Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7125482, 16 pages
https://doi.org/10.1155/2021/7125482

Research Article

A Deep Learning-Based Power Control and Consensus Performance of Spectrum Sharing in the CR Network

Muhammad Muzamil Aslam,[1] **Liping Du**,[1,2] **Zahoor Ahmed**,[3,4] **Muhammad Nauman Irshad**,[1] **and Hassan Azeem**[1]

[1]*School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China*
[2]*Shunde Graduate School of University of Science and Technology Beijing, Foshan 528000, China*
[3]*Department of Automation, Shanghai Jiaotong University, Shanghai 200240, China*
[4]*Department of Electronics, GC University Lahore, 54000, Pakistan*

Correspondence should be addressed to Liping Du; dlp2001@ies.ustb.edu.cn

Received 21 December 2019; Revised 5 November 2020; Accepted 23 January 2021; Published 19 February 2021

Academic Editor: Longzhe Han

The cognitive radio network (CRN) is aimed at strengthening the system through learning and adjusting by observing and measuring the available resources. Due to spectrum sensing capability in CRN, it should be feasible and fast. The capability to observe and reconfigure is the key feature of CRN, while current machine learning techniques work great when incorporated with system adaption algorithms. This paper describes the consensus performance and power control of spectrum sharing in CRN. (1) CRN users are considered noncooperative users such that the power control policy of a primary user (PU) is predefined keeping the secondary user (SU) unaware of PU's power control policy. For a more efficient spectrum sharing performance, a deep learning power control strategy has been developed. This algorithm is based on the received signal strength at CRN nodes. (2) An agent-based approach is introduced for the CR user's consensus performance. (3) All agents reached their steady-state value after nearly 100 seconds. However, the settling time is large. Sensing delay of 0.4 second inside whole operation is identical. The assumed method is enough for the representation of large-scale sensing delay in the CR network.

1. Introduction

Intelligent processing is one of the major advantages of CRN [1]. Due to this feature, these systems possess the capability to learn their environment, increase awareness, and reconfigure themselves accordingly. High environmental awareness of these systems, due to their ability to perceive, gives them an enormous advantage in a wireless communication environment. Therefore, a device's wireless operations can be effectively adapted to its surroundings, and thus, it uses the maximum resource to deliver the best results. Some techniques, e.g., intelligent reflecting surface (IRS), for secure communication of multiantenna have been studied in [2] which contain reflector elements that are reconfigurable and controlled by software that is communication-oriented [3, 4]. In most cases, spectrum sensing is used to achieve perception capability in these systems [5]. Therefore, weak spectrum sensing ability would limit their operations and, as a result, decrease the efficiency. In this context, many researchers have proposed techniques to improve spectrum sensing [6, 7], which contain, but are not limited to, wideband spectrum sensing, cooperative spectrum sensing [8], and sequential spectrum sensing. It is a concept widely used for perception improvement in a wireless networking environment [9]. In this context, spectrum measurements, multidimensional spectrum sensing, and interference sensing [10] have been studied extensively.

Primary users' (PU) role is considered an active and passive user model in academia for spectrum sharing. There was a cooperative or noncooperative relationship between PU and SU in the active models [11]. To enhance the transmission performance of the system, information interaction

has been performed, although, in research, SU performs spectrum sensing in a passive user model for finding power allocation or idle spectrum. When there is a passive PU role, then PU assigns its transmit power relying on its power controlling scheme [12].

CRN transmit power selection must be properly controlled for the transmissions of the CR users [13, 14]. This is the factor that can help to achieve high efficiency of the spectrum through reuse of the spectrum bands of the power units which are affected by the PU limitations. For spectrum sensing, resources are required at all nodes [15]. Hence, effective development of spectrum sensing is of high significance to achieve optimal bandwidth, time, and power spent in between the transmission and sensing [16–18]. Intelligent [19] wireless communication devices must be adaptive to their environment to be able to perceive their surroundings. This can be achieved through the optimization of working limitations and dynamic spectrum access.

In recent years, fast developments in machine learning have been noted. Applied in different problems of wireless communication [20], both supervised and unsupervised learning [21] have been considered. Different from supervised learning, *Q*-learning or reinforcement learning [22] is found to be valuable to maximize the long-term performance of a system and to achieve a balance between exploitation and exploration. Furthermore, deep learning has proven to be a predominant method of achieving sound and higher performance, which is unaffected by massive data sets, and also seems promising in wireless communication [22]. In this context, reinforcement learning is a machine learning technique inspired by biological phenomena, where knowledge is acquired to an agent by repeated trial-and-error communication with its surroundings. Environment feedback to an agent's action is used to optimize the machine for future behaviour. Dynamic interaction between the agent and its surroundings and the resulting response are two of the key topographies which make reinforcement learning techniques appealing for cognitive radio ad hoc network applications, mostly for the tasks of routing and spectrum decision [23]. In several cases, operations based on reinforcement learning are better than traditional solutions [24].

For better performance of CRN, awareness about surroundings (radiofrequency (RF)) is a must for CR. CR should sense and check all around the location for keeping an eye on the RF activities. There was a basic identification of spectrum sensing in CRs. Some sensing techniques, e.g., energy detection, based on a matched filter and cyclostationary process, have been studied in [25]. For better results of sensing accuracy, cooperative sensing was studied in [26] to focus on wireless network hidden terminal problems. Recently researchers studied cooperative CRS [27].

Cooperative communication is an advanced communication-in-future technology that optimizes signal transmissions w.r.t both medium contact control and physical coating as presented in [28]. There is an important role of *Q*-learning in CR application, for instance, dynamic spectrum access. Users iteratively improve their planning to attain their tasks through knowledge and recompense from their surroundings. In these days, there are several research techniques to solve the problems of power control and power allocation [29], for example, game theory [30], optimization theory [31], and machine learning; among these, deep reinforcement learning which is a kind of machine learning has got a lot of attention because of its fast speed over the complex problem. The contribution of its challenging ability in many applications can be found such as Atari Games [32]. Agents [15] that are trained through reinforcement learning are intelligent enough to absorb their act value policies from high-dimensional rare data. An example of such applications would be videos or images [33] also shown in our tentative result.

Deep ReLU (rectified linear unit) learning can assist in learning an operative action charge policy even when the state notes are contaminated by measurement errors or arbitrary noise. Using deep reinforcement learning power control-based spectrum sharing in wireless networks showing the relation between PU and SU, wireless sensors have been studied in [34] and channel selection policy for PU and SU has been studied in [35]. On the other hand, the conventional ReLU method is unfeasible for such issues because of an inadequate number of states in the presence of arbitrary noise. Therefore, deep reinforcement learning is appropriate for wireless communication applications, where municipal dimensions are usually arbitrary. In this paper, we described the consensus performance [4, 36] and the power control of spectrum sharing in CRN. We also introduced a sensing delay. In CR, power control is investigated for spectrum sharing to assure the quality of service for PU and SU. First of all, CRN users are considered noncooperative users such that the power control policy of primary users (PUs) is predefined, while the secondary user (SU) is unaware of PU's power control policy. Secondly, a model was constructed for cooperation between wireless sensors, PU, and SU. In this model, wireless sensors' received signal strength is spatially distributed to assist PU for power transmission information with SU. The performance criterion of sharing in CRN has been improved in the presence of sensing delay in a communication network. The studied model can get minibatch updates for several iterations. The network becomes able to converge quickly and meet a user's quality of services. The overall contributions of this work or difference with other works can be summarized as follows:

(i) The proposed method is different from the traditional method in the sense that there is a performance criterion of sharing, and consensus of CRN has been improved in the presence of sensing delay (communication delay) through the communication topology

(ii) An agent-based approach is introduced for the CR user's consensus performance for the first time

(iii) Besides, a deep learning power control strategy (reinforcement learning) has been developed along with an agent-based approach altogether for a more efficient performance of sharing control

In the rest of the paper, we discussed the following: overview of CRN, intelligent power control of spectrum sharing,

delay performance using a primary sensor network, and CRN consensus criteria under communication delay.

2. An Overview of CRN

We suppose a CRN which contains licensed users and unlicensed users; in this, the unlicensed user is willing to segment spectrum resources which are common with the licensed user, without any danger to harm the licensed user activity. The licensed user contains the primary transmitter (P_{T1}) and receiver P_{R1}; similarly, the unlicensed user consists of a secondary transmitter S_{T2} and receiver S_{R2}. The receiver and transmitter of both licensed and unlicensed users are shown in Figure 1 in a multicluster network number of sensors that are present for direct communication with the central station; the central station can be the Cluster Head (CH), the name cluster is because of network similarity with neural network. In our supposed work, there is no working cooperation between the licensed and unlicensed users, where a licensed user does not know the presence of the unlicensed user in the same surroundings even licensed users adjust its conduct power built on its particular power policy. The next step of the licensed user will be affected in a couched way by the sudden action of the unlicensed user; it should be highlighted that for the licensed user, the power control policy depends on the network of licensed users and the number of sensors used. Yet licensed and unlicensed users are without communication, and both have no information about the power transmitter and power control policy of each other. For an easy understanding, we consider that licensed and unlicensed users synchronously update their transmit power, where it is based on the time frame base.

