Multi-regional Anti-jamming Communication Scheme Based on Transfer Learning and Q Learning

Chen Han[1] and Yingtao Niu[2*]
[1] College of Communications Engineering, Army Engineering University of PLA
Nanjing, China
[e-mail: chenhan2017lgd@163.com]
[2] Nanjing Telecommunication Technology Institute
Nanjing, China
[e-mail: niuyingtao78@hotmail.com]
*Corresponding author: Yingtao Niu

Received January 31, 2018; revised September 18, 2018; revised December 21, 2018; accepted January 30, 2019; published July 31, 2019

Abstract

The smart jammer launches jamming attacks which degrade the transmission reliability. In this paper, smart jamming attacks based on the communication probability over different channels is considered, and an anti-jamming Q learning algorithm (AQLA) is developed to obtain anti-jamming knowledge for the local region. To accelerate the learning process across multiple regions, a multi-regional intelligent anti-jamming learning algorithm (MIALA) which utilizes transferred knowledge from neighboring regions is proposed. The MIALA algorithm is evaluated through simulations, and the results show that the it is capable of learning the jamming rules and effectively speed up the learning rate of the whole communication region when the jamming rules are similar in the neighboring regions.

Keywords: Anti-jamming, Reinforcement learning, Transfer learning, Q-Learning algorithm, Decision-making

http://doi.org/10.3837/tiis.2019.07.001 **ISSN : 1976-7277**

1. Introduction

Due to the broadcast nature of wireless communication, users are vulnerable to malicious attacks [1]. Thus, the anti-jamming issue has been one of the most significant tasks of wireless networks in the past decade. Traditionally, various measures have been put forward for anti-jamming defense. For instance, the spread spectrum techniques are used to improve anti-jamming performance at the expense of spectrum resource [2]. These methods are regarded as spectrally inefficient [3]. Therefore, it is of great significance to seek for the anti-jamming scheme with high spectrum efficiency, especially with a scarce spectrum resource. Moreover, with the development of artificial intelligence, smart jammers can adaptively launch jamming attacks on users according to the reconnaissance results, which could severely decrease the reliability of transmission [4, 5]. It is difficult to counter this jamming by conventional anti-jamming methods, such as frequency-hopping communication schemes, whose anti-jamming strategies are fixed and independent [6].

To address this problem, some artificial intelligence technologies, such as reinforcement learning technology and deep learning technology [6, 7], have also been widely applied for wireless communications, due to its dynamic adaptability on anti-jamming [8]. One of the most commonly used method in AI fields is the reinforcement learning method, which has been utilized to analyze the jamming policy [9], and make anti-jamming decision [10]. So that users adaptively adjust actions according to the jamming policy and establish secure communication. Resource allocation techniques for anti-jamming, such as power control and channel allocation, could achieve better anti-jamming performance through efficient allocation of the communication resources. In terms of power control, a hierarchical learning approach based on reinforcement learning was formulated in [11] to solve the anti-jamming power game. The anti-jamming channel allocation problem is considered as a Markov decision process [6], which is generally solved using a reinforcement learning algorithm, such as Q learning algorithm [12]. In [13], a cross-layer resource allocation approach based on Q learning was proposed to effectively utilize unused spectrum opportunities. In [10], an anti-jamming Q learning algorithm was proposed to avoid sweeping jamming.

Although the anti-jamming performance in the local region has been improved significantly, the transfer learning methods across neighboring regions has never been applied to anti-jamming learning tasks. Experience gained in one learning task can be transferred to help improve learning performance in a related, but different, task [14, 15]. Combining transfer learning methods with the anti-jamming reinforcement learning algorithm, the anti-jamming knowledge acquired from the local region is transferred to neighboring regions to accelerate the learning process and improve the multi-regional anti-jamming performance.

In [16], based on knowledge sharing, a “docitive” paradigm was proposed to improve the learning ability and accuracy of the cognitive radio network and to determine policies for action selection in unvisited states. In [17], the author proposed distributed methods by employing reinforcement learning algorithms to address the problem of cooperative communication within the MAC layer. In [18], a transfer learning scheme was proposed to change the experience base from the recent phase by multi-agent coordination. In [19, 20], a transfer actor-critic method was proposed to minimize energy consumption of radio access networks and accelerate the ongoing learning process.

In existing literature, applications of reinforcement learning algorithms and transfer learning methods in the field of wireless communications have mainly focused on impacting the user’s decision-making in centralized networks [21, 22]. Few of them have employed the

Q-Learning and Transfer-Learning concepts to speed up the learning processes of communication users for anti-jamming, especially in the presence of smart jammers. For specific jamming environment, anti-jamming technology based on Q learning can achieve better performance, but if the wireless environment changes, the method only based on Q learning has to restart learning process to obtain new anti-jamming policies. When there are correlations among the jamming patterns or jamming environment, the methods only based on Q learning may increase the additional consumption of communication resource. However, the anti-jamming technologies, combining Q learning with transfer learning, can solve this problem. In this paper, smart jamming attacks based on the communication probability over different channels is considered. The user utilizes Q learning techniques to analyze the jamming rules and determine corresponding anti-jamming strategy. Then, anti-jamming knowledge is transferred to the neighboring regions to speed up their learning processes.

The structure of this paper is given as follow. Section 2 elaborates the system model and problem formulation. A multi-regional intelligent anti-jamming learning algorithm (MIALA) is proposed in Section 3. Then, Section 4 shows the experimental results and analyzes anti-jamming performance. Section 5 draws the final conclusions.

2. System Model and Problem Formulation

2.1 System Model

The system model concerning the user and jammer are presented in this section. For convenience, **Table 1** lists the used notations.

Table 1. Summation of used notations.

