J Supercomput (2015) 71:3237–3257
DOI 10.1007/s11227-015-1420-1

Power control with reinforcement learning in cooperative cognitive radio networks against jamming

Liang Xiao[1] · Yan Li[1] · Jinliang Liu[1] · Yifeng Zhao[1]

Published online: 9 April 2015

Abstract In this paper, we study the anti-jamming power control problem of secondary users (SUs) in a large-scale cooperative cognitive radio network attacked by a smart jammer with the capability to sense the ongoing transmission power. The interactions between cooperative SUs and a jammer are investigated with game theory. We derive the Stackelberg equilibrium of the anti-jamming power control game consisting of a source node, a relay node and a jammer and compare it with the Nash equilibrium of the game. Power control strategies with reinforcement learning methods such as Q-learning and WoLF-PHC are proposed for SUs without knowing network parameters (i.e., the channel gains and transmission costs of others and so on) to achieve the optimal powers against jamming in this cooperative anti-jamming game. Simulation results indicate that the proposed power control strategies can efficiently improve the anti-jamming performance of SUs.

Keywords Cooperative cognitive radio networks · Anti-jamming · Cooperative transmission game · Q-learning · WoLF-PHC

1 Introduction

In cognitive radio networks (CRNs), secondary users (SUs) are allowed to use the temporarily unused licensed spectrum, while avoiding any interference on the communications of primary users (PUs). However, CRNs are especially vulnerable to jamming attacks, due to SUs' opportunistic access and the emergence of smart jam-

✉ Liang Xiao
lxiao@xmu.edu.cn

[1] Department of Communication Engineering, Xiamen University, Xiamen 361005, China

mers that can choose their jamming frequencies and signal strengths according to the transmission strategies of SUs.

In recent years, collaborative anti-jamming techniques were proposed to exploit the node cooperation to improve the communication efficiency against smart jammers [1]. As both jammers and SUs can choose their transmissions with autonomy, game theory is a natural tool to analyze jamming attacks in CRNs [2–5]. Yang et al. formulated the power control interaction between a user and a jammer that can quickly learn the user's power as a Stackelberg game in [2]. However, they considered an ideal scenario by assuming that the players have full knowledge about the others' channel gains and strategies, which are difficult to be obtained in reality. The power allocation between a SU and a jammer was modeled as an incomplete and imperfect information game in [5], in which the jammer has complete information about the SU's location, while the SU does not have such an information of the jammer. Learning techniques have been applied for SUs to further achieve the optimal strategy with unknown information [6–9]. In [6], SUs apply the minimax-Q principle to decide the number of channels for transmitting control and data messages and how to switch between the different channels against jamming. Delayed learning with rewards determined by state transitions rather than the states themselves was developed in [8]. In addition, using stochastic learning, the transmission characteristic that a attacker aims to gain the highest impact out of jamming is learned from the history of attacks in [9].

In this paper, we investigate the anti-jamming power control problem of SUs in a large-scale cooperative cognitive radio network in the existence of a smart jammer to extend the approach proposed in [10]. The smart jammer can quickly learn the transmission strategies of SUs before making a jamming decision. Considering the transmission costs, SUs have to weigh the costs and the utilities to make decisions. The same problem also arises for the jammer. Therefore, our goal is to derive the optimal power control strategies for SUs in the presence of such a smart jammer. First, the interaction between SUs and a smart jammer is formulated as a cooperative transmission game, a Stackelberg game, where the source node is the leader; the relay nodes are the vice leaders and the jammer is the follower. Then, we analyze the optimal strategies for both SUs and the jammer and thus derive the Stackelberg equilibrium (SE) of the game with single relay. The Nash equilibrium (NE) scheme, in which the three players take actions simultaneously, is also derived to compare with the SE scheme. As shown later in the paper, the optimal power control strategies obtained from the SE can minimize the worst-case damage caused by the jammer. Moreover, considering the SUs without knowing the underlying game model, we introduce reinforcement learning methods such as Q-learning [11] and WoLF-PHC [12] for SUs to determine their own transmission strategies in a dynamic environment. Using power control strategies with reinforcement learning methods, the SUs can choose transmission powers based on the observation of the dynamic network environment and thus achieve their own maximum utilities and improve the anti-jamming performance after their learning is quite enough.

The rest of this paper is organized as follows. We briefly review related work in Sect. 2. In Sect. 3, we describe the cooperative transmission game against jamming. Then, the equilibrium in the cooperative transmission game is derived in Sect. 4. Next, the anti-jamming power control strategies with Q-learning and WoLF-PHC are

proposed in Sect. 5. In addition, simulation results are presented in Sect. 6. Finally, Sect. 7 concludes our work.

2 Related work

Game theory has been used to solve many communication problems, such as anti-jamming transmission [3–5,13,14], packet relaying [15] and resource allocation [16]. In [3], the jamming and anti-jamming process between a jammer and SUs has been formulated as a zero-sum game, in which the jammer tries to find the channels of SUs through spectrum sensing and then launches jamming attacks on these channels. Considering the interactions between SUs and jammers, a stochastic zero-sum game model has been introduced to study the channel selection strategies of a jammer and a SU in [4]. The jammer can emit sufficient power over channels to deceive SUs the communication channel in use. In addition, a non-zero-sum game between transmitter and jammer has been formulated in [13], which considers the transmission cost and proves the existence and uniqueness of the Nash equilibrium. In [14], the energy-constrained jammer–defender interaction has been modeled as a zero-sum finite-horizon stochastic game, where a jammer and a sender choose the power and channel to transmit and whether to transmit or sleep. In this paper, we have considered a more smart jammer that can quickly learn the transmission powers of SUs and then adjust its jamming power to attack the SUs.

In addition, reinforcement learning algorithms have been used to legitimate users for defending jamming attacks [17–21]. In particular, two learning schemes have been provided in [17] for SUs to gain knowledge of jammers in the situations without perfect knowledge and thus reduce the probability of being jammed. Reinforcement Q-learning techniques have been applied to solve the optimal channel accessing schemes for legitimate nodes and jammers to cope with a hostile environment in [18]. In [19], a jamming resilient control channel algorithm based on WoLF principle has been proposed to facilitate control channel allocations. In addition, in dynamic environments, channel selection strategies of a jammer and a SU based on reinforcement learning have been proposed in [21].

3 Cooperative transmission game against jamming

We consider a large-scale cooperative CRN, in which R secondary relay nodes help a secondary source node broadcast packets on a single channel to a secondary receiver, as the size of data traffic from the secondary source node is large. However, the CRN is attacked by a smart jammer, which can quickly learn the powers of the source node and relay nodes before making a jamming decision. The legitimate SUs (the source node and relay nodes) and jammer can freely control their transmission powers to achieve their maximum utilities. In CRNs, SUs are allowed to access the channel only when PUs are absent. Thus, at the beginning of each time slot, the source node and relay nodes observe the presence of the PU to avoid interference to the PU. To assist the source node efficiently, when the PU is absent, the relay nodes decide their own transmission powers after rapidly learning the source node's transmission power.

Meanwhile, since the jammer is interested in jamming the SUs but not the PU, it has to first sense the channel to determine the presence of PU. After quickly learning the powers of the source node and relay nodes in the absence of PU, the jammer chooses its jamming power on the channel to make the damaging effect maximized. We assume the PU accesses the channel with a probability p in each time slot. Denote the presence of indicator of PU as δ, which equals zero in the presence of PU and equals one otherwise, that is,

$$\delta = \begin{cases} 0, & \text{PU exists on the channel,} \\ 1, & \text{o.w.} \end{cases} \tag{1}$$

Obviously, Stackelberg game, which is good at dealing with a situation where players take actions sequentially, is a really great tool to model the cooperative power control problem in our system. Under the above analysis, we formulate the power control problem as a Stackelberg game. We consider the source node as a leader, the relay nodes as vice leaders and the jammer as a follower in this game. The action of each player is its own transmission power. The powers of the source node and jammer can be denoted by $x \in [0, \infty)$ and $z \in [0, \infty)$, respectively. In addition, let $h_s > 0$ and $h_j > 0$ denote the fading channel gains of the source node and jammer, respectively. Similarly, let $y_w \in [0, \infty)$ and $h_r^w > 0$ $(1 \le w \le R)$ denote the transmission power and channel gain of the w-th relay node, respectively. Moreover, we consider the transmission cost of each player and let $C_s > 0$ and $C_j > 0$ be the transmission costs per unit power of the source node and jammer, respectively. The transmission costs per unit power of the w-th relay node are denoted by $C_r^w > 0$ as well. The signal to interference plus noise ratio (SINR) at the receiver is given by

$$\text{SINR} = \frac{h_s x + \sum_{w=1}^{R} h_r^w y_w}{N + h_j z}, \tag{2}$$

where N is the noise power.

