfactory (in game theory, equilibrium) policy for action through repeated interaction with their common environment. Moreover, these procedures are completely decentralized; each component interacts with others only through the effects of the environment, so explicit coordination is not necessary.

We outline the most well-known adaptive game theoretic learning schemes. In what follows, let $n = 0, 1, 2, \ldots$ be discrete time, let $\boldsymbol{X}_n^l$ denote the action of player l at time n, and let $\boldsymbol{X}_n^{-l}$ denote the joint actions of all players but l at time n.

1. *Best response:* If the common interaction between players is ignored, each player will simply attempt to maximize its performance, assuming the environment will remain the same. In the best response scheme, each player simply takes action

$$\boldsymbol{X}_{n+1}^{l} = \text{argmax}_{x \in \boldsymbol{S}^{l}} \{u^{l}(x, \boldsymbol{X}_{n}^{-l})\}.$$

That is, each player l acts optimally, assuming the other players will repeat their previous actions. Although it fails to account for simultaneous adaptation from multiple players, this approach can be shown to converge in some special cases, such as two-player zero-sum games, supermodular games, potential games and certain types of submodular games.

2. *Fictitious play:* The most well-known procedure, fictitious play was introduced in [27] and has been extensively studied since, see [28]. In this scheme, each player calculates a best response assuming the historical distribution of play is a good predictor of future actions. That is,

$$\boldsymbol{X}_{n+1}^{l} = \text{argmax}_{x \in \boldsymbol{S}^{l}} \{u^{l}(x, \bar{z}_{n}^{-l})\}$$

where $\bar{z}_{n}^{-l}$ is the empirical joint distribution of play up to time n. Fictitious play enjoys good convergence properties in practice, although convergence to Nash equilibrium is known to be false in general. One drawback is the need to explicitly observe and model the behavior of all opponents, which may not be appropriate for cognitive radios with limited awareness.

3. *Regret-based algorithms:* More recently, a general class of algorithms has been proposed in the form of regret-based learning [28–31]. Regret-based algorithms are, in a sense, a generalization fictitious play, which replace explicit opponent modeling with an implicit "regret matrix," θ_{n}^{l}. This tracks, for every pair of actions $j, k \in S^{l}$, the difference in utility if l had taken action k in the past everywhere it took action j. Given $\boldsymbol{X}_{n}^{l} = j$, the probability of $\boldsymbol{X}_{n+1}^{l} = k$ is proportional to $\theta_{n,jk}^{l}$, the regret from j to k. Learning proceeds by exploring and switching to actions that are perceived as "better" according to this regret measure.

In this paper we focus on regret-based procedures, as they are simple to implement and have well-understood convergence properties. Maintenance of θ_{n}^{l} requires minimal computation and no explicit awareness of other players. The main disadvantage is that players are required to know $u^{l}(k, \boldsymbol{X}_{n}^{-l})$ for all possible $k \in \boldsymbol{S}^{l}$ at each n. This requirement is removed in modified regret matching [29], which is presented (modified for our purposes) in Algorithm 11.1.

11.3 Decentralized Dynamic Spectrum Access Through Adaptive Reinforcement Learning

We consider a system of L cognitive radios, competing for access to $\mathcal{C}$ wireless communication channels which may be occupied at any time by primary users, who have priority in access. At successive time intervals of length Λ, each radio determines which of the $\mathcal{C}$ channels are unoccupied by primary users, and of these, the transmission rate (quality) sustainable by each channel. The objective of each radio $l = 1, 2, \ldots, L$ is to select a subset of unoccupied channels for use, in order to satisfy

its current demand level. However, since there is competition, there is no guarantee that the selected channels will be captured for exclusive use by l. Instead, we consider a simple slotted CSMA scheme for sharing each channel among users who select it, and propose a decentralized reinforcement learning scheme that allows each radio to find a satisfactory channel allocation through repeated channel selections and performance measurements.

Once each radio selects a subset of channels for use in a particular time interval, repeated competition takes place for each selected channel, as follows. Divide time into K subintervals of length λ/K. In each subinterval $k = 1, 2, \ldots, K$, all radios l active on channel i generate a backoff time $\tau_k^l(i)$; the smallest backoff time captures the channel for transmission for the remainder of the subinterval, as in a typical CSMA MAC protocol.

The history of successes and failures over these K channel capture attempts is used to give performance feedback to each user, i.e., as a sample of how much data it can expect to transmit over each selected channel. However, we can get more information out of these attempts. Specifically, we show how to couple the success/failure history with the history of backoff times used to estimate the number of users competing for these channels in Sect. 11.3.2. This extra information allows us to increase the level of cooperation in the cognitive radio problem; instead of merely trying to satisfy their own demand, users can attempt to minimize their interference with each other by explicitly favoring uncrowded channels over crowded ones. The complete radio utility function to accomplish this is formulated in Sect. 11.3.1.

We assume that the environment of the cognitive radio users varies slowly in time relative to the decision interval length λ. The variation we consider here is in terms of the channel occupancy of primary users, and the traffic demand level of individual cognitive radios. Furthermore, the cognitive radio utility function may be periodically updated by a central base station (see below). These slow variations in parameters motivate us to consider an *adaptive* reinforcement learning strategy, which allows radios to respond to changes in their environment without discarding everything they have learned to date. This adaptive strategy is based on the decentralized, game theoretic learning procedure of modified regret matching [29], and is outlined in Sect. 11.3.3.

Finally, even when radios act in the decentralized fashion described above, we may be able to improve performance by occasionally adjusting the behavior of each radio from a central controller. Since each radio acts to maximize a utility in our framework, we propose a scheme which parameterizes the radio utility, and periodically broadcasts parameter updates from a central controller, or base station, so as to improve global system performance.

We formulate this parameter adjustment scheme as an optimal "pricing" problem for the system, which we approach through stochastic optimization techniques. Suppose that a parameter (price) ϕ can be periodically broadcast (on a slow time scale) to each radio. Upon receiving the price update, each radio adjusts its utility as a function of ϕ and continues its usual behavior under the new utility. The aim of the central controller is to discover that ϕ which maximizes a global utility $G(\pi(\phi))$,

where $\pi(\phi)$ is the long-run (equilibrium) behavior of the radios under the utility priced by ϕ.

