Performance of Cognitive Radio Reinforcement Spectrum Sharing Using Different Weighting Factors

Tao Jiang, David Grace, Yiming Liu
Communication Research Group, Department of Electronics,
University of York, York, YO10 5DD, United Kingdom
Email: {tj511|dg| yl127} @ohm.york.ac.uk

Abstract— **This paper introduces a distributed spectrum sharing scheme in the context of cognitive radio which enables efficient usage of spectrum. This is achieved by using the past experience based on reinforcement learning. It shows that reinforcement spectrum sharing provides a good solution and has the potential to significantly improve the system performance. Several learning strategies based on different sets of weighting factors are investigated. Comparisons of system performance using different learning strategies are given to illustrate the importance of weighting factors in the spectrum sharing process.**

Keywords- Cognitive Radio, Spectrum Sharing, Reinforcement Learning, Weighting Factor

I. INTRODUCTION

One of the main methods to improve spectrum usage in wireless communication is spectrum sharing. Conventional licensed spectrum allocation strategy by radio regulatory bodies can be overly restrictive, making a large part of radio spectrum underutilized. According to Federal Communications Commission (FCC), 15% to 85% assigned spectrum is utilized with large temporal and geographical variations [1], [2]. Since the physical spectrum resource is limited and the demand of radio spectrum for wireless communication is increasing dramatically, efficient utilization of radio spectrum has attracted significant attention. Cognitive radio (CR), has been proposed as a novel approach for dynamically using the shared radio frequencies, thereby enabling efficient utilization of the radio spectrum [3],[4],[5].

The definition of cognitive radio used in this paper is suggested in [6] as: 'a radio that is aware of and can sense its environment, learn from its environment and adjust its operation according to some objective function'. One vital element of cognitive radio is learning from interaction between environment and itself.

Reinforcement learning is a computational approach of learning used to maximize some notion of long-term reward. More specifically, the reinforcement learning technique uses a mathematical way to define the success level of the interaction between a learning agent and its environment [7], [8]. Its emphasis on individual learning from direct interaction with environment makes it perfectly suited to distributed spectrum sharing scenarios [5], [9]. In this paper, we implement the computational method by using a reward function and reward values. Based on the results of the reward function, the action policy of the agent is modified accordingly. In other words, agents adjust their operation according to the reward function feedback.

The purpose of this paper is to introduce our reinforcement learning based distributed spectrum sharing scheme which enables efficient usage of spectrum by exploiting users' past experience. In our spectrum sharing scheme, a reward weight is assigned to the used resource based on the result of the reward function. Cognitive radio users select spectrum resources to use based on the weights assigned to the spectral resources - resources with higher weights are considered higher priority. Furthermore we investigate and compare the system performance of different sets of reward values which effectively are the weighting factors in the reward function. In fact, we will show how different weighting factor values have significant impact on the system performance, and that inappropriate weighting factor setting may cause some specific problems. We will provide results and more details in the following sections.

This paper is organized as follows. In section II, we introduce our reinforcement spectrum sharing schemes. Then the reward function which is used in this paper is given in section III. Simulation results are discussed in section IV. Finally, conclusions are drawn in section V.

II. REINFORCEMENT LEARING BASED SPECTRUM SHARING SCHEMES

Spectrum sharing is one of the main challenges in cognitive radio: between different CR users or even between secondary CR users and primary licensed users. Reinforcement learning, an approach to learn from interaction, provides an ideal method to resolve the dynamic spectrum sharing problem [5]. By using reinforcement-based learning, CR users will assess the success level of a particular action. This in our scenario is whether the target spectrum is suitable for the considered communication request. According to the previous judgments, a reward is assigned in order to reinforce the weight of the physical resource. The concept of 'weight' in this paper is a number assigned to a resource, and the number reflects the importance of the resource to a certain CR user. The proposed reinforcement learning based channel assignment schemes are discussed in more detail in the following paragraphs.

The following is the basic rule: the CR users always choose the spectrum with the highest weight to communicate, and the weights of the resource for these users will be modified based

on the assessment of the degree of success. In other words, CR users are learning from the interaction between themselves and the environment. Initially, all CR users have equal access to the entire available spectrum pool. After each activation, the weight of the successfully used spectrum for a user is increased by a certain weighting factor. When the attempt fails, the weight is reduced.