Here, in terms of SNR_c, it is used for showing the quality of the system of licensed and unlicensed users. ∂_a and ∂_b show the transmit power of licensed and unlicensed users. The receiver value of SNR_c is given as

$$\mathrm{SNR}_c = \frac{|H_{xx}|^2 \partial_a}{\sum_{y \neq x} |H_{yx}|^2 \partial_b + N_x}, \quad \forall\, x = 1, 2. \tag{1}$$

In Equation (1), H_{xy} is the frequency expansion from P_{T1} (primary transmitter) to S_{T2} (secondary transmitter), while N_x is the noise power from P_{R1} (primary receiver). Here, the purpose is to assist the unlicensed users in acquiring sufficient control power policy, so that after some time of power adjustment, both licensed and unlicensed users are intelligent for the successful transmission of data with the essential quality of services. If unlicensed or licensed users are only aware of their transmit power, then such a task is complicated. To assist unlicensed users, in a wireless communication network using different locations, sensor nodes are hired for the dimension of received signal strength (RSS). Received signal strength measurements are related to both (licensed and unlicensed) users' transmit power, close-fitting the information system from the state. Here, we suppose that RSS information is helpful for the unlicensed user. Zigbee technology [7] and CRNs generally function with different frequencies, and there is no interference in transmission because of sensor node transmission in CRNs. Tocollect the received signal strength information from spatially scattered sensors, node is a basic condition, for example, source localization [17]. In the proposednumerical method, only once per time frame is needed to report RSS informationwhichconsists of low data rate.

We consider P_{R1} and S_{R2}; here, S_{R2} mollify the lowest SNR_c aimed at an effective response. For getting the quality of system condition, the licensed operator or user is considered to deceptively regulate its power transmission that is power control policy-based. We used "2" controlling power policies that supposed for licensed users, even if the licensed user (PU) adopts any other power control policy, our supposed method will also work. For the first policy, the transmit power of the licensed user is updated following the classical power control algorithm. Transmit power is adjusted on a time framed bases K(x) that maps the continuous values level, known as discretization operation.

$$\partial_a(r+1) = K\left(\frac{\omega'_c \partial_a(r)}{\mathrm{SNR}_c(r)}\right). \tag{2}$$

Here, $\mathrm{SNR}_c(r)$ is the SNR_c measured at P_{R1} at the r^{th} time frame. $\partial_a(r)$ is the transmit power at the r^{th} time frame.

Now, we explain the PU policy of our supposed work. Let us see

$$\partial_a(r+1) = \begin{cases} \partial^{\partial}_{y+1}, & \text{if } \partial_y{}^{\partial} \leq \tilde{\eta} \leq \partial^{\partial}_{y+1} \text{ and } y+1 \leq E_1, \\ \partial^{\partial}_{y-1}, & \text{else } \tilde{\eta} \leq \partial^{\partial}_{y-1} \text{ and } y-1 > 1, \\ \partial^{\partial}_{y}, & \text{otherwise,} \end{cases} \tag{3}$$

Here, $\tilde{\eta} := \acute{\omega}_c\, \partial_a(k)/\mathrm{SNR}_c(r)$; compared with Equation (2), Equation (3) has more unadventurous behaviour for power control policy. Transmit power is updated step by step. The power is increased by a single step when $\mathrm{SNR}_c(r) \leq \acute{\omega}_c$ and $\acute{\omega} \geq \acute{\omega}_c$; also, the power is reduced by a single phase, when $\mathrm{SNR}_c(r) \geq \acute{\omega}_c$ and $\acute{\omega} \geq \acute{\omega}_c$. If not increasing or decreasing, then it will stay at the present power level. And here, $\acute{\omega} := \mathrm{SNR}_c(r)\, \partial_a(r+1)/\partial_a(r)$ is the forecast SNR at the $(r+1)$ time frame. Now,

$$P_1 \approx \left\{ \partial^{\partial}_a \cdots \partial^{\partial}_{E_1} \right\}. \tag{4}$$

Here, $\partial^{\sigma}_a \leq \cdots \partial$. Furthermore, precisely we suppose that $K(x)$ is very equal to a discrete step, which is not lesser than x, and suppose that $k(x) = \partial$ if $x > \partial$. Now, assume the second power control policy and consider that the transmit power at the r^{th} time frame is $\partial_a(r) = \partial^{\sigma}_y$, where $\partial^{\sigma}_y \in P$. Equation (3) shows the transmit power of PUs updating.

A suggested control channel is used to notify the sensors about the available users. Scheming a CRN without a suggested control user can be studied in [9]. Sensor node switch between various users depends upon the available user; also,

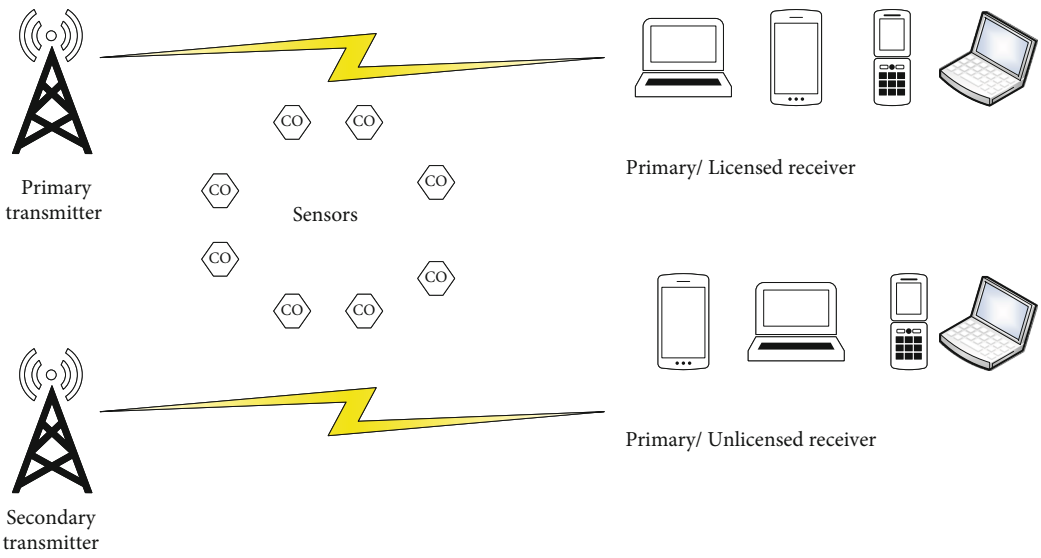


Figure 1: Spectrum sharing in CRNs (sharing the same channel).

there is a possibility of the sensor to operate on different users. Further info about this can be found in [30]. A multi-radio wireless sensor network has been explained in some literature which can be found in [18]. In this research article, consider that N sensors are installed to sample the received signal strength info. Consider that S_n represents node n and $\partial_n^m(r)$ represents the receiving power at the r^{th} frame at sensor n. We simulate the state of received signal strength observations.

And now, the SU policy is explained in the form of the following equation:

$$\partial_n^m(r) = \partial_a(r)J_{cn} + \partial_b(r)J_{dn} + L_n(r). \tag{5}$$

In Equation (5), $\partial_a(r)$ represents the transmit power of PUs, $\partial_b(r)$ represents the transmit power of SUs, J_{cn} is the path loss between sensor n and the primary transmitter, J_{dn} is the path loss between the unlicensed transmitter and sensor n, and $L_n(r)$ is the zero mean Gaussian random variables with variance φ_n^2.

Here, according to free-space broadcast, with regard to the Friis law, J_{cn} and J_{dn} are given below:

$$\begin{aligned} J_{cn} &= \left(\frac{\gamma}{4\pi D1_n}\right)^2, \\ J_{dn} &= \left(\frac{\gamma}{4\pi D2_n}\right)^2. \end{aligned} \tag{6}$$

Usually, the Friis equation is used for the calculation of ideal power which was received at an antenna from elementary info about the transmission. Here, "γ" is the wavelength of the signal $D1_n$ and $D2_n$ showing detachment between nodes and primary and secondary transmitters. The Friis equation is used only for the calculation of single frequency, even though transmission characteristically contained many. Here, also to suppose, from a finite set, we choose the transmit power of the unlicensed user:

$$P_2 \coloneqq \left\{\partial_a^m \cdots \partial_{E_2}^m\right\}, \tag{7}$$

where $\partial_a^m \leq \cdots \leq \partial_{E2}^m$; the SU goal is to see how power transmission can be settled depending on the received signal strength info $\{\partial_n^m(r)\}_n$; after rare rounds of power transmission at every time frame, both PU and SU can see their respective quality of system necessities. At least transmit power is present in our supposed work, which transmits power $\{\partial_{s_1}^{\sigma}, \partial_{s_2}^m\}$; here, P_{R1} and S_{R2} mollify their quality of system SNR necessities.

2.1. Intelligent Power Control of Spectrum Sharing. The distribution of system time is uniform, denoted as channel switching intermission. The presently employed channel develops unattainable culmination of the cooperative spectrum interval. When the sensor senses the second user, the channel might delay till the start of the next cooperative spectrum pause (if at least one user is available). In that case, both the user and sensor jump to the new channel. This process is termed as periodic switching (PS). Another cause may be that the central head may highlight the sensor if a new channel is available, as the prior channel is misplaced; this channel loss is stated as triggered switching (TS). The number of channel switching may be zero or one for periodic switching (PS); it depends whether the channel is present or not at the start of the cooperative spectrum interval. For a satisfactory transmission, the central station sends back an Automatic Control Key (ACK) to the sensors. If the sensor does not receive an ACK after a data packet transmission, it is assumed that the present channel is unavailable, and the transmission is stopped immediately. However, this kind of failure may happen due to channel fading which would cause a transmission failure on the CR sensor net. It is not a good practice to stop the transmission in this case, which would instigate needless

interference with the PN (Primary Network). Therefore, BE (Best-Effort) data traffic and actual-time data traffic can be assisted in this case. A good example of this channel switching and traffic resource allocation can be found in [2].

