Aligning Spectrum-User Objectives for Maximum Inelastic-Traffic Reward

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 objective functions for large-scale distributed dynamic spectrum access (DSA) networks that, by means of any learning algorithm, enable DSA users to locate and exploit spectrum opportunities effectively, thereby increasing their achieved throughput (or "rewards" to be more general). We show that the proposed functions are: (i) optimal 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) distributed by being implementable in a decentralized manner.**

I. Introduction

Federal Communications Commission (FCC)'s foreseeable approach for solving the spectrum shortage problem [1] is to promote dynamic spectrum access (DSA). The basic idea behind DSA is allow spectrum users (SUs) to seek and exploit the available spectrum bands (or channels) dynamically, thereby improving spectrum efficiency.

Due to its potentials, DSA has attracted the focus of many researchers during these past years, resulting in numerous works ranging from spectrum sensing techniques [2, 3] to protocol design and management strategies [4–7]. There have also been some research efforts on developing adaptive techniques that also promote DSA [8–10]. These mainly consist of first constructing channel/spectrum availability prediction models, and then, using these models to find the best spectrum opportunities. The challenge, 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 availability.

As a result, learning-based techniques which do not require prediction models, yet can still perform well by learning directly from interactions with the environment, are of a particular interest to DSA, and consequently, they have recently been the focus of many researchers [11–13]. Instead of using models, these techniques rely on learning algorithms (e.g., reinforcement learners [14] and evolving neuro-controllers [15]) to learn from past and present interaction experience to decide what to do best in the future. In essence, learning algorithms allow SUs to use their knowledge acquired from interaction with the environment to take 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 poorly designed objective functions, their collective behavior often leads to worsening each other's long-term cumulative rewards. Therefore, it is 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 work, we derive efficient objective functions that SUs can aim to maximize, and that by doing so, their collective behaviors also lead to good overall system performance, resulting in maximizing each SU's long-term cumulative received rewards. We show that our derived objective functions are: (i) *near-optimal*, in that they enable SUs to achieve high rewards; (ii) *very scalable*, in that they perform well in systems with a small as well as a large number of SUs; (iii) *highly learnable*, in that they allow SUs to reach up high rewards very quickly; and (iv) *distributive*, in that they can be implemented 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. We evaluate the performances of the proposed functions in Section IV, and finally conclude the paper in Section V.

II. Problem Statement

We consider a DSA network with m non-overlapping spectrum bands. We also consider a time-slotted system, where SUs 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 reinforcement learning algorithm [14] 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, these groups will also be referred to as *DSA agents* or simply *agents*.

At each time step, each agent using a band receives a service that is passed to it from that band. The service that the band offers can be measured in terms of, for example, amount of throughput, reliability of the communication, the signal to noise ratio, the packet success rates, etc. Let S_j represent the total amount of service that spectrum band j offers. We assume that once the agent switches to a particular band, it can immediately quantify and measure the service level that it receives from using such a band. 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.

A. *Inelastic Traffic Model*

In this work, we consider studying the *inelastic traffic model*. In this model, an agent receives a constant reward if it switches to a band that offers 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:

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

where $n_j(t)$ is the number of agents using band j at episode (time step) t, and β is a decaying factor. Note that when the number of agents using band j is greater than $c_j \equiv S_j/Q$, the reward decreases exponentially. This means that none of the agents will be satisfied with the amount of service they receive from band j if the band has more agents than c_j (c_j here represents the maximum number of agents band j can support while satisfying agents' required QoS levels).

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

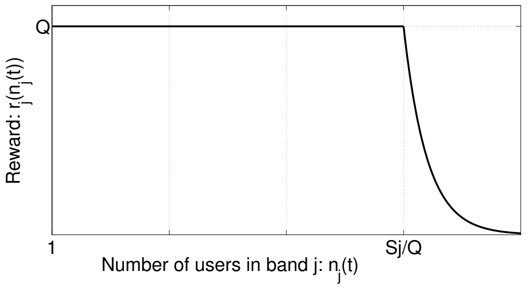


Fig. 1. Reward function: $\beta = 20$ and $S_j/Q = 4$ for all $j = 1, 2, \ldots, m$.

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

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

where again m is the number of spectrum 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 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.

We want to emphasize that the focus of this paper is not on learning, but rather on the design of efficient coordination techniques that can be used by any learning algorithms. However, for the purpose of evaluating our proposed techniques, we choose to use throughout this work the ϵ-greedy Q-learner [14] with a discount rate of 0 and an ϵ value of 0.05. Each agent is then assumed to use the Q-learner to implement and 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 (i.e., as a result of using the selected band), 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 [14] for more details on the Q-learner.