For the source node, its purpose is to achieve the maximum SINR with the minimum cost. Thus, according to Eq. (2), the utility function of the source node denoted as u_s is given by

$$u_s(x, \mathbf{y}, z) = \delta \frac{h_s x + \sum_{w=1}^{R} h_r^w y_w}{N + h_j z} - C_s x. \tag{3}$$

Similarly, the total utility function of the R relay nodes denoted as u_r is given by

$$u_r(x, \mathbf{y}, z) = \delta \frac{h_s x + \sum_{w=1}^{R} h_r^w y_w}{N + h_j z} - \sum_{w=1}^{R} C_r^w y_w. \tag{4}$$

However, the jammer aims at jamming the legitimate SUs as soon as possible by initiating a jamming attack. Thus, any gains of the SUs result in a corresponding loss of the jammer. On the contrary, any costs of the SUs are the jamming gains. Accordingly, the utility function of jammer denoted by u_j is expressed as

$$u_j(x, \mathbf{y}, z) = -\delta \frac{h_s x + \sum_{w=1}^{R} h_r^w y_w}{N + h_j z} + C_s x + \sum_{w=1}^{R} C_r^w y_w - C_j z. \tag{5}$$

In summary, this cooperative anti-jamming transmission game denoted as $\mathbf{G} = \langle \{s, r_1, r_2, \ldots, r_R, j\}, \{x, \mathbf{y}, z\}, \{u_s, u_r, u_j\} \rangle$ consists of a source node as a leader, R relay nodes as vice leaders and a jammer as a follower; the actions of these players are their own transmission powers; the utilities of these players are shown as Eqs. (3), (4) and (5). Each player takes its optimal transmission power to maximum its utility, respectively. In addition, for simplicity, we assume the energy of these wireless radios is limitless. Even so, they would not choose oversize power to emit signal, because their utilities become reduced with the increasing of the cost to a certain degree. For ease of reference, the usual used notations are summarized in Table 1.

4 Equilibrium in the cooperative transmission game

In this section, Stackelberg equilibrium (SE) in the cooperative transmission game is analyzed. For simplicity, we assume that there is a relay selection algorithm for the secondary source node to choose a best secondary relay node from the R available secondary relay nodes, which may has the shortest path or the lowest transmission cost and so on. Thus, there is only one relay in the cooperative transmission game. The optimal power control strategies in the game are given by

$$x^* = \arg \max_{x \geq 0} \; u_s(x, y^*, z^*) \tag{6}$$

$$y^* = \arg \max_{y \geq 0} \; u_r(x^*, y, z^*) \tag{7}$$

$$z^* = \arg \max_{z \geq 0} \; u_j(x^*, y^*, z). \tag{8}$$

Through the above analysis, the SE in the cooperative transmission game between the source node, relay node and jammer involves solving three optimization problems: two from the viewpoint of SUs and another from the viewpoint of the jammer. In the game, based on the impact on the other players by its power, the source node first chooses the optimal transmission power to maximize its utility as Eq. (3). Then, considering the source node's strategy and the impact on the jammer by its transmission power, the relay node sends the message with the optimal power to achieve the maximum utility as Eq. (4). Finally, after observing the transmission powers of the source node and relay node, the jammer emits the optimal jamming signal to get the highest utility as Eq. (5). As a consequence, the game $\mathbf{G}$ reaches its Stackelberg equilibrium, which consists of the optimal transmission powers of the three players. To derive the SE, we first analyze the impacts of the source node's power on the jammer and the relay node.

4.1 Optimal power allocation of the jammer

To maximize the jammer's utility defined by Eq. (5), the optimal jamming power of the jammer denoted by z^{SE} is given from the following optimization problem:

Table 1 Summary of symbols and notations

Symbols	Notations
R	Number of relay nodes
p	Access probability of PU
δ	Presence indicator of PU
$i(=s, r, j)$	Player i (source node, relay node or jammer)
x	Transmission power of the source node
y_w	Transmission power of the w-th relay node
z	Jamming power
$C_{s/j}$	Transmission cost per unit power of the source node or jammer
C_r^w	Transmission cost per unit power of the w-th relay node
$h_{s/j}$	Channel gain of the source node or jammer
h_r^w	Channel gain of the w-th relay node
N	Noise power
u_i	Utility function of player i
$(x^{\mathrm{SE}}, y^{\mathrm{SE}}, z^{\mathrm{SE}})$	Stackelberg equilibrium strategy
$(x^{\mathrm{NE}}, y^{\mathrm{NE}}, z^{\mathrm{NE}})$	Nash equilibrium strategy
M_i	Total power level number of player i
$\mathbf{A}_i$	Power action set of player i
$k > 0$	Quantitative factor of power
$\lambda \in \mathbf{A}_s$	Action of the source node
$\mu \in \mathbf{A}_r$	Action of the relay node
$\nu \in \mathbf{A}_j$	Action of the jammer
$\mathbf{s}_i^n$	State observed by player i in the n-th time slot
$\alpha_i \in (0, 1]$	Learning rate of player i
$\beta_i \in (0, 1]$	Discount factor of player i
$Q_i\left(\mathbf{s}_i^n, a\right)$	Q-function of player i choosing power a in $\mathbf{s}_i^n$
$V_i\left(\mathbf{s}_i^n\right)$	Maximum Q value of player i in the n-th time slot
ε_i	Probability that the optimal power is not chosen by player i
$\pi_i\left(\mathbf{s}_i, a\right)$	Probability that player i chooses power a in $\mathbf{s}_i$
$\bar{\pi}_i\left(\mathbf{s}_i, a\right)$	Average probability that player i chooses power a in $\mathbf{s}_i$
$\theta_i^{\mathrm{win}} \in [0, 1]$	Learning parameter of player i in WoLF-PHC when winning
$\theta_i^{\mathrm{lose}} \in [0, 1]$	Learning parameter of player i in WoLF-PHC when losing
$K_i\left(\mathbf{s}_i^n\right)$	Occurrence count vector of state $\mathbf{s}_i^n$ observed by player i

$$\max_{z\geq 0} u_j(x, y, z) = \max_{z\geq 0} -\delta\frac{h_s x + h_r y}{N + h_j z} + C_s x + C_r y - C_j z. \tag{9}$$

Lemma 1 *The optimal jamming power allocation of the jammer is given by:*

$$z^{\mathrm{SE}} = \begin{cases} 0, & \delta = 0 \quad \text{or} \quad h_s x + h_r y \leq \frac{C_j N^2}{h_j}, \\ \frac{1}{h_j}\left[\sqrt{\frac{h_j(h_s x + h_r y)}{C_j}} - N\right], & \text{o.w.} \end{cases} \tag{10}$$

Proof First, we analyze the property of the utility function of the jammer u_j as Eq. (5) by differentiating it,

$$\frac{\partial u_j(x, y, z)}{\partial z} = \frac{\delta h_j(h_s x + h_r y)}{(N + h_j z)^2} - C_j \tag{11}$$

$$\frac{\partial^2 u_j(x, y, z)}{\partial z^2} = -\frac{2\delta {h_j}^2(h_s x + h_r y)}{(N + h_j z)^3}. \tag{12}$$

If $\delta = 0$, we see that $u_j(x, y, z) = C_s x + C_r y - C_j z$. $u_j(x, y, z)$ decreases with z and thus $z^{\mathrm{SE}} = 0$.

In the absence of PU, i.e., $\delta = 1$, by Eq. (12), $u_j(x, y, z)$ is a concave function with respect to z. By setting Eq. (11) to 0, we can obtain $\tilde{z} = \sqrt{(h_s x + h_r y)/(h_j C_j)} - N/h_j$. Thus, $z^{\mathrm{SE}} = \tilde{z}$ if $\tilde{z} > 0$ (i.e, $h_s x + h_r y > (C_j N^2)/h_j$). If $\tilde{z} \le 0$, $u_j(x, y, z)$ decreases with z for $z \ge 0$, yielding $z^{\mathrm{SE}} = 0$, and thus we have Eq. (10). □

Therefore, the optimal jamming power varies with the ongoing transmission of SUs and the action of the PU. If the PU is present (i.e., $\delta = 0$) or the jamming gain is less than the jamming cost caused by the ongoing low transmission power (i.e, $h_s x + h_r y \le (C_j N^2)/h_j$), the optimal power of the jammer is to keep silence. In the absence of the PU (i.e, $\delta = 1$), if the current transmission power exceeds a certain threshold (i.e, $h_s x + h_r y > (C_j N^2)/h_j$), the optimal jamming power is to adjust the jamming power according to the current transmission powers of SUs.

