I.J.Modern Education and Computer Science, **2013, 9, 11-18**
Published Online October 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2013.09.02

Collaborative Anti-jamming in Cognitive Radio Networks Using Minimax-Q Learning

Sangeeta Singh
Department of Electronics & Communication Engineering PDPM-IIITDM, Jabalpur, 482005, India
Email: sangeetasingh.1409@gmail.com

Aditya Trivedi
Department of Digital Communication ABV-IIITM, Gwalior, 474010, India
Email: atrivedi@iiitm.ac.in

Navneet Garg
Department of Electrical Engineering IIT Kanpur, 208016, India
Email: navneet.garg4@gmail.com

Abstract — Cognitive radio is an efficient technique for realization of dynamic spectrum access. Since in the cognitive radio network (CRN) environment, the secondary users (SUs) are susceptible to the random jammers, the security issue of the SU's channel access becomes crucial for the CRN framework. The rapidly varying spectrum dynamics of CRN along with the jammer's actions leads to challenging scenario. Stochastic zero-sum game and Markov decision process (MDP) are generally used to model the scenario concerned. To learn the channel dynamics and the jammer's strategy the SUs use reinforcement learning (RL) algorithms, like Minimax-Q learning. In this paper, we have proposed the multi-agent multi-band collaborative anti-jamming among the SUs to combat single jammer using the Minimax-Q learning algorithm. The SUs collaborate via sharing the policies or episodes. Here, we have shown that the sharing of the learned policies or episodes enhances the learning probability of SUs about the jammer's strategies but reward reduces as the cost of communication increases. Simulation results show improvement in learning probability of SU by using collaborative anti-jamming using Minimax-Q learning over single SU fighting the jammer scenario.

Index Terms **—** Cognitive radio networks, Stochastic game theory, Collaborative games, Markov decision process, Reinforcement learning

I. INTRODUCTION

Cognitive radio (CR) concept was proposed in [1] to resolve the problem of spectrum scarcity by exploiting the spectrum holes by the secondary users (SUs). The cognitive radio network (CRN) as proposed in [1] and [2] solves the conflicting situation between limited spectrum utilization and the increasing demand for spectrum resources. It exploits the spectrum holes by enabling the SUs to sense, select the free channel, collaborate with the other SUs, access the free channels and free the channels whenever the primary user (PU) needs those channels. Main research concerns till now were spectrum sensing, sharing and accessing procedures.

These works have assumed that SUs are greedy for spectrum holes and cooperate among themselves to fulfil their common objective. This assumption ignores the jammer's attack on SU scenario. In order to provide secure spectrum sharing in CRN, the random jammer's attack has to considered and modelled. Markov Decision Process (MDP) in CRN was introduced in [3] as it can easily model the competitive behaviour of SUs in the limited spectrum scenario of CRN. Stochastic games in CRN, is given in [4], [5] and [6], where a game was designed between the jammers and the SUs and zero-sum game condition also fetched the games' boundary conditions. The same framework is extended for MDP in [7] and [8], where competitive interaction among agents was considered in detail. In [8] and [9] MDP is used for the reinforcement learning (RL), this RL technique make the SUs learn the policies adopted by jammer. So, after learning the jammers' policy the SUs can predict jammer next action and plan their next course of action to combat the jammers. The reinforcement learning concept as introduced in (RL) [10], [11] and [12] has been used in the anti-jamming scenario, was introduced in [13] and [14]. Jammers attack in CRN can be modelled as zero-sum stochastic game framework. A zero-sum anti-jamming game is developed in [15] and the extension of QV learning is covered in [16] and [12]. One more advanced and online reinforcement algorithm is Minimax-Q learning is coined in [17] where there is an improvement in the learning probability of the SUs can be achieved as compared with the simple QV reinforcement learning algorithm. In the framework as developed in [15] quality of channel, availability of spectrum and the observation of attackers' actions define the state of game. The SU's actions, jammer's actions PU presence or absence, channels utilization gains and switching between jammed and un-jammed channels are modelled. This work has considered the SUs as independent agents learning independently without any collaboration with the other SUs. An improvement in the learning probability can be achieved by using the collaboration concept given in [18] and [19]. Here, collaboration is achieved via sharing the learned policies or episodes. In this paper, we propose the collaborative multi-agent multi-band anti-jamming game that involves the sharing of the local statistics, i.e., the number of jammed data or control channels or the number of unjammed data or control channels. To achieve the collaboration, this information is shared in the CRN with the neighbouring agents. The independent SUs will use the same decision policy by using Minimax-Q learning algorithm. In the proposed game each agent updates the Q-matrix for the same policy independently but now the rate of update gets multiplied by the number of collaborating SUs simultaneously. This sharing of the learned policies or episodes enhances the learning probability of SUs about the jammer’s strategies.

