Coordinating Secondary-User Behaviors for Inelastic Traffic Reward Maximization in Large-Scale DSA Networks

Bechir Hamdaoui, MohammadJavad NoroozOliaee, Kagan Tumer, and Ammar Rayes[†]

Oregon State University, Corvallis, OR 97331
{hamdaoub, noroozom}@onid.orst.edu; kagan.tumer@oregonstate.edu
[†] Cisco Systems, San Jose, CA 95134
[†] rayes@cisco.com

***Abstract*—We develop efficient coordination techniques that support inelastic traffic in large-scale distributed dynamic spectrum access (DSA) networks. By means of any learning algorithm, the proposed techniques enable DSA users to locate and exploit spectrum opportunities effectively, thereby increasing their achieved throughput (or "rewards" to be more general). Basically, learning algorithms allow DSA users to learn by interacting with the environment, and use their acquired knowledge to select the proper actions that maximize their own objectives, thereby "hopefully" maximizing their long-term cumulative received reward. However, when DSA users' objectives are not carefully coordinated, learning algorithms can lead to poor overall system performance, resulting in lesser per-user average achieved rewards. In this paper, we derive efficient objective functions that DSA users can aim to maximize, and that by doing so, users' collective behavior also leads to good overall system performance, thus maximizing each user's long-term cumulative received rewards. We show that the proposed techniques are: (i) efficient by enabling users to achieve high rewards, (ii) scalable by performing well in systems with a small as well as a large number of users, (iii) learnable by allowing users to reach up high rewards very quickly, and (iv) distributive by being implementable in a decentralized manner.**

Index Terms: distributed resource allocation and management, cooperative and coordinated learning, dynamic and opportunistic spectrum access.

I. Introduction

Federal Communications Commission (FCC)'s foreseeable approach for solving the spectrum shortage problem [1, 2] is dynamic spectrum access (DSA). Essentially, DSA improves spectrum efficiency by allowing unlicensed or secondary users (SUs) to exploit unused licensed spectrum, but in a manner that limits interference to licensed or primary users (PUs). DSA requires that SUs sense any licensed spectrum band prior to using it to check whether the band is vacant. When the band is vacant, SUs can then use it opportunistically, in that upon the return of any PUs, they must immediately vacate the band.

DSA has created tremendous research interests that resulted in numerous works on protocol and algorithm design [3–6], architectures and management strategies [7–12], and spectrum sensing techniques [13–16]. Research efforts have also been given to the development of adaptive techniques that also promote effective DSA first by constructing models that can predict the dynamics of the environment, and then, by using these models to adjust to the environment's behaviors so as to maximize the performance [17–22]. The issue, however, is that DSA gives rise to unique characteristics, making it too difficult to construct models that can predict its environment's dynamics without making assumptions about the environment itself. These assumptions are often unrealistic, leading to an inaccurate prediction of spectrum availabilities.

Learning-based techniques that do not require models but can still learn through interactions with the environment are particularly well suited to DSA, and consequently, have recently attracted the focus of many researchers [23–28]. Instead of using models, learning-based techniques rely on learning algorithms (e.g., reinforcement learners [29, 30] and evolving neuro-controllers [31, 32]) to learn from past and present interaction experience to decide what to do best in the future. In essence, learning algorithms allow SUs to learn by interacting with the environment, and use their acquired knowledge to select the proper actions that maximize their own (often selfish) objective functions, thereby "hopefully" maximizing their long-term cumulative received rewards.

However, when SUs' objective functions are not carefully coordinated, learning algorithms can lead to poor performance in terms of the SUs' long-term received rewards. In other words, when SUs aim to maximize these not so carefully designed objective functions, their collective behavior often leads to worsening each other's long-term cumulative rewards. It is, therefore, imperative that objective functions be designed carefully so that when SUs maximize them, their collective behavior does not result in worsening each other's performance.

In this paper, we develop coordination techniques that maximize the achievable rewards of SUs' inelastic traffic in large-scale DSA networks. We investigate the use of *difference objective functions*, which have been successfully applied to other system domains, such as controlling multi-robot systems [33] and regulating air traffic flow [34], and are shown to perform well in these systems. For our DSA system, we specifically derive distributed and scalable objective functions that SUs can aim to maximize, and that by doing so, SUs' collective behavior also leads to good overall system performance, thus maximizing each SU's long-term cumulative received rewards.

We consider a DSA network with several spectrum bands and a large number of SUs, all continuously seeking and using unused spectrum bands. By means of any learning algorithm, SUs can maximize the derived objective functions to ensure that they achieve high performances in terms of the long-term average received rewards. We show that the proposed objective functions (i) allow SUs to achieve high rewards, (ii) perform well in systems with a small as well as a large number of SUs, (iii) allow SUs to reach up high rewards very quickly, and (iv) are implementable in a decentralized manner by relying on local information only.

The rest of the paper is organized as follows. In Section II, we present the model, describe the motivation, and state the objective of this work. In Section III, we present our proposed objective functions. In Section IV, we derive the optimal/theoretical distribution of SUs across the available bands. We evaluate the performances of the proposed functions in Section V. Section VI discusses some practical/implementation aspects of the proposed techniques. We present the related works in Section VII. Finally, we conclude the paper in Section VIII.

II. Model, Motivation, and Objective

We consider m non-overlapping spectrum bands, where each band is associated with many PUs. We also consider a distributed DSA system, where PUs are assumed to arrive and leave at the beginning and at the end of time slots. We assume that each SU implements and uses a learning algorithm (e.g., a reinforcement learner [29, 30]) to allow it to locate and select the best available band. When a group of two or more SUs want to communicate with each other, all members of the group must first select and switch to the same spectrum band to be able to carry out a communication among them. Throughout this paper, these groups will also be referred to as *agents*.

At each time step, each agent using a band receives an amount of service that is passed to it from that band. The reward that the agent receives as result of using the DSA system is a function of the amount of service the agent receives from the band. Although the service the system offers can be perceived/quantified in various forms (e.g., data rates, reliability of the communication, signal to noise ratio, packet success rates, etc), in this work, we consider the agent's "received throughput" as the service metric. Therefore, we can safely assume that once the agent switches to a particular band, it can easily quantify the service level that it receives from using such a band by measuring the amount of throughput it receives. The methods that agents use to quantify and measure the service received as a result of using any particular band are beyond the scope of this work. Throughout, let S_j represent the total amount of service that spectrum band j offers.

A. Inelastic Traffic Model

This work studies the *inelastic traffic model*. In this model, an agent receives a constant reward if it switches to a band offering a quality-of-service (QoS) level equal to or greater than a certain required threshold Q, and receives a zero (or close to zero) reward when the offered QoS level is below the threshold. This model suits well inelastic applications, such as multimedia applications, in which receiving a QoS level less than what is required (i.e., Q) is not acceptable, thus yielding a zero (or almost zero) reward. But also, receiving a QoS level higher than what is required is not beneficial either, which explains why the reward is kept constant. Formally, the inelastic reward, $r_j(n_j(t))$ or simply $r_j(t)$, the spectrum band j contributes to any agent using it at time step t can be written as:

$$r_j(t) = \begin{cases} Q & \text{if} \quad n_j(t) \le S_j/Q \\ Qe^{-\beta\frac{n_j(t)Q-S_j}{S_j}} & \text{otherwise} \end{cases} \tag{1}$$

where $n_j(t)$ is the number of users/agents using band j at episode (time step) t, and β is a decaying factor. Note that when the number of users using band j is greater than[1] $c_j \equiv S_j/Q$, the reward decreases exponentially. This means that none of the users will be satisfied with the amount of service they receive from band j if band j contains more than c_j users.

For illustration purposes, we show in Fig. 1 the inelastic traffic reward $r_j(t)$ contributed by band j as a function of the number of users $n_j(t)$ using band j for $\beta = 20$ and $S_j/Q = 4$.

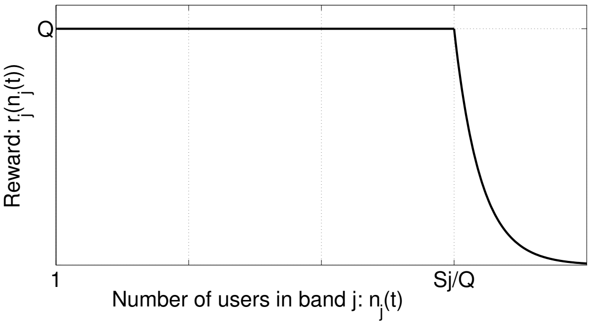


Fig. 1. Reward function: $\beta = 20$ and $S_j/Q = 4$ for all $j = 1, 2, \ldots, m$.

