A Collaborative Multi-Agent Reinforcement Learning Anti-Jamming Algorithm in Wireless Networks

Fuqiang Yao and Luliang Jia

***Abstract*—In this letter, we investigate the anti-jamming defense problem in multi-user scenarios, where the coordination among users is taken into consideration. The Markov game framework is employed to model and analyze the anti-jamming defense problem, and a collaborative multi-agent anti-jamming algorithm (CMAA) is proposed to obtain the optimal anti-jamming strategy. In sweep jamming scenarios, on the one hand, the proposed CMAA can tackle the external malicious jamming. On the other hand, it can effectively cope with the mutual interference among users. Moreover, we consider the impact of sensing errors due to miss detection and false alarm. Simulation results show that the proposed CMAA is superior to both sensing-based method and independent *Q*-learning method, and has the highest normalized rate.**

***Index Terms*—Anti-jamming, multi-agent reinforcement learning, Q-learning, Markov game.**

I. Introduction

JAMMING attack is a serious threat in wireless networks, and various anti-jamming methods have been developed in recent years [1]–[5]. Due to factors of the jammers' activities, the quality of channels varies between "good" and "poor" dynamically. The Markov decision process (MDP) [6] is a suitable paradigm to model and analyze the anti-jamming defense problem. Unfortunately, it is difficult to obtain the state transition probability function in an adversarial environment. In these scenarios, reinforcement learning (RL) techniques are available, such as the classic Q-learning method [7]. Based on the Q-learning method, the anti-jamming decision-making problem in single-user scenarios were investigated in [8]–[10]. Then, Aref *et al.* [11]–[13] extended it to the multi-user scenarios, and they resorted to the Markov game framework [14], which is the extension of the Markov decision process and can characterize the relationship among multiple users. Moreover, the corresponding multi-user reinforcement learning anti-jamming algorithm was designed. However, each user employed a standard Q-learning method in [11]–[13], and the coordination among users was not considered. Based on our previous work [15], the cooperative anti-jamming problem was investigated in UAV communications networks in [16], and the mobility of UAV was exploited. It is noted that miss detection and false alarm were not considered in [15] and [16], and they are common in practical wireless networks.

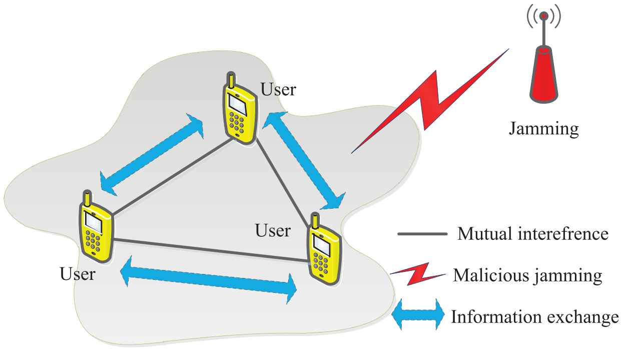

Fig. 1. System model.

In order to achieve better anti-jamming performance, the coordination among users is necessary. Through collaborative learning, on the one hand, it can tackle the external malicious jamming, and on the other hand, it can effectively cope with the mutual interference caused by competition among users. In this letter, a collaborative anti-jamming framework is formulated, in which the "coordination" and "competition" are simultaneously considered. To model and analyze the anti-jamming defense problem, the Markov game framework is adopted, and a collaborative multi-agent reinforcement learning anti-jamming algorithm is proposed. The main contributions of this letter are given as follows:

- Based on the Markov game, the anti-jamming defense problem is investigated in multi-user scenarios, and the coordination among users is considered.
- We develop a collaborative multi-agent reinforcement learning anti-jamming algorithm to obtain the optimal anti-jamming strategy.

II. System Model and Problem Formulation

A. System Model

As illustrated in Fig. 1, there are N users and one jammer in the considered model. The users have intelligent capabilities that include spectrum sensing, learning and decision-making, and thus, they are referred to as cognitive users in this letter. The cognitive user set is denoted as $\mathcal{N} = \{1, \ldots, N\}$, and the available channel set is defined as $\mathcal{M} = \{1, \ldots, M\}$. The number of available channels is M $(N < M)$. The coordination among cognitive users can be achieved through information exchange. It is noted that the jamming pattern is the sweep jamming, and one channel is jammed at each time slot. The

Manuscript received December 18, 2018; accepted March 3, 2019. Date of publication March 11, 2019; date of current version August 21, 2019. The associate editor coordinating the review of this paper and approving it for publication was P. Li. *(Corresponding author: Luliang Jia.)*

F. Yao is with the Sixty-Third/63rd Research Institute, National University of Defense Technology, Nanjing 210007, China (e-mail: yfq2030@163.com).