This paper is organized as follows. In section II, we have covered the system model along with the basic assumptions involved. In section III, we have given the Minimax-Q learning algorithm and collaborative multi-agent multi-band anti-jamming game that involves the sharing of the local statistics, i.e., the number of jammed data or control channels or the number of un-jammed data or control channels. In section IV, we have presented the simulation results and in section V conclusion is given.

II. SYSTEM MODEL

In this section, we give all the assumptions and notations for the given stochastic game model in brief. Further details of the game scenario are given in [15]. We assume that all the SUs are under the control of a single secondary base station and the jammer can only jam the SUs. Moreover, the jammer can jam at most N channels in each time slot due to limited number of antenna channels and transmit power. Here, the dynamics of channel, PU’s presence or absence, SUs actions and channel utilization gain has been modelled as in [15] and we have used the same developed system model. The basic analytical expressions involved are as follows [15]. The SUs’ motto is to get an optimal policy with maximum expected summation of discounted reward

$$max[E[\sum_{t=0}^{\infty}\gamma^t r(s^t,a^t,a_j^t)]] \qquad (1)$$

In the anti-jamming stochastic game the value of state V (s^t) is given by

$$V(s^t) = \underset{\pi(a^t)}{max}\ \underset{\pi_j(a_j^t)}{min}\sum_{a^t \in A} Q(s^t,a^t,a_j^t)\pi(a^t) \qquad (2)$$

where, $Q(s^t, a^t, a^t_j)$ stands for the Q-value of state and is updated by

$$\begin{aligned} Q(s^t,a^t,a_j^t) &= r(s^t,a^t,a_j^t) + \\ \gamma \times \sum_{s}^{t+1} & p(s^{t+1} \mid s^t,a^t,a_j^t) \times V(s^{t+1}) \end{aligned} \qquad (3)$$

To reduce the complexity, equation for updating the Q-function has been modified as

$$\begin{aligned} Q(s^t,a^t,a_j^t) &= (1-\alpha^t)\times Q(s^t,a^t,a_j^t) \\ &+(\alpha^t)\times[r(s^t,a^t,a_j^t)+\gamma\times V(s^{t+1})] \end{aligned} \qquad (4)$$

α^t stands for the learning rate decays for the time by $\alpha^{t+1} = \mu\alpha^t$ with $0 < \mu < 1$. The action set $a^t=\{a^t_1, a^t_2, ..., a^t_L\}$. The actions of the jammer are formulated as $a^t_J=\{a^t_{1,J}, a^t_{2,J}, ..., a^t_{L,J}\}$. The states of the anti-jamming game at time t is defined as $s^t=\{s^t_1, s^t_2, \ldots, s^t_L\}$ and $s_l^t=\{P_l^t, g_l^t, J_{l,C}^{t,} J_{l,D}^t\}$. The transition probability is expressed

$$p\left(s^{t+1} \mid s^t,\ a^t,\ a^t{}_J\right) = \prod_{l=1}^{L} p\left(s_l{}^{t+1} \mid s_l{}^t,\ a_l{}^t,\ a_l{}^t{}_J\right) \qquad (5)$$

The transition probability $p(s_l{}^{t+1} \mid s_l{}^t, a_l{}^t, a_l{}^t{}_J)$ can also be expressed as

$$\begin{aligned} p\left(s^{t+1} \mid s^t,\ a^t,\ a^t{}_J\right) &= p(J_{l,C}^{t+1}, J_{l,D}^{t+1} \mid J_{l,C}^t, J_{l,D}^t, a_l{}^t, a_{lJ}{}^t,)\times \\ p(P_l^{t+1}, g_l^{t+1} \mid P_l^t, g_l^t) \end{aligned} \qquad (6)$$

The cumulative average reward per iteration as given by the equation

$$\overline{r}t' = 1/t'\{\sum_{t=1}^{t'} r(s^t,a^t,a_j^t)\} \qquad (7)$$

$a^t{}_l = (a^t{}_{l,C1}, a^t{}_{l,D1}, a^t{}_{l,C2}, a^t{}_{l,C2})$ where action $a^t{}_{l,C1}$ (or $a^t{}_{l,\ D1}$) stands for the fact the secondary network will transmit control (or data) messages in $a^t{}_{l,C1}$ (or $a^t{}_{l,D1}$) channels by uniformly selecting from the earlier un-jammed channels, and action $a^t{}_{l,C2}$ (or $a^t{}_{l,D2}$) means that the secondary network will transmit control (or data) messages in $a^t{}_{l,C2}$ (or $a^t{}_{l,C2}$) channels uniformly selected from the previously jammed channels with($a^t{}_{l,J1}$) or ($a^t{}_{l,J2}$) means that the attackers will jam ($a^t{}_{l,J1}$)or ($a^t{}_{l,J2}$) channels uniformly selected from the previously un-attacked (or attacked) channels at current time t. Detailed mathematical formulation is given in [15].