From the system's perspective, the system or global reward, $G(t)$, at time step t is the sum of all agents' received rewards, and can formally be expressed as

$$G(t) = \sum_{j=1}^{m} n_j(t) r_j(t) \tag{2}$$

where again m is the number of bands. The per-agent average received reward $\bar{r}(t)$ at time step t can then be written as

$$\bar{r}(t) = \frac{G(t)}{\sum_{j=1}^{m} n_j(t)} \tag{3}$$

B. Learning Algorithm

Our objective in this work is to derive distributive and scalable objective functions for SUs that are aligned with global system objective, so that when SUs (i.e., agents) aim to maximize them, they indeed lead to the maximization of the agents' long-term cumulative received rewards. Basically, by

[1] c_j here represents band j's capacity; i.e., the maximum number of users that the band can support while meeting the users' required QoS levels.

means of any learning algorithm, these objective functions will enable SUs to efficiently find and locate spectrum opportunities, thus increasing the long-term achievable rewards that each SU can receive from accessing the DSA network.

Even though the focus of this work is on the design of efficient objective functions and not on the development of learning algorithms, we choose to use throughout this work the ϵ-greedy Q-learner [29] with a discount rate of 0 and an ϵ value of 0.05 for the purpose of evaluating our proposed techniques. Each agent is then assumed to implement and rely on the Q-learner to maximize the proposed objective function. At the end of every episode, each agent selects and takes the action with the highest entry value with probability $1-\epsilon$, and selects and takes a random action among all possible actions with probability ϵ. After taking an action, the agent then computes the reward that it receives as a result of taking such an action, and uses it to update its Q-table. A table entry $Q(a)$ corresponding to action a is updated via $Q(a) \leftarrow (1-\alpha)Q(a) + \alpha u$, where α (here, the value of α is set to 0.5) is the learning rate, and u is the received reward from taking action a. All the results presented in this paper are based on this Q-learner. Readers are referred to [29] for more details on the Q-learner.

C. Motivation and Objective

The key question that arises naturally and that we address in this work is which objective function g_i should each DSA agent i aim to maximize so that its received reward is maximized? There are two intuitive choices. One possible choice for g_i is for each agent i using band j to selfishly go after the intrinsic reward r_j contributed by the band j as defined in Eq. (1); i.e., $g_i = r_j$ for each agent i using band j. A second also intuitive choice is for each agent to maximize the global (i.e., total) rewards received by all agents; i.e., $g_i = G$ for each agent i as defined in Eq. (2), hoping that maximizing the overall received rewards will eventually lead to maximizing every agent's long-term average received rewards.

For illustration purposes, we measure and show in Fig. 2 the average reward $\bar{r}(t)$ (measured and calculated via Eq. (3)) that each agent receives under each of these two private objective function choices. In this experiment, we consider a DSA network with 500 agents and 10 bands. There are two important observations that we want to make regarding the performance behaviors of these two objective functions, and that constitute the main motivation of this work. First, note that when agents aim to maximize their own intrinsic rewards (i.e., $g_i = r_j$ for each agent i using band j), the per-agent average received reward goes up quickly at first but then drops down rapidly too, and then starts to ramp up quickly and drop down rapidly again, and so on. With the intrinsic function, an agent's reward, by design, is sensitive to its own actions, which enables it to quickly determine the proper actions to select by limiting the impact of other agents' actions, thus learning about good spectrum opportunities fast enough. However, agents' intrinsic objectives are likely not to be aligned with one another, which explains the sudden drop in their received reward right after learning about good opportunities; i.e., right after their received reward becomes high.

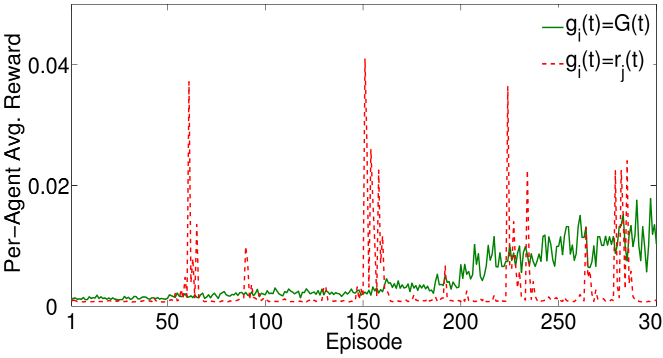


Fig. 2. Per-agent average achieved reward $\bar{r}(t)$ as a function of episode t under the two private objective functions: intrinsic choice ($g_i = r_j$) and global choice ($g_i = G$) for $Q = 2$, $\beta = 2$, $S_j = 20$ for $j = 1, 2, \ldots, 10$.

The second observation is regarding the second objective function choice, G. Observe that, unlike the intrinsic function, when each agent i sets its objective function g_i to the global reward function G, this results in a steadier performance behavior where the per-agent average received reward increases continuously, but slowly. With this function choice, agents' rewards are aligned with one another by accounting for each other's actions, and thus are less (or not likely to be) sensitive to the actions of any particular agents. The alignedness feature of this function is the reason behind the observed monotonic increase in the average received reward. However, this monotonic increase is relatively slow due to the function's insensitivity to one's actions, leading to slow learning rates.

To recap, objective functions must be designed with two requirements in mind: (i) *alignedness*; when agents maximize their own private objectives, their collective behavior should indeed result in increasing each agent's long-term received rewards, and (ii) *sensitivity*; objective functions should be sensitive to the agents' own actions so that proper selections of actions allow agents to learn about good spectrum opportunities fast enough.

With this in mind, the objective of this work is then to derive private objective functions for supporting inelastic traffic in large-scale, distributed DSA networks that meet the following design requirements. First, they should be efficient in that they should enable agents to achieve high rewards. Second, they should be scalable in that they should perform well in DSA networks with a small as well as a large number of agents. Third, they should be learnable in that they should enable agents to find and locate spectrum opportunities quickly. Fourth, they should be distributive in that they should be implementable in a decentralized manner. The objective functions that we derive in this work meet all of these design requirements.

III. Secondary-User Objective Coordination

In this section, we first begin by presenting the factoredness and learnability concepts, both of which are essential for capturing as well as ensuring the two required design properties: alignedness and sensitivity. Then, we propose efficient objective functions that meet the above design requirements.

A. Properties of Objective Functions

Again, let g_i denote the function that DSA agent i aims to maximize as its objective, and that we want to derive. Let z characterize the joint move of all DSA agents in the system. Here, the global reward, G, is a function of z, which specifies the full system state (G can then precisely be written as $G(z)$). Hereafter, we use the notation $-i$ to specify all agents other than agent i, and z_i and z_{-i} to specify the parts of the system state controlled respectively by agent i and agents $-i$. The system state z can then be written as $z = z_i + z_{-i}$.

For the joint actions of multiple DSA agents to lead to good overall average reward, two (often conflicting) requirements must be met. First, we must ensure that a DSA agent aiming to maximize its own private objective function also leads to maximizing the global (total achievable) rewards, so that its long-term average received rewards are indeed maximized. This means that the agents' private objective functions ($g_i(z)$ for agent i) need to be “aligned" or “factored" with the global reward function ($G(z)$) for a given system state z. Formally, for systems with discrete states, the degree of *factoredness* of a given private objective function g_i is defined as [35]:

$$\mathcal{F}_{g_i} = \frac{\sum_z \sum_{z'} h[(g_i(z) - g_i(z'))\ (G(z) - G(z'))]}{\sum_z \sum_{z'} 1} \tag{4}$$

for all z' such that $z_{-i} = z'_{-i}$, where $h[x]$ is the unit step function, equal to 1 if $x > 0$, and zero otherwise. Intuitively, the higher the degree of factoredness of an agent's objective function g_i, the more likely it is that a change of state will have the same impact on both the agent's (i.e., local) and the total (i.e., global) received rewards. A system is fully factored when $\mathcal{F}_{g_i} = 1$.

Second, we must ensure that each agent can discern the impact of its own actions on its private objective function, so that a proper action selection allows the agent to quickly learn about good spectrum opportunities. This means that the agent's private objective function should be more sensitive to its own actions than the actions of other agents. Formally, the level of sensitivity or *learnability* of a private objective function g_i, for agent i at z, can be quantified as [35]:

$$\mathcal{L}_{i,g_i}(z) = \frac{E_{z'_i}[|g_i(z) - g_i(z_{-i} + z'_i)|]}{E_{z'_{-i}}[|g_i(z) - g_i(z'_{-i} + z_i)|]} \tag{5}$$

where $E[\cdot]$ is the expectation operator, z'_i's are parts of the system states, controlled only by agent i, that are resulting from agent i's alternative actions at z, and z'_{-i}'s are parts of the system states, controlled by agent $-i$, that are resulting from agent $-i$'s alternative joint actions. So, at a given state z, the higher the learnability, the more $g_i(z)$ depends on the move of agent i. Intuitively, higher learnability means that it is easier for an agent to achieve higher rewards.

B. Proposed Objective Functions

The challenge in designing objective functions is to find the best tradeoff/balance between the two properties: factoredness and learnability (discussed in Section III-A). Doing so ensures that agents can learn to maximize their own objectives while doing so also leads to good overall system performance, resulting then in increasing each agent's long-term received rewards.