In the proposed work, SUs are assumed to take any action at each time step, such as for choosing transmit power from the earliest quantified power set P_2 based on its present state. To analyse this, the Markov decision process (MDP) [37] is considered. Particularly, for instance, consider an action $a(r) = \partial(r+1)$ by SU in the state $m(r)$. Then, the next new state $m(r+1)$ is dependent on the present state $m(r)$ and action of a decision-maker $a(r)$. Here, $m(r)$ and $a(r)$ are known. The following states will be independent of all earlier states and actions with the condition that

$$m(k) \approx \left[\overset{\vee}{\partial_a^n}(r) \cdots \overset{\vee}{\partial_N^n}(r) \right]^T. \tag{8}$$

For a new state, a corresponding reward $n(r) := n(m(r))$ is assumed for the decision-maker $a(r)$ which can be mathematically expressed as

$$\overset{\vee}{n}(r) \approx \begin{cases} z, & \text{if } \mathrm{SNR}_c(r+1) \geq \mathrm{SNR}_d(r+1) \geq \dot{\omega}_d, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In this work, "z" is assumed to be 10. It can be seen through simulation that the value of z should be sufficiently large to avoid the learning disturbance. Communication between a SU, a PU, and the sensors is shown in Figure 2(a). It is to be highlighted that the unlicensed user is supposed to distinguish whether communication between P_{T1} and P_{R1} is successful or not. This kind of information can be gained to monitor an acknowledged signal transferred P_{R1} for an indication of successful transmission from P_{T1}. Figure 2(b) [15] shows the detached control architecture and is an object control system design in which there is a linear time-invariant (LTI) and multiple input-multiple output (MIMO) system.

The basic problem for the Markov decision process is to understand the policy for the decision-maker.

$$U^{\pi}(m(r)) \coloneqq \sum_{a=r}^{T-1} \gamma^{a-r} \overset{\vee}{n}(a), \tag{10}$$

where π stipulates the action $\pi(m)$ from the decision-maker. It is important to say that the purpose of SU is to understand a strategy π for the selection of its action $a(r)$ which is constructed on the present state $m(r)$ in a method that discounted cumulative reward reduced maximum as studied in [38]. In Equation (10), "γ" is the discount factor and T^{-1} is representing the time frame, at which the targeted state is touched. In this case, the targeted state is well defined, where $\mathrm{SNR}_c(r) \geq \dot{\omega}_i$ for $i = 1, 2$. The target is for studying optimal policy π^* that maximizes U^{π}, i.e.,

$$\pi^* = \arg\max U^{\pi}(m), \quad \forall m, \tag{11}$$

It is complicated to directly learn π^*. In the case of reinforcement learning, there is an alternative approach to solve Equation (11) through Q-learning. This alternative approach has been studied in [13]. Q-function (action value) is defined to determine the predictable reduced growing reward after the action $a(k+n)$, for state m. In such cases, the optimal policy would be built by choosing the action by the largest value in each state.

By the updated Bellman equation,

$$Q(m, a) = \overset{\vee}{n}(m, a) + \gamma \max\ Q\left(m', a'\right) m \tag{12}$$

The elementary purpose of Q-learning and several other reinforcement learning algorithms is to iteratively update the action value according to the value updating instruction. In Equation (12), "m'" is the state followed by smearing of action "a" to the present state "m." The value iteration algorithm defined in Equation (12) approaches the best action value purpose. The number of states is limited for Q-learning, and at each state, the action value function is guessed separately. The Q-table is designed such that rows represent the states and columns represent the probable action state. One can choose an action from the Q-table as the maximum value of $Q(m, a)$ through an optimal action state. However, the value of state function m is continuous due to arbitrary variation in the received signal. Therefore, the Q-learning method is unfeasible in this work since it is not convenient to consider an unlimited figure of states. To overcome this issue, DQN (Deep Q-Network) is opted. A brief study and proposed work of DQN are presented in [39].

Different from the traditional Q-learning technique, which is used to generate unlimited action value functions, "ϑ" is used here to express weights of Q-learning, while the Q-table is substituted by the deep neural network $Q(m, a : \vartheta)$ that would act as the action value function. Particularly, for an input "m," DQN produces an E_2-dimensional vector for selecting action $a = \mathcal{A}_i^m$ from $\mathcal{A}_2$. When $m(r)$ is given, the data used for Q-network training is generated as follows: we explore amorphous selected action with ϵ_k having the largest output $Q(m(r), a(r): \vartheta_0)$, where ϑ_0 is used for denotation of the present iteration. Subsequently, the action an (r), SU gets a reward Υ(r), a new state$m(r+1)$is obtained. This transit, $d(r) := \{m(r), a(r), m(r), m(r+1)\}$, is held in the reiteration memory K. When enough transitions are collected in K, the training of the Q-network begins. Say $O =$ 300 evolutions, we casually choose a minibatch of transitions $\{d(i) i \in \Omega_r\}$ from K, and with adjustment of ϑ, DQN can be trained as given loss function is minimized:

$$E(\vartheta) \approx \frac{1}{\Omega_r} \sum_{i \varepsilon \mho_r} Q'(i) - Q\left(m(i), \dddot{a}(i); \vartheta\right)^2 m. \tag{13}$$

Ω_r is the index set of the irregular minibatch used at the r^{th} iteration. The estimated value of the Bellman equation $Q'(i)$ for the i^{th} iteration is given by

$$Q'(i) = \overset{\vee}{n}(i) + \gamma \max\ Q\left(s(i+1), \ddot{a}; \vartheta_0\right), \quad \forall i \varepsilon \Omega_r m. \tag{14}$$

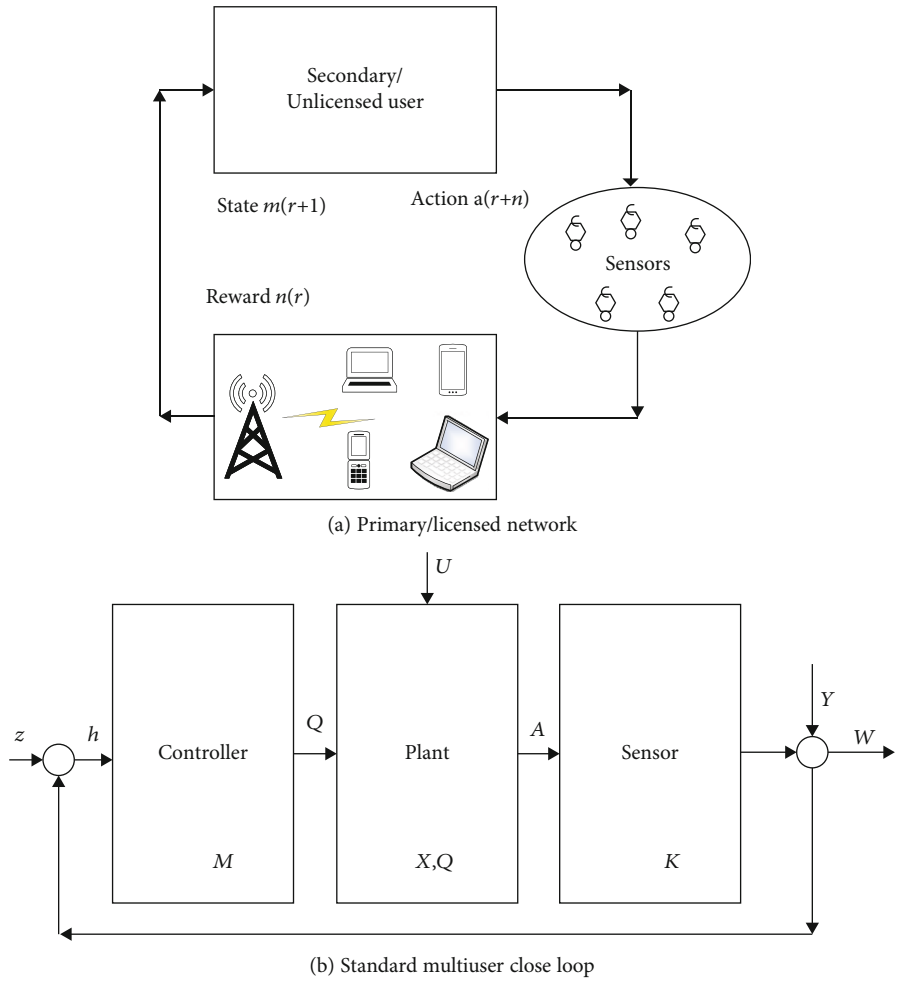


Figure 2: Unlicensed user and licensed network interactions and standard multiuser close loop.

Different from the conventional supervised learning, the goal for DQN knowledge is different from that of the traditional approach.