III. MINIMAX-Q LEARNING ALGORITHM & COLLABORATIVE MULTI-AGENT MULTI-BAND ANTIJAMMING

Minimax-Q learning algorithm for the single independent SU combatting the jammer as given in [15].

1) STEP 1

At state s^t, $t = 0,1, \ldots$

– if state s^t has not been observed previously, add s^t to s_{hist},

– generate action set $A(s^t)$, and $A_J(s^t)$ of the attackers;
– initialize $Q(s^t, a^t, a_j^t) \leftarrow 1$, for all a ε A(s^t), a_J ε A_J (s^t)
– initialize $V(s^t) \leftarrow 1$;
– initialize $\pi(s^t, a^t) \leftarrow 1/|A(s^t)|$, for all a ε $A(s^t)$;
otherwise, use previously generated $A(s^t)$, $A_J(s^t)$, $Q(s^t, a^t, a_j^t)$, $V(s^t)$, and $\pi(s^t)$;

2) STEP 2 Choose an action at a^t time *t*:

– with probability p_{exp}, return an action uniformly at random;
– otherwise, return action a^t with probability $\pi(s^t, a)$ under current state s^t.

3) STEP 3 Learn:

Assume the attackers take action $a^t{}_J$, after receiving reward $r(s^t, a^t, a_j^t)$ for moving from state s^t to s^{t+1} by taking action a^t
– Update Q-function $Q(s^t, a^t, a_j^t)$ according to (3:9);
– Update the optimal strategy $\pi^*(s^t, a)$ by
$\pi^*(s^t) \leftarrow \arg\max_{\pi(s^t)} \min_{\pi_j(s^t)} \sum_a \pi(s^t, a) Q(s^t, a^t, a_j^t)$
– Update $V(s^t) \leftarrow \min_{s^t} \sum_a \pi^*(s^t, a) Q(s^t, a^t, a_j^t)$;
–Update $\alpha^{t+1} \leftarrow \alpha^t * \mu$;
– Go to step 1 until converge.
where, $\pi(s^t)$denotes state policy, $r(s^t, a^t, a_j^t)$ is reward of the game. $A(s^t)$ and $A_J(s^t)$ denote SU's and jammer's actions set.

A. Collaborative Anti-jamming Game

This framework has been used with the collaborative learning where SUs communicate with each other to combat the jammer. The collaboration can be achieved by the three approaches via sharing the test statistics or sensation, via sharing the iterative episodes or via sharing the learned optimal policy. The additional statistics shared by the agents are useful when used efficiently to speed up the learning, sharing of the learned optimal policy can be judicial, but this improvement is at the cost of the communication. Here, the collaboration has been achieved by sharing the learned optimal policy. Although these joint tasks slow the learning process initially, but it outperforms the independent agents. The proposed multi-agent multi-band collaborative anti-jamming game in which each SU uses the single user Minimax-Q reinforcement learning algorithm and same decision policy. Although each SU updates same policy independently, the rate of updating the policy is multiplied by the number of SUs collaborating. It involves the sharing of the local statistics, i.e., the number of jammed data or control channels or the number of un-jammed data or control channels. This information is shared in the CRN with the neighbouring agents for the collaboration. Each agent updates the Q-matrix for the same policy independently but the rate of updating the Q-matrix gets multiplied by the number of collaborating agents simultaneously. Agents performing same task can differ as the exploration of the state space differs. In this way they complement each other, i.e., policy learnt by one can be beneficial for other. It is an independent decision process. This collaborative game between the jammer and the SUs are as depicted. In Fig. 1 Single SU game without collaboration is clearly illustrated. In Fig. 2 two SUs game with collaboration between two agents is shown. Finally, in Fig. 3 three SUs game with collaboration between three agents.

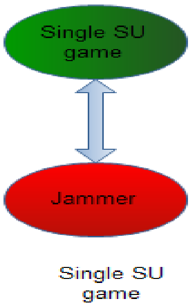

Fig. 1 Single SU game without collaboration

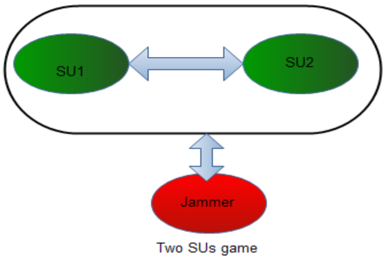

Fig. 2 Two SUs game with collaboration between two agents

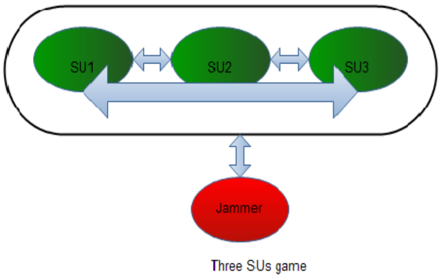

Fig. 3 Three SUs game with collaboration between three agents

IV. SIMULATION RESULTS

Now, we give the simulation results to evaluate the performance of the proposed collaborative anti-jamming strategy of the SU.