Notations	Explanation
$\mathbf{SJNR}_i$	The SJNR of all the channels in the *i*th region
$C_{i,j}$	The channel capacity of channel *j*
φ_{ij}	The active time of the *j*th channel
ϕ_i	The active time of all channels in the *i*th region
$\mathbf{Pr}_J^{i,t}$	The jamming probability
$\mathbf{S}$	The set of possible states
$\mathbf{A}$	The set of possible actions
$r(s,a)$	Immediate transmission reward
π	Selection strategy
$\Pr_{s,s'}(a)$	The state transition probability
$\mathbf{Q}(s,a)$	Q-value
Tl	Time label
p_u	Transmitting power
p_j	Jammer power
n_0	Noise power
λ	Channel switching cost
W_{ch}	Channel bandwidth
W_J	Jamming signal's bandwidth

T_i	Jamming time for each region
T	Jamming signal's pulse period
K	Iteration count
γ	Discount factor
α_0	Initial learning rate
τ υ ξ_0	Boltzmann model coefficients
ψ	Normalization constant
δ	Time constant
θ	Transfer coefficient

As indicated in **Fig. 1**, the system model includes one jammer and M users which are located in M neighboring regions. The available spectrum consists of N non-overlapping channels whose bandwidths are W_{ch}. Channel capacity is adopted for anti-jamming performance evaluation, which is highly related to the Signal-to-Jamming-plus-Noise Ratio (SJNR). The SJNR of channels in the ith region is denoted by $\mathbf{SJNR}_i = [SJNR_{i,1}, SJNR_{i,2} \ldots, SJNR_{i,N}]$. The channel capacity of channel j is expressed as:

$$C_{i,j} = W_{ch} \log_2(1 + SJNR_{i,j}) \quad j = 1, 2 \ldots N \tag{1}$$

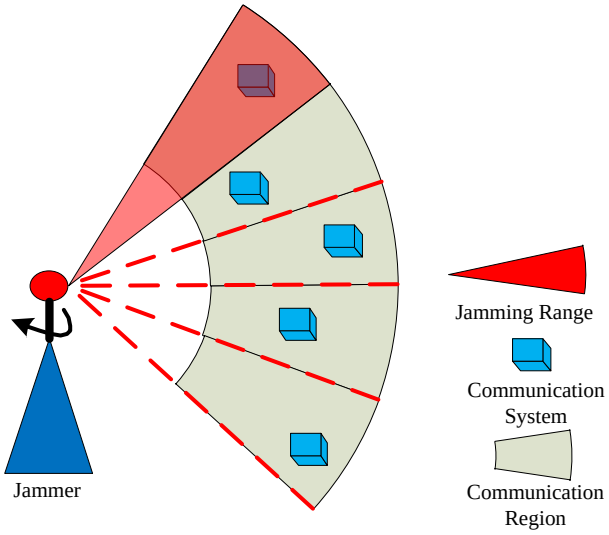

Fig. 1. System model

It is assumed that the smart jammer, limited by power, launches jamming attacks on each region in turn. The jammer detects the active time of all the channels in the ith region, where the term "active" refers to the period of time in which data is being transmitted on the current channel. Then, based on the detection results, the jammer launches jamming attack every T seconds. The jamming time for every region is T_i. W_J is denoted as the bandwidth of jamming signal. The jamming probability is given by:

$$\begin{aligned}\mathbf{Pr}_J^{i,t} &= [\Pr_{J,1}^{i,t}, \Pr_{J,2}^{i,t} \ldots, \Pr_{J,N}^{i,t}], i = 1 \ldots M, t = 1 \ldots T_i \\ \Pr_{J,j}^{i,t} &= \frac{\varphi_{ij}}{\phi_i}, j = 1 \ldots N\end{aligned} \tag{2}$$

where φ_{ij} is denoted as the active time of the *j*th channel and ϕ_i is expressed as the active time of all channels in the *i*th region.

2.2 Problem Formulation

The problem of multi-regional anti-jamming channel allocation is considered as multiple independent Markov decision processes [23]. Let s denote the state, or the current channel, and let $\mathbf{S}$ denote the possible states set, such that $\mathbf{S} = [1, 2 \ldots N], s \in \mathbf{S}$. Let a denote the action, or the next transmission channel to be accessed, and let $\mathbf{A}$ denote the set of possible actions, such that $\mathbf{A} = [1, 2 \ldots N], a \in \mathbf{A}$. An immediate reward for every state-action pair (s, a) is expressed as $r(s, a) = C_a$. The objective of the user is to obtain an optimal policy $\pi^* = \Pr(a \mid s)$ which probabilistically maps state s to action a, in order to maximize the long-term reward [21]. The long-term reward is defined as [23]:

$$V^{\pi}(s) = E\left(\sum_{l=0}^{\infty} \gamma^l r(s_l, a_l) \mid s_0 = s \right) \tag{3}$$

where $\gamma \in (0,1)$ is a factor which represents the important degree of the long-term reward to the current state. s_0 and a_l are expressed as the initial state and the current action, respectively. π is the strategy of channel selection in the state s, whereby $\pi(s_l) \to a_l$.

$\Pr_{s,s'}(a)$ is denoted as the transition probability from state s to s' by implementing action a. Then, the $V^{\pi}(s)$ can be rewritten as:

$$\begin{aligned} V^{\pi}(s) &= E\left(r(s_0, \pi(s_0)) \mid s_0 = s\right) + \\ &\quad E\left(\sum_{l=1}^{\infty} \gamma^l r(s_l, \pi(s_l)) \mid s_0 = s \right) \\ &= E\left(r(s, \pi(s))\right) + \gamma \sum_{s'} \Pr_{s,s'}(a) \times \\ &\quad E\left(\sum_{l=1}^{\infty} \gamma^{l-1} r(s_l, \pi(s_l)) \mid s_0 = s' \right) \\ &= R(s, \pi(s)) + \gamma \sum_{s'} \Pr_{s,s'}(a) \times V^{\pi}(s') \end{aligned} \tag{4}$$

The Q-value is defined as $\mathbf{Q}(s, a)$:

$$\mathbf{Q}^{\pi}(s, a) = R(s, a) + \gamma \sum_{s' \in S} \Pr_{s,s'}(a) \mathbf{Q}^{\pi}(s', a') \tag{5}$$

where $R(s,a)$ is expressed as the mean value of $r(s,a)$. According to Bellman optimality theory, the optimal strategy is given as:

$$\begin{aligned}\mathbf{Q}^*(s,a) &= R(s,a)+\gamma\sum_{s'\in S}\Pr_{s,s'}(a)\max_{a'}[\mathbf{Q}(s',a')]\\ \mathbf{Q}^{\pi^*}(s,a) &= \max_a[\mathbf{Q}^*(s,a)]\end{aligned} \tag{6}$$

where the intermediate maximization of $\mathbf{Q}^{\pi}(s,a)$ is expressed as $\mathbf{Q}^*(s,a)$, and the intermediate Q-value for every next (s',a') is maximized [24]. For every next state s', the optimal action a' is executed [25]. Thus, the optimal policy π^* concerning the current state *s* can be obtained and $\mathbf{Q}^{\pi}(s,a)$ is maximal:

$$\pi^* = \arg\ \max_{\pi}\left\{\mathbf{Q}^{\pi}(s,a)\right\} \tag{7}$$

Once the learning process within the region m_i is completed, the anti-jamming knowledge can be transferred to a neighboring region m_{i+1} to speed up the learning rate within that region.