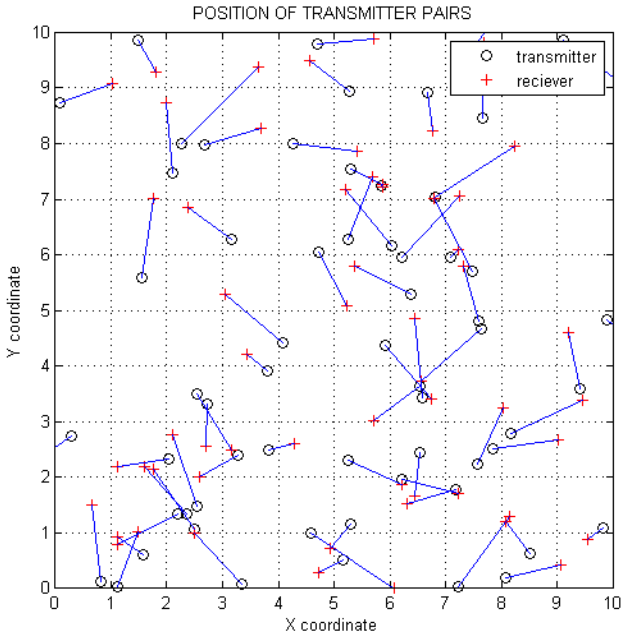

Fig.1. Sample of spatial layout of cognitive radio pairs for simulation

In this paper we consider the CR users are a set of transmitting-receiving pairs of nodes, denoted as U, uniformly distributed in a square area, and all the pairs $U_i \in U$ are spatially fixed. Fig.1 is an example layout of the nodes. When a pair U_i tries to set up a communication link from its transmitter Tx_i to the intended receiver Rx_i, it performs according to the following steps:

- *Step 1*: Spectrum selection. At the beginning of each activation, U_i chooses a channel to communicate according to the weights of the available resources. It starts with the spectrum with the highest weight or picks up a channel randomly if all resources have same priority. The selected channel is denoted as C_k where $C_k \in C$ and C is the available channel set.
- *Step 2*: Spectrum sensing. U_i senses the interference level on C_k. If the interference level I of C_k is below the interference threshold I_{thr}, U_i is activated. Otherwise if $I > I_{\text{thr}}$, the weight of C_k for U_i is decreased by a punishment weighting factor and U_i returns back to step 1.
- *Step 3*: *SINR* measuring. In this step, all the existing users within the channel C_k can measure the Signal-to-Interference-plus-Noise Ratio (*SINR*) at their receivers. The *SINR* at Rx_i can be expressed as:

$$SINR_i = \frac{S_i}{\sum_{j=1, j\neq i}^{n} I_{ji} + N} \quad (1)$$

where S_i is the receiving signal power at Rx_i , I_{ji} is the interference power from Tx_j to Rx_i , n is the number of existing users in C_k. N denotes the noise power. The purpose of measuring *SINR* is to maintain the communication quality of channels. We set up a *SINR* threshold $SINR_{thr}$. If the *SINR* of the activated pair U_i is greater than the threshold ($SINR_i > SINR_{thr}$), U_i successfully uses the spectrum and the weight of U_i for C_k will be increased by a weighting factor f. If $SINR_i < SINR_{thr}$, U_i is blocked by the channel and the weight is updated with a punishment weighting factor. In addition, according to the measurement of *SINR* of the existing users, the existing users whose *SINR* is decreased below the *SINR* threshold is dropped.

The CR users follow the above steps at each communication request. This is on one condition that $N(U_i) < N_{max}$, $N(U_i)$ denotes the number of sensed channels of U_i in each activation and N_{max} is the maximum number of channels which a CR user is allowed to scan in a single activation. If $N(U_i) > N_{max}$, and U_i is still searching for an unoccupied resource, it is blocked and waits for the next activation. It is unrealistic to allow users to keep sensing and searching for a better resource without a time limit, because sensing is a power-intensive and time-consuming process.

