Reinforcement learning based distributed multiagent sensing policy for cognitive radio networks

Jarmo Lundén*†, Visa Koivunen*†, Sanjeev R. Kulkarni†, and H. Vincent Poor†

*Aalto University, SMARAD CoE,
Department of Signal Processing and Acoustics, Finland
Email: {jrlunden,visa}@wooster.hut.fi

†Princeton University,
Department of Electrical Engineering, Princeton, NJ, USA
Email: {jlunden,koivunen,kulkarni,poor}@princeton.edu

***Abstract*—In this paper a distributed multiagent, multiband reinforcement learning based sensing policy for cognitive radio ad hoc networks is proposed. The proposed sensing policy employs secondary user (SU) collaboration through local interactions. The goal is to maximize the amount of available spectrum found for secondary use given a desired diversity order, i.e. a desired number of SUs sensing simultaneously each frequency band. The SUs in the cognitive radio network make local decisions based on their own and their neighbors' local test statistics or decisions to identify unused spectrum locally. Thus, the network builds a locally available map of spectrum occupancy of its geographical area. Simulation results show that the proposed sensing policy provides a significant increase in the amount of available spectrum found for secondary use compared to a random sensing policy.**

I. Introduction

In a cognitive radio ad hoc network [1] there is no pre-established infrastructure to centrally manage the network operations, such as spectrum sensing and access. Moreover, the network topology may change rapidly because of mobility or nodes joining or leaving the network. Hence, distributed algorithms are extremely well suited for identifying and exploiting free spectrum and managing the network. In this paper a distributed multiagent, multiband reinforcement learning based sensing policy for cognitive radio ad hoc networks is proposed.

Reinforcement learning has shown great potential in many applications for improving performance through experiment and experience [3], [11]. Reinforcement learning based spectrum sensing and access policies have been proposed in [2], [5], [8]–[10], and [13]. The sensing and access policies proposed in [2] and [5] aim at balancing between sensing, transmission, and switching the frequency band that a secondary user (SU) exploits. In [10], a two-stage reinforcement learning based collaborative fusion center controlled multiband sensing policy is proposed. The proposed policy optimizes the frequency bands to be sensed as well as the SUs sensing each frequency band. A multiagent, multiband spectrum access policy based on Q-learning is proposed in [9]. The proposed policy is independently employed by each SU without any communication among the SUs. The other SUs are considered as part of the environment. In [8] multiagent machine learning policies based on collaborative filtering are proposed for learning the primary user (PU) occupancy probabilities and for maximizing the data rate of spectrum access. The SUs collaborate with neighboring users experiencing highly correlated channels by sharing their local estimates. In [13], a distributed multiagent learning based approach, in which the SUs optimize their joint actions through payoff propagation [7], is proposed for dynamic spectrum access. Many other spectrum sensing and access policies not necessarily based on learning have been proposed as well; see [4] for a relatively recent survey.

The main contribution of this paper is to propose a linear function approximation based approach for reducing the dimensionality of the spectrum sensing state-action space in a multiagent reinforcement learning scenario. The proposed approach allows computationally efficient learning also in networks with high numbers of SUs and different frequency bands. Moreover, the tradeoff between finding more available spectrum and sensing reliability is addressed in this paper in a convenient manner using a single design parameter. This parameter controls the local number of SUs sensing each band simultaneously. That is, the proposed sensing policy employs user collaboration to sense more spectrum simultaneously as well as to achieve diversity gain. Diversity gain allows mitigating propagation effects such as shadowing and fading. The collaboration among the SUs is achieved through local interactions. That is, the neighboring users share their local test statistics or local decisions with others in order to exploit the benefits of collaborative sensing.

The paper is organized as follows. In Section II the system and network model is described. Section III describes the proposed reinforcement learning algorithm employed by the individual SUs. In Section IV the proposed single-agent algorithm is put into the context of multiagent learning and SU cooperation. Simulation results are presented in Section V. Finally, concluding remarks are given in Section VI.

II. System model

In this work, we are considering a cognitive radio ad hoc network consisting of a group of SUs in different spatial locations. Fig. 1 illustrates an example network topology. It is

J. Lundén's and H. V. Poor's work was supported by the Qatar National Research Fund under grant NPRP 08-522-2-211. J. Lundén's work was supported also by the Finnish Cultural Foundation.

