Dynamic Spectrum Sharing Among Repeatedly Interacting Selfish Users With Imperfect Monitoring

Yuanzhang Xiao and Mihaela van der Schaar

Abstract

In this paper, we develop a novel design framework for dynamic spectrum sharing among secondary users who adjust their power allocation while satisfying under interference temperature (IT) constraints imposed by primary users. The considered interaction among the secondary users is characterized by the following three unique features. First, secondary users are selfish and aim to maximize their own long-term payoffs from utilizing the network rather than obeying the prescribed allocation of a centralized controller. Second, the secondary users are interacting with each other repeatedly and they can coexist in the system for a long time as long as the IT constraints are not violated. Third, the secondary users have imperfect and limited monitoring ability: they only observe whether the IT constraints are violated, and their observation is imperfect due to the erroneous measurements. To capture these unique features, we model the interaction of secondary users as a repeated game with imperfect monitoring. We first characterize the set of Pareto optimal operating points that can be achieved by deviation-proof spectrum sharing policies, which are policies that the selfish users find it in their interest to comply with. Next, for any given operating point in this set, we show how to construct a deviation-proof policy (protocol) to achieve it. The constructed deviation-proof policy is amenable to distributed implementation, and allows users to transmit in a time-division multiple-access (TDMA) fashion. In the presence of strong multi-user interference, our policy outperforms existing spectrum sharing policies that dictate users to transmit at constant power levels simultaneously. Moreover, our policy can achieve Pareto optimality even when secondary users have imperfect and limited monitoring ability, as opposed to existing solutions based on repeated game models, which require perfect monitoring abilities. Simulation results validate our analytical results and quantify the performance gains enabled by the proposed spectrum sharing policies.

Y. Xiao and M. van der Schaar are with Department of Electrical Engineering, UCLA. Email: {yxiao,mihaela}@ee.ucla.edu.
I. INTRODUCTION

Cognitive radios have increased in popularity in recent years because they enable unlicensed users (or secondary users) to share the spectrum with licensed users (or primary users), as long as the primary users’ quality of service (QoS), such as the signal-to-interference-and-noise ratio (SINR) and the throughput, is not affected by the addition of secondary users [1]–[4]. A common approach to guarantee primary users’ QoS requirements is to impose interference temperature (IT) constraints [1][5]–[16]; that is, secondary users (SU) cannot generate an interference level higher than the interference temperature limit set by the primary users (PU). One of the major challenges in designing cognitive radio systems is the spectrum sharing policy that enables efficient spectrum usage among SUs while maintaining the IT constraints set by PUs.

The spectrum sharing policy, which specifies the SUs’ transmit power levels, is essential in order to improve spectrum efficiency and protect the primary users’ QoS. Since SUs can use the spectrum as long as they do not degrade the QoS of the PUs, they can use the spectrum and coexist in the system for long periods of time. In general, the optimal spectrum sharing policy should allow SUs to transmit at different power levels temporally even when the environment (e.g. the number of SUs, the channel gains) remains unchanged. However, most existing spectrum sharing policies require the SUs to transmit at constant power levels over the time horizon in which they interact\(^1\) [5]–[16]. These policies with constant power levels are inefficient in many spectrum sharing scenarios where the interference among the SUs is strong. When multi-user interference is strong, increasing one user’s power level significantly degrades the other users’ QoS. Hence, when the cross channel gains are large, the feasible QoS region is nonconvex [18][19]. When the feasible QoS region is nonconvex, a spectrum sharing policy with constant power levels is inferior to a policy with time-varying power levels in which the users transmit in a time-division multiple-access (TDMA) fashion, because the latter can achieve operating points on the Pareto boundary of the convex hull of the nonconvex feasible QoS region, which are not achievable by the former.

Another important feature neglected in the design of spectrum sharing policies in recent works [6]–[10][12]–[16] is the selfishness of SUs, which aim to maximize their own utility and may

\(^1\)Although some spectrum sharing policies go through a transient period of adjusting the power levels before the convergence to the optimal power levels, the users maintain constant power levels after the convergence.
deviate from the prescribed spectrum sharing policy, if by doing so their QoS can be improved. Hence, the spectrum sharing policy should be \textit{deviation-proof}, which means that selfish SUs cannot improve their QoS by deviating from the policy. In this way, selfish SUs will find it in their self-interest to follow the policy.

Given the fact that the SUs will interact with each other repeatedly when sharing the spectrum, we model the interaction among secondary users as a repeated game. In a repeated game, the stage game is played repeatedly, and a user’s payoff in the repeated game is the discounted average of the stage-game payoffs (QoS in the stage games). Users can choose different actions (power levels) in different stage games, and the repeated-game payoff is a convex combination of different stage-game payoffs. (Note that this may not be achievable by a spectrum sharing policy with constant power levels). A repeated-game strategy prescribes what action to take given past observations, and hence, it provides a spectrum sharing policy. If a repeated game strategy constitutes an equilibrium, then no user can gain from deviation at any occasion. Hence, an equilibrium strategy is a deviation-proof policy.

The spectrum sharing policy in a repeated game framework was studied in [20][21], under the assumption of \textit{perfect} monitoring. Specifically, they require each secondary user to be able to perfectly monitor the transmit power levels of all the secondary users. In practice, the monitoring cannot be perfect. With \textit{imperfect} monitoring, the punishment-based policies in [20][21] must have performance loss, namely they cannot achieve Pareto optimal operating points. This is because even if no user deviates, the punishment, in which all the users receive low payoffs, will occur with a positive probability due to imperfect monitoring. Thus, the repeated-game payoff, when averaged over all the expected stage-game payoffs, cannot be Pareto optimal because of the low payoffs received in the punishment phase.

Repeated games with imperfect monitoring have been studied extensively in the game theory literature. In [22], it is shown that if certain sufficient conditions are satisfied, Pareto optimal operating points can be asymptotically achieved when the users are sufficiently patient. The users’ patience reflects how they discount future payoffs, and is represented by their discount factors. The users being sufficiently patient means that their discount factors can be arbitrarily close one. This requirement results from the punishment-based equilibrium strategies for the similar reason mentioned in the discussion of [20][21]. In addition, the sufficient conditions in [22] require the users to be able to statistically distinguish the outcomes from sufficiently many
different actions. Roughly speaking, in the spectrum sharing scenarios, it requires that any pair of two users can statistically distinguish the outcomes from a certain number of power levels, where this number is at least equal to the total number of power levels this pair of users can choose. Hence, [22] requires a sufficiently good, although imperfect, monitoring ability of the secondary users.