Since $\pi(\phi)$ and hence $G(\pi(\phi))$ is difficult or impossible to calculate a priori, a stochastic approximation approach is necessary for the discovery of the optimal ϕ. We propose to investigate Robbins–Monro type algorithms for this purpose [32]. By estimating the derivative $g(\phi) \approx \mathrm{d}G(\pi(\phi))/mathrm\mathrm{d}\phi$, we can use for example the steepest ascent method

$$\widehat{\phi}_{k+1} = \widehat{\phi}_k + \alpha_k \widehat{g}(\widehat{\phi}_k) \tag{11.6}$$

to successively approach an optimal ϕ, where $\alpha_k > 0$.

We propose to use spectral efficiency for our performance measure $G(\pi(\phi))$, which measures the time average proportion of available channels actually used by cognitive radios during a given period. Since radio decisions are decentralized, we do not expect the spectral efficiency to be 100%, but we hope to make incremental improvements through our pricing procedure.

A block diagram of our system is given in Fig. 11.1. In total, there are four time scales in our problem formulation. At the slowest time scale, the base station sets pricing parameters. Next is the time scale of variation of primary user activity and demand levels. Third, and much faster, is the decision time scale (intervals of length λ) of the cognitive radios themselves, and fourth, the fastest time scale (intervals of length λ/K), are the CSMA channel access attempts. For definiteness, we assume that pricing changes are on the order of hours, while primary user and demand variations are on the order of seconds. We take λ to be approximately 1 ms, and $K \approx 10$ CSMA attempts per channel allocation decision.

11.3.1 Decentralized Dynamic Spectrum Access Model and Radio Utility

In this section we present a mathematical outline of the decentralized dynamic spectrum access problem, which is used to formulate a utility function which each cognitive radio user attempts to maximize.

As above, we divide time into equal slots of length λ, and label each slot by $n = 1, 2, \ldots$. At the beginning of the nth time slot, we assume each cognitive radio $l = 1, 2, \ldots, L$ has the following information:

1. $\mathcal{C}$, the number of channels available for transmission use in the radio system.
2. $C \in \mathbb{R}^{\mathcal{C}}$, a vector giving the quality (bits transmissible per time slot) of each available channel.
3. $Y_n \in \Psi = \{x \in \mathbb{R}^{\mathcal{C}} : x(i) \in \{0, 1\} \text{ for all } i = 1, 2, \ldots, \mathcal{C}\}$, a vector showing the current channel usage pattern of primary users; channel i is in use if $Y_n(i) = 1$.
4. $d_n^l \in \mathbb{R}$, the current demand level of the cognitive radio user l (in bits per time slot).
5. A pricing parameter $\phi(i)_n$ for each channel $i = 1, 2, \ldots, \mathcal{C}$, obtained from the base station.

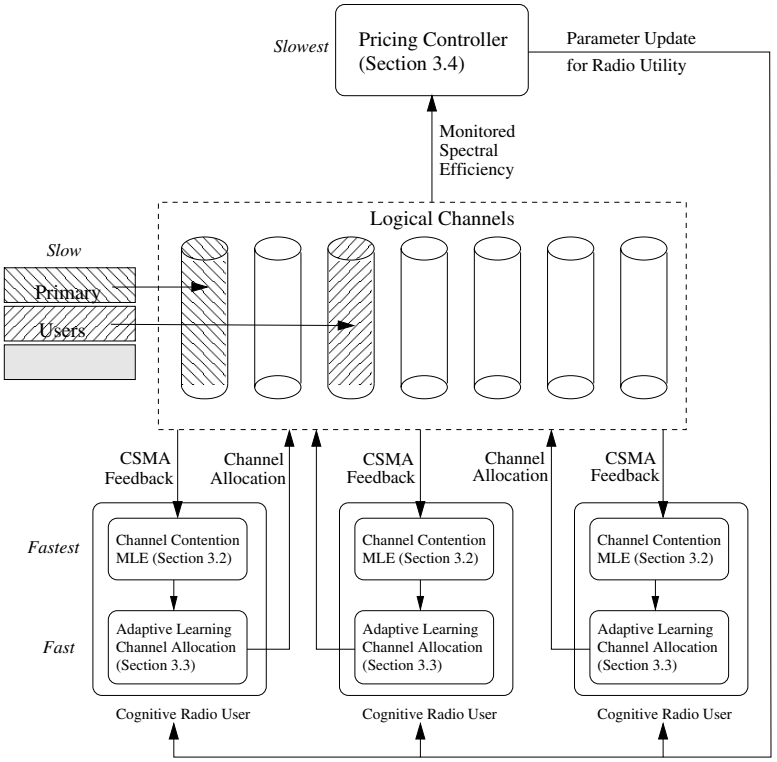

Fig. 11.1. Block diagram of decentralized learning system for cognitive radio dynamic spectrum access.

All these quantities are static or vary slowly in time, hence each radio knows their value before a channel allocation decision is made. For example, we will suppose that the primary user activity Y_n evolves according to a Markov chain with transition matrix $I+\varepsilon Q$, where $0 < \varepsilon \ll 1$ and Q is a generator matrix with each row summing to zero.

Next, each radio l chooses a channel allocation action $\boldsymbol{X}_n^l$, according to the learning scheme outlined in Sect. 11.3.3. For any $y \in \Psi$, define

$$\Psi_\perp(y) = \{x \in \Psi : x \cdot y = 0\} \tag{11.7}$$

to be the set of vectors in Ψ orthogonal to y. Action $\boldsymbol{X}_n^l$ then belongs to the slowly varying space $\boldsymbol{S}_n^l = \Psi_\perp(Y_n)$. That is, each player can select any collection of unused channels. For notational convenience, we adopt the following definition:

Definition 11.3. *For any index $i = 1, 2, \ldots, \mathcal{C}$ and any vector $x \in \Psi$, we say that $i \in x$ if and only if $x(i) = 1$.*

We also denote the joint action of all l decision makers by $\boldsymbol{X}_n$.

The joint channel allocation action $\boldsymbol{X}_n$ is then fixed for K successive CSMA transmission slots $n_1, n_2, \ldots, n_K$, each of length Λ/K. In each transmission slot n_k, each radio l generates a backoff time $\tau^l_{n_k}(i)$ for each selected channel $i \in \boldsymbol{X}^l_n$. Backoff times are generated according to a uniform distribution on the interval $(0, \tau_{\max})$ for some fixed parameter $\tau_{\max}$. Each radio waits until its backoff time expires then transmits data in the remainder of the slot only if the channel is sensed clear. If the smallest backoff time is sufficiently smaller than the next smallest backoff time (allowing time to sense the channel clear and switch from receive to transmit mode), then the radio with the smallest backoff time transmits successfully. Otherwise, there is a collision since two radios will have sensed the channel to be clear and transmitted data. Thus, each transmission slot is used at most by one radio. For each n, $i \in \boldsymbol{X}^l_n$, and $k = 1, 2, \ldots, K$, define

$$\gamma^l_{n_k}(i) = I\{\text{ channel } i \text{ captured by } l \text{ in slot } n_k\} \tag{11.8}$$

where $I\{\cdot\}$ is the usual indicator function.