S. R. Kulkarni's work was supported in part by the NSF Science & Technology Center under grant CCF-0939370, the U.S. Army Research Office under grant W911 NF-07-1-0185, and by Deutsche Telekom AG under grant RES AGMT.

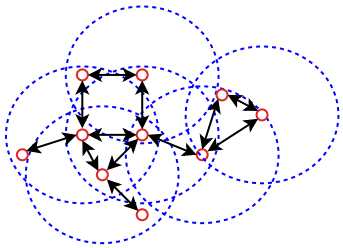


Figure 1. A wireless cognitive radio ad hoc network in which neighboring users collaborate with each other by sharing their local sensing information. The large dashed circles indicate the local interaction radii of the SUs.

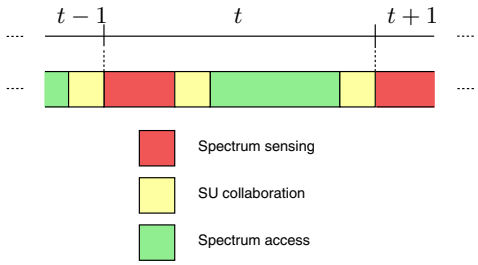


Figure 2. Time is divided into time slots. Each time slot is further divided into three different kind of minislots: spectrum sensing, SU collaboration, and spectrum access. There are two SU collaboration slots in each time slot. During the SU collaboration slots the SUs share their local sensing information with the neighboring users to enable collaborative sensing and SU action coordination.

assumed that each user communicates through local interaction with the users within its communication distance indicated by the large dashed circles in Fig 1. The users within the communication distance of a given SU are called the neighbors or the neighboring users of this particular SU. Let N_{SU} denote the number of SUs in the network and let $G_n \subset \{1, \ldots, N_{SU}\}$ denote the group of neighbors of SU n where n is the index of the SU in question. For simplicity, we assume a fixed network topology, i.e., N_{SU} and G_n do not vary as a function of time.

The SUs are assumed to be synchronized with sufficient accuracy within the cognitive radio network. This guarantees that the spectrum sensing and access tasks are performed simultaneously by all of the users. Time is divided into time slots comprised of minislots in which spectrum sensing, spectrum access, and SU information exchange may take place. Fig. 2 depicts the slotted nature of the SU network operation. There are two SU collaboration slots in each time slot. These slots are used for local information sharing between neighboring SUs. It is assumed that there is a control channel that can be used for the local communication among the SUs.

The spectrum of interest may be very wide and possibly non-contiguous. Consequently, it is assumed that the spectrum of interest is divided into N_{fb} frequency bands, indexed by $\{1, \ldots, N_{fb}\}$, and each SU is able to sense only one frequency band in each time slot. Hence, the spectrum occupancy may be only partially observed at each time instant. Thus, it is important to design a sensing policy that focuses on sensing the frequency bands that provide persistent spectrum opportunities more frequently and the ones with high PU occupancy less frequently. To this end a reinforcement multiagent learning based sensing policy is proposed in the following. We start by first describing the main reinforcement learning algorithm employed by the individual users. After that we will describe how the multiagent learning is coordinated within the network.

III. Reinforcement Learning with Function Approximation

The learning algorithm formulated in this section is employed by each SU to update the local action values. The goal of our learning algorithm is to maximize the expected sum of discounted future rewards for each SU n, $n = 1, \ldots, N_{SU}$,

$$R_t^n = E\left[\sum_{k=0}^{\infty} \gamma^k r_{n,t+k}\right] \quad (1)$$

where $r_{n,t}$ is the reward at SU n in time slot t and γ, $0 < \gamma < 1$, is the discount rate. In this context, the reward $r_{n,t}$ is defined as the number of frequency bands identified as being free by SU n in time slot t.