4.2 Optimal power allocation of relay

Similarly, by Eq. (4), the optimal transmission power of the relay denoted as y^{SE} is derived from the following optimization problem:

$$\max_{y \ge 0} u_r(x, y) = \max_{y \ge 0} \delta \frac{h_s x + h_r y}{N + h_j z^{\mathrm{SE}}} - C_r y. \tag{13}$$

Lemma 2 *The optimal transmission power of the relay node is given by:*

$$y^{\mathrm{SE}} = \begin{cases} 0, & \Omega_1, \\ \frac{h_r C_j}{4 h_j {C_r}^2} - \frac{h_s x}{h_r}, & \Omega_2, \\ \frac{1}{h_r}\left(\frac{C_j N^2}{h_j} - h_s x\right), & \text{o.w.} \end{cases} \tag{14}$$

where

$$\Omega_1 : \delta = 0 \quad \text{or} \quad x \ge \max\left(\frac{C_j N^2}{h_s h_j}, \frac{{h_r}^2 C_j}{4 h_s h_j {C_r}^2}\right) \quad \text{or} \quad x < \frac{C_j N^2}{h_s h_j}, \frac{h_r}{N} \le C_r,$$

$$\Omega_2 : \delta = 1, \frac{C_j N^2}{h_s h_j} \le x < \frac{{h_r}^2 C_j}{4 h_s h_j {C_r}^2} \quad \text{or} \quad \delta = 1, \quad x < \frac{C_j N^2}{h_s h_j}, \frac{h_r}{N} \ge 2C_r.$$

Proof According to Eq. (4), the utility function of the relay node, u_r, can be written as:

$$u_r(x, y) = \begin{cases} -C_r y, & \delta = 0, \\ \left(\frac{h_r}{N} - C_r\right) y + \frac{h_s}{N} x, & \delta = 1, \quad y \le \gamma, \\ \sqrt{\frac{C_j}{h_j}(h_s x + h_r y)} - C_r y, & \delta = 1, \quad y > \gamma, \end{cases} \tag{15}$$

where $\gamma = [(C_j N^2)/h_j - h_s x]/h_r$. When $\delta = 0$, i.e., the PU accesses the channel, $u_r(x, y, z)$ decreases with y and we have $y^{\mathrm{SE}} = 0$. □

If $\delta = 1$, the property of the utility function of the relay node as Eq. (4) is analyzed as follows. First, u_r is a linear function for $y \le \gamma$. When $\delta = 1$ and $y > \gamma$, we have

$$\frac{\partial u_r(x, y)}{\partial y} = \frac{h_r}{2}\sqrt{\frac{C_j}{h_j(h_s x + h_r y)}} - C_r \tag{16}$$

$$\frac{\partial^2 u_r(x, y)}{\partial y^2} = \frac{-{h_r}^2}{4(h_s x + h_r y)}\sqrt{\frac{C_j}{h_j(h_s x + h_r y)}}. \tag{17}$$

By Eq. (17), u_r is a concave function for $y > \gamma$ and maximized by $\tilde{y} = (h_r C_j)/(4h_j C_r^2) - (h_s x)/h_r$ if $\tilde{y} \ge 0$. To find the optimal y to maximize u_r when $\delta = 1$, we consider the following two cases.

1. $x \ge (C_j N^2)/(h_s h_j)$ (i.e, $\gamma \le 0$): u_r is only a concave function for $y \ge 0$. As $\tilde{y} \le 0$ (i.e, $x \ge ({h_r}^2 C_j)/(4h_s h_j {C_r}^2)$), u_r decreases with y, and thus $y^{\mathrm{SE}} = 0$. Otherwise, if $\tilde{y} > 0$, u_r is maximized on $\tilde{y}$ and we have $y^{\mathrm{SE}} = \tilde{y}$.

2. $x < (C_j N^2)/(h_s h_j)$: u_r is a decreasing concave function for $y > \gamma$. As $\gamma \le \tilde{y}$ (i.e, $h_r/C_r \ge 2N$), u_r is increasing for $0 \le y \le \gamma$. Thus, $u_r(x, \tilde{y})$ is the maximum value of u_r, that is, $y^{\mathrm{SE}} = \tilde{y}$. However, when $\gamma > \tilde{y}$, if $h_r/N > C_r$, u_r increases with y for $0 \le y \le \gamma$ and thus $y^{\mathrm{SE}} = \gamma$, but if $h_r/N \le C_r$, u_r decreases with y for $0 \le y \le \gamma$, so $y^{\mathrm{SE}} = 0$. To sum up, we have Eq. (14).

The relay node's optimal transmission power depends on the source node's transmission power, the action of the PU and its own transmission cost. It can be observed that when the current power of the source node learned by the relay node is large enough (i.e, $x \ge \max[(C_j N^2)/(h_s h_j), ({h_r}^2 C_j)/(4h_s h_j {C_r}^2)]$), the optimal transmission power of the relay node is not sending any more. When the power of the source node is low and the transmission cost of the relay node is large enough (i.e, $x < (C_j N^2)/(h_s h_j), h_r/N \le C_r$), the relay is too powerless to help the source. Otherwise, the relay node's optimal transmission strategy is to adjust its power based on the current transmission power of the source node. In addition, the relay node stops transmitting in the presence of the PU (i.e., $\delta = 0$) to avoid interfering with the PU.

4.3 Optimal power allocation of the source node

To find the source node's optimal transmission power denoted as x^{SE}, based on Eq. (3), we solve the following optimization problem:

$$\max_{x\geq 0} u_s(x) = \max_{x\geq 0} \delta \frac{h_s x + h_r y^{\mathrm{SE}}}{N + h_j z^{\mathrm{SE}}} - C_s x. \tag{18}$$

Lemma 3 *The optimal transmission power of the source node is given by:*

$$x^{\mathrm{SE}} = \begin{cases} 0, & \Phi, \\ \frac{C_j N^2}{h_s h_j}, & \delta = 1,\ C_r \geq \frac{h_r}{N},\ \frac{h_s}{2N} \leq C_s < \frac{h_s}{N}, \\ \frac{h_s C_j}{4 h_j {C_s}^2}, & \text{o.w.} \end{cases} \tag{19}$$

where Φ : $\delta = 0$ or $\delta = 1,\ C_r \leq \frac{h_r}{2N},\ \frac{C_s}{C_r} \geq \frac{h_s}{2h_r}$ or $\delta = 1,\ \frac{h_r}{2N} < C_r < \frac{h_r}{N}, C_s \geq \frac{h_s}{4N}$ or $\delta = 1,\ C_r \geq \frac{h_r}{N},\ C_s \geq \frac{h_s}{N}$.

Proof By substituting z^{SE} and y^{SE} into Eq. (3), the utility function of the source, u_s, is revised to:

$$u_s(x) = \begin{cases} -C_s x, & \delta = 0, \\ \sqrt{\frac{h_s x C_j}{h_j}} - C_s x, & \delta = 1, \quad x \geq \max(\gamma_1, \gamma_2), \\ -C_s x + \frac{C_j N}{h_j}, & \delta = 1, \quad x < \gamma_1, \quad C_r < \frac{h_r}{N} < 2C_r, \\ \left(\frac{h_s}{N} - C_s\right) x, & \delta = 1, \quad x < \gamma_1, \quad \frac{h_r}{N} \leq C_r, \\ \frac{h_r C_j}{2 h_j C_r} - C_s x, & \text{o.w.} \end{cases} \tag{20}$$

where $\gamma_1 = (C_j N^2)/(h_s h_j)$ and $\gamma_2 = ({h_r}^2 C_j)/(4 h_s h_j {C_r}^2)$. First, if $\delta = 0$, u_s decreases with the power of the source node, x, and we have $x^{\mathrm{SE}} = 0$. Then, we analyze the nature of u_s in the case $\delta = 1$ as follows. For $x < \max(\gamma_1, \gamma_2)$, u_s is a linear function. But when $x \geq \max(\gamma_1, \gamma_2)$, we have

$$\frac{\partial u_s(x)}{\partial x} = \frac{1}{2}\sqrt{\frac{h_s C_j}{h_j x}} - C_s \tag{21}$$

$$\frac{\partial^2 u_s(x)}{\partial x^2} = -\frac{1}{4x}\sqrt{\frac{h_s C_j}{h_j x}}. \tag{22}$$

and u_s is a concave function and maximized by $\tilde{x} = (h_s C_j)/(4 h_j {C_s}^2)$. To find the optimal x to maximize u_s, we consider the following three cases:

1. $h_r/N \geq 2C_r$ (i.e, $\gamma_1 \leq \gamma_2$): u_s is a decreasing linear function for $0 \leq x \leq \gamma_2$. 1.1) If $h_s/C_s \leq h_r/C_r$ (i.e, $\tilde{x} \leq \gamma_2$), u_s decreases with x for $x > \gamma_2$ and thus $x^{\mathrm{SE}} = 0$. 1.2) If $h_s/C_s > h_r/C_r$, u_s is maximized by $\tilde{x}$ for $x > \gamma_2$. we compare $u_s(0)$ with $u_s(\tilde{x})$ to find the maximum u_s. If $u_s(0) \geq u_s(\tilde{x})$, we have $x^{\mathrm{SE}} = 0$, otherwise, $x^{\mathrm{SE}} = \tilde{x}$.