A. Anti-jamming for single licensed band

Here, one licensed band is available to the SU, i.e, L = 1 and the other simulation parameters are taken from [15]. Fig. 4 depicts the learning probability of the jammer for no collaboration condition in the state (0, 1, 0, 2). The learning probability is about 0.35 and the number of iterations required to learn are about 430. The different coloured curves show the different strategies of the jammer. Here, the jammer has eight different strategies as listed in the Fig. 4. Out of these eight strategies (2, 0) is having the highest learning probability.

In Fig. 5 the learning probability of single SU is shown for the state (0, 1, 0, 2) for no collaboration scenario. The learning probability of SU is 0.5 and the number of iterations required to learn the jammer's policy are 430. The different coloured curves show the different strategies of the SU. Here, the SU has 52 different strategies so cannot be listed in the Fig. 5. Strategy (0, 3, 2, 0) is having the highest learning probability. Fig. 6 depicts the cumulative average reward curve of SU for the state (0, 1, 0, 2) for no collaboration scenario. This reward is highest of all three scenario considered because reward decreases as the cost of communication required for the collaborative anti-jamming game increases. Fig. 7 shows the learning probability of the jammer, where two SUs are collaborating for the state (0, 1, 0, 2). The learning probability is about 0.2 and the number of iterations required to learn are about 350.

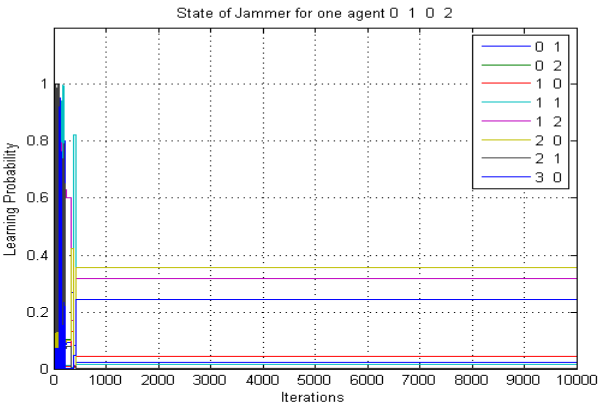

Fig. 4 Learning probability curve for jammer of the state (0,1,0,2), single SU, no collaboration, L=1

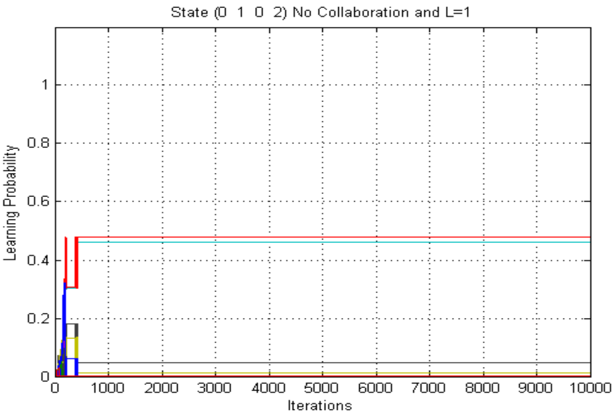

Fig. 5 Learning probability curve for SU of the state (0,1,0,2), single SU, no collaboration, L=1

Out of eight strategies of the jammer (1, 2) is having the highest learning probability. In Fig. 8 the learning probability of the SUs is shown for the state (0, 1, 0, 2), where two SUs are collaborating. The learning probability is 0.96 and the number of iterations required to learn the jammer's policy are 340. Out of the 52 strategies of SU (0, 1, 2, 0) is having the highest learning probability. Fig. 9 depicts the cumulative average reward curve for SU for the state (0, 1, 0, 2) where the collaboration between two SUs is employed. The reward is less than the no collaboration case and more than the three SUs collaborating scenario considered. Fig. 10 depicts the learning probability of the jammer, where three SUs are collaborating for the state (0, 1, 0, 2). The learning probability is about 0.33 and the number of iterations required to learn are about 430. Out of the eight strategies of SU (1, 1) is having the highest learning probability. In Fig. 11 the learning probability of the SUs is shown for the state (0, 1, 0, 2).

