IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2021, accepted March 19, 2021, date of publication March 23, 2021, date of current version April 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068129

A Transfer Games Actor–Critic Learning Framework for Anti-Jamming in Multi-Channel Cognitive Radio Networks

HUYNH THANH THIEN[1], VAN-HIEP VU[2], AND INSOO KOO[1]
[1]Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, South Korea
[2]NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City 70000, Vietnam

Corresponding author: Insoo Koo (iskoo@ulsan.ac.kr)

This work was supported in part by the National Research Foundation of Korea through the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2021R1A2B5B01001721.

ABSTRACT A cognitive radio network (CRN) is a novel solution that promises to solve the spectrum scarcity problem and enhance spectrum utilization. However, unsecured CRN can easily be manipulated in order to attack legacy users on the communication channel. As a result, the network's performance significantly degrades. Therefore, communication channel security is an important issue that needs to be addressed in a CRN. In this work, we focus on improving the security of multi-channel communication in a CRN, while various jammers try to access channels of interest to prevent SUs from using them. By using game-theoretic concepts and by defining states, actions, and players' rewards, we propose game–based schemes that find the best channel for the secondary users (SUs) in order to avoid jammer's attacks on communication channels. Accordingly, the problem is finding the optimal channel to maximize the long-term reward of the SU where communication channels are not used by the primary users (PUs) and are not jammed by attackers. In addition, the idea of transfer learning might be applied to the problem under consideration, and thus, a transfer Game-Actor-Critic (TGACT) scheme is proposed, which uses the transferred knowledge in a double-game period to accelerate the learning process and provide performance improvement in channel selection. Finally, the performance of the proposed schemes is simulated with different configurations. The simulation results show that the proposed schemes are quite resistant to jammer attacks, and achieve better performance compared to other channel selection schemes.

INDEX TERMS Actor–critic, cognitive radio networks, game theory, jammer, reinforcement learning, transfer learning.

I. INTRODUCTION

Nowadays, the demand for communication and entertainment of users is increasing, leading to a significant increase in wireless applications and services. As a result, issues such as spectral scarcity and increasing demand for spectrum sources pose enormous challenges for network operators. To address existing issues, the cognitive radio network (CRN) was developed [1], [2] and is considered one of the most promising technologies for improving spectrum efficiency. The basic idea of a CRN is to exploit spectrum holes by enabling secondary users (SUs) (also called unlicensed users) to sense, select, and access free channels which are not occupied by the primary users (PUs) (also called licensed users). However, whenever a PU needs those channels, the SU has to vacate them. For the implementation of efficient spectrum exploitation in CRNs, selection by users of the appropriate access channel has a great influence on the performance of the network. Many channel selection schemes have been investigated in the previous works [3]–[8]. Although the proposed solutions can utilized the spectrum effectively, they are all based on the assumption that SUs exploit spectrum holes, while coordinating together to achieve their common target. This assumption ignores the scenario in which different random attackers could attack communication channels between SUs that could threaten network security and interfere with CRN. Physical or media access control layers are vulnerable to attacks that are a security threat to communication in CRN.

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Gaggero.

These threats are not only harmful to commercial networks but also threaten national defense and national security [9]. Hence, along with the challenges in spectrum management, secure spectrum utilizing also plays a crucial role for the development of CRN architecture. For that reason, a considerable amount of research on security techniques has been investigated for the CRNs [10], [11]. However, the influence of jammers on spectrum sharing has been still little considered. Some previous work proposed resource allocation and intelligent jamming to avoid security threats from jammers [12], [13]. In [14], the authors proposed an anti-jamming game in CRNs with multiple channels by modeling the interaction between a SU and attackers. Moreover, the anti-jamming game is redefined as the defense strategy with randomized power allocation. Most of the researchers only consider resource allocation and intelligent jamming to counter jammers attacks.

Recently, the Markov decision process (MDP) and the game theory approach in CRN have been investigated [15]–[17]. A stochastic game in [18] considered a competition and interaction among players, which is an extension of the MDP proposed in [19]. However, these proposed game approaches do not exploit the knowledge about the PU status on the channel, which can be collected via spectrum sensing on a pre-selected channel. In this work, we also solve the anti-jamming problem using the game theory approach. First, we propose a single-game scheme by formulating the problem of channel selection as a game framework, solving the problem for finding the best channel by using value iteration–based dynamic programming. In the paper, anti-jamming means that there is the absence of PU on the channels under consideration and further the jammers are not accessing these channels. Subsequently, jammers do not jam PU but only jam on the channels of interest. Besides, jammers could be considered as malicious SUs that try to access channels to prevent other normal SUs from using them. Malicious SUs can forge the spectral characteristics of the PU to gain priority access to wireless channels, known as primary user emulation (PUE) attacks. To match with the scenario in this paper, it is assumed that the CRN can easily detect PUE attacks based on several detection mechanisms such as channel parameters and spectrum decision, feature detection with filter and cyclostationary and many other detection mechanisms are mentioned [20]. Therefore, the jamming ability of attackers on PU is not considered in our work. Second, for the purpose of improving the performance of the single-game scheme through gathering the knowledge about the PU status on the channel, we can use the double-game scheme in our previous work [21]. In double-game scheme, the first game is solved to find pre-selected channel for the SU. Then, based on this channel, the SU performs spectrum sensing to collect PU status information on the channel. From the sensing results, the belief and state of the system are updated and used for the second game to find the final best channel. Note that, our best channel selection problem against jammer's attacks is different from the previous work [21]. In this paper, the problem of best channel selection against jammer's attacks is investigated in the scenario where communication channels are not used by the PUs and are not jammed by attackers. Meanwhile, the problem of best channel selection in [21] is based on maximize the secrecy rate of the SU.

Furthermore, the dynamic game solutions assume that the environment's dynamics (e.g., the jammer's strategies) is known in advance, which is rarely true due to the random attack nature of jammers. Since accurate information about the dynamics of the environment is sometimes not available, the problem of stochastic optimization is usually formulated as the MDP framework [22]. Afterward, the problem was formulated with MDP could be solved using reinforcement learning (RL) approaches [23]. In RL, the agent makes the optimal policy through environmental interactions and requires no prior knowledge of the dynamic of the environment [24]. Because of the advantages of a reinforcement learning approach, a series of studies have been carried out using the combination of anti-jamming and reinforcement learning techniques [25]–[29]. Wang *et al.* [25] proposed an anti-jamming defense mechanism in CRN based on a stochastic game framework in which SUs can decide how many channels are used for a given purpose based on observations of the jammer's attack strategy, channel quality, and the spectrum availability. To learn the optimal policy, the spectrum-efficient throughput is maximized using the minimax-Q learning. Singh and Trivedi in [26] have proposed the anti-jamming approach using the State-action-reward-state-action (SARSA) and QV RL algorithm in which the SU can learn the jammer's strategy and the characteristics of the channel. The results show an improvement in the performance of QV and SARSA algorithm when compared with the minimax–Q learning algorithm. In these studies, the Q–learning algorithm is used for most of the anti-jamming mechanisms due to the advantage of not knowing the model of this algorithm. However, with high-dimensional or continuous inputs, anti-jamming problems can face challenges when using traditional Q–learning algorithms. Therefore, several algorithms have been proposed to overcome this weakness such as the deep Q–network (DQN) [27] and double DQN [28], [29] which leverage a deep neural network to approximate the Q table. Specifically, Han *et al.* [27] proposed a two-dimensional anti-jamming mechanism for CRNs in which the SINR of the SU signals can be improved based on the exploitation of user mobility and spread spectrum. Besides, the anti-jamming scheme used a DQN-based approach to find the optimal policy of the network. The authors in [28] used the double DQN algorithm to counter the jammer in a multi-user manner with frequency hopping strategy attacks. Xu *et al.* [29] modeled the encounter between the jammer and the CRN based on a double DQN design to maximize the users' transmission rate. In this paper, for a performance comparison with our proposed schemes, we can also solve the problem of channel selection to avoid jammer's attacks using an RL approach, called an actor–critic (AC) algorithm. Specifically, based on the state of the system, the long-term network performance

is maximized to find the optimal channel policy that can be used against jammer's attacks.

In the case of RL, agents must get the information under a trial-and-error process to find an action in each state, because in the beginning they have no prior information on the environment [30]. Therefore, the procedure could take a considerable amount of time for learning in the AC algorithm to reach an optimal policy. To address this problem, we use transfer learning (TL) technique [31]. Regarding to transfer learning techniques, problems in target task can be effectively solved through the application of information obtained from source task [32]. Consequently, TL has attracted a lot of interest from researchers [31]–[36]. Additionally, several studies of anti-jamming by combining RL and TL have been investigated recently [37]–[39]. Chen *et al.* [37] proposed a RL-based power control scheme in which the WBAN coordinator and the in-body sensors can communicate with each other to defend against attacks. The Q-learning algorithm and the transfer learning method are used to obtain an optimal policy and accelerate the learning speed, respectively. Dai *et al.* [38] provided a safe version of deep RL for network security in which the risk level is estimated and the transfer learning technique is used to reduce initial random exploration. An anti-jamming scheme with the help of an unmanned aerial vehicle (UAV) in a cellular network is proposed in [39] where the deep RL algorithm is used to find the optimal relay policy. Furthermore, transfer learning is also used to help cellular networks battle jammers without knowing system models as well as observed communication states. In general, in the above-mentioned anti-jamming jobs, the RL algorithm is used quite commonly, however, these studies are either considered on basic wireless networks that are not CRNs or not considered combining with transfer learning technique. Therefore, the problem of anti-jamming by combining RL and TL in CRNs is considered in this paper. Furthermore, the transfer learning technique used to transfer knowledge from source task where the anti-jamming problem is solved based on game theory is also a highlight in this paper. By using the learned knowledge about channel selection from historical period (the double-game period), the ongoing learning process can be accelerated in the target task during the classic AC period, and provide additional improvements to the channel selection problem. As a result, the problem of channel selection with the help of transfer learning technology is proposed based on the transfer of knowledge learned from double-game scheme into a classic AC algorithm, which is denoted as the Transfer Game-Actor-Critic (TGACT) scheme in this work.