Unfortunately, there are many difficulties to determine $\Pr_{s,s'}(a)$ and $R(s,a)$. Therefore, this paper proposes a multi-regional intelligent anti-jamming learning algorithm to obtain anti-jamming strategy π^* without priori $\Pr_{s,s'}(a)$ and $R(s,a)$, and share the anti-jamming knowledge with neighboring regions to accelerate the learning rate of the whole communication region.

3. Multi-regional Intelligent Anti-jamming Learning Algorithm

3.1 Anti-jamming Q-Learning Algorithm

The agent using Q-learning is capable of improving the performance by keeping a continuous observation of the state changes and the action rewards within the operational environment.

As indicated in **Fig. 2**, at the state *s*, the user chooses an action *a* and obtains a corresponding reward *r* from the unknown environment. Using the Temporal Difference (TD) method, the agent updates the Q-value according to equation (8), after each execution of an action.

$$\mathbf{Q}_{t+1}(s,a) = \mathbf{Q}_t(s,a)+\alpha(r_t+\gamma\max_{a'}\mathbf{Q}_t(s',a')-\mathbf{Q}_t(s,a)) \tag{8}$$

where $\alpha\in(0,1)$ is used to denote the learning rate.

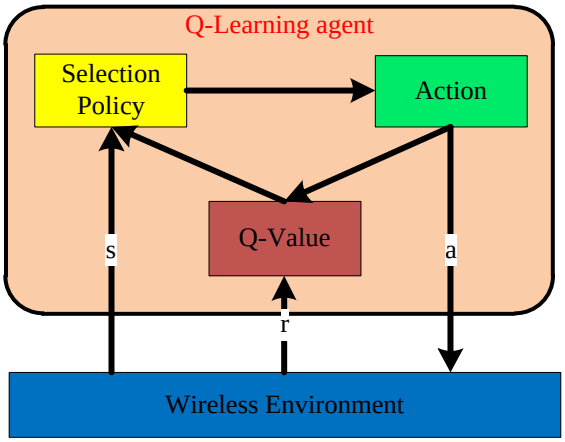

Fig. 2. Q-Learning algorithm

Classical Q-Learning usually utilizes the ε-greedy method to achieve a trade-off between exploring and exploiting [26]. ε is sensitive to the balance between exploration and exploitation. The learning agent can faster explore the environment with a higher ε, but it may result in unsatisfactory performance because of excessive exploration and insufficient exploitation. However, the agent with a low ε maybe converges to the local optimum [26].

Motivated by the Boltzmann model, the modified Q learning algorithm updates the user's policy according to equation (9) to make a smooth transition from exploring to exploiting. For $\xi \to 0$, the learning agent tends to select the action with the maximal Q-value, however, if $\xi \to \infty$, the learning agent's policy is completely random [27].

$$\pi(a \mid s, Tl) = \frac{\tau^{\mathbf{Q}(s,a,Tl)/\xi}}{\tau^{\mathbf{Q}(s,a,Tl)/\xi} + \sum_{a^-} \tau^{\mathbf{Q}(s,a^-,Tl)/\xi}} \tag{9}$$

$$\xi = \xi_0 \tau^{(-\upsilon t)} \tag{10}$$

where $\{Tl \in \mathbf{Tlabel} \mid \mathbf{Tlabel} = [1, 2 \ldots T_i]\}$ is "time label" which marks the time-dependent variables during the learning process in order to analyze the jamming period. a^- represents all channels except the next one to be accessed, ξ_0 impacts exploring time and υ, τ are the coefficients of the Boltzmann model, which affect the transition from exploring to exploiting [6].

Motivated by classical Q-Learning theory, an anti-jamming Q learning algorithm (AQLA) is put forward to solve the anti-jamming channel selection problem [24]. The user chooses a channel a for communication according to the current strategy $\pi(a \mid s, Tl)$, then, the corresponding reward r is given as follows.

$$r_t(s_j, a_j, Tl_t) = C_t - \lambda \left| s_j - a_j \right| \qquad s_j \in \mathbf{S}, a_j \in \mathbf{A} \tag{11}$$

where λ is expressed as the switching cost from current channel s_j to the next channel a_j. Then, the selection strategy is updated according to equation (9), and the Q-value is

updated as follows.

$$\mathbf{Q}_{t+1}(s,a,Tl)=(1-\alpha)\mathbf{Q}_t(s,a,Tl)+\alpha(r_t+\gamma\max_{a'}\mathbf{Q}_t(s',a',Tl)) \tag{12}$$

Afterwards, the Boltzmann coefficients and learning rate are updated according to equations (10) and (13), respectively.

$$\alpha=\alpha_0 / \mu(s,a) \tag{13}$$

where, α_0 is the initial step size, and $\mu(s,a)$ is denoted as the access times to (s,a). The proposed AQLA algorithm is elaborated in Algorithm 1.

Algorithm 1: Anti-jamming Q-Learning Algorithm (AQLA)

Step 1: Initialize $\mathbf{Q}[s,a,Tl]=\mathbf{0}$ and $\pi(a\mid s,Tl)=1/|\mathbf{S}|$. Set $t=0$.
Step 2: Select a_t according to $\pi(a\mid s,Tl)$ and obtain r_t.
Step 3: Update Q-value and $\pi(a\mid s,Tl)$.
Step 4: Set $t=t+1$ and update the learning rate and Boltzmann coefficients.
Step 5: Repeat the process starting from Step 2 until converge.