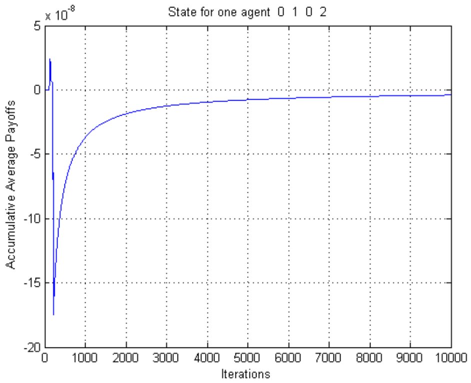

Fig. 6 Cumulative average reward curve for SU for the state (0,1 ,0, 2), single SU, no collaboration, L=1

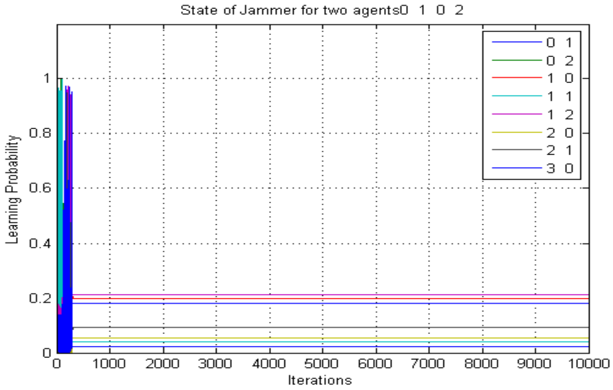

Fig. 7 Learning probability curve for jammer of the state (0, 1, 0, 2), two SUs collaborating, L=1

The learning probability is 0.99 and the number of iterations required to learn the jammer's policy are 200. Out of the 52 strategies of the SU (0, 4, 2, 0) is having the highest learning probability. Fig. 12 depicts the cumulative average reward curve for SU for the state (0, 1, 0, 2) where three SUs are collaborating. This reward is the smallest of the three scenario considered. So, more the number of agents collaborating more the cost of communication and lesser is the cumulative average reward.

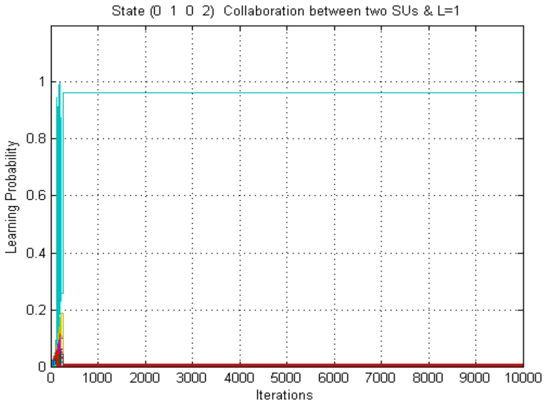

Fig. 8 Learning probability curve for SU of the state (0, 1, 0, 2), two SUs collaborating, L=1

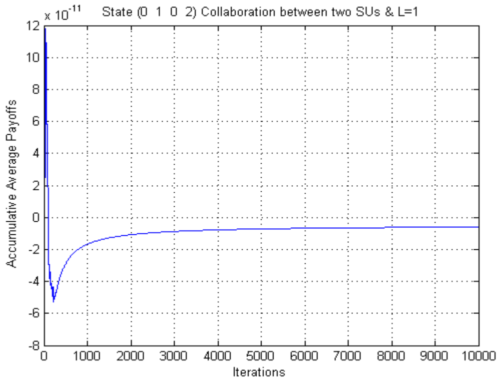

Fig. 9 Cumulative average reward curve for SU for the state (0, 1, 0, 2), two SUs collaborating, L=1

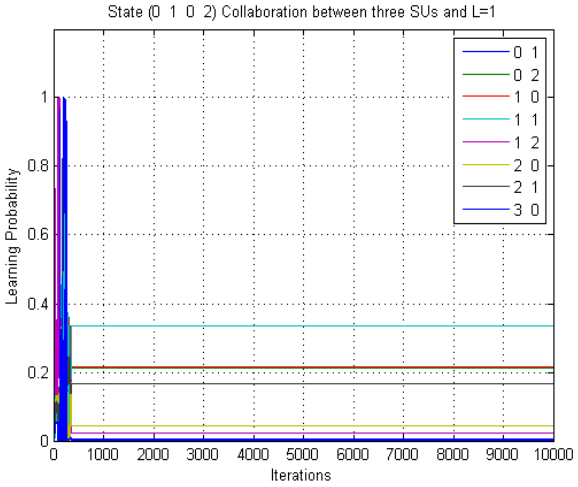

Fig. 10 Learning probability curve for jammer of the state (0, 1, 0, 2), three SUs collaborating, L=1

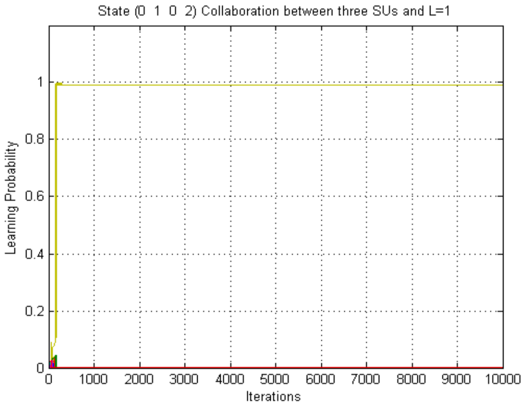

Fig. 11 Learning probability curve for SU of the state (0, 1, 0, 2), three SUs collaborating, L=1

