Anti-Jamming Game to Combat Intelligent Jamming for Cognitive Radio Networks

KHALID IBRAHIM, SOON XIN NG (Senior Member, IEEE), IJAZ MANSOOR QURESHI, AQDAS NAVEED MALIK, AND SAMI MUHAIDAT (Senior Member, IEEE)

1Department of Electrical Engineering, International Islamic University, Islamabad 44000, Pakistan
2Department of Electrical and Computer Engineering, COMSATS University Islamabad, Wah Campus, Wah Cantt 47040, Pakistan
3Next Generation Wireless Group, ECS, University of Southampton, Southampton SO17 1BJ, U.K.
4Center for Cyber-Physical Systems, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE
5Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

Corresponding author: Sami Muhaidat (sami.muhaidat@ku.ac.ae)

This work was supported in part by Khalifa University under Grant EX2020-037-8434000382 and Grant EX2020-038-8434000383.

ABSTRACT Cognitive Radio (CR) provides a promising solution to the spectrum scarcity problem in dense wireless networks, where the sensing ability of cognitive users helps acquire knowledge of the environment. However, cognitive users are vulnerable to different types of attacks, due to its shared medium. In particular, jamming is considered as one of the most challenging security threats in CR networks. In jamming, an attacker jams the communication by transmitting a high power noise signal in the vicinity of the targeted node. The jammer could be an intelligent entity that is capable of exploiting the dynamics of the environment. In this work, we provide a machine-learning-based anti-jamming technique for CR networks to avoid a hostile jammer, where both the jamming and anti-jamming processes are formulated based on the Markov game framework. In our framework, secondary users avoid the jammer by maximizing its payoff function using an online, model-free reinforcement learning technique called Q-learning. We consider a realistic mathematical model, where the channel conditions are time-varying and differ from one sub-channel to another, as in practical scenarios. Simulation results show that our proposed approach outperforms existing approaches to combat jamming over a wide range of scenarios.

INDEX TERMS Cognitive radio networks, anti-jamming communication, stochastic game, machine learning, multi-agent reinforcement learning, Q learning.

LIST OF SYMBOLS

- $\alpha$: Learning rate.
- $\beta$: Regret factor of jammer.
- $\eta$: Bandwidth efficiency (bps/Hz).
- $\gamma$: Discount factor.
- $a^n_l$: Action of the $n^{th}$ secondary user at $l^{th}$ channel.
- $c^m_l$: Action of the $m^{th}$ jammer at $l^{th}$ channel.
- $C^n_l(a^n_l, c^m_l)$: Channel capacity of the $n^{th}$ secondary user at $l^{th}$ channel at time slot $t$.
- $C^n_l$: Transmission cost of the $n^{th}$ secondary user at $l^{th}$ channel.
- $J^n_l$: Transmission cost of the $m^{th}$ jammer at $l^{th}$ channel.
- $T^n_l$: Transmission gain of $n^{th}$ SU at $l^{th}$ channel.
- $J^n_l$: Jamming gain of $m^{th}$ jammer at $l^{th}$ channel.
- $L$: No. of sub-channels in the network.
- $H$: No. of primary users in the network.
- $M$: No. of jammers in the network.
- $N$: No. of secondary users in the network.
- $Q(s^k, a^n_l)$: Q table entries for the secondary user at $l^{th}$ channel in $k^{th}$ time slot.
- $Q(s^k, c^m_l)$: Q table entries for jammer at $l^{th}$ channel in $k^{th}$ time slot.
- $A(s)$: Action space of secondary users.
- $S$: State space of the game.
- $U^n_l(a^n_l, c^m_l)$: Utility of the $n^{th}$ secondary user at $l^{th}$ channel at time slot $t$.

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

VOLUME 9, 2021 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Utility of the jammer at $l^{th}$ channel at time slot $t$.

Switching function of the $n^{th}$ secondary user at $l^{th}$ channel at time slot $t$.

Average received power at $l^{th}$ channel by $n^{th}$ secondary user at time slot $t$.

LIST OF ABBREVIATIONS

AWGN Additive White Gaussian Noise.
CH Channel Hopping.
CR Cognitive Radio.
CRN Cognitive Radio Network.
DoS Denial of Service.
DHSS Direct Sequence Spread Spectrum.
FH Frequency Hopping.
FHSS Frequency Hopping Spread Spectrum.
GSG General Sum Game.
HLA Hierarchical Learning Algorithm.
IL Independent Learner.
JAL Joint Action Learner.
MDP Markov Decision Process.
MARL Multi-Agent Reinforcement Learning.
MLE Maximum Likelihood Estimation.
NE Nash Equilibrium.
OLOFS One leader One Follower Stackelberg Game.
PSD Power Spectral Density.
PU Primary User.
PUEA Primary User Emulation Attack.
RF Radio Frequency.
SSDFA Spectrum Sensing Data Falsification Attack.
SU Secondary User.
SG Stochastic Game.
SLMFS Single Leader Multiple Follower Stochastic Game.
SE Stackelberg Equilibrium.
SNR Signal to Noise Ratio.
WSN Wireless Sensor Network.

I. INTRODUCTION

Cognitive Radio (CR) is a promising technology to cope with the scarcity issue of the electromagnetic spectrum, which is a natural resource. Traditional wireless radio communication works on fixed frequency slots, which results in overcrowding in certain portions of the electromagnetic spectrum while other portions of the spectrum are underutilized. CR is aware of its surrounding Radio Frequency (RF) environment. It learns, reasons, decides and adapts to external conditions with the aim of efficiently utilizing the radio spectrum and carrying out reliable and uninterrupted wireless communication [1], [2]. Furthermore, CR could provide opportunistic access to spectrum holes to solve intermittent use of radio spectrum using machine learning algorithms [3]–[7].

The adaptability of CR could provide intelligence to spectrum sensing and spectrum decision [5]. On the other hand, the adversary can also manoeuvre the same features intelligently to create more harm to the underlying CR Network (CRN) [8]–[10]. Therefore, ensuring security is of paramount importance to the successful deployment of cognitive radio networks. More explicitly, jamming attacks, Denial of Service (DoS) attacks [11], Primary User Emulation Attack (PUEA) [12]–[14], Spectrum Sensing Data Falsification Attack (SSDFA) [15], exploitation of common control channel security [16] and collaborative jamming [17] are well known attacks in cognitive radio networks. However, our major concern is to combat jamming attacks in cognitive radio networks.

The model of a jammer in a CRN is shown in Fig. 1, where the jammer disrupts the wireless communication by generating high-power noise causing narrowband interference on a single sub-channel at a time near the transmitting and receiving nodes [18]. Intensive jamming could result in either total disruption of the wireless communication or a very low signal to noise ratio (SNR) that does not allow secondary users (SUs) to communicate successfully. In traditional wireless communication systems, Frequency Hoping Spread Spectrum (FHSS) and Direct Sequence Spread Spectrum (DSSS) are widely used to thwart jammers [19], [20]. Due to dynamic spectrum mobility [21], it is not possible to directly apply these techniques in cognitive radio technology to combat the hostile jammers.

II. LITERATURE REVIEW

Game theory is widely used in wireless communications to solve communication problems like resource allocation [22]–[24], packet relaying [5] and anti-jamming communication [8], [10], [25]–[27], as shown in Fig. 2. Different types of jammers discussed in the literature [28], namely, random jammer [29], constant jammer [20], reactive jammer [30], [31], sweep jammer [32], [33] and intelligent jammer [8], [34]. In recent years, game theory has been extensively used to model the dynamic interaction between legitimate users and the jammer. Anti-jamming games could be based on power domain [31], [35]–[41], code domain [20], [42], frequency
Game theory has also been used to investigate the frequency hopping anti-jamming communication in wireless communication networks. For instance, anti-jamming communication in CRNs with unknown channel statistics has been studied in [55]. The authors formulated the problem of anti-jamming multi-channel access in CRN as a non-stochastic multi-armed bandit problem, where both secondary sender and receiver chooses their common operating channels by minimizing the probability of being jammed. Another interference avoidance based channel-hopping stochastic game was investigated in a multi-agent environment in [32], where game-theoretic based reinforcement learning mechanism is used to avoid jamming.