In summary, the main contributions of this paper are presented as follows:

- We investigate anti-jamming approaches for CRN with a multi-channel and multiple jammer, where an SU is transmitting data to a receiver SU while multiple jammers independently perform jamming on transmitter SU–receiver SU (SUtx–SUrx) transmissions. Each jammer attacks a random channel of interest. To optimize the security of a CRN, we propose an anti-jamming scheme by using game-theoretic concepts through definitions of states, actions, and players' rewards. The network scenario is modeled as a dynamic game, namely, a single-game scheme that finds the optimal channel for the SU in order to protect communication channels from jamming attacks. By using the optimal channel, the SU can receive maximum long-term reward which can reduce jammers' impact on channels. Then, we propose a double game–based anti-jamming scheme based on a repeat game algorithm in our previous work [21], which has demonstrated an improvement in performance compared to the single-game scheme. After that, the network performance with the proposed double-game scheme can be compared with a random-attack, single-game, and no-jammer schemes.
- Besides, the best channel selection problem with anti-jamming can be reformulated as an MDP framework. For a performance comparison with our proposed schemes, we consider the solution to the formulated MDP by using the classic AC algorithm, an RL approach where there is no need to know the jammers' access strategy in advance.
- Moreover, TL technology is also applied (namely, the TGACT scheme), which uses the transferred knowledge in the double-game period to accelerate the learning process and to provide performance improvements in channel selection, compared with a classic AC scheme and a transfer Actor-Critic (TACT) algorithm [35].
- To evaluate the performance of the proposed schemes, we use the average reward metric (also called the security level in this paper) of the SU in different configurations. The simulation results show that the proposed schemes are quite resistant to jammer attacks, and achieve better performance compared to other conventional channel selection schemes. Specifically, the double-game scheme provides better performance in comparison with random-attack and single-game schemes. Furthermore, the performance of the proposed TGACT scheme is also better than the dynamic game, classic AC, and TACT schemes.

The remaining of this paper is arranged as follows. Section II presents the system model and local spectrum sensing. Section III describes game formulation for channel selection with anti-jamming for single- and double-game schemes. The reinforcement learning approach–based anti-jamming schemes are described in Section IV, which introduce the classic AC, the TACT [35], and the proposed TGACT schemes. In Section V, we present the simulation results and discussions. Finally, Section VI provides a conclusion.

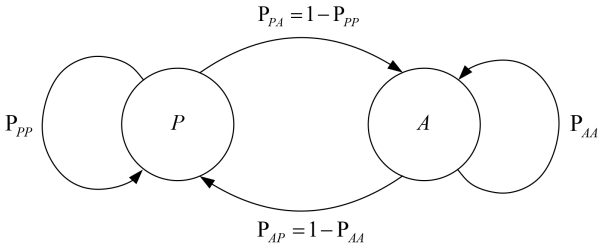

FIGURE 1. Markov chain for the PU states.

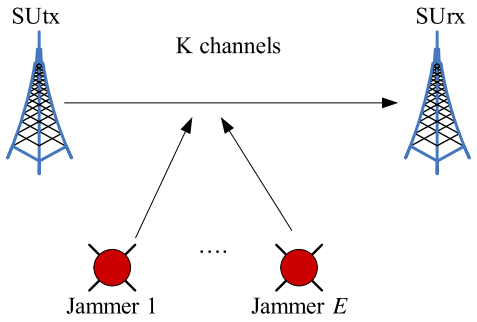

FIGURE 2. The system model.

II. THE SYSTEM MODEL AND LOCAL SPECTRUM SENSING

A. SYSTEM MODEL

Consider a CRN system where a transmitter SU tries to access the licensed channel of the PUs (K licensed channels) by using local spectrum sensing and send data to a receiver SU while jammers (E) independently perform jamming on a random channel of interest, as shown in Fig. 2. The channels are assumed to be an additive white Gaussian noise (AWGN) channel. Let $\mathcal{K} = \{1, \ldots, k, \ldots, K\}$ and $\mathcal{E} = \{1, \ldots, e, \ldots, E\}$ denote the set of channels and jammers, respectively. We assume that the SUtx always has data to transmit to the SUrx. For the convenience of tracking and formulation terminology, *SU* will be used instead of *SUtx* in the remainder of this paper. The operation of the system is in a time-slotted manner with slots of equal length and non-overlap, which are represented with the letter t. In this work, the operation of a PU on channel k is assumed to follow a two-state Markov discrete-time process. Let $S_{PU}(k) = \{A, P\}$ denote the PU state, in which the notations A and P represent the absence and presence of the PU, respectively. The operation of Markov chain states of the PU is shown in Fig. 1, in which P_{PA} and P_{AA} represent the state transition probabilities between the two absence and presence states of the PU. Let S^k denote the PU state on channel k, and we define $P^k_{AA} = \Pr\left(S^k_{(t+1)} = A | S^k_{(t)} = A\right)$ and $P^k_{PA} = \Pr\left(S^k_{(t+1)} = P | S^k_{(t)} = A\right)$ as the transition probability of the PU from state A to itself and from state P to state A, respectively.

First, the SU perform local spectrum sensing on a particular channel by using energy detector. Then, the SU will perform the data transmission process on this channel when it is free. On the contrary, the SU is not allowed to occupy the channel for data transmission, and will wait until the next time slot to repeat the whole process.

In time slot t, the SU selects a channel for its communication, $x \in \mathcal{K}$, and a denotes the action of the SU with $a = \{x | x \in \mathcal{K}\}$. Action a will be a distribution over set $\mathcal{K}$, which is given as:

$$\begin{aligned} &P_{SU}(k) = \Pr\{a = k\} \\ &\text{s.t.} \sum_k P_{SU}(k) = \sum_k \Pr\{a = k\} = 1, \quad k \in \mathcal{K}, \end{aligned} \tag{1}$$

where $P_{SU}(k)$ denotes the probability that the SU accesses channel k.

In the same way, the jammers select channels for jamming, $Y = \{y_1, y_2, \ldots, y_E\}, y_e \in \mathcal{K}$, and b_e denotes the action of jammer e where $b_e = \{y_e | y_e \in \mathcal{K}\}$. Then, actions by all jammers in the system are given as $b = \{b_1, b_2, \ldots, b_E\}$. Action b_e will be a distribution over set $\mathcal{K}$, which is given as:

$$\begin{aligned} &P_e(k) = \Pr\{b_e = k\} \\ &\text{s.t.} \sum_k P_e(k) = \sum_k \Pr\{b_e = k\} = 1, \quad k \in \mathcal{K};\ e \in \mathcal{E}, \end{aligned} \tag{2}$$

where $P_e(k)$ denotes the probability that jammer e attacks channel k.

Next, it is necessary to define the payoff function that characterizes the level of jamming. In this paper, based on the exploitation of spectrum holes in CRNs, the payoff function is determined based on the characteristics of channel access behavior (related to user and jammers) and occupancy status of spectrum holes (related to PU), not the jamming intensity. Therefore, the payoff function is determined according to whether a particular channel is not under jammer's attacks and is not occupied by PU. Specifically, when the SU accesses channel k that is not under jammer's attacks, and this channel is not occupied by PU, the payoff function of the SU is given as:

$$\begin{aligned} &R(a = k, b, S_{PU}(k)) \\ &\quad = \begin{cases} 1, & \text{if } S_{PU}(k) = A \text{ and } a \neq b_e, \forall e, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3}$$

B. LOCAL SPECTRUM SENSING

In the CRN considered, we assume the network includes a transmitter/receiver SU pair. The SU may use an energy detection method to perform local spectrum sensing. The binary hypothesis test of the SU is given as follows [40]:

$$\begin{cases} P: & x(t) = hu(t) + w(t), \\ A: & x(t) = w(t), \end{cases} \tag{4}$$

where $x(t)$ is the received signal by SU, h is the channel gain of the communication link between PU and SU, $u(t)$ is the signal transmitted by PU, and $w(t)$ is zero-mean AWGN.

The obtained energy at the SU [41]:

$$yE = \sum_{j=1}^{I} |x\,(j)|^2, \tag{5}$$

where I is the number of sensing samples during each detection interval, and $x\,(j)$ is the received PU signal at the SU in the j^{th} sample. When I is adequately large (e.g., $I > 10$ in practice), we approximate xE as a Gaussian random variable under the binary hypothesis (P and A) with mean μ_P, μ_A and variance σ_P^2, σ_A^2 as [41]:

$$yE \sim \begin{cases} P: & \mathcal{N}\left(\mu_P = I\,(1+\phi)\,, \sigma_P^2 = 2I\,(1+2\phi)\right), \\ A: & \mathcal{N}\left(\mu_A = I, \sigma_A^2 = 2I\right), \end{cases} \tag{6}$$

where ϕ is the sensed channel's signal-to-noise ratio (SNR) in decibels (dB). After that, two states of the PU can be made a decision as follows:

$$D\,(t) = \begin{cases} 1, & \text{when } yE\,(t) \geq \lambda, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

where 0 and 1, respectively, are binary bits that denote two states of the PU, absence and presence; and λ denotes a predefined threshold of decision energy.