In this paper, we design deviation-proof spectrum sharing policies with time-varying power levels to achieve Pareto optimal operating points that are not achievable by existing policies with constant power levels [5]–[16]. We provide a systematic design approach, which first characterizes the set of Pareto optimal operating points achievable by deviation-proof policies, and then for any operating point in this set, constructs a deviation-proof policy to achieve it. The proposed policy can be easily implemented in a distributed manner. Moreover, we prove that the proposed policy can achieve Pareto optimal operating points, even when the SUs are impatient (their discount factor are strictly smaller than and bounded away from one), and have imperfect and limited monitoring ability. Actually, they only need to observe whether the IT constraints are violated, and their observation is imperfect due to the erroneous measurements of the interference temperature. This requirement on the users’ monitoring ability is significantly relaxed compared to existing works on repeated games, which require either perfect monitoring on all the users’ power levels [20][21] or sufficiently good monitoring that can imperfectly distinguish different outcomes from sufficiently many power levels [22].

Finally, we summarize the comparison of our work with the existing works in dynamic spectrum sharing in Table I. We distinguish our work from existing works in the following categories: the power levels prescribed by the spectrum sharing policy are constant or time-varying, whether the policy can be implemented distributedly or not, whether the policy is deviation-proof or not, and what are the requirements on the secondary users’ monitoring ability. The “monitoring” category is only discussed with respect to the existing works on repeated games and does not refer to the feedback ability in general.

The rest of the paper is organized as follows. In Section II, we describe the system model for dynamic spectrum sharing. Then, in Section III, we formulate the policy design problem using repeated games. We characterize the set of Pareto optimal equilibrium payoffs in Section ?? and discusses related design and implementation issues. Simulation results are presented in
Section V. Finally, Section VI concludes the paper.

II. SYSTEM MODEL FOR DYNAMIC SPECTRUM SHARING

We consider a system with one primary user and \( N \) secondary users (see Fig 1 for an illustrating example of two secondary users). The set of (secondary) users is denoted by \( \mathcal{N} \triangleq \{1, 2, \ldots, N\} \). For simplicity, we use “secondary user” and “user” interchangeably hereafter. Each user has a transmitter and a receiver. The channel gain from user \( j \)’s transmitter to user \( i \)’s receiver is \( g_{ij} \). Each user \( i \) chooses a power level \( p_i \) from a finite set \( \mathcal{P}_i \). In other words, each user choose from discrete power levels. We assume that \( 0 \in \mathcal{P}_i \), namely user \( i \) can choose not to transmit. We define user \( i \)’s maximum transmit power as \( P_{i,\text{max}} = \max_{p_i \in \mathcal{P}_i} p_i \). The set of joint power profiles is denoted by \( \mathcal{P} = \prod_{i=1}^{N} \mathcal{P}_i \), and the joint power profile of all the users is denoted by \( \mathbf{p} = (p_1, \ldots, p_N) \in \mathcal{P} \). Let \( \mathbf{p}_{-i} \) be the power profile of all the users other than user \( i \). Each user \( i \)’s instantaneous QoS (reward) is a function of the joint power profile, namely \( u_i : \mathcal{P} \to \mathbb{R} \). The reward can be, for example, the user’s throughput defined as

\[
    u_i(\mathbf{p}) = \log_2(1 + \text{SINR}_i(\mathbf{p})) = \log_2 \left( 1 + \frac{g_{ii}p_i}{\sum_{j \neq i} g_{ij}p_j + n_i} \right). 
\]

(1)

where \( n_i \) is the noise power at user \( i \)’s receiver.

As in [12]–[16], there is a local spectrum server (LSS) serving as a mediating entity among the secondary users. The LSS has a transmitter and a receiver and can measure the interference temperature, but it cannot control the actions of the autonomous secondary users. The LSS could be a device deployed by the primary user or simply the primary user itself, if the primary user manages by itself the spectrum leased to the secondary users. Even when the primary user is the LSS, it is beneficial to consider the LSS as a separate logical entity that performs the functionality of spectrum management. The LSS could also be a device deployed by some regulatory agency such as Federal Communications Commission (FCC), who uses it for spectrum management in that local geographic area. In both cases, the LSS aims to improve the spectrum efficiency and ensure fairness, e.g. the sum throughput of all the secondary users, while ensuring that the interference power is lower than the interference temperature limit \( \bar{I} \) set by the primary user. Note also that the primary user may want to maximize the spectrum efficiency to maximize its revenue obtained from spectrum leasing.
The LSS measures the interference power at its receiver imperfectly. The measurement can be written as \[ \sum_{i \in \mathcal{N}} g_{0i} p_i + \varepsilon, \] where \( g_{0i} \) is the channel gain from user \( i \)'s transmitter to the LSS’s receiver, and \( \varepsilon \) is the additive measurement error. We assume that the measurement error has zero mean and a probability distribution function (p.d.f.) \( f_\varepsilon \) known to the LSS. When the measurement \( \sum_{i \in \mathcal{N}} g_{0i} p_i + \varepsilon \) exceeds the interference temperature limit \( \bar{I} \), the LSS will broadcast a distress signal to all the secondary users. Due to measurement errors, there may be false alarms. To reduce the false alarm probability and the frequency of broadcasting distress signals, the LSS may set an intermediate interference temperature limit \( I \) lower than \( \bar{I} \). Hence, the interference temperature constraint imposed can be written as

\[
\sum_{i \in \mathcal{N}} g_{0i} p_i \leq I, \tag{2}
\]

and the false alarm probability can be defined as

\[
\Gamma(p) = \text{Prob} \left( \sum_{i \in \mathcal{N}} g_{0i} p_i + \varepsilon > \bar{I} \mid \sum_{i \in \mathcal{N}} g_{0i} p_i \leq \bar{I} \right). \tag{3}
\]

A user’s reward is affected by the multi-user interference \( \sum_{j \neq i} g_{ij} p_j \), which is dependent on the cross channel gains among different users. When the multi-user interference is weak due to small cross channel gains, power control becomes less important, since one user’s power level does not affect the others’ rewards. Hence, in this paper, we focus on the more interesting scenario when the multi-user interference is strong and power control is essential for efficient interference management. We quantify the strength of multi-user interference as follows. First, we write \( \tilde{p}^i = (\tilde{p}^i_1, \ldots, \tilde{p}^i_N) \) as the action profile that maximizes user \( i \)'s reward, namely

\[
\tilde{p}^i = \arg\max_{p \in \mathcal{P}} u_i(p), \text{ subject to } \sum_{i \in \mathcal{N}} g_{0i} p_i \leq I. \tag{4}
\]

Since \( u_i \) is increasing in \( p_i \) and decreasing in \( p_j, \forall j \neq i \), \( \tilde{p}^i \) can be written explicitly as

\[
\tilde{p}^i = \min\{P_{i}^{\text{max}}, \frac{I}{g_{0i}}\}, \quad \tilde{p}^i_j = 0, \quad \forall j \neq i. \tag{5}
\]

We define the maximum reward achievable by user \( i \) as \( \bar{v}_i \triangleq u_i(\tilde{p}^i) \). Note that \( u_j(\tilde{p}^i) = 0, \quad \forall i \in \mathcal{N}, \quad \forall j \neq i \), because a user’s reward should be zero when it does not transmit. Then, we define the scenario with strong multi-user interference as the one that satisfies the following property.