Recently, the authors in [56] presented bandwidth-efficient frequency hopping game in wireless sensor networks. The authors in [57] presented a brief overview of anti-jamming communication in the context of dynamic spectrum access. Two typical ways of thwarting jammer are by adaptation of transmission rate and Frequency Hopping (FH). These two are jointly adopted by [58] in order to improve the average throughput and to provide better jamming resiliency against reactive sweep jammer. Specifically, the interaction between the jammer and the legitimate user is modeled in [58] as a Zero Sum Markov Game (ZSMG) and a constrained NE is derived. The authors in [45] utilized game theoretic framework to access an optimal channel in the presence of attacker, hence maximizing the channel payoff. Channel Hopping (CH) based rendezvous scheme is adopted for the SU to meet and to make connection for further communications [44]. This bounded time rendezvous scheme neither uses pre-shared secrets nor is role pre-assignment needed for bringing the SUs on commonly available channel.

The authors in [46], presented a mobility-based Single Leader and Multiple Follower Stackelberg game (SLMFG) to avoid jamming for increasing the network life in the WSN. Anti-jamming games in multi-channel cognitive radio networks were presented in [47], where the SU hops to another channel to avoid the jamming. A zero sum game is played between the attacker and the SU based on a Markov Decision Process (MDP). Maximum Likelihood Estimation (MLE) and Q learning are used for SU to learn from their environment. In [43], the authors proposed a Hierarchical Learning Algorithm (HLA) for anti-jamming channel selection strategies in the presence of co-channel interference as a Stackelberg game. They considered jammer and users as independent learners (ILs), which choose their strategies independently and selfishly.

In [48], anti-jamming FH game is constructed using a bi-matrix game between the jammer and the legitimate user. In [49], game theoretic stochastic learning approach is used for anti-jamming communication in dense wireless networks. In [33], the authors have considered joint multi agent learners in stochastic game settings against a sweep jammer. They presented a collaborative multi-agent anti-jamming algorithm based on reinforcement learning in wireless networks. Markov game is formulated to model and
analyze the anti-jamming problem in a multi-user environment. A time domain countermeasure against random pulse jamming using MDP and reinforcement learning was presented in [29]. In [32], MARL is used as independent Q learning for each agent against a sweep jammer as a common practice. Another game theoretic anti-jamming scheme for CRN is presented in [34], where the SU uses Q learning to learn the dynamics of the jammer and to reduce the complexity of value iteration based learning. We will consider this scheme as our benchmark scheme. However, they only consider the anti-jamming in ideal channel conditions with no noise present. Secondly, they did not consider the time variations in the wireless channel. We have improved the framework presented in [34] by considering time varying variable channels, which is a more realistic approach in CRN. Moreover, in our case, utilities of both players are dependent on the channel qualities, the better the channel is, the higher is the reward and vice versa. We differentiate the sub-channels based on the received SNR, which results in the varying maximum channel capacities.

A. MOTIVATION AND CONTRIBUTION

From the above discussion, it is evident that the research community has contributed a lot of research towards anti-jamming for CRN in the frequency domain. However, most of literature has assumed fixed strategy for the jammer, which is not changing with time. With the development and technological advancement in the cognitive radio networks, it is highly conceivable that a jammer will also manoeuvre its attacking strategies intelligently. Hence, there is a need for intelligent anti-jamming strategy for CRN. To fulfill the need, we develop a mathematical modelling of the system that incorporates intelligence in the SU to cope with intelligent jammer. Intelligent jammer is cognitive in nature, having the ability to learn, to reason and to adjust its strategies against SU for maximum damage to the CRN.

Moreover, the authors in [35] have shown that for discrete problems like selection of frequencies in anti-jamming problems, it is difficult to manage using convex optimization. Hence, learning theory is needed in the decision process. The learning algorithm should be capable of coping with uncertain dynamics and incomplete information, whereas game theory can adequately model and analyze the mutual interactions among adversarial users. Therefore, it is promising to incorporate the learning algorithm into game theory.

Against this backdrop, we devise a game-theoretic optimal frequency hopping scheme between SU and intelligent jammer in a dynamic environment using the Q-learning approach to pick the optimal sub-channel as shown in Fig. 4. The emphasis of this study is on a game theoretic frequency hopping technique for avoiding the jammer. We refer the readers to [31], [35]–[41] for power adaptation based anti-jamming techniques. We develop a game model with both players as Independent Learners (ILs), where they selfishly and independently select their optimal sub-channels in a Multi-Agent Reinforcement Learning (MARL) setting to increase their individual utilities. The proposed game theoretic model in conjunction with learning based FH algorithm helps SU to avoid the attacker and hence to reduce the probability of jamming and to increase the bandwidth efficiency of the cognitive system. Our novel contribution can be summarized as follows:

1) We consider a cognitive adversarial jammer, which is an intelligent attacker that can adopt the dynamics of the sub-channels and the strategies of SU.

2) We extend the framework of [34], which assumed that all sub-channels have the same quality of service, by proposing a more realistic and practical channel model. More specifically, our channel conditions may change over time and differ from one sub-channel to another.

3) The proposed framework considers various factors and parameters that capture the near practical channel dynamics, e.g. SNR, variable channel capacity, jamming gain as well as transmission cost and jamming cost of each player in the game.

The rest of the paper is organized as follows. Section III explains the system and adversary models. Section IV provides an anti-jamming game formulation against intelligent jammer. Section V describes the proposed solution mechanism, while section VI describes the evaluation and results. Finally, Section VII concludes the paper.

III. SYSTEM AND ADVERSARY MODEL

A. SYSTEM MODEL

The interweave paradigm for the time-slotted system is assumed in our CRN, where SU can access the spectrum only if it is not used by the PU [59]. In cognitive radio-based communication systems, the presence of timing and frequency asynchronization affects the spectrum sensing performance and may result in false detection of PUs by the SUs. Assuming imperfect synchronization is more realistic for the better realization and analysis of communication systems. However, as a first step we assume perfect time and frequency synchronization between all SUs for the sake of simplicity as in [60], [61].

Every user scans the available sub-channels and starts transmission after a white space is found. We assume that there are $\mathcal{H}$ PUs, $\mathcal{N}$ SUs and $\mathcal{M}$ jammers in the network. Furthermore, $\mathcal{W}$ is the bandwidth of the channel. Each sub-channel may have different channel capacity, based on the received SNR. It is assumed that the jamming attack is the
only source of channel deterioration in the network and we neglect any other source of interference including the effects of multipath fading. Each sub-channel can be in one of the two states, namely the idle state and the busy state. The relationship between the PU and an SU can be described by one of the two possible states of the sub-channel as follow:

- **IDLE STATE**: The channel is idle if it is not being used by any PU. The SU and the jammer are allowed to utilize an idle channel. The idle state of the sub-channel is represented by \( P = 1 \)
- **BUSY STATE**: The channel is considered busy if it is being used by any PU. Both SU and jammer are not allowed to transmit over a busy channel. This state is represented by \( P = 0 \).

The channel states (idle or busy) are not known before the sensing action is taken place. We considered \( L \) sub-channels, where the quality of each sub-channel is different. Each sub-channel has its maximum capacity limit based on its received SNR, given by:

\[
C_{l,t}(a^n, c^m) = \frac{W}{L} \log_2(1 + SNR_{l,t}(a^n, c^m)),
\forall n \in N, \forall m \in M \tag{1}
\]

where \( C_{l,t}(a^n, c^m) \) represents the capacity of the \( l \)th sub-channel for the \( n \)th SU at time slot \( t \), where \( a^n \) and \( c^m \) are the actions of the \( n \)th SU and the \( m \)th jammer, respectively. Moreover, \( W/L \) is the bandwidth in Hz for each of the \( L \) sub-channels and \( SNR_{l,t}(a^n, c^m) \) is the received SNR of the \( l \)th channel for the \( n \)th SU. Let us first consider the case where there is no jammer present in the system and the SNR is defined as:

\[
SNR_{l,t}(a^n, c^m) = \frac{P_{l,t}^{n}}{N_0 \cdot W/L}, \forall n \in N, \forall m \in M \tag{2}
\]

where \( P_{l,t}^{n} \) is the average signal power received by the \( n \)th SU at the \( l \)th sub-channel at time slot \( t \) and \( N_0 \) is the power spectral density (PSD) of the Additive White Gaussian Noise (AWGN). A high \( SNR_{l,t}(a^n, c^m) \) would give a high channel capacity \( C_{l,t}(a^n, c^m) \) and hence a higher channel quality. The channel capacity of the \( n \)th SU at the \( l \)th sub-channel can be computed as:

\[
C_{l,t}(a^n, c^m) = \frac{W}{L} \log_2(1 + \frac{P_{l,t}^{n}}{N_0 \cdot W/L}), \forall n \in N, \forall m \in M \tag{3}
\]