2. $C_r < h_r/N < 2C_r$: u_s decreases with x for $0 \leq x \leq \gamma_1$. 2.1) If $h_s/C_s \leq 2N$ (i.e, $\tilde{x} \leq \gamma_1$) , u_s is a decreasing concave function for $x > \gamma_1$ and thus $x^{\mathrm{SE}} = 0$. 2.2) If $h_s/C_s > 2N$, $u_s(\tilde{x})$ is the maximum value of u_s for $x > \gamma_1$. Thus, if $u_s(0) \geq u_s(\tilde{x})$, we have $x^{\mathrm{SE}} = 0$, otherwise, $x^{\mathrm{SE}} = \tilde{x}$.

3. $h_r/N \leq C_r$: 3.1) if $h_s/C_s \leq 2N$ (i.e, $\tilde{x} \leq \gamma_1$), for $x > \gamma_1$, u_s is a decreasing concave function. In addition, if $h_s/C_s \leq N$, u_s is decreasing for $0 \leq x \leq \gamma_1$ and thus $x^{\mathrm{SE}} = 0$, but if not, u_s decreases with x for $0 \leq x \leq \gamma_1$, thus $x^{\mathrm{SE}} = \gamma_1$. 3.2) If $h_s/C_s > 2N$, $u_s(\tilde{x})$ is the maximum value of u_s for $x > \gamma_1$. As $h_s/C_s > N$, u_s is increasing for $0 \leq x \leq \gamma_1$, thus $x^{\mathrm{SE}} = \tilde{x}$. □

In conclusion, the SE of the cooperative transmission game **G** denoted as (x^{SE}, y^{SE}, z^{SE}) is given by Lemma 1, 2, and 3. Based on Eq. (19), the source node as the leader chooses its transmission power in overall consideration of both the impacts on the relay node and the jammer and the channel conditions for all of them. If the transmission costs of the source node and relay node are both too large (i.e, $C_r \geq h_r/N$, $C_s \geq h_s/N$) or the transmission gain of the relay node is better than that of the source node, the source node's optimal strategy is stopping transmission. Otherwise, the optimal power of the source is to adjust its power based on all channel gains and the jamming cost of the jammer. Similarly, the source node stops sending if the PU accesses the channel (i.e., $\delta = 0$) to avoid interfering with the PU.

4.4 Nash equilibrium of the anti-jamming transmission game

For comparison with the SE scheme, in this section, we consider an anti-jamming transmission game denoted as $\mathbf{G}' = \left\langle \{s, r, j\}, \{x, y, z\}, \left\{u_s, u_r, u_j\right\} \right\rangle$, in which all of players take actions simultaneously. The Nash equilibrium (NE) of anti-jamming transmission game denoted by (x^{NE}, y^{NE}, z^{NE}) is derived in this section. Different from the smart jammer in the SE scheme with the capability to sense the ongoing transmission power before it makes a jamming decision, the jammer in the NE does not have this capability and it makes a jamming at the same time with the transmitters. The NE of the transmission game is the optimal transmission power strategy of the three nodes.

Lemma 4 *The NE of the anti-jamming cooperative transmission game is given by:*

$$(x^{\mathrm{NE}},\ y^{\mathrm{NE}},\ z^{\mathrm{NE}}) = \begin{cases} \left(\frac{h_s C_j}{h_j {C_s}^2}, 0, \frac{1}{h_j}\left(\frac{h_s}{C_s} - N\right)\right), & \Gamma_1, \\ \left(0, \frac{h_r C_j}{h_j {C_r}^2}, \frac{1}{h_j}\left(\frac{h_r}{C_r} - N\right)\right), & \Gamma_2, \\ (0, 0, 0), & \Gamma_3, \end{cases} \tag{23}$$

where $\Gamma_1:\ \delta = 1,\quad C_s \leq \frac{h_s}{N},\quad \frac{C_s}{C_r} \leq \frac{h_s}{h_r},\quad \Gamma_2:\ \delta = 1,\quad C_r \leq \frac{h_r}{N},\quad \frac{C_s}{C_r} > \frac{h_s}{h_r}$ and $\Gamma_3:\ \delta = 0$ or $\delta = 1,\quad C_s > \frac{h_s}{N},\quad C_r > \frac{h_r}{N}$.

Proof First, if the PU accesses the channel (i.e., $\delta = 0$), it is obvious that u_s, u_r and u_j decrease with x, y and z, respectively, and thus (x^{NE}, y^{NE}, z^{NE}) = (0, 0, 0). However, when $\delta = 1$, the derivatives of the utility function Eqs. (3), (4) and (5) are, respectively, given by:

$$\frac{\partial u_s(x,y,z)}{\partial x} = \frac{h_s}{N + h_j z} - C_s \tag{24}$$

$$\frac{\partial u_r(x,y,z)}{\partial y} = \frac{h_r}{N + h_j z} - C_r \tag{25}$$

$$\frac{\partial u_j(x,y,z)}{\partial z} = \frac{h_j(h_s x + h_r y)}{(N + h_j z)^2} - C_j. \tag{26}$$

To compute the NE when $\delta = 1$, we consider the following three cases.

1. $h_s/C_s \geq N$ and $h_s/C_s \geq h_r/C_r$: we set $\partial u_s(x,y,z)/\partial x = 0$ and thus $\hat{z} = (h_s/C_s - N)/h_j \geq 0$ and the utility of the source node as Eq. (3) is a certain number for any x. Let $z^{\mathrm{NE}} = \hat{z}$. As $\partial u_r(x,y,z^{\mathrm{NE}})/\partial y \leq 0$, u_r decreases with y, and thus $y^{\mathrm{NE}} = 0$. To make $z^{\mathrm{NE}} = \hat{z}$, we must have $\hat{z} = \tilde{z}$ and thus $x^{\mathrm{NE}} = (h_s C_j)/(h_j {C_s}^2)$.

2. $h_r/C_r \geq N$ and $h_r/C_r > h_s/C_s$: let $\partial u_r(x,y,z)/\partial y = 0$ and we have $\bar{z} = (h_r/C_r - N)/h_j \geq 0$ and u_r is a certain value for any y. Set $z^{\mathrm{NE}} = \bar{z}$. As $\partial u_s(x,y,z^{\mathrm{NE}})/\partial x \leq 0$, u_s decreases with x, and thus $x^{\mathrm{NE}} = 0$. To make $z^{\mathrm{NE}} = \bar{z}$, we must have $\bar{z} = \tilde{z}$ and thus $y^{\mathrm{NE}} = (h_r C_j)/(h_j {C_r}^2)$.

3. $N > \max(h_s/C_s, h_r/C_r)$: as $\partial u_s(x,y,z)/\partial x \leq 0$, u_s decreases with x, and thus $x^{\mathrm{NE}} = 0$. Similarly, we can get $y^{\mathrm{NE}} = 0$. By integrating x^{NE} and y^{NE} into Eq. (26), we have $\partial u_j(x,y,z)/\partial z < 0$ and thus $z^{\mathrm{NE}} = 0$. Ultimately, we obtain Eq. (23). □

From Eq. (23), in the anti-jamming transmission game, the SUs choose their transmission powers based on their channel conditions and the action of the PU. In the absence of the PU, if the transmission cost of the source node is small and its channel gain is large (i.e., $C_s \leq h_s/N, h_s \geq (h_r C_s)/C_r$), the source node's optimal strategy is to adjust its power based on its channel gain and the transmission cost and that of the jammer while the optimal relay strategy is stopping transmission; otherwise, if the transmission cost of the relay node is small and its channel gain is large (i.e., $C_r \leq h_r/N$, $h_r > (h_s C_r)/C_s$), the optimal relay strategy is to adjust its power based on its channel gain and the transmission cost and that of the jammer while the source node's optimal strategy is not sending any more. In addition, if the transmission costs of the SUs are both too large (i.e, $C_s \geq h_s/N, C_r \geq h_r/N$), the SUs' optimal strategies are stopping transmission. Similarly, the SUs stop sending as well if the PU accesses the channel (i.e., $\delta = 0$).