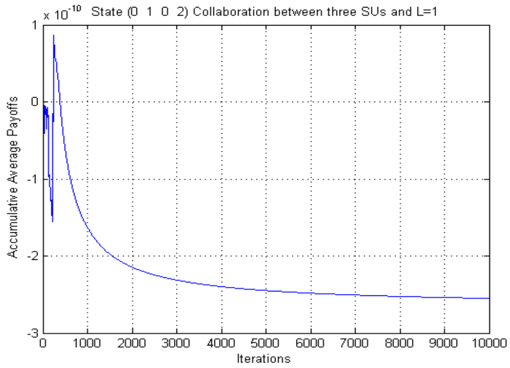

Fig. 12 Cumulative average reward curve for SU for the state (0, 1, 0, 2), three SUs collaborating, L=1

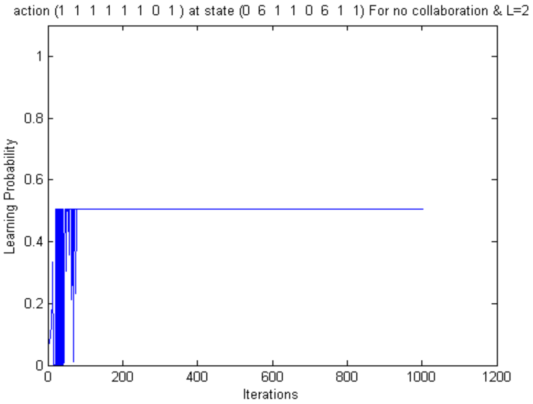

Fig. 13 Learning probability curve for SU of the state (0, 6, 1, 1 ,0 ,6 ,1, 1), single SU, no collaboration, L=2

B. Anti-jamming for two licensed band

Here, L = 2 and other parameters are taken from [15]. Fig. 13 the learning probability of single SU is shown for the state (0, 6, 1, 1, 0, 6, 1, 1) for no collaboration case. The learning probability is 0.5 and the number of iterations required to learn the jammer's policy are 100. Fig. 14 and Fig. 16 shows the cumulative average reward curve of SU for the, state (0, 6, 1, 1, 0, 6, 1, 1), for no collaboration and three SUs collaborating scenario respectively. The reward for no collaboration is more as compared to the collaboration case as the cost of communication required for the collaborative anti-jamming game is more. In Fig. 15 the learning probability of the SUs is shown for the state (0, 6, 1, 1, 0, 6, 1, 1), where three SUs, are collaborating. The learning probability of the SUs is 0.97 and the number of iterations required to learn the jammer's strategies are about 80.

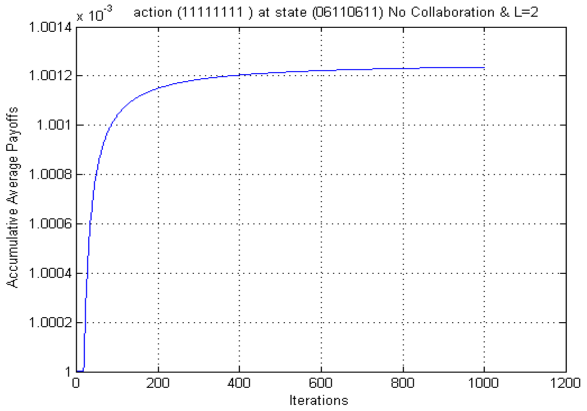

Fig. 14 Cumulative average reward curve for SU for the state(0, 6, 1, 1, 0, 6, 1, 1), single SU, no collaboration, L=2.

Fig. 17 depicts the consolidated view of multi-agent scenario for the multi-band collaborative anti-jamming game so that a clear comparison can be done and results can be analysed.

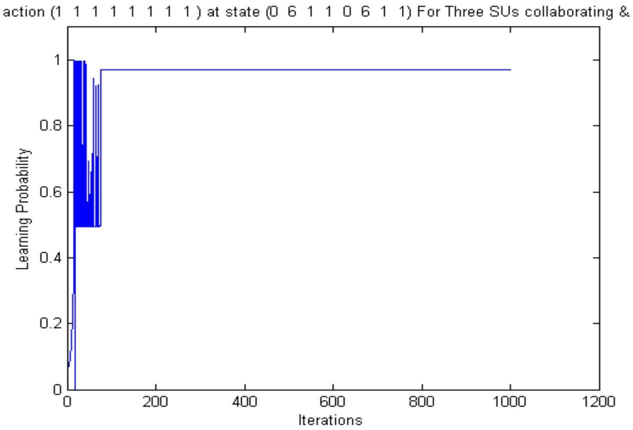

Fig. 15 Learning probability curve for SU of the state (0 ,6, 1, 1, 0, 6, 1, 1), three SUs collaborating, L=2