Moreover, the Signal to Interference plus Noise Ration (SINR) in the presence of a jammer can be calculated as:

\[
SINR_{l,t}(a^n, c^m) = \frac{P_{l,t}^{n}}{N_0 \cdot W/L + N_j \cdot W/L}, \forall n \in N, \forall m \in M \tag{4}
\]

where \( N_j \) is PSD of the jamming signal and \( B_j \) is the bandwidth of the jammed channel. Since all sub-channels have identical bandwidth of \( B_j = \frac{W}{L} \), the SNR becomes:

\[
SNR_{l,t}(a^n, c^m) = \frac{P_{l,t}^{n}}{N_0 \cdot W/L + N_j \cdot W/L}, \forall n \in N, \forall m \in M \tag{4}
\]

We only get (4) when \( a^n = c^m \), i.e. both the SU and the jammer are on the same channel. This results in a severe degradation of the SNR for the SU. The objective of the SU is to carefully switch to the available high-capacity channel to maximize the spectrum utilization, while at the same time to avoid the potential jamming.

### B. SECONDARY USER MODEL

The SU senses its environment during its sensing period, before initiating any data transmission. Contention based channel selection algorithm uses a structure called Preferable Channel List (PCL) to initiate data transmission. The algorithm avoids collision and perform Request To Send / Clear To Send (RTS-CTS) contention for data transmission. Moreover, a sensing-assisted access (SAA) protocol may be used as a complete random access mechanism for CRN to initiate data transmission. In this mechanism, the contention-based access is designed based on the integration of the backoff process and spectrum sensing [62], [63]. However, in this contribution, this aspect is not investigated. During the sensing period, each SU would try to sense for the presence of any PU in the available sub-channels. However, the SU cannot be able to detect the presence of a jammer at the beginning of the time slot. Nonetheless, the SU is able to realize the presence of the jammer at the end of each time slot. More specifically, at the end of each time slot, the SU would know if its transmission was successful or was jammed by a malicious jammer. The interested readers may refer to [64], [65] for more details concerning jamming detection. A successful transmission yield a positive payoff to the SU, while a jammed transmission would yield a negative payoff to the SU. The utility of the \( n \)th SU in the \( l \)th sub-channel based on the actions of SU (represented as \( a^n \)) and the action of the jammer (denoted as \( c^m \)) at time slot \( t \) can be derived as:

\[
U_{l,t}^{n}(a^n, c^m) = C_{l,t}(a^n, c^m) x_{l,t}(a^n, c^m)(T_{l}^{n} - \xi_{l}^{n}) - (1 - x_{l,t}(a^n, c^m))(J_{t}^{l} + \xi_{l}^{l}), \tag{5}
\]

where \( \xi_{l}^{n} \) is the cost of transmission of the \( n \)th SU, \( T_{l}^{n} \) is the SU gain factor for a successful communication, \( J_{t}^{l} \) is the loss factor for SU when the SU transmission in the \( l \)th sub-channel is jammed. The impact of the sub-channel SNR and the decision of each player on the utilities of both players are shown in Fig. 5. The received SNR at each sub-channel is increasing from sub-channel 1 to sub-channel 10. The utilities earned by each player is opposite to each other. The missing utility at sub-channel 4 in Fig. 5 indicates that the PU is transmitting in this sub-channel and neither SU nor jammer can use this sub-channel. Furthermore, sub-channel 9 was jammed in the previous time slot, if the jammer stays there, then the SU would have negative utility to use sub-channel 9. Similarly, for the jammer, if SU was at sub-channel 9 and if it stays there, then the jammer would have positive utility at sub-channel 9. Additionally, the PU could change to a different sub-channel in each time slot, but we assume that
both the SU and the jammer are able to detect the sub-channel used by the PU.

The utility of our proposed system in (5) can be compared with the utility function in of the benchmark system [34] given by \( G_i(s, a) = \sum_{l=1}^{L} G_i(s, a) \), where the gain of the SU at the \( l^{th} \) sub-channel is computed as:

\[
G_i(s, a) = x_i(s, a) \times U - y_i(s, a) \times C,
\]

while \( x_i(s, a) \) and \( y_i(s, a) \) are binary switching functions. Furthermore, \( U \) and \( C \) denote the utility earned by the SU and the jamming cost of SU, respectively. The authors in [34] assume that the values for the utility and for the cost in every sub-channel are identical, since their sub-channels have the same quality. The left hand side of both (5) and (6) denotes the utility function of the SU, although different notations were used. The right hand side of (5) and (6) have the following differences:

- We introduce the factor \( C^n_i(d^n, c^m_j) \) in (5) in order to differentiate the sub-channels based on the sub-channel capacities. Hence, successful transmission in a good quality sub-channel would yield better utility for the SU. This quality factor is missing in [34] and in (6).
- We also considered the factors \( E^n_i \) and \( E^m_j \) to account for the transmission costs for SU and for the jammer, respectively, in terms of the battery utilization and the power transmitted.
- The two binary switching functions \( x_i(s, a) \) and \( y_i(s, a) \) in (6), are used in [34] such that \( x_i(s, a) + y_i(s, a) = 1 \). To simplify the mathematical notation, we used only one binary switching function \( x_{l,t}(d^n, c^m_j) \) instead of two, such that \( x_{l,t}(d^n, c^m_j) + (1 - x_{l,t}(d^n, c^m_j)) = 1 \).
- Furthermore, we have more detailed utility function for the jammer compared to that of [34] as will be explained later in (11) and (13).

Combining (5) and (3) gives (7), as shown at the bottom of the page, where \( x_{l,t}(d^n, c^m_j) \in [1, 0] \) is a binary switching function used to indicate successful/jammed SU communication:

\[
x_{l,t}(d^n, c^m_j) = \begin{cases} 
1, & d^n \neq c^m_j, \forall n \in N, \forall m \in M, \\
0, & d^n = c^m_j, \forall n \in N, \forall m \in M.
\end{cases}
\]

Note that \( x_{l,t}(d^n, c^m_j) \) is 1 for successful SU transmission and is 0 for jammed SU transmission. Specifically, \( C^n_i(d^n, c^m_j) = 0 \) if the SNR is below a certain threshold value \( \text{SNR}_{th} \), i.e. \( \text{SNR}_{th}^n(d^n, c^m_j) \leq \text{SNR}_{th} \) and the value of the switching function \( x_{l,t}(d^n, c^m_j) \) would also become 0.

Equations (2), (4) and (8) are related in the sense that:

\[
x_{l,t}(d^n, c^m_j) = \begin{cases} 
1, & \text{SNR}_{th}^n(d^n, c^m_j) > \text{SNR}_{th}, \\
0, & \text{SNR}_{th}^n(d^n, c^m_j) \leq \text{SNR}_{th}.
\end{cases}
\]

In other words, the SU utility function in (7) incorporates the practical channel condition in terms of both the channel capacity and the jamming conditions. The goal of the SU is to maximize the expected sum of the discounted payoff by choosing a good quality channel that is not jammed by the jammer.