III. GAME APPROACH–BASED ANTI-JAMMING SCHEME

We model the problem of channel selection for the interaction between the SU and the jammers as a game framework by using game-theoretic concepts through definitions of states, actions, and players' rewards [42], [43]. The game formulation of channel selection problem to avoid jammer's attacks is represented as follows.

- *Players:* the number of players joining the game is $(1 + E)$ players (i.e., an SU and E jammers).
- *State:* the system state is defined as:

$$\mathcal{S} = \{\mathrm{P}_0\,(k)\,, \mathrm{P}_e\,(k)\,|k \in \mathcal{K}; e \in \mathcal{E}\}\,, \tag{8}$$

  where $\mathrm{P}_0\,(k)$ is the probability (also called the belief) that the state of channel k is A (i.e., not used by the PU). If we consider the operation of the SU, the state will be approximated as:

$$\mathcal{S}_{ch} = \{\mathrm{P}_{ch}\,(k)\,|k \in \mathcal{K}\}\,, \tag{9}$$

  where $\mathrm{P}_{ch}\,(k)$ denotes the probability that the state of channel k ($\mathcal{S}_{ch}$) is free (i.e., not jammed and not being used by the PU), which is given as:

$$\mathrm{P}_{ch}\,(k) = \mathrm{P}_0\,(k)\left(1 - \prod_e \mathrm{P}_e\,(k)\right). \tag{10}$$

- *Action:* in each time slot t, the SU should select a channel for its communication, and a denotes the action of the SU with $a = \{x\,|x \in \mathcal{K}\}$.
- *Reward:* the reward for the SU, $\overline{R}\,(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch})$, is given by:

$$\begin{aligned} \overline{R}\,(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch}) &= E\,[R\,(a, b,\ S_{PU}\,(a))] \\ &= \mathrm{P}_{SU}\,(a)\,\mathrm{P}_{ch}\,(a)\,R\,(a, -a,\ A)\,, \end{aligned} \tag{11}$$

  where $E\,[R]$ denotes the expected value of the SU's payoff function R.

In this paper, the goal of choosing the best channel for SU is to maximize long-term reward (also called the accumulated reward of the SU) of the system, $\overline{aR}\left(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch}^0\right)$, which is defined as follows:

$$\overline{aR}\left(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch}^0\right) = \sum_{t=m}^{\infty} \gamma^t \overline{R}\left(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch}^t | \mathcal{S}_{ch}^m = \mathcal{S}_{ch}^0\right), \tag{12}$$

where m is the current time slot, t is the t^{th} time slot, $\mathcal{S}_{ch}^m$ is the system state in time slot m, γ is a discount constant ($\gamma \in (0, 1)$).

Then, the problem of choosing the optimal channel for SU to protect communication channels from jammers attacks is identified as follows:

$$a_{opt} = \arg\max_a \left(\overline{aR}\left(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch}^0\right)\right). \tag{13}$$

A. SINGLE GAME–BASED ANTI-JAMMING SCHEME

The problem in (13) can be solved by maximizing the accumulated reward of the SU. Through SU and jammers' channel access strategies, we can determine the state of the system. Therefore, the accumulated reward of the SU can be easily calculated based on its state action space. Besides, using a value iteration-based dynamic programming (DP) approach can obtain closed-form solutions for the value function [44]–[46]. Therefore, we consider our game model in which value iteration-based dynamic programming can be employed to come up with optimal strategies for the SU in order to find the optimal channel and protect communication channels from jamming attacks. The single game–based anti-jamming scheme is represented in Algorithm 1.

Algorithm 1 Single Game–Based Anti-Jamming Scheme

Input: $K, E, \mathrm{P}_0\,(k)\,, T, \mathrm{P}_{AA}, \mathrm{P}_{PA}, \gamma$

Output: the optimal channel for the SU, a_{opt}.

1: Given the system state:
   $\mathcal{S} = \{\mathrm{P}_0\,(k)\,, \mathrm{P}_e\,(k)\,|k \in \mathcal{K}; e \in \mathcal{E}\}$, as expressed in (8)
2: Determine the local decision for the state of the PU based on Section II-B, local spectrum sensing.
3: **for** $t = 1$ *to* T **do**
4: **for** $a = 1$ *to* K **do**
5: Calculate:
6: The payoff function: $R^t\,(a) \leftarrow (3)$
7: The probability: $\mathrm{P}_{ch}^t\,(a) \leftarrow (10)$
8: The reward for the SU: $\overline{R^t}\left(\mathrm{P}_{SU}^t\,(a)\,, \mathcal{S}_{ch}^t\right) \leftarrow (11)$
9: **end for**
10: Calculate the accumulated reward:
   $\overline{aR}\,(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch}) \leftarrow (12)$
11: Find the optimal channel, a_{opt}^t:
   $a_{opt}^t = \arg\max_a \left(\overline{aR}\,(\mathrm{P}_{SU}\,(a)\,, \mathcal{S}_{ch})\right)$
12: **end for**

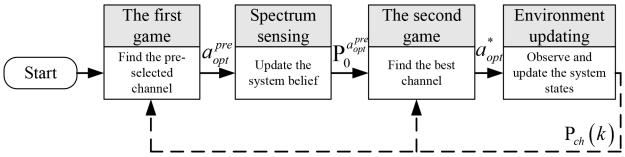

FIGURE 3. A block diagram of the proposed double-game scheme.

B. DOUBLE GAME–BASED ANTI-JAMMING SCHEME

In this section, we propose a double game–based anti-jamming scheme for a CRN. The basic idea of the double-game scheme is to exploit the knowledge about the PU status on the channel [21], which can be determined by solving the best channel selection problem from the single-game scheme. Specifically, the double-game scheme utilizes single-game scheme two times. One is for pre-selected channel, and the other one is for the best channel after spectrum sensing. That is, we first choose a most preferable channel, which is called as the pre-selected channel, without spectrum sensing using the Algorithm 1. After performing spectrum sensing on the pre-selected channel, we choose the best channel using the Algorithm 1, again. The double-game scheme is composed of 4 phases of the first game, spectrum sensing, the second game and environment updating, which are shown in Fig 3. Specifically, the four phases of the proposed double-game scheme are presented as follows.

- *The First Game Phase (Line 3 in Algorithm 2):* solving the problem in (13) based on initial state of the system to find the pre-selected channel, a_{opt}^{pre}.
- *Spectrum Sensing Phase (Lines 4-6):* based on pre-selected channel, spectrum sensing is carried out by SU to exploit the knowledge about the PU status on channel. Depending on the sensing result for the state of the PU signal is A or P (i.e., the channel is free or busy, respectively), the system belief will be updated accordingly using Bayes' rule [47]. Specifically, if the channel is free, the belief, $\mathrm{P}_0^{a_{opt}^{pre}}$, is updated as follows:

$$\mathrm{P}_0^{a_{opt}^{pre}} = \frac{\mathrm{P}_0^{a_{opt}^{pre}}\left(1-\mathrm{P}_f\left(a_{opt}^{pre}\right)\right)}{\mathrm{P}_0^{a_{opt}^{pre}}\left(1-\mathrm{P}_f\left(a_{opt}^{pre}\right)\right)+\left(1-\mathrm{P}_0^{a_{opt}^{pre}}\right)\left(1-\mathrm{P}_d\left(a_{opt}^{pre}\right)\right)}, \tag{14}$$

where P_d and P_f are the probabilities of correct detection and false alarm, respectively. Otherwise, the belief is updated as follows:

$$\mathrm{P}_0^{a_{opt}^{pre}} = \frac{\mathrm{P}_0^{a_{opt}^{pre}}\mathrm{P}_f\left(a_{opt}^{pre}\right)}{\mathrm{P}_0^{a_{opt}^{pre}}\mathrm{P}_f\left(a_{opt}^{pre}\right)+\left(1-\mathrm{P}_0^{a_{opt}^{pre}}\right)\mathrm{P}_d\left(a_{opt}^{pre}\right)}. \tag{15}$$

The state of the system is updated based on the estimated belief from either (14) or (15), which is denoted by $\mathcal{S}^u$. According to the updated state, $\mathcal{S}^u$, the system update the rewards and the accumulated rewards by (11) and (12), respectively.
- *The Second Game Phase (Line 7):* solving the problem in (13) based on the updated accumulated reward from spectrum sensing phase to find the final best channel, a_{opt}^{*}.
- *Environment Updating Phase (Line 8):* the optimal reward can be obtained by using optimal channel of the SU, a_{opt}^{*}. Based on the observation of channel status, we need to update the system state for use in the next time slot. Specifically, if communication with channel a_{opt}^{*} is successful (i.e., the channel is not occupied by PU), the belief $\mathrm{P}_0^{a_{opt}^{*}}$ is updated as:

$$\mathrm{P}_0^{a_{opt}^{*}} = \mathrm{P}_{AA}^{a_{opt}^{*}}. \tag{16}$$

Otherwise, if communication fails (i.e., the channel is occupied by the PU), the belief is updated as:

$$\mathrm{P}_0^{a_{opt}^{*}} = \mathrm{P}_{PA}^{a_{opt}^{*}}. \tag{17}$$

The system state in (9) will be updated using the updated belief, $\mathrm{P}_0^{a_{opt}^{*}}$, for use in the next time slot.

In short, the double game–based anti-jamming scheme is represented in Algorithm 2.