At the end of the decision time slot n (of length Λ), each radio l has collected the following information on its CSMA attempts:

$$\gamma^l_n = \left\{\gamma^l_{n_k}(i) : i \in \boldsymbol{X}^l_n, k = 1, 2, \ldots, K\right\} \tag{11.9}$$

$$\tau^l_n = \left\{\tau^l_{n_k}(i) : i \in \boldsymbol{X}^l_n, k = 1, 2, \ldots, K\right\}. \tag{11.10}$$

This information is used for performance feedback. For each $i \in \boldsymbol{X}^l_n$, we calculate the proportional throughput achieved:

$$R^l_n(i) = \frac{1}{K}\sum_{k=1}^{K} \gamma^l_{n_k}(i). \tag{11.11}$$

Since the CSMA MAC is random, (11.11) is a random function of the joint decision $\boldsymbol{X}_n$. We note that we can take $R^l_n(i) = 0$ for all $i \notin \boldsymbol{X}^l_n$, and that $E[R^l_n(i)]$ clearly decreases in the contention level $\sum_{l=1}^{L} \boldsymbol{X}^l_n(i)$.

Section 11.3.2 also shows how to use (γ^l_n, τ^l_n) to obtain an estimate $\widehat{N}^l_n(i)$ for the number of users contending for channel i in decision time slot n. We show there that the maximum likelihood estimate for the contention level is given by $\widehat{N}^l_n(i) = 1 + \theta$, where θ solves

$$\sum_{k:\gamma^l_{n_k}(i)=0} \frac{a^\theta_k \log(a_k)}{1 - a^\theta_k} = \sum_{k:\gamma^l_{n_k}(i)=1} \log(a_k) \tag{11.12}$$

where $a_k = 1 - (\tau^l_k(i) + \delta)/\tau_{\max}$, for CSMA parameters $(\delta, \tau_{\max})$. We will also give an approximate solution to (11.12).

Given the information from (11.11) and (11.12), we propose a utility function to guide the reinforcement learning procedure. The utility for radio user l is given by:

$$\hat{u}^l(Y_n, d^l, X^l_n) = -(d^l - \sum_{i=1}^{\mathcal{C}} C(i) R^l_n(i))^2 - \sum_{i=1}^{\mathcal{C}} \phi(i) \widehat{N}^l_n(i) R^l_n(i). \tag{11.13}$$

The following remarks on (11.13) are in order:

1. The utility is implicitly a function of X_n, the actions of all players, through $\widehat{N}_n^l(i)$ and $R_n^l(i)$.
2. It is negative; to maximize (11.13), a radio must match its resources to its demand (first term), and simultaneously avoid designated crowded channels (second term).
3. The objective of avoiding other users as directed by the base station, and not exceeding the demand level d^l, enables cooperation between cognitive radio users.

The observed utility (11.13) is used as feedback to guide future channel allocation decisions in a decentralized fashion. To accomplish this, each radio takes a sequence of actions $\{X_1^l, X_2^l, \ldots, X_n^l\}$ and observes corresponding rewards $\{u_1^l, u_2^l, \ldots, u_n^l\}$. This data is used to generate a new action X_{n+1}^l through a decentralized, adaptive, regret-based reinforcement learning procedure, as described in Sect. 11.3.3. This procedure is game theoretic in nature, that is, it converges even when other cognitive radio users are simultaneously adapting their behavior. This is a critical observation, since naive, single-agent reinforcement learning procedures rely heavily on a static environment for convergence, which is not present in a multiagent situation. Game theoretic algorithms such as the one studied here enables cognitive radio activity to converge to an equilibrium (specifically a correlated equilibrium), which implies that each radio adopts a channel allocation that maximizes its own utility in response to the actions of others. This allows the cognitive radio system to learn, in a completely decentralized manner, to equitably share the available radio channels.

11.3.2 Channel Contention Estimate

In this section we show how to use the information obtained from repeated CSMA attempts to estimate the number of cognitive radio users competing for a given channel. This estimate is required for computing the utility (11.13) for reinforcement learning, and is based solely upon the history of successes and failures of repeated CSMA channel access attempts over a fixed period, along with the associated backoff times used in each attempt. This information is given in (11.9) and (11.10).

Consider a fixed channel i and a specific active user l. We wish to estimate $\widehat{N}_n^l(i)$, the number of users competing for resource i during decision slot n, based on K CSMA channel access attempts within that slot.

First, consider a single, general CSMA channel access attempt on channel i, and suppose there are $\theta(i)$ other active users competing for that channel. Each of these users $m \neq l$ chooses a random backoff time $\tau^m(i)$ uniformly on $(0, \tau_{\max})$. If l chooses $\tau^l(i) = t$, it captures the channel if $t < \tau^m(i) + \delta$ for all $\theta(i)$ users $m \neq l$, where δ is the time required to sense the channel clear and switch from RX to Tx mode. The probability of this event is given according to the order statistic $\tau_{\theta(i)}^{(1)}$, by

$$\Pr(l \text{ captures channel}) = \Pr\left(\tau_{\theta(i)}^{(1)} > t + \delta\right). \tag{11.14}$$

Likewise, we have

$$\Pr(l \text{ fails to capture channel}) = \Pr\left(\tau_{\theta(i)}^{(1)} < t + \delta\right). \tag{11.15}$$

It is well known that the order statistics for uniform random variables are given by the beta distribution. For the first order statistic $\tau_{\theta}^{(1)}$ on the interval $(0, \tau_{\max})$, the distribution simplifies to:

$$\Pr\left(\tau_{\theta(i)}^{(1)} > t + \delta\right) = \begin{cases} \left(1 - \frac{t+\delta}{\tau_{\max}}\right)^{\theta(i)}, & t \leq \tau_{\max} - \delta \\ 0, & t > \tau_{\max} - \delta. \end{cases} \tag{11.16}$$

A bit of reflection reveals that this indeed satisfies probabilistic intuition.