C. JAMMER MODEL

Jamming is a hostile attack in the CRN, where it disrupts the wireless communication by generating high-power noise at the targeted sub-channel as shown in Fig. 6. We considered two types of jammers i.e. random jammer and intelligent jammer. A random jammer would randomly jam a sub-channel in different time slot. Inspired from [34] and [8], when an intelligent jammer with cognitive capabilities is assumed, it adopts the best strategy in response to the observation of the channel dynamics and the SU strategies. The jammer senses the RF environment for a given sensing duration and then transmits its jamming signals based on the channel conditions and the strategy of the SU. If a PU is detected in a sub-channel, the jammer would switch to other available sub-channels to

\[
U^n_{l,t}(d^n, c^m_j) = \frac{W}{L} \log_2(1 + \frac{P^n_{l,t}}{N_o \cdot W / L})(x_{l,t}(d^n, c^m_j)(T^n_I^l - E^n_i) - (1 - x_{l,t}(d^n, c^m_j))(J^n_I^l + E^n_i)), \quad \forall n \in N, \forall m \in M
\]
avoid the heavy penalty imposed by law-enforcement agencies and start sensing again [8]. The utility function of the jammer in the $l^{th}$ sub-channel is based on the actions of the SU and of the jammer, which is represented by:

$$U_{jl}^m(a^n, c_m^n) = C_{jl}^m(a^n, c_m^n)(1 - x_{jl}(a^n, c_m^n)(\mathcal{T}_{jl}^m - c_m^n) - x_{jl}(a^n, c_m^n)(\beta \mathcal{E}_{jl}^m)), \quad \forall n \in \mathcal{N}, \quad \forall m \in \mathcal{M},$$

where $C_{jl}^m(a^n, c_m^n)$ is the channel capacity of the $l^{th}$ sub-channel and $\mathcal{T}_{jl}^m$ is the jammer gain factor when a SU was successfully jammed, while $\mathcal{E}_{jl}^m$ is the cost of transmitting the jamming signals. Furthermore, $\beta$ is the jammer regret factor when the jamming was not successful, which is the negative reward earned by the jammer when the jammer sends a jamming signal to a sub-channel that was not used by the SU. As mentioned in (8), $x_{jl}(a^n, c_m^n)$ is a switching function having $x_{jl}(a^n, c_m^n) = 0$, when the jammer successfully jams a channel (zero regrets), while $x_{jl}(a^n, c_m^n) = 1$ when the jammer fails to jam the SU. Hence, an intelligent jammer is more inclined to jam a SU that operates in a high-capacity sub-channel compared to that in a low-capacity sub-channel. The objective of the jammer is to maximize the probability of successful jamming.

### IV. GAME THEORETIC ANTI-JAMMING MECHANISM

#### A. PRELIMINARIES

A Stochastic Game (SG) is the natural extension and generalization of the Markov Decision Process (MDP) to multi-agent systems [7], [66]. A SG provide a framework for multi-agents in multiagent reinforcement learning (MARL). In this contribution, a stochastic anti-jamming game is developed between two players of conflicting interests.

**Definition 1 [67]:** A two-player stochastic game is defined as $\mathcal{G} = (X, S, A_i, U_i)$, where $X = \{1, 2\}$ is the index of the players, $S$ is the discrete strategy space of the game, $A_i$ is the discrete action space of player $i$, while $U_i : S \times A_i$ is the utility/payoff of player $i, \forall i \in X$.

Stochastic games are also called Markov games, which essentially are $n$-agent MDP [67].

**Definition 2 [68]:** A pair of strategies $(i^*, j^*)$ for the row player, and $j$ for the column player) yields a non-cooperative Nash equilibrium solution to a bimatrix game $(A = \{A_i\}, B = \{B_{ij}\})$, where $A$ and $B$ are payoff matrices for each player, if the following two inequalities are satisfied:

$$A_{ij} \geq A_{il}, \quad \forall i, j = 1, 2, 3, \ldots, P,$$

$$B_{ij} \geq B_{il}, \quad \forall i, j = 1, 2, 3, \ldots, P,$$

where $P$ is the total number of pure strategies. Each stage of a stochastic game can be viewed as a bimatrix game.

The basic assumption of a stochastic game between two interacting players is that the actions of each player will have an impact on the utility of other player. The same assumption is valid in our case. The SU obtain its utility based on the past actions of jammer in the previous time slot. If the sub-channel to be accessed by the SU receives an SNR lower than a threshold, it implies that the sub-channel was successfully jammed by the jammer and SU will get lower utility at that sub-channel at time slot $t$.

There are two types of learners in the MARL setting, namely the independent learner (IL) and the joint action learner (JAL) [69]. IL uses Q-learning in a classical setting, ignoring the other agents. More specifically, it assumes that the other agents are part of the environment. A MARL algorithm is an IL algorithm if the learner would take action individually and would not consider the actions taken by other agents. The IL algorithm is an appropriate method of learning if the agent is unaware of the other agents in the system and their actions [70], [71].

A JAL is an agent that learns the environment in the presence of other agents and then updates its Q values based on the joint actions of all the agents in the system. In fact, even though JALs have much more information at their disposal, they do not perform much different from ILs in the straightforward application of Q-learning to multi-agent systems [69]. We take both the SU and jammer players as ILs, where each IL would apply the Q-learning algorithm in the classical setting, while ignoring the actions of the other agents.

**Theorem 1 [72]:** An IL agent in a MARL setting, which follows the Q-learning update rule, will converge to the optimal Q-function with unit probability.

### B. GAME FORMULATION

Based on the knowledge about the channel, the system and the attacker, the objective of the SU is to carefully choose a sub-channel to maximize its spectrum utilization, while avoiding the jamming. On the other hand, the jammer aims to forbid the SU from effective channel utilization by a strategic jamming approach. The objectives of the two players, namely the jammer and the SU, are opposite to each other. Therefore, the dynamic interaction between them is well formulated as a non-cooperative game, where the gain of one player is the loss of another player. Furthermore, spectrum availability,
quality of the channel and strategies of both SU and jammer can be time-varying. Players of the game are assumed to be intelligent and would exhibit rational behaviour to maximize their own payoffs according to their individual objectives.

We have formulated a two-player SG between the SU and the jammer as described below:

**Players:** There are two non-cooperative players in the game namely the SU and the jammer.

**States:** Every sub-channel occupation is considered as the state $S$ of the game. For example, if we have $L$ sub-channels, then we have $L$ states. The number of available states to the SU and jammer is given by $L - H$, where $L$ is the total number of sub-channels and $H$ is the number of PUUs in the network.

**Actions:** An action $A(s)$ at each state has $L - H$ hopping possibilities. For each $L - H$ available states, the possible action set is $A(s) = \{a_1, a_2, a_3, \ldots, a_l, \ldots, a_{L-H}\}$, where $a_i$ is the action to hop to the $i_{th}$ sub-channel in the $L - H$ available sub-channels. Both players choose actions to hop to any of the available sub-channels, which are not occupied by the PU. Since the available frequency slots are the same for both players, the action set $A(s)$ is the same for both players. Every action results in a change of state. Both jammer and SU sense the channel during the sensing period and hence the channel states and channel quality are assumed to be a common knowledge in the game.

**Payoff:** The immediate payoffs of both players in a bimatrix game at each stage are given by (7) and (11), respectively. The total utility of the whole secondary network is given by

$$U_{T,r}(a, c_j) = \sum_{n=1}^{N} \sum_{l=1}^{L} U_{nl}(a^n, c_j^n),$$

where $U_{nl}(a^n, c_j^n)$ is given by (7). The long-term objective of SU is to maximize the expected sum of discounted payoff, which can be written as [8], [66]:

$$\max E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{T,r}(a, c_j) \right\},$$

where $\gamma^t$ is a time decaying discount factor, $0 < \gamma^t < 1$, that determines the significance of future payoffs and $U_{T,r}(a, c_j)$ is the utility of the secondary network at time $t$, which is given by (13). The frequency hopping strategy of the SU is to maximize its utility by taking an optimal action that is given by:

$$a^* = \arg \max_{a \in A(s)} E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{T,r}(a, c_j) \right\}. \quad (15)$$

Similarly, the frequency hopping strategy of an intelligent jammer is to maximize its expected utility of $\sum_{t=0}^{\infty} \gamma^t U_{l,T,r}(a, c_j)$ by taking an optimal action of:

$$c_j^* = \arg \max_{c_j \in A(s)} E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{l,T,r}(a, c_j) \right\}. \quad (16)$$

The pair $(a^*, c_j^*)$ is said to be an equilibrium pair, if (15) and (16) follow the following inequalities:

$$E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{l,T,r}(a^*, c_j^*) \right\} \geq E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{l,T,r}(a^*, c_j) \right\},$$

$$E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{l,T,r}(a^*, c_j^*) \right\} \geq E \left\{ \sum_{t=0}^{\infty} \gamma^t U_{l,T,r}(a^*, c_j) \right\}. \quad (17)$$