The reward depends on the state and actions taken in each state. Let $\boldsymbol{s}_t^n$ and $\boldsymbol{a}_t^n$ denote the state of the environment observed by the SU and the actions taken by SU n and its neighbors in G_n in time slot t, respectively. In this cognitive radio multiagent learning problem the state vector is formulated as $\boldsymbol{s}_t^n = [b_{n,1}, b_{n,2}, \ldots, b_{n,N_{fb}}]^T$ where $b_{n,i} \in [0, 1]$ is SU n's belief that the frequency band i is vacant. The values of $b_{n,i}$ are updated after each time instant regardless of whether the frequency band has been sensed or not. Algorithm 1 summarizes the update procedure. The value of $b_{n,i}$ is moved gradually towards 0.5 from the previous sensed state using a small step size as time passes. The value 0.5 represents the highest uncertainty in the frequency band occupancy.

The proposed sensing policy is based on the learned action values denoted by $Q_t^n(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n)$. Now, in order to reduce the dimensionality of the problem, we employ function approximation to approximate the action values. The action values $Q_t^n(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n)$ are approximated by a linear function as follows:

$$Q_t^n(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n) = (\boldsymbol{\theta}_t^n)^T \boldsymbol{f}(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n) = \sum_{i=1}^{N_{fb}} \theta_{t,i}^n f_i(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n), \quad (2)$$

where $\boldsymbol{\theta}_t^n$ is a parameter vector and $\boldsymbol{f}(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n)$ is a feature vector depending on the state and joint action by SU n and its neighbors in G_n. Thus, the learning problem is transformed to the problem of learning the parameter vector $\boldsymbol{\theta}_t^n$. The feature vector is constructed as follows. The number of features is equal to the number of frequency bands N_{fb}. That is, there is one feature corresponding to each frequency band. The feature value for each frequency band is given by

$$f_i(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n) = b_{n,i} \cdot h\left(I_{[a_t^n=i]} + \sum_{j \in G_n} I_{[a_t^j=i]}\right), \quad (3)$$

$$h(m) = \begin{cases} m, & m \leq N_D, \\ 2N_D - m + 1, & N_D + 1 \leq m \leq 2N_D \\ 0, & m > 2N_D \end{cases} \quad (4)$$

where $I_{[a_t=i]}$ is an indicator function having value 1 if action $a_t = i$ and 0 otherwise. Here, the h-function and the parameter

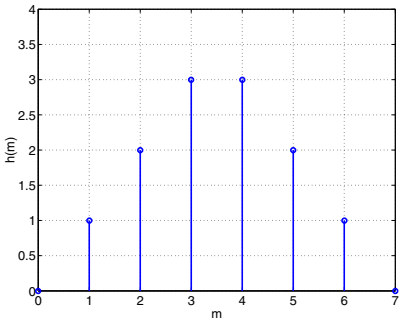


Figure 3. The h-function for $N_D = 3$. The h-function values are designed to favor collaborative sensing by N_D SUs.

Algorithm 1: Update of the SU's knowledge of the occupancy state of the frequency band i. $b_{n,i}$ may be initialized to 0.5 and δ is a small step size parameter (e.g., $\delta = 0.01$).

if *Frequency band i was sensed by SU n or its neighbors in G_n* **then**
  if *Frequency band i was found vacant* **then**
    $b_{n,i} = 1$;
  else
    $b_{n,i} = 0$;
  end
else if $b_{n,i} \geq 0.5$ **then**
  $b_{n,i} = \max(0.5, b_{n,i} - \delta)$
else
  $b_{n,i} = \min(0.5, b_{n,i} + \delta)$
end

N_D are employed to favor a desired number of SUs sensing the same band simultaneously. Fig. 3 illustrates the h-function for a value of $N_D = 3$. The goal of N_D and the h-function is to give incentive to the secondary network to employ diversity to improve the detection performance. However, it does not guarantee that a certain number of users are sensing each frequency band in a given time slot. This issue will be addressed later in the design of action selection methods.

The value inserted into the h-function is the estimated number of SUs sensing frequency band i in time slot t. The estimate is formed from the action a_t^n of SU n in time slot t and from the user's knowledge of what the neighboring users are sensing in time slot t. The knowledge of the frequency band sensed by the neighboring user $j \in G_n$ at time t is either the actual frequency band sensed in time slot t, i.e. a_t^j, or the frequency band sensed in the previous time instant $t-1$, i.e. a_{t-1}^j. The latter information, i.e. a_{t-1}^j, is employed only if a_t^j is not available.