Suppose now that during decision interval n, l has recorded the success or failure of K CSMA attempts, along with the backoff time used in each attempt. These attempts are labeled $n_1, n_2, \ldots, n_K$. Note that $\theta(i)$ is held fixed over the K attempts by the decision structure. Then l can obtain a maximum likelihood estimate for the $\theta(i)$ by maximizing the quantity:

$$L(\theta(i)) = \prod_{k:\gamma_{n_k}^l=1} \Pr\left(\tau_{\theta(i)}^{(1)} > \tau_{n_k}^l(i) + \delta\right) \cdot \prod_{k:\gamma_{n_k}^l=0} \Pr\left(\tau_{\theta(i)}^{(1)} < \tau_{n_k}^l(i) + \delta\right) \tag{11.17}$$

$$= \prod_{k:\gamma_{n_k}^l=1} \left(1 - \frac{\tau_{n_k}^l(i) + \delta}{\tau_{\max}}\right)^{\theta(i)} \cdot \prod_{k:\gamma_{n_k}^l=0} \left(1 - \left(1 - \frac{\tau_{n_k}^l(i) + \delta}{\tau_{max}}\right)^{\theta(i)}\right) \tag{11.18}$$

where $\tau_{n_k}^l(i)$ is the backoff time of user l at time index k on channel i and $\gamma_{n_k}^l(i)$ denotes success or failure of the corresponding CSMA attempt, as in (11.8). The MLE is simply $\widehat{N}_n^l(i) = 1 + \arg\max_{\theta} L(\theta(i))$.

Differentiating the likelihood (or log likelihood) with respect to θ, we obtain that the maximizing θ must solve

$$\sum_{k:\gamma_{n_k}^l(i)=0} \frac{a_k^{\theta} \log(a_k)}{1 - a_k^{\theta}} = \sum_{k:\gamma_{n_k}^l(i)=1} \log(a_k) \tag{11.19}$$

where $a_k = 1 - (\tau_k^l(i) + \delta)/\tau_{\max}$.

Equation (11.19) is difficult to solve analytically. Numerically, we can state the following general properties:

1. $\widehat{N}_n^l(i)$ increases with the number of channel access failures.
2. $\widehat{N}_n^l(i)$ increases on average with the maximum successful backoff time.
3. $\widehat{N}_n^l(i)$ increases on average with the minimum unsuccessful backoff time.

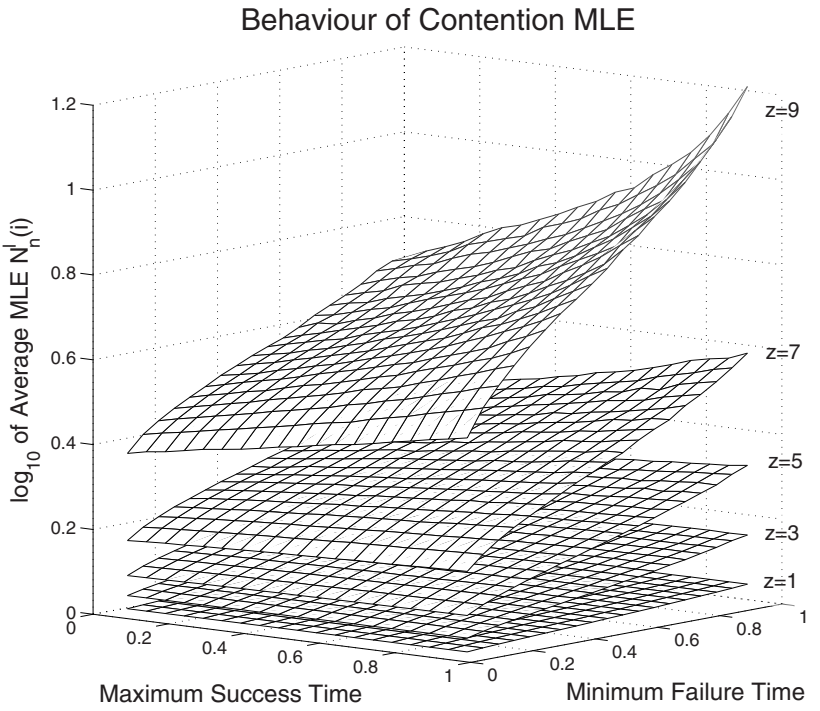

Fig. 11.2. Numerical plot of the (log) average of estimates $\widehat{N}_n^l(i)$. z denotes the number of failed CSMA attempts out of $K = 10$, and the maximum backoff time is $\tau_{\max} = 1$. Each data point represents an average over 5000 randomly generated observations satisfying the given limits for success and failure times.

A plot of $\widehat{N}_n^l(i)$ is given in Fig. 11.2 for $K = 10$ CSMA attempts and $\delta = 0$. For each data point, we specified the number of CSMA channel access failures $z = 1, 3, 5, 7, 9$ as well as the maximum and minimum backoff times of the successful and unsuccessful CSMA attempts, respectively ($\tau_{\max}$ is normalized to one). We then generated 5000 data samples corresponding to the specified limits and plotted an average of the results on a logarithmic scale, to emphasize the importance of the number of failures z on the estimate.

If we approximate a_k on the left-hand side by

$$\bar{a}_0 = \frac{1}{|I_0|} \sum_{k:\gamma_{n_k}^l(i)=0} a_k$$

where $|I_0| = \sum_{k=1}^{K}(1 - \gamma_{n_k}^l(i))$ is the number of terms in that summation (the number of channel access failures), (11.19) becomes:

$$|I_0|\frac{\bar{a}_0^\theta \log(\bar{a}_0)}{1-\bar{a}_0^\theta} = \sum_{k:\gamma^l_{n_k}(i)=1} \log(a_k). \tag{11.20}$$

The approximation in (11.20) corresponds to replacing the backoff times of failed channel access attempts by their average. From this, we can obtain an analytic solution:

$$\theta = -\log\left(1 + \frac{|I_0|\log(\bar{a}_0)}{\sum_{k\in I_1}\log(a_k)}\right) / \log(\bar{a}_0). \tag{11.21}$$

Numerical studies show that the approximation (11.21) is quite accurate on average, but can have a large variance in unfavorable conditions. Experimentally, it can be shown that the approximation error in (11.21) is small when either the number of channel access failures $|I_0|$ is small, or when the successful backoff times are small. In other cases, it may be preferable to use (11.21) to generate an initial guess, which may be refined by the Newton–Raphson method.

