Achieving Optimal Elastic Traffic Rewards in Dynamic Multichannel Access

M. NoroozOliaee, B. Hamdaoui, and K. Tumer
Oregon State University
noroozom,hamdaoub@onid.orst.edu ; kagan.tumer@oregonstate.edu

ABSTRACT

*This paper proposes objective functions for dynamic multichannel access (DMA) networks that enable spectrum users (SUs) to assess, locate, and exploit available spectrum opportunities effectively, thereby maximizing the SU's rewards measured in terms of the average received throughput. We show that the proposed objective functions are: near-*optimal*, as they achieve high rewards;* scalable*, as they perform well in small- as well as large-scale DMA networks;* learnable*, as they allow SUs to reach up near-optimal rewards very quickly; and* distributive*, as they are implementable by requiring local information sharing only.*

KEYWORDS: *Multichannel access; dynamic network resource sharing; elastic traffic; wireless networks.*

1. INTRODUCTION

Dynamic multichannel access (DMA) capability empowers spectrum users (SUs) to seek, locate, and use available spectrum bands (or channels) dynamically and opportunistically. DMA is promoted by FCC as a potential solution to the spectrum shortage problem [1] that FCC has recently observed [2]. As a result, there have been numerous publications addressing various DMA challenges, ranging from protocol and algorithm design [3–5] to channel selection and prediction technique development [6–8]. More recently, there have also been some research efforts on the development of learning-based techniques that also enable effective DMA by learning directly from interaction with the environment [9–11]. These techniques rely on learning algorithms (e.g., reinforcement learners [12]) 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 objective functions, thereby "hopefully" maximizing their long-term cumulative received rewards.

The key challenge that we address in this work is that when SUs' private objective functions are not carefully coordinated, learning algorithms can lead to poor overall performance. In other words, when SUs aim at maximizing their intrinsic (not carefully designed) objective functions, their collective behavior often leads to worsening each other's long-term cumulative rewards, a phenomenon known as the "tragedy of the commons" [13]. 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 paper, we derive efficient SU objective functions that are aligned with system objective, so that when SUs maximize them, their collective behaviors also lead to good system-level performance, thereby resulting in increasing each SU's long-term received rewards. Specifically, we propose objective functions that are (i) near-*optimal*, in that they allow SUs to achieve rewards close to the maximal/optimal achievable rewards, (ii) *scalable*, in that they perform well in systems with a small as well as a large number of SUs, (iii) *learnable*, in that they allow SUs to reach up near-optimal rewards very quickly, and (iv) *distributive*, in that they are implementable in a decentralized manner by relying on local information only.

The rest of the paper goes as follows. In Section 2, we describe our system model. Section 3 states our motivation and objective. In Section 4, we present our proposed techniques. In Section 5, we derive upper bounds on the maximal achievable rewards. In Section 6, we evaluate the proposed functions. Finally, Section 7 concludes the paper.

2. SYSTEM MODEL

We assume that spectrum is divided into m non-overlapping spectrum bands (or channels). We consider a time-slotted system, where SUs are assumed to arrive and leave at the beginning and at the end of time slots. An *agent* is a group of two or more SUs who want to communicate together.

In order to communicate with each other, all SUs in the group must be tuned to the same band. At the end of each time step, by means of a reinforcement learning algorithm [12], each agent selects the "best" available spectrum band, and uses it during the next time step. At each time step, each agent receives a service that is passed to it from the DMA system. One possible service metric is the amount of throughput that the visited spectrum band offers the agent. Another possible metric is the reliability of the communication carried on the spectrum band, which can be measured through, for example, SNR (signal to noise ratio), PSR (packet success rate), etc. What service metric to use and how to quantify it are beyond the scope of this work. Here, we assume that once the agent switches to a particular band, the received service level can immediately be quantified by monitoring the metric in question. Hereafter, we then assume that each band j is characterized by a value V_j that represents the maximum/total service level that the band can offer.