L. Jia is with the School of Space Information, Space Engineering University, Beijing 101416, China (e-mail: jiallts@163.com).

Digital Object Identifier 10.1109/LWC.2019.2904486

jamming channel set is represented as $\mathcal{C} = \{1, \ldots, C\}$. In this letter, we assume that the available channel set is the same as the jamming channel set. If two or more users select the same channel, it will lead to the mutual interference. In order to realize the reliable transmission, it is necessary to simultaneously consider the external malicious jamming and mutual interference due to competition among users. In this letter, the mutual interference refers to the co-channel interference among users, and the strategy is the selection of available channels.

B. Problem Formulation

The anti-jamming defense problem can be formulated as a Markov game [14], which is the extension of the Markov decision process in multi-user scenarios. Mathematically, it can be expressed as $\mathcal{G} = \{\mathcal{S}, \mathcal{A}_1, \ldots, \mathcal{A}_N, f, r_1, \ldots, r_N\}$, where $\mathcal{S}$ denotes the set of states, $\mathcal{A}_n, n = 1, \ldots, N$ is the set of the strategies, f represents the state transition model, and $r_n, n = 1, \ldots, N$ is the reward. In this letter, we consider sensing errors due to miss detection and false alarm, and the miss detection probability and false alarm probability of channel m are denoted as $P_{m,md}$ and $P_{m,fa}$, respectively. Referring to [8] and [9], the state can be defined as $s = \{\mathbf{a}, \tilde{f}_{jx}\}$, where $\tilde{f}_{jx}$ is the observed jamming channel, $\mathbf{a} = (a_1, a_2, \ldots, a_N)$ represents a joint action profile, and the set of the joint action profiles is $\mathcal{A} = \otimes \mathcal{A}_n, n = 1, \ldots, N$, where $\otimes$ represents the Cartesian product. Similar to [18], the global reward can be defined as:

$$R = \sum\nolimits_{n=1}^{N} r_n(s, \mathbf{a}), \tag{1}$$

where $s \in \mathcal{S}$ denotes a state. It is assumed that the jamming channel is denoted as f_{jx}, the selected channel of cognitive user n is represented as $f_{n,x}$, and the reward of cognitive user n at time slot t can be expressed as:

$$r_n(s, \mathbf{a}, t) = \begin{cases} 1, & \text{if } f_{n,x} \neq f_{jx} \,\&\, f_{n,x} \neq f_{m,x} (m \in \mathcal{N}/n), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

III. Collaborative Multi-Agent Anti-Jamming Algorithm

In this letter, we consider the two characteristics "coordination" and "competition" among users simultaneously. In wireless network, the coordination has various meanings, such as relay and information exchange. Here, the coordination is realized by information exchange through a common control channel [17]. Based on the coordination among users, the method of "decision-feedback-adjustment" is applied to obtain the optimal anti-jamming strategy.

To solve the formulated anti-jamming Markov game, a multi-agent Q-learning algorithm is proposed. Similar to [18], cognitive user n updates its Q values according to the following rules:

$$\begin{cases} Q_n(s, \mathbf{a}) = (1 - \lambda) Q_n(s, \mathbf{a}) + \lambda \left[r_n + \gamma V_n(s') \right], & (3) \\ V_n(s') = Q_n\left(s', \mathbf{a}^*\right), & (4) \end{cases}$$

where λ is the learning rate, and s' represents the next state $s(t+1)$ after executing action $\mathbf{a}$ at state $s(t)$, $\mathbf{a}^*$ denotes the best joint action, and it can be given by:

$$\mathbf{a}^* \in \arg\max_{\mathbf{a}'} \sum_{n=1}^{N} Q_n\left(s', \mathbf{a}'\right), \tag{5}$$

Algorithm 1 Collaborative Multi-Agent Anti-Jamming Algorithm (CMAA)

```
Initiate: $\mathcal{S}$, $Q_n$, $n \in N$.
Loop: $t = 0, \cdots, T$
    Each user observes its current state $s(t) = \{\mathbf{a}(t), \tilde{f}_{jx}(t)\}$, and selects a channel according to the following rules:
    • User $n$ randomly chooses a channel profile $\mathbf{a} \in \mathcal{A}$ with probability $\varepsilon$;
    • User $n$ chooses a channel profile $\mathbf{a}^* \in argmax \sum_{n=1}^{N} Q_n\left(s', \mathbf{a}'\right)$ with probability $1 - \varepsilon$.
    Each user measures its payoff $r_n(s, \mathbf{a})$.
    The state is transferred into $s(t+1) = \{\mathbf{a}(t+1), \tilde{f}_{jx}(t+1)\}$, and Q values are updated according to the rules in (3).
End loop
```