Algorithm 2 Double Game–Based Anti-Jamming Scheme

Input: $K, E, \mathrm{P}_0(k), T, \mathrm{P}_{AA}, \mathrm{P}_{PA}, \gamma$

Output: the optimal channel for the SU, a_{opt}^{*}.

1: Given the system state: $\mathcal{S} = \{\mathrm{P}_0(k), \mathrm{P}_e(k) | k \in \mathcal{K}; e \in \mathcal{E}\}$, as expressed in (8)
2: Determine the local decision for the state of the PU based on Section II-B, local spectrum sensing.
3: Find the optimal pre-selected action (the channel) of the game, a_{opt}^{pre}: $a_{opt}^{pre} = \arg\max_a \left(\overline{aR}\left(\mathrm{P}_{SU}(a), S_{ch}\right)\right)$, which can be solved with Algorithm 1.
4: Implement spectrum sensing and update the belief about the system: $\mathrm{P}_0^{a_{opt}^{pre}} \leftarrow$ (14) or (15)
5: According to updated belief $\mathrm{P}_0^{a_{opt}^{pre}}$, update the state of the system: $\mathcal{S}^u \leftarrow$ (8)
6: Update the accumulated reward, $\overline{aR_u}$, based on the updated state, $\mathcal{S}^u$: $\overline{aR_u} \leftarrow$ (12)
7: Solve the problem in (13) with updated accumulated reward $\overline{aR_u}$ to find optimal channel a_{opt}^{*} for the SU.
8: The optimal reward can be obtained by using optimal channel of the SU, a_{opt}^{*}. According to the observation of the communications link in the channel, update the system state for use in the next time slot by using (16) or (17).

IV. REINFORCEMENT LEARNING APPROACH–BASED ANTI-JAMMING SCHEMES

We reformulated the best channel selection problem in a multi-channel CRN system as the framework of an MDP.

Since the strategies of the jammers on communication channels are unknown, we employ the RL approach, which finds the optimal channel selection policy to reduce the jammer's influence and enhances the long-run network performance. In a model-free RL framework, RL agents can learn the optimal policy through trial-and-error learning during their interaction with the environment.

A. MARKOV DECISION PROCESS

A basic RL model is composed of two factors, environment and agent, in which these two elements interact over time. Furthermore, based on an environment states, the agent does a process of trial-and-error learning, and then the agent can make a suitable action and maximize the accumulated rewards. Regarding the MDP framework, we need to consider objects like state space ($\mathcal{S}_{ch}$), action space ($\mathcal{A}$), the state-transition probability function ($\mathcal{P}$), and the reward function (R). Therefore, the MDP framework of the channel selection problem for anti-jamming can be defined as a tuple $\langle \mathcal{S}_{ch}, \mathcal{A}, \mathcal{P}, R \rangle$.

- *States:* for the operation of the SU, the state is defined as $\mathcal{S}_{ch} = \{\mathrm{P}_{ch}(k) | k \in \mathcal{K}\}$, where $\mathrm{P}_{ch}(k)$ as defined in (10).
- *Actions:* at the time slot t, the agent observes state $\mathcal{S}_{ch}^{t}$ in state space $\mathcal{S}_{ch}$ of the environment, and then chooses action a^t in action space $\mathcal{A}$ following a probability of taking action, π. In this work, the SU (the network agent) chooses the best channel that it can access (i.e., the channel which is not occupied by the PU and not being jammed). Therefore, action a^t is set as $a^t = \{x\}_{x \in \mathcal{K}}$, which is defined as explained in Section II-A.
- *Rewards:* then, the environment will return a reward to the agent, $R\left(\mathcal{S}_{ch}^{t}, a^t\right)$. The reward of the network can be defined as in (11), and transforms to the new state $\mathcal{S}_{ch}^{t+1}$. The next state, $\mathcal{S}_{ch}^{t+1}$, is updated following (9), which is based on the action (channel k).
- *The State-Transition Probability Function:* once the SU selects an action, the system changes from the current state, $\mathcal{S}_{ch}^{t}$, to the new state, $\mathcal{S}_{ch}{}'$, based on the probability of state-transition as follows:

$$\mathcal{P}\left(\mathcal{S}_{ch}{}' | \mathcal{S}_{ch}^{t}, a^t\right) = \begin{cases} 1, & \text{if } \mathcal{S}_{ch}{}' = \mathcal{S}_{ch}^{t+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The purpose of the RL approach is to learn to select actions based on the states of the system through learning from experience to maximize the accumulated reward (also called the state-value function) of the system. The state-value function is expressed as follows [24]:

$$\begin{aligned} V(\mathcal{S}_{ch}) &= \sum_{t=0}^{\infty} \gamma^t R\left(\mathcal{S}_{ch}^{t}, \pi\left(\mathcal{S}_{ch}^{t}\right) \middle| \mathcal{S}_{ch}^{0} = \mathcal{S}_{ch}\right) \\ &= R(\mathcal{S}_{ch}, \pi(\mathcal{S}_{ch})) \\ &\quad + \gamma \sum_{\mathcal{S}_{ch}{}' \in \mathcal{S}} \mathcal{P}\left(\mathcal{S}_{ch}{}' | \mathcal{S}_{ch}, \pi(\mathcal{S}_{ch})\right) V\left(\mathcal{S}_{ch}{}'\right), \end{aligned} \tag{19}$$

where $\pi(\mathcal{S}_{ch}) : \mathcal{S}_{ch} \mapsto a$ denotes the stochastic policy which SU can take an action, a, based on the state of the environment, $\mathcal{S}_{ch}$, and $\mathcal{P}\left(\mathcal{S}_{ch}{}' | \mathcal{S}_{ch}, \pi(\mathcal{S}_{ch})\right)$ denotes the state-transition probability from the current state $\mathcal{S}_{ch}$ to the next state $\mathcal{S}_{ch}{}'$. The Bellman equation is used to maximize the state-value function, and find the optimal policy, π^*, which is given as follows [24]:

$$\begin{aligned} &\pi^*(\mathcal{S}_{ch}) \\ &= \arg\max_{a} \left(R(\mathcal{S}_{ch}, a) + \gamma \sum_{\mathcal{S}_{ch}{}' \in \mathcal{S}} \mathcal{P}\left(\mathcal{S}_{ch}{}' | \mathcal{S}_{ch}, a\right) V^*\left(\mathcal{S}_{ch}{}'\right) \right). \end{aligned} \tag{20}$$

Through determining the optimal policy, we can find the optimal channel for SU which can avoid jamming from attackers in a multi-channel CRN.

B. THE AC–BASED CHANNEL SELECTION SCHEME

Traditionally, the MDP problem can be solved with a value iteration-based dynamic programming approach. However, this approach needs to know the dynamic environment in advance. In addition, the agent will face more challenges in the process of finding the optimal policy when using dynamic programming approach to solve the Bellman equation in a high-dimensional space of state and action. Therefore, for a performance comparison with our proposed schemes, we also consider using an RL approach, called the classic AC algorithm which requires no prior knowledge of the environment's dynamics. Regarding this approach, the agent can learn the optimal policy through trial-and-error learning during their interaction with the environment. Basically, an agent for the AC algorithm consists of two separate components [23]: the actor, which observes the environment state and selects an action by stochastic policy π; and the critic, which evaluate an actor's action based on the value function and reward, as shown in Fig. 4.

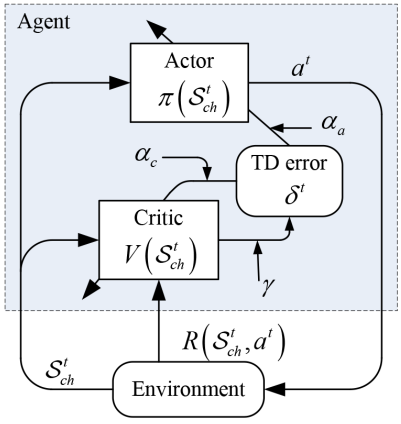

FIGURE 4. A block diagram of the classic actor–critic algorithm.

When an SU and jammers connect to the network, the initial state of the system is $\mathcal{S}$. In order to optimize performance

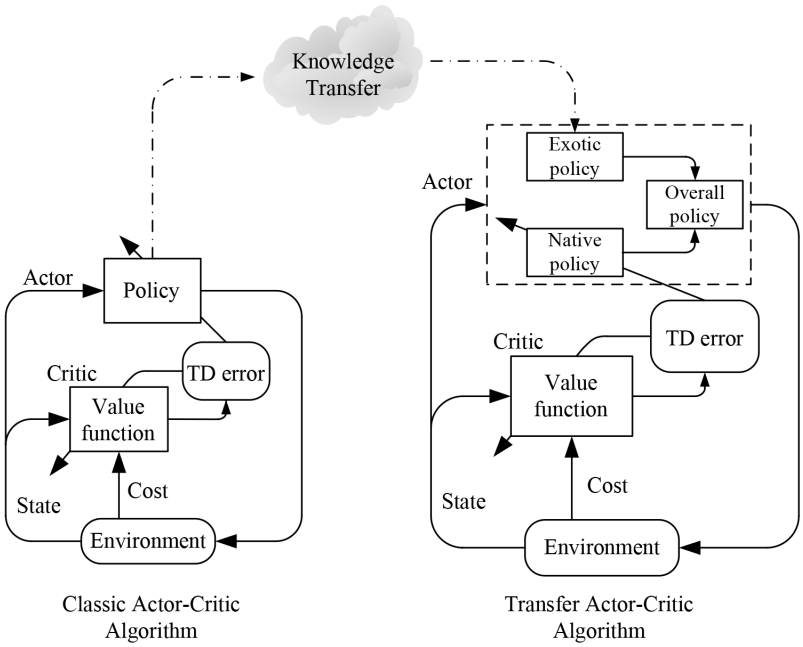

FIGURE 5. A block diagram of the TACT scheme [35].

