Self-Coexistence in Cognitive Radio Networks using Multi-Stage Perception Learning

Deepak K Tosh
Department of Computer Science
Graduate Center, CUNY
New York, NY 10016
Email: dtosh@gc.cuny.edu

Shamik Sengupta
Department of Math. & Comp.Sc.
John Jay College, CUNY
New York, NY 10019
Email: ssengupta@jjay.cuny.edu

***Abstract*—In this paper, we study self-coexistence problem the competitive Cognitive Radio networks in an uncoordinated distributed environment from the perspective of homogeneous and heterogeneous bands. This problem imitates the characteristics of famous optimal foraging theory, where the humming birds forage to explore an island of maximum food source for survival. The learning technique from observations leads them all to reach an island of optimal resources. In the self-coexistence problem of homogeneous bands, the individual networks are incorporated with a perception based learning mechanism to strategize their actions based on rewards obtained from the accessed spectrum bands and eventually get a clear spectrum band to operate. However, in case of heterogeneous bands, rationality of networks lead them to collision, so we propose a regret minimization heuristic which incorporate the perception learning to resolve the contention among networks and achieve maximum system utility. Experimental results conclude that the networks can quickly acquire a free spectrum band using perception based learning in homogeneous bands and achieve maximum system utility by applying strategies from regret minimization heuristic in heterogeneous bands.**

I. Introduction

Over the past decade, wireless services and applications are exponentially expanding their horizon along with population of users. Thus the scarce spectrum resources need to be managed intelligently for maximizing the spectrum utilization and ensure better quality of service (QoS) to the users. Because a significant amount of licensed spectrum is underutilized in most of the regions due to fixed spectrum allocation policy, dynamic spectrum allocation (DSA) [2] technology offers a better spectrum sharing paradigm to alleviate spectrum scarcity problem. The DSA technology allows unlicensed wireless users (secondary users) to access the licensed spectrum bands of legacy owners opportunistically in a non-interfering basis. This functionality is best exploited by the cognitive radios (CR) [1] which have been introduced to alleviate the spectrum scarcity problem and maximize the spectrum efficiency. The ability to sense, learn, and tune the transmission parameters in CR helps to achieve better QoS by accessing best available spectrum bands opportunistically.

In a distributed uncoordinated wireless environment, the cognitive radio networks fight for spectrum resources aiming to minimizing interference with neighboring CR networks which ultimately improves the overall quality of service (QoS) of the system. To coordinate each CR networks along with their rationality towards finding best available spectrum band in this bazaar environment is promising as well as challenging. Thus strategic thinking in CR agents can help them to coordinate and maintain self-coexistence. Most of the recent works on CR has been concentrating on primary–secondary spectrum etiquette, spectrum sensing, primary user detection techniques etc., but not much work has been done on the self-coexistence issues, and more importantly self-coexistence in heterogeneous spectrum scenario. In [4] the issues of self-coexistence in IEEE 802.22 networks are addressed and a utility graph coloring technique was proposed to allocate spectrum bands to the base stations for the purpose of coexistence. The distributed self-coexistence problem was solved using techniques of game theory in [5] where the problem was modeled as modified game theory and ultimately an approximate close form solution was found to the mixed strategy Nash Equilibrium for homogeneous bands. A Coexistence-Aware Spectrum Sharing (CASS) protocol was proposed in [7] which minimizes the self-interference with minimum control overhead. To improve spectrum utilization a round-robin based resource allocation algorithm was proposed for IEEE 802.22 WRAN in [8] which maintains the fairness of resource allocation and improves spectrum utilization.

In this paper, the problem of self-coexistence among CR networks is studied in a competitive environment where network competes with each other to acquire a contention free spectrum band and maximize the overall system utility assuming no primary user activity in the bands. This problem can be correlated to the famous optimal foraging model [13][12], where a species of birds forage over different islands to find food source by investing their energy in flying. The birds face a trade-off between energy gain from the islands and total foraging/scavenging period so that it can stay alive for long. As there might be more species scavenging for food in the same island, the bird's foraging period is affected with this contention too. Here the network players (bird) scavenge for spectrum bands (islands) in an uncoordinated manner to maintain QoS. In this distributed environment, each network can only observe its own actions and payoffs, and should learn to adapt in such situation so that its overall gain is maximized over a long run. The islands of food source can be of two types: homogeneous islands have same amount of food source, heterogeneous islands have variable amount of food source.