**Definition 1 (Strong Multi-user Interference):** A spectrum sharing scenario has strong multi-user interference, if the set of feasible payoffs, defined as the convex hull of the set of feasible
rewards $V^+ = \text{co}\{u(p) = (u_1(p), \ldots, u_N(p)) : p \in P\}$, has $N+1$ extremal points: $(0, \ldots, 0) \in \mathbb{R}^N$, $u(\tilde{p}^1), \ldots, u(\tilde{p}^N)$. This definition characterizes the strong interference among the users: the increase of one user's payoff comes at such an expense of the other users' payoffs that the set of feasible rewards is nonconvex. The spectrum sharing game satisfies this property when the cross channel gains among users are large [18][19]. In the extreme case of strong multi-user interference, simultaneous transmissions from different users result in packet loss, as captured in the collision model [23]–[25]. According to this definition, the set of feasible payoffs can be written as $V^+ = \text{co}\{(0, \ldots, 0), u(\tilde{p}^1), \ldots, u(\tilde{p}^N)\}$. Moreover, its Pareto boundary is $\{v : \sum_{i=1}^N v_i/\bar{v}_i = 1, v_i \geq 0, \forall i\}$ as part of a hyperplane.

III. FORMULATION OF THE POLICY DESIGN PROBLEM

In this section, we first formulate the interaction among secondary users as a repeated game with imperfect monitoring, and define the deviation-proof spectrum sharing policy. Then, we formally define the policy design problem and outline our design framework to solve it.

A. Formulation of The Repeated Game

Similar to [5]–[16], we assume that the system parameters, such as the number of secondary users and the channel gains, remain fixed during the considered time horizon. The system is time slotted at $t = 0, 1, \ldots$. We assume that the users are synchronized as in [5]–[16]. At the beginning of time slot $t$, each secondary user $i$ chooses its action $p_t^i$, and receives the reward $u_i(p_t^i)$. The LSS obtains the measurement $\sum_{i \in N} g_0 p_t^i + \varepsilon_t$, where $\varepsilon_t$ is the realization of the error $\varepsilon$ at time slot $t$, and compare the measurement with the interference temperature limit. The set of outcomes of the comparison $Y$ has two elements, namely $Y = \{y_0, y_1\}$. The outcome $y_t$ is determined according to

$$y_t = \begin{cases} y_0, & \text{if } \sum_{i \in N} g_0 p_t^i + \varepsilon_t > \bar{I} \\ y_1, & \text{otherwise} \end{cases}.$$  

(6)

$^2$The extremal points of a convex set are those that are not convex combinations of other points in the set.
We write the conditional probability distribution of the outcome $y$ given the action profile $p$ as $\rho(y|p)$, which can be calculated as

$$\rho(y_1|p) = \int_{x \in I_g} \int_{y \in \mathcal{Y}} f(x)dx, \ \rho(y_0|p) = 1 - \rho(y_1|a).$$ (7)

At the end of time slot $t$, the LSS sends the distress signal if the outcome $y^t = y_0$. Note that the LSS does not send signals when the outcome is $y_1$, and the users know that the interference temperature constraint is satisfied by default when they do not receive the distress signal. Each user keeps track of all the actions it has taken and all the outcomes. Hence, the history of user $i$ up to time slot $t \geq 1$ is $h_i^t = \{p_i^0, y_i^0; \ldots; p_i^{t-1}, y_i^{t-1}\} \in (\mathcal{P}_i \times \mathcal{Y})^t$, and that at time slot 0 is $h_i^0 = \emptyset$. The history of user $i$ contains private information about user $i$’s actions that is unknown to the other users; in contrast, we define the public history of all the users up to $t$, $h^t$, as the collection of outcomes, namely $h^t = \{y^0; \ldots; y^{t-1}\} \in Y^t$ for $t \geq 1$ and $h^0 = \emptyset$. Note that in repeated games with perfect monitoring [20][21], the outcome available to each user at time slot $t$ is exactly the action profile taken by all the users, i.e. $y^t = p^i$. In this paper, since the users can only observe an outcome that imperfectly indicates whether the IT constraint is violated, the interaction is modeled as a repeated game with imperfect monitoring.

Each user $i$’s strategy $\sigma_i$ is a mapping from the set of all possible histories $\mathcal{H}_i \triangleq \cup_{t=0}^\infty (\mathcal{P}_i \times \mathcal{Y})^t$ to its action set $\mathcal{P}_i$. Then the action taken by user $i$ at time slot $t$ is determined by $p_i^t = \sigma_i(h_i^t)$. The spectrum sharing policy is a collection of all the users’ strategies, $\sigma = (\sigma_1, \ldots, \sigma_N)$, which is also called joint strategy profile of all the users. The users are selfish and maximize their own payoffs, where the payoff is defined as the expected discounted average reward per time slot. Assuming the same discount factor $\delta \in [0, 1)$ for all the users, user $i$’s payoff can be written as

$$U_i(\sigma) = (1 - \delta) \left[ u_i(p_i^0) + \sum_{t=1}^\infty \delta^t \sum_{y^t \in \mathcal{Y}} \rho(y^t|p_i^{t-1})u_i(p_i^t) \right],$$

where $p_i^0 = (p_i^0; \ldots; p_i^N)$ is determined by $p_i^0 = \sigma_i(\emptyset)$ for all $i$, and $p_i^t$ for $t \geq 1$ is determined by $p_i^t = \sigma_i(h_i^t) = \sigma_i(h_i^{t-1}; p_i^{t-1}, y_i^{t-1})$ for all $i$.