and maximize the accumulated reward, the SU chooses suitable actions in which the selected channels are not used by the PU and are not being attacked by jammers. The learning process of the AC algorithm to find the optimal channel selection policy is presented as follows. In the time slot t, the SU selects an action, a^t, following a policy, $\pi^t\left(S_{ch}^t\right)$. The probability of taking action a^t in state S_{ch}^t is given as follows [24]:

$$\pi^t\left(\mathcal{S}_{ch}^t, a^t\right) = \Pr\left(a^t \middle| \mathcal{S}_{ch}^t\right) = \frac{e^{h^t\left(\mathcal{S}_{ch}^t, a^t\right)}}{\sum_{a'} e^{h^t\left(\mathcal{S}_{ch}^t, a'\right)}}, \tag{21}$$

where $h^t\left(\mathcal{S}_{ch}^t, a^t\right)$ is the tendency to select action a^t in state $\mathcal{S}_{ch}^t$. Once the SU selects action a^t, the current state, $\mathcal{S}_{ch}^t$, will transit to the next state, $\mathcal{S}_{ch}^{t+1}$, according to the state-transition probability, which is given in (18), and returns an immediate reward, $R\left(\mathcal{S}_{ch}^t, a^t\right)$. Afterward, based on the calculation of the temporal difference (TD) error value, the critic will evaluate the selected action from the actor. The TD error value is calculated from the value of $R\left(\mathcal{S}_{ch}^t, a^t\right) + \gamma V^t\left(\mathcal{S}_{ch}^{t+1}\right)$ at the critic and the state-value function in the previous state, $V^t\left(\mathcal{S}_{ch}^t\right)$, which is given as follows:

$$\delta^t = R\left(\mathcal{S}_{ch}^t, a^t\right) + \gamma V^t\left(\mathcal{S}_{ch}^{t+1}\right) - V^t\left(\mathcal{S}_{ch}^t\right). \tag{22}$$

Thereafter, based on the TD error, the critic will update its state-value function in the next time slot to improve the state-value function and policy. The state-value function is updated as follows:

$$V^{t+1}\left(\mathcal{S}_{ch}^t\right) = V^t\left(\mathcal{S}_{ch}^t\right) + \alpha_c \delta^t, \tag{23}$$

where α_c denotes the step-size parameter of the critic. Besides, the policy at the actor is also updated as follows:

$$h^{t+1}\left(\mathcal{S}_{ch}^t, a^t\right) = h^t\left(\mathcal{S}_{ch}^t, a^t\right) + \alpha_a \delta^t, \tag{24}$$

where α_a denotes the step-size parameter of the actor. Overall, the system performance can be improved by updating functions of the state-value and policy based on the TD error with appropriate step-size parameters by the actor and critic.

C. THE TACT–BASED CHANNEL SELECTION SCHEME

The previous section addresses the problem of finding the best anti-jamming channel using the classic AC algorithm. In this section, we present a methodology where the controller utilizes information on the strategies learned during the historical period to find the best anti-jamming channel. First, *state, action, reward* and *value function* definitions are also defined as described in Section IV-A. For a performance comparison with our proposed schemes, a TACT algorithm in [35] can be applied to our channel selection problem. The block diagram of the TACT scheme is shown in Fig. 5. For a TACT-based approach, the information on the policy, $h\left(\mathcal{S}_{ch}, a\right)$, from a source task (left side in Fig. 5) is transferred to a target task (right side in Fig. 5). However, there might be some differences although the target task and the source task have similarities. For example, the source task has a higher reward than the target task even though these two tasks use the same state. Therefore, action a can be taken by the controller in the target task in an aggressive direction for channel selection. As a result, the transferred policy can have a negative effect on the action selection process. Hence, by reducing the effects of the transferred policy, we can

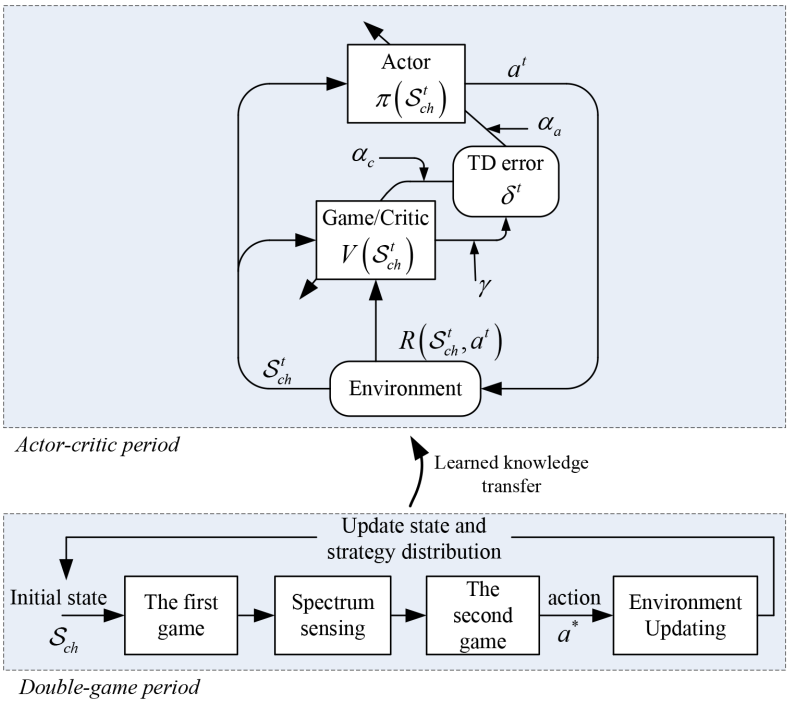

FIGURE 6. A block diagram of the transfer Game–AC scheme.

mitigate these negative effects. In general, the basic idea of the TACT algorithm is to avoid the negative effect of the transferred policy on the action selection process [35]. From this idea, we can update the overall policy with transfer rate where the transfer rate should be decreased over time to reduce the impact of transferred policy on the overall policy. As can be seen in Fig. 5, the overall policy, h_o, is a combination of an exotic policy (also called transferred policy), h_e, and a native policy, h_n. The overall policy is updated as follows [35], [36]:

$$h_o^{t+1}\left(\mathcal{S}_{ch}^{t}, a^t\right) = (1-\zeta)\, h_n^{t+1}\left(\mathcal{S}_{ch}^{t}, a^t\right) + \zeta h_e\left(\mathcal{S}_{ch}^{t}, a^t\right), \tag{25}$$

where $\zeta \in (0, 1)$ denotes the transfer rate which represents the exotic policy contribution to the overall policy. During the initial training process, the overall policy update strategy with the dominance of the exotic policy over the native policy, so the performance of the system can be improved. However, the goal is still learning at the target task, so we need to reduce the impact of transferred policy on the overall policy. Therefore, the transfer rate should be decreased over time with decay factor d_ζ, and thus, $\zeta \mapsto 0$ as the number of iterations reaches infinity. Besides that, the native policy updates itself according to the classic AC algorithm as defined in (24).

D. THE PROPOSED TGACT–BASED CHANNEL SELECTION SCHEME

Transfer learning method in the TGACT–based channel selection scheme consists of two phases: i) transferring information from using the optimal channel, which can be obtained from the double-game period; and ii) training the target task based on the updated state and strategy distribution. More specifically, in the first phase of transfer learning, the Algorithm 2 is exploited to get the knowledge about the PU status on the channel to find the optimal channel. The communication link status on this channel can then be determined to be either occupied or not occupied by the PU. As a result, the PU state, P_0, should be updated according to (16) and (17) and will be transferred to the second phase of transfer learning. In the second phase, the learning process is implemented by using the classic AC algorithm as described in Section IV-A and Section IV-B with the updated state, P_0, from the first phase. The TGACT block diagram is shown in Fig. 6, and the proposed TGACT-based anti-jamming scheme is presented in the Algorithm 3 in which the first phase of transfer learning is from line 1 to line 4 and the learning process of the classic AC algorithm is from line 5 to line 14.

V. SIMULATION RESULTS AND DISCUSSION

In this section, we show simulation results to demonstrate the efficiency of the proposed schemes for anti-jamming in multi-channel CRNs. We also compare the performance of our proposed schemes, which include single-game, double-game, and TGACT schemes, against the performance of other baseline schemes, such as the classic AC scheme, the TACT scheme [35], a random-attack scheme, and a no-jammers scheme. We sometimes use terms like learning and non-learning. The learning schemes include classic AC, TACT, TGACT, and double-game schemes. The double-game

Algorithm 3 The Proposed TGACT–Based Anti-Jamming Scheme

Input: $K, E, \mathrm{P}_0(k), T, \mathrm{P}_{AA}, \mathrm{P}_{PA}, \gamma, \alpha_c, \alpha_a, \zeta, d_\zeta$
Output: the optimal channel selection policy, π^*_{opt}.