In this paper, we consider the elastic traffic model, where the agent's received reward (i.e., satisfaction) increases proportionally to the service it receives from using the spectrum band so long as the received quality-of-service (QoS) level is higher than a certain (typically low) threshold R. But when the received QoS level is below the threshold R, the agent's reward decreases rapidly (e.g., exponentially) with the received QoS level; i.e., the reward/satisfaction goes almost immediately to zero when the received QoS level is below R. This traffic model is suitable for elastic applications, such as file transfer and web browsing, where the higher the received service quality level, the better the quality/reward perceived by these applications. But when the received QoS level is below a certain low threshold (i.e., R), the quality of these applications becomes unacceptable. Formally, the reward $r_j[n_j[t]]$ (also often referred to simply as $r_j[t]$ for simplicity of notation) contributed by band j at time step t can be written as:

$$r_j[n_j[t]] = \begin{cases} V_j/n_j[t] & \text{if} \quad n_j[t] \leq V_j/R \\ Re^{-\beta\frac{n_j[t]R-V_j}{V_j}} & \text{otherwise} \end{cases} \quad (1)$$

where $n_j[t]$ denotes the number of agents that choose band j at time step t, and β is a reward decaying factor. Note that here we assume that the total service level V_j offered by any band j is split equally among all the $n_j[t]$ agents that use band j at time t.

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

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

where m is the number of spectrum bands. The per-agent average reward $\bar{r}[t]$ at time step t is then

$$\bar{r}[t] = \frac{\sum_{j=1}^{m} n_j[t] r_j[t]}{\sum_{j=1}^{m} n_j[t]} \quad (3)$$

3. MOTIVATION AND OBJECTIVE

The goal of this work is to design efficient objective functions for agents, so that when agents aim to maximize them, their collective behaviors lead to good system-level performance, thereby resulting in increasing each agent's long-term received rewards. Hereafter, let g_i denote agent i's objective function. Although the objective functions (g_i for agent i) that we derive in this paper are designed to be used by any learning algorithm, throughout this work, we choose to use the ϵ-greedy Q-learner [12] (with a discount rate of 0 and an ϵ value of 0.05) to evaluate the effectiveness of our developed functions. At each episode (or time step) t, each agent i aims then at maximizing its own private objective function $g_i[t]$ using its own Q-learner.

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 [12] for more details on the Q-learner.

3.1. Motivation

The key question that arises naturally is which objective function g_i should each 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 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.

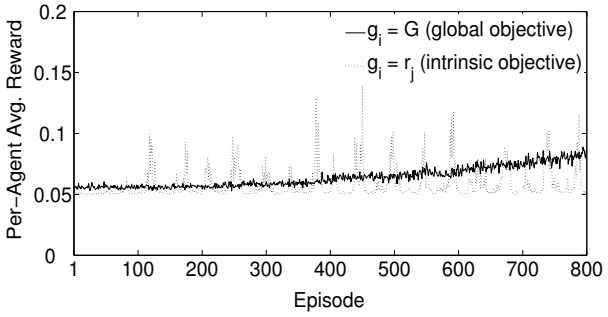


Figure 1. Per-agent average achieved reward $\bar{r}[t]$ as a function of episode t under the two private objective functions: intrinsic choice ($g_i = r_j$) and global choice ($g_i = G$) for $R = 2$, $\beta = 2$, $V_j = 20$ for $j = 1, 2, \ldots, 10$.

For illustration purposes, we plot in Fig. 1 the per-agent average received reward $\bar{r}[t]$ (measured and calculated via Eq. (3)) under each of these two private objective function choices. In this experiment, we consider a DMA system with a total number of agents equal to 500 and a total number of bands m equal to 10. 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 explains 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.

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, it is imperative that private objective functions be designed with two (usually conflicting) requirements in mind: (i) *alignedness*; when agents maximize their own private objectives, they should not end up working against one another; instead, their collective behaviors should result in increasing each agent's long-term 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.