It is noted that the multi-agent Q-learning algorithm in (3) is decentralized, and each user updates its Q values separately. However, for the problem in (5), it is necessary to solve a global coordination game, which has common payoff [18]:

$$Q(s, \mathbf{a}) = \sum_{n=1}^{N} Q_n(s, \mathbf{a}). \tag{6}$$

The exchanged information between users is only the Q-value, and each user broadcasts its current Q value to other users at the end of each time slot. The exploration rate $\varepsilon \in (0, 1)$ is introduced to avoid falling into a local optimum. Users randomly select a joint action $\mathbf{a} \in \mathcal{A}$ with probability ε, and users select the joint action $\mathbf{a}^* \in argmax \sum_{n=1}^{N} Q_n(s', \mathbf{a}')$ with probability $1 - \varepsilon$. Based on the above analysis, a collaborative multi-agent anti-jamming algorithm (CMAA) is proposed, and its implementation procedure is shown in Algorithm 1. The convergence condition of the proposed Algorithm 1 is either that the Q-table is no longer updated, or that its change is small. Motivated by [19], we analyze the computational complexity of Algorithm 1. In one time slot, according to equation (5), each cognitive user selects the joint action, and its computational complexity can be expressed as $\mathcal{O}(F_1)$, in which F_1 denotes a constant. For each cognitive user, it updates its Q values according to equation (3), and its computational complexity can be given by $\mathcal{O}(F_2)$, in which F_2 represents another constant. In our system, it is assumed that there are N users, and the number of time slots is T_{num}. Then, the computational complexity of Algorithm 1 can be given by:

$$C = \sum_{n=1}^{N} T_{num}[\mathcal{O}(F_1) + \mathcal{O}(F_2)]. \tag{7}$$

Similar to [8], the wideband spectrum sensing is adopted to sense the jammer's activities, and all Q values are updated simultaneously. A transmission slot structure diagram is

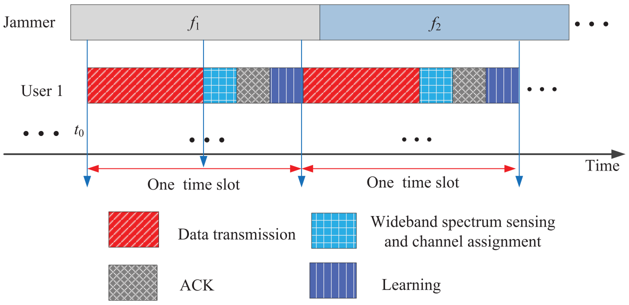

Fig. 2. Illustration of the transmission slot structure.

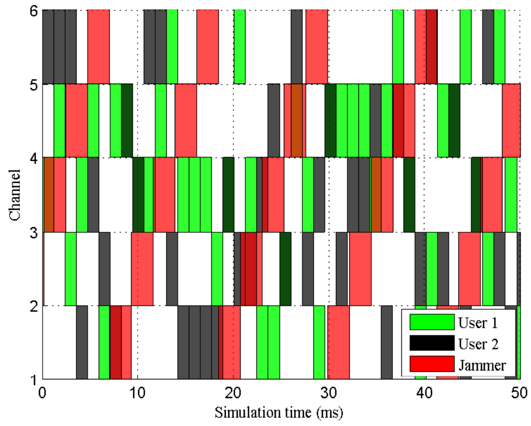

Fig. 3. Time-frequency information at initial state.

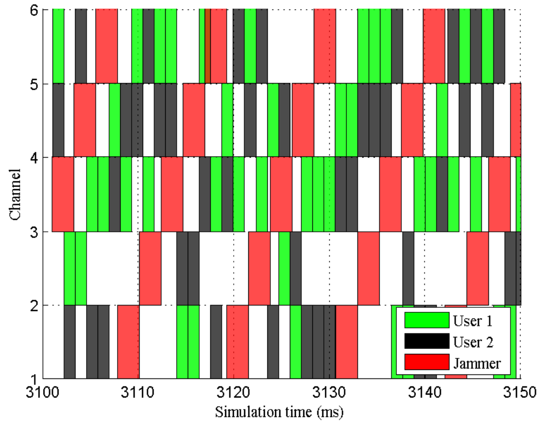

Fig. 4. Time-frequency information at convergent state.