11.3.3 Adaptive Learning for Channel Allocation

In this section we describe our decentralized learning approach to the cognitive radio dynamic spectrum assignment problem. Our approach is based on the modified regret matching procedure of [29], which is formulated here as a distributed stochastic approximation algorithm. This formulation allows us to specify an adaptive variant of the original procedure, called "modified regret tracking," which uses a constant stepsize to dynamically adapt to time varying conditions, thus allowing users to function in a changing environment.

As is usual in reinforcement learning, each user takes a sequence of actions $\{X_n^l \in \boldsymbol{S}_n^l : n = 0,1,2,\ldots\}$ and observes a sequence of rewards $\{u_n^l \in \mathbb{R} : n = 0,1,2,\ldots\}$. The action at time $n+1$ is a random function of this history of actions and rewards.

At each decision period n, users take joint action $X_n \in \mathbb{S}$, with user l taking action $X_n^l \in \boldsymbol{S}^l$. To implement the algorithm, each user l uses the observed utilities associated with past joint actions $\{X_n : n = 1,2,\ldots\}$ to derive regret values $\theta^l_{n,jk} : j,k \in \boldsymbol{S}^l$, according to:

$$\theta^l_{n,jk} = \sum_{\tau\le n: X^l_\tau = k} \varepsilon_{\tau-1}\Big(\prod_{\sigma=\tau}^{n-1}(1-\varepsilon_\sigma)\Big)\frac{p^l_\tau(j)}{p^l_\tau(k)}u^l(k, X^{-l}_\tau) - \sum_{\tau\le n: X^l_\tau = j} \varepsilon_{\tau-1}\Big(\prod_{\sigma=\tau}^{n-1}(1-\varepsilon_\sigma)\Big)u^l(j, X^{-l}_\tau). \tag{11.22}$$

If $\varepsilon_\tau = 1/(\tau+1)$, this is simply the average:

$$\theta^l_{n,jk} = \frac{1}{n}\sum_{\tau\le n: X^l_\tau = k}\frac{p^l_\tau(j)}{p^l_\tau(k)}u^l(k, X^{-l}_\tau) - \frac{1}{n}\sum_{\tau\le n: X^l_\tau = j} u^l(j, X^{-l}_\tau). \tag{11.23}$$

If $\varepsilon_\tau = \varepsilon$, it is the exponentially weighted moving average:

$$\theta^l_{n,jk} = \sum_{\tau \le n: X^l_\tau = k} \varepsilon(1-\varepsilon)^{n-\tau} \frac{p^l_\tau(j)}{p^l_\tau(k)} u^l(k, X^{-l}_\tau) - \sum_{\tau \le n: X^l_\tau = j} \varepsilon(1-\varepsilon)^{n-\tau} u^l(j, X^{-l}_\tau). \tag{11.24}$$

These values are computed recursively in the algorithm below.

To gain some intuition, we refer back to the original regret matching algorithm of [28]. Here, the regret value is taken as the simple time average

$$\theta^l_{n,jk} = \frac{1}{n} \sum_{\tau \le n: X^l_\tau = j} \left(u^l(k, X^{-l}_\tau) - u^l(j, X^{-l}_\tau) \right). \tag{11.25}$$

That is, the regret measures the average gain that l would have received had he played k in the past instead of j. If the gain is positive, then clearly l should be more likely to switch to action k in the future, and in fact regret matching does exactly this by switching to each action k at time $n+1$ with probability proportional to the positive component of $\theta^l_{n,jk}$. Note, however, that (11.25) requires that l knows what utility he would have received for each action, *even if that action was not taken*. To overcome this difficulty, [29] approximates the first term of the summation (11.25) by the first summation in (11.23).

The complete procedure, including the exact formulation of action probabilities, is summarized in Algorithm 11.1, which is carried out independently by each user.

Algorithm 11.3.1 Adaptive Learning for Channel Allocation: The regret-based algorithm for user activation has parameters $(u^l, \mu, \delta, \{\varepsilon_n : n = 1, 2, \ldots\}, \theta^l_0, X^l_0)$, where u^l are the user utilities, μ is a function of the utilities as in (11.29), δ is a small probability with which actions are chosen from a uniform distribution, $\{\varepsilon_n\}$ is a small stepsize, and θ^l_0, X^l_0 are arbitrary initial regrets and actions.

Define the $\boldsymbol{S}^l \times \boldsymbol{S}^l$ matrix with entries:

$$\boldsymbol{H}^l_{jk}(X_n) = I\{X^l_n = k\} \frac{p^l_n(j)}{p^l_n(k)} u^l(k, X^{-l}_n) - I\{X^l_n = j\} u^l(j, X^{-l}_n). \tag{11.26}$$

The Procedure Is As Follows:

1. *Initialization:* Set $n = 0$ and take action X^l_0. Initialize regret $\theta^l_0 = \boldsymbol{H}^l(X_0)$. Repeat for $n = 0, 1, 2, \ldots$:
   Action update: Choose $X^l_{n+1} = k$ with probability

$$\Pr(X^l_{n+1} = k | X^l_n = j, \theta^l_n = \theta^l) = \begin{cases} (1-\delta) \min\left(\max\{\theta^l_{jk}, 0\}/\mu, \frac{1}{\boldsymbol{S}^l - 1} \right) + \frac{\delta}{\boldsymbol{S}^l}, & k \ne j, \\ 1 - \sum_{i \ne j} \left[(1-\delta) \min\left(\max\{\theta^l_{ji}, 0\}/\mu, \frac{1}{\boldsymbol{S}^l - 1} \right) + \frac{\delta}{\boldsymbol{S}^l} \right] & k = j. \end{cases} \tag{11.27}$$

Regret value update: Calculate $\boldsymbol{H}^l(X_{n+1})$, and update θ_{n+1} using the stochastic approximation (SA):

$$\theta^l_{n+1} = \theta^l_n + \varepsilon_n(\boldsymbol{H}^l(X_{n+1}) - \theta^l_n). \tag{11.28}$$

In (11.27), μ is a normalization constant, which is chosen

$$\mu > (S^l - 1)(u^l_{\max} - u^l_{\min}) \tag{11.29}$$

over all $l = 1, 2, \ldots, L$, where $(u^l_{\max}, u^l_{\min})$ are obtained from (11.13).

Note that θ^l_n is a moving average of the updates $\{H^l(X_k) : k = 1, 2, \ldots, n\}$. Because of this, Algorithm 11.1 can be viewed as a stochastic approximation with a constant stepsize $\varepsilon_n \equiv \varepsilon > 0$; actions are chosen with probability proportional to their (moving) average potential performance in the past. (This differs from best response, which would base action choices on the immediately previous result, essentially setting $\varepsilon = 1$.) For the original modified regret matching algorithm of [29], one would instead use $\varepsilon_n = 1/(n + 1)$.