Similarly, the spectrum resources can also be categorized as homogeneous or heterogeneous types.

The contributions in the paper are as follows: (1) For homogeneous bands, we have presented a perception based learning model which helps networks building perception about spectrum bands by observing the payoffs and maps their perception vector to corresponding strategy vector to decide whether to explore or exploit in the next scavenging period. With this model, the networks could ultimately coexist within minimum game stages of interaction without interfering with each other. (2) For heterogeneous bands, the networks aim to achieve two goals simultaneously: get a contention free band and the average utility reward over the stages must be maximized in the non-cooperative simultaneous move game. The individual players[1] must play the game in distributed manner towards maximization of individual payoff where no central system exists to provide any kind of feedback rather they should learn from their previous history. To achieve optimal system utility in this scenario of heterogeneous bands, we present a regret minimization heuristic.

The rest of the paper is organized as follows. The system assumption and description along with mathematical model of self-coexistence problem for homogeneous bands with mutual exclusive access is elaborated in the section 2. We also present the perception based learning model in this section. In section 3, we study the self-coexistence problem in heterogeneous bands and propose a regret minimization based heuristic to achieve optimal system utility. The experimental details and results of the conducted simulations are analyzed in section 4. Finally concluding remarks are presented in the last section.

II. Self-Coexistence in Homogeneous Bands

A. System Description

In this work, the self-coexistence problem is modeled as a dynamic multi-stage interaction game where N cognitive radio networks act as players of the game and they compete for accessing M distinct orthogonal spectrum bands. The players are assumed rational and homogeneous in nature, i.e., their strategy space is same. The individual networks are aiming to access exclusively one out of M spectrum bands which are assumed to be free of contention from primary (licensed) users in the foraging period. In this work, we study this foraging game based on homogeneous and heterogeneous resources, because, when islands are homogeneous and provide identical amount of food source, rational birds will try to find any island where no other entity is scavenging, however when islands provide non-identical amount of food source, birds will optimally forage to find the best possible island which will tend to satiate the bird's need. The heterogeneous bands are distinguished based on factors like bandwidth, data rate, operating frequency range etc. So self-coexistence in heterogeneous bands will be interesting to analyze because the rational player's foraging behavior will be different from homogeneous case.

The networks in the game play in a non-cooperative manner towards acquiring a contention-free spectrum band along with maximizing the system utility which can be achieved by maximizing individual utility over the total scavenging period. Initially each network randomly selects one out of M spectrum bands to scavenge, and observes the payoff out of the band. This incentive is used to build perception about the spectrum bands foraged by the network. As per our assumption, if more than one network forages the same spectrum band simultaneously, they fight to acquire it which rewards nothing to the networks. Hence interference in the spectrum bands leads them all to play subsequent stage of the game where each network must choose one of the M possible actions according to observed perception about the spectrum bands. In the following subsections, we formally describe the game settings for homogeneous band based self-coexistence problem and a perception based learning model for all networks to optimally scavenge to acquire a band. Later, in section III, we will enhance our model to present the self-coexistence problem from heterogeneous perspective.

B. Game Settings for homogeneous band based self-coexistence problem

For the homogeneous scenario, it is considered that each network tries to maximize its own average utility by acquiring a free band of fair utility reward. Accessing a particular band by multiple networks results in no reward to each of the contending networks. Each network i has a mixed strategy space to play in this game: $p_i = (p_1^{(i)}, p_2^{(i)}, ..., p_M^{(i)})$, where $0 \leq p_j^{(i)} \leq 1$ is the probability of network i choosing the band j to operate and $\sum_{j=1}^{M} p_j^{(i)} = 1$. Because the bands are assumed to be homogeneous in nature, there will be a constant utility out of it on each access. Thus the utility function for player i can be defined as follows:

$$U_i(a_i, a_{-i}) = \begin{cases} \alpha & \text{if } a \neq a_i,\ \forall a \in a_{-i} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where α is a constant utility for all homogeneous spectrum bands.

According to above utility function, if a network i chooses band j to operate, then it obtains a constant utility of α provided no other network has chosen the same band j. If two or more networks choose the same band j out of M bands, then all networks receive no reward, rather there exist some cost for initiating communication and sensing the band for availability. In foraging process, it is assumed that the birds do not have any information on actions or strategies of other birds while deciding which island to forage. Similarly each network takes independent action based on its perception about spectrum bands. The decision of a network at the end of a stage can be either to stick with currently chosen band and exploit it, if the network player is satisfied with the current gain, or choose another band to explore more. The criteria for

[1]We use the words "player", "network" interchangeably throughout the paper.

choosing a particular band at the end of a game stage is based on the following described perception based learning model.