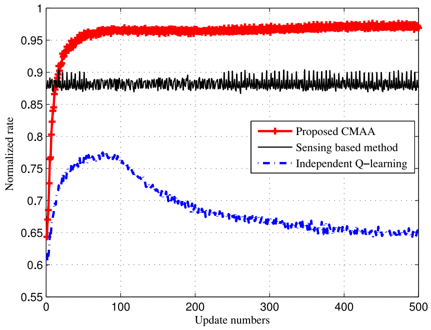

Fig. 5. Performance comparison of the normalized rate.

presented in Fig. 2. At the end of current slot, each user obtains a reward, and updates its strategy according to the received reward.

IV. Numerical Results and Discussions

In this subsection, we present some simulation results. A system with two cognitive users and one jammer is considered, in which five channels are available. t_{Rx}, t_{WBSS}, t_{ACK}, and $t_{Learning}$ denote the transmission time, wideband sensing time, ACK time, and learning time, respectively. In this letter, we have $P_{1,md} = \cdots = P_{M,md} = P_{md}$ and $P_{1,fa} = \cdots = P_{M,fa} = P_{fa}$ to analyze the impact of false alarm and miss detection. The jammer begins to jam the transmission at time slot $t = 0.2ms$. Referring to [9], the simulation parameters are given as: $\lambda = 0.8$, $\gamma = 0.6$, $\varepsilon = 0.2$, $t_{Rx} = 0.98ms$, $t_{WBSS} + t_{ACK} + t_{Learning} = 0.2ms$. Moreover, the dwelling time of the sweeping jammer on each channel is $t_{dwell} = 2.28ms$, the number of time slots for simulations is $K = 10000$, and the simulation time is $T = K * (t_{Rx} + t_{WBSS} + t_{ACK} + t_{Learning})$.

Fig. 3 and Fig. 4 respectively show the time-frequency information at the initial and convergent state. As indicated in Fig. 3, at the initial stage, the cognitive users employ random actions, and the signals of cognitive users and jammer are overlapped. Moreover, the signals among users are also overlapped. Fig. 4 shows the time-frequency information of the proposed CMAA at convergent stage. As can be seen from Fig. 4, at convergent stage, the signals of cognitive users can avoid the signal of the jammer. Meanwhile, the signals among cognitive users can effectively cope with the mutual interference, and the actions of cognitive users are coordinated.

To validate the proposed CMAA, we compare it with the following two methods.

- *Sensing Based Method:* In this method, the users cannot learn the actions of the jammer, and channels are selected based on the sensing results. Moreover, we resort to a coordination approach, as in [18], in which user n ($n > 1$) selects its channel a_n^* until the previous users $1, \ldots, n-1$ broadcast their chosen channels in the ordering. Then, user n broadcasts its channel.
- *Independent Q-Learning [11]:* Each user adopts a standard Q-learning method. The coordination among users is not considered, and other users are treated as part of its environment.

In this section, the normalized rate is introduced to validate the performance of the proposed CMAA, and it can be defined as $\rho = PK_{succ}/PN_0$, where PK_{succ} represents the number of packets for successful transmission, and PN_0 denotes the length of packet statistics, which means that the normalized rate ρ is calculated after PN_0 packets are transmitted. In this simulation, we have $PN_0 = 20$. Then, the following results are obtained by making 200 independent runs and then taking the mean.

Fig. 5 shows the performance of the normalized rate, it can be seen that the proposed CMAA can converge in about 80 update numbers. Also, the proposed CMAA is superior to both sensing based method and independent Q-learning method.

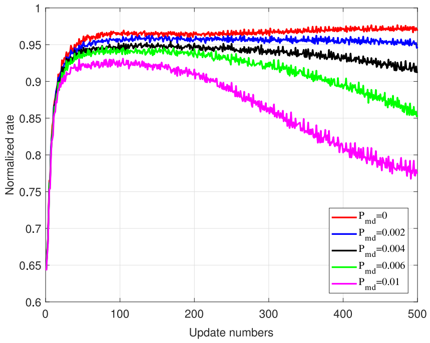

Fig. 6. The impact of miss detection probability.

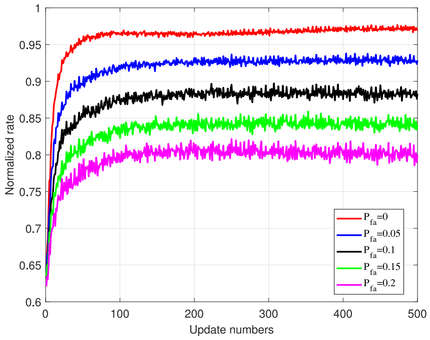

Fig. 7. The impact of false alarm probability.