SU can select action after training, which gives a maximum projected value $Q(m, a, \vartheta^*)$. For simplicity, the DQN scheme is summarized in the power control scheme for SU. It would be clearer to highlight that during the training process of DQN, an SU needs the knowledge that satisfies the QoS (quality-of-service) necessities for the PU, and therefore, the SU is mollified. However, the only purpose of SU, after DQN training completion, is feedback from the sensors for the decision of the next transmit power. The junction problem of the supposed policy for power control is hence studied. Let "m" be a target state; then, if SU transmit power stays unchanged, it will be informal to say that the next state m' is also a goalmouth state. Transmit power can result from either Equation (2) or (3), and as a result, PU will be updated. In another way, SU will ultimately acquire to select a diffuse power as m' persists the goalmouth state. It can be concluded that once s achieves the target state, it will be in the goalmouth state till broadcast information is done. Now assume that there is a possibility of data loss because of the irregularity of data transmission and SU aims to take up for an innovative transmission; no long learning is required in this case. Rendering to control power policy, SU can select transmit power.

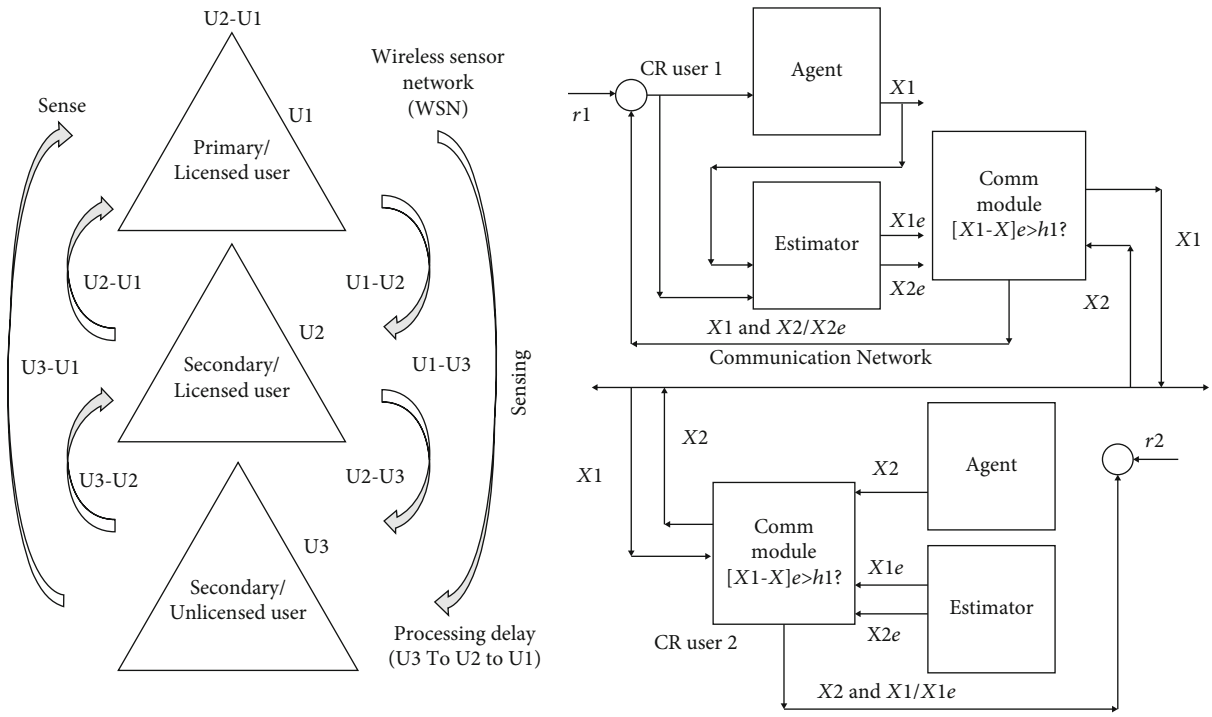
Figure 2(b) shows a standard multiuser closed loop representing power control in this case. Standard discrete time is considered for control system design. For a multi-input multioutput system called the MIMO system with linear time-invariant known as LTI, there are z input, w inputs, and h states.

$$\begin{aligned} Z(q+1) &= X_A(l) + Qv(l) + W(l)m, \\ C(q) &= K_A(q) + H(q)z, \end{aligned} \tag{15}$$

where "q" represents the time index linked with T time sampling in the discrete time domain. $X \in , K \in \ddot{R}^{z*n}$ and $\ddot{R}^{n*n}$, $Q \in \ddot{R}^{n*m}$ are output matrices and system input, and $W(.)$ and $H(.)$ are system measurements and disturbances, respectively. $U(.) = [u1(.), \cdots, uf(.)]^T$ and $Y(.) = [y1(.), \cdots, yz(.)]^T$ are expected to be barred, for example, $|ui(.)| \leq g_{iu}$, $i = 1, \cdots, f$ and $|yj(.)| \leq g_{jv}$, $j = 1, \cdots, z$, in which $g_{iu}{}^{'}s$ and $g_{jv}{}^{'}s$ are positive constant. For simplicity, all states are considered measurable, i.e., $K = I$. There is a possibility to design a

Initialize D with capacity O
Initialize network $Q(m, a, \vartheta)$ with variable weights $\vartheta = \vartheta_0$
Initialize $\mathcal{O}_a$ and $\mathcal{O}_b$ and then obtain m (1)
For $r = 1$, r do
Update $\mathcal{O}_a(r+1)$ via power control policy of PU (2) or (3)
With iterations ε_k, choose an arbitrary action $a(k)$; otherwise, select $a(k) = \max_a Q(m(r), a; \vartheta_0$
Obtain $m(r+1)$ via arbitrary model (5) and detect reward $n(r)$
Store transition $d(r) := \{m(r), a(r), m(r), m(r+1)\}$, in D
Sensing delay
Repeat sensing delay
If $r \geq O$ then
Sample random minibatch of iterations from $D\{d(i)i \in \Omega_r\}$,
Here, the Ω_r index is uniformly selected at independent
Minimize loss function of 12, in which goal is given by (15) $Q'(i)$
Adjust $\vartheta_0 = \arg\ \min_\vartheta E_\vartheta$
End if
$m(r)$ is the target state and then initialize $\partial_a(r+1)$ and $\partial_b(r+1)$ and then gain $m(r+1)$
end if
end for

Algorithm 1: Power control policy-deep learning training based.



(a) Sensing time delay

(b) Spectrum between users

Figure 3: Sensing delay (sensing time) and spectrum sharing between users.

feedback controller at state (1) through maximum input maximum output, which is shown as follows.

$$V(M) = M_k(q)m,$$
$$M[(M(q) - A(C) + H(q)]m. \tag{16}$$

That is, Figure 3(b) shows the spectrum sharing of the control system, which is well designed. So, the stability and performance of the system (z) could be definite through an approximate selection of V sampling time in Equation (16) and K as gain feedback design in Equation (15). A dynamic system is to be supposed (Equation (15)), there is a designed controller that works as a designed framework baseline.

In this paper, there is a consideration that the PU and SU update their power of transmission synchronously. However, we would like to highlight that our proposed scheme's synchronous assumption is not a must. Consider that there is

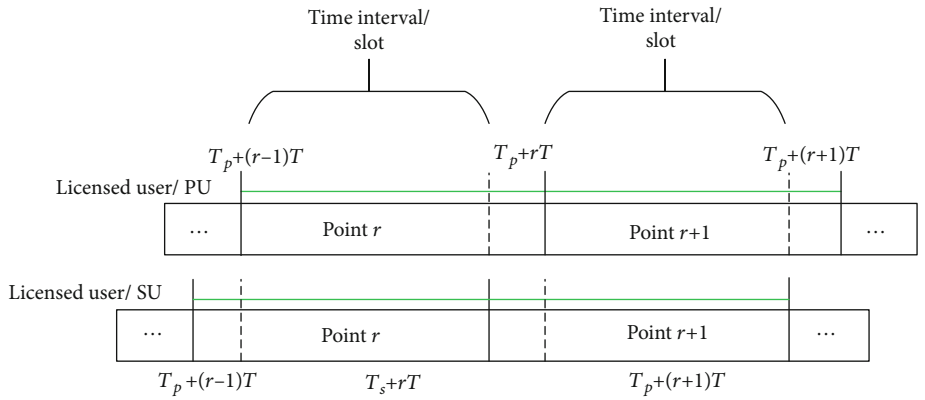


Figure 4: Licensed and unlicensed users updated transmit power.

no strict synchronous occurrence in the time frame between PU and SU; both synchronize transmit power at the start of a timeframe. Figure 4 represents the transmit power of a licensed and unlicensed user. PU adjusts its transmit power at time $t_p, t_p + T, t_p + 2T \cdots$, whereas SU's transmit power is updated at $t_s, t_s + T, t_s + 2T \cdots$, where T is representing the time of each setting. Deprived of loss simplification, we consider that $T > t_p - t_s > 0$. Our scheme, termed intelligent power control scheme, would work in matching through the in-synchronous form if both PU and SU achieve their respective goal. The essential info for PU can be $\mathrm{SNR}_c(r)$, and for SU, it can be $\mathrm{SNR}_c(r)$ and $\mathrm{SNR}_d(r)$, while $m(r)$ is the decision made during the time window $[t_p + (r-1)T.t_s + rt]$.