C. *Motivation and Objective*

The key question that arises naturally 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 that one can think of. One possible objective function choice 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

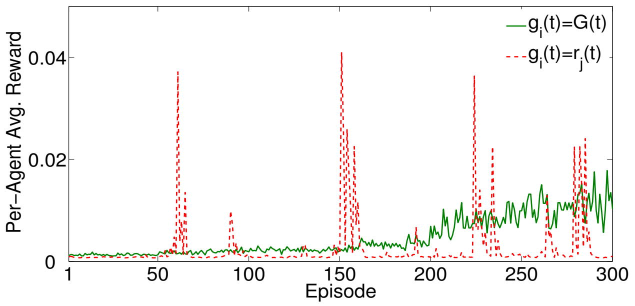


Fig. 2. Per-agent average achieved reward $\bar{r}(t)$ as a function of episode t under the two private objective functions: intrinsic choice ($g_i = r_j$) and global choice ($g_i = G$) for $Q = 2$, $\beta = 2$, $S_j = 20$ for $j = 1, 2, \dots, 10$.

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 presents an oscillating behavior: it ramps 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, which can be explained as follows. With the intrinsic objective 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 reward becomes high.

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, the increase in the received reward is relatively slow due to the function's insensitivity to one's actions, leading to slow learning rates.

Therefore, objective functions must be designed with two requirements in mind: (i) *alignedness*; when agents maximize their own private objectives, their collective behaviors should indeed result in increasing each agent's long-term received rewards, and not in worsening each other's received rewards, and (ii) *sensitivity*; objective functions should be sensitive to agents' own actions so that proper action selections allow agents to learn about good opportunities fast enough.

With this in mind, the objective of this work is 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 optimal 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. Objective Function Design

In this section, we first begin by presenting the factoredness and learnability concepts, both 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 be the function that DSA agent i aims to maximize, and that we want to derive. Let z characterize the joint move of all DSA agents in the system. Here, the global (total) 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 [16]:

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

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 private 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 DSA agent can discern the impact of its own actions on its private objective function, so that a proper action selection allows it to quickly learn about good spectrum opportunities. This means that the agent's objective function should be more sensitive to its own actions than the actions of other agents. Formally, the level of sensitivity or *learnability* of an objective function g_i, for agent i at z, can be quantified as [16]:

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

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. Objective Functions

The key challenge when 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. In general, a highly factored objective function has low learnability, and a highly learnable function has low factoredness [16].

To provide some intuition on how we designed our objective functions, we will visit the behaviors of the global reward function, illustrated earlier in Section II-C. Recall that when agents set the global reward G as their objectives (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.

Note that by removing the effects of all agents other than agent i from 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 at all; i.e., while still being fully factored. Formally, these private objective functions can be written

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

where z_{-i} again represents the parts of the state on which agent i has no effect. These difference functions have been successfully applied to other domains (e.g., multi-robot control [17] and air traffic flow regulation [18]). 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 state z, the objective function D_i for agent i selecting band j at time t can then be written as:

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

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 constitutes one important property among few others (to be described later) that this proposed function has.

C. Optimal User Distribution

In order to help us understand and explain the intuition behind the achievable performance of our proposed functions (to be presented later in Section IV), we will derive in this section the optimal/ideal 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/maximum 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 this work, we assume that $n > m\frac{S}{Q}$, and let $c = \frac{S}{Q}$, which denotes the capacity (i.e., maximum number of agents) of each spectrum band. We now present our result on the optimal agent distribution (the proof can be seen in [19]).

Proposition 3.1: 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.

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

IV. 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 consider 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. 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.

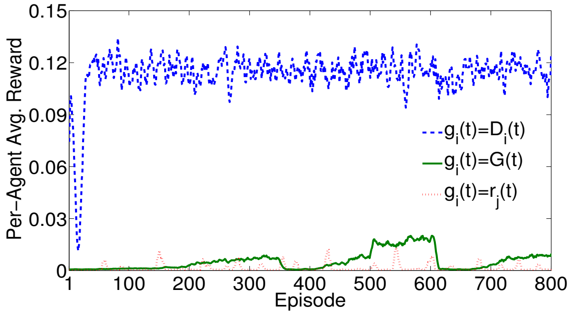


Fig. 3. Per-agent average achieved reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j.

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.