C. Perception based Learning Model

This learning model helps the players to build belief/perception about the accessed spectrum bands based on utility out of it. The perception of a network can be interpreted as a metric for categorizing the spectrum bands based on the returned utility reward. By looking to the perception values, a network can infer about utility rewards as well as probability of interference of a spectrum band. Each network $i \in N$ maintains a perception vector, $P^{(i)} = (P_1^{(i)}, P_2^{(i)}, P_3^{(i)}, ..., P_M^{(i)})$, about all M bands, and updates j^{th} entry after accessing to band j. The perception $P_j^{(i)}$ of a player i about a band j is mapped to player i's mixed strategy. Based on the generated strategy, the network takes stochastic decision about choosing any band $j \in M$, for next game stage. Based on observed reward from band j by network i, perception value band j, $P_j^{(i)}$ is updated. All networks aim to get a band free of contention for the purpose of co-existence with less number of game interactions.

In the starting of foraging stage $t = 0$, network $i \in N$ chooses a band $j \in M$ randomly to operate. The observed utility in stage t is defined to be $U_{i,a_i(t)}(t)$. As there exist no history about game until stage $t = 0$, we set the initial perception vector of player i $(P^{(i)}(0))$ to a small constant which implies no biasness towards any particular bands. In many past literatures on reinforcement learning[10][11], the Q-parameters for actions are estimated from the player's experience. And the Q-values are mapped to the player's mixed strategy based on Boltzmann distribution which is a common softmax method for controlling the exploration in a large search space using a controlling parameter named temperature ($\gamma > 0$). In our model, the perception vector is obtained according to the player's experience on actions. Thus the perception about choosing action in next stage can be mapped to corresponding mixed strategy using Boltzmann distribution policy. Hence the perception vector $(P^{(i)}(t))$ of network i is mapped to its corresponding mixed strategy, $p_i(t) = (p_1^{(i)}(t), p_2^{(i)}(t), ..., p_M^{(i)}(t))$ according to equation 2.

$$p_j^{(i)}(t) = \frac{e^{\gamma P_j^{(i)}(t)}}{\sum_{i=1}^{M} e^{\gamma P_j^{(i)}(t)}}, \forall j \in M \tag{2}$$

where γ is the temperature parameter in Boltzmann's distribution. γ controls the exploration of strategy space of a player. The value of γ changes over game stages as the experience of player increases. Initially the γ can be set to low value which emphasize each actions to be chosen with equal probability. Later the value of γ can be increased so that the stochastic exploration will be reduced and networks will incline more towards exploitation of bands that have high perception value.

After mapping the perception vector to their corresponding mixed strategy, a stochastic action is taken for each network i, to decide the operating spectrum band for the next game stage. The players observe the utility reward for the previously taken action and update their perception vector based on the reward obtained in the current period. The updated perception for network i about a band j for stage $t+1$ is given in equation (3). According to the expression (3), the network i has already played t stages and recorded the perception vector over the stages. At the end of each stage t, the networks update their perception vector which will be used as decision criteria for choosing a band in the next game stage $(t + 1)$. If network i has successfully acquired a band j by taking action $a_i(t) = j$ in game stage t, then the perception $(P_j^{(i)}(t))$ about the band j must increase proportionate to utility reward from that band for stage $(t + 1)$. And for unsuccessful possession of band j will lead to decrease in perception value of that band which is expressed in the first case of eqn. (3). The perceptions of those bands which are not accessed in stage t remain unaltered. Algorithm 1 summarizes the distributed procedure for self-coexistence with homogeneous bands using perception based learning model.

$$P_j^{(i)}(t+1) = \begin{cases} (1-\mu_t)P_j^{(i)}(t) + \mu_t U_{i,j}(t) & \textbf{if } a_i(t) = j \\ P_j^{(i)}(t) & \textbf{otherwise} \end{cases} \tag{3}$$

where $\mu_t \in (0, 1)$ is the smoothing variable factor which changes over the stages. Initially the value of μ is set to be high by the system which allows the network players to explore the strategy space. Gradually, the value of μ is decreased so that the networks will settle down on the particular band whose perception value is high.