3.2. Objective

Our goal here is to design efficient coordination techniques for large-scale DMA networks. Specifically, we devise private objective functions with 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 DMA systems 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.

Before delving into our function design, we want to emphasize that the focus of this work is not on learning, but rather on designing objective functions that can be used by any learners.

4. PRIVATE OBJECTIVE FUNCTIONS

For a private objective function to lead to a good overall system performance, two requirements must be met. First, we must ensure that an 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 need to be *aligned* or *factored* with the global reward function G. Intuitively, the more aligned an agent's objective function, 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.

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. Intuitively, more sensitive or *learnable* objective function means that it is easier for an agent to achieve higher rewards.

The challenge in designing objective functions for large-scale DMA systems is then to find the best tradeoff/balance between alignedness and sensitivity. Doing so will ensure that agents can learn to maximize their own objectives while doing so will also lead to good overall system performance; i.e., their collective behaviors will not result in worsening each other's received rewards. Throughout, let g_i denote the objective function of agent i that we aim to derive in this work.

In general, a highly aligned (or factored) private objective function will experience low sensitivity (or learnability), and a highly learnable function will have low factoredness [14]. Let us visit again the observed behaviors of the global reward function, illustrated in Section , to understand the intuition behind the design of our proposed functions. Recall that (as observed earlier in Section) 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 (though very slowly, see Fig. 1), because agents' private objectives are aligned with system objective. The issue, however, is that because G depends on (is impacted by) all agents, it is too difficult for an agent (using G as its objective function) to discern the effects of its own actions on its private objective, resulting then in low learnability.

The key observation leading to the design of our 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 higher learnability than G, yet without compromising its alignedness quality. Formally, these functions can be written as

$$D_i(z) \quad \equiv \quad G(z) - G(z_{-i}) \tag{4}$$

where z represents the full system state (i.e., joint move of all agents in the system), and z_{-i} specifies the parts of the system state controlled all agents other than agent i; i.e., z_{-i} represents the parts of the state on which agent i has no effect. These difference functions have also been shown to lead to good system performance in other domains, such as multi-robot control [15] and air traffic flow regulation [16]. First, note that these proposed functions (D_i for agent i) are fully factored, because the second term of Eq. (4) 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. (4), explicitly noting the time dependence t, and for clarity, removing the implicit dependence on the full state z, the objective function D_i for agent i selecting band j at time t can then be written as:

$$D_i[t] = n_j[t] r_j[n_j[t]] - (n_j[t] - 1) r_j[n_j[t] - 1] \tag{5}$$

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. (5)) depends on band j only. Therefore, the proposed function D_i is simpler to compute than the global function G. More 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 others (to be described later) that this proposed function has.

5. OPTIMAL ACHIEVABLE REWARDS

In this section, we derive a theoretical upper bound on the maximum/optimal achievable rewards. This upper bound will serve as a means of assessing how well the developed objection functions perform when compared not only with the two intuitive objective functions (intrinsic r_j and global G), but also with the optimal achievable performances.

Without loss of generality and for simplicity, let us assume that $V_j = V$ for $j = 1, 2, \cdots, m$. Let n denote the total number of agents in the system at any time. First, note that when $n \leq m\frac{V}{R}$, the maximum global achievable reward is simply equal to mV (assume $n \geq m$), which corresponds to having each band contain no more than $\frac{V}{R}$ agents. Therefore, in what follows, we assume that $n > m\frac{V}{R}$, and let $c = \frac{V}{R}$, which denotes the capacity (in terms of number of supported agents) of each spectrum band.

Now, we start by proving the following lemma, which will later be used for proving our main result.

LEMMA 1: The global received reward of an DMA system reduces less when a new agent joins a more crowded spectrum band than when it joins a less crowded band.