2.2. Delay Performance Using a Primary Sensor Network. A promising approach for CRN design can be the use of transmission data with the quality of services in wireless sensor networks. Designing a proper physical CRN, with effective and efficient spectrum allocation, sensing, and minimum potential interference, poses several challenges. Since license-free spectra are crowded in CRN, they are affected by uncontrolled interference, and hence, real-time traffic is recommended with the CRNs for the future. Figure 3(a) presents the sensing delay that occurred in the sensing network. It can be seen that there are several SUs on a single PU node. The sensing delay for a U2-U1 connection is less than that for a U3-U2-U1 connection. This is due to the extra sensing delay to sense and get information for an extra link.

The direct U3-U1 link is not obvious for the system. It checks the optimal link for U3 to connect with U1, and during that process, it depends on the response from U2. Similarly, the sensing delay will increase when there are several users on the primary node, and hence, information may be late.

Figure 3(b) represents a network of two cognitive radio users with a communication module. Here, we supposed that CR users use an estimated state for action and communicate its present state to the other user in case of an unsuccessful state. Both estimator 1 and estimator 2 estimate their corresponding $X1e$ and $X2e$. In a normal situation, there is no need for communication. CR user 1 is working based on its

Initialize $\mathcal{F}_b(1)$ and then obtain m (1)
for $r = 1, r$ do
Sensing delay
choose $a(k) = \max_a Q(m(r), a; \vartheta_0$
obtain $m(r+1)$
end for

Algorithm 2: Control power policy (DQN based).

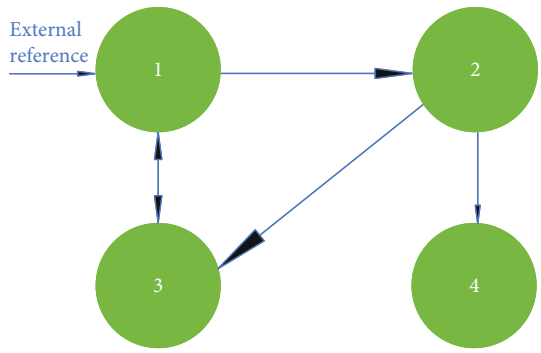


Figure 5: CR users in CRN.

own state $X1$, and CR user 2 estimated state $X2e$. When $X2$ is received from CR user 2, comm module 1 broadcasts X1 to CR user 2 if $|X1 - X1e| > h1$, where $h1$ is the defined threshold. Similarly, a communication mechanism and estimation are designed at CR user 2 and other CR users.

2.3. CRN Consensus Criteria under Communication Delay. Let transfer of message by the nearest agent m be received by agent n with a sensing delay T, which is similar to a network with a sensing delay of fixed one-hop communication. The algorithm of such consensus is given in the equation below.

$$\chi_m(t) = \sum_{n \varepsilon Z_m} h_{mn} \left(\chi_m \left(\tau - \dot{T} \right) - \chi_m \left(\tau - \dot{T} \right) \right) m. \quad (17)$$

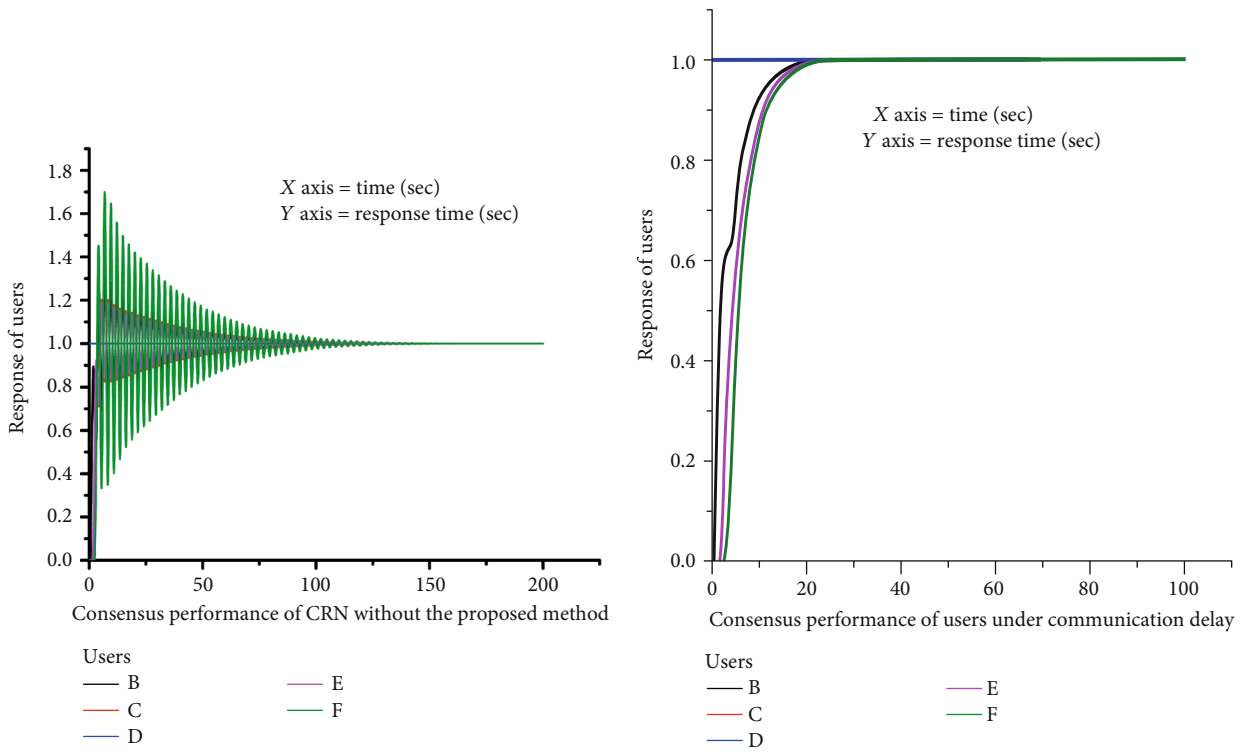


Figure 6: Consensus performance of CRN without the proposed method and with communication delay.

For getting an average consensus, a weight undirected graph G has been supposed in Equation (17). Therefore, the algorithm is

$$\chi_m(t) = \sum_{n\varepsilon Z_m} h_{mn}\left(\chi_n\left(\tau - \dot{T}\right) - \chi_m(\tau)\right)m. \tag{18}$$

The following form shows the collective dynamic of the network:

$$\chi(t) = -L_x\left(t - \dot{T}\right)m. \tag{19}$$

Taking Laplace to transform,

$$A(s) = \frac{W(s)}{s}\chi(0)m. \tag{20}$$

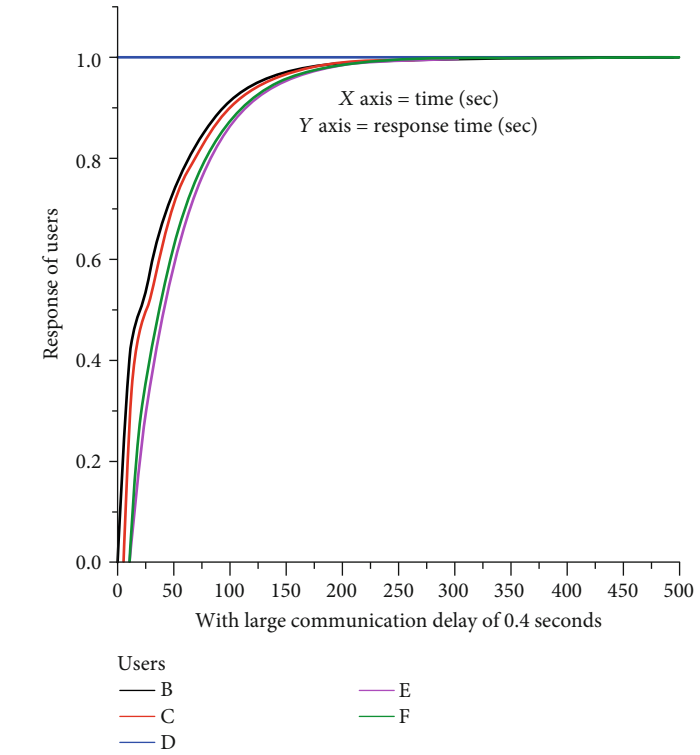
Here, function $W(s) = ((I_n + (I/s))\exp(-sT)L)^{-1}$. For stability verification of $H(s)$, the Nyquist criterion can be used. We can refer to the study as an approach that is similar in [40]. There is an upper bound sensing delay method for network stability that can maintain during time delay. The Equation (10) algorithm clarifies the problem of average consensus asymptotically for all early states or levels, if and only if $0 \le \dot{T} < \pi/2\gamma n$, where $\gamma n < 2\underline{\Delta}$. Refer to [36] for proof. A sufficient condition for the average consensus algorithm (17) is $\tau < \pi/4\underline{\Delta}$. There may be a trade-off between sturdiness to sensing time delay and having a large max degree. A network having a great degree is normally considered a free-scale network. To study [41], small networks and random graphs are impartially active for sensing delay, in the absence of several degrees. consensus achievement of a buildup engineering network is a good example.

2.4. Simulation. Suppose four CR users in CRN after their traffic management; the communication topology in their information sharing is shown in Figure 5.