Let us first visit the behavior of the global reward function, illustrated earlier in Section II-C, so as to provide some intuition on our proposed function design. Recall that when agents set the global reward G as their objective functions (i.e., $g_i = G$ for each agent i), their collective behaviors did indeed result in increasing the total system achievable rewards (i.e., did result in a fully factored system), as agents' private objectives are aligned with system objective. The issue, however, is that because G depends on all the components of the system (i.e., all agents), it is too difficult for agents (using G as their objective functions) to discern the effects of their own actions on their objectives, resulting then in low learnability rates.

The key observation that leads to the proposed functions is that by removing the effects of all agents other than agent i from the function G, the resulting agent i's private objective function will have a much higher learnability level than G does, yet without compromising its degree of factoredness. These objective functions can formally be written as

$$D_i(z) \quad \equiv \quad G(z) - G(z_{-i}) \tag{6}$$

where z_{-i} again represents the parts of the state on which agent i has no effect. These difference functions have been applied to other domains (e.g., multi-robot control [33] and air traffic flow regulation [34]), and are shown to perform well.

First, note that these proposed functions (D_i for agent i) are fully factored, because the second term of Eq. (6) does not depend on agent i's actions. On the other hand, they also have higher learnability than G, because subtracting this second term from G removes most of other agents' effects from agent i's objective function. Intuitively, since the second term evaluates the value of the system without agent i, subtracting it from G provides an objective function (i.e., D_i) that essentially measures agent i's contribution to the total system received rewards, making it more learnable without compromising its factoredness quality.

By substituting Eq. (2) into Eq. (6), explicitly noting the time dependence t, and for clarity, removing the implicit dependence on the full state z, the function D_i for agent i selecting band j at time t can then be written as:

$$\begin{aligned} D_i(t) = & \sum_{k=1}^{m} n_k(t) r_k(n_k(t)) \\ & - \left(\sum_{k=1, k\neq j}^{m} n_k(t) r_k(n_k(t)) + (n_j(t)-1) r_j(n_j(t)-1) \right) \\ = & \quad n_j(t) r_j(n_j(t)) - (n_j(t)-1) r_j(n_j(t)-1) \end{aligned} \tag{7}$$

It is important to note that, by taking away agent i from the second term of the function D_i, the terms corresponding to all spectrum bands k, except the band j that agent i is using, cancel out. This explains why D_i, as shown in Eq. (7), depends on band j only. Therefore, the proposed function D_i is simpler to compute than the global function G. More specifically and importantly, it is fully decentralized as agents implementing/using it as their objectives need to gather and share information only with the agents that belong to the same band. This is one important property among few others (to be described later) that this proposed function has.

Let us now formally prove the claims that we made regarding the performances of the proposed objective functions.

Proposition 3.1: D_i is fully factored.

Proof: Differentiating both sides of Eq. (6) w.r.t. agent i's state z_i yields $\frac{\partial}{\partial_{z_i}} D_i(z) = \frac{\partial}{\partial_{z_i}} G(z) - \frac{\partial}{\partial_{z_i}} G(z_{-i})$, which also yields $\frac{\partial}{\partial_{z_i}} D_i(z) = \frac{\partial}{\partial_{z_i}} G(z)$ since $\frac{\partial}{\partial_{z_i}} G(z_{-i}) = 0$. ■

Proposition 3.2: The expected learnability of D_i is higher than the expected learnability of G.

Proof: We now sketch this proof. From Eq (5), the inner term of the numerator of D_i's learnability is equal to $D_i(z) - D_i(z_{-i} + z'_i)$, which, from Eq. (6), can also be written as $G(z) - G(z_{-i}) - (G(z_{-i} + z'_i) - G(z_{-i}))$ or equivalently as $G(z) - G(z_{-i} + z'_i)$. Hence, the numerator of the learnability is the same for D_i and G. Therefore, any gains in learnability must come from the denominator. Now, for a state z where agent i picked band j and a state z' where it did not, the inner term of the denominator of D_i's learnability is:

$$\begin{aligned} DEN_{\mathcal{L},D} &= D_i(z) - D_i(z'_{-i} + z_i) \\ &= n_j g_j(n_j) - (n_j - 1) g_j(n_j - 1) \\ &\quad - \left((n'_j + 1) g_j(n'_j + 1) - n'_j g_j(n'_j)\right) \end{aligned}$$

where we dropped the t terms for clarity and where n'_k is the number of agents that choose band k in the alternate state z'. That is the denominator consists of two terms, representing two bands *that differ by only one user*. Now, let us focus on the denominator for the learnability of G for a state z where agent i picked band j and a state z' where it picked band k:

$$\begin{aligned} DEN_{\mathcal{L},G} &= G(z) - G(z'_{-i} + z_i) \\ &= \sum_{l=1, l\neq j, l\neq k}^{m} n_l g_l(n_l) - n'_l g_l(n'_l) \\ &+ \quad n_k g_k(n_k) - (n'_k - 1) g_k(n'_k - 1) \\ &+ \quad n_j g_j(n_j) + (n'_j + 1) g_j(n'_j + 1) \end{aligned}$$

Now, here, there are also two terms, representing two bands (j and k) *that differ by only one user*. The expected magnitude of these values will be the same as those for the *only* two terms for $DEN_{\mathcal{L},D}$. However, there are $m-2$ terms that differ by as many as the total number of agents minus 1. As a consequence, we have $E[DEN_{\mathcal{L},G}] >> E[DEN_{\mathcal{L},D}]$ leading to D having much higher learnability on average than G. ■

IV. Optimal Agent Distribution

In order to help understand the behaviors and explain the intuition behind the achievable performances of our proposed functions (to be presented later in Section V), we will begin by deriving in this section the optimal behaviors of the DSA agents. Specifically, we will derive the optimal distribution of agents across the m available spectrum bands that leads to the optimal overall achievable rewards.

Without loss of generality and for simplicity, let us assume that $S_j = S$ for $j = 1, 2, \cdots, m$. Let n denote the total number of agents in the system at any time. First, note that when $n \leq m\frac{S}{Q}$, the optimal agent distribution is trivial, which basically corresponds to having each band contain no more than $\frac{S}{Q}$ agents, leading to the maximum possible overall achievable rewards (which equals mS when $n = m\frac{S}{Q}$). Therefore, in what follows, we assume that $n > m\frac{S}{Q}$, and we let $c = \frac{S}{Q}$, which denotes the capacity (i.e., maximum number of agents) of each spectrum band.

Next, we first begin by proving the following lemma, which will later be used for proving our main result.

Lemma 4.1: The global/total received rewards of a loaded[2] DSA network with a given number of agents reduces less when a new agent joins a more crowded band than when joining a less crowded band.

Proof: Recall that when a band j has $n' > c$ agents, its reward is $G_j(n') = n'Qe^{-\beta(\frac{n'}{c}-1)}$. If a new agent joins this band, the new reward becomes $G_j(n'+1) = (n'+1)Qe^{-\beta(\frac{n'+1}{c}-1)}$. First, it can easily be shown that when $n' > c \geq 1$, $G_j(n') > G_j(n'+1)$; i.e., the reward when joining band j decreases by $\Delta_j(n') \equiv G_j(n') - G_j(n'+1)$. Now we can easily see that $\Delta_j(n')$ decreases when n' increases. Hence, the greater the number n' (i.e., the more crowded the band), the smaller the decrease in reward. ■

Proposition 4.2: The optimal agent distribution corresponds to when $m-1$ bands each has exactly c agents and the m-th band has the remaining $n - c(m-1)$ agents.

Proof: Let $k = n - mc$, and let's refer to the agent distribution stated in the proposition as C. Note that C corresponds to when $m-1$ bands each has exactly c agents and the other m-th band has the remaining $c+k$ agents (since $n - c(m-1) = c+k$).

We proceed with the proof by comparing C with any possible distribution C' among all possible distributions. Let $c + k_1$ be the number of agents in the most crowded band in C', $c + k_2$ be the number of agents in the second most crowded band in C', and so forth. We just need to deal with the bands that each contains more than c agents. If there are p bands each containing more than c agents, then we know that $\sum_{i=1}^{p} k_i \geq k$.

For each band having $c + k'$ agents, let ϵ_i be the amount by which the global reward is reduced when agent i joins the band for $i = 1, 2, \cdots, k'$. From Lemma 4.1, it follows that $\epsilon_i > \epsilon_{i+1} > 0$, for all $i = 1, 2, \cdots, k'-1$.

Note that for the distribution C, the global reward is reduced by $t = \sum_{i=1}^{k} \epsilon_i$, and for C', it is reduced by $t' = \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i$. It remains to show that $t' - t > 0$ for any $C' \neq C$. We consider three different scenarios:

- $k_1 > k$: Here, we have

$$\begin{aligned} t' - t &= \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i - \sum_{i=1}^{k} \epsilon_i \\ &= \sum_{i=k}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i \end{aligned}$$

  which is greater than zero.