The action values are updated using the Sarsa-algorithm [11]. Sarsa is an on-policy temporal-difference (TD) learning algorithm. The update of the parameter vector $\boldsymbol{\theta}^n$ of Sarsa with linear function approximation is given by [11]

$$\boldsymbol{\theta}_{t+1}^n = \boldsymbol{\theta}_t^n + \alpha_{n,t}\Big[r_{n,t} + \gamma Q_t^n(\boldsymbol{s}_{t+1}^n, \boldsymbol{a}_{t+1}^n) - Q_t^n(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n)\Big]\boldsymbol{f}(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n), \tag{5}$$

where $r_{n,t}$ is the reward in time slot t and $\alpha_{n,t}$ $(0 < \alpha_{n,t} \leq 1)$ is a step size parameter. This is a gradient-descent method where $\boldsymbol{f}(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n)$ is the gradient of the action value function $Q_t^n(\boldsymbol{s}_t^n, \boldsymbol{a}_t^n)$ with respect to $\boldsymbol{\theta}_t^n$.

The above summarizes the main components of the reinforcement learning method employed by the SUs. However, since we are considering a cooperative multiagent learning problem, the cooperation among the SUs has to be considered as well. Moreover, the algorithms or methods for selecting the actions given the learned action values have to be also considered. These issues will be addressed in the following section.

IV. Multiagent learning

The proposed multiagent learning algorithm in which each SU employs the single-user reinforcement algorithm introduced in the previous section is summarized in Algorithm 2.

Collaborative sensing involves sharing the local test statistics or decisions as well as the indices of the sensed frequency bands by each user with its neighbors. The information is shared in the network using two different types of data packets $D_{n,t}^1$ and $D_{n,t}^2$. The $D_{n,t}^1$ packets are transmitted during the first SU collaboration slot. Each packet includes the index of the sensed frequency band and the local test statistic or decision of the corresponding SU for that frequency band, i.e., $D_{n,t}^1 = \{a_t^n, \mathcal{T}_{n,t}\}$ where $\mathcal{T}_{n,t}$ denotes the local test statistic or decision of SU n at time t. The $D_{n,t}^2$ packets are transmitted during the second SU collaboration slot and each packet includes only the index of the frequency band to be sensed in the next time slot by the corresponding SU, i.e., $D_{n,t}^2 = \{a_{t+1}^n\}$.

During the SU collaboration slots the SUs are assumed to be transmitting their data packets in a sequential order that may be random and vary between different time slots. During the first SU collaboration slot the ordering does not affect the end result since the fused decisions are obtained only after all data packets have been received. However, during the second SU collaboration slot the SU transmission order affects the action selection. That is, the SUs that are later in the sequential order have the possibility to use the information obtained from the SUs preceding them in the sequential order to make more informed decisions about which frequency band to sense in the next time slot. Sharing this information allows action updates to be made after each time slot without jeopardizing the stability of the learning processes of the individual SUs.

The proposed learning algorithm employs the action selection method described in Algorithm 3. The proposed action selection algorithm has two goals. The first goal is to effectively achieve a desired number of SUs sensing each frequency band if that is possible given the number of SUs. The parameter N_D is employed in this context as well. The second goal of the action selection algorithm is to balance between exploitation and exploration. To this end we employ the ϵ-greedy action selection method [11]. The ϵ-greedy action selection is a simple, yet effective method that balances between exploitation and exploration by selecting the action that has the highest estimated state-action value, i.e. $a_t^* = \max_a Q_t(\boldsymbol{s}_t, \boldsymbol{a}_t)$, with probability $1 - \epsilon$, or a random

Algorithm 2: The proposed multiagent reinforcement learning based sensing policy is based on an on-policy Sarsa control algorithm with linear function approximation. Note that although the *for*-loop over the SUs indicates a sequential order from 1 to N_{SU} this is just a convenience of notation. The order may be random and vary between different time slots. The data packets are $D^1_{n,t} = \{a^n_t, \mathcal{T}_{n,t}\}$ and $D^2_{n,t} = \{a^n_{t+1}\}$ where $\mathcal{T}_{n,t}$ is the local test statistic or decision of SU n in time slot t. The action selection algorithm is described in Algorithm 3.