Proof. Recall that when a band j has $n' > c$ agents, its reward is $G_j(n') = n'Re^{-\beta(\frac{n'}{c}-1)}$. If a new agent joins this band, the new reward becomes $G_j(n'+1) = (n'+1)Re^{-\beta(\frac{n'+1}{c}-1)}$. 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. □

THEOREM 1: When there are n agents in the system, the global reward reaches its maximal only when $m-1$ bands (out of the total m 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 us refer to the agent distribution stated in the theorem 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 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

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

  which is greater than zero.

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

$$t' - t = \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i$$

  which is also greater than zero.

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

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

  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 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. □

COROLLARY 1: The per-agent average achievable reward is at most $(m-1)V/n + (R-(m-1)V/n)e^{-\beta(\frac{nR}{V}-m)}$.

Proof. The proof follows straightforwardly from THEOREM 1 by calculating the global achievable reward for the derived optimal agent distribution. □

Note that this upper bound (that we derived and stated in COROLLARY 1 is the maximum/optimal average reward that an agent can achieve (it is a theoretical upper bound). In the next section, we will evaluate the performances of the proposed objective functions in terms of their achievable rewards, and compare them against these optimal achievable performances.

6. PERFORMANCE EVALUATION

We now compare the performances of the proposed objective functions in terms of the per-agent average achievable rewards with the optimal achievable rewards calculated through COROLLARY 1 as well as with those achievable under each of the two intuitive functions r_j and G. In what follows, we set $R = 2$, $\beta = 2$, and $V = 20$.

6.1. Optimality

We first begin by considering the same experiment, conducted in Section , where the total number of agents is set to 500, and that of bands is set to 10. In Fig. 2, we show the per-agent average achievable reward normalized w.r.t. the optimal achievable reward under each of the three functions: intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed

($g_i = D_i$). The figure clearly shows that the proposed func-

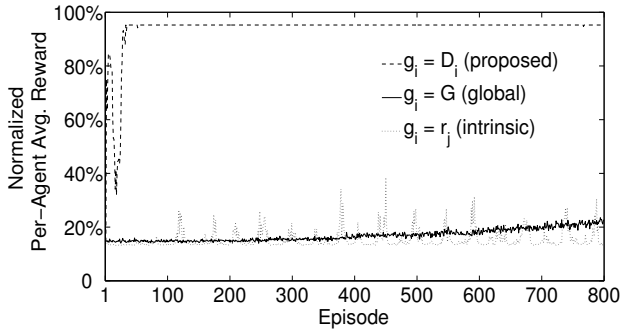


Figure 2. Per-agent average achieved reward normalized w.r.t. maximum achievable reward under intrinsic function ($g_i = r_j$), global function ($g_i = G$), and proposed function ($g_i = D_i$).

tion D_i achieves substantially much better performances than the other two. In fact, when using D_i, an agent can achieve up to about 90% of the total possible, achievable reward, whereas it only can achieve up to about 20% when using any of the other two functions. Another distinguishing feature of the proposed D_i function lies in its learnability; that is, not only does D_i achieve good rewards, but also does so quite fast, as the received rewards ramp up rapidly, quickly reaching near-optimal performance.

6.2. Scalability

We also study the proposed function with regard to another performance metric: scalability. For this, we plot in Fig. 3 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 $m = 10$ the same. Observe that D_i outperforms the other two functions substantially when it also comes to scalability. Note that D_i achieves high rewards, even for large numbers of agents, whereas the achievable reward under either of the other two functions drops dramatically with the number of agents. We therefore conclude that the proposed function D_i is very scalable, and works well in systems with small as well as large numbers of agents.