[2] Here, we consider that n is large enough to assume that no band contains less than the capacity c.

- $k_1 = k$: In this scenario, we have

$$\begin{aligned} t' - t &= \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i - \sum_{i=1}^{k} \epsilon_i \\ &= \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i \end{aligned}$$

  which is also greater than zero.
- $k_1 < k$: In this scenario, we have

$$\begin{aligned} t' - t &= \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i - \sum_{i=1}^{k} \epsilon_i \\ &= \underbrace{\sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i}_{part\ a} - \underbrace{\sum_{i=k_1}^{k} \epsilon_i}_{part\ b} \end{aligned}$$

  Since $k_1 + k_2 + \cdots + k_p \geq k$, the number of ϵ_i terms in *part a* is greater than the number of terms in *part b*. From Lemma 4.1, we know that the largest term in *part b* is ϵ_{k_1}, which is smaller than the smallest term ϵ_{k_2} in *part a*. Hence, *part a* is greater than *part b*, and thus $t' - t$ is greater than zero.

In all scenarios, we showed that $t' - t > 0$. Therefore, the global reward for any distribution C' is smaller than that for the distribution C; i.e., C is the distribution that corresponds to the maximal global achievable reward. ■

The optimal agent distribution that we derived in this section leads to the maximum/optimal per-agent average achievable rewards. This optimal distribution will help us, as will be shown in the next section, understand and evaluate the performance of our proposed objective functions.

V. Performance Evaluation

In this section, we evaluate the effectiveness of the proposed objective functions by measuring their achievable rewards, and comparing them with those achievable under each of the two intuitive functions r_j and G.

A. Optimality

We first begin by considering the same experiment that we conducted in Section II-C, where again the total number of agents is set to 500 and the number of bands is set to 10. Here, we assume that all agents enter and leave the DSA network at the same time. Also, we ignore the PUs' activities in this section; these activities will be considered in Section V-E.

Fig. 3 shows the per-agent average achievable reward under each of the three functions: intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$). Our results show that the proposed function D_i outperforms substantially the other two functions. Observe that D_i achieves a per-agent average reward of about 0.12, whereas, each of the other two functions achieves a reward of no more than approximately 0.02. That is, D_i achieves almost 6 times as much as each of the other two functions does. Another property that D_i has, and that requires attention is learnability. Observe how quickly the rewards achievable under D_i reach up their high value. To recap, these obtained results show that the proposed function outperforms the other two functions in terms of both optimality and learnability.

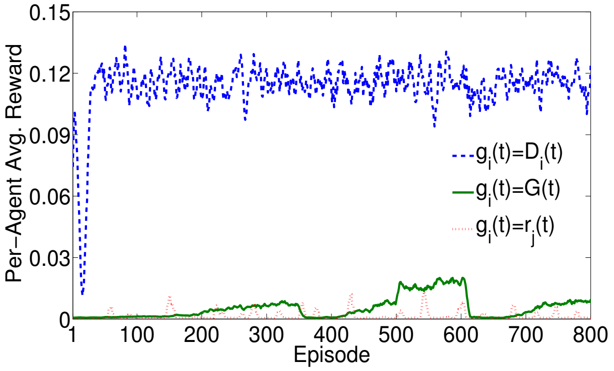


Fig. 3. Per-agent average achieved reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j.

B. Scalability

We now study the proposed function with regard to scalability. For this, we plot in Fig. 4 the per-agent average achievable reward under each of the three studied objective functions when varying the number of agents, n, from 100 to 800 while keeping the number of bands m equal to 10. Observe that unlike the functions r_j and G, the proposed function D_i is highly scalable. Note that as the number of agents increases, D_i maintains high achievable rewards, whereas the achievable reward under either of the other two functions drops dramatically with the number of agents.

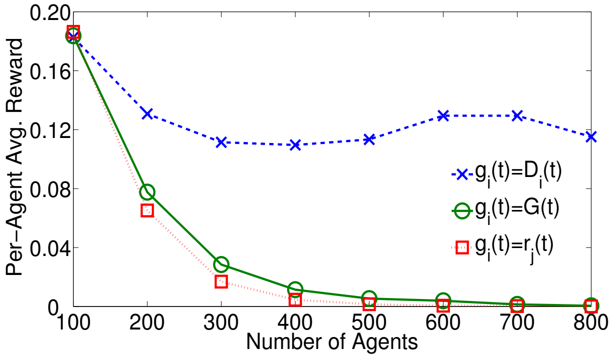
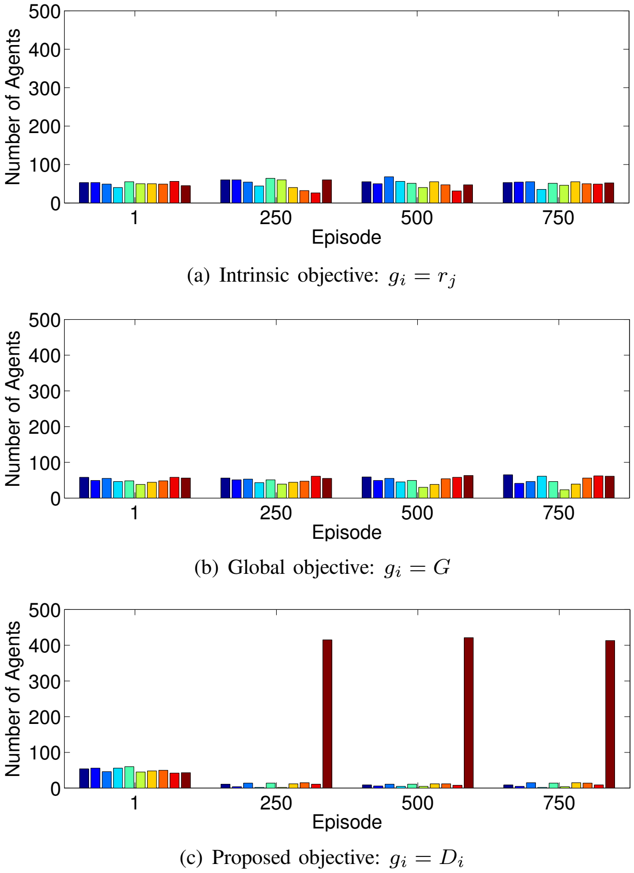


Fig. 4. Per-agent average achieved reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various numbers of agents: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j.

C. Agent Distribution

In this section, we want to further investigate the behaviors of agents in terms of their distribution/repartition across the m available spectrum bands. More specifically, we compare the actual/measured distribution of agents as a result of using the proposed objective functions with that ideal/theoretical distribution derived in Section IV. Recall that the ideal/theoretical agent distribution, as stated in Proposition 4.2, corresponds to the repartition that leads to the maximum achievable rewards.

To illustrate, we plot in Fig. 5 the actual, measured distribution of the $n = 500$ agents across the $m = 10$ bands at different times (i.e., every 250 episodes) under the three studied objective functions. Note that in the case of r_j (Fig. 5(a)) and G (Fig. 5(b)), agents are (approximately) equally distributed among the 10 bands (≈ 50 agents/band), and at all times. But when using D_i (Fig. 5(c)), 9 bands out of 10 each contains about 10 agents, which represent the capacity $c = \frac{S}{Q}$, and the rest (≈ 410 agents) are in the 10^{th} band. It is important to note that this corresponds to (or very close to) the optimal agent distribution that we derived in Proposition 4.2. Thus, the proposed objective function, D_i, when used as an objective function, leads then to a distribution of agents across the available bands that is very close to the optimal agent distribution stated through Proposition 4.2, which explains the high performances that it achieves.



(a) Intrinsic objective: $g_i = r_j$

(b) Global objective: $g_i = G$

(c) Proposed objective: $g_i = D_i$

Fig. 5. Distribution of the 500 agents across the $m = 10$ different bands: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j. Each bar corresponds to one band.

It is important to mention, as it will become clearer in next sections, that the most crowded band (led to under D_i) does not always contain the same set of agents. That is, agents belonging to this crowded band (which of course offers the least per-agent reward) change over time, since agents move across bands at different time steps. The fact that agents do not get stuck in the crowded band is an important property of D_i, as it ensures fairness among agents by allowing different agents to receive approximately equal amounts of rewards. This is studied thoroughly in the next section.

D. Fairness

Fairness is also another important performance metric that we want to evaluate. We want to assess how fair D_i is when compared with the other two functions. For this, we plot in Fig. 6 the coefficient of variations (defined as the ratio of the standard deviation to the mean of the agents' received rewards; we use this metric as a means of assessing the fairness, which reflects how close agents' received rewards are to one another) of the per-agent average received rewards under the three studied functions for various numbers of agents. First, observe that D_i achieves small coefficient of variation values, and this is independent of the number of agents in the system; i.e., even when the number of agents is large, coefficient of variation values are maintained low. Second, note that as the number of agents increases, while the coefficient of variation values are maintained low under D_i, they increase under each of the other two functions (but worse under the intrinsic function than under the global function).