5 Power control with reinforcement learning against jamming

Reinforcement learning can address the problem how agents ought to take actions in a dynamic environment so as to maximize their cumulative reward. Specially, as the update rule of the Q-function does not require knowledge about the transition and reward functions, Q-learning [11] is model-free algorithm in single-agent case. In multiagent case, by combining the "Win or Learn Fast" principle with hill-climbing principle, WoLF-PHC [12] varies the learning rate used by the algorithm to encourage convergence without sacrificing rationality and thus makes agents learning and adapting to other agents' behaviors [22].

Through above analysis, based on the assumption that the three nodes have the full location knowledge of each other, the SE strategies are derived. The three nodes have the channel gains and transmission costs of each other. However, in practice, the transmission parameters (i.e., channel gain and transmission cost) of a node are usually unknown by other nodes. Considering the above reality, we introduce reinforcement learning methods such as Q-learning and WoLF-PHC for the source node and the relay node to determine their own transmission powers in a dynamic environment without knowing the underlying game model. In addition, the worst-case damage caused by the jammer is considered, where the smart jammer has the full knowledge about the two SUs.

We first define three important components for the three players. Let $i = s, r$ and j denote the source node, the relay node and the jammer, respectively. We assume that the transmission power of player i can be chosen from M_i levels. To simplify the calculation, the power action set of player i is denoted by $\mathbf{A}_i = [P_m]_{1\times M_i}/k$ where $k > 0$ is the quantitative factor of power. Let $\lambda \in \mathbf{A}_s$, $\mu \in \mathbf{A}_r$, $\nu \in \mathbf{A}_j$ denote the transmission power actions taken by the source node, the relay node and the jammer, respectively. Meanwhile, the state observed by player i is denoted by $\mathbf{s}_i$.

In each time slot, the source node, the relay node and the jammer take actions sequentially. At the beginning of the n-th time slot, the source node first sends packets and the decision making of its power λ_n is based on the transmission state in the previous time slot, i.e., $\mathbf{s}_s^n = (\delta_n, \mu_{n-1}, \nu_{n-1})$, where δ_n indicates whether the PU exists on the channel in the n-th time slot. Similarly, after observing the transmission power of the source node, the relay node decides its transmission power μ_n based on the state $\mathbf{s}_r^n = (\delta_n, \lambda_n, \nu_{n-1})$. Finally, based on the observed state $\mathbf{s}_j^n = (\delta_n, \lambda_n, \mu_n)$, the jammer chooses its optimal power ν_n given by Eq. (9). At the end of the n-th time slot, player i can acquire an immediate payoff u_i^n (i.e., the utility value of player i) as shown Eqs. (3), (4) and (5).

5.1 Anti-jamming relay strategy with Q-learning

As a well-known reinforcement learning method based on dynamic programming, Q-learning [11] enables agents to learn how to act optimally in a dynamic environment. The anti-jamming power control strategy based on Q-learning for the relay node is showed as follows. Let $\alpha_r \in (0, 1]$ denote the learning rate of the relay node and $\beta_r \in [0, 1]$ denote the discount factor of the relay node. The Q-function of the relay node with the transmission power μ in the state $\mathbf{s}_r$ is denoted by $Q_r(\mathbf{s}_r, \mu)$. We adopt the update rule of the Q-function in the n-th time slot as follows,

$$Q_r\left(\mathbf{s}_r^n, \mu_n\right) \leftarrow (1-\alpha_r)\, Q_r\left(\mathbf{s}_r^n, \mu_n\right) + \alpha_r\left[u_r^n + \beta_r V_r\left(\mathbf{s}_r^{n+1}\right)\right] \tag{27}$$

$$V_r\left(\mathbf{s}_r^n\right) \leftarrow \max_{\mu\in\mathbf{A}_r} Q_r\left(\mathbf{s}_r^n, \mu\right), \tag{28}$$

where $V_r(\mathbf{s}_r)$ denotes the maximum Q value of the relay node in the state $\mathbf{s}_r$.

In the power control strategy based on Q-learning, the relay node is assumed to use ε-greedy policy to choose its transmission power, where the power with the maximum

Q value in the state $\mathbf{s}_r$ is chosen with a high probability $1-\varepsilon_r$ while other power is taken with an equal low probability $\varepsilon_r/(M_r-1)$. Thus, the probability of power action a taken by the relay node is given by,

$$\Pr(\mu = a) = \begin{cases} 1-\varepsilon_r, & a = \mu^*, \\ \frac{\varepsilon_r}{M_r-1}, & \text{o.w.} \end{cases} \tag{29}$$

where

$$\mu^* = \arg\max_{\mu \in \mathbf{A}_r} Q_r(\mathbf{s}_r, \mu), \tag{30}$$

which is the optimal transmission power of the relay node in the state $\mathbf{s}_r$. The power control strategy of the relay node with Q-learning is shown in detail as Algorithm 1.

Algorithm 1. Power control strategy of the relay node with Q-learning

Set α_r, β_r, $\mathbf{s}_r$, $\mathbf{A}_r$.
Initialization: $Q_r(\mathbf{s}_r, \mu) = \mathbf{0}$, $V_r(\mathbf{s}_r) = \mathbf{0}$, $\forall \mathbf{s}_r, \mu \in \mathbf{A}_r$.
Repeat (for each episode)
  For $n = 1, 2, 3, \ldots$
    Observe the current state $\mathbf{s}_r^n$.
    Select power μ_n at random with the probability by Eq. (29).
    Observe the next state $\mathbf{s}_r^{n+1}$ and immediate payoff u_r^n.
    Update $Q_r(\mathbf{s}_r^n, \mu_n)$ by Eq. (27).
    Update $V_r(\mathbf{s}_r^n)$ by Eq. (28).
  End for
End repeat

5.2 Anti-jamming power control strategy with WoLF-PHC

In a multiagent dynamic environment, WoLF-PHC algorithm as an extension of Q-learning algorithm enables players to learn a moving target, which caused by the fact that the optimal policy of an agent at all time depends on the strategies of the other agents [12]. Thus, we apply the WoLF-PHC algorithm for the source node and relay node to learn their own transmission strategies in a dynamic environment. The anti-jamming learning algorithm based on WoLF-PHC combines the "Win or Learn Fast" principle with hill-climbing principle and thus increases the probability that a player selects the action with the highest Q value.

The anti-jamming relay power control strategy based on WoLF-PHC is described as follows. In the learning algorithm, the updating rule of the Q-function is the same as that in the Q-learning algorithm, i.e., Eqs. (27) and (28). The relay node is assumed to choose its transmission power following a transmission policy $\boldsymbol{\pi}_r : \mathbf{s}_r \rightarrow \Pr(\mathbf{A}_r)$ mapping from the state space to distributions on action which can maximize the expected sum of the discounted reward. Let $\pi_r(\mathbf{s}_r, \mu) \in \boldsymbol{\pi}_r$ denote the probability that the relay node chooses the transmission power μ in the state $\mathbf{s}_r$. To update the transmission policy $\boldsymbol{\pi}_r$ for the relay node, the learning algorithm requires two learning parameters, θ_r^{win} and θ_r^{lose}, where $\theta_r^{\text{win}}, \theta_r^{\text{lose}} \in [0, 1]$ and $\theta_r^{\text{win}} < \theta_r^{\text{lose}}$. By comparing whether the expected Q value of the current transmission policy is greater

than the expected Q value of the current average transmission policy denoted by $\bar{\boldsymbol{\pi}}_r$, the relay node is estimated to be winning or losing. If the expected Q value of $\boldsymbol{\pi}_r$ is higher, the player is winning and thus θ_r^{win} is used to update the transmission policy, otherwise, θ_r^{lose} is used. In the learning process, the occurrence count vector of states observed by the relay node denoted by $\mathbf{K}_r$ is recorded and updated by

$$K_r\left(\mathbf{s}_r^n\right) \leftarrow K_r\left(\mathbf{s}_r^n\right) + 1. \tag{31}$$