III. REWARD FUNCTION

Reinforcement learning is a computational approach to learn how to map situations to actions, and it is well suited to problems which include a long-term versus short-term reward trade-off [7]. One of the distinguishing features of reinforcement learning is the concept of reward value and reward function. The action policy of CR users is updated according to the reward function feedback, therefore the reward function of reinforcement learning is also the system objective function in our work.

The following linear reward equation is used as the reward function to determine the weights of the resource in this paper:

$$W_t = f_1 \cdot W_{t-1} + f_2 \quad (2)$$

where f_1 and f_2 are weighting factors that have different values depending on the localized judgment of current system states and the environment. W_{t-1} is the weight of a channel at time $t-1$, and W_t is the weight at time t according to previous weight W_{t-1} and the updated feedback from system. Furthermore, the weighting factor f is effectively the reward value in function (2). Based on the evaluation of the success level of CR users' action, either reward values or punishment values are assigned to f_1 and f_2 by the system. Choosing an appropriate value for f is the main issue of our work. In this paper, the values of f_1 and f_2 are assigned according to TABLE I.

IV. SIMULATION AND RESULTS

In this paper, we focus on the behavior of the nodes in our cognitive radio system. In order to achieve a deep understanding of such behavior, a basic transmitter-receiver pair system model and a free space propagation model[1] are used in our simulation. 1000 cognitive radio pairs are uniformly distributed on a square simulation area of $1000km^2$. An event-based scenario is employed in our work, at each event a random subset of pairs are activated. A number of 400 is assigned to define the maximum size of the subset. The link length is uniformly distributed between $200m$-$1500m$. 100 channels are available for communication. The following Friis transmission formula is used to determine the received signal power:

$$P_{rx} = P_{tx} \cdot G_{rx} \cdot G_{tx} \left(\frac{\lambda}{4\pi r} \right)^2 \qquad (3)$$

where P_{tx} is the transmitter power, fixed at $30dBm$, G_{tx} and G_{rx} are the gains of the transmit and receive antennas respectively both fixed at $0dBi$. A noise floor of $-137dBm$ is used, which corresponds to a noise bandwidth of $20kHz$ and a receiver noise temperature of $300K$.

An interference threshold of $-40dBm$ is used. The *SINR* threshold is set to $10dB$, and the maximum channel sensing number N_{max} of 3 is used which means the CR user is allowed to scan maximum 3% of available resources at the beginning of each communication in all schemes.

The values of weighting factors are shown in TABLE I. Based on the degree of success, either a reward or a punishment is assigned to the weight of the used spectrum.

TABLE I. WEIGHTING FACTOR VALUES

SCHEMES	f_1		f_2	
	Reward	**Punishment**	**Reward**	**Punishment**
Mild Punishment Scheme	1	1	1	-1
Harsh Punishment Scheme	1	0	1	0
Discounted Scheme	1	0.5	1	0

The reward value of 1 is used in all of the three schemes in TABLE I. The main difference between these schemes is the values assigned to punishment factors. In the first scheme, the absolute values of the reward value and the punishment value are equal. In other words the weight is increased or decreased by the same step size. This scheme is also named the '*mild punishment scheme*' in this paper. In the second scheme, if the attempt for communication fails, the weight is directly reduced to zero. Therefore we call it the '*harsh punishment scheme*'. Practically, the second scheme is a low complexity learning scheme where the CR users remember the last successful spectrum and keep using it at new activation until the request for that resource is declined. Then the user picks up a channel randomly and keeps using it as long as the quality of communication in that channel is above the requirement. Weights are reduced by a certain percentage in the third scheme, and a percentage of 50% is used to reduce the weight of an unsuccessful channel. We can refer to the scheme as the '*discounted scheme*'.