3.2 Multi-regional Intelligent Anti-jamming Learning Algorithm

Combining transfer learning methods with the AQLA algorithm, this paper proposes a multi-regional intelligent anti-jamming learning algorithm (MIALA) to solve the anti-jamming channel allocation problem across multiple regions. The MIALA algorithm, illustrated in **Fig. 3**, is described as follows.

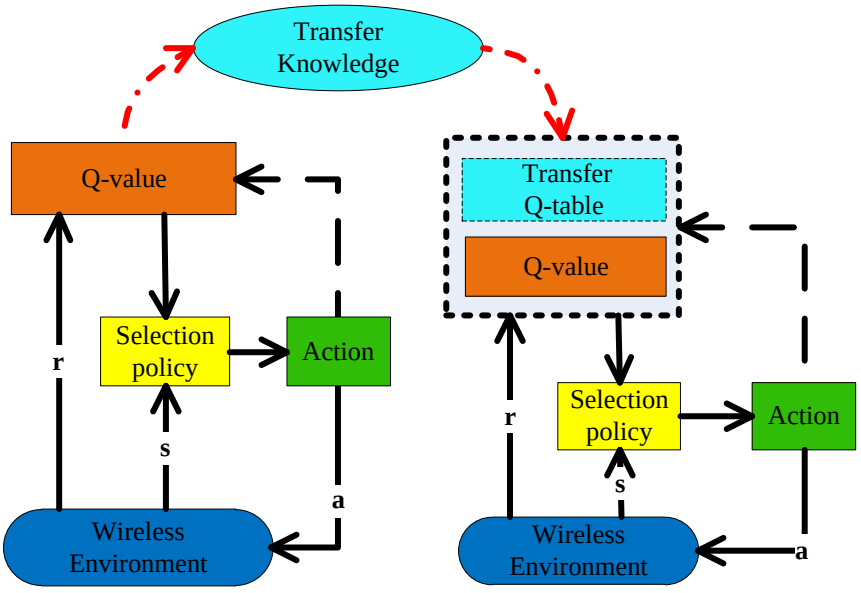

Fig. 3. Multi-regional intelligent anti-jamming learning algorithm

First, the wireless communication user within the region m_i analyzes the jamming rules and determines the anti-jamming strategy using AQLA algorithm. Then, the jamming signal

and anti-jamming knowledge $\mathbf{Q}_i$ learned locally within m_i are transferred to the neighboring region m_{i+1}. The sensing module within m_{i+1} compares the deviation in jamming signals within m_i and m_{i+1}, which is expressed as $\eta_{i,i+1}$. The deviation of the jamming signal is determined using the K-L divergence of the jamming probabilities [28]. The K-L divergence between $p(x)$ and $q(x)$ is denoted as follows:

$$D(p \| q) = \frac{1}{2}\left(\sum_{h=1}^{H} p(x) \log_2 \frac{p(x)}{q(x)} + \sum_{h=1}^{H} q(x) \log_2 \frac{q(x)}{p(x)} \right) \tag{14}$$

Thus, the deviation between $p(x)$ and $q(x)$ is expressed as follows.

$$\eta_{p,q} = \frac{D(p \| q)}{\psi} \tag{15}$$

where ψ is a normalization constant. Afterwards, the user within m_{i+1} initializes $\mathbf{Q}_{i+1} = \mathbf{Q}_i$, $\xi_0^{i+1} = \eta_{i,i+1} \times \xi_0^i$ and begins to learn the anti-jamming policy for m_{i+1}. The $\mathbf{Q}_{i+1}$ is updated as follows:

$$\mathbf{Q}_{i+1}^{t+1} = \beta \mathbf{Q}_i + (1-\beta)\hat{\mathbf{Q}}_{i+1}^{t+1} \tag{16}$$

$$\beta = \theta^{t/\delta} \tag{17}$$

where $\hat{\mathbf{Q}}_{i+1}^{t+1}$ represents the Q-value learned locally within m_{i+1}, β is the transfer rate $\theta \in (0,1]$ is the transfer coefficient and δ is the time constant. The proposed MIALA algorithm is elaborated in Algorithm 2.

Algorithm 2: Multi-regional Intelligent Anti-jamming Learning Algorithm (MIALA)

Step 1: For the current region m_i, initialize $\mathbf{Q}_i$ and $\pi_i(a \mid s, Tl)$. Set $t = 0$, $i = 1$.

Step 2: Implement AQLA algorithm to obtain the anti-jamming knowledge $\mathbf{Q}_i$. Then, transfer $\mathbf{Q}_i$ and the jamming signal in m_i to neighboring region m_{i+1}. Initialize $\mathbf{Q}_{i+1} = \mathbf{Q}_i$, $\hat{\mathbf{Q}}_{i+1} = \mathbf{0}$ and $\xi_0^{i+1} = \eta_{i,i+1} \times \xi_0^i$.

Step 3: Select an action b_{i+1} according to the policy π_{i+1} and obtain the reward r_t^{i+1}.

Step 4: Update selection policy π_{i+1}.

Step 5: Combined with the previous observations and the current reward r_t^{i+1}, update the Q-value and the transfer coefficient.

Step 6: Iterate through AQLA algorithm to obtain the anti-jamming policy within m_{i+1}. Update $i = i + 1$.

Step 7: If each region has obtained the anti-jamming policy of that particular region, stop. Otherwise repeat algorithm starting from Step 2.

The current $\mathbf{Q}(s,a)$ is adjusted towards $r_t + \gamma \max_{a'} Q_t(s',a')$ under the control of learning rate α. Q learning algorithm has been proven to converge to the optimal policy, provided it satisfies Equation (18) [19, 24]. As $\alpha \to 0$, each state is visited enough times. Singh showed that the Boltzmann method is greedy limited by the infinite exploration, given a sufficiently large ξ [29]. So the AQLA algorithm will converge to the optimal policy.

$$\sum_{l=0}^{\infty} \alpha = \infty, \ \sum_{l=0}^{\infty} \alpha^2 < \infty \tag{18}$$

Because of the independence of neighboring regions, each user independently utilizes the AQLA algorithm to obtain the anti-jamming policy of that particular region, but with Q-value initialization based on the anti-jamming knowledge transferred [30, 31]. This transfer of knowledge provides a possible performance jumpstart at the beginning of the learning process within the neighboring region. However, the transfer rate $\beta \to 0$ as $t \to \infty$, so the impact of the transfer knowledge continuously decreases. Inspired by [16, 19], the MIALA algorithm can be shown to converge provided the learning rate α satisfies Equation (18) and the transfer rate β satisfies Equation (19).