Since the utility varies, a constant stepsize in Algorithm 11.3.1 is needed to keep users responsive to the changes.

11.3.3.1 Convergence of Regret-Based Learning

When a decreasing stepsize $\varepsilon_n = 1/(n+1)$ is used in Algorithm 11.1, it is proven in [29] that the global empirical distribution of play (defined below) converges almost surely to the set of ε-correlated equilibria. If, in addition, the "tremble" term δ in (11.27) is decreased sufficiently slowly, convergence is to the set of correlated equilibria proper (11.5).

It is therefore reasonable to expect similar convergence results of the constant stepsize version of Algorithm 11.3.1, with fixed small $\varepsilon_n = \varepsilon$ and a fixed small tremble δ. The general relation between decreasing and constant stepsize stochastic approximation (SA) algorithms is well known [32]. Essentially, when a decreasing stepsize SA converges almost surely, it can be shown that the constant stepsize version converges weakly, as the stepsize $\varepsilon \to 0$. Intuitively then, our adaptive version of Algorithm 11.3.1 should track the set of correlated equilibria, with the benefit that changes to the utility functions are handled smoothly by the constant stepsize.

We now describe in detail what is meant by this type of convergence. First, convergence is stated in terms of the empirical distribution of play, which can be viewed as a diagnostic that monitors the performance of the entire cognitive radio network. This is defined as follows:

Definition 11.4. *The empirical distribution of play up to time n is:*

$$\bar{z}_n = \sum_{\tau \le n} \varepsilon_{\tau-1}\Big(\prod_{\sigma=\tau}^{n-1}(1 - \varepsilon_\sigma)\Big)e_{X_\tau} \tag{11.30}$$

where $e_x = [0, 0, \ldots, 1, 0, \ldots, 0]$ with the one in the xth position.

Here ε_n is a weighting factor. If $\varepsilon_n = 1/(n+1)$, the empirical distribution is simply

$$\bar{z}_n = 1/n \sum_{\tau} e_{X_\tau}. \tag{11.31}$$

If $\varepsilon_n \equiv \varepsilon > 0$ is constant, it is the exponentially weighted moving average

$$\bar{z}_n = \varepsilon \sum_{\tau} (1-\varepsilon)^{n-\tau} e_{X_\tau}. \tag{11.32}$$

Note that in both cases $\bar{z}_n$ is an empirical frequency, since $\sum_i \bar{z}_n(i) = 1$.

We point out here that $\bar{z}$ satisfies the following recursion:

$$\bar{z}_{n+1} = \bar{z}_n + \varepsilon_n (e_{X_{n+1}} - \bar{z}_n) \tag{11.33}$$

where X_{n+1} is constructed according to Step (2a) of Algorithm 11.3.1. When a decreasing stepsize $\varepsilon_n = 1/(n+1)$ is used, (11.33) directly yields (11.31). With a constant stepsize $\varepsilon_n = \varepsilon$, (11.33) directly yields (11.32).

Second, in contrast to most convergence results, convergence of the empirical distribution of play for Algorithm 11.3.1 is not to a specific point, but to the *set* of correlated equilibria ($\mathcal{CE}$). This property is as follows:

Definition 11.5. *$\bar{z}_n$ converges to the set $\mathcal{CE}$ if for any $\varepsilon > 0$ there exists $N_0(\varepsilon)$ such that for all $n > N_0$ we can find $\psi \in \mathcal{CE}$ at a distance less than ε from $\bar{z}_n$.*

The actual proof of weak convergence for the adaptive modified regret tracking algorithm can be approached in two ways. First, one can attempt to adapt the original proof in [29] for a constant stepsize. Second, one can take a differential inclusion approach, similar to that found in [31,33]. The first approach appears plausible, but technically difficult. The proof in [29] is based on the idea of Blackwell approachability [34], to which the existence of a decreasing stepsize is central. One would therefore be forced to begin by modifying Blackwell's 1956 result, then proceed to carry the modifications through the proof in [29]. The differential inclusion approach therefore appears more promising. Although [31,33] still assumes here a decreasing stepsize, it treats the convergence of the original (non-modified) regret matching algorithm of [28] in such a way that the constant stepsize result can easily be obtained through the methods of [32]. Since the modified procedure (used here) of [29] is obtained from [28], it should not be too difficult to use similar methods here.

11.3.4 Stochastic Optimization of Spectral Efficiency via Centralized Pricing

In this section we describe a simple stochastic optimization approach for improving spectral efficiency in the decentralized channel access learning environment. The approach relies on a base station, which monitors only the outcome of cognitive radio activities. That is, the base station is not aware of the actions of individual cognitive radios, but only of how often free channels are used by the group for data

transmission. It attempts to influence the behavior by periodically broadcasting a common $C\text{-}valued$ parameter vector to each radio, which is used by the cognitive radios to update their utility function. The results of this section are not integral to the decentralized learning scheme; the pricing scheme describes a way to improve the equilibrium behavior obtained in the previous sections from a global perspective, but is not necessary to the operation of the cognitive radio system as already described.

Recall the utility function (11.13) of cognitive radio users, which is parameterized by pricing vector ϕ. For each channel i, $\phi(i)$ represents a unit interference penalty; user l essentially pays a cost $\phi(i)$ for each portion of channel i it uses and each user present on channel i. This is meant as a simple disincentive so that a base station, seeing that channel i is too crowded, can encourage users to move to other channels by imposing a high cost $\phi(i)$. Conversely, users may be attracted to low-cost channels in order to balance load across the spectrum.

Although it is possible to devise much more sophisticated incentive rules, possibly through mechanism design theory, than the one presented here, we feel that at least elementary control can be imposed through our formulation, and that it provides a sufficient proof of concept of stochastic optimization-based pricing for tuning cognitive radio networks. Moreover, the basic stochastic optimization approach, which we outline here, will remain the same regardless of the particular pricing parametrization used.

The objective of the base station is to discover, through experimentation, a pricing parameter ϕ which maximizes the spectral efficiency of the cognitive radio system. The spectral efficiency is defined as the average proportion of available radio channels that are used by the cognitive radios, which may be sampled over T decision intervals as:

$$\widehat{\mathrm{SE}}(\phi) = \frac{\sum_{n=1}^{T}\sum_{l=1}^{L}\min\left\{\sum_{i=1}^{\mathcal{C}} C(i)R_n^l(i), d^l\right\}}{\sum_{n=1}^{T}\sum_{i=1}^{\mathcal{C}}[C(i)(1-Y_n(i))]}. \tag{11.34}$$

Note that (11.34) is based only on the observable channel usage.