6.3. 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

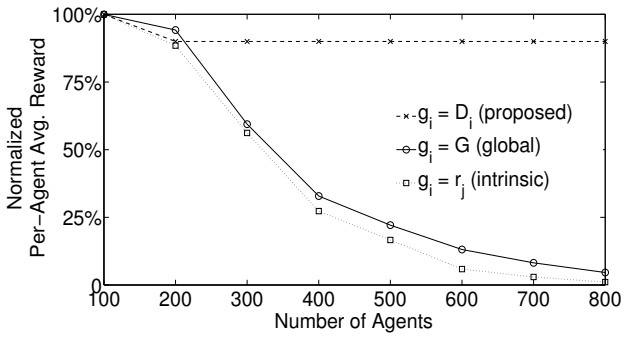


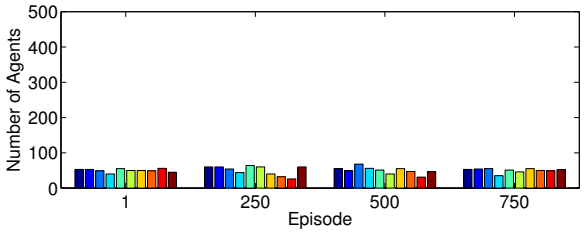
Figure 3. Per-agent average achieved reward normalized w.r.t. maximum achievable reward under intrinsic ($g_i = r_j$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various numbers of agents.

ideal/theoretical distribution derived in Section . Recall that the ideal/theoretical agent distribution, as stated in THEOREM 1, corresponds to the repartition that leads to the maximum achievable rewards. Therefore, comparing the agent distribution led to under D_i to the theoretical one reflects on how well D_i performs.

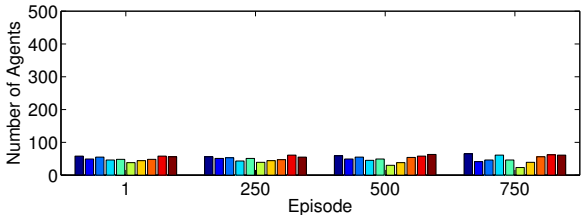
To illustrate, we plot in Fig. 4 the actual, measured distribution of the $n = 500$ agents across the $m = 10$ bands at different times (i.e., every 250 episodes) under the three studied objective functions. Note that in the case of r_j (Fig. 4(a)) and G (Fig. 4(b)), agents are (approximately) equally distributed among the 10 bands (≈ 50 agents/band), and at all times. But when using D_i (Fig. 4(c)), 9 bands out of 10 each contains about 10 agents, which represent the capacity $c = \frac{V}{R}$, 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 THEOREM 1. Thus, the proposed function, D_i, when used as an objective function, leads to a near-optimal agent distribution, yielding then near-optimal achievable rewards (as observed in previous sections).

7. CONCLUSION

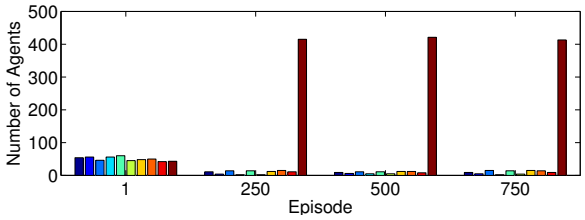
This paper derives scalable and distributed private objective functions for supporting elastic traffic in multichannel access networks. Spectrum users can rely on any learning algorithms to maximize these proposed objective functions, thereby ensuring near-optimal performances in terms of the long-term average received rewards. We showed that these proposed functions (i) receive near-*optimal* rewards, (ii) are highly *scalable* as they perform well for small- as well as large-scale systems, (iii) are highly *learnable* as rewards



(a) Objective function: r_j



(b) Objective function: G



(c) Objective function: D_i

Figure 4. Distribution of the 500 agents among the $m = 10$ different bands. Each bar corresponds to one band.

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] *FCC, Spectrum Policy Task Force (SPTF), Report of the Spectrum Efficiency WG, Report ET Docet no. 02-135, November*, 2002.

[3] 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.

[4] 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 Technology*, vol. 59, no. 1, 2010.

[5] N. Chakchouk and B. Hamdaoui, "Traffic and interference aware scheduling for multi-radio multi-channel wireless mesh networks," *IEEE Tran. on Vehicular Technology*, vol. 60, no. 2, pp. 555–565, Feb. 2011.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, pp. 1243–1248, 1968.

[14] 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.

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

[16] 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.