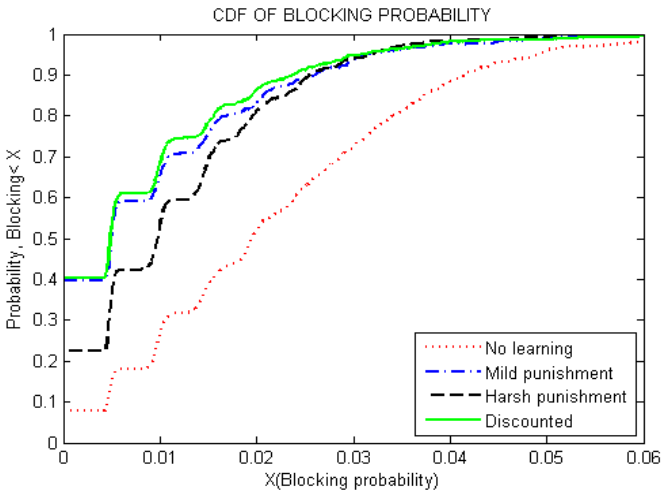

Fig.2. Cumulative distribution function of system blocking probability at discrete points over the service area

Fig.2. - Fig.5. illustrate the performance of schemes which we discussed above. Blocking probability is measured at regular points in the service area and a Cumulative Distribution Function (CDF) of system blocking probability at these points is derived. In order to analyse the level of system interruption, a CDF of dropping probability is calculated at the same time. All CR users' parameters are exactly the same for each scheme evaluation, with different system performance being caused only by different weighting factor values.

Fig.2. shows the CDF of system blocking probability of the three learning schemes along with a lower bound performance of random spectrum sharing without reinforcement learning. Comparing with the red dotted line which is the CDF of the no learning scheme, the blocking probability of our reinforcement learning spectrum sharing schemes are much lower than the scheme without learning. About 90% users blocking probability in the discounted scheme are below 0.02, but in the no learning scheme only 50% users are able to meet this requirement. By using a reinforcement learning way to share spectrum, the blocking probability can be significantly reduced. It can be seen that the discounted scheme has the best performance in Fig.2. The overall blocking probability of the discounted scheme is about 40% of that of the no learning

[1] However, the technique is widely applicable for other propagation scenarios.

scheme. The blocking probability of the mild punishment scheme is slightly higher than the discounted scheme. This is because of the setting of punishment value. We believe that the value of weighting factor reflects the degree of the reaction of a user to a specific action. The higher the value is, the higher the degree is. In the discounted scheme, the weight of an unsuccessful channel is reduced by a certain percentage at each time. According to function (2) if the request for a channel has been refused n times, the weight of that channel is:

$$W_t = f_1^n \cdot W_{t-n} \tag{4}$$

If a user in the mild punishment scheme is in the same situation, the weight of the unsuccessful channel will be:

$$W_t = W_{t-n} - n \tag{5}$$

Take n = 3, W_{t-n} = 100 for example, we assume that 100 is the highest weight of all available spectrum for a CR user. After the best channel has failed to communicate for three times, the weight of that channel W_t in the discounted scheme is 12.5, the channel probably no longer at the top of the priority list for the CR user. But in the mild punishment scheme the weight W_t is 97, it still high enough to maintain its position as a good channel for the user. Since the reaction of the discounted scheme towards a communication failure is stronger and quicker than that of the mild punishment scheme, the performance of the discounted scheme is better.

Nevertheless the punishment factor is not the higher the better. The black dashed line is the CDF of the harsh punishment scheme. In this scheme the weight of the unsuccessful spectrum is directly decreased to zero but the system blocking probability is still higher than the discounted scheme. This is because the 'over-reactive' behavior of the harsh punishment scheme. If a spectrum sharing scheme sets a punishment factor overly severe, the results of learning could be significantly changed by a rare occurrence. In the results of simulation, the best performance is achieved by the discounted scheme.

It can be seen that in every reinforcement learning scheme there are about 5% of users whose blocking probability is above 0.03. The performance of blocking probability of these users is difficult to improve no matter how the system defines the weighting factors, because these users are located at an extremely high user density area and the opportunity for these users to successfully set up a communication link is limited.

Some excellent learning algorithms for dynamic channel assignment (DCA) can be found in previous work. Nie and Comaniciou investigate a no regret learning algorithm in [10]. In order to achieve the best performance, the agents in such algorithm not only need to explore the space of actions by playing all possible actions, but also have to update the weights of all possible strategies at each activation. This is completely different in our scheme where the nodes do not necessarily examine all available spectrum. The CR users in our scheme always directly start with the spectrum which successfully used in the past. In other words, the nodes will never take a new action unless they have no prior experience of the available resources. Moreover, unlike the no regret learning, our learning scheme only updates the weight of the strategy currently performed. From this point of view, the complexity of our learning scheme is lower.