Initialize $t = 0$;
Initialize s^n_t, $n = 1, \ldots, N_{SU}$;
Initialize $\boldsymbol{\theta}^n_t$ arbitrarily, $n = 1, \ldots, N_{SU}$;
Choose action a^n_t randomly, $n = 1, \ldots, N_{SU}$;
repeat
  // Sensing slot
  Take action a^n_t, $n = 1, \ldots, N_{SU}$;
  // First SU collaboration slot
  for $n = 1$ ***to*** N_{SU} **do**
    Transmit $D^1_{n,t}$ to the neighboring SUs in G_n;
    for $j \in G_n$ **do**
      Receive $D^1_{n,t}$ and thus observe action a^n_t and $\mathcal{T}_{n,t}$;
    end
  end
  Combine the test statistics $\mathcal{T}_{j,t}, j \in G_n$, to find vacant frequency bands and observe reward r^n_t, $n = 1, \ldots, N_{SU}$;
  Obtain $\boldsymbol{s}^n_{t+1}$ using the Algorithm 1, $n = 1, \ldots, N_{SU}$;
  // Spectrum access slot
  Access the vacant frequency bands according to some access policy;
  // Second SU collaboration slot
  for $n = 1$ *to* N_{SU} **do**
    Choose action a^n_{t+1} using the action selection algorithm in Algorithm 3;
    Transmit $D^2_{n,t}$ to the neighboring SUs in G_n;
    for $j \in G_n$ **do**
      Receive $D^2_{n,t}$ and thus observe action a^n_{t+1};
    end
  end
  $\boldsymbol{\theta}^n_{t+1} = \boldsymbol{\theta}^n_t + \alpha_{n,t}[r_{n,t} + \gamma Q^n_t(\boldsymbol{s}^n_{t+1}, \boldsymbol{a}^n_{t+1}) - Q^n_t(\boldsymbol{s}^n_t, \boldsymbol{a}^n_t)]\boldsymbol{f}^n(\boldsymbol{s}^n_t, \boldsymbol{a}^n_t)$, $n = 1, \ldots, N_{SU}$;
  $t = t + 1$;
until *the network is terminated*;

Algorithm 3: The action selection algorithm.

input : Future actions of the neighbors preceding SU n in the sequential order, i.e., $a^j_{t+1}, j \in G_n, j < n$, and the previous actions of the SUs following SU n in the sequential order, i.e., $a^k_t, k \in G_n, k > n$
output: Action of SU n, i.e., a^n_{t+1}

Count the number of neighboring SUs $j \in G_n, j < n$, (i.e., only the SUs preceding SU n in the sequential order) sensing each frequency band. Let D_i, $i = 1, \ldots, N_{fb}$, denote the obtained values;
if $N_D == 1$ **then**
  Let $\mathcal{C} = \{i : D_i == 0\}$ denote the set of frequency bands not sensed by any SU $j \in G_n, j < n$;
else
  Let $\mathcal{C} = \{i : 1 \leq D_i < N_D\}$ denote the set of frequency bands sensed by at least 1 but less than N_D SUs $j \in G_n, j < n$;
  if $\mathcal{C} == \varnothing$ **then** $\mathcal{C} = \{i : D_i == 0\}$;
end
// If set $\mathcal{C}$ is empty then choose any frequency band
if $\mathcal{C} == \varnothing$ **then** $\mathcal{C} = \{1, \ldots, N_{fb}\}$;
Choose action a^n_{t+1} from the set $\mathcal{C}$ using a policy derived from $Q^n_t(\boldsymbol{s}^n_{t+1}, \boldsymbol{a}^n_{t+1})$ (e.g., ϵ-greedy);

action, uniformly, with probability ϵ regardless of the action value estimates [11].

V. Simulation Results

We consider a simulation scenario of 8 PUs all operating in the same location but on different frequency bands. In addition, there are 40 SUs spatially dispersed around the PUs. The SUs' locations have been randomly selected from a uniform distribution inside a square of size 0.2×0.2 centered at origin. All of the SUs are inside the protected radius of the PUs. Thus, they are not allowed to transmit if the PU is already transmitting. The number of neighboring SUs varies between 2 and 8 with a mean of approximately 5. Fig. 4 illustrates the scenario.