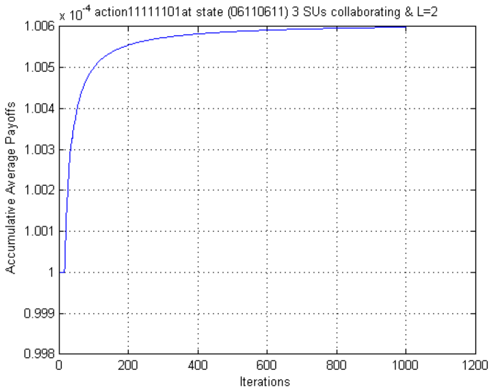

Fig. 16 Cumulative average reward curve for SU for the state (0, 6, 1, 1, 0, 6, 1, 1), three SUs collaborating, L=2

V. CONCLUSION

In this paper, we have considered the random jammer's attack on secondary users (SUs) in cognitive radio network (CRN). We have proposed the multi-agent multi-band collaborative anti-jamming using reinforcement learning, where SUs collaborate with each other to combat single random jammer. Minimax-Q learning is used by the SUs independently to make individual decision then they collaborate to learn about the single jammer's strategies. This paper demonstrates that in the multi-agent multi-band collaborative reinforcement learning agents (SUs) can learn faster about the jammer's strategies and converge sooner than independent agents via sharing the learned policies. But this improvement in the learning probability is at the cost of increased communication.

The proposed collaborative game framework can be extended to model various anti-jamming mechanisms in other layers of a CRN, as it can model the dynamics because of the environment and the cognitive attackers as well. This collaborative approach can be advantageous for the other layers defence mechanism as well.

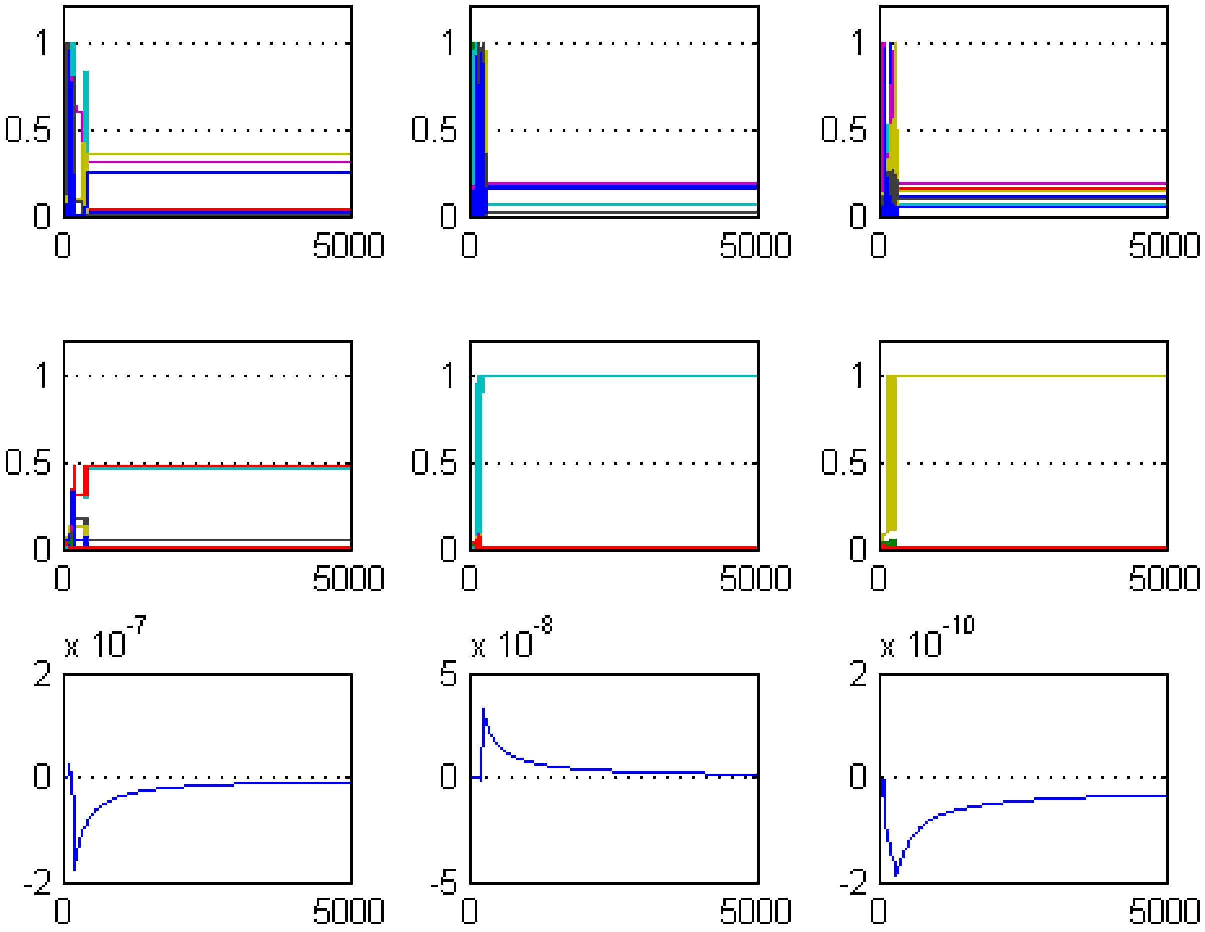

Fig. 17 Consolidated view of multi-agent scenario, L=1

REFERENCES

[1] S. Haykin, “Cognitive radio: brain-empowered wireless communications,” *IEEE Journal on, Selected Areas in Communications, vol. 23, no. 2, pp. 201–220, 2005.*