**V. SOLVING OPTIMAL STRATEGIES OF THE GAME**

In the previous section, we have described the anti-jamming game formulation. In this section, we will describe the defending strategy of the SU i.e. how to defend the SU from being jammed by the jammer. The Q-learning is a value-based reinforcement learning algorithm, which uses a Q table to maximize the utility [73]. It is a mechanism to effectively learn the sub-channel selection strategy. The Q function $Q(s^k, a^k)$ at stage $k$, is the expected discounted payoff when the SU takes the action $a^k$ at the state $s^k$. More specifically, the Q value is the estimation of the expected sum of the discounted payoffs [8]. Hence, an SU can consider the Q value in a bimatrix game at stage $k$ as the expected sum of the discounted payoffs given in (14). Given the Q function $Q(s^k, a^k)$, the SU can find the value of the game from:

$$V(s^k) = \max_{a^k} Q(s^k, a^k). \quad (18)$$

After an action $a^k$ is taken, the SU would receive an immediate payoff $R(s^k, a^k)$, which is then used to update the Q table. Specifically, the Q function can be approximated as:

$$Q(s^k, a^k) = R(s^k, a^k) + \sum_{s^{k+1} \in S} P(s^{k+1} | s^k, a^k) V(s^{k+1}), \quad (19)$$

where $P(s^{k+1} | s^k, a^k)$ is the transition probability from state $s^k$ to $s^{k+1}$. The value of $Q(s^k, a^k)$ in (19) can be updated recursively without having to estimate the transition probabilities [74], as follows:

$$Q(s^k, a^k) = Q(s^k, a^k)(1 - \alpha^k) + \alpha^k [R(s^k, a^k) + \gamma \max_{a^{k+1}} Q(s^{k+1}, a^{k+1})]], \quad (20)$$

where $\alpha^k$ is the linear learning rate satisfying $0 < \alpha^k < 1$. For $\alpha^k$ to be time decaying it must satisfy the conditions $\sum_{k=0}^{\infty} \alpha^k = \infty$ and $\sum_{k=0}^{\infty} (\alpha^k)^2 < \infty$. Note that $\gamma$ is the discount factor satisfying $0 < \gamma < 1$. Similarly, the intelligent jammer can also adopt the Q-learning algorithm to learn the dynamics of the SU and the sub-channels [75]. After updating the Q table, an intelligent jammer could take a decision based on its own updated Q table. The update rule for the jammer’s Q value is given by:

$$Q_l(s^k, c_j^k) = Q_l(s^k, c_j^k)(1 - \alpha^k) + \alpha^k [R_l(s^k, c_j^k) + \gamma \max_{c_j^{k+1}} Q_l(s^{k+1}, c_j^{k+1})]], \quad (21)$$
where $Q(s^k, a^k)$ and $Q_j(s^k, c_j^k)$ are the estimates of the expected sums of the discounted payoffs for both the SU and the jammer, respectively, which could evolve over time. The rewards of the SU and the jammer after choosing their respective actions at state $s^k$ are given by $R(s^k, a^k)$ and $R_j(s^k, c_j^k)$, respectively. These immediate rewards are calculated using (7) and (11) for the SU and the jammer, respectively. The SU would stay on the current sub-channel if the reward it receives on the current sub-channel is good enough to contribute positively in the Q value update. A negative instant reward in a certain sub-channel indicates that the sub-channel has been jammed and the SU should avoid that sub-channel by hopping to another available sub-channel in the next round. The SU’s optimal frequency hopping strategy to avoid the jammer is the action that maximizes its Q-value in state $s$ and is given by:

$$a^* = \arg \max_{a \in A(s)} Q(s, a).$$ (22)

Algorithm 1 summarizes the game theoretic frequency hopping algorithm for SU. The jammer’s optimal frequency hopping strategy to jam the SU is a greedy policy that chooses the action with maximum Q-value in state $s$ and is given by:

$$c_j^* = \arg \max_{c_j \in A(s)} Q_j(s, c_j).$$ (23)

Algorithm 1 Game Theoretic Frequency Hopping Algorithm

1: Initialize $k = 0$, $K = 100$, $a^k \in [0,1]$ and $\gamma \in [0,1]$, $\forall s^k \in S$, $a^k \in A(s)$ and $c_j^k \in A(s)$.
2: Let $Q(s^k, a^k) = 0$.
3: while $k \rightarrow K$ do
4:   Execute action $a^k$, get immediate reward $R(s, a^k)$ according to (7) and observe $s^k+1$.
5:   Choose $a^k+1$ from $s^k+1$ using policy derived from $Q(s^k, a^k)$ as given in (22).
6:   Update $Q(s^k, a^k)$ for SU according to (20).
7:   $s^k \leftarrow s^k+1$ and $a^k \leftarrow a^k+1$.
8: end while
9: if $a = c_j$ then
10:   The channel is jammed.
11: else
12:   The SU transmission is successful.
13: end if

For jammer, the procedure is summarized in Algorithm 2.

Algorithm 2 Game Theoretic Algorithm for Jammer

1: Initialize $k = 0$, $K = 100$, $a^k \in [0,1]$ and $\gamma \in [0,1]$, $\forall s^k \in S$, $a^k \in A(s)$ and $c_j^k \in A(s)$.
2: Let $Q_j(s^k, c_j^k) = 0$.
3: while $k \rightarrow K$ do
4:   Execute action $c_j^k$, get immediate reward $R_j(s, c_j^k)$ according to (11) and observe $s^k+1$.
5:   Choose $c_j^k+1$ from $s^k+1$ using policy derived from $Q_j(s^k, c_j^k)$ as given in (23).
6:   Update $Q_j(s^k, c_j^k)$ for jammer according to (21).
7:   $s^k \leftarrow s^k+1$ and $c_j^k \leftarrow c_j^k+1$.
8: end while
9: if $c_j = a$ then
10:   The jammer is successful.
11: else
12:   The SU transmission is successful.
13: end if

linear equations. The complexity of policy phase is given by $O(S)$. On the other hand, complexity of Q learning phase is calculated as $O(|A|.|S|)$. By combing both, we get $O(|A|.|S| + |S|)$. As the algorithms run for $K$ number of iterations, hence the overall complexity may be represented by $O(K.|A|.|S| + |S|)$, where $S$ and $A$ represent states and actions, respectively.

VI. RESULTS AND DISCUSSION

In our simulation study, we have considered $N = 10$ sub-channels, one SU, one PU and up to four jammers. When a sub-channel is occupied by the PU, neither the SU nor the jammers can access that channel. The SU chooses a high-capacity sub-channel that is potentially jamming-free, while the jammers predict and choose the sub-channel used by the SU.

A. THE EFFECT OF USING DIFFERENT CHANNEL TYPES

The capacity of a sub-channel is a measure of the highest information rate that can be achieved with a very small error rate. The channel capacity $C_{n,a}(a^n, c_j^m)$ is represented by (1) and (10), while the bandwidth efficiency in bits per second per Hertz (bps/Hz) can be computed as:

$$\eta^n = \frac{C_{n,a}(a^n, c_j^m)}{W/\mathcal{L}} = \log_2 \left(1 + \frac{P_{n,a}^I}{N_0 W/\mathcal{L}}\right), \quad \forall n \in \mathcal{N}$$ (24)

In each epoch the simulations are run $A = 2000$ times to get the average bandwidth efficiency for each SU:

$$\eta = \frac{1}{A} \sum_{n=1}^{A} \eta^n.$$ (25)

We considered two cases, where case 1 refers to the situation when all the $\mathcal{L}$ sub-channels have different SNRs,
TABLE 1. Parameters used in the simulations, as well as different SNR values for all sub-channels in case I.

| $l$ | $\text{SNR}_{\text{dB}}$ | Other Parameters |
|-----|-----------------|-----------------|
| 1   | 0               | $\mathcal{L}$   | 10             |
| 2   | 10              | $\mathcal{H}$   | 1              |
| 3   | 13              | $N$             | 1              |
| 4   | 14              | $M$             | 2              |
| 5   | 15              | $T_{c_1}$       | -100           |
| 6   | 16              | $\varepsilon_l$ | 10             |
| 7   | 17              | $\varepsilon_l$ | 10             |
| 8   | 18              | $T_{c_1}$       | 100            |
| 9   | 19              | $\alpha$       | 0.1            |
| 10  | 20              | $\beta$        | -100           |
|     | $\gamma$       | 0.8             |

$\text{SNR}_{\text{dB}} = 16.2$ dB

and hence, different channel capacities as given by Table 1. By contrast, all $\mathcal{L}$ sub-channels have the same SNR in case II. More specifically, case II is related to the idealistic scenario considered in [34]. The mean SNR in dB is calculated by:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left( \frac{1}{\mathcal{L}} \sum_{l=1}^{\mathcal{L}} 10^{\frac{\text{SNR}_{\text{dB},l}}{10}} \right),$$

where $\text{SNR}_{\text{dB},l} = 10 \log_{10}(\text{SNR}_{l,t})$ is the SNR of $l^{th}$ sub-channel. The SNR for each sub-channel in case II is the same as the average SNR of case I ($\text{SNR}_{\text{dB}} = 16.2$ dB). For a fair comparison, the means SNRs for both cases are equal, which is $\text{SNR}_{\text{dB}} = 16.2$ dB as shown in Table 1.