Moreover, the proposed CMAA has the highest normalized rate ρ. The reason is that the sensing based method cannot learn the actions of the jammer, and channels are chosen based on the sensing results. Meanwhile, the independent Q-learning method does not consider the coordination among users, and each user chooses its channel independently. For the proposed CMAA, it can not only learn the actions of the jammer, but also consider the coordination among users.

Fig. 6 and Fig. 7 show the impact of false alarm probability and miss detection probability. Fig. 6 shows that the normalized rate decreases with growing miss detection probability. The reason is that higher miss detection probability leads to more invalid transmission. In addition, miss detection makes it more difficult for cognitive users to learn jamming rules. As shown in Fig. 7, as the false alarm probability increases, the normalized rate decreases. The reason is that false alarm causes the waste of channels. Moreover, false alarm also makes it more difficult for cognitive users to learn jamming rules.

V. Conclusion

In this letter, we consider the "coordination" and "competition" simultaneously, and the Markov game framework is employed to model and analyze the anti-jamming defense problem. Then, a collaborative multi-agent anti-jamming algorithm (CMAA) is proposed to obtain the optimal anti-jamming strategy. Through collaborative learning, it can cope with the external malicious jamming and the mutual interference caused by competition among users simultaneously. In addition, we consider the impact of miss detection and false alarm. Finally, simulation results are presented. Compared with the sensing based method and independent Q-learning method, the proposed CMAA has the highest normalized rate. Note that we consider a system with two users and one jammer in our simulation. For a system with more users, it is also applicable. In future work, the deep reinforcement learning approach may be a good candidate to obtain faster convergence speed in multi-user scenarios.

References

[1] K. Grover, A. Lim, and Q. Yang, "Jamming and anti-jamming techniques in wireless networks: A survey," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 17, no. 4, pp. 197–215, Dec. 2014.

[2] L. Jia, Y. Xu, Y. Sun, S. Feng, and A. Anpalagan, "Stackelberg game approaches for anti-jamming defence in wireless networks," *IEEE Wireless Commun.*, vol. 25, no. 6, pp. 120–128, Dec. 2018.

[3] D. Yang, G. Xue, J. Zhang, A. Richa, and X. Fang, "Coping with a smart jammer in wireless networks: A Stackelberg game approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4038–4047, Aug. 2013.

[4] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[5] L. Jia *et al.*, "A game-theoretical learning approach for anti-jamming dynamic spectrum access in dense wireless networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1646–1656, Feb. 2019.

[6] Q. Hu and W. Yue, *Markov Decision Processes With Their Applications*. New York, NY, USA: Springer, 2007.

[7] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, May 1992.

[8] F. Slimeni, B. Scheers, Z. Chtourou, and V. Le Nir, "Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, 2015, pp. 1–7.

[9] F. Slimeni, Z. Chtourou, B. Scheers, V. Le Nir, and R. Attia, "Cooperative Q-learning based channel selection for cognitive radio networks," *Wireless Netw.*, to be published. doi: 10.1007/s11276-018-1737-9.

[10] S. Machuzak and S. K. Jayaweera, "Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," in *Proc. IEEE Int. Conf. Commun. China (ICCC)*, 2016, pp. 1–5.

[11] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent reinforcement learning based cognitive anti-jamming," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2017, pp. 1–6.

[12] M. A. Aref and S. K. Jayaweera, "A novel cognitive anti-jamming stochastic game," in *Proc. Cogn. Commun. Aerosp. Appl. Workshop (CCAA)*, 2017, pp. 1–4.

[13] M. A. Aref and S. K. Jayaweera, "A cognitive anti-jamming and interference-avoidance stochastic game," in *Proc. IEEE Int. Conf. Cogn. Informat. Comput. (ICCI*CC)*, Oxford, U.K., 2017, pp. 520–527.

[14] L. Busoniu, R. Babuska, and B. D. Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.

[15] F. Yao and L. Jia, "A collaborative multi-agent reinforcement learning anti-jamming algorithm in wireless networks," *arXiv 2018, arXiv:1809.04374*. [Online]. Available: https://arxiv.org/

[16] Y. Xu *et al.*, "Interference-aware cooperative anti-jamming distributed channel selection in UAV communication networks," *Appl. Sci.*, vol. 8, no. 10, pp. 1–20, Oct. 2018.

[17] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE J. Sel. Topics Signal Pocess.*, vol. 6, no. 2, pp. 180–194, Apr. 2012.

[18] N. Vlassis, *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*. San Rafael, CA, USA: Morgan and Claypool, 2007.

[19] Y. Xu, Q. Wu, L. Shen, J. Wang, and A. Anpalagan, "Opportunistic spectrum access with spatial reuse: Graphical game and uncoupled learning solutions," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4814–4826, Oct. 2013.