Next, the estimated average transmission policy of the relay node, $\bar{\boldsymbol{\pi}}_r$, is updated by

$$\bar{\pi}_r\left(\mathbf{s}_r^n, \mu\right) \leftarrow \bar{\pi}_r\left(\mathbf{s}_r^n, \mu\right) + \frac{\pi_r\left(\mathbf{s}_r^n, \mu\right) - \bar{\pi}_r\left(\mathbf{s}_r^n, \mu\right)}{K_r\left(\mathbf{s}_r^n\right)}, \quad \forall \mu \in \mathbf{A}_r. \tag{32}$$

The relay node in the learning process is assumed to use WoLF-PHC policy to choose its transmission power, where the probability that the relay node chooses the power action with the maximum Q value in a state $\mathbf{s}_r^n$ is gradually increased while the probability that the relay node chooses the other actions is gradually decreased. Thus, the updated rule of the transmission policy of the relay node is given by

$$\pi_r\left(\mathbf{s}_r^n, \mu\right) \leftarrow \pi_r\left(\mathbf{s}_r^n, \mu\right) + \Delta_{\mathbf{s}_r^n, \mu}, \quad \forall \mu \in \mathbf{A}_r, \tag{33}$$

where

$$\Delta_{\mathbf{s}_r^n, \mu} = \begin{cases} -\min\left(\pi_r\left(\mathbf{s}_r^n, \mu\right), \frac{\theta_r}{M_r - 1}\right), & \text{if } \mu \neq \arg\max\limits_{\mu' \in \mathbf{A}_r} Q_r\left(\mathbf{s}_r^n, \mu'\right), \\ \sum\limits_{\mu' \neq \mu} \min\left(\pi_r\left(\mathbf{s}_r^n, \mu'\right), \frac{\theta_r}{M_r - 1}\right), & \text{o.w.} \end{cases} \tag{34}$$

Notice that, the relay node chooses its learning parameter θ_r from parameters θ_r^{win} and θ_r^{lose}, based on the result whether the current expected Q value of π_r is higher than the current expected Q value of $\bar{\pi}_r$, that is,

$$\theta_r = \begin{cases} \theta_r^{\text{win}}, & \Pi, \\ \theta_r^{\text{lose}}, & \text{o.w.} \end{cases} \tag{35}$$

where $\Pi : \sum_{\mu \in \mathbf{A}_r} \pi_r\left(\mathbf{s}_r^n, \mu\right) Q_r\left(\mathbf{s}_r^n, \mu\right) > \sum_{\mu \in \mathbf{A}_r} \bar{\pi}_r\left(\mathbf{s}_r^n, \mu\right) Q_r\left(\mathbf{s}_r^n, \mu\right)$.

The power control strategy of the relay node with WoLF-PHC is shown in detail as Algorithm 2. Similarly, the power control strategy of the source node with WoLF-PHC can be presented as Algorithm 3.

Algorithm 2. Power control strategy of the relay node with WoLF-PHC

Set α_r, β_r, $\mathbf{s}_r$, $\mathbf{A}_r$, θ_r^{win}, θ_r^{lose}.
Initialization: $Q_r(\mathbf{s}_r, \mu) = \mathbf{0}$, $V_r(\mathbf{s}_r) = \mathbf{0}$, $\pi_r(\mathbf{s}_r, \mu) = \mathbf{1}/M_r$, $K_r(\mathbf{s}_r) = \mathbf{0}$, $\forall \mathbf{s}_r, \mu \in \mathbf{A}_r$.
Repeat (for each episode)
  For $n = 1, 2, 3, ...$
    Observe the current state $\mathbf{s}_r^n$.
    Select power μ_n at random with the probability policy $\pi_r(\mathbf{s}_r^n, \mu)$.
    Observe the next state $\mathbf{s}_r^{n+1}$ and immediate payoff u_r^n.
    Update $Q_r(\mathbf{s}_r^n, \mu_n)$ by Eq. (27).
    Update $V_r(\mathbf{s}_r^n)$ by Eq. (28).
    Update $K_r(\mathbf{s}_r^n)$, by Eq. (31).
    Update $\bar{\pi}_r(\mathbf{s}_r^n, \mu)$, by Eq. (32).
    Update $\pi_r(\mathbf{s}_r^n, \mu)$, by Eq. (33).
  End for
End repeat

Algorithm 3. Power control strategy of the source node with WoLF-PHC

Set α_s, β_s, $\mathbf{s}_s$, $\mathbf{A}_s$, θ_s^{win}, θ_s^{lose}.
Initialization: $Q_s(\mathbf{s}_s, \lambda) = \mathbf{0}$, $V_s(\mathbf{s}_s) = \mathbf{0}$, $\pi_s(\mathbf{s}_s, \lambda) = \mathbf{1}/M_s$, $K_s(\mathbf{s}_s) = \mathbf{0}$, $\forall \mathbf{s}_s, \lambda \in \mathbf{A}_s$.
Repeat (for each episode)
  For $n = 1, 2, 3, ...$
    Observe the current state $\mathbf{s}_s^n$.
    Select power λ_n at random with the probability policy $\pi_s^n(\mathbf{s}_s^n, \lambda)$.
    Observe the next state $\mathbf{s}_s^{n+1}$ and immediate payoff u_s^n.
    Update $Q_s(\mathbf{s}_s^n, \lambda_n)$ and $V_s(\mathbf{s}_s^n)$:
      $Q_s(\mathbf{s}_s^n, \lambda_n) \leftarrow (1 - \alpha_s) Q_s(\mathbf{s}_s^n, \lambda_n) + \alpha_s \left[u_s^n + \beta_s V_s \left(\mathbf{s}_s^{n+1} \right) \right]$;
      $V_s(\mathbf{s}_s^n) \leftarrow \max_{m \in \mathbf{A}_s} Q_s(\mathbf{s}_s^n, m)$.
    Update $\bar{\pi}_s(\mathbf{s}_s^n, \lambda)$ and $\pi_s(\mathbf{s}_s^n, \lambda)$:
      $K_s(\mathbf{s}_s^n) \leftarrow K_s(\mathbf{s}_s^n) + 1$;
      $\bar{\pi}_s(\mathbf{s}_s^n, \lambda) \leftarrow \bar{\pi}_s(\mathbf{s}_s^n, \lambda) + \frac{\pi_s(\mathbf{s}_s^n, \lambda) - \bar{\pi}_s(\mathbf{s}_s^n, \lambda)}{K_s(\mathbf{s}_s^n)}, \forall \lambda \in \mathbf{A}_s$;
      $\pi_s(\mathbf{s}_s^n, \lambda) \leftarrow \pi_s(\mathbf{s}_s^n, \lambda) + \Delta_{\mathbf{s}_s^n, \lambda}, \forall \lambda \in \mathbf{A}_s$.
  End for
End repeat

6 Simulation results

In this section, we first show some simulation results to evaluate the system performance of the proposed SE strategy in the anti-jamming cooperative transmission game. In addition, by performing simulations in different dynamic environments without knowing the underlying game model, we evaluate the anti-jamming performance of the proposed power control with reinforcement learning methods against a smart jammer which can sense the power of SUs in advance.

6.1 SE scheme's performance

Some simulations are performed to evaluate the performance of the proposed SE strategy in the anti-jamming cooperative transmission game. The utilities of all players,

u_s, u_r and u_j given by Eqs. (3), (4) and (5), are presented. In addition, the SINR of the SE strategy showed by Eq. (2) is also computed. In these simulations, we assume the transmission cost per unit power of player i, $C_i = 1$ and the noise power, $N = 0.2$. In addition, the PU is assumed to be absent (i.e., its access probability $p = 0$) to present the anti-jamming performance of the SUs adequately.

Figure 1 indicates the impacts of the fading channel gain of the jammer h_j on the utilities of all players with the proposed SE strategy compared with the NE strategy with $h_s = 0.5$ and $h_r = 0.5$. The utilities of the source and relay node decrease with h_j, while the utility of the jammer as Eq. (3) increases with h_j. The reason is that the higher value of h_j indicates the better channel condition for the jammer. Note that, from Eqs. (23) and (3), the utility of the source node with the NE strategy is 0 as shown in Fig. 1, if the transmission cost of the source node is small and its channel gain is large (i.e., $C_s \leq h_s/N, h_s \geq (h_r C_s)/C_r$). In addition, the jammer's utility as Eq. (5) in the SE strategy is higher than the NE strategy, because the jammer can rapidly learn the ongoing power before making a decision in the SE strategy, which is much smarter than the jammer in the NE strategy. A smart jammer with the capability to sense the ongoing transmission power before making a jamming decision can destroy a cooperative CRN more seriously than a common jammer without such a sensing capability. For example, when the fading channel gains of the source and relay nodes are both 0.5 and the jammer's channel gain is 0.1, the total utility of the source and relay nodes in NE strategy is 5, while the total utility in SE strategy is only 3.75.