1: Determine the local decision for the state of the PU based on Section II-B, local spectrum sensing.
2: Find the optimal channel for the SU, a^*_{opt}, which can be determined with Algorithm 2.
3: Based on the optimal access channel, update the state of the system as seen in (16) or (17).
4: Determine the system state in (8) based on the updated state.
5: Initialize the lookup table for policy $\pi(\mathcal{S}_{ch}, a)$, tendency $h(\mathcal{S}_{ch}, a)$, and state-value function $V(\mathcal{S}_{ch})$.
6: **for** $t = 1$ *to* T **do**
7: Select action a^t based on temporal policy $\pi^t\left(\mathcal{S}^t_{ch}, a^t\right)$
8: Calculate immediate reward: $R\left(\mathcal{S}^t_{ch}, a^t\right) \leftarrow (11)$
9: Update the state of the system from $\mathcal{S}^t_{ch}$ to $\mathcal{S}^{t+1}_{ch}$, and calculate TD error: $\delta^t \leftarrow (22)$
10: Update the state-value function: $V\left(\mathcal{S}^t_{ch}\right) \leftarrow (23)$
11: Update the tendency to select an action: $h^{t+1}\left(\mathcal{S}^t_{ch}, a^t\right) \leftarrow (24)$
12: Update the policy: $\pi\left(\mathcal{S}^t_{ch}, a^t\right) \leftarrow (21)$
13: **end for**
14: Return the optimal policy: $\pi^*(\mathcal{S}_{ch}) = \arg\max_{a\in\mathcal{A}} \{\pi(\mathcal{S}_{ch}, a)\}$

scheme is seen as an improved method of the single-game scheme; it exploits the action in the current time slot to update the system belief when using in the next time slot. Therefore, we consider the double-game scheme as one of the learning schemes. The non-learning schemes include the single-game, the no-jammers, and the random-attack schemes. With a single-game scheme, in the current time slot, the player selects an action and tries to maximize the accumulated reward of the system. When using the no-jammers scheme, there are no jammer attacks on the channels. With the random-attack scheme, jammers randomly attack the channels, and there are no anti-jamming efforts applied to the system.

A. SIMULATION SETTINGS

In this paper, we compare the performance of the proposed schemes under various configurations. First, we show the convergence property for each of the proposed schemes in terms of the average rewards metric, *ave_R*, which are calculated as follows:

$$ave\_R = \frac{1}{T}\sum_{t=1}^{T} Re(t), \tag{26}$$

where T is the number of time slots, $Re(t)$ is the reward for the SU in the t^{th} time slot which is calculated by (11) without effect of the channel selection probability of SU. Then, we validate the network performance in terms of average rewards under three conditions: varying the number of channels, varying the number of jammers, and varying SNR value of the sensed channel. In the first scenario, simulations are performed when the number of jammers was fixed at $E = 5$ while the number of channels changed from three to 11. In the second scenario, we consider the performance of the proposed schemes when the number of channels is $K = 5$ while the number of jammers changes from one to five. The SNR of the sensed channel is $\phi = -6$ dB in both first and second scenarios. In the last one, the SNR of the sensed channel changes from -18 dB to -2 dB, while the number of jammers and the number of channels are each fixed at 5. In all cases, we assume that the initial values of the state transition probabilities are $\mathrm{P}_{AA} = 0.8$ and $\mathrm{P}_{PA} = 0.2$. The value of discount factor, $\gamma = 0.99$. To provide the best performance from the proposed schemes, the simulations are performed several times to achieve the most suitable step-size parameters (α_a and α_c). Then, we set $\alpha_a = 0.1$ and $\alpha_c = 0.1$. As seen in previous work [35], [36], higher transfer rates resulted in a faster convergence rate and better performance. Therefore, the transfer rate is set to $\zeta = 0.9$ with a decay rate of $d_\zeta = 0.99$. Simulations are performed with $T = 2,000$ timeslots. All of the schemes are implemented using Matlab.

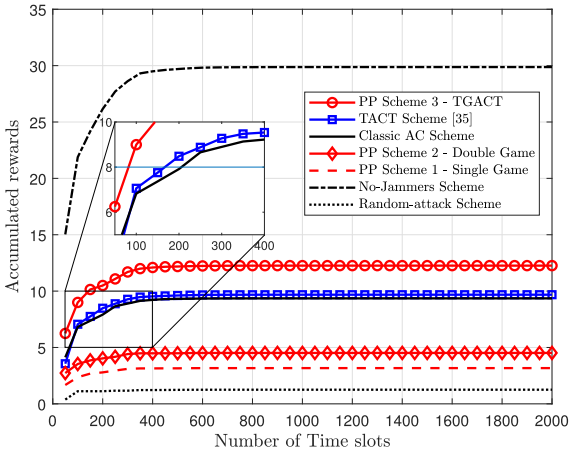

FIGURE 7. Accumulated rewards with five channels ($K = 5$) and five jammers ($E = 5$) when the SNR of the sensed channel is -6 dB ($\phi = -6$ dB).

B. CONVERGENCE PROPERTY

In this section, we check the convergence property in terms of the accumulated rewards from our proposed schemes when the number of time slots, T, increases gradually from 1 to 2,000. The number of channels and jammers are fixed at $K = E = 5$, while the SNR of the sensed channel is set to $\phi = -6$ dB. Fig. 7 shows the improvement of the accumulated reward as the number of time slots increases. We observe that the accumulated rewards from the schemes increase rapidly over the first 400 time slots, and reach optimal value with more time slots. The convergence speed in the game schemes is faster than the random-attack scheme

with no anti-jamming. The convergence speed of the double-game scheme is faster than the single-game scheme owing to information exploitation about the PU status on the channel that is collected via sensing on the optimal channel obtained from the single-game scheme. However, with the single-game scheme, the selected action is not affect the future reward, the player only try to choose the action in the current time slot which maximize the accumulated reward of the system. Therefore, the convergence speed in dynamic programming schemes like the random and single-game schemes is much lower than in reinforcement learning schemes like the AC scheme and the transfer learning schemes (TACT and our proposed TGACT). The convergence speed of transfer learning–based reinforcement learning schemes is faster than the AC scheme owing to the advantage of transfer learning, which transfers information from the source task to the target task. Fig. 7 also shows that schemes using transfer learning (TGACT and TACT) can accelerate learning process of conventional RL algorithm, AC scheme. Specifically, to reach the value of the accumulated rewards of 8, the TGACT and TACT schemes need 70 and 160 time slots, respectively. Meanwhile, to get this value of the accumulated rewards, the classic AC scheme needs about 210 time slots. From this result, we can see transfer learning can accelerate reinforcement learning process.

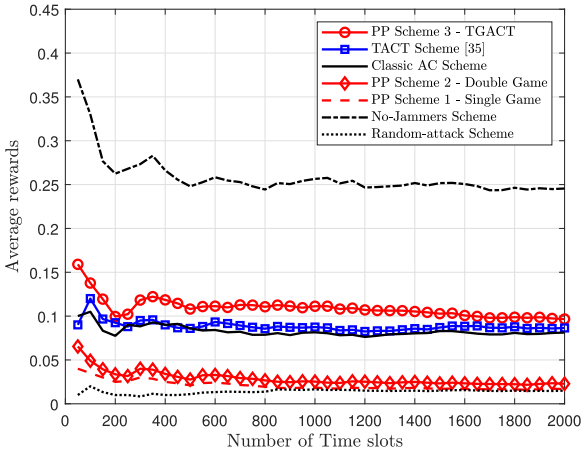

FIGURE 8. Average rewards with five channels ($K = 5$) and five jammers ($E = 5$) when the SNR of the sensed channel is -6 dB ($\phi = -6$ dB).

Afterward, we verify the convergence property of the proposed schemes in terms of the average reward. As seen in Fig. 8, the convergence rate of the schemes significantly decreases over the first 200 time slots, then, the average reward continues to decrease but at a slower rate. Finally, the schemes reach to an optimal reward to use for channel selection after about 1,000 time slots. The reward for the SU that used the classic AC scheme is lower than for SUs using the transfer learning schemes. This is because the agent of the transfer learning schemes can learn faster by exploiting transferred knowledge from source task. In addition, the agent in classic AC scheme needs to be trained from scratch, and therefore, it needs more trials-and-errors to learn. The convergence rate of the TGACT scheme outperforms in most learning schemes. The random-attack scheme provides the lowest convergence speed, and thus, got the smallest rewards. For most of the schemes, the performance is the best in a favorable environment with no jammer attacks on the system. In the convergence process, if the agent uses too many time slots for training, a local optimal policy might be obtained. However, the training process might take a very long time. Therefore, the total number of time slots for training should be neither too large nor too small.

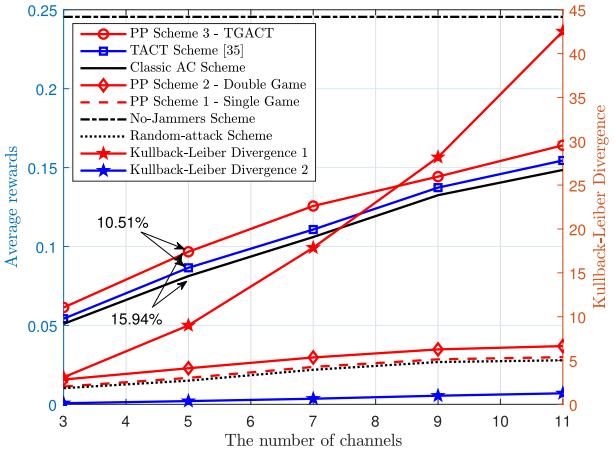

FIGURE 9. The left Y-axis shows average rewards according to the number of channels when the number of jammers is $E = 5$ and the SNR value of the sensed channel is $\phi = -6$ dB. The right Y-axis represents the Kullback-Leibler (KL) divergence in which KL divergence 1 represents the KL divergence of TGACT scheme over Double-game scheme and KL divergence 2 represents the KL divergence of TACT scheme over classic AC scheme.