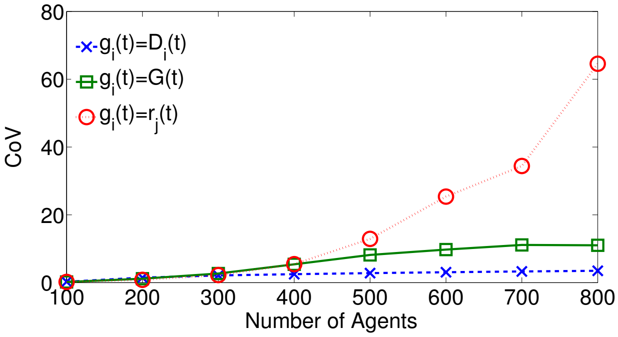


Fig. 6. Coefficient of variation (CoV) of per-agent average received reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various numbers of agents: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j.

Our fairness results, therefore, show that the most crowded band does not always contain the same set of agents. In other words, agents do move across bands as time goes on. Agent movement across bands is triggered by the exploration nature of the learning algorithm, which forces agents to constantly rotate across the different bands, thereby ensuring fairness among agents by allowing them to receive approximately equal amounts of rewards.

Now to further study the fairness of the proposed technique, we now look at how an individual agent's reward behaves over time. For this, we randomly picked three agents/users (out of 500) and for each of the three users, we tracked the instantaneous received rewards that the user receives as time progresses. These new results are shown in Fig. 7 for different time scales (1, 2, and 5 aggregated slots). The figure shows that each user does indeed rotate among the bands; that is, users do not get stuck in the worst band all times. First, observe that users' received rewards fluctuate (because users move across different bands), but on average all receive roughly equal rewards. This can be seen clearly in Fig. 7(c) when the rewards of each individual user are aggregated over each 5 consecutive slots. The figure also shows that the received reward of a given individual user stays above the threshold for several consecutive slots but then drops below also for some consecutive slots, then goes above and down again, and so on. For example, as shown in Fig. 7(a), user 2's reward stays below the threshold for 2 consecutive slots, then above for 3 consecutive slots, then below for 1 slot, then above for 1 slot, then below for 1 slot, then above for 3 slots, etc. For completeness, we also measured the maximum number of consecutive time slots during which the received rewards stay below the threshold; this maximum number is about 4 slots.

What our proposed technique tries to do is to make the best use of the system in terms of the per-user average received rewards as well as fairness, and more importantly, does so distributively. Observe that agents' received rewards fluctuate above and below the threshold as time goes on. That is, the instantaneous throughput that agents receive may not always be above the required throughput threshold. However, even when the instantaneous throughput fluctuates, as long as the average throughput (even at higher time scales, not necessarily on a per time slot basis) is above threshold, the perceived quality can still be acceptable for applications like video and audio. Multimedia applications can overcome these short-term throughput fluctuations by relying on existing techniques such as resolution adaptation techniques, which adjust playing resolution based on achievable rates.

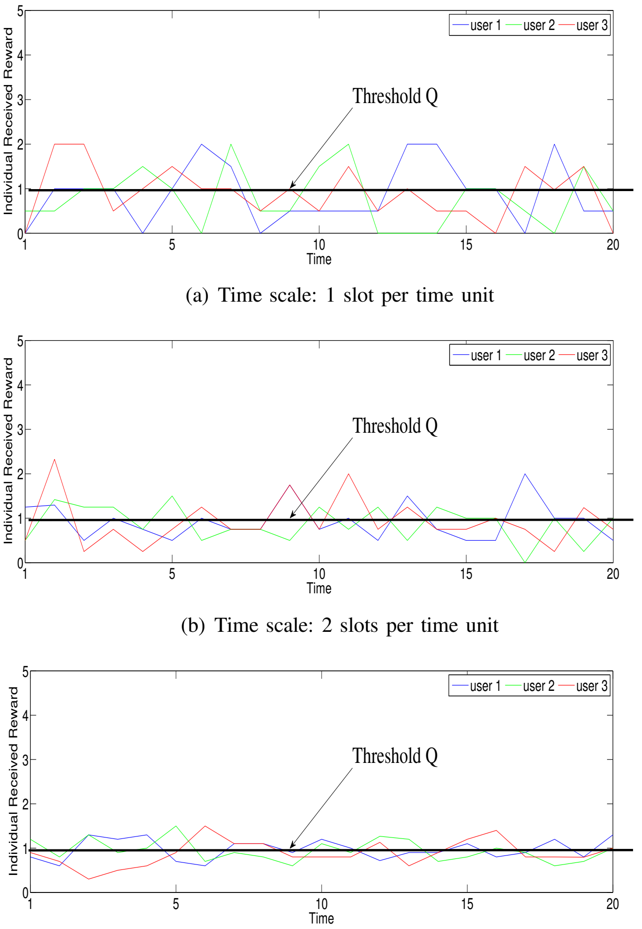
To summarize, we showed that the proposed functions (i) achieve high per-agent average rewards, (ii) are scalable as they perform well in small- as well as large-scale systems, (iii) are learnable as rewards reach up high values very quickly, (iv) are distributive as they require information sharing only among agents belonging to the same band, and (v) are fair as they ensure that agents receive approximately equal rewards.

In the next section, we show that these performance qualities still hold when considering primary users (PUs)' activities too.

E. PUs' Activities

In this section, we want to investigate how well these obtained results hold when considering the presence of PUs. We also consider that agents can choose to enter and leave the network at different independent times. Recall that agents here refer to SUs or DSA users, and can be viewed as data sessions/flows that are initiated by SUs, which can start and finish at different times, and independently from one another.

1) Impact of SUs' traffic behaviors: We first begin by studying the impact of SUs' traffic behaviors without considering the presence of primary users. The impact of the presence of primary users will be investigated in the next section. To mimic the agents' dynamic behaviors, we assume that agents (e.g., data sessions) arrive according to a Poisson process with arrival rate λ, and stay in the network for an exponentially distributed duration of mean τ. We use $\kappa = \lambda\tau$ to designate the *SU load*, which essentially represents the average number of agents in the system. In this section, we study the impact of the average



(a) Time scale: 1 slot per time unit

(b) Time scale: 2 slots per time unit

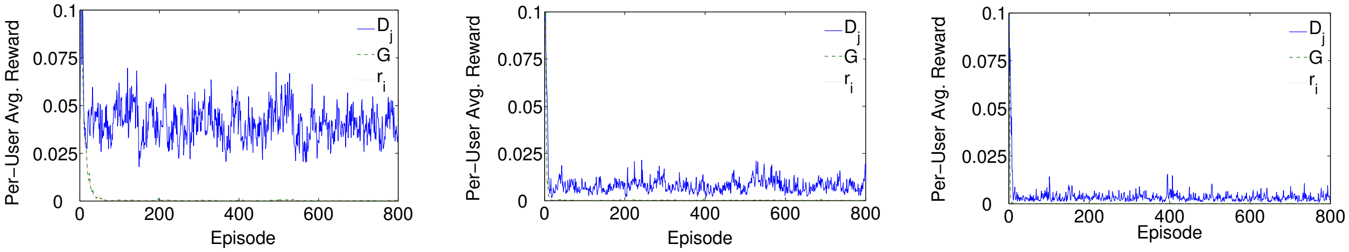
(c) Time scale: 5 slots per time unit

Fig. 7. Individual agent reward behaviors.

length of the period SUs spend in the system. For this, we fix the average number of agents (i.e., SU load) to $\kappa = 500$ and evaluate the performance of the proposed function for different values of the ratio λ/τ. Recall that for a fixed SU load, the higher the ratio $\frac{\lambda}{\tau}$, the shorter the sessions' durations. For example, when $\kappa = 500$, $\frac{\lambda}{\tau} = 1$ implies that the sessions' average duration τ and arrival rate λ are both equal to ≈ 22.3, whereas $\frac{\lambda}{\tau} = 5$ implies that $\tau = 10$ and $\lambda = 50$.

Fig. 8 shows the per-user average received reward when κ is fixed to 500 but for various combinations of λ and τ. Fig. 8(a) for $\lambda = 22.3$ and $\tau = 22.3$ (i.e., $\frac{\lambda}{\tau} = 1$); Fig. 8(b) for $\lambda = 50$ and $\tau = 10$ (i.e., $\frac{\lambda}{\tau} = 5$); and Fig. 8(c) for $\lambda = 100$ and $\tau = 5$ (i.e., $\frac{\lambda}{\tau} = 20$). First, the figure shows that the proposed function outperforms the other two regardless of the average length of SUs' staying periods. Second and more importantly, we observe that as the average duration decreases (that is, as agents spend less and less time on average in the system), the proposed functions' achievable performance decreases as well. This is merely because short periods of times will not be enough for the agents learn, and by the time they start to learn where the best opportunities are, their sessions end.