The SUs use energy detection with soft combining [6], [12] and a false alarm rate of $P_{fa} = 0.01$. The detection time is long enough to obtain a probability of detection equal to one at signal-to-noise ratio (SNR) of -20 dB for single-user detection in an additive white Gaussian noise (AWGN) channel. The AWGN model is used in the simulations. These selections were made in order to isolate the influence of the physical layer sensing algorithm from the evaluation of the sensing policy

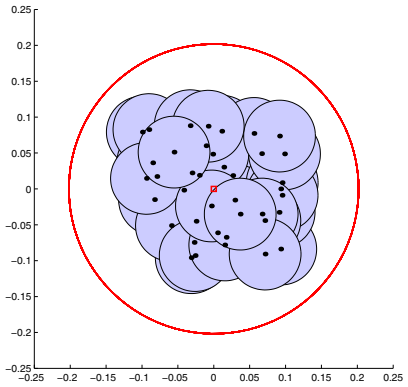


Figure 4. Multiagent spectrum sensing scenario. The small square at the origin denotes the location of the PUs and the large circle denotes the protected radius of the PUs. Small dots indicate the locations of the 40 SUs. Each shaded circle denotes the corresponding SU's (i.e., the one in the middle of the circle) communication range.

performance. This allows better comparison and analysis of the performance of the proposed sensing policies.

Fig. 5 depicts the occupancy of the PU systems. It is observed that there is a noticeable difference in the availability of different frequency bands at different times.

In the following, two different sensing policies are compared: the proposed reinforcement learning based sensing policy and a random sensing policy. The reinforcement learning based sensing policy employs ϵ-greedy action selection with $\epsilon = 0.1$. The other parameter values of the learning algorithm are $\alpha = 0.1$, $\gamma = 0.9$, and $\delta = 0.01$. The random policy employs the proposed action selection algorithm in Algorithm 3 with ϵ-greedy action selection with $\epsilon = 1$. Thus, the action selections are totally random except for the fact that the parameter N_D and what the neighboring SUs are sensing affect the set of available frequency bands.

Figs. 6 (a) and (b) illustrate the sensing performance for the two sensing policies with $N_D = 1$ and $N_D = 3$, respectively. For $N_D = 1$, the performance of an optimum genie-aided policy is shown as well. It is observed that the

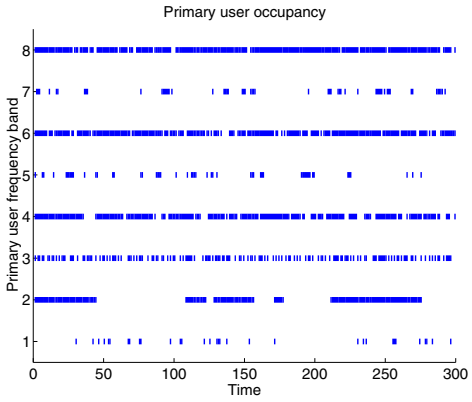
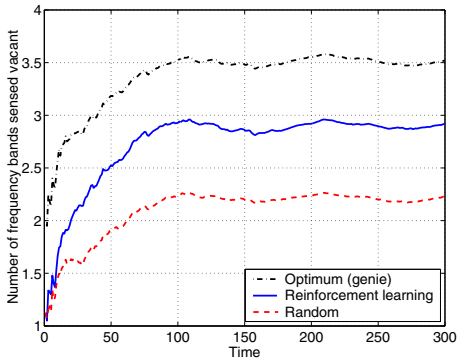


Figure 5. Primary user (PU) occupancy. Shading indicates that the corresponding frequency band is occupied by the PU. Especially frequency bands 1, 5, and 7 are unoccupied most of the time.