Algorithm 1: Algorithm for self-coexistence based on Perception learning

```
Initialize the temperature γ ;
Initialize P_j^(i)(0) = 1/M for all networks i ∈ N and
  j ∈ M;
while stage t ≤ MaxT do
    for all network i ∈ N do
        Select a band j ∈ M based on its mixed strategy
        equation (3);
        Observe the utility reward for the stage t,
        U_{i,a_i(t)}(t);
        Update the perception (P_j^(i)(t+1)) for all bands
        j ∈ M according to equation (4);
        t ← t + 1;
    end
end
```

III. Self-Coexistence in Heterogeneous Bands

A. Game Settings for heterogeneous band based self-coexistence problem

The game settings for this scenario is similar to the case of homogeneous band based self-coexistence but here the utility of each band is assumed to be distinct. Unlike all bands provide constant utility, here bands present distinct utility reward $u_1, u_2, ..., u_M$, to the networks provided there exist

no contention. Hence the utility function for network i can be defined as

$$U_i(a_i, a_{-i}) = \begin{cases} u_j & \text{if } a \neq a_i,\ \forall a \in a_{-i} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Though this problem seems to be fairly similar to previously discussed problem, it has more importance from the rationality perspective of network players in a distributed uncoordinated wireless environment. As all network players in the game are rational, they always look for self-betterment and try to achieve a band that gives high utility reward. Thus every network will always eye upon the highest utility band which will lead them all to collision resulting in failed transmission instead of any reward and incurring unnecessary switching cost which eventually reduces the payoff. Hence the networks need to adopt some strategy so that they will get a contention free band of fair utility within few game stages and the system utility is maximized. In case of homogeneous bands, if networks could able to find an unused band then system utility is maximized because each band rewards constant utility. However, for heterogeneous scenario, to achieve maximum system utility in the non-cooperative simultaneous game is challenging. It can only be achieved when networks will explore all available spectrum bands and use some experience based heuristic to learn the availability of bands with high reward. Ultimately the networks will neglect the bands that return low utility reward and acquire a band with fairly high utility reward by which system utility can be maximized. To study this problem we propose a heuristic that uses the previously described perception learning model along with a regret minimization technique.

B. Regret minimization model

As utility rewards out of the spectrum bands are different, the rational network players will tend to choose the band that returns highest reward. Thus no single network will be able to successfully grab the channel and get penalized. Hence each network must strategize intelligently to decide which band to access so that the payoff over long run will be maximized and reach to optimal convergence point where the overall system utility is maximized. Before reaching optimal state, networks may get contention-free band where overall system utility is not necessarily maximum, we call it as sub-optimal convergence point.

The following regret minimization [9] heuristic is proposed to achieve optimal system utility by N networks by acquiring best N bands out of M-heterogeneous bands assuming all networks have information about the total available bands and the utility reward from each of the bands. To achieve this, the networks use regret matching technique to maintain the regret difference of actions that would have given more utility reward than currently taken action. Thus the strategy of choosing an action $\bar{\alpha}$ by network i in stage $(t+1)$, must be a function of average regret accumulated for the current action. This regret function will be able to lead the networks to choose actions that result high reward. In the same way, all other networks will choose for the highest utility band to operate, but due to collision no one will be able to operate on that band. To control these collisions, we must use the perception vector which will reduce the probability of choosing the same action over number of collisions. This will finally lead all networks to stick to the bands that return fairly high utility reward after certain stages. Thus strategy of choosing an action $\bar{\alpha}$ for stage $(t+1)$ should be a function of the regret difference $(R_i^{\bar{\alpha}}(t))$ and perception $(P_{\bar{\alpha}}^{(i)}(t))$ of network i up to game stage t which is presented in equation 5.

The average regret $(R_i^{\bar{\alpha}})$ accumulated for network i, for all actions, $\bar{\alpha} \in A_i$ up to stage t is given by

$$R_i^{\bar{\alpha}}(t) = \left(\frac{1}{t}\right) \sum_{t'=1}^{t} \left[U_{i,\bar{\alpha}}(t') - U_{i,\alpha}(t')\right]$$

where $U_{i,\alpha}(t)$ is the utility reward to network i by choosing action $\alpha \in A_i$ at stage t.