In this paper, we restrict our attention to public strategies defined as follows [26, Definition 7.1.1]

**Definition 2 (Public Strategies):** A strategy $\sigma_i$ is a public strategy if it only depends on the public history $h^t$ for any $t \geq 0$: for any $h_i^t, \hat{h}_i^t \in \mathcal{H}_i$ that satisfy $y^\tau = \hat{y}^\tau$ for all $0 \leq \tau \leq t - 1$, we have $\sigma_i(h_i^t) = \sigma_i(\hat{h}_i^t)$. 

January 17, 2012
Due to realization equivalence principle [26, Lemma 7.1.2], we lose nothing by only considering public strategies, in terms of the achievable Pareto optimal operating points. Moreover, it is easier for the users to play public strategies, which require only public histories to determine the action. In the rest of this paper, we mean “public strategy” by “strategy” and define a strategy as \( \sigma_i : \bigcup_{t=0}^{\infty} Y^t \to A_i \). It appears that a user has to store all the past outcomes to determine which action to choose. However, we will see in Section IV that a user only needs a memory of a fixed size to implement the proposed policy.

We define the deviation-proof policy as the perfect public equilibrium (PPE) of the game. The perfect public equilibrium prescribes a strategy profile \( \sigma \) from which no user has incentive to deviate at any time slot and after any given history, and thus can be considered as a deviation-proof policy. It is normally more strict than Nash equilibrium, because it requires that the users have no incentive to deviate at any given history, while Nash equilibrium only guarantees this at the histories that possibly arise from the equilibrium strategy. These two concepts are equivalent, when the distribution of the outcome has full support given any action profile (\( \rho(y|p) > 0 \) for any \( y \) and \( p \)), because any history may arise from the equilibrium strategy. We can also consider PPE in repeated games with imperfect monitoring as the counterpart of subgame perfect equilibrium defined in repeated games with perfect monitoring [26].

Before the definition of PPE, we introduce the concept of continuation strategy: user \( i \)'s continuation strategy induced by any history \( h^t \in H \), denoted \( \sigma_i|_{h^t} \), is defined by \( \sigma_i|_{h^t}(h^\tau) = \sigma_i(h^t h^\tau) \), \( \forall h^\tau \in H \), where \( h^t h^\tau \) is the concatenation of the history \( h^t \) followed by the history \( h^\tau \). By convention, we denote \( \sigma|_{h^t} \) and \( \sigma_{-i}|_{h^t} \) the continuation strategy profile of all the users and the continuation strategy profile of all the users other than user \( i \), induced by \( h^t \), respectively. Then the perfect public equilibrium is defined as follows [26, Definition 7.1.2]

**Definition 3 (Perfect Public Equilibrium):** A public strategy profile \( \sigma \) is a perfect public equilibrium if for any public history \( h^t \in H \), the induced continuation strategy \( \sigma|_{h^t} \) is a Nash equilibrium of the game, namely

\[
U_i(\sigma|_{h^t}) \geq U_i(\sigma'_i|_{h^t}, \sigma_{-i}|_{h^t}), \quad \text{for all } \sigma'_i \text{ and for all } i \in N.
\]

We define the equilibrium payoff as a vector of payoffs \( v = (U_1(\sigma), \ldots, U_N(\sigma)) \) achieved at the equilibrium.


B. The Policy Design Problem

The primary user or the regulatory agency aims to maximize an objective function defined on the secondary users’ payoffs, \( W(U_1(\sigma), \ldots, U_N(\sigma)) \). An example of the objective function is the weighted sum payoff \( \sum_{i=1}^{N} w_i U_i \), where \( \{w_i\}_{i=1}^{N} \) are the weights satisfying \( w_i \in [0, 1], \forall i \) and \( \sum_{i=1}^{N} w_i = 1 \). The primary user (respectively, the regulatory agency) maximizes the objective function for the revenue (the spectrum efficiency), while maintaining the interference temperature constraint (2). To reduce the transmission of distress signals, a constraint on the false alarm probability may also be imposed as \( \Gamma(p) \leq \bar{\Gamma} \), where \( \bar{\Gamma} \) is the maximum false alarm probability allowed. At the maximum of the welfare function, some users may have extremely low payoffs. To avoid this, a minimum payoff guarantee \( \gamma_i \geq 0 \) is imposed for each user \( i \). To sum up, we can formally define the policy design problem as follows

\[
\max_{\sigma} W(U_1(\sigma), \ldots, U_N(\sigma)) \tag{9}
\]

\[
s.t. \quad \sigma \text{ is public perfect equilibrium, } \quad \sum_{i \in \mathcal{N}} g_{0i} \cdot \sigma_i(h^t) \leq I, \forall t, \forall h^t \in Y^t, \\
\Gamma(\sigma(h^t)) \leq \bar{\Gamma}, \forall t, \forall h^t \in Y^t, \\
U_i(\sigma) \geq \gamma_i, \forall i \in \mathcal{N}.
\]

We outline the design framework we proposed to solve the policy design problem, which is a procedure consisting of three phases as illustrated in Fig. 2. First, the LSS identifies the set of Pareto optimal equilibrium payoffs, which requires some information exchange between the LSS and SU’s in Phase I. We will describe what information to exchange explicitly in Section IV. Once the set is identified, the LSS finds the optimal equilibrium payoff that maximizes the objective function \( W \) under the constraints in Phase II, which is easy given the analytical characterization of the feasible set of the design problem. Finally in Phase III, the secondary users follow the deviation-proof spectrum sharing policy in a distributed way to achieve the optimal operating point.

Note that the information exchange in Phase I is necessary for the system to determine and operate at a Pareto optimal operating point. A similar information exchange phase is proposed in [20]. The information exchange phase can be also considered as a substitute for the convergence process needed by the algorithms in [5]–[8]. Since the secondary users determine
and distributively implement the policy immediately after obtaining the information in the first two phases, the information exchange phase is advantageous in that its duration and the amount of information to exchange are predetermined. On the other hand, the convergence time in the algorithms in [5]–[8] is not bounded.

IV. SOLVING THE POLICY DESIGN PROBLEM

In this section, we solve the policy design problem following the three phases outlined in Fig. 2.