$$\lim_{t \to \infty} \frac{\beta}{\alpha} = 0 \tag{19}$$

4. Simulation Results

There are 5 regions $m_i, i = 0,1,2,3,4$ and $N = 5$ channels available for communication within each region. The jammer launches pulse jamming attack on each region in turn according to the communication probabilities of different channels. At the very start, the initial jamming channel is channel 5 and the initial communication channel is channel 2. The jamming probability of the 5 channels within the 5 regions is taken as follows:

$$\mathbf{Pr_0} = [0.05, 0.6, 0.05, 0.3, 0]$$
$$\mathbf{Pr_1} = [0.05, 0.55, 0.05, 0.3, 0.05]$$
$$\mathbf{Pr_2} = [0.05, 0.5, 0.05, 0.3, 0.1]$$
$$\mathbf{Pr_3} = [0.05, 0.45, 0.05, 0.3, 0.15]$$
$$\mathbf{Pr_4} = [0.05, 0.4, 0.05, 0.3, 0.2]$$

The deviation of the jamming signals between m_0 and $m_h, h = 1,2,3,4$ is taken as follows: $\eta_{0,1} = 0.4$, $\eta_{0,2} = 0.5$, $\eta_{0,3} = 0.7$, $\eta_{0,4} = 1$. Other experimental parameters are given in **Table 2**.

Table 2. Experimental parameters

Parameters	Value
Modulation mode	QPSK
Fading coefficients of the 5 channels	$[0.3, 0.9, 0.1, 0.5, 0.7]$
Transmitting power p_u	4dBmW
Jammer power p_j	8dBmW
Noise power n_0	−12dBmW
Channel switching cost λ	0.01
Channel bandwidth W_{ch}	1MHz
Jamming signal's bandwidth W_J	1MHz
Jamming time for each region T_i	9s
Jamming signal's pulse period T	3s
Iteration count K	3000
Discount factor γ	0.01
Initial learning rate α_0	1
Boltzmann model coefficients τ 、 υ ξ_0	1.5 -0.02 10^7
Normalization constant ψ	3
Time constant δ	100
Transfer coefficient θ	0.9

The state set, action set and reward function within m_i are expressed as follows:

$$\mathbf{S}_i = [1,2,3,4,5]$$
$$\mathbf{A}_i = [1,2,3,4,5]$$
$$r_i = W_{ch} \log_2(1 + SJNR_i)$$

The channel quality across all 5 channels within m_0, as measured by average reward, is shown in **Fig. 4**. It shows 3 jamming periods, each of which contains 3 time slots denoted as Time slot 1, Time slot 2 and Time slot 3. At Time slot 1 and 2, the channel sequence 2, 5, 4, 1, 3 is arranged in descending order of average rewards. At Time slot 3, the jammer launches jamming attacks. Channel 2 suffers the greatest reduction in channel quality, channels 1 and 4 are affected to a certain extent, while channels 3 and 5 are insignificantly affected.

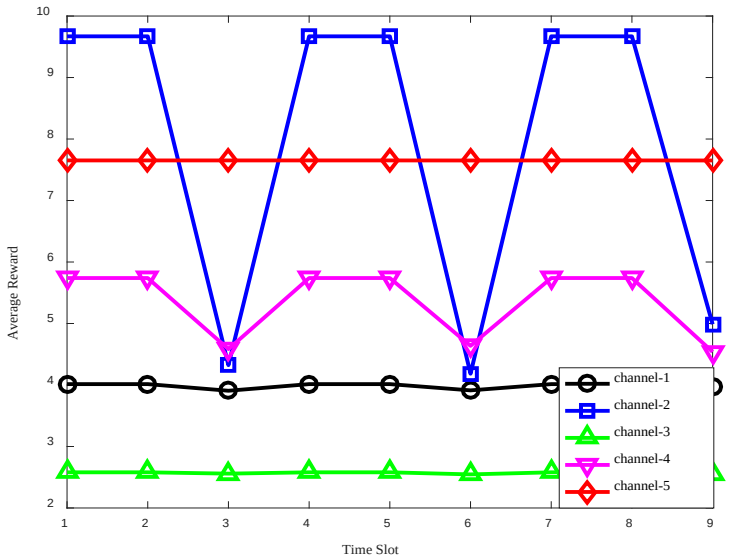

Fig. 4. Channel quality within m_0

Therefore, in each jamming period, it can be seen that the optimal channel selection strategy is selecting channel 2 at Time slot 1 and 2 and selecting channel 5 at Time slot 3. Fig. 5 indicates the convergence of the AQLA algorithm. As indicated in **Fig. 5 (a)**, at Time slot 1 and 2 in every jamming period during the whole learning process, each channel has equal selection probability $P = 0.2$ at the beginning of the algorithm. Then, the selecting probability for channel 2 converges to 1 after 2100 iterations and others converge to 0 as expected. **Fig. 5 (b)** indicates convergence at each Time slot 3 in every jamming period during the whole learning process. Along with the increasing of iteration times, the selecting probability for channel 5 converges to 1 after 2100 iterations as expected.