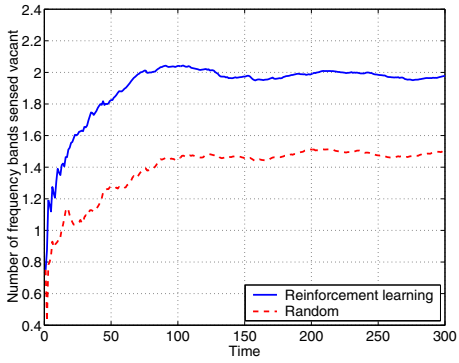
proposed reinforcement learning based sensing policy finds on average a significantly higher number of available frequency bands than the random sensing policy. The difference between the two policies is on average approximately 0.69 and 0.47 frequency bands for $N_D = 1$ and $N_D = 3$, respectively, which corresponds to an increase of roughly 31% in available bandwidth in both cases. Moreover, for $N_D = 1$ the number of available frequency bands found by the reinforcement learning based policy is on average approximately 83 % of the number of available frequency bands obtained by the optimum genie-aided policy. For $N_D = 3$, the average number of SUs sensing each band is on average 2.22 and 1.92 SUs for the reinforcement learning based and random sensing policies, respectively. Thus, the reinforcement learning based sensing policy provides better sensing reliability as well.

VI. Conclusion

In this paper, a distributed multiagent reinforcement learning based sensing policy for cognitive radio ad hoc networks has been proposed. The proposed sensing policy employs SU collaboration through local interactions to obtain a geographical map of spectrum occupancy. Moreover, it provides a simple way of controlling the tradeoff between sensing more available spectrum and sensing reliability. Simulation results show that the proposed approach provides an efficient way of finding available spectrum in a multiuser, multiband scenario.



(a) $N_D = 1$



(b) $N_D = 3$

Figure 6. Average cumulative number of frequency bands sensed vacant as a function of time for (a) $N_D = 1$ and (b) $N_D = 3$. Compared to a random policy the proposed reinforcement learning based policy improves the performance by allowing the SUs to find more available frequency bands. The optimum policy corresponds to a genie that knows at every time instant which frequency bands are available for secondary use.

References

[1] I. F. Akyildiz, W.-Y. Lee, and K. R. Chowdhury, "CRAHNs: Cognitive radio ad hoc networks," *Ad Hoc Networks*, vol. 7, no. 5, pp. 810–836, Jul. 2009.

[2] U. Berthold, F. Fu, M. van der Schaar, and F. K. Jondral, "Detection of spectral resources in cognitive radios using reinforcement learning," in *Proc. 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Chicago, IL, Oct. 14–17, 2008.

[3] L. Busoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews,* vol. 38, no. 2, pp. 156–172, Mar. 2008.

[4] C. Cormio and K. R. Chowdhury, "A survey on MAC protocols for cognitive radio networks," *Ad Hoc Networks*, vol. 7, no. 7, pp. 1315–1329, Sep. 2009.

[5] M. Di Felice, K. R. Chowdhury, W. Meleis, and L. Bononi, "To sense or to transmit: A learning-based spectrum management scheme for cognitive radio mesh networks," in *Proc. Fifth IEEE Workshop on Wireless Mesh Networks (WIMESH 2010)*, Boston, MA, Jun. 21, 2010.

[6] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *IEEE Transactions on Communications*, vol. 55, no. 1, pp. 21–24, Jan. 2007.

[7] J. R. Kok and N. Vlassis, "Collaborative multiagent reinforcement learning by payoff propagation," *Journal of Machine Learning Research*, vol. 7, pp. 1789–1828, Dec. 2006.

[8] H. Li, "Learning the spectrum via collaborative filtering in cognitive radio networks," in *Proc. IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Singapore, Apr. 6–9, 2010.

[9] H. Li, "Multiagent Q-learning for Aloha-like spectrum access in cognitive radio systems," *EURASIP Journal on Wireless Communications and Networking*, Volume 2010, Article ID 876216, 2010.

[10] J. Oksanen, J. Lundén, and V. Koivunen, "Reinforcement learning method for energy efficient cooperative multiband spectrum sensing," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, Kittilä, Finland, Aug. 29–Sep. 1, 2010.

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

[12] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, Apr. 1967.

[13] K.-L. A. Yau, P. Komisarczuk, and P. D. Teal, "Achieving efficient and optimal joint action in distributed cognitive radio networks using payoff propagation," in *Proc. IEEE International Conference on Communications*, Capetown, South Africa, May 23–27, 2010.