A. Characterizing The Set of Pareto Optimal Equilibrium Payoffs

The first step in solving the design problem (9) is to characterize the set of Pareto optimal equilibrium payoffs for the dynamic spectrum sharing game. In particular, we are interested in the case when the secondary users are impatient (their discount factor is fixed and smaller than 1), as opposed to the case when the secondary users are patient (their discount factor can be arbitrarily close to 1) in [20][22]. The characterization of Pareto optimal equilibrium payoffs with impatient users is provided in [21] for repeated games with perfect monitoring. Our result in Theorem 1 is the first one that analytically characterizes the set of Pareto optimal equilibrium payoffs for repeated games with imperfect monitoring and impatient users. This characterization requires some information exchange between the LSS and the SU’s in Phase I, which will be described after we state Theorem 1.

First, for the spectrum sharing games with strong multi-user interference, we know from Definition 1 that the set of feasible payoffs can be written as \( \mathcal{V}^\dagger = \text{co}\{(0, \ldots, 0), u(\hat{p}^1), \ldots, u(\hat{p}^N)\} \).

Moreover, its Pareto boundary is \( \{v : \sum_{i=1}^{N} v_i/\hat{v}_i = 1, \; v_i \geq 0, \; \forall i\} \).

Given the system parameters, we can determine which portion of the Pareto boundary can be achieved as equilibrium payoffs, i.e. operating points that can be achieved by deviation-proof policies.

Before stating Theorem 1, we define the benefit of deviation as follows.

**Definition 4 (Benefit of Deviation):** We define user \( j \)’s benefit of deviation from user \( i \)’s reward maximizing action profile \( \hat{p}^i \) as

\[
 b_{ij} = \sup_{p_j \in \mathcal{P}_j, p_j \neq \hat{p}_j} \frac{\rho(y_0|\hat{p}^i) - \rho(y_0|p_j, \hat{p}^i_{-j})}{u_j(p_j, \hat{p}^i_{-j})/\hat{v}_j}, \quad \forall i \in \mathcal{N}, \; \forall j \neq i. \quad (10)
\]
As we will see in Theorem 1, if we want to achieve any Pareto optimal equilibrium payoff, the benefits of deviation $b_{ij}$ for all $i$ and $j \neq i$ must be strictly smaller than 0. Since the reward $u_j$ is always larger than 0, $b_{ij} < 0$ is equivalent to $\rho(y_0|p_j, \tilde{p}_{-j}) > \rho(y_0|\tilde{p})$ for all $p_j \neq \tilde{p}_j$, which means that the probability of the outcome $y_0$ that indicates deviation increases when deviation happens. This guarantees that any deviation from $\tilde{p}$ by user $j$ ($\forall j \neq i$) can be statistically identified. We can observe that the benefit of deviation is also related to the reward user $j$ obtains by deviation, $u_j(p_j, \tilde{p}_{-j})$. If the reward obtained by deviation is smaller, the benefit of deviation is smaller.

Now we state Theorem 1, which characterizes the set of Pareto optimal equilibrium payoffs.

**Theorem 1:** There exist Pareto optimal equilibrium payoffs, if and only if the following two sets of conditions are satisfied:

- **Condition 1:** benefit of deviation $b_{ij} < 0$ for all $i$ and $j \neq i$.
- **Condition 2:** for all $i \in N$, we have

$$
\bar{v}_i - u_i(p_i, \tilde{p}_{-i}) + \bar{v}_i \sum_{j \neq i} \frac{\rho(y_0|\tilde{p}_j) - \rho(y_0|p_i, \tilde{p}_i)}{-b_{ij}} \geq 0, \forall p_i \in \mathcal{P}_i.
$$

(11)

When the above conditions are satisfied, the set of Pareto optimal equilibrium payoffs is

$$
\mathcal{B}_\mu = \left\{ v : \sum_{i=1}^{N} \frac{v_i}{\bar{v}_i} = 1, \frac{v_i}{\bar{v}_i} \geq \mu_i, \forall i \in N \right\},
$$

(12)

where

$$
\mu_i \triangleq \max_{j \neq i} \frac{1 - \rho(y_0|\tilde{p}_j)}{-b_{ij}},
$$

(13)

and any payoff in $\mathcal{B}_\mu$ can be achieved if the discount factor $\delta$ satisfies

$$
\delta \geq \delta \triangleq \frac{1}{1 + \frac{1 - \sum_{i \in N} \mu_i}{N - 1 + \sum_{i \in N} \sum_{j \neq i} (\rho(y_0|\tilde{p})/b_{ij})}}.
$$

(14)

**Proof:** Due to space limit, we only provide the outline of the proof here. Please refer to [28, Appendix A] for the complete proof.

The proof heavily replies on the concept of self-generating sets. Simply put, a self-generating set, associated with a discount factor, is a set in which every payoff is an PPE payoff under the associated discount factor [26, Section 2.5.1]. Any self-generating set has a minimum discount factor associated with it; any discount factor that is larger than the minimum one can be associated with that self-generating set. The idea of the proof is to find the largest self-generating set and the
associated minimum discount factor. Since we focus on the Pareto optimal equilibrium payoffs, we restrict to the self-generating sets on the Pareto boundary. This restriction allows us to obtain the analytical expression of the largest self-generating set, which is the set of Pareto optimal equilibrium payoffs $B_{\mu}$. Meanwhile, the sufficient and necessary conditions for $B_{\mu}$ to be self-generating are obtained.

Theorem 1 first provides the sufficient and necessary conditions for the existence of Pareto optimal equilibrium payoffs. Condition 1 (respectively, Condition 2) ensures that at action profile $\tilde{p}^i$, user $j$ for any $j \neq i$ (respectively, user $i$) has no incentive to deviate. If any of the above conditions is violated, no Pareto optimal equilibrium payoff can be achieved. When the conditions are satisfied, Theorem 1 gives us the set of Pareto optimal equilibrium payoffs $B_{\mu}$, under given system parameters. We can choose any payoff in $B_{\mu}$ as the deviation-proof operating point.

We can determine the maximum level of impatience the users can have in order to achieve any payoff in $B_{\mu}$.

**Information Exchange In Phase I:** We describe the information exchange phase for the LSS to identify the set of Pareto optimal equilibrium payoffs. The key quantities needed are $\{\rho(y_0 | \tilde{p}^i)\}_{i=1}^N$, $\{\rho(y_0 | p_j, \tilde{p}^i_j)\}_{i \neq j}$, and $\{b_{ij}\}_{i \neq j}$, which can be obtained at the end of the information exchange phase. We list the information obtained by the LSS and SU’s during this phase in Table II.