B. COMPARISON WITH THE BENCHMARK SYSTEM

Here, we compare the proposed scheme and the benchmark scheme in case I, where the sub-channels have varying qualities. However, the benchmark scheme assumes that all sub-channels have a fixed quality (6). As seen in Fig. 7, our proposed system outperforms the benchmark scheme both in terms of jamming probabilities and bandwidth efficiency. Please note that when the channel qualities of all sub-channels are fixed, the proposed scheme will perform similarly to the benchmark scheme. In other words, our proposed scheme is the generalization of the benchmark scheme to the general case, where the sub-channel qualities are varying.

Fig. 8 shows the probability of successful jamming by the jammer and the bandwidth efficiency of the SU, for both case I and case II, when one or two random jammers are considered. With increasing epochs, the intelligent SU could learn the environment in a better way and the probability of successful jamming is expected to be reduced, while the bandwidth efficiency would increase. The probability of successful jamming for the two-jammer scenario is slightly higher than that of the single-jammer scenario, but the probabilities converge to zero after 60 epochs as shown in Fig. 8a. Furthermore, in the more challenging case I, where the channel quality varies across the sub-channels, our algorithm still works well despite requiring a longer training period (or epochs) to reach the convergence point, as seen in Fig. 8b.

Our proposed algorithm allows SU to intelligently choose sub-channels with higher channel capacities, when the channels are varying as in case I. Hence, the average bandwidth efficiency of the system is improved. It can also be seen from Fig. 8b that the average bandwidth efficiency in the single-jammer scenario of case I (solid line) is higher than that of the single-jammer scenario of case II (dotted line) after 40 epochs. A similar pattern can be seen for the two-jammer scenario in Fig. 8b, after 45 epochs. In other words, our algorithm works better for case I after a sufficient training period. Hence, the proposed scheme that operates in variable-quality channels (in case II) outperforms the benchmark scheme of [34] that works in fixed-quality channels (in case-I). Furthermore, the average bandwidth efficiency of the SU in the two-jammer scenario is almost equal to that of single-jammer scenario for both case I and case II. This indicates that our algorithm performs equally well when it has to work against two random jammers.

C. THE EFFECT OF HAVING DIFFERENT TYPES OF ATTACKS

In this scenario, we investigate the effect of having an intelligent jammer in the system. Keep in mind that the intelligent
Our proposed scheme works better against intelligent jammer in variable sub-channels case (case-I, dash dot red) as compared to the benchmark scheme (case-II, dash indigo). Our proposed scheme (case-I, dash dot red) reduces the successful jamming probability to zero after 70 epochs, while benchmark scheme (case-II, dash indigo) is not capable of doing so.

As seen in Fig. 9b, the corresponding average SU bandwidth efficiency in the presence of an intelligent jammer (solid line) is almost equal to that when having a random jammer (dashed line) after 40 epochs. Hence, our intelligent SU is capable of avoiding the intelligent jammer after a certain training period. Furthermore, the SU bandwidth efficiency in case I is higher than that in case II. Hence, the SU is also capable of choosing intelligently sub-channels with higher capacity, in the case I, for increasing the average bandwidth efficiency of the system, while successfully avoiding the intelligent jammer.

D. THE EFFECT OF USING DIFFERENT DEFENSE STRATEGIES

Here, we discuss all four possible intelligent/random SU against intelligent/random jammer scenarios, based on case I.
As seen in Fig. 10, when a SU chooses a random sub-channel strategy in the presence of a random jammer, then both the successful jamming probability and the average bandwidth efficiency of SU remain almost constant (dotted lines). The performance of SU improves remarkably when it behaves intelligently against the random jammer (dashed lines).

It is visible that a SU using random strategy against an intelligent jammer will result in a severe jamming and the SU bandwidth efficiency degrades drastically (dashed dotted lines). Hence, the SU must be intelligent to combat against an intelligent jammer, in order to reduce the successful jamming probability and to increase the average SU bandwidth efficiency (solid lines).

E. THE EFFECT OF MULTIPLE INTELLIGENT SUs AGAINST AN INTELLIGENT JAMMER

In Fig. 11 we compared the performances when increasing the number of intelligent SUs in the secondary network. In Fig. 11a, it is shown that the jamming probability is higher for two intelligent SUs in the presence of a single intelligent jammer compared with the situation when only one intelligent SU is transmitting. The huge impact of increase in bandwidth efficiency is shown in Fig. 11b. Bandwidth efficiency of the secondary network is almost doubled when we have two intelligent SUs against an intelligent jammer.

F. TIME SLOTTED VIEW

In Fig. 12, the x-axis shows the time slot index and the height of the bar shows the utility earned based on the decision of each player after a training period of 100 epochs. As we have described in the proposed model, each sub-channel has different quality based on the received SNR. Without loss of generality it is assumed that \( \text{SNR}_1 \leq \text{SNR}_2 \leq \cdots \leq \text{SNR}_i \leq \cdots \leq \text{SNR}_N \), \( i \in N \) i.e. the SNR is increasing from sub-channel 1 to sub-channel 10. The impact of changes in SNR and the decision of each player on the utilities are shown in Fig. 5, where the utilities earned by each player are opposite to each other.
When the players are trained adequately, then the corresponding Q tables would be updated properly, which would result in good decisions for all players. The intelligent SU has more choices in terms of choosing an optimal sub-channel. More explicitly, the SU can choose any of the available sub-channels, while there is only one sub-channel that the jammer can choose 'correctly' for a successful jamming. Hence, it is less probable for the intelligent SU to get jammed based on the updated Q table values. Both intelligent SU and intelligent jammer would opt for high quality channels as depicted by the height of each bar in Fig. 12a. As seen in Fig. 12a, the SU chooses sub-channel 10, while the jammer chooses sub-channel 4 at time slot 1. Hence the SU has a positive utility, while the jammer has a negative utility. Also shown in Fig. 12a, the jammer only manages to jam the SU at time slot 10, over the 15 time slots considered. Hence, the intelligent SU manages to avoid the jammer, while choosing high-capacity sub-channels.

**H. INTELLIGENT SU AND RANDOM JAMMER**

Fig. 12b shows the decision patterns for the case when intelligent SU adopts the Q learning for strategy update, while the jammer uses random strategy. As seen in Fig. 12b, the intelligent SU manages to avoid the random jammer in all of the 15 time slots considered, while at the same time capable of choosing high-capacity sub-channels.

**I. RANDOM SU AND INTELLIGENT JAMMER**

Fig. 12c depicts the decision patterns for the case when the SU uses a random strategy against an intelligent jammer that invokes the Q learning for strategy update. As depicted in Fig. 12c, the jamming rate is high, which is 11 successful jamming out of the 15 time slots considered.

**J. THE EFFECT OF INCREASING THE NUMBER OF SUB-CHANNELS AND JAMMERS**

Fig. 13 shows the impact of increasing the number of sub-channels and the number of jammers. As the number of sub-channel increases, the intelligent SU will have a higher chances of avoiding the jammer. It is shown that the bandwidth efficiency of SU increases as the number of sub-channels increases due to the increase of the choices in the sub-channel space. We find that increasing the number of random jammers does not effect much the bandwidth efficiency of the intelligent SU. It seems that frequency hopping is a very good strategy for the SU to avoid the jammers, especially when the number of sub-channels is high.
VII. CONCLUSION
In this paper, we have investigated an anti-jamming stochastic game in conjunction with multi-agent reinforcement learning algorithm. Both random and intelligent jammers were considered. The anti-jamming game was designed as a Markov game based on the Q-learning algorithm. We devised a game-theoretic optimal frequency hopping scheme in a dynamic environment in the presence of adversarial jammers by using Q-learning approach to pick high-capacity sub-channels while avoiding the jammer. We developed a game model with both players as independent defenders, where SU and jammer selfishly and independently select their optimal sub-channels in a multi-agent reinforcement learning setting to increase their individual utilities. The proposed game theoretic model in conjunction with learning based frequency hopping algorithm helps the SU to avoid the attacker, hence reduces the probability of jamming and increases bandwidth efficiency of SU. It was shown in our simulation results that the proposed method outperforms the benchmark system in terms of both the bandwidth utilization and the jamming probability. Furthermore, when the channel exhibits variable channel quality (as in case I), our intelligent SU is capable of intelligently choosing sub-channels with higher capacity, while avoiding the intelligent jammer. Moreover, the bandwidth efficiency of the SU does not decrease significantly when the number of random jammers increases.