The action for network i for stage $t+1$ can be taken based on the following probability distribution $p_i^{\bar{\alpha}}(t+1)$ which rely on the accumulated regret difference and perception about all actions over previous t stages. In eq. 5, the normalized regret difference contributes in leading the networks to choose higher utility bands, however the normalized perception value will control the number of collisions by reducing the probability of action $\bar{\alpha}$.

$$p_i^{\bar{\alpha}}(t+1) = \frac{R_i^{\bar{\alpha},+}(t)}{\sum_{\bar{\alpha} \in A_i} \left[R_i^{\bar{\alpha},+}(t)\right]} * \frac{P_{\bar{\alpha}}^{(i)}(t)}{\sum_{\bar{\alpha} \in A_i} \left[P_{\bar{\alpha}}^{(i)}(t)\right]} \quad (5)$$

where $R_i^{\bar{\alpha},+}(t) = max(R_i^{\bar{\alpha}}(t), 0)$

IV. Simulation and Results

In this section, we report the simulation results for the problem of self-coexistence in homogeneous as well as heterogeneous networks using perception based learning model, and regret based heuristic. The simulations are carried out using Matlab version 7.9 with the following parameter settings. The total number of networks in the game and spectrum bands is assumed to be 150. The utility reward (α) for homogeneous bands is assumed to be 1. To achieve a good convergence we have allowed all networks to play for 300 stages at max. The trade-off analysis between exploration parameter (γ) and the system utility is presented in figure 1 where the number of networks and bands is set as 100. The value of γ is varied from 0 to 20 to observe the effect of exploration parameter over the system utility. It can be observed that for small values of γ the networks explore the bands to get knowledge about interference in the band and gradually settle down on the highest perception band. Thus with small values of exploration parameter (γ) the system utility is not necessarily maximum.

To analyze the effect of switching cost (c) on overall system utility, we simulated for different values of switching cost and plotted the result in figure 2. In competitive scenario, some networks which did not find a free band, switches many times because of collision. The switching cost will decrease the average utility and perception vector of each action taken in

the game stages. Thus it takes more number of stages required to find free spectrum band with high perception value. With increasing switching costs, the system utility achieved is not necessarily maximum all the time.

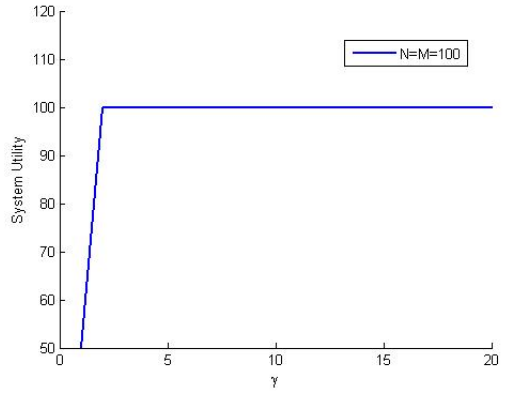

Fig. 1: Trade-off between System utility Vs γ

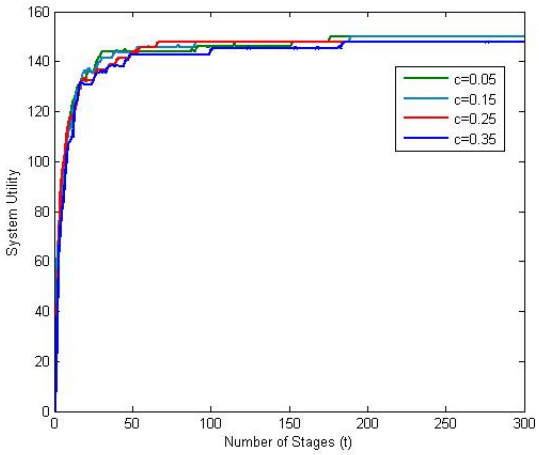

Fig. 2: Trade-off between System utility Vs switching cost(c)

To show the variability of system utility as the number of networks (N) in the game varies, we experimented by fixing the number of available spectrum bands (M) to be 150, utility reward of the spectrum bands is assumed to be unity. For variable number of network players, we run the algorithm for 1000 times and the average result is reported in figure 3. It is found that as the when number networks (N) is small compared to number of free spectrum bands (M), the networks can get a contention-free spectrum band easily within few stages of interaction because few networks compete for many available resources. But when N is close to M, some networks find difficulty in acquiring a contention-free band, therefore the convergence to highest system utility takes more number of stages in the game.