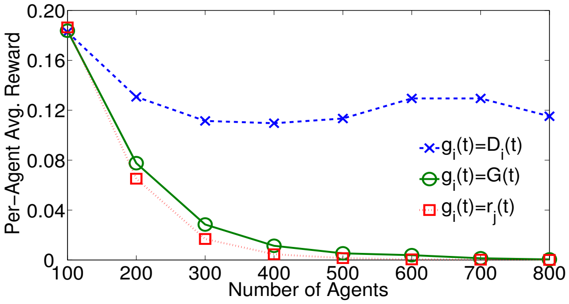
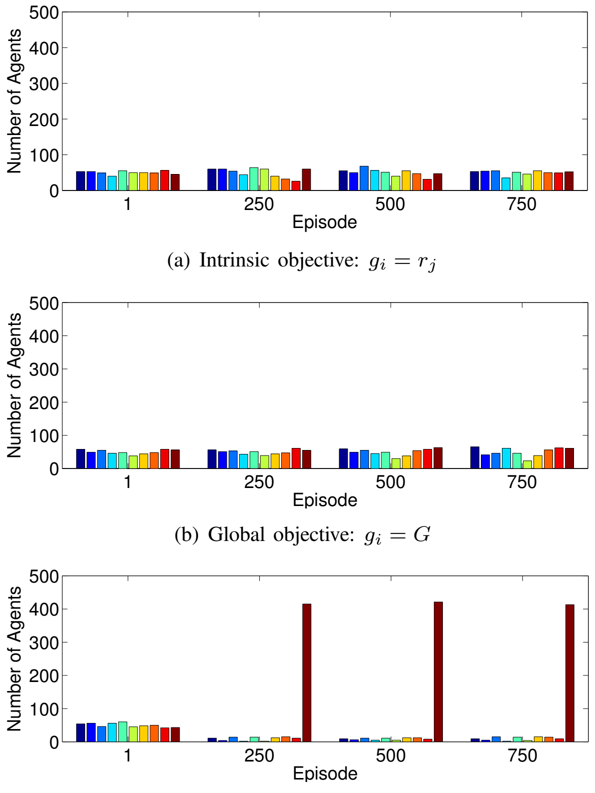


Fig. 4. Per-agent average achieved reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various numbers of agents: $Q = 2$, $\beta = 2$, $S_j = 20$ for all j.

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-C. Recall that the ideal/theoretical agent distribution, as stated in Proposition 3.1, 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 ($\approx$ 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 ($\approx$ 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 3.1. 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 3.1, yielding then near-optimal achievable rewards (as observed in Section IV-A).



(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 that, during this study, we observed 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.

V. Conclusion

In this paper, we propose scalable and distributive objective functions that DSA users can use to locate and exploit the best spectrum opportunities. We show that these proposed functions (i) achieve near-*optimal* rewards as they enable DSA users to receive high rewards, (ii) are highly *scalable* as they perform well for small- as well as large-scale DSA networks, (iii) are highly *learnable* as rewards reach up near-optimal values very quickly, and (iv) are *distributive* as they require information sharing only among users belonging to the same spectrum band.

References

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

[2] R. Fan and H. Jiang, "Optimal multi-channel cooperative sensing in cognitive radio networks," *IEEE Tran. on Wireless Comm.*, vol. 9, no. 3, March 2010.

[3] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal selection of channel sensing order in cognitive radio," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, Jan. 2009.

[4] B. Hamdaoui and K. G. Shin, "OS-MAC: An efficient MAC protocol for spectrum-agile wireless networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 8, pp. 915–930, August 2008.

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

[6] N. Chakchouk and B. Hamdaoui, "Traffic and interference aware scheduling for multi-radio multi-channel wireless mesh networks," *IEEE Tran. on Vehicular Tech.*, vol. 60, no. 2, Feb. 2011.

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

[8] 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.

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

[10] B. Hamdaoui, "Adaptive spectrum assessment for opportunistic access in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 922–930, Feb. 2009.

[11] 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.

[12] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. on Signal Processing*, vol. 58, no. 11, November 2010.

[13] P. Venkatraman, B. Hamdaoui, and M. Guizani, "Opportunistic bandwidth sharing through reinforcement learning," *IEEE Tran. on Vehicular Technology*, vol. 59, no. 6, pp. 3148–3153, July 2010.

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

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

[16] 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.

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

[18] 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.

[19] M. NoroozOliaee, B. Hamdaoui, and K. Tumer, "Achieving optimal elastic traffic rewards in dynamic multichannel access," in *Proceedings of IEEE Conference on High Performance Computing and Simulation*, July 2011.