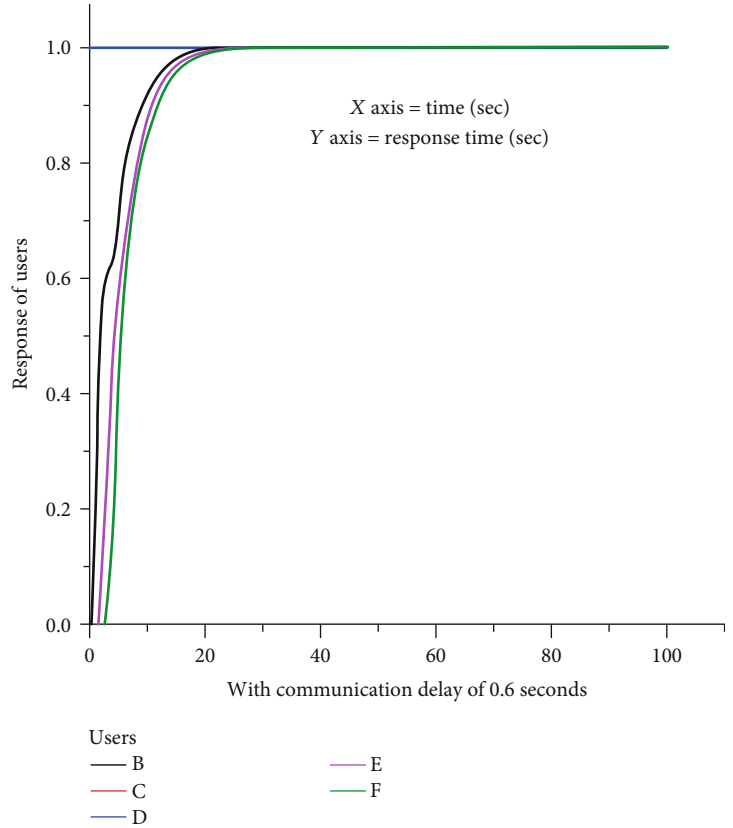
If we consider that all users have integrator dynamics, then by using $m_x = -L_x$, in which values of m changes regarding network dynamics $m_x(t)$, where $L = L_{xy}$ is the graph Laplacian of network, the performance without delay is shown in Figure 6(a). When delay has seemed to be in the network, it affects the performance of the CRN. Hence, consensus performance of the CRN could be seen in Figure 6(b) after communication delay without the proposed method. Although these responses are in consensus, their transient responses are very poor. Hence, the settling time is very large. So, for better performance, the designer should follow this proposed method to minimize the effects of time delay and sensing delay as well. Let a communication delay of 0.4 second between all users be the same; then, their consensus performance can be calculated by using the algorithm shown in Figure 7(a). This proposed method is also sufficient for large arbitrary communication delay of 0.4 sec. Thus, the consensus performance of the CR users with a large communication delay of 0.6 sec is shown in Figure 7(b).

3. Results and Discussion

The transmit power (in watts) for PU and SU is selected from a predefined set $P_c = P_d = \{1/20, 0.1/1, \cdots, 0.4/1\}$, and the

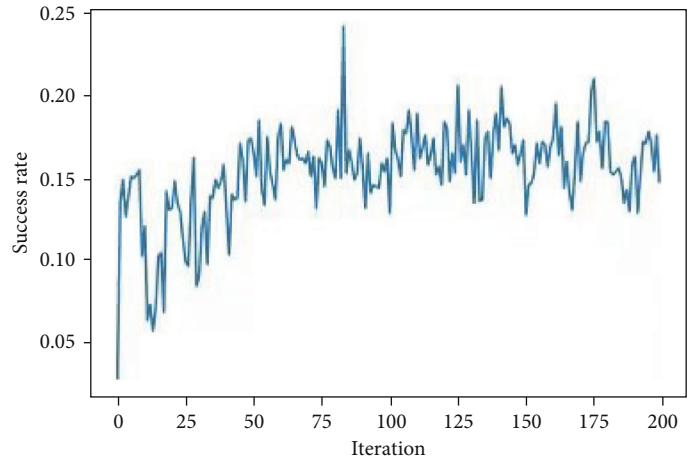


(a) With large communication delay of 0.4 second

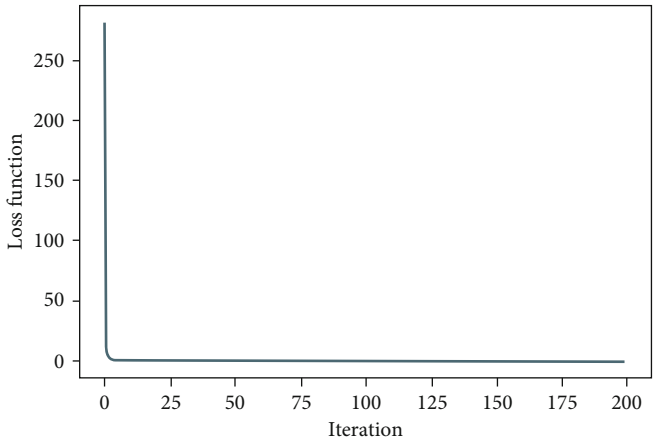


(b) With communication delay of 0.6 sec

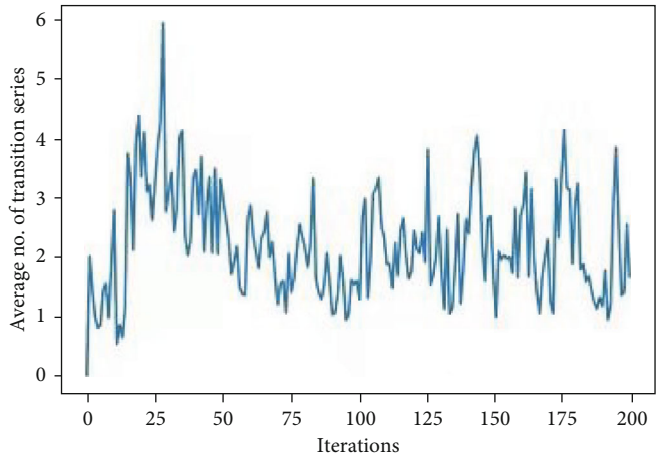
Figure 7: Communication delay in CRN.



(a) Success rate vs. no. of iterations



(b) Loss functions vs. no. of iterations



(c) Average no. of transition series vs. no. of iterations

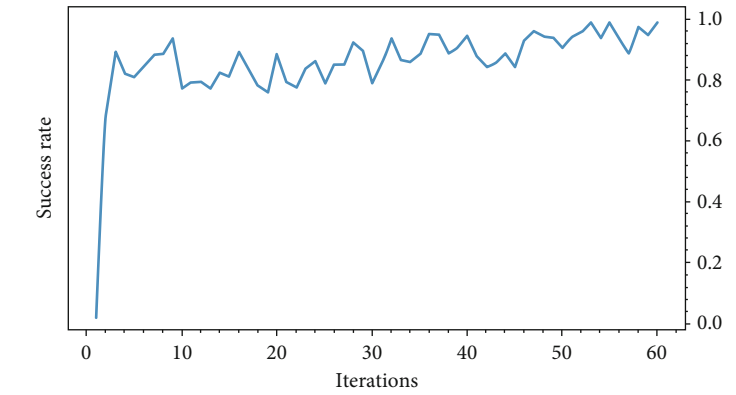
Figure 8: Success rate, loss function, and average number of transitions vs. iterations using no. of sensors (3).

noise power N_1 and N_2 at P_{R1} and S_{R2}, respectively, are set to 0.01 W. For simplicity, channel advancement from the primary transmitter and secondary transmitter to P_{R1} and S_{R2} is considered to be $h_{xy} = 1, \forall x, y$. To achieve the lowest SNR (signal-to-noise ratio), productive reception for PU and SU is adjusted to $\acute{\omega}_c = 11/9$ and $\acute{\omega}_d = 7/10$, respectively. The presence of transmit power $\{\partial_a, \partial_b\}$ can be easily noted, which means that the quality of the PU-SU system is satisfactory. Additionally, several active sensors remain N, which would accept the acknowledged signal strength information to help SU to adapt to the power control policy. There is uniformly distributed distance d_{xy} between the sensor node and T_{xi} transmitter in a pause range of [100,300] in meters.

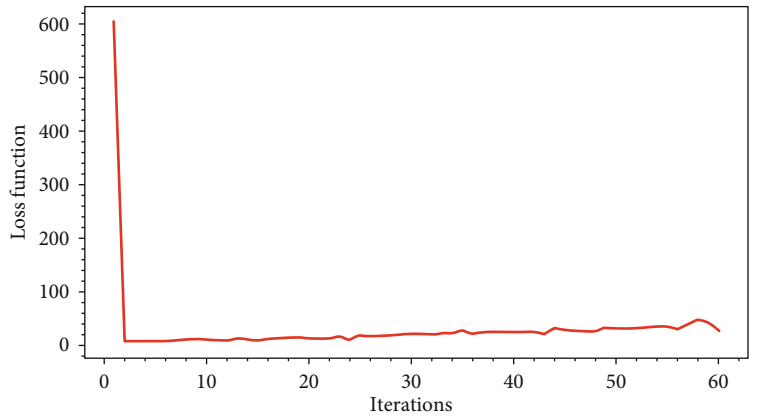
For the action value approximation function, we used the deep neural network (DNN), which contains feedforward layers. The number of these layers is three and is entirely linked. These hidden layers contain neurons 256, 256, and 512, respectively. For the 1st and 2nd layers, rectified linear units are working as an action function, where the ReLU layer output will be 0 or raw for negative input. The activation function is used in the case of the 3rd layer, and the Adam algorithm is used to update weight φ, while the mini-batch size is set to 256. Here, it is considered that rerun memory K holds$N_K = 400$. In each iteration, training of φ starts, and 300 evolutions/transitions are considered for each K. The number of transitions is adjusted to $r = 10^5$. Exploring the probability of a new-fangled action, action reduces with an increase in the number of iterations, and as the iterations increase to a sufficiently large value, it reduces from 0.8 to 0. Additionally, at reiteration r, it is supposed that $\underline{\epsilon}_r = 0.8(1 - r)/K$.