For self-coexistence in heterogeneous bands, we have simulated the regret minimization based heuristic to achieve optimal system utility where all networks aim to occupy a

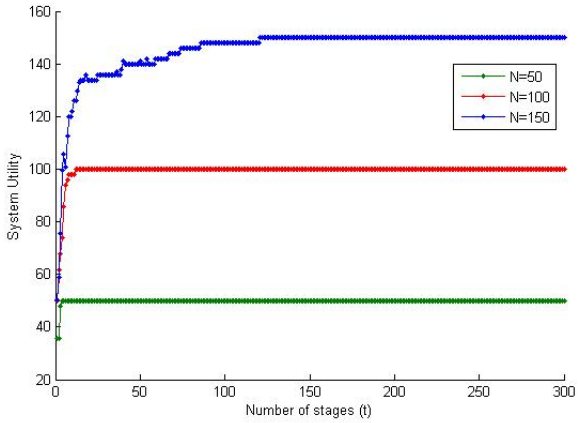

Fig. 3: System utility Vs Number of stages for varying Networks

band with fairly high utility reward. So we experimented with variable number of networks and spectrum bands to analyze the performance of our heuristic. All the simulations are executed for 100 times, and the average results are presented.

To show how convergence is affected with variable of number of networks by fixing available bands (M) to 20, we experimented with various N values starting from 5. From figure 4, it can be observed that when the number of networks (N) in the game is fewer than available bands (M), it is very easy to find a free spectrum band. Thus for sub-optimal convergence, the number of stages required is comparatively less. To achieve optimal system utility, the networks must compete hard with other networks in fairly large strategy space for enough number of stages until they have build a better perception value on a higher utility bands. But the important achievement here is that, the networks will ultimately acquire a band such that overall system utility is maximized. As the number of networks approaches the number of available bands, the number of game stages for convergence to optimal system utility must rise to resolve the competition among each other and adapt to a band of preferably high utility reward.

In figure 5, we fixed the number of networks (N) to 10 and varied the number of available spectrum bands (M) from 15 to 40. The average optimal and sub-optimal convergence period is reported here. When number of bands surpasses the number of networks, the CR network find more free resources to use without contending others. So the probability of finding a free band is more with increasing number of spectrum bands, thus sub-optimal convergence period decreases. However with increase in more spectral resources, networks have to scavenge more to build their belief about the resources. Thus the foraging period to optimally select a band that provides fairly high utility increases with number of available bands.

Finally we analyzed the number of stages for convergence with the following ratio mix: ratio of number of networks and number of spectrum bands is 0.5 and 0.75. From the plot presented in figure 6, it can be observed that, with ratio mix of

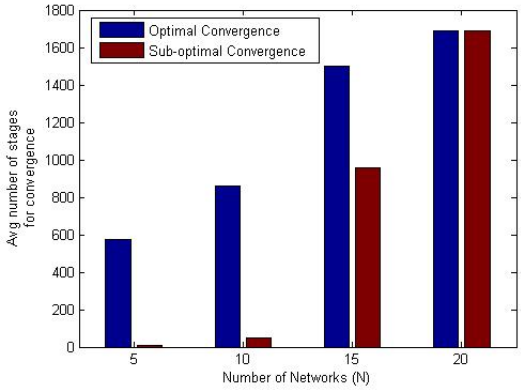

Fig. 4: Average number of stages for convergence Vs Number of Networks (N)

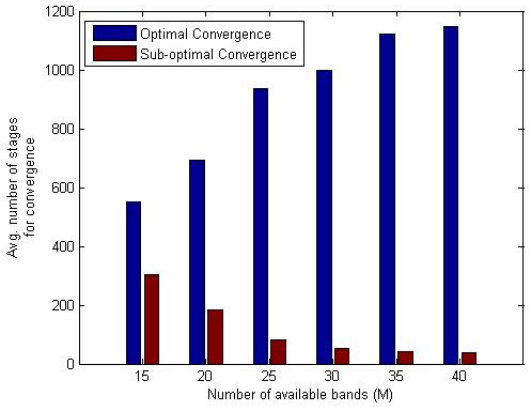

Fig. 5: Average number of stages for convergence Vs Number of Bands (M)

50%, the number of stages required to optimal convergence is less compared to ratio mix of 75% because competition among networks increases with increase in ratio mix. However, to find sub-optimal convergence by networks in case of 50% ratio mix, they require a few game stages. In case of 75% ratio mix, to achieve sub-optimal system utility, networks need to compete more hence require more number of stages to find a contention free band.