**B. Determining The Optimal Operating Point**

Since we have identified the set of Pareto optimal equilibrium payoffs $B_{\mu}$, the problem of find the optimal operating point that solves the policy design problem can be written as

$$\max_v W(v_1, \ldots, v_N)$$

s.t. \hspace{1em} $$(v_1/\bar{v}_1, \ldots, v_N/\bar{v}_N) \in B_{\mu},$$

\hspace{1em} $v_i \geq \gamma_i, \ \forall i \in \mathcal{N}.$

The linear constraints in the above problem can be further simplified as

$$v_i \geq \max\{\mu_i \cdot \bar{v}_i, \gamma_i\}, \ \forall i \in \mathcal{N}. \hspace{1em} (16)$$
Hence, we get the sufficient and necessary conditions under which the optimization problem (15) is feasible as follows

\[
\sum_{i=1}^{N} \max \{ \mu_i \cdot \bar{v}_i, \gamma_i \} \leq 1.
\]  

(17)

The optimization problem (15) is easy to solve when \( W \) is a convex function in \( (v_1, \ldots, v_N) \). For example, if the objective function is the weighted sum of the users’ payoffs, namely \( W = \sum_{i=1}^{N} w_i v_i \), the solution can be obtained analytically as

\[
v_i^* = \begin{cases}
(1 - \sum_{j \neq i} \max \{ \mu_j, \gamma_j / \bar{v}_j \}) \cdot \bar{v}_i, & \text{if } i = \arg \max_{j \in N} w_j \bar{v}_j \\
\max \{ \mu_i, \gamma_i / \bar{v}_i \} \cdot \bar{v}_i, & \text{otherwise}
\end{cases}.
\]  

(18)

After obtaining the optimal operating point \( v^* \), the LSS will send the target payoff \( v_i^* \) for the corresponding user \( i \).

C. Constructing The Deviation-Proof Policy

For a given payoff profile \( v^* \in B_\mu \), we can construct the deviation-proof policy, i.e. the equilibrium strategy, that achieves the payoff \( v^* \). According to Definition 1, any payoff \( v^* \in B_\mu \) should be achieved by alternating among the \( N \) non-zero extremal points, \( u(\bar{p}^1), \ldots, u(\bar{p}^N) \). Hence, the deviation-proof policy \( \sigma^* \) satisfies \( \sigma^*(h^t) \in \{ \bar{p}^1, \ldots, \bar{p}^N \} \) for any \( t \geq 0 \) and for any public history \( h^t \in Y^t \). Since only user \( i \) consumes the resources at the action profile \( \bar{p}^i \), the deviation-proof policy can also be regarded as a scheduling in a TDMA fashion. By judiciously deciding which user can transmit in each time slot, each user \( i \) receives a discounted expected average payoff \( v_i^* \) and has no incentive to deviate from the policy. The deviation-proof policy can be implemented by each user in a distributed manner. The algorithm run by user \( i \) is described in the algorithm in Table III.

At each time slot \( t \), each SU \( i \) calculates and broadcasts the index

\[
\alpha_i(t) = \frac{v_i(t)/\bar{v}_i - \mu}{\bar{v}_i \sum_{j \neq i}(w_{ij} + v_j(t)/\bar{v}_j)}.
\]  

(19)

The user that has the largest index will transmit in time slot \( t \). Then all the users calculate \( v_i(t+1) \) for the index to broadcast at the next time slot. Except from the indices from other users, all information for each user to run the algorithm in Table III is available locally. Note that the overhead of broadcasting is necessary to achieve a Pareto optimal operating point. Distributed
algorithms with limited message passing (broadcasting) have been proposed in [5]–[8][27] for spectrum sharing policies with constant power levels.

Theorem 2 ensures that if all the users run the algorithm in Table III locally, they will achieve the optimal operating point $v^*$, and will have no incentive to deviate.

Theorem 2: For any target payoff $v^* \in B_\mu$, and any discount factor $\delta \geq \delta_0$, the strategy generated by each user running the algorithm in Table III is PPE and achieves $v^*$.

Proof: Due to space limit, we only provide the outline of the proof here. Please refer to [28, Appendix B] for the complete proof. The key to the proof is to demonstrate that all the payoffs $v(t), \forall t \geq 0$ generated in the algorithm in Table III are in the self-generating set (the set of Pareto optimal equilibrium payoffs) $B_\mu$. $lacksquare$

V. SIMULATION RESULTS

In this section, we demonstrate the performance gain of our spectrum sharing policy over existing policies, and validate our theoretical analysis through numerical results. Throughout this section, we use the following system parameters by default unless we change some of them explicitly. The noise powers at all the users’ receivers are normalized as 0 dB. The maximum transmit powers of all the users are $P_i = 10$ dB, $\forall i$. For simplicity, we assume that the direct channel gains have the same distribution $g_{ii} \sim \mathcal{CN}(0, 1), \forall i$, and the cross channel gains have the same distribution $g_{ij} \sim \mathcal{CN}(0, \beta), \forall i \neq j$, where $\beta$ is defined as the cross interference level. The channel gain from each SU to the LSS also satisfies $g_{0i} \sim \mathcal{CN}(0, 1), \forall i$. The interference temperature limit is $\bar{I} = 10$ dB. The measurement error $\varepsilon$ is Gaussian distributed with zeros mean and variance 0.1. The maximum false alarm probability is $\bar{\Gamma} = 10\%$. The objective function is the average payoff, i.e. $W = \sum_{i=1}^{N} \frac{1}{N} U_i$, where $U_i$ is the expected discounted average throughput of user $i$. The minimum payoff guarantee is 10% of the maximum achievable payoff, i.e. $\gamma_i = 0.1 \cdot \bar{v}_i, \forall i$.

A. Performance Evaluation

1) Comparison with policies with constant power levels: We first compare the performance of the proposed policy with that of the optimal policy with constant power levels. Specifically, we define the optimal policy with constant power levels (or “the optimal stationary policy”) as
the policy that satisfies $\sigma(h^t) = p^*$ for all $t \geq 0$ and for all $h^t \in Y^t$, where the optimal power level $p^*$ is the solution to the following problem

$$\max_{p} W(u_1(p), \ldots, u_N(p))$$

s.t. $\sum_{i \in N} g_{0i} \cdot p_i \leq I$,

$\Gamma(p) \leq \bar{\Gamma}$,

$u_i(p) \geq \gamma_i, \forall i \in N$.

Note that we drop the incentive constraint that $\sigma$ is PPE from the original policy design problem. Hence, the performance of the optimal stationary policy is the best that can be achieved by existing stationary policies [6]–[10], and is an upper bound for the deviation-proof stationary policies [5][11].