To incrementally improve the spectral efficiency, we propose the following algorithm:

Algorithm 11.3.2 Stochastic Optimization-Based Pricing: For large T, set pricing interval length $T\Lambda$. A decision time n of the cognitive radios (on time scale Λ) is said to belong to pricing interval m if $mT + 1 \leq n \leq (m+1)T$. For pricing intervals $m = 0, 1, 2, \ldots$, repeat the following:

1. Monitor the decisions in pricing interval m. That is, collect data $Y_n(i)$ and $R_n^l(i)$ for $i = 1, 2, \ldots, \mathcal{C}$, $l = 1, 2, \ldots, L$, and n in pricing interval m.
2. At the end of interval m, calculate the spectral efficiency according to (11.34) using the data gathered. Estimate the derivative $\widehat{\mathrm{dSE}}/\mathrm{d}\phi$.
3. Broadcast a new pricing vector for pricing interval $m+1$, according to

$$\widehat{\phi}_{m+1} = \widehat{\phi}_m + \alpha_m \frac{\widehat{\mathrm{dSE}}}{\mathrm{d}\phi}(\widehat{\phi}_m). \tag{11.35}$$

The derivative may be estimated using standard approximation methods, for example the finite difference or simultaneous perturbation methods [36].

Conclusion

In this chapter we have presented an iterative, decentralized method for discovering efficient dynamic spectrum access policies for cognitive radio. Under the spectrum overlay model, we have shown how the spectrum access problem can be treated as a game theoretic problem and given algorithms that allow cognitive radios to independently assess and adapt to their environment in real time.

The key advantage of our approach is complete decentralization, that is, the lack of requirement for any collaboration or communication between cognitive radios. We do require, in the centralized pricing scheme of Sect.11.3.4, the ability to receive occasional updates from a central base station, but this feature is meant only as an optional enhancement to the decentralized system. We are able to obtain effective performance from the decentralized scheme essentially for two reasons. First, since radios are aware of the presence of competitors, they are able to estimate and adapt to channel competition by leveraging game theoretic algorithms specifically designed to converge in a multiuser setting. Second, we have built cooperative tendencies into the utility function (11.13) itself; radios are penalized for obtaining more resources than they require and are bound to obey direction from the base station through the pricing function ϕ. While this second consideration might be negated by selfish design, the structural results would not change; the equilibrium obtained would simply be less efficient than that obtained through cooperative design. Let us emphasize: cooperative design in essentially decentralized systems allows us to achieve many of the benefits of a completely integrated architecture without the same costly infrastructure.

The constant step size learning algorithm presented in this chapter converges weakly to the set of correlated equilibria of a non-cooperative game. Moreover, the algorithm can be used to track a slowly time varying correlated equilibrium set caused due to changing activity of primary users, with the limiting behavior of the algorithm captured by a differential inclusion. Suppose we were to assume that primary user activity evolves according to a slow Markov chain with transition probability matrix $I + \epsilon Q$ (where $\epsilon > 0$ is a small parameter and Q is a generator matrix with each row summing to zero). With this assumption, how can one analyze the tracking performance of the learning algorithm with step size ϵ? Note that the adaptation speed (step size ϵ) of the algorithm matches the speed at which the correlated equilibrium set changes (transition matrix $(I + \epsilon Q)$). In our recent work [36,37], we have shown that the limiting behavior of the stochastic approximation algorithm for tracking a parameter evolving according to a Markov chain is captured by a Markovian switched ordinary differential equation. This result was somewhat remarkable, since typically the limiting process of a stochastic approximation algorithm is a deterministic ordinary differential equation. We conjecture that the limiting behavior of Algorithm 11.3.1 is captured by a Markovian switched differential inclu-

sion (see [38]). This analysis requires use of yet another extremely powerful tool in stochastic analysis namely, the so-called "martingale problem" of Strook and Varadhan, see [41,42] for comprehensive treatments of this area.

There are many other interesting avenues for continuation of this research. Aside from improving and validating the algorithms presented here, one can modify the problem to consider the case of partial channel observation. This is especially important when the number of channels becomes too large for simultaneous monitoring. Moreover, for this situation, it is important to identify initial methods for eliminating a large number of channels from consideration, in order to improve the convergence rate and memory requirements of the adaptive learning approach considered here. Finally, we can expand our scope from static games to stochastic games, in which the player actions not only determine their immediate utility, but also give a probability distribution over new games to be played in future rounds. A stochastic game approach for similar sensor-based systems has been carried out in [41,42].

References

1. R. Aumann, "Correlated equilibrium as an expression of Bayesian rationality," *Econometrica*, vol. 55, no. 1, pp. 1–18, 1987.
2. R. Aumann, "Subjectivity and correlation in randomized strategies," *J. Math. Econ.*, vol. 1, pp. 67–96, 1974.
3. Q. Zhao, L. Tong, and A. Swami, "Decentralized cognitive MAC for dynamic spectrum access," in *Proc. IEEE DySPAN 2005*, pp. 224–232, 2005.
4. H. Zheng and L. Cao, "Device-centric spectrum management," in *Proc. IEEE DySPAN 2005*, pp. 56–65, 2005.
5. N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," in *Proc. IEEE DySPAN 2005*, pp. 269–278, 2005.
6. J. Robinson, "An iterative method of solving a game," *Ann. Math.*, vol. 54, pp. 298–301, 1951.
7. J. Huang, R. Berry, and M. Honig, "Auction-based spectrum sharing," *Springer, Mobile Netw. Appl.*, vol. 11, no. 3, pp. 405–418, 2006.
8. M. McHenry, "Spectrum white space measurements," June 2003. Presented to New America Foundation Broadband Forum; Measurements by Shared Spectrum Company, Available at http://www.newamerica.net/Download Docs/pdfs/Doc File 185 1.pdf.
9. D. Hatfield and P. Weiser, "Property rights in spectrum: Taking the next step," in *Proc. First IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Nov. 2005.
10. L. Xu, R. Tonjes, T. Paila,W. Hansmann, M. Frank, and M. Albrecht, "DRiVE-ing to the Internet: Dynamic radio for IP services in vehicular environments," in *Proc. 25th Annual IEEE Conference on Local Computer Networks*, pp. 281–289, Nov. 2000.
11. Y. Benkler, "Overcoming agoraphobia: Building the commons of the digitally networked environment," *Harvard J. Law Technol.*, Winter 1997–1998.
12. J. Mitola, "Cognitive radio for flexible mobile multimedia communications," in *Proc. IEEE International Workshop on Mobile Multimedia Communications*, pp. 3–10, 1999.
13. "DARPA: the next generation (XG) program." http://www.darpa.mil/ato/programs/xg/index.htm.

14. Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access: signal processing, networking, and regulatory policy," *IEEE Signal Processing Magazine*, vol. 55, no. 5, pp. 2294–2309, May, 2007.
15. Q. Zhao, "Spectrum opportunity and interference constraint in opportunistic spectrum access," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2007.
16. D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in *Proc. 38th. Asilomar Conference on Signals, Systems, and Computers*, pp. 772–776, 2004.
17. W. Gardner, "Signal interception: A unifying theoretical framework for feature detection," *IEEE Trans. Commun.*, vol. 36, pp. 897–906, Aug. 1988.
18. A. Sahai, N. Hoven, and R. Tandra, "Some fundamental limits on cognitive radio," in *Proc. Allerton Conference on Communication, Control, and Computing*, Oct. 2004.
19. Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Select. Areas Commun.,* Special Issue on Adaptive, Spectrum Agile and Cognitive Wireless Networks, Apr. 2007.
20. Y. Chen, Q. Zhao, and A. Swami, "Joint PHY/MAC design of opportunistic spectrum access in the presence of sensing errors," submitted to *IEEE Trans. Signal Process.* in Jan. 2007.
21. T. Weiss and F. Jondral, "Spectrum pooling: An innovative strategy for enhancement of spectrum efficiency," *IEEE Commun. Mag.*, vol. 42, pp. 8–14, Mar. 2004.
22. U. Berthold and F. K. Jondral, "Guidelines for designing OFDM overlay systems," in *Proc. First IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Nov. 2005.
23. H. Tang, "Some physical layer issues of wide-band cognitive radio systems," in *Proc. First IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Nov. 2005.
24. J. Nash, "Non-cooperative games," *Ann. Math.*, vol. 54, no. 2, pp. 286–295, 1951.
25. R. Nau, S. Canovas, and P. Hansen, "On the geometry of Nash equilibria and correlated equilibria," *Int. J. Game Theory*, vol. 32, no. 4, pp. 443–453, 2004.
26. S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to Nash equilibrium," *Am. Econ. Rev.*, vol. 93, no. 5, pp. 1830–1836, Dec. 2003.
27. D. Fudenberg and D. Levine, *The theory of learning in games*. MIT Press, 1999.
28. S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.
29. S. Hart and A. Mas-Colell, "A reinforcement procedure leading to correlated equilibrium," *Economic Essays*, Springer, 2001, pp. 181–200.
30. A. Cahn, "General procedures leading to correlated equilibria," *Int. J. Game Theory*, vol. 33, no. 1, pp. 21–40, 2004.
31. M. Benaim, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions ii: Applications," *UCLA Department of Economics, Levine's Bibliography*, May 2005.
32. H. Kushner and G. Yin, *Stochastic approximation and recursive algorithms and applications*, 2nd ed. New York, NY: Springer-Verlag, 2003.
33. M. Benaim, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM J. Control Optim.*, vol. 44, no. 1, pp. 328–348, 2005.
34. D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific J. Math.*, vol. 6, pp. 1–8, 1956.

35. J. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control*. Wiley Press, 2003.
36. G. Yin and V. Krishnamurthy "Least mean square algorithms with Markov regime switching limit," *IEEE Trans. Autom. Control,* vol. 50, no. 5, pp. 577–593, 2005.
37. G. Yin, V. Krishnamurthy, and C. Ion, "Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization," *SIAM J. Optim*., vol. 14, no. 4, pp. 1187–1215, 2004.
38. A. Benveniste, M. Metivier, and P. Priouret "Adaptive Algorithms and Stochastic Approximations," in *Applications of Mathematics*, vol. 22, Springer-Verlag, 1990.
39. S. Ethier and T. Kurtz, *Markov processes–characterization and convergence.* Wiley, 1986.
40. H. Kushner, *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory.* Cambridge, MA: MIT Press, 1984.
41. M. Maskery and V. Krishnamurthy, "Decentralized algorithms for netcentric force protection against anti-ship missiles," (preprint) *IEEE Trans. Aerospace Electr. Syst.*, 2007.
42. M. Maskery and V. Krishnamurthy, "Network enabled missile deflection: Games and correlated equilibrium," (preprint), *IEEE Trans. Aerospace Electr. Syst.*, 2007.

Additional Reading

1. S. Sankaranarayanan, P. Papadimitratos, A. Mishra, and S. Hershey, "A bandwidth sharing approach to improve licensed spectrum utilization," in *Proc. First IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2005.
2. Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, 2006.
3. H. Zheng and C. Peng, "Collaboration and fairness in opportunistic spectrum access," in *Proc. IEEE International Conference on Communications (ICC)*, 2005.
4. W. Wang and X. Liu, "List-coloring based channel allocation for open-spectrum wireless networks," in *Proc. IEEE VTC*, 2005.
5. M. Steenstrup, "Opportunistic use of radio-frequency spectrum: A network perspective," in *Proc. First IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, 2005.
6. M. Maskery and V. Krishnamurthy, "Decentralized activation in a ZigBee-enabled unattended ground sensor network: A correlated equilibrium game theoretic analysis," submitted to *IEEE/ACM Trans. Netw.*, 2006.
7. V. Krishnamurthy, G. Yin, and M. Maskery "Stochastic approximation based tracking of correlated equilibria for game-theoretic reconfigurable sensor network deployment," in *Proc. IEEE Conference on Decision and Control*, 2006.
8. V. Krishnamurthy, M. Maskery, and M. Hanh Ngo, "Scalable sensor activation and transmission scheduling in sensor networks over Markovian fading channels," in *Wireless sensor networks. Signal processing and communications perspectives*, Wiley Press, 2007.
9. M. Maskery and V. Krishnamurthy, "Decentralized activation in a ZigBee-enabled unattended ground sensor network: A correlated equilibrium game theoretic analysis," in *Proc. IEEE International Conference on Communications*, 2007.
10. M. Maskery and V. Krishnamurthy, "Decentralized management of sensors in a multiattribute environment under weak network congestion," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.