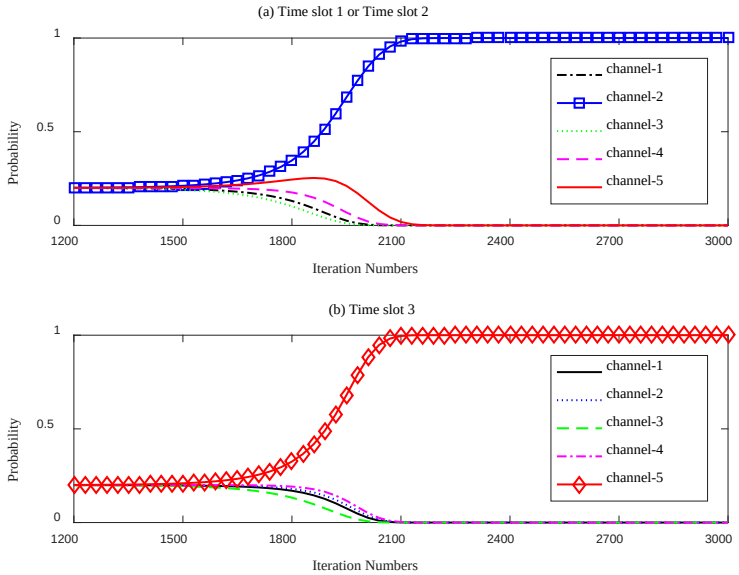

Fig. 5. Convergence of channel selection probabilities within m_0

The AQLA algorithm is compared to the random selecting algorithm (RSA) to evaluate anti-jamming performance. The user utilizes RSA algorithm to randomly selects one channel in a slot. As shown in **Fig. 6**, comparing with RSA, AQLA algorithm always selects the best channel which has the highest SJNR. The average bit error rates of AQLA and RSA, measured every 150 iterations, are compared in **Fig. 7**, which shows that the AQLA algorithm surpasses the RSA algorithm in that it yields lower average bit error rates.

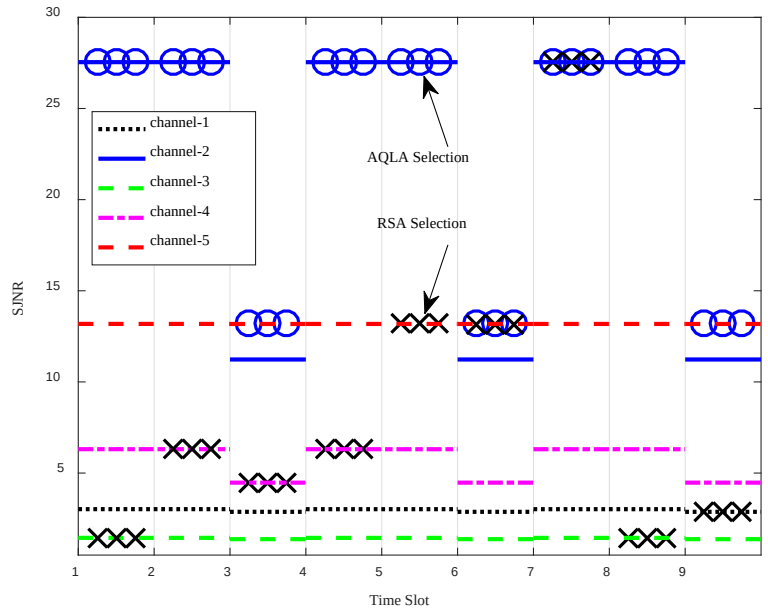

Fig. 6. SJNR of AQLA and RSA

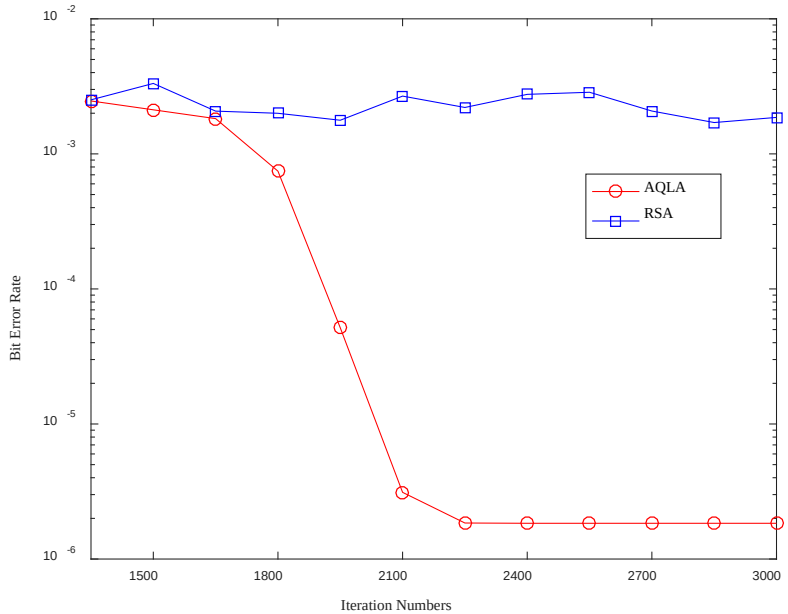

Fig. 7. Average bit error rate of AQLA and RSA

The channel qualities within m_h are shown in **Fig. 8**. The jammer launches similar jamming attacks on the regions m_1, m_2, m_3, m_4, whereby channel 2 is selected prior to jamming and channel 5 is selected after jamming.

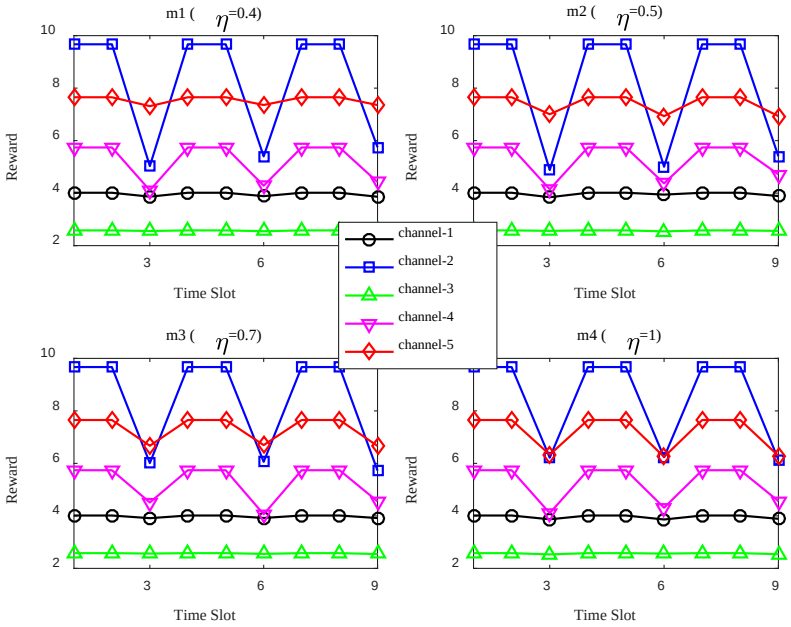

Fig. 8. Channel quality within regions m_1, m_2, m_3, m_4.