There is access to deep reinforcement learning in our article, specifically, success rate, loss function, and average number. The transition step is being used as transition number function r has been analysed. Here, the presentation is judged through two metrics, which are transition steps and average number. Success rate computation is in standings of the figure of successful trails to total figure of easily turn's a ratio. Here, assume a productive trail if "m" transfers to a target in 20 time frames. In the case of a productive trail, the average number of time frames needs to reach a target which termed as transition step average number.

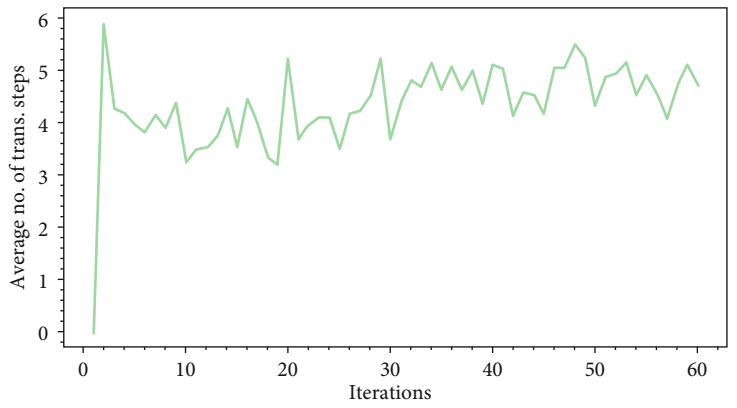
During the exercise process, loss function can be calculated as presented in [14]. After learning iteration r, SU



(a) Success rate vs. iterations



(b) Loss functions vs. iterations



(c) Average no. of transition series vs. iterations

Figure 9: Success rate, loss function, and average number of transitions vs. iterations using no. of sensors (5) and iterations (10000).

may use the trained network for interaction with PU. Network performance is determined standings of an average number of transitions and success rates. Outcomes are average over 10^3 freely turns, where the chance start state is chosen for each track. The numerical results have been divided into two cases.

3.1. Case 1 (without Delay). Case one shows that by the guidance of learned power control policy, transmit power of SU/unlicensed user can intelligently be adjusted like final target may be touched since somewhat preliminary target indoors rare figure of transmission steps. In Figure 8(a) (success rate vs. no. of iterations for case 1), the number of sensors used is 3, in which we set according to the number of iterations k, $\sigma_n = (\partial_a^{\sigma} J_{cn} + \partial_b^m) J_{dn}/3$. With the increase in σ_n to reduce the number of sensors, the loss function value becomes larger. It is good policy to achieve the average number of transition steps and loss function vs. no. of iterations similar to those shown in Figures 8(b) and 8(c). Figure 8 reveals the strength of the deep reinforcement learning approach.

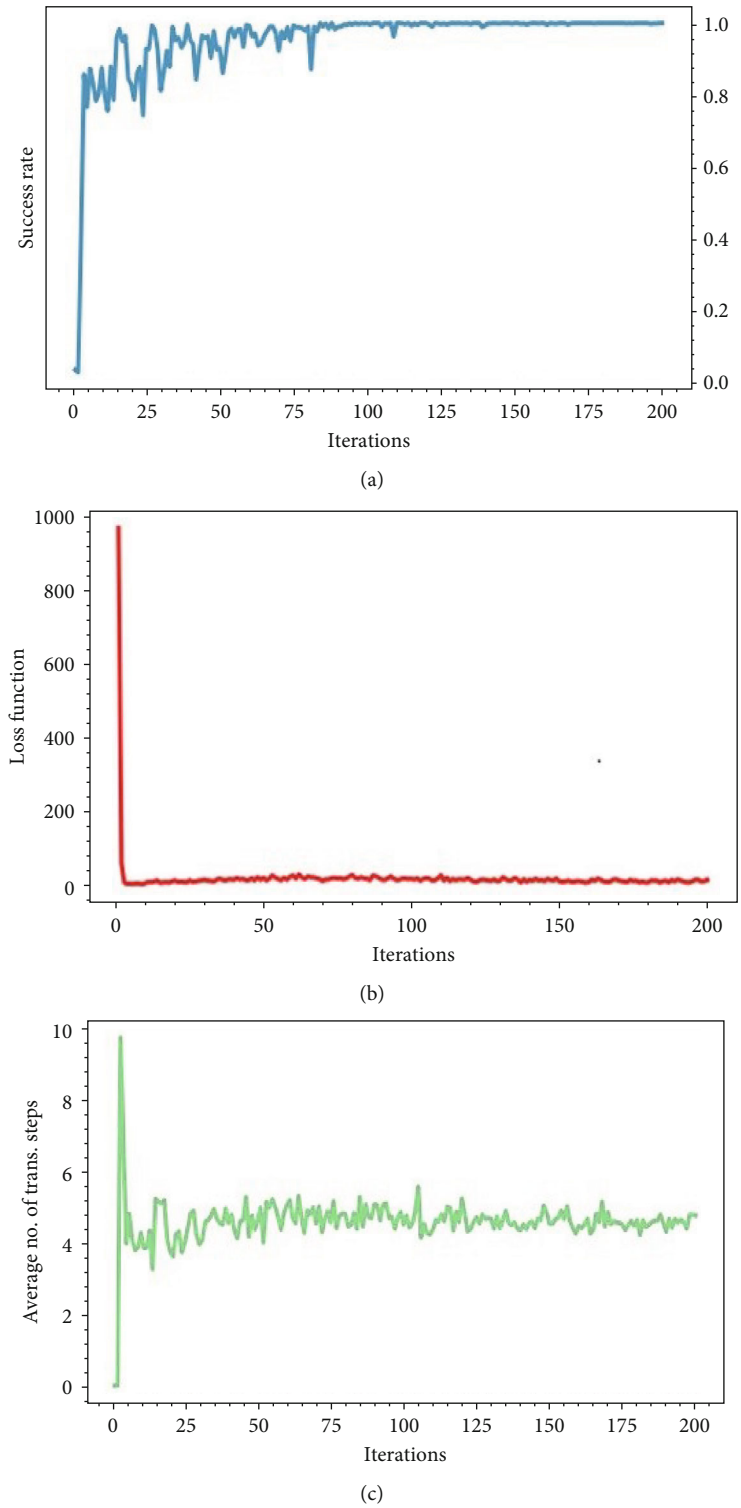


Figure 10: Success rate, loss function, and an average number of transitions vs. iterations using several sensors (10) and iterations (30000).

3.2. Case 2 (with Delay). Case 2 shows the sensing time delay (r+k) as shown in Figure 2. This sensing time delay has been found in the SU policy as in Equation (5). This sensing time delay occurs when PU shares transmit power or spectrum with SU; then, there is the penetration of SUs on the primary node that causes sensing delay during the sharing of transmit power or spectrum. In this case, power control performance reduces. Figure 9(a) shows the success rate vs. no. of iterations. Figure 9(b) shows the loss function vs. no. of iterations, and Figure 9(c) shows the average no. of transitions vs. no. of iterations. We showed the recital of the DQN-based power control method in which a secondary power control policy

is employed by the PU as in Equation (4) which seems additionally traditional. The target of learning a suitable power control policy is in Figure 9 which shows the loss function, success rate, and an average number of transition steps vs. k. Here, we used the number of sensors which is 10. $\sigma_n = (\partial_a^{\sigma} J_{cn} + \partial_b^m) J_{dn}/10$. It is to be observed that a greater number of iterations are required for getting a success rate. In our policy, the average number of transition steps is 2.5 for achieving the target state. Here, the PU used a second policy which only permits transmit power for an increase or decrease in the signal level at each step. For this purpose, we need more steps. The explained method sustains a supposed performance loss in the case of fewer sensors deployed; this happened due to random variations in observation states which built various conditions less unique from one another and save the agents from learning of an active strategy. Random variation effect can be neutralized by the use of more sensors.

Figure 10(a) shows the success rate vs. iterations, Figure 10(b) shows the loss function vs. iterations, and Figure 10(c) shows the average no. of transition steps vs. no. of iterations. For the purpose of training, we used a number of sensors (10). $\sigma_n = (\partial_a^{\sigma} J_{cn} + \partial_b^m) J_{dn}/10$. In this sample, the plot is run with 30000 iterations and a delay of 2 samples. Here, in the success rate plot, the effect of delay is clear. This sensing delay is because of the penetration of SUs on the primary node during the sharing of spectrum. In this case, power control performance will reduce.

Figure 6(a) shows the consensus performance of CRN showing results without the proposed model sensing delay. However, these are poor transit responses of consensus performance. All agents reached their steady-state value after nearly 100 seconds. However, the settling time is large. So, for progressing performance, one should track this projected technique to lessen the effects of communication and sensing time delay as well. If there is the consideration of communication and sensing delay amid entirely CR operators, suppose that sensing delay of 0.4 second inside whole operations is identical. The assumed method is enough for the representation of large-scale sensing delay in the CR network. Figures 7(a) and (7)(b) show the communication delay of 0.4 sec and 0.6 sec, respectively, in a CRN.