The convergence behaviour of our reinforcement learning scheme is also quite different from other DCA learning algorithms. The centralised Q-learning approach proposed in [11] and the no regret learning scheme we mentioned before all need a sufficiently long stage to converge to their optimal state. Even if Q-learning is guaranteed to converge with probability 1, the convergence theorem still requires all possible actions in every state to be performed repeatedly and infinitely. Here we have a decentralised approach in our scheme, and also it is not necessary for a CR user to go through multiple stages. Actually, it is possible for CR nodes to find their good channel at the first activation and keep using it. It is also possible that the system converges to a spectrum sharing equilibrium after all users have just activated once. Practically, in our system the ability CR users to perform limited spectrum sensing enables an efficient system convergence and ensures a better system performance because the users can find clean resources at the beginning of communication. It can be seen in Fig.2. that about 40% users in the mild punishment scheme and the discounted scheme are never blocked by the system. This result indicates that about 40% CR users in these schemes have found their ideal spectrum immediately after the first activation.

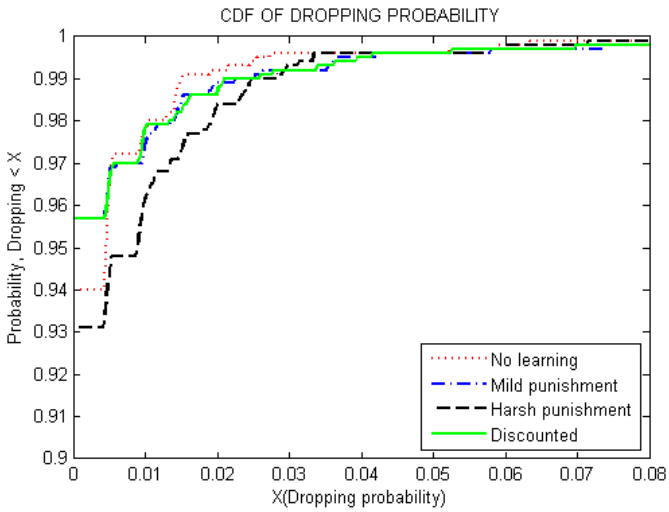

Fig.3. Cumulative distribution function of system dropping probability at discrete points over the service area

Fig.3. illustrates the CDF of dropping probability which demonstrates the level of system interruption. It shows that about 93% users are never dropped by system throughout the simulation. Since our schemes only take advantage of localized information to update the weights of spectrum, the performance of reinforcement learning-based scheme is no longer better than the no learning scheme. On the contrary, the dropping probability of the no learning scheme is lower than learning schemes. This is because a few CR users regard the channels with high dropping probability as their preferred resources and

keep using these channels as long as their blocking probability is low. Using the information of system dropping along with blocking to adjust weights may be a potential method to achieve a better system performance. Further work needs to be done to examine this argument.

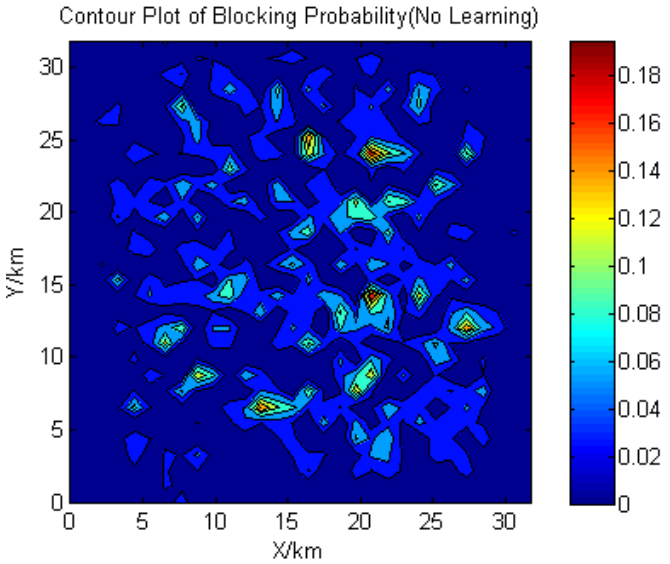

Fig.4. Contour plot of blocking probability of no learning scheme

Fig.4. and Fig.5. show the spatial plots of the no learning and discounted schemes respectively. Since the users in our scenario are spatially fixed, the blocking probability is strongly connected to the user density in a certain area. From Fig.4. and Fig.5. we can clearly see the improvement of system performance by applying the reinforcement learning. Not only the 'high blocking' area of no learning scheme is significantly reduced by the discounted scheme, the blocking probability of some 'red hotspots' are also decreased.