C. THE PERFORMANCE OF THE SYSTEM ACCORDING TO THE NUMBER OF CHANNELS, THE NUMBER OF JAMMERS, AND THE SNR OF THE SENSED CHANNEL

In Fig. 9, we observe the performance of the proposed schemes under the influence of the number of channels. In this case, the number of channels is set at $K \in \{3, 5, 7, 9, 11\}$ while the number of jammers and the SNR of the sensed channel are fixed at $E = 5$ and $\phi = -6$, respectively. As seen in Fig. 9, the average reward increases as the number of channels increases. In fact, the more channels are used, the weaker the ability to attack a particular channel, and thus obtain higher system rewards. The average reward from the single-game scheme dominated the random-attack scheme. To explain this, with the single-game scheme, the SU maximizes the accumulated reward based on the action selection at the current time slot, whereas there are no anti-jamming solutions used in the system with the random-attack scheme. The double-game scheme is better than the single-game scheme owing to exploitation of PU status information, which can be collected via sensing based on the optimal channel from the single-game scheme. Moreover, the average reward of the proposed TGACT scheme outperforms

the classic AC and TACT schemes. In particular, when the number of channels is five, the average reward of the TGACT scheme provides improvements of 10.51 % and 15.94 % over TACT and classic AC schemes, respectively. This is because, in the proposed TGACT scheme, agent can exploit information transferred from double-game period, and thus, learn effectively the optimal policy. Therefore, the TGACT scheme provides the best performance in comparison with the remaining schemes, except for the no-jammers scheme. With the no-jammers scheme, system performance is the best compared to most other schemes. However, in this scheme, jammers are not allowed to attack the channels. Moreover, although the number of channels changes, the local decision for the state of the PU is specified only once for the initial parameter. Hence, the system reward from this scheme remains unchanged. The average reward is lowest in case of the random-attack scheme. This is because the SU does not use channel selection schemes with anti-jamming and channels can be randomly attacked by jammers.

Fig. 9 also shows the Kullback-Leibler (KL) divergence [48] in which KL divergence 1 represents the KL divergence of TGACT scheme over Double-game scheme and KL divergence 2 represents the KL divergence of TACT scheme over classic AC scheme. Comparing KL divergence 1 and KL divergence 2, the implementation of a transfer learning from a double-game scheme to classic AC scheme (i.e., TGACT scheme) provides a significant improvement in performance over performing a transfer learning from an AC scheme to AC scheme (i.e., TACT). Furthermore, the properties of KL divergence show that the smaller the DL divergence value, the more similar the two distributions are. Therefore, with a small improvement in average rewards, the KL divergence 2 achieves a relatively low gain when the TACT scheme and the classic AC scheme are compared. Likewise, KL divergence 1 with a significantly large value can be explained. As a result, the higher KL divergence between the target task and the source task, the more efficient it is in performing transfer learning. The simulation results also show that the average rewards improve with increasing the number of channels, the value of KL divergence also increases.

In the same way, we evaluate the efficiency of proposed schemes under varying numbers of jammers, E, and compare the results with the classic AC, the TACT, the random-attack, and the no-jammers schemes, as shown in Fig. 10. While the number of jammers ranges from one to five, the number of channels and the SNR of the sensed channel are fixed at $K = 5$ and $\phi = -6$ dB, respectively. When the number of jammers is high, the channel becomes more vulnerable to attack due to the large number of jammers. Therefore, with an increase in the number of jammers, the average reward in the system decreases significantly. In addition, the performance of proposed schemes is dominant than the conventional channel selection schemes because the channel can be selected effectively by maximizing the system reward in the current time slot, as in the game schemes, or by learning the variations in the environment and transferring information

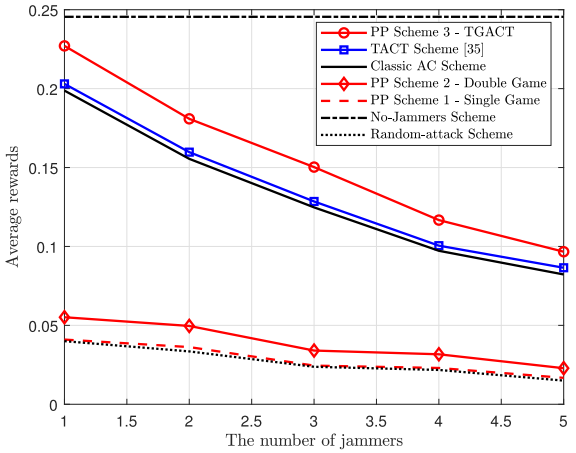

FIGURE 10. Average rewards according to the number of jammers when the number of channels is $K = 5$ and the SNR of the sensed channel is $\phi = -6$ dB.

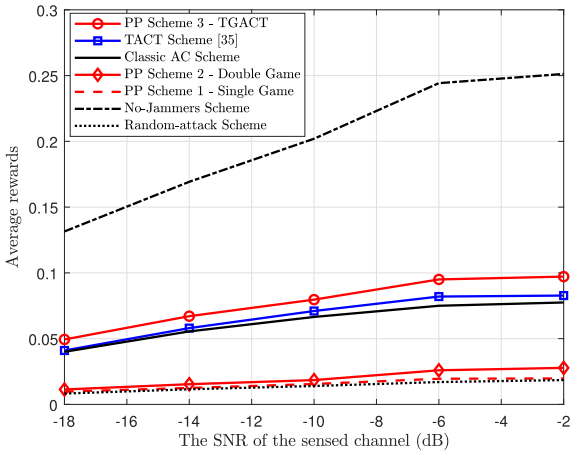

FIGURE 11. Average rewards according to the SNR of the sensed channel when the number of channels and jammers are $K = 5$ and $E = 5$, respectively.

from double-game period, as done in the TGACT scheme. Again, the system performance is also remained unchanged in no-jammers scheme because jammers are not allowed to attack the channels in this scheme.

We further inspect the impact of the SNR of the sensed channel on the security level of the channel selection schemes, which is shown in Fig. 11. To verify this, we evaluate the results based on the following SNR values (in decibels), $\phi \in \{-18, -14, -10, -6, -2\}$, while keeping the number of channels and jammers at $K = 5$ and $E = 5$, respectively. As observed in Fig. 11, the achieved average reward increases with an increase in the SNR of the sensed channel, which enables SU to effectively spectrum sensing and local decision-making. Obviously, a better SNR provides better detection accuracy. The result is that the larger SNR of the sensed channel may provide a better overall performance. Again, the TGACT scheme provides the highest average reward, whereas the random-attack scheme shows the lowest average reward. This is because the TGACT scheme is able

to choose the effective channel in each time slot by the combination of estimating the future reward and the exploitation of the transferred information from double-game period. Meanwhile, the random-attack scheme does not use a channel selection scheme against jammer's attacks to enhance the security level. Consequently, we verify that the TGACT scheme can provide effective communication channels in terms of security level.

VI. CONCLUSION

In this work, we proposed anti-jamming approaches for a CRN in which the SU works multi-channel communications, and various numbers of jammers randomly attack. We first designed a single game–based anti-jamming scheme that solves the problem of maximizing the accumulated reward for the SU in order to find the optimal channel. Then, a double game–based anti-jamming scheme is considered, in which the pre-selected channel is determined by using a single-game scheme. Afterward, through the pre-selected channel, the SU performs spectrum sensing to collect the PU status information, then, the second game will be solved using the updated accumulated reward. In addition, we adopted the transfer learning technique into the double-game scheme to accelerate the learning speed and improve network performance by exploiting the information learned in the double-game period. The simulation results show the efficiency of the proposed solutions in improving the long-term performance of the network. Through the proposed schemes, the optimal channel will be provided for SU to avoid jamming from attackers and significantly improve the security level of the CRN. The channel selection problem with anti-jamming can be extended with multiple SUs in the future work. Consequently, the model and learning parameters would need to be modified. Even the state and action spaces will be larger, and thus, the problem becomes more complicated. For this reason, combining both transfer learning technology and a deep RL approach could be considered, in which deep neural network can be used as an approximation function for mapping the system input (e.g., the system state) and the output in the RL task (e.g., the optimal policy).

REFERENCES

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, Aug. 1999, doi: 10.1109/98.788210.

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Comput. Netw.*, vol. 50, no. 13, pp. 2127–2159, Sep. 2006, doi: 10.1016/j.comnet.2006.05.001.

[3] B. Mumey, J. Tang, I. R. Judson, and D. Stevens, "On routing and channel selection in cognitive radio mesh networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 9, pp. 4118–4128, Nov. 2012, doi: 10.1109/TVT.2012.2213310.

[4] A. Jamal, C.-K. Tham, and W.-C. Wong, "Dynamic packet size optimization and channel selection for cognitive radio sensor networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 1, no. 4, pp. 394–405, Dec. 2015, doi: 10.1109/TCCN.2016.2531082.

[5] M. Ju and K.-M. Kang, "Cognitive radio networks with secondary network selection," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 966–972, Feb. 2016, doi: 10.1109/TVT.2015.2400433.

[6] M. Xu, M. Jin, Q. Guo, and Y. Li, "Multichannel selection for cognitive radio networks with RF energy harvesting," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 178–181, Apr. 2018, doi: 10.1109/LWC.2017.2763591.

[7] A. Sabbah, M. Ibnkahla, O. Issa, and B. Doray, "Control channel selection techniques in cognitive radio networks: A comparative performance analysis," *J. Commun. Netw.*, vol. 20, no. 1, pp. 57–68, Feb. 2018, doi: 10.1109/JCN.2018.000006.