As shown in Fig. 2, the SINRs in the SE and NE strategies decrease with the fading channel gain of the jammer, h_j, due to the jammer is more damaging when h_j increases. As mentioned before, the jammer in the SE strategy is more intelligent than that in the NE strategy. Consequently, the SINR achieved by the SE strategy is less than that achieved by the NE strategy. In addition, for a fixed h_s and h_j, for example

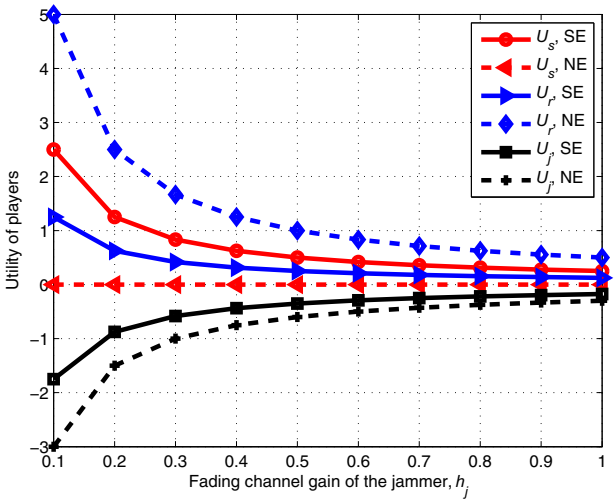

Fig. 1 The utilities of all players vs. the fading channel gain of the jammer, h_j, in a cooperative CRN with $C_s = C_r = C_j = 1$, $h_s = 0.5$, $h_r = 0.5$ and $N = 0.2$

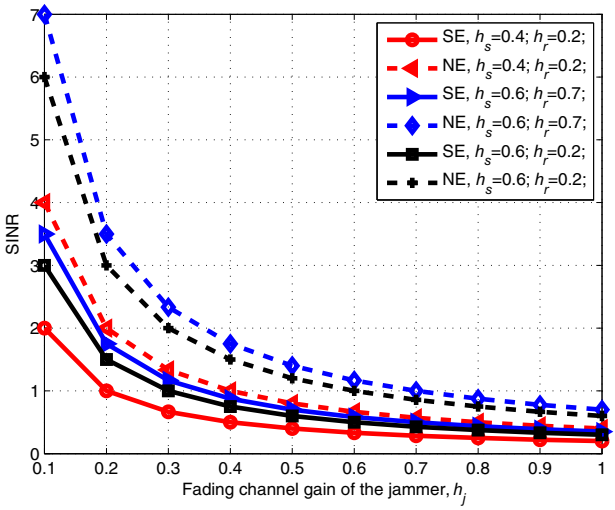

Fig. 2 The SINR vs. the fading channel gain of the jammer, h_j, in a cooperative CRN with $C_s = C_r = C_j = 1$ and $N = 0.2$

$h_s = 0.6$ and $h_j = 0.1$, the SINR increases from 3 to 3.5 in the SE strategy while increases from 6 to 7 in the NE strategy when the fading channel gain of the relay node h_r increases from 0.2 to 0.7. Similarly, if h_r and h_j are fixed, the SINR also increases with h_s. This is because the increases of h_s and h_r indicate a better channel condition for the source and relay node, respectively, which directly results in a higher transmission power received by the receiver.

6.2 Anti-jamming power control learning simulation results

We consider both the scenario with the relay node using power control strategy based on Q-learning and the scenario with the source and relay node using power control strategies based on WoLF-PHC. In the simulations, for simplicity, we set the power action sets of the three nodes, $\mathbf{A}_s = \mathbf{A}_r = \mathbf{A}_j = (0, 1, \ldots, 6)$, the quantitative factor of power $k = 1$ and the PU's access probability $p = 0$. The maximum episode numbers in the learning based on both Q-learning and WoLF-PHC are 1000 to ensure the agent can learn an optimal action. The learning rate of the source node $\alpha_s = 0.7$ which indicates how far the current estimate value of Q is adjusted toward the update target value of Q. In addition, the discount factor of the source $\beta_s = 0.8$ that indicates the increasing uncertainty about rewards that will be received in the future. Similarly, the learning rate of the relay node $\alpha_r = 0.7$ and the discount factor of the relay node $\beta_r = 0.8$.

First, we assume the relay node without full knowledge about the dynamic environment, such as the channel gains and transmission costs of the other nodes, while the source node and jammer have these knowledge. To achieve its optimal utility as Eq. (4), the relay node adopts power control strategy based on Q-learning with the probability that the optimal power is not chosen by the relay node $\varepsilon_r = 0.1$ to ensure

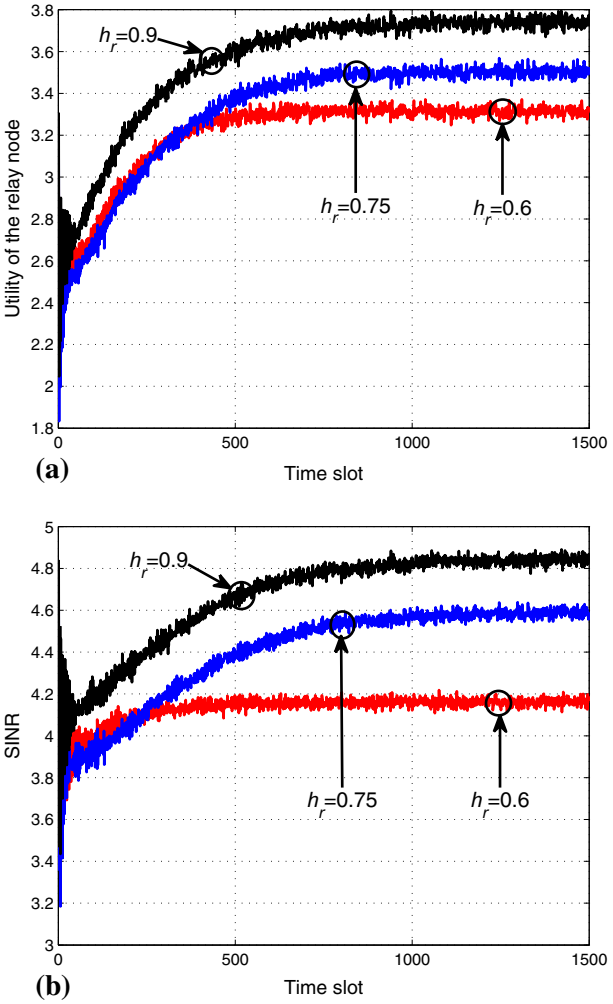

Fig. 3 Anti-jamming performance of a cooperative CRN, where the relay node chooses its transmission power based on Q-learning and $C_s = C_r = 0.5$, $h_s = 0.7$, $N = 0.6$, $C_j = 1$ and $h_j = 0.3$. **a** Utility of the relay node. **b** SINR

that the relay node can try all actions in all states repeatedly. Based on the transmission powers of others in the previous time slot, the source node and the jammer choose their optimal powers of optimization problems as Eqs. (9) and (18), respectively. The utility of the relay node and SINR received by the receiver with the learning time are shown in Fig. 3. Note that, as the relay node is gradually aware of the dynamic environment with the learning time increasing, the utility of the relay node increases as shown in Fig. 3a. In addition, the utility of the relay node increases with the relay channel gain h_r. Meanwhile, it is also shown in Fig. 3b that the SINR increases with the learning time passed, which indicates a well anti-jamming performance. This reason is that the relay node chooses a more proper power after has a well knowledge about the envi-