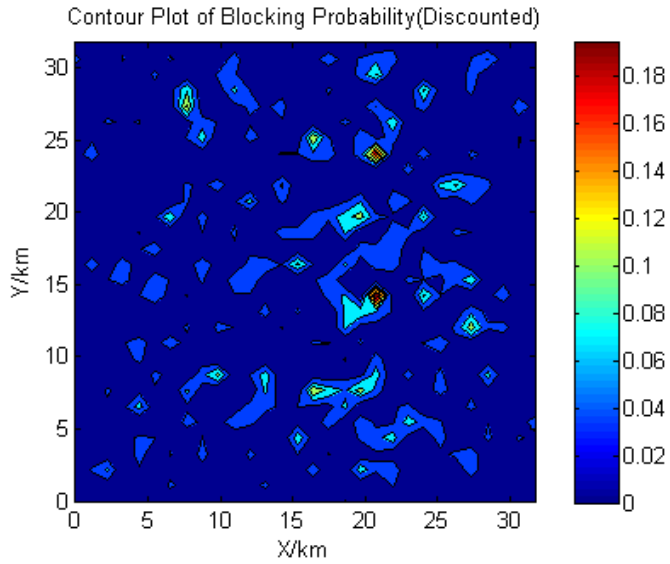

Fig.5. Contour plot of blocking probability of discounted punishment scheme

V. CONCLUSIONS

In this paper, we introduced a reinforcement learning spectrum sharing scheme for cognitive radio, which can enable efficient usage of the radio spectrum. Simulation results show that weighting factors have significant impact on the performance of the communication system. How to set the reward value is the key issue in the reinforcement learning scheme. The system achieves better performance only if the reward value is assigned appropriately. From the measurements of system blocking and dropping probability, the performance improvements of applying our reinforcement learning scheme can be clearly seen. About 90% of users have a blocking probability below 0.02 in the discounted scheme, compared with a situation of 50% with the no learning scheme. The overall blocking probability of the discounted scheme is 60% lower than that of the no learning scheme. In addition, we compare the system performance of different sets of reward values. About 90% users perform better in the discounted scheme than in the harsh punishment scheme. In our scenario, the scheme with a discounted punishment factor achieves the best performance. In addition, our spectrum sharing scheme can reduce the need for spectrum sensing which effectively save the power and time for sensing.

REFERENCES

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, pp. 2127-2159, 2006.

[2] FCC, "Notice of proposed rule making and order," ET Docket No 03-222, December 2003.

[3] J. Mitola, "Cognitive radio: Making software radios more personal," *IEEE Personal Communication*, vol. 6, pp. 13-18, Aug. 1999.

[4] S. Haykin, "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE Journal on selected areas in communications*, vol. 23, pp. 201-220, Feb. 2005

[5] B. Fette, *Cognitive Radio Technology*: Newnes, 2006.

[6] L. Dasilva and A. Mackenzie, "Cognitive Networks: Tutorial," in *CrownCom* Orlando, FL, July 2007.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning : an introduction*: The MIT Press, 1998.

[8] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *Journal of artificial intelligence Research*, vol. 4, pp. 237-285, May. 1996.

[9] M. Bublin, J. Pan, I. Kambourov, and P. Slanina, "Distributed spectrum sharing by reinforcement and game theory," presented at 5th Karlsruhe workshop on software radio, Karlsruhe, Germany, March. 2008.

[10] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," *Mobile Networks and Applications*, vol. 11, pp. 779-797, December, 2006.

[11] J. Nie and S. Haykin, "A Dynamic Channel Assignment Policy Through Q-Learning," *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 10, pp. 1443-1455, NOV. 1999.