REFERENCES
[1] L. Gavrilovska, V. Atanasovski, I. Macaluso, and L. A. DaSilva, “Learning and reasoning in cognitive radio networks,” IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 1761–1777, 4th Quart., 2013.
[2] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” IEEE J. Sel. Areas Commun., vol. 23, no. 2, pp. 201–220, Feb. 2005.
[3] W. Liang, S. X. Ng, J. Feng, and L. Hanzo, “Prognostic distributed algorithm for spectral access in cooperative cognitive radio networks,” IEEE Trans. Commun., vol. 62, no. 4, pp. 1188–1200, Apr. 2014.
[4] W. Liang, S. X. Ng, and L. Hanzo, “Cooperative overlay spectrum access in cognitive radio networks,” IEEE Commun. Surveys Tuts., vol. 19, no. 3, pp. 1924–1944, 3rd Quart., 2017.
[5] A. H. Bastami and P. Kazemi, “Cognitive multi-hop multi-branch relaying: Spectrum leasing and optimal power allocation,” IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 4075–4088, Aug. 2019.
[6] T. M. Chiwewe, C. F. Mbuya, and G. P. Hancke, “Using cognitive radio for interference-resistant industrial wireless sensor networks: An overview,” IEEE Trans. Ind. Informat., vol. 11, no. 6, pp. 1466–1481, Dec. 2015.
[7] M. Bkassiny, Y. Li, and S. K. Jayaweera, “A survey on machine-learning techniques in cognitive radios,” IEEE Commun. Surveys Tuts., vol. 15, no. 3, pp. 1136–1159, Jul. 2013.
[8] 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.
[9] F. Salahdine and N. Kaabouch, “Security threats, detection, and countermeasures for physical layer in cognitive radio networks: A survey,” Phys. Commun., vol. 39, Apr. 2020, Art. no. 101001.
[10] M. A. Aref, S. K. Jayaweera, and E. Yepez, “Survey on cognitive anti-jamming communications,” IET Commun., vol. 14, no. 18, pp. 3110–3127, Nov. 2020.
[11] S. Sadagari, A. Attar, V. C. M. Leung, and S. G. Bilen, “Denial of service attacks in cognitive radio networks through channel eviction triggering,” in Proc. IEEE Global Telecommun. Conf. (GLOBECOM), Dec. 2010, pp. 1–5.
[12] A. Anand, Z. Jin, and K. P. Subbalakshmi, “An analytical model for primary user emulation attacks in cognitive radio networks,” in Proc. 3rd IEEE Symp. Frontiers Dyn. Spectr. Access Netw., Oct. 2008, pp. 1–6.
[13] H. Li and Z. Han, “Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part I: Known channel statistics,” IEEE Trans. Wireless Commun., vol. 9, no. 11, pp. 3566–3577, Nov. 2010.
[14] K. Li and J. Wang, “Optimal joining strategies in cognitive radio networks under primary user emulation attacks,” IEEE Access, vol. 7, 2018, 183812–183822, 2018.
[15] A. Ahmadfard, A. Jamshidi, and A. Keshavarz-Haddad, “Probabilistic spectrum sensing data falsification attack in cognitive radio networks,” Signal Process., vol. 137, pp. 1–9, Aug. 2017.
[16] B. Rashid, M. H. Rehmani, and A. Ahmad, “Broadcasting strategies for cognitive radio networks: Taxonomy, issues, and open challenges,” Comput. Electr. Eng., vol. 52, pp. 349–361, May 2016.
[17] W. Wang, S. Bhattacharjee, M. Chatterjee, and K. Kwiat, “Collaborative jamming and collaborative defense in cognitive radio networks,” Pervasive Mobile Comput., vol. 9, no. 4, pp. 572–587, Aug. 2013.
[18] H. Zhang, H. Xing, J. Cheng, A. Nallanathan, and V. C. M. Leung, “Secure resource allocation for OFDMA two-way relay wireless sensor networks without and with cooperative jamming,” IEEE Trans. Ind. Informat., vol. 12, no. 5, pp. 1714–1725, Oct. 2016.
[19] R. Pickholz, D. Schilling, and L. Milstein, “Theory of spread-spectrum communications—a tutorial,” IEEE Trans. Commun., vol. COM-30, no. 5, pp. 855–884, May 1982.
[20] K. Pelechrinis, M. Iliofotou, and S. Krishnamurthy, “Denial of service attacks in wireless networks: The case of jammers,” IEEE Commun. Surveys Tuts., vol. 13, no. 2, pp. 245–257, May 2010.
[21] M. Grissa, B. Hamdaoui, and A. A. Yavuzu, “Location privacy in cognitive radio networks: A survey,” IEEE Commun. Surveys Tuts., vol. 19, no. 3, pp. 1726–1760, 3rd Quart., 2017.
[22] A. Ahmad, S. Ahmad, M. H. Rehmani, and N. U. Hassan, “A survey on radio resource allocation in cognitive radio sensor networks,” IEEE Commun. Surveys Tuts., vol. 17, no. 2, pp. 888–917, 2nd Quart., 2015.
[23] S. Airabuee, A. Agarwal, D. Anand, and M. Khasawneh, “Game theory for security in cognitive radio networks,” in Proc. Int. Conf. Adv. Mobile Netw., Commun. Appl., Aug. 2012, pp. 60–63.
[24] Z. Song, X. Wang, Y. Liu, and Z. Zhang, “Joint spectrum resource allocation in nona-based cognitive radio network with SWEPT,” IEEE Access, vol. 7, pp. 89594–89603, 2019.
[25] B. R. Deepak, P. S. Bharathi, and D. Kumar, “Radio frequency anti-jamming capability improvement for cognitive radio networks: An evolutionary game theoretical approach,” in Proc. 4th Int. Conf. Signal Process., Commun. Netw. (ICSCN), Mar. 2017, pp. 1–6.
[26] J. He, C. Chen, S. Zhu, B. Yang, and X. Guan, “Anti-jamming game framework for secure state estimation in power systems,” IEEE Trans. Ind. Informat., vol. 15, no. 5, pp. 2628–2637, May 2019.
[27] I. K. Ahmed and A. O. Fapojuwo, “Stackelberg equilibria of an anti-jamming game in cooperative cognitive radio networks,” IEEE Trans. Cognit. Commun. Netw., vol. 4, no. 1, pp. 121–134, Mar. 2018.
[28] M. Lichtman, J. D. Poston, S. Amuru, C. Shahriar, T. C. Clancy, R. M. Buehrer, and J. H. Reed, “A communications jamming taxonomy,” Proc. IEEE, vol. 106, no. 1, pp. 47–54, Jan. 2018.
[29] Q. Zhou, Y. Li, and Y. Niu, “A countermeasure against random pulse jamming in time domain based on reinforcement learning,” IEEE Access, vol. 8, pp. 97164–97174, 2020.
[30] N. Van Huynh, D. T. Hoang, D. N. Nguyen, and E. Dutkiewicz, “DeepFaka: Deep duel-based deception strategy to defeat reactive jammers,” 2020, arXiv:2005.07034. [Online]. Available: https://arxiv.org/abs/2005.07034.
[31] D. Yang, G. Xue, J. Zhang, A. Richa, and X. Fang, “Coping with a smart jammer in wireless networks: A Stackelberg game approach,” 2020, arXiv:1907.07982. [Online]. Available: https://arxiv.org/abs/1907.07982.
[32] M. A. Aref and S. K. Jayaweera, “A cognitive anti-jamming and interference-avoidance stochastic game,” in Proc. IEEE 16th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC), Jul. 2017, pp. 520–527.
[33] J. Baek, S. J. Han, and Y. Han, “Energy-efficient UAV routing for wireless sensor networks,” IEEE Trans. Veh. Technol., vol. 69, no. 2, pp. 1741–1750, Feb. 2020.
[34] C. Chen, M. Song, C. Xin, and J. Backens, “A game-theoretical anti-jamming scheme for cognitive radio networks,” IEEE Netw., vol. 27, no. 3, pp. 22–27, May 2013.
[35] L. Jia, Y. Xu, Y. Sun, S. Feng, and A. Anpalagan, “Stackelberg game approaches for anti-jamming defence in wireless networks,” IEEE Wireless Commun., vol. 25, no. 6, pp. 120–128, 2018.
X. Wang, J. Wang, Y. Xu, J. Chen, L. Jia, X. Liu, and Y. Yang, “Dynamic