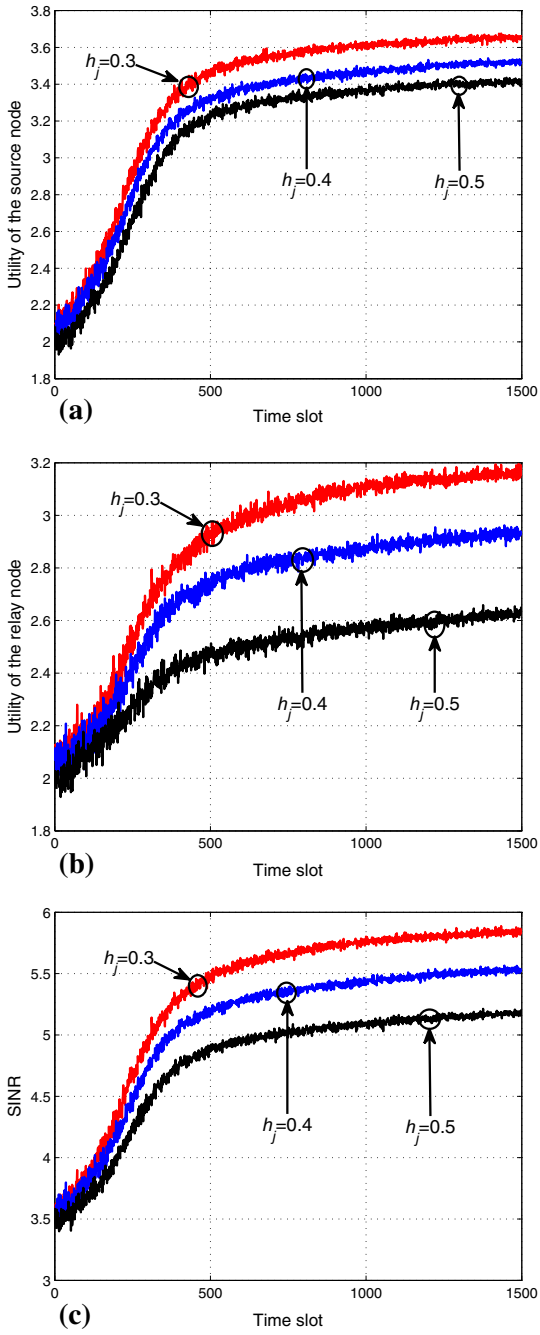

Fig. 4 Anti-jamming performance of a cooperative CRN, where both the source node and relay choose their transmission powers with WoLF-PHC, with $C_s = C_r = 0.5$, $h_s = h_r = 0.9$, $N = 1.5$ and $C_j = 1.5$. **a** Utility of the source node. **b** Utility of the relay node. **c** SINR

ronment. The SINR is also increasing with the relay channel gain h_r, which directly results in a higher transmission power received by the receiver.

Next, we evaluate the anti-jamming performance in the scenario that the source node and the relay node perform their own power control strategies based on WoLF-PHC, against the smart jammer with full knowledge about the dynamic environment. The source node and the relay node set learning parameters $\theta_s^{\text{win}} = 0.01, \theta_s^{\text{lose}} = 0.02$, $\theta_r^{\text{win}} = 0.05$ and $\theta_r^{\text{lose}} = 0.1$, which indicate that the agents should escape fast from losing situations, while adapting cautiously when it is winning, to encourage convergence. Similarly as Fig. 3, the jammer chooses its optimal power of optimization problem as Eq. (9) based on the transmission powers of the SUs in the previous time slot. As shown in Fig. 4, the utilities of the source and relay nodes and the SINR increase with the learning time. The reason is that the source and relay nodes can utilize better strategies to achieve higher SINRs and utilities over time, as they obtain more information about the radio networks via the previous learning process. In addition, the utilities of the source and relay nodes and the SINR deteriorate with the increase of the jamming channel gain h_j, which directly strengthens the jamming quality. In summary, the power control strategies based on WoLF-PHC can efficiently improve the anti-jamming performance of SUs. For example, when the channel gain of the smart jammer is 0.5, the SINR increases from about 3.5 to 5.2 as shown in Fig. 4c.

7 Conclusion

In this paper, we have investigated the cooperative power control problem of SUs in a large-scale cooperative CRN, where relay nodes help the source counteract a smart jammer. The power interaction among two SUs and a jammer is formulated as an anti-jamming transmission game. The Stackelberg equilibrium of the game with a relay has been presented and compared with the Nash equilibrium. Reinforcement learning can be applied by SUs to determine their transmission powers against a jammer in a dynamic environment without knowing the underlying game model. Simulation results have verified that the proposed power control strategies can efficiently improve the anti-jamming performance. For example, if both the source node and relay choose their power with WoLF-PHC against a smart jammer on one channel, the SINR increases from about 3.5 to 5.2.

Acknowledgments This work was supported in part by NSFC (No. 61271242, 61301097, 61440002) and the Fundamental Research Funds for the Central Universities (2013121023).

References

1. Xiao L, Dai H, Ning P (2012) Jamming-resistant collaborative broadcast using uncoordinated frequency hopping. IEEE Trans Inf Forensics Secur 7(1):297–309
2. Yang D, Xue G, Zhang J, Richa A, Fang X (2013) Coping with a smart jammer in wireless networks: a stackelberg game approach. IEEE Trans Wirel Commun 12(8):4038–4047
3. Chen C, Song M, Xin C, Backens J (2013) A game-theoretical anti-jamming scheme for cognitive radio networks. IEEE Netw 27(3):22–27
4. Zhu Q, Li H, Han Z, Basar T (2010) A stochastic game model for jamming in multi-channel cognitive radio systems. In: Proceedings of the IEEE international conference on communications (ICC), pp 1–6

5. El-Bardan R, Brahma S, Varshney P (2014) Power control with jammer location uncertainty: a game theoretic perspective. In: Proceedings of the annual conference on information sciences and systems (CISS), pp 1–6
6. Wang B, Wu Y, Liu K, Clancy T (2011) An anti-jamming stochastic game for cognitive radio networks. IEEE J Sel Areas Commun 29(4):877–889
7. Bkassiny M, Li Y, Jayaweera S (2013) A survey on machine-learning techniques in cognitive radios. IEEE Commun Surv Tutor 15(3):1136–1159
8. Amuru S, Buehrer RM (2014) Optimal jamming using delayed learning. In: Proceedings of the IEEE military communications conference (MILCOM), pp 1528–1533
9. Bhunia S, Sengupta S, Vazquez-Abad F (2014) Cr-honeynet: a learning and decoy based sustenance mechanism against jamming attack in CRN. In: Proceedings of the IEEE military communications conference (MILCOM), pp 1173–1180
10. Li Y, Xiao L, Liu J, Tang Y (2014) Power control stackelberg game in cooperative anti-jamming communications. In: Proceedings of the international conference on game theory for networks, pp 93–98
11. Watkins C, Dayan P (1992) Q-learning. Mach Learn 8(3–4):279–292
12. Bowling M, Veloso M (2002) Multiagent learning using a variable learning rate. Artif Intell 136(2):215–250
13. Altman E, Avrachenkov K, Garnaev A (2007) A jamming game in wireless networks with transmission cost. Netw Control Optim Lect Notes Comput Sci 4465:1–12
14. DeBruhl B, Kroer C, Datta A, Sandholm T, Tague P (2014) Power napping with loud neighbors: optimal energy-constrained jamming and anti-jamming. In: Proceedings of the ACM conference on security and privacy in wireless and mobile networks, pp 117–128
15. Seredynski M, Bouvry P (2013) Analysing the development of cooperation in manets using evolutionary game theory. J Supercomput 63(3):854–870
16. Khan S (2011) Mosaic-net: a game theoretical method for selection and allocation of replicas in ad hoc networks. J Supercomput 55(3):321–366
17. Wu Y, Wang B, Liu K, Clancy T (2012) Anti-jamming games in multi-channel cognitive radio networks. IEEE J Sel Areas Commun 30(1):4–15
18. Gwon Y, Dastangoo S, Fossa C, Kung H (2013) Competing mobile network game: embracing anti-jamming and jamming strategies with reinforcement learning. In: Proceedings of the IEEE conference on communications and network security, pp 28–36
19. Lo B, Akyildiz I (2012) Multiagent jamming-resilient control channel game for cognitive radio ad hoc networks. In: Proceedings of the IEEE international conference on communications, pp 1821–1826
20. Dastangoo S, Fossa C, Gwon Y (2014) Competing cognitive tactical networks. Linc Lab J 20(2):16–35
21. Conley W, Miller A (2013) Cognitive jamming game for dynamically countering ad hoc cognitive radio networks. In: Proceedings of the IEEE military communications conference (MILCOM), pp 1176–1182
22. Busoniu L, Babuska R, Schutter B (2008) A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man Cybern Part C: Appl Rev 38(2):156–172