2) Impact of PUs' traffic behaviors: To mimic PUs' activities, we consider that each band is associated with a set of



(a) $\lambda = 22.3$, $\tau = 22.3$

(b) $\lambda = 50$, $\tau = 10$

(c) $\lambda = 100$, $\tau = 5$

Fig. 8. Per-user average received reward under Poisson arrival traffic model for $\kappa = 500$ and with no PUs traffic ($\eta = 0\%$).

PUs that enter and leave the band at random times. We model PUs' activities on each band as a renewal process alternating between ON and OFF periods, representing the time during which PUs are respectively present (ON) and absent (OFF). For each spectrum band j, we assume that ON and OFF durations are exponentially distributed with means ν_j^{ON} and ν_j^{OFF}, respectively[3]. In what follows, we use $\eta_j \equiv \nu_j^{ON}/(\nu_j^{OFF} + \nu_j^{ON})$ to denote the *PU traffic load* on spectrum band j. In this experiment, we fix λ/τ to 1, and vary the average number of agents κ from 250 to 1000; i.e., $\kappa = \lambda\tau$ is varied from 250 to 1000.

Figs. 9 and 10 show the per-agent average received reward under each of the three studied objective functions for a DSA network with, respectively, 10% and 50% PU traffic load (i.e., $\eta = \eta_j = 50\% \,\forall j$) while considering various numbers of agents in the network: $\kappa = 500$, $\kappa = 750$, and $\kappa = 1000$. There are three observations that we can make out of these results. First, observe that regardless of the number of agents in the network, the proposed function D_i achieves (on average) higher rewards than those achieved under the other two functions, and the performance gain increases with the number of agents (i.e., the greater the number of agents κ, the greater the performance gain between the proposed function and any of the other two functions). Second, note that the achievable rewards quickly reach up high values when PUs are not present, but also quickly drop down to zero as soon as the PUs return to their bands. This explains the observed ups and downs behavior of the achievable rewards. Third, note that as the number of agents (i.e., κ) in the network increases, the achievable rewards under the global function G or the intrinsic function r_j decreases substantially. For example, when the number of agents equals $\kappa = 1000$ (Fig. 10(c)), the achievable rewards under each of these two functions (G or r_j) is almost zero. However, unlike these two functions, the proposed function D_i yields much higher per-agent average received rewards than what the other two functions achieve, and does so even under large numbers of agents (e.g., when the number of agents equals 750 or 1000). In other words, the proposed function is still highly scalable even in the presence of PUs' activities.

In this work, real data traces [36] are also used to evaluate the effectiveness of the proposed techniques. This data is measured over 60 channels each having a bandwidth of 25kHz, and over a 100-minute time period through spectral measurements of PUs' activities in the 850-870MHz band at every 0.01 second with a frequency resolution of 8.333kHz [36]. Our obtained results show that the proposed functions outperform substantially the other methods when also considering real PU traffic behaviors, and these performances are as good as those obtained when PUs' activities are mimicked via simulation.

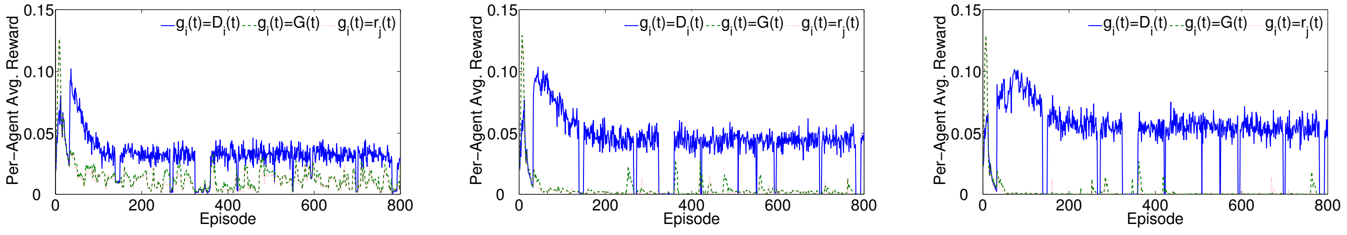
Therefore, these results confirm that our proposed coordination techniques perform well in the presence of PUs' activities too, and also in terms of achievable rewards, scalability, and learnability.

Next, we show that these proposed functions perform well in the presence of PUs' activities when it also comes to fairness. In Fig. 11, we show coefficient of variation of the per-agent average received rewards under the three functions when varying the number of agents from 250 to 1000 for various PU traffic loads η: Fig. 11(a) for $\eta = 10\%$, Fig. 11(b) for $\eta = 30\%$, and Fig. 11(c) for $\eta = 50\%$. Observe that when PUs are present, the proposed objective function achieves coefficient of variation values also lower than those achievable under any of the other two functions, especially when the number of agents present in the network is large. This is true, and independent of the PU traffic load. We also observe that coefficient of variation increases with the number of agents when the global function G or the intrinsic function r_j is used; whereas, it remains low when the proposed function is used.

Therefore, in terms of fairness, our results show that the band having the largest number of agents does not always contain the same set of agents. That is, agents belonging to the crowded band change over time, as agents move across bands at different time steps. Agent movement across bands is triggered by the exploration nature of the learning algorithm by exploring new opportunities every once in a while. This constant rotation of agents across the different bands is what ensures fairness among agents by allowing them to receive approximately equal amounts of rewards. This is independent of the number of agents. This explains why the coefficient of variation under the proposed function does not change much with the number of agents.

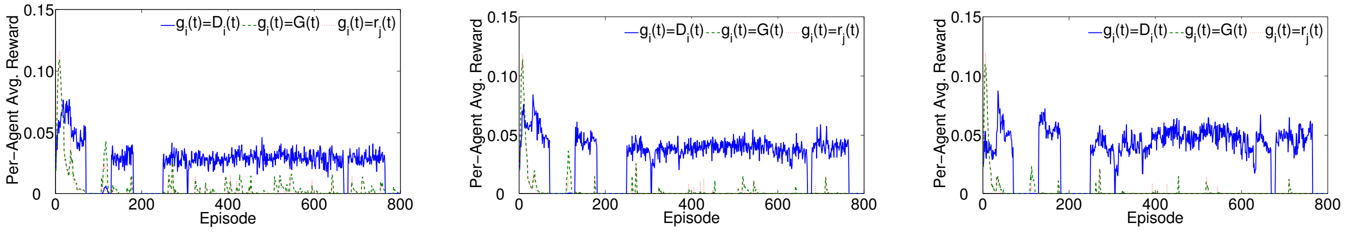
As for the existing functions, because the agents are equally distributed among the bands under these functions, all agents will receive low, but almost equal rewards, yielding then very low variability (e.g., low variance) of received rewards. Now

[3] Recall that learners do not actually need prior knowledge of PUs' traffic behavior. Here, the exponential distributions will be used to generate samples so as to be able to mimic the DSA environment.



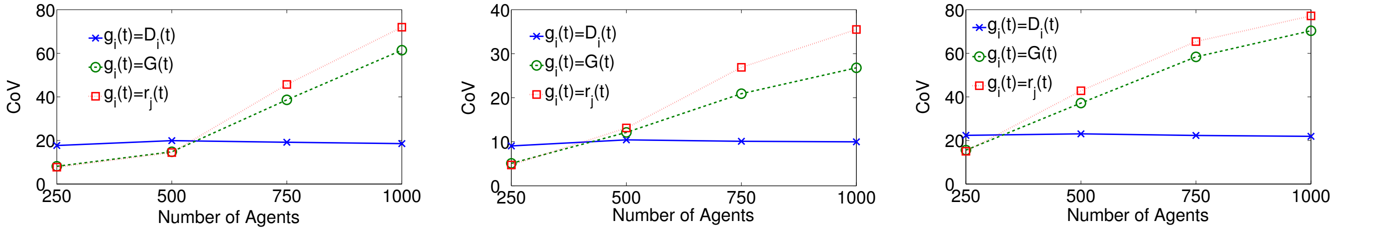
(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 9. Per-user average reward under DSA agent traffic with Poisson arrival of $\lambda\tau = 1$ and with PUs traffic load of $\eta = 10\%$.



(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 10. Per-user average reward under DSA agent traffic with Poisson arrival of $\lambda\tau = 1$ and with PUs traffic load of $\eta = 50\%$.



(a) $\eta = 10\%$ (b) $\eta = 30\%$ (c) $\eta = 50\%$

Fig. 11. Coefficient of variation (CoV) of per-agent average achieved reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various PU traffic loads: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j.

when the number of agents is low, the average received reward is high (less agents will have to share the same bands), thereby resulting in a higher coefficient of variation values. But as the number of agents increases, the per-user average received reward decreases, explaining then the observed increase in the coefficient of variation values.