S. Misra, A. Mondal, P. Bhavathankar, and M.-S. Alouini, “M-JAW:

S. Mneimneh, S. Bhunia, F. Vázquez-Abad, and S. Sengupta, “A game-

K. Ibrahim, I. M. Qureshi, A. N. Malik, and S. X. Ng, “Bandwidth-efficient

G. K. Kurt and Ö. Cepheli, “Physical layer security of cognitive IoT net-

W. Xu, T. Wood, W. Trappe, and Y. Zhang, “Channel surfing and spatial

L. Xiao, J. Liu, Q. Li, N. B. Mandayam, and H. V. Poor, “User-centric view

G. Han, L. Xiao, and H. V. Poor, “Two-dimensional anti-jamming communi-

S. Misra, A. Mondal, P. Bhavathankar, and M.-S. Alouni, “M-JAW:

F. Yao, L. Jia, Y. Sun, X. Xu, S. Feng, and Y. Zhu, “A hierarchical learning

J.-F. Huang, G.-Y. Chang, and J.-X. Huang, “Anti-jamming rendezvous

G. Dubosarski, S. Primak, and X. Wang, “Multichannel power allocation

K. Ibrahim

IEEE Access, vol. 8, pp. 33479–33488, 2020.

IEEE Trans. Inf. Forensics Secur., vol. 10, no. 12, pp. 2578–2590, Dec. 2015.

IEEE Trans. Mobile Comput., vol. 15, no. 9, pp. 2247–2259, Apr. 2016.

IEEE Trans. Mobile Comput., vol. 13, no. 8, pp. 1746–1759, Aug. 2014.

IEEE Wireless Commun. Mag., vol. 98, no. 1, pp. 1646–1656, Feb. 2018.

IEEE J. Sel. Areas Commun., vol. 30, no. 1, pp. 4–15, Jan. 2012.

IEEE Trans. Wireless Commun., vol. 17, no. 8, pp. 5314–5326, Aug. 2018.

IEEE Trans. Wireless Commun., vol. 16, no. 3, pp. 648–661, Mar. 2017.

IEEE Trans. Mobile Comput., vol. 13, no. 4, pp. 689–703, Apr. 2011.

IEEE J. Sel. Areas Commun., vol. 29, no. 4, pp. 297–309, Feb. 2011.

IEEE Trans. Inf. Forensics Secur., vol. 9, no. 6, pp. 101–123, Sep. 2014.

IEEE Trans. Mobile Comput., vol. 71, no. 9, pp. 3237–3257, Sep. 2015.

IEEE Access, vol. 6, no. 7, pp. 818–821, Sep. 2015.

IEEE J. Sel. Areas Commun., vol. 25, no. 1, pp. 201–213, Jan. 2019.

IEEE Trans. Mobile Comput., vol. 16, no. 3, pp. 648–661, Mar. 2017.

IEEE Trans. Veh. Technol., vol. 58, no. 2, pp. 79–85, Feb. 2019.

IEEE J. Sel. Areas Commun., vol. 30, no. 1, pp. 1–5, Dec. 2011.

IEEE Trans. Veh. Technol., vol. 69, no. 6, pp. 818–821, Mar. 2020.

IEEE Trans. Veh. Technol., vol. 68, no. 2, pp. 1646–1656, Feb. 2018.

IEEE Trans. Inf. Forensics Secur., vol. 7, no. 1, pp. 297–309, Feb. 2011.

IEEE Trans. Veh. Technol., vol. 60, no. 7, pp. 3386–3407, Sep. 2011.

IEEE J. Sel. Areas Commun., vol. 8, pp. 33479–33488, 2020.

IEEE Trans. Inf. Forensics Secur., vol. 8, pp. 33479–33488, 2020.

IEEE Trans. Inf. Forensics Secur., vol. 9, no. 6, pp. 101–123, Sep. 2014.

IEEE Trans. Inf. Forensics Secur., vol. 7, no. 1, pp. 297–309, Feb. 2011.

IEEE Trans. Inf. Forensics Secur., vol. 17, no. 8, pp. 5314–5326, Aug. 2018.

IEEE J. Sel. Areas Commun., vol. 30, no. 1, pp. 4–15, Jan. 2012.

IEEE Trans. Wireless Commun., vol. 29, no. 4, pp. 297–309, Feb. 2011.

IEEE Trans. Mobile Comput., vol. 13, no. 4, pp. 689–703, Apr. 2011.
SOON XIN NG (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electronic engineering and the Ph.D. degree in telecommunications from the University of Southampton, Southampton, U.K., in 1999 and 2002, respectively. From 2003 to 2006, he was a Postdoctoral Research Fellow working on collaborative European research projects known as SCOUT, NEWCOM, and PHOENIX. Since August 2006, he has been a member of Academic Staff with the School of Electronics and Computer Science, University of Southampton, where he is currently an Associate Professor of telecommunications. He is involved in the OPTIMIX and CONCERTO European projects as well as the IU-ATC and UC4G projects. His research interests include adaptive coded modulation, coded modulation, channel coding, space-time coding, joint source and channel coding, iterative detection, OFDM, MIMO, cooperative communications, distributed coding, quantum error correction codes, and joint wireless-and-optical-fibre communications. He has published over 200 articles and coauthored two John Wiley/IEEE Press books in this field. He is also a Chartered Engineer and a fellow of the Higher Education Academy, U.K.

IJAZ MANSOOR QURESHI received the B.E. degree in avionic engineering from NED University, Karachi, Pakistan, the M.S. degree from the Department of Electrical Engineering, METU, Ankara, Turkey, in 1980, and the Ph.D. degree from the University of Toronto, Canada, in 1985. He has worked as a Professor with various universities in Pakistan, including Quaid-i-Azam University, from 1987 to 2002; Mohammad Ali Jinnah University, from 2002 to 2007; and the International Islamic University (IIU), from 2007 to 2009. Since 2009, he has been working as a Professor and an Incharge of graduate program with the Department of Electrical Engineering, Air University, Islamabad, Pakistan. He is also the Director of the Institute of Signals, Systems and Soft Computing (ISSS), Islamabad. He has around 300 publications in the various fields of engineering. His research interests include digital communication, radar signal processing, image processing, soft computing, and engineering computational mathematics.

AQDAS NAVEED MALIK received the Ph.D. degree from Muhammad Ali Jinnah University, Pakistan, in 2006. He has more than 15 years post Ph.D. experience in teaching and research at different intuitions of good repute in Pakistan. More than three Ph.D. students have been produced in his supervision. He is currently the Vice President of Higher Studies & Research (HS&R) and the Dean of the Faculty of Engineering and Technology, International Islamic University, Islamabad, Pakistan. His research interests include digital/wireless communications, digital signal processing, and information and power electronics.

SAMI MUHAIDAT (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2006. From 2007 to 2008, he was an NSERC Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Canada. From 2008 to 2012, he was an Assistant Professor with the School of Engineering Science, Simon Fraser University, BC, Canada. He is currently a Professor at Khalifa University and an Adjunct Professor with Carleton University, ON. His research interests include advanced digital signal processing techniques for wireless communications, RIS, 5G and beyond, MIMO, optical communications, the IoT with emphasis on battery-free devices, and machine learning. He served as a Senior Editor and an Editor for the IEEE COMMUNICATIONS LETTERS, an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, and an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He is an Area Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and a Lead Guest Editor of the IEEE Open Journal of the Communications Society “Large-Scale Wireless Powered Networks with Backscatter Communications” Special Issue.