To evaluate the anti-jamming performance, the MIALA algorithm is compared to the independent learning algorithm (ILA), in which the user in each region independently learns the optimal channel selection policy using the AQLA algorithm. The convergence of channel selection policies and average bit error rates of both algorithms are indicated in **Fig. 9** and **Fig. 10**.

As shown in **Fig. 9**, the proposed MIALA algorithm converges to the optimal channel selection policy in fewer iterations than ILA algorithm. In **Fig. 10**, it is shown that the MIALA algorithm surpasses the ILA algorithm in that it yields lower average bit error rates. Because of the similarity of jamming rules within m_1, m_2, m_3, m_4 and those within m_0, the anti-jamming channel selection policy learned within $m_h, h=1,2,3,4$ is expedited by the transfer of knowledge from m_0. Furthermore, it can be seen from **Fig. 9** and **Fig. 10** that the impact of transferred knowledge is related to the deviation in jamming rules η, the smaller the deviation, the greater the performance improvement.

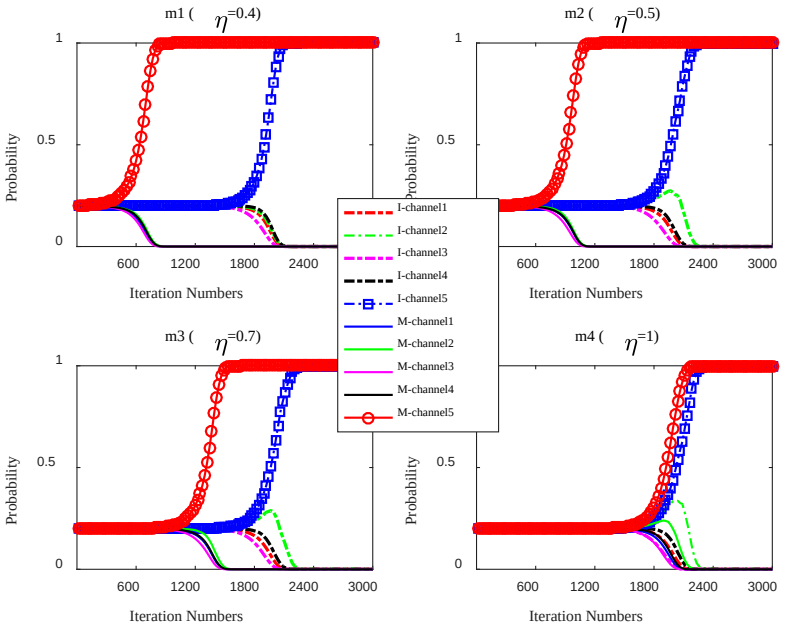

Fig. 9. Convergence of channel selection probabilities within m_1, m_2, m_3, m_4. (I denotes the ILA algorithm and M denotes the MIALA algorithm)

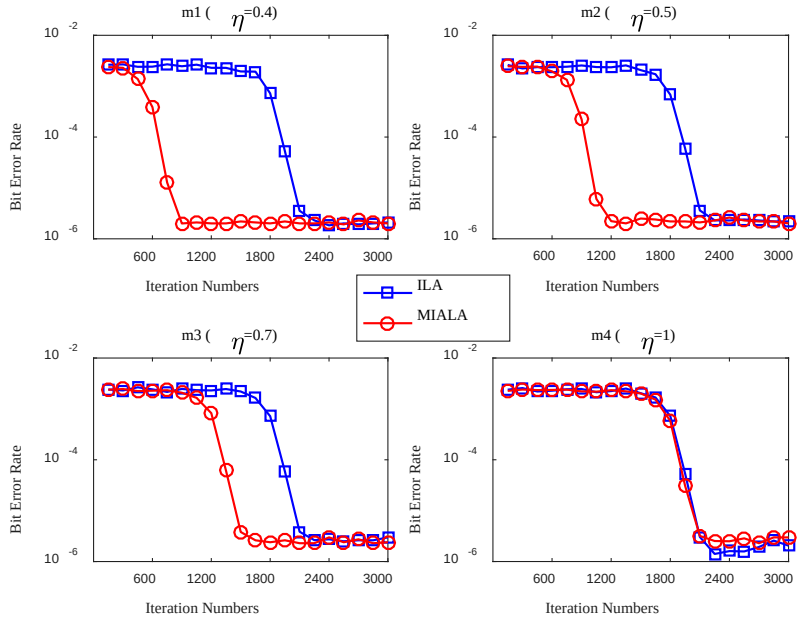

Fig. 10. Average bit error rates of ILA and MIALA within m_1, m_2, m_3, m_4.

5. Conclusion

Based on Transfer-Learning and modified Q-Learning, this paper proposes a multi-regional intelligent anti-jamming learning algorithm (MIALA) to solve the problem of anti-jamming channel selection across multiple regions. The simulation results have proven that, based on the local environment and transferred knowledge, the MIALA algorithm is capable of learning the jamming rules of the smart jammers and effectively speed up the learning rate of the whole communication region when the jamming rules are similar in the neighboring regions.

References

[1] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends," in *Proc. of the IEEE*, vol. 104, no. 9, pp. 1727-1765, September, 2016. Article (CrossRef Link)

[2] L. Zhang, Z. Guan, and T. Melodia, "United against the enemy: anti-jamming based on cross-layer cooperation in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5733-5747, August. 2016. Article (CrossRef Link)

[3] L. Jia, Y. Xu Y, Y. Sun, A. Anpalagan. "Stackelberg Game Approaches for Anti-jamming Defence in Wireless Networks," *IEEE Wireless Communications*, vol. 25, no. 6, pp. 120-128, 2018. Article (CrossRef Link)

[4] L. Xiao, H. Dai, and P. Ning, "MAC Design of Uncoordinated FH-Based Collaborative Broadcast," *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 261-264, June, 2012. Article (CrossRef Link)

[5] X. He, H. Dai, and P. Ning, "A Byzantine Attack Defender in Cognitive Radio Networks: The Conditional Frequency Check," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2512-2523, October, 2013. Article (CrossRef Link)

[6] C. Han, Y. Niu. "Cross-Layer Anti-Jamming Scheme: A Hierarchical Learning Approach," *IEEE Access*, vol. 6, pp. 34874-34883, June, 2018. Article (CrossRef Link)