[2] I. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, “Next generation/ dynamic spectrum access/cognitive radio wireless networks: a survey,” Computer Networks, vol. 50, no. 13, pp. 2127–2159, 2006.

[3] M. Littman and C. Szepesv´ari, “A generalized reinforcement-learning model: Convergence and applications,” in MACHINE LEARNINGINTERNATIONAL WORKSHOP THEN CONFERENCE. Citeseer, pp. 310–318, 1996.

[4] J. Mertens and A. Neyman, “Stochastic games,” International Journal of Game Theory, vol. 10, no. 2, pp. 53–66, 1981.

[5] A. Neyman and S. Sorin, Stochastic games and applications. Springer Netherlands, vol. 570, 2003.

[6] G. Rummery and M. Niranjan, On-line Q-learning using connectionist systems. Univ. of Cambridge, Department of Engineering, 1994.

[7] M. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in Proceedings of the eleventh international conference on machine learning. Citeseer vol.157163, 1994.

[8] J. Filar and K. Vrieze, Competitive Markov decision processes. Springer Verlag, 1997.

[9] M. Wiering, “QV (lambda)-learning: A new on-policy reinforcement learning algorithm,” In D. Leone, editor, Proceedings of the 7th European Workshop on Reinforcement Learning, pages 2930, 2005.

[10] C. Claus and C. Boutilier, “The dynamics of reinforcement learning in cooperative multiagent systems,” in Proceedings of the National Conference on Artificial Intelligence. JOHN WILEY & SONS LTD, pp. 746–752, 1998.

[11] R. Sutton and A. Barto, Introduction to reinforcement learning. MIT Press, 1998.

[12] L. Matignon, G. Laurent, and N. Le Fort-Piat,“Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems,” The Knowledge Engineering Review, vol. 27, no. 01, pp. 1–31, 2012.

[13] K. Liu and B.Wang, Cognitive Radio Networking and Security: A Game theoretic View. Cambridge Univ Pr, 2010.

[14] B. Wang, Y. Wu, and K. Liu, “Game theory for cognitive radio networks: An overview,” Computer Networks, vol. 54, no. 14, pp. 2537–2561, 2010.

[15] B. Wang, Y. Wu, K. Liu, and T. Clancy, “An anti-jamming stochastic game for cognitive radio networks,” IEEE Journal on, Selected Areas in Communications, vol. 29, no. 4, pp. 877–889, 2011.

[16] M. Wiering and H. van Hasselt, “The QV family compared to other reinforcement learning algorithms,” in IEEE Symposium on, Adaptive Dynamic Programming and Reinforcement Learning, ADPRL, pp. 101– 108, 2009.

[17] M. Wiering and H. Van Hasselt, “Two novel on-policy reinforcement learning algorithms based on TD (λ)-methods,” in IEEE International Symposium on, Approximate Dynamic Programming and Reinforcement Learning, ADPRL., pp. 280–287, 2007.

[18] M. Tan, “Multi-agent reinforcement learning: Independent v/s. cooperative agents,” in Proceedings of the tenth international conference on machine learning, vol. 337. Amherst, MA, 1993.

[19] M. Veloso, “An analysis of stochastic game theory for multiagent reinforcement learning.” ICML, 2000.

Sangeeta Singh: Post-graduate student for doctor degree in the department of Electronics & Communication at PDPM-Indian Institute of Information Technology Design and Manufacturing (IIITDM) Jabalpur. She has received her M. Tech degree from ABV- Indian Institute of Information Technology & Management (IIITM) Gwalior in Digital communication.

Aditya Trivedi: Professor in the department of the Information and Communication Technology (ICT) at ABV -Indian Institute of Information Technology and Management, (IIITM) Gwalior, India. He obtained his doctorate (Ph.D) from IIT Roorkee in the area of Wireless Communication Engineering. He has published around 80 papers in various national and international journals/conferences. He is a fellow member of the Institution of Electronics and Telecommunication Engineers (IETE).

Navneet Garg: Post-graduate student for doctor degree in the department of Electrical Engineering of Indian Institute of Technology (IIT) Kanpur. He has received his M. Tech degree from ABV- Indian Institute of Information Technology & Management (IIITM) Gwalior in Digital communication.