To recap, our results show that the proposed functions, when used in practical, dynamic network settings, achieve good performances in terms of scalability, learnability, and fairness.

VI. Discussions

A. Distributed Function Computation Method

One key feature of the proposed techniques is that agents can implement them by relying only on information that can be observed and gathered locally. Although the design of distributed methods for computing these functions is in itself a different problem, we shed some light on it here for completeness.

What makes the proposed function implementable in a decentralized manner is the following observation: by taking away agent i from the second term of Eq. (6), the terms corresponding to all spectrum bands, except the band that agent i uses, cancel out. This leads to Eq. (7). From this equation, note that $D_i(t)$ depends only on $n_j(t)$, the number of agents that happen to be contending with agent i for band j. Hence, in order to compute D_i, one needs to estimate $n_j(t)$ given the information that agent i observes locally. Now, an agent i using band j can easily measure (without needing any collaboration) the amount of throughput/data rates, $a_i(t)$, it receives from using the DSA system. Thus, assuming that all agents sharing a band will roughly receive the same amount of throughput, the number of agents, $n_j(t)$, using band j can be estimated to $S_j/a_i(t)$, which can then be used to estimate/compute D_i. Now when different agents receive different amounts of throughput, the problem can indeed be very challenging without any cooperation from the primary network providers/owners. We believe that economical incentives (such as in [37]) can be used in this case to encourage the primary network providers to reveal such information.

The design of distributed methods for computing the proposed functions is in itself a different, challenging problem. A more thorough study of these methods requires further assessment of the tradeoffs between the estimation accuracy and the incurred overhead due to cooperation. This is left for future investigation.

B. Band Switching Overhead

Switching from a spectrum band to another often incurs delay. Such switching delay mainly depends on how far the frequency band to be switched to is from the current band, as well as on the sensing delay needed for discovering and locating an available spectrum band. This delay is incurred every time an agent switches to a new band due to, for example, the return of primary users to their bands. During this switching process/delay, an agent has to cease its communication and hence will not be able to send data, thereby impacting/reducing the amount of its received throughput. This amount of received throughput is what is used to update the reward function. Therefore, the switching and sensing delays are indirectly incorporated in the proposed reward model through the received throughput. However, other types of overhead like energy overhead resulting from the spectrum sensing process have not been investigated, and are left as a future work.

C. Resource Access and Sharing Methods

Numerous medium access control (MAC) protocols have been proposed during the last few years to enable multiple access in cognitive radio/dynamic spectrum access networks; [38, 39] are just a couple of (among many) surveys on MAC designs that can be found in literature. Our proposed coordination technique assumes a CSMA/CA-like multiple access technique. More specifically, we rely on our recently proposed and implemented IEEE 802.11-like multichannel MAC protocol [40]. Here, we want to mention that even though we propose to use a CSMA/CA-like approach, our proposed technique is independent of the MAC protocol being adopted by the users, and can be used regardless of the MAC being used.

That is said, we now want to bring up the following point for the sake of discussion and completeness. We believe that in order to truly enable successful dynamic spectrum access, more sophisticated medium access and sharing approaches are needed. Although so many MAC design ideas have been proposed over the years for alleviating medium contention in cognitive radio networks, most of (if not all) these protocols require/assume that all SUs deploy the same communication/modulation/medium access strategy and policy (e.g., TDMA, FDMA, CSMA, etc). This is of course needed so that multiple users can access and share the medium among themselves. However, we believe that spectrum users (or DSA users), as envisioned by the cognitive radio paradigm, will (or at least are very likely to) be deploying different communication/software techniques, belonging to various different technology platforms, and using different architectures. Therefore, the assumption that all SUs will be using the same MAC, though seems needed to enable resource sharing, is somehow unrealistic. Having an universal MAC that SUs ought to use if they want to use spectrum opportunistically seems essential and necessary. SUs can still deploy their MAC technique when communicating in their home networks, but when wanting to exploit spectrum opportunities, they are required to conform to the universal MAC policies.

VII. Related Work

There have been significant research efforts on the development of learning-based techniques that promote effective DSA [23, 24, 26, 27]. In [23], the authors propose an approach for the detection of spectral resources based on reinforcement learning, allowing the cognitive radio to select the most available channels. In [24], a greedy channel-selection strategy and access policy are introduced. The proposed techniques maximize the instantaneous reward that SUs receive from the DSA system, where DSA is modeled as a partially observable Markov decision process (POMDP). Liu et al. [26] propose a cooperative multiuser approach based on explicit communication between the secondary users, which is basically a learning-based approach that involves and relies on the use of collision feedback information to locate good opportunities.

In [41], an auction-based framework is developed that allows spectrum users to bid for primary and secondary access based on their utilities and traffic demands, and uses the bids to solve the access allocation problem. These auctions can be set up to maximize revenue, utilization, and/or efficiency. In [28], the authors propose a game-theoretical approach with a new solution concept, the correlated equilibrium. To achieve this correlated equilibrium, they construct an adaptive algorithm based on no-regret learning that guarantees convergence. Some proposed solutions for enabling effective DSA in cognitive radio networks adopt market-based approaches in order to effectively regulate the available spectrum resources. It is shown how various centralized and decentralized spectrum access market strategies can be designed based on a stochastic game framework, where SUs can compete over time for the dynamically available transmission opportunities [42].

In [43], multiagent reinforcement leaning (MARL) is used to allow SUs to learn good strategies of channel selection to avoid collisions incurred by the lack of coordination, where each SU learns how to select channels based on its past experience. Shetty et al. [44] propose a non-cooperative and learning-based approach to allow multiple SUs to achieve maximal throughput in an unslotted DSA network. They also consider collisions among SUs while making channel sensing decisions. Unlike these works where the main focus was on learning- and/or market-based approaches, this work (as well as [45], a conference version) focuses instead on the design of coordination techniques (e.g., efficient objective functions) that can be used by these learning schemes to promote efficient DSA; i.e., our work complements these proposed learning-based approaches.

VIII. Conclusion

In this paper, we propose and study scalable and distributive objective functions that DSA users can use to locate and exploit the best spectrum opportunities. DSA users can rely on any learning algorithms to maximize these proposed objective functions, thereby ensuring high performances in terms of the long-term average received rewards. Our results show that these proposed objective functions (i) enable DSA users to receive *high* rewards (optimal), (ii) perform well in small- as well as large-scale DSA networks (scalable), (iii) reach up optimal reward values very quickly (learnable), and (iv) require information sharing only among users belonging to the same spectrum band (distributed).

IX. Acknowledgment

This work was supported in part by National Science Foundation through NSF CAREER award CNS-0846044.

References

[1] M. McHenry, "Reports on spectrum occupancy measurements, shared spectrum company," in *www.sharedspectrum.com/?section=nsf_summary*.

[2] M. McHenry and D. McCloskey, "New York city spectrum occupancy measurements," *Shared Spectrum Conf.*, Sept. 2004.

[3] B. Hamdaoui and K. G. Shin, "OS-MAC: An efficient MAC protocol for spectrum-agile wireless networks," *IEEE Transactions on Mobile Computing*, August 2008.

[4] C. Zou and C. Chigan, "On game theoretic DSA-driven MAC for cognitive radio networks," *Computer Communications*, vol. 32, no. 18, 2009.

[5] M. Ma and D. H. K. Tsang, "Joint design of spectrum sharing and routing with channel heterogeneity in cognitive radio networks," *Physical Communication*, vol. 2, no. 1-2, 2009.

[6] M. Timmers, S. Pollin, A. Dejonghe, L. Van der Perre, and F. Catthoor, "A distributed multichannel MAC protocol for multihop cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 1, 2010.

[7] L. Chen, S. Iellamo, M. Coupechoux, and P. Godlewski, "An auction framework for spectrum allocation with interference constraint in cognitive radio networks," in *Proceedings of IEEE INFOCOM*, 2010.

[8] H. Xu, J. Jin, and B. Li, "A secondary market for spectrum," in *Proceedings of IEEE INFOCOM*, 2010.

[9] L. Duan, J. Huang, and B. Shou, "Competition with dynamic spectrum leasing," in *Proceedings of IEEE DySPAN*, 2010.

[10] G. S. Kasbekar and S. Sarkar, "Spectrum auction framework for access allocation in cognitive radio networks," in *Proceedings of ACM MobiHoc*, 2009.

[11] J. Jia, Q. Zhang, and M. Liu, "Revenue generation for truthful spectrum auction in dynamic spectrum access," in *Proceedings of ACM MobiHoc*, 2009.

[12] T. R. Newman, S. M. S. Hasan, D. Depoy, T. Bose, and J. H. Reed, "Designing and deploying a building-wide cognitive radio network testbed," *IEEE Communications Magazine*, vol. 48, no. 9, 2010.

[13] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, vol. 11, no. 1, May 2009.