[7] J. Gazda, E. Slapak, G. Bugar, D. Horvath, T. Maksymyuk, and M. Jo. "Unsupervised Learning Algorithm for Intelligent Coverage Planning and Performance Optimization of Multitier Heterogeneous Network," *IEEE Access*, Vol. 6, pp. 39807-39819, June, 2018. Article (CrossRef Link)

[8] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A Survey on Machine-Learning Techniques in Cognitive Radios," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1136-1159, October, 2013. Article (CrossRef Link)

[9] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 877-889, April, 2011. Article (CrossRef Link)

[10] S. Machuzak and S. K. Jayaweera, "Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," in *Proc. of IEEE/CIC Conf. on Communications in China*, pp. 1-5, July 27-29, 2016. Article (CrossRef Link)

[11] L. Jia, F. Yao, Y. Sun, Y. Xu, S. Feng, and A. Anpalagan, "A Hierarchical Learning Solution for Anti-jamming Stackelberg Game with Discrete Power Strategies," *IEEE Wireless Communications Letters*, vol. PP, pp. 1-1, August, 2017. Article (CrossRef Link)

[12] M. L. Littman, "Value-function reinforcement learning in Markov games," *Cognitive Systems Research*, vol. 2, no. 1, pp. 55-66, April, 2001. Article (CrossRef Link)

[13] M. B. Ghorbel, B. Hamdaoui, M. Guizani, and B. Khalfi, "Distributed Learning-Based Cross-Layer Technique for Energy-Efficient Multicarrier Dynamic Spectrum Access With Adaptive Power Allocation," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1665-1674, March, 2016. Article (CrossRef Link)

[14] M. E. Taylor, and P. Stone, "Transfer Learning for Reinforcement Learning Domains: A Survey,"

Journal of Machine Learning Research, vol. 10, pp. 1633-1685, December, 2009. Article (CrossRef Link)

[15] M. N. Ahmadabadi, and M. Asadpour, "Expertness based cooperative Q-learning," *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*, vol. 32, no. 1, pp. 66-76, February, 2002. Article (CrossRef Link)

[16] A. Galindo-Serrano, L. Giupponi, P. Blasco, and M. Dohler, "Learning from experts in cognitive radio networks: The docitive paradigm," in *Proc. of 5th Int. Conf. on Cognitive Radio Oriented Wireless Networks and Communications*, pp. 1-6. June 9-11, 2010. Article (CrossRef Link)

[17] G. Naddafzadeh-Shirazi, P. Y. Kong, and C. K. Tham, "Distributed Reinforcement Learning Frameworks for Cooperative Retransmission in Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 4157-4162, October, 2010. Article (CrossRef Link)

[18] Q. Zhao, T. Jiang, N. Morozs, D. Grace, and T. Clarke, "Transfer Learning: A Paradigm for Dynamic Spectrum and Topology Management in Flexible Architectures," in *Proc. of IEEE Conf. on Vehicular Technology*, pp. 1-5, September, 2013. Article (CrossRef Link)

[19] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2000-2011, April, 2014. Article (CrossRef Link)

[20] S. Sharma, S. J. Darak, A. Srivastava, and H. Zhang, "A transfer learning framework for energy efficient Wi-Fi networks and performance analysis using real data," in *Proc. of IEEE Conf. on Advanced Networks and Telecommunications Systems*, pp. 1-6, November 6-9, 2016. Article (CrossRef Link)

[21] W. Wang, A. Kwasinski, D. Niyato, and Z. Han, "A Survey on Applications of Model-Free Strategy Learning in Cognitive Wireless Networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1717-1757, March, 2016. Article (CrossRef Link)

[22] K.-L. A. Yau, P. Komisarczuk, and P. D. Teal, "Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues," *Journal of Network and Computer Applications*, vol. 35, no. 1, pp. 253-267, January, 2012. Article (CrossRef Link)

[23] N. Junhong, and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 5, pp. 1676-1687, September, 1999. Article (CrossRef Link)

[24] C. Han, Y. Niu, T. Pang, and Z. Xia, "Intelligent Anti-jamming Communication Based on the Modified Q-Learning," *Procedia Computer Science*, vol. 131, pp. 1023-1031, May, 2018. Article (CrossRef Link)

[25] C. Yih-Shen, C. Chung-Ju, and R. Fang-Chin, "Q-learning-based multirate transmission control scheme for RRM in multimedia WCDMA systems," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 38-48, January, 2004. Article (CrossRef Link)

[26] S. K. Jayaweera, *Signal Processing for Cognitive Radios*, 1nd Edition, Wiley, New York, 2014. Article (CrossRef Link)

[27] H. R. Berenji, and D. Vengerov, "A convergent actor-critic-based FRL algorithm with application to power management of wireless transmitters," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 478-485, August, 2003. Article (CrossRef Link)

[28] S. Kullback, and R. A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79-86, March, 1951. Article (CrossRef Link)

[29] S. Singh, T. Jaakkola, and M. L. Littman, "Convergence Results for Single-Step On-Policy-Reinforcement-Learning Algorithms," *Machine Learning*, vol. 38, no. 3, pp. 287-308, March, 2000. Article (CrossRef Link)

[30] A. Galindo-Serrano, and L. Giupponi, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823-1834, May, 2010. Article (CrossRef Link)

[31] L. Panait, and S. Luke, "Cooperative Multi-Agent Learning: The State of the Art," *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 3, pp. 387-434, November, 2005. Article (CrossRef Link)

Chen Han received his B.S. degree in Electronic Information Engineering from Beihang University, Beijing, China, in 2016, and he is currently working toward the Ph.D. degree in College of Communication Engineering, Army Engineering University of PLA, Nanjing, China. His research interests include learning theory, satellite communication, and communication anti-jamming technology.

Yingtao Niu received his M.S. degree from PLA Commanding Communication Academy, China, in 2005, and received his Ph.D. degree from Institute of Communication Engineering, PLA University of Science and Technology Institute, China. He has authored more than 30 journal and conference papers. His main research interests are spread-spectrum communication, cognitive radio theory and techniques, with particular emphasis on algorithms of wireless communication signal processing and decision-making in cognitive radio systems.