In Fig. 3, we compare the performance of the proposed policy and that of the optimal stationary policy under different cross interference levels and different numbers of users. As expected, the proposed policy outperforms the optimal stationary policy in medium to high cross interference levels, which approximately correspond to the cases when $\beta \geq 1$. In the cases of high cross interference levels ($\beta \geq 2$) and many users ($N = 5$), the stationary policy fails to meet the minimum payoff guarantees due to strong interference (indicated by zero average throughput in the figure). On the other hand, the desirable feature of the proposed policy is that the average throughput does not decrease with the increase of the cross interference level, because users transmit in a TDMA fashion.

Note that the proposed policy has zero average throughput when the cross interference level is very small. This is because it cannot be deviation-proof in this scenario. When the interference level is very small, user $j$ can deviate from $\tilde{p}_i$ and receives a high reward $u_j(p_j, \tilde{p}_{-j})$ because the interference from user $i$, $h_{ji} \tilde{p}_i$, is small. Hence, the benefit of deviation $b_{ij}$ is large, and the deviation is inevitable. This observation could lead to an efficient way for the LSS to check the cross interference level without knowing the actual channel gains of the users. If the proposed policy is infeasible, the LSS knows that the cross interference level is low, and can switch to stationary policies.

2) Comparison with “punish-forgive” policies proposed for perfect monitoring: We also compare the proposed policy with existing policies that work for repeated games with perfect
monitoring [20][21]. Specifically, we consider the “punish-forgive” policy in [20][21], which requires users to switch to the punishment phase of $L$ time slots once a deviation is detected. In the punishment phase, all the users transmit at the maximum power levels to create high interference to the deviator. A special case of the punish-forgive policy when the punishment length $L = \infty$ is the celebrated “grim-trigger” strategy in game theory literature [26].

As discussed before, the punish-forgive policy works well if the users can perfectly monitor the power levels of all the users, because the punishment serves as a threat to deter the users from deviating, and it will never happen in perfect monitoring case if no user deviates. However, when the users have imperfect monitoring ability, the punishment will happen with some positive probability, which decreases all the users’ expected payoffs.

Fig. 4 compares the performance of the proposed policy with the punish-forgive policy with different punishment lengths. We can see that the longer the punishment length, the larger the performance loss due to mistakenly triggered punishments in punish-forgive policies. Note that although a shorter punishment length is preferable in terms of performance loss, it requires higher patience of the users (a high discount factor) for the punish-forgive policy with a shorter punishment length to be deviation-proof (see [20, Fig. 5] for an illustration).

Fig. 5 shows that the proposed policy outperforms the punish-forgive policies under different variances of measurement errors and different false alarm probabilities. For each combination of the error variance and the false alarm probability, we choose the punish-forgive policy with the optimal punishment length. The performance of punish-forgive polices degrades with the increase of the error variance and the false alarm probability, because of the increasing probability of mistakenly triggered punishments. Some interesting observation on how the performance of the proposed policy changes with the error variance and the false alarm probability is explained in details in the following subsections.

**B. Impacts of Variances of Measurement Errors**

Fig. 6 shows that with the increase of the variance of measurement errors, the average throughput decreases, and the users’ patience (the discount factor) required to achieve Pareto

---

3 Note that all the users transmitting at the maximum power levels may violate the IT constraint, which was not considered in [20][11]. To maintain the efficiency of the punish-forgive policy, we allow the violation of the IT constraint in the punishment phase. Note that the IT constraint is never violated in the proposed policy.
optimal equilibrium payoffs increases. First, when the error variance increases, the intermediate IT limit $I$ must decrease to maintain the constraint on the false alarm probability. The decrease of $I$ leads to the decrease of users’ maximum transmit power levels allowed, which results in the decrease of the average throughput. Another impact of the decrease in the error variance is that $\rho(y_0|p_j, \tilde{P}^i_{j-1}) = \int_{x > I - h_0 p_j - h_0 \tilde{p}_i} f_\varepsilon(x) dx$ decreases, which leads to the increase of benefit of deviation $b_{ij}$. Hence, the minimum discount factor $\delta$ increases according to Theorem 1.

C. Impacts of Constraints on The False Alarm Probability

Fig. 7 shows that with the increase of the false alarm probability limit $\bar{\Gamma}$, both the average throughput and the users’ patience (the discount factor) required to achieve Pareto optimal equilibrium payoffs increase. First, an increased false alarm probability limit, the intermediate IT limit $I$ can increase, which leads to an increase of the users’ maximum transmit power levels and thus the increase of the users’ throughput. Meanwhile, since

$$\rho(y_0|\tilde{P}^i) - \rho(y_0|p_j, \tilde{P}^i_{j-1}) = \int_{x > I - h_0 \tilde{p}_i} f_\varepsilon(x) dx - \int_{x > I - h_0 p_j - h_0 \tilde{p}_i} f_\varepsilon(x) dx = \int_{x > I - I} f_\varepsilon(x) dx - \int_{x > I - I - h_0 p_j} f_\varepsilon(x) dx = -\int_{I - h_0 p_j}^{I - I} f_\varepsilon(x) dx$$

increases when $I$ increases, the benefit of deviation $b_{ij}$ increases. This leads to an increase of the minimum discount factor.

This observation indicates an interesting design tradeoff. On one hand, the LSS wants to reduce the false alarm probability and the overhead of sending distress signals associated with false alarms. This will also relax the requirement on users’ patience. On the other hand, the LSS wants to increase the false alarm probability in order to increase the average throughput, such that the spectrum efficiency or the revenue obtained from users’ transmission can increase. Our theoretical results characterize such a tradeoff, which can be used by the LSS to choose the optimal intermediate IT limit $I$.

VI. CONCLUSION

In this paper, we studied power control in dynamic spectrum sharing among secondary users under the interference temperature constraint, and proposed a dynamic spectrum sharing policy
that allows secondary users to transmit in a TDMA fashion. The proposed policy can achieve Pareto optimal operating points in the convex hull of the nonconvex feasible QoS region, which are not achievable under existing dynamic spectrum sharing policies in which users transmit at constant power levels simultaneously. The proposed policy is amenable to distributed implementation and is deviation-proof, in that the secondary users are in their self-interests, i.e. maximizing the expected discounted average QoS, to follow the policy. The proposed policy can achieve Pareto optimality even with limited and imperfect monitoring, namely the secondary users only observe distress signals that erroneously indicate the violation of the interference temperature constraint. Simulation results validate our analytical results on the policy design and demonstrate the performance gains enabled by the proposed policy.