4. Conclusions

Machine learning and CR intelligence have a good capability to understand and adapt to the wireless environment. In wireless communication, machine learning techniques are linked with CR technology. The cognitive radio system which consists of PU and SU is well studied; in this paper, we explained the problem of spectrum sharing of PU and SU in CRNs, and also, we introduced sensing delay in communication. Because there was a common concept that there is no cooperation between PU and SU, PUs adjust their transmit powers on their own transmit control power policy. In this article, we discussed the consensus performance and introduced a Q-learning or deep reinforcement learning method for SU to study for the adjustment of its transmit power so that ultimately both the PU and SU have the ability to transmit respective data fruitfully with the essential qualities of services. Our numerical results showed that the considered learning method is healthy against the random variations in the state variation and within few numbers of steps. We can get our target from an initial state, and also, in numerical results, we showed sensing delay that is because of the number of SUs when PU transmits spectrum with the SU; then, there is a rush of SU on the primary node that caused sensing delay; it is because of a lot of several users. All agents reached their steady-state value after nearly 100 seconds. However, the settling time is large. The given method is enough for the representation of large-scale sensing delay in the CR network.

Data Availability

Data can be provided on request.

Conflicts of Interest

We hereby confirm that there is no conflict of interest between authors to declare.

Acknowledgments

This work was supported by the Scientific and Technological Innovation Foundation of Shunde Graduate School, University of Science and Technology Beijing (the major project no. is BK19CF002) of China.

References

[1] Z. Qin, X. Zhou, L. Zhang, Y. Gao, Y. Liang, and G. Y. Li, "20 years of evolution from cognitive to intelligent communications," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 6–20, 2020.

[2] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," *IEEE Wireless Communications Letters*, vol. 9, no. 1, pp. 108–112, 2019.

[3] M. Di Renzo, K. Ntontin, J. Song et al., "Reconfigurable intelligent surfaces vs. relaying: differences, similarities, and performance comparison," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 798–807, 2020.

[4] Z. Ahmed, M. M. Khan, M. A. Saeed, and W. Zhang, "Consensus control of multi-agent systems with input and communication delay: a frequency domain perspective," *ISA Transactions*, vol. 101, pp. 69–77, 2020.

[5] N. Tadayon and S. Aïssa, "A multichannel spectrum sensing fusion mechanism for cognitive radio networks: design and application to IEEE 802.22 WRANs," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 4, pp. 359–371, 2015.

[6] J. Adu Ansere, G. Han, H. Wang, C. Choi, and C. Wu, "A reliable energy efficient dynamic spectrum sensing for cognitive radio IoT networks," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6748–6759, 2019.

[7] A. Kumar and K. Kumar, "Multiple access schemes for cognitive radio networks: a survey," *Physical Communication*, vol. 38, article 100953, 2020.

[8] M. A. Al-Jarrah, A. Al-Dweik, S. S. Ikki, and E. Alsusa, "Spectrum-occupancy aware cooperative spectrum sensing using

adaptive detection," *IEEE Systems Journal*, vol. 14, no. 2, pp. 2225–2236, 2020.

[9] Y. R. Kondareddy, P. Agrawal, and K. Sivalingam, "Cognitive radio network setup without a common control channel," in *MILCOM 2008 - 2008 IEEE Military Communications Conference*, pp. 1–6, San Diego, CA, 2008.

[10] A. A. Khan, M. H. Rehmani, and M. Reisslein, "Cognitive radio for smart grids: survey of architectures, spectrum sensing mechanisms, and networking protocols," *IEEE Communication Surveys and Tutorials*, vol. 18, no. 1, pp. 860–898, 2016.

[11] I. Mitliagkas, N. D. Sidiropoulos, and A. Swami, "Joint power and admission control for ad-hoc and cognitive underlay networks: convex approximation and distributed implementation," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4110–4121, 2011.

[12] Z. Xiao, X. Shen, F. Zeng et al., "Spectrum resource sharing in heterogeneous vehicular networks: a noncooperative game-theoretic approach with correlated equilibrium," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9449–9458, 2018.

[13] M. H. Islam, Y. Liang, and A. T. Hoang, "Distributed power and admission control for cognitive radio networks using antenna arrays," in *2007 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 250–253, Dublin, 2007.

[14] H. V. Vu, N. H. Tran, and T. Le-Ngoc, "Full-duplex device-to-device cellular networks: power control and performance analysis," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3952–3966, 2019.

[15] M. M. Aslam, L. Du, Z. Ahmed, H. Azeem, and M. Ikram, "Consensus performance of traffic management system for cognitive radio network: an agent control approach," in *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*, vol. 1, Springer, Singapore, 2019.

[16] Y. Kuo, J. Yang, and J. Chen, "Efficient swarm intelligent algorithm for power control game in cognitive radio networks," *IET Communications*, vol. 7, no. 11, pp. 1089–1098, 2013.

[17] S. Tomic, M. Beko, and R. Dinis, "RSS-based localization in wireless sensor networks using convex relaxation: noncooperative and cooperative schemes," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 5, pp. 2037–2050, 2015.

[18] J. Gummeson, D. Ganesan, M. D. Corner, and P. Shenoy, "An adaptive link layer for range diversity in multi-radio mobile sensor networks," in *IEEE INFOCOM 2009*, pp. 154–162, Rio De Janeiro, Brazil, 2009.

[19] M. M. Aslam, M. N. Irshad, H. Azeem, and M. Ikram, "Cost effective & energy efficient intelligent smart home system based on IoT," *Afyon Kocatepe University International Journal of Engineering Technology and Applied Sciences*, vol. 3, pp. 10–20, 2020.

[20] F. Azmat, Y. Chen, and N. Stocks, "Analysis of spectrum occupancy using machine learning algorithms," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 6853–6860, 2016.

[21] N. Sizochenko, M. Syzochenko, N. Fjodorova, B. Rasulev, and J. Leszczynski, "Evaluating genotoxicity of metal oxide nanoparticles: application of advanced supervised and unsupervised machine learning techniques," *Ecotoxicology and Environmental Safety*, vol. 185, article 109733, 2019.

[22] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: a survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020.

[23] X. Zhou, M. Sun, G. Y. Li, and B. H. Fred Juang, "Intelligent wireless communications enabled by cognitive radio and machine learning," *China Communications*, vol. 15, no. 12, pp. 16–48, 2018.

[24] R. Barto and A. G. Sutton, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, 1998.

[25] G. Ding, Y. Jiao, J. Wang et al., "Spectrum inference in cognitive radio networks: algorithms and applications," *IEEE Communication Surveys and Tutorials*, vol. 20, no. 1, pp. 150–182, 2018.

[26] G. Ganesan and Y. Li, "Cooperative spectrum sensing in cognitive radio, part II: multiuser networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 6, pp. 2214–2222, 2007.

[27] Z. Liu, M. Zhao, K. Y. Chan, Y. Liu, and K. Ma, "Resource allocation strategy against selfishness in cognitive radio ad-hoc network based on Stackelberg game," *IET Communications*, vol. 13, no. 13, pp. 1962–1970, 2019.

[28] E. C. van der Meulen, "A survey of multi-way channels in information theory: 1961-1976," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 1–37, 1977.

[29] H. Zhang, N. Yang, K. Long, M. Pan, G. K. Karagiannidis, and V. C. M. Leung, "Secure communications in NOMA system: subcarrier assignment and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 7, pp. 1441–1452, 2018.

[30] Q. Yu, J. Chen, Y. Fan, X. Shen, and Y. Sun, "Multi-channel assignment in wireless sensor networks: a game theoretic approach," in *2010 Proceedings IEEE INFOCOM*, pp. 1–9, San Diego, CA, 2010.

[31] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3722–3732, 2016.

[32] V. Mnih et al., "Playing Atari with deep reinforcement learning," pp. 1–9, 2013, https://arxiv.org/abs/1312.5602.

[33] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," *International Conference on Machine Learning, PMLR*, vol. 48, pp. 1329–1338, 2016.

[34] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. M. Leung, "Power control based on deep reinforcement learning for spectrum sharing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4209–4219, 2020.

[35] F. Hu, B. Chen, X. Zhai, and C. Zhu, "Channel selection policy in multi-SU and multi-PU cognitive radio networks with energy harvesting for Internet of everything," *Mobile Information Systems*, vol. 2016, 12 pages, 2016.

[36] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[37] B. Givan and R. Parr, "Introduction POMDP.pdf," pp. 1–23, 2001.

[38] Z. Ahmed, M. A. Saeed, A. Jenabzadeh, and Z. Weidong, "Frequency domain analysis of resilient consensus in multi-agent systems subject to an integrity attack," *ISA Transactions*, 2020.

[39] D. Choi, "Model-based-RL-deepmind," pp. 1–13, 2016.

[40] J. A. Fax and R. M. Murray, "Information flow and cooperative control of vehicle formations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1465–1476, 2004.

[41] P. Erdos and A. Rényi, "On the evolution of random graphs," *Structure in Dynamic Networks*, vol. 9781400841, pp. 38–82, 2011.