[14] H. Kim and K. G. Shin, "Efficient discovery of spectrum oppotunities with MAC-layer sensing in cognitive radio networks," *IEEE Transactions on Mobile Computing*, May 2008.

[15] K. Kim, I. A. Akbar, K. K. Bae, J.-S. Um, C. M. Spooner, and J. H. Reed, "Cyclostationary approaches to signal detection and classification in cognitive radio," in *Proceedings of IEEE DySPAN*, 2007.

[16] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, Februray 2008.

[17] X. Li, Q. C. Zhao, X. Guan, and L. Tong, "Optimal cognitive access of markovian channels under tight collision constraints," *IEEE Journal on Selected Areas in Communications, Special Issue on Advances in Cognitive Radio Networks and Communications*, To Appear 2011.

[18] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Trans. on Signal Processing*, vol. 58, no. 5, May 2010.

[19] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. on Inf. Theory*, May 2008.

[20] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Transactions on Information Theory*, 2009.

[21] J. Unnikrishnan and V. V. Veeravalli, "Cooperative sensing for primary detection in cognitive radio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, 2008.

[22] X. Liu, B. Krishnamachari, and H. Liu, "Channel selection in multichannel opportunistic spectrum access networks with perfect sensing," in *Proceedings of IEEE DySPAN*, 2010.

[23] U. Berthold, M. Van Der Schaar, and F. K. Jondral, "Detection of spectral resources in cognitive radios using reinforcement learning," in *Proceedings of IEEE DySPAN*, 2008, pp. 1–5.

[24] J. Unnikrishnan and V. V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, August 2010.

[25] J. Unnikrishnan and V. V. Veeravalli, "Dynamic spectrum access with learning for cognitive radio," in *Proc. of Asilomar Conference on Signals Systems and Computers*, Oct. 2008.

[26] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *Proceedings of IEEE ICC*, 2008.

[27] K. Liu and Q. Zhao, "Distributed learning in cognitive radio networks: multi-armed brandit with distributed multiple players," in *Submitted to IEEE Int. Conf. on Acousitcs, Speech, and Signal Processing*, 2010.

[28] Z. Han, C. Pandana, and K. J. R. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," in *Proceedings of IEEE WCNC*, 2007.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[30] G. Tesauro, "Practical issues in temporal difference learning," *MLJ*, vol. 8, pp. 257–277, 1992.

[31] A. Agogino and K. Tumer, "Unifying temporal and structural credit assignment problems," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*, New York, NY, July 2004.

[32] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[33] A. K. Agogino and K. Tumer, "Efficient evaluation functions for evolving coordination," *Evolutionary Computation*, vol. 16, no. 2, pp. 257–288, 2008.

[34] K. Tumer and A. Agogino, "Distributed agent-based air traffic flow management," in *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Honolulu, HI, May 2007, pp. 330–337.

[35] A. K. Agogino and K. Tumer, "Analyzing and visualizing multiagent rewards in dynamic and stochastic environments," *Journal of Autonomous Agents and Multi Agent Systems*, vol. 17, no. 2, pp. 320–338, 2008.

[36] D. Xu, E. Jung, and X. Liu, "Optimal bandwidth selection in multi-channel cognitive radio networks: how much is too much?," in *Proceedings of IEEE DySPAN*, 2008.

[37] S. Sengupta and M. Chatterjee, "An economic framework for dynamic spectrum access and service pricing," *ACM/IEEE Transactions on Networking*, August 2009.

[38] C. Cormiob and K. R. Chowdhurya, "A survey on MAC protocols for cognitive radio networks," *Ad Hoc Networks*, vol. 7, no. 7, 2009.

[39] A. De Domenico, E. Calvanese Strinati, and M.-G. Di Benedetto, "A survey on MAC strategies for cognitive radio networks," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, vol. 14, no. 1, 2012.

[40] M. Maiya and B. Hamdaoui, "iMAC: Improved medium access control for multi-channel multi-hop wireless networks," *Wireless Communications and Mobile Computing*, July 2011.

[41] G. S. Kasbekar and S. Sarkar, "Spectrum auction framework for access allocation in cognitive radio networks," *Proceedings of ACM MobiHoc*, 2009.

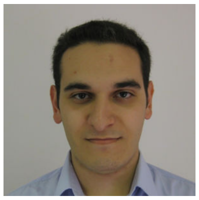
[42] M. van der Schaar and Fangwen Fu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," *Proceedings of IEEE*, vol. 97, April 2007.

[43] Husheng Li, "Multiagent q-learning for aloha-like spectrum access in cognitive radio systems," *EURASIP Journal on Wireless Communications and Networking*, 2010.

[44] S. Shetty, Min Song, Chunsheng Xin, and E.K. Park, "A learning-based multiuser opportunistic spectrum access approach in unslotted primary networks," *IEEE INFOCOM proceedings*, 2009.

[45] B. Hamdaoui, M. NoroozOliaee, K. Tumer, and A. Rayes, "Aligning spectrum-user objectives for maximum inelastic-traffic reward," in *Proceedings of IEEE Int'l Conference on Computer Communication Networks*, August 2011.

Bechir Hamdaoui received the Diploma of Graduate Engineer (1997) from the National School of Engineers at Tunis, Tunisia. He also received M.S. degrees in both Electrical and Computer Engineering (2002) and Computer Sciences (2004), and the Ph.D. degree in Computer Engineering (2005) all from the University of Wisconsin at Madison. In September of 2005, he joined the RTCL Lab at the University of Michigan at Ann Arbor as a postdoctoral researcher. Since September of 2007, he has been with the School of EECS at Oregon State University as an assistant professor. His research focus is on cross-layer protocol design, system performance modeling and analysis, adaptive and learning technique development, and resource and service management for next-generation wireless networks and communications systems. He has won the NSF CAREER Award (2009), and the IWCMC Best Paper Award (2011). He is presently an Associate Editor for IEEE Transactions on Vehicular Technology (2009-present) and Wireless Communications and Computing Journal (2009-present). He also served as an Associate Editor for Journal of Computer Systems, Networks, and Communications (2007-2009). He served as the chair for the 2011 ACM MobiCom's Student Research Competition, and as the program chair/co-chair of the Pervasive Wireless Networking Workshop (PERCOM 2009), the WiMAX/WiBro Services and QoS Management Symposium (IWCMC 2009), the Broadband Wireless Access Symposium (IWCMC 2010), the Cooperative and Cognitive Networks Workshop (IWCMC 2011 and 2012), and the Internet of Things, Machine to Machine, and Smart Services Applications Workshop (CTS 2012). He also served on program committees of several IEEE conferences. He is a member of IEEE, IEEE Computer Society, IEEE Vehicular Society, and IEEE Communications Society.



MohammadJavad NoroozOliaee received the BS degree in Computer Engineering from Sherif University of Technology, Iran, in 2009. He is currently working toward the PhD degree in Computer Science at Oregon State University. His research focus is on the design and development of distributed coordination techniques for wireless networking systems with dynamic spectrum access capabilities.

Kagan Tumer is a professor of robotics and control at Oregon State University, where he leads the Adaptive Agents and Distributed Intelligence Lab. Dr. Tumer's research interests are learning and control in large autonomous systems with a particular emphasis on multiagent coordination. Applications of his work include coordinating multiple robots, controlling unmanned aerial vehicles, reducing traffic congestion and managing air traffic.

His work has led to over one hundred and thirty publications, including three edited books, one patent, several best paper awards (2007 Conference on Autonomous Agents and Multiagent Systems, 2012 Genetic and Evolutionary Computation Conference). He is an associate editor of the Journal on Autonomous Agents and Multiagent Systems, and was the program co-chair of the 2011 Autonomous Agents and Multiagent Systems Conference (AAMAS 2011). Dr. Tumer received his PhD (1996) in Electrical and Computer Engineering at The University of Texas, Austin, and is a member of AAAI and a senior member of IEEE.

Ammar Rayes is a Distinguished Engineer at Cisco Systems. He has been at the core of developing IP-based network and service management solutions for over 20 years. His main areas of expertise include Smart Services, NMS/OSS, Big Data, Could, Mobility, Performance and Traffic Engineering, and Embedded management. Prior to joining Cisco Systems, he was a Director in the Traffic Capacity Management and Planning Department at Telcordia Technologies (formally Bell Labs).

Dr. Rayes has authored / co-authored over a hundred papers and patents on advances in communications-related technologies, including a book on Network Modeling and Simulation and one on ATM switching and network design. He the president of the newly formed International Society of Service Innovation Professionals (www.issip.org), an Associate Editor of ACM "Transactions on Internet Technology" and Editor-in-Chief for "Advances of Internet of Things" Journal.

Dr. Rayes received his BS and MS Degrees in Electrical Engineering from the University of Illinois at Urbana in 1986 and 1988, respectively. He received his Doctor of Science degree in Electrical Engineering from Washington University in St. Louis, Missouri, in 1994 where he received the Outstanding Graduate Student Award in Telecommunications