REFERENCES


TABLE I
COMPARISON WITH RELATED WORKS IN DYNAMIC SPECTRUM SHARING.

<table>
<thead>
<tr>
<th></th>
<th>[9][10]</th>
<th>[6]–[8] [12]–[16]</th>
<th>[5][11]</th>
<th>[20][21]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power levels</td>
<td>Constant</td>
<td>Constant</td>
<td>Constant</td>
<td>Time-varying</td>
<td>Time-varying</td>
</tr>
<tr>
<td>Distributed</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Deviation-proof</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Monitoring</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Perfect</td>
<td>Imperfect</td>
</tr>
</tbody>
</table>

Fig. 1. An example system model with two secondary users. The solid line represents a link for data transmission, and the dashed line indicate a link for control signals. The channel gains for the corresponding data link are written in the figure. The primary user (PU) specifies the interference temperature (IT) limit to the local spectrum server (LSS). The LSS sets the intermediate IT limit to the secondary users and send distress signals if the estimated interference power exceeds the IT limit.
Fig. 2. An illustration of the design framework. In Phase I, the local spectrum server (LSS) and the secondary users (SU) exchange information. Based on the information exchanged, LSS determines the optimal operating point in Phase II, and tells the payoff values at the operating points to corresponding SU’s. In Phase III, SU’s implement the spectrum sharing policy in a distributed way.

### TABLE II

**INFORMATION OBTAINED BY THE LSS AND SU’S DURING THE INFORMATION EXCHANGE PHASE.**

<table>
<thead>
<tr>
<th>Events</th>
<th>Information obtained by LSS</th>
<th>Information obtained by SU $i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization</td>
<td>$I$ (obtained according to $I$, $\bar{E}$)</td>
<td>$P_i$ (known initially)</td>
</tr>
<tr>
<td>Each SU $i$ transmits at $P_i^\text{max}$</td>
<td>$(h_0, P_i^\text{max})_{i=1}^N, {\rho(y_0</td>
<td>\bar{p}^i)}_{i=1}^N$</td>
</tr>
<tr>
<td>LSS broadcasts ${\rho(y_0</td>
<td>\bar{p}^i)}_{i=1}^N$, min${1, \frac{\bar{I}}{P_i^\text{max}}}, \forall i$</td>
<td>$\bar{p}_i^j, {\rho(y_0</td>
</tr>
<tr>
<td>Each SU $i$ transmits at $\forall p_i \in P_i \setminus {0, P_i^\text{max}}$</td>
<td>${\rho(y_0</td>
<td>\bar{p}^i) - \rho(y_0</td>
</tr>
<tr>
<td>LSS broadcasts ${\rho(y_0</td>
<td>\bar{p}^i) - \rho(y_0</td>
<td>p_i, \bar{p}<em>{-i}^j)}</em>{j\neq i}$</td>
</tr>
<tr>
<td>Each SU $i$ broadcasts ${b_{ki}}_{k\neq i}$</td>
<td>$b_{ij}, \forall i, \forall j \neq i$</td>
<td>$b_{ij}, \forall i, \forall j \neq i$</td>
</tr>
<tr>
<td>Each SU $i$ transmits $\bar{v}_i$ to LSS</td>
<td>${\bar{v}<em>i}</em>{i=1}^N$</td>
<td>${\bar{v}<em>i}</em>{i=1}^N$</td>
</tr>
</tbody>
</table>
### TABLE III

**THE ALGORITHM RUN BY USER $i$.**

**Require:** The target payoff $v^*_i$ obtained from the LSS, the discount factor $\delta \geq \delta$

**Initialization:** Set $t = 0$, $v_i(0) = v^*_i$.

**Repeat**

1. Calculates the index $\alpha_i(t) = \frac{v_i(t)/\bar{v}_i - \mu_{i}}{1 - v_i(t)/\bar{v}_i + \sum_{j \neq i}(-\rho(y_0|\tilde{p}_i^*))/b_{ij}}$
2. Broadcasts the index $\alpha_i(t)$ to the other SUs
3. Receives $\alpha_j(t)$, $\forall j \neq i$ from the other SUs
4. **If** $i = i^* \triangleq \arg \max_{j \in N} \alpha_j(t)$ **Then**
   - Transmits at the power level $\tilde{p}_i^*$
   - **If** No Distress Signal Received At Time Slot $t$ **Then**
     
     $v_i(t+1) = v_i(t) - \left(1 - \frac{1}{2}\right) \cdot \left(1 - \frac{v_i(t)}{\bar{v}_i} \right) + \sum_{j \neq i} \frac{\rho(y_0|\tilde{p}_i^*)}{b_{ij}} \cdot \bar{v}_i$

   **Else**
   
   $v_i(t+1) = v_i(t) - \left(1 - \frac{1}{2}\right) \cdot \left(1 - \frac{1}{2} \cdot \frac{v_i(t)}{\bar{v}_i} \right) - \sum_{j \neq i} \frac{\rho(y_1|\tilde{p}_i^*)}{b_{ij}} \cdot \bar{v}_i$

**Endif**

**Else**

**If** No Distress Signal Received At Time Slot $t$ **Then**

$v_i(t+1) = \left(1 - \frac{1}{2}\right) \cdot \frac{\rho(y_0|\tilde{p}_i^*)}{b_{i^*i}} \cdot \bar{v}_i + \frac{1}{2} \cdot v_i(t)$

**Else**

$v_i(t+1) = \left(1 - \frac{1}{2}\right) \cdot \frac{\rho(y_1|\tilde{p}_i^*)}{b_{i^*i}} \cdot \bar{v}_i + \frac{1}{2} \cdot v_i(t)$

**Endif**

**Endif**

$t \leftarrow t + 1$

**Until** $v_i(t) = v^*_i$
Fig. 3. Performance comparison of the proposed policy and the optimal policy with constant power levels (‘stationary’ in the legend) under different numbers of users and different cross interference levels. A zero average throughput indicates that there exists no feasible policy that satisfies all the constraints in the policy design problem.

Fig. 4. Performance comparison of the proposed policy and the punish-forgive policy under different punishment lengths.
Fig. 5. Performance comparison of the proposed policy and the punish-forgive policy with the optimal punishment length under different error variances and different false alarm probabilities.

Fig. 6. Properties of the proposed policy: variance of the measurement error versus the average throughput and the discount factor.
Fig. 7. Properties of the proposed policy: the false alarm probability versus the average throughput and the discount factor.