Hindawi
Wireless Communications and Mobile Computing
Volume 2019, Article ID 6051201, 12 pages
https://doi.org/10.1155/2019/6051201

Research Article

Actor–Critic-Algorithm-Based Accurate Spectrum Sensing and Transmission Framework and Energy Conservation in Energy-Constrained Wireless Sensor Network-Based Cognitive Radios

Hurmat Ali Shah,[1] **Insoo Koo**,[2] **and Kyung Sup Kwak**[1]

[1]*Department of Information and Communication Engineering, Inha University, Incheon, Republic of Korea*
[2]*School of Electrical and Computer Engineering, University of Ulsan, Ulsan, Republic of Korea*

Correspondence should be addressed to Insoo Koo; iskoo@ulsan.ac.kr

Received 22 March 2019; Revised 30 June 2019; Accepted 8 July 2019; Published 14 August 2019

Guest Editor: Bojan Dimitrijevic

Spectrum sensing is of the utmost importance to the workings of a cognitive radio network (CRN). The spectrum has to be sensed to decide whether the cognitive radio (CR) user can transmit or not. Transmitting on unoccupied spectrum becomes a hard task if energy-constrained networks are considered. CRNs are ad hoc networks, and thus, they are energy-limited, but energy harvesting can ensure that enough energy is available for transmission, thus enabling the CRN to have a theoretically infinite lifetime. The residual energy, along with the sensing decision, determines the action in the current time slot. The transmission decision has to be grounded on the sensing outcome, and thus, a combined sensing–transmission framework for the CRN has to be considered. The sensing–transmission framework forms a Markov decision process (MDP), and solving the MDP problem exhaustively through conventional methods cannot be a plausible solution for ad hoc networks such as a CRN. In this paper, to solve the MDP problem, an actor–critic-algorithm-based solution for optimizing the action taken in a sensing–transmission framework is proposed. The