V. Conclusions and Future Work

In this work, we have studied the issues of self-coexistence in cognitive radio network, which is an important aspect for maximizing spectrum utilization. We modeled the problem of self-coexistence as a standard multi-stage interaction game between N networks, competing for M homogeneous or heterogeneous bands, which was motivated from the famous optimal foraging model. We presented a perception based learning model which uses the past belief and utility reward to take decision to choose bands in future stages. As shown in simulation results, this learning model helps networks to quickly learn and stick to best possible band according

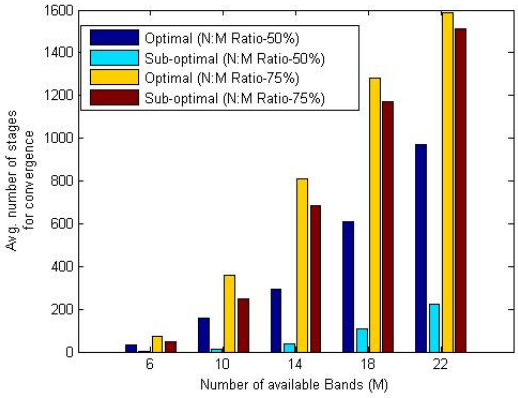

Fig. 6: Average number of stages for convergence Vs Number of Bands for 50% and 75% N:M ratio

to the perception vector about all bands. We also define the importance of self-coexistence in heterogeneous bands. Aiming to achieve optimal system utility, a regret minimization heuristic was proposed that applies regret matching as well as perception model to maximize system utility. In future, we will work on the regret minimization heuristic to make it scalable. And we aim to apply exploration technique to this heuristic which will enable the networks to achieve maximum system utility in a variable spectrum opportunity environment and able to converge with minimum possible scavenging period.

References

[1] J. Mitola III. An integrated agent architecture for software defined radio. Ph.D. dissertation, Royal Institute of Technology (KTH) (May 2000)

[2] I. Akyildiz, W. Lee, M. Vuran, and S. Mohanty. Next generation/dynamic spectrum access/ cognitive radio wireless networks: a survey. Computer Networks, pp. 2127–2159, 2006.

[3] ETRI, Channel Management in IEEE 802.22 WRAN Systems, 2010.

[4] S. Sengupta, S. Brahma, M. Chatterjee, and S. Shankar, Enhancements to Cognitive Radio Based IEEE 802.22 Air-Interface. IEEE ICC, pp. 5155–5160, 2007.

[5] S. Sengupta, R. Chandramouli, S. Brahma, and M. Chatterjee, A game theoretic framework for distributed self-coexistence among IEEE 802.22 networks. In proceedings of IEEE Global Communications Conference (GLOBECOM), pp. 1–6, Nov.-Dec. 2008.

[6] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. Econometrica, pp. 1127–1150, 2000.

[7] K. Bian and J. Park, A Coexistence-Aware Spectrum Sharing Protocol for 802.22 WRANs. In Proceedings of International Conference on Computer Communications and Networks (ICCCN '09), pp. 1–6, 2009.

[8] M. Yoo and S. Hwang, A Self-Coexistence Method for the IEEE 802.22 Cognitive WRAN. In Recent Researches in Automatic Control, Systems Science and Communications.

[9] J. R. Marden, G. Arslan, and J. S. Shamma, Regret based dynamics: convergence in weakly acyclic games. In Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems (AAMAS '07), pp. 42:1–8, 2007.

[10] C.J.Watkins, Models of Delayed Reinforcement Learning. PhD thesis, Psychology Department, Cambridge University, 1989.

[11] A. Bab, R. Brafman, Multi-agent reinforcement learning in common interest and fixed sum stochastic games: an experimental study. Journal of Machine Learning Research, 9: 2635-2675, 2008.

[12] C. L. Gass, J. S. E. Garrison, Energy regulation by traplining hummingbirds. Functional Ecology, 13: 483492, 1999.

[13] E. L. O'Brien, A. E. Burger, R. D. Dawson, Foraging Decision Rules and Prey Species Preferences of Northwestern Crows (Corvus caurinus). Ethology, 111: 77–87, 2005