[8] J. Ren, H. Zhang, X. Liu, and Y. Qin, "Energy efficiency-centric channel selecting in energy harvesting cognitive radio sensor network," in *Proc. IEEE 4th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Chengdu, China, Dec. 2019, pp. 2736–2739, doi: 10.1109/IAEAC47372.2019.8997673.

[9] Arbor Networks, Burlington, Massachusetts, United States. (Jan. 24, 2017). *12th Annual Worldwide Infrastructure Security Report*. [Online]. Available: https://www.netscout.com/news/press-release/worldwide-infrastructure-security-report

[10] J. Li, Z. Feng, Z. Feng, and P. Zhang, "A survey of security issues in cognitive radio networks," *China Commun.*, vol. 12, no. 3, pp. 132–150, Mar. 2015, doi: 10.1109/CC.2015.7084371.

[11] A. G. Fragkiadakis, E. Z. Tragos, and I. G. Askoxylakis, "A survey on security threats and detection techniques in cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 428–445, 1st Quart., 2013, doi: 10.1109/SURV.2011.122211.00162.

[12] Z. Bai, L. Ma, Y. Dong, P. Ma, and Y. Ma, "Energy-efficient resource allocation for secure cognitive radio network with delay QoS guarantee," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2795–2805, Sep. 2019, doi: 10.1109/JSYST.2018.2875835.

[13] D. T. Hoang, D. Niyato, P. Wang, and D. I. Kim, "Performance analysis of wireless energy harvesting cognitive radio networks under smart jamming attacks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 1, no. 2, pp. 200–216, Jun. 2015, doi: 10.1109/TCCN.2015.2488620.

[14] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 4–15, Jan. 2012, doi: 10.1109/JSAC.2012.120102.

[15] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005, doi: 10.1109/JSAC.2004.839380.

[16] L. Xiao, J. Liu, Q. Li, N. B. Mandayam, and H. V. Poor, "User-centric view of jamming games in cognitive radio networks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2578–2590, Dec. 2015, doi: 10.1109/TIFS.2015.2467593.

[17] B. Wang, Y. Wu, and K. J. R. Liu, "Game theory for cognitive radio networks: An overview," *Comput. Netw.*, vol. 54, no. 14, pp. 2537–2561, Oct. 2010, doi: 10.1016/j.comnet.2010.04.004.

[18] A. Neyman and S. Sorin, *Stochastic Games and Applications*. Stony Brook, NY, USA: Springer, 2003, pp. 9–25.

[19] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York, NY, USA: Springer, 1997, pp. 9–84.

[20] X. Xie and W. Wang, "Detecting primary user emulation attacks in cognitive radio networks via physical layer network coding," *Procedia Comput. Sci.*, vol. 21, no. 1, pp. 430–435, Dec. 2013, doi: 10.1016/j.procs.2013.09.057.

[21] V.-H. Vu, H. Thien, and I. Koo, "A repeated games-based secure multiple-channels communications scheme for secondary users with randomly attacking eavesdroppers," *Appl. Sci.*, vol. 9, no. 5, p. 868, Feb. 2019, doi: 10.3390/app9050868.

[22] D. Niyato, P. Wang, D. I. Kim, W. Saad, and Z. Han, "Mobile energy sharing networks: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3519–3535, May 2016, doi: 10.1109/TVT.2015.2437386.

[23] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012, doi: 10.1109/TSMCC.2012.2218595.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (A Bradford Book). London, U.K.: MIT Press, 2018.

[25] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 877–889, Apr. 2011, doi: 10.1109/JSAC.2011.110418.

[26] S. Singh and A. Trivedi, "Anti-jamming in cognitive radio networks using reinforcement learning algorithms," in *Proc. 9th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Indore, India, Sep. 2012, pp. 1–5, doi: 10.1109/WOCN.2012.6331885.

[27] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2087–2091, doi: 10.1109/ICASSP.2017.7952524.
[28] Y. Bi, Y. Wu, and C. Hua, "Deep reinforcement learning based multi-user anti-jamming strategy," in *Proc. ICC-IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6, doi: 10.1109/ICC.2019.8761848.
[29] J. Xu, H. Lou, W. Zhang, and G. Sang, "An intelligent anti-jamming scheme for cognitive radio based on deep reinforcement learning," *IEEE Access*, vol. 8, pp. 202563–202572, 2020, doi: 10.1109/ACCESS.2020.3036027.
[30] T. Takano, H. Takase, H. Kawanaka, H. Kita, T. Hayashi, and S. Tsuruoka, "Transfer learning based on forbidden rule set in actor-critic method," *Int. J. Innov. Comput. Inf. control*, vol. 7, no. 5, pp. 2907–2917, May 2011.
[31] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Res.*, vol. 10, pp. 1633–1685, Jul. 2009.
[32] D. Aha, M. Molineaux, and G. Sukthankar, "Case-based reasoning in transfer learning," in *Proc. Int. Conf. Case-Based Reasoning*. Berlin, Germany: Springer, 2009, pp. 29–44.
[33] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.
[34] L. A. Celiberto, J. P. Matsuura, R. L. D. Mantaras, and R. A. C. Bianchi, "Using transfer learning to speed-up reinforcement learning: A cased-based approach," in *Proc. Latin Amer. Robot. Symp. Intell. Robot. Meeting*, Sao Bernardo do Campo, Brazil, Oct. 2010, pp. 55–60, doi: 10.1109/LARS.2010.24.
[35] 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 Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014, doi: 10.1109/TWC.2014.022014.130840.
[36] K. A. M., F. Hu, and S. Kumar, "Intelligent spectrum management based on transfer actor-critic learning for rateless transmissions in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1204–1215, May 2018, doi: 10.1109/TMC.2017.2744620.
[37] G. Chen, Y. Zhan, Y. Chen, L. Xiao, Y. Wang, and N. An, "Reinforcement learning based power control for in-body sensors in WBANs against jamming," *IEEE Access*, vol. 6, pp. 37403–37412, 2018, doi: 10.1109/ACCESS.2018.2850659.
[38] C. Dai, L. Xiao, X. Wan, and Y. Chen, "Reinforcement learning with safe exploration for network security," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 3057–3061, doi: 10.1109/ICASSP.2019.8682983.
[39] X. Lu, L. Xiao, C. Dai, and H. Dai, "UAV-aided cellular communications with deep reinforcement learning against jamming," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 48–53, Aug. 2020, doi: 10.1109/MWC.001.1900207.
[40] W. Zhang, R. Mallik, and K. Letaief, "Optimization of cooperative spectrum sensing with energy detection in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5761–5766, Dec. 2009, doi: 10.1109/TWC.2009.12.081710.
[41] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 28–40, Feb. 2008, doi: 10.1109/JSTSP.2007.914882.
[42] M. G. Oskoui, P. Khorramshahi, and J. A. Salehi, "Using game theory to battle jammer in control channels of cognitive radio ad hoc networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5, doi: 10.1109/ICC.2016.7511334.
[43] M. T. Goodrich and R. Tamassia, *Algorithm Design and Applications*. Hoboken, NJ, USA: Wiley, 2015, p. 349.
[44] M. Hajimirsaadeghi and N. B. Mandayam, "A dynamic colonel blotto game model for spectrum sharing in wireless networks," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2017, pp. 287–294, doi: 10.1109/ALLERTON.2017.8262750.
[45] C. Jiang, Y. Chen, K. J. R. Liu, and Y. Ren, "Optimal pricing strategy for operators in cognitive femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5288–5301, Sep. 2014, doi: 10.1109/TWC.2014.2327970.
[46] Y. Zhu, D. Zhao, and X. Li, "Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 714–725, Mar. 2017, doi: 10.1109/TNNLS.2016.2561300.
[47] J. V. Stone, *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*, 2013 ed. Sheffield, U.K.: Sebtel Press, 2013, p. 174.
[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

HUYNH THANH THIEN received the B.E. degree in electronics and telecommunications engineering from Ton Duc Thang University, Ho Chi Minh, Vietnam, in 2012, and the M.S. degree in mechatronics technology from Chinese Culture University, Taipei, Taiwan, in 2014. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Ulsan, Ulsan, South Korea. His research interests include cognitive radio and next generation wireless communications systems, game theory, deep learning, and reinforcement learning.

VAN-HIEP VU received the B.E. degree in electronics and telecommunications engineering from Ton Duc Thang University, Ho Chi Minh, Vietnam, in 2005, the B.B.A. degree from the University of Economy, Ho Chi Minh City, in 2007, and the Ph.D. degree in electrical engineering from the University of Ulsan, Ulsan, South Korea, in 2013.

From 2013 to 2018, he was a Research Fellow with the Multimedia Communication System Laboratory, University of Ulsan. Since 2018, he has been a Researcher with the NTT Hi-Tech Institute, Nguyen Tat Thanh University, Vietnam. His current research interests include cognitive radio and next-generation wireless communications systems.

INSOO KOO received the B.E. degree from Konkuk University, Seoul, South Korea, in 1996, and the M.S. and Ph.D. degrees from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 1998 and 2002, respectively.

From 2002 to 2004, he was a Research Professor with the Ultrafast Fiber-Optic Networks Research Center, GIST. In 2003, he was a Visiting Scholar with the Royal Institute of Science and Technology, Stockholm, Sweden. Since 2005, he has been a Full Professor with the University of Ulsan, Ulsan, South Korea. His research interests include next-generation wireless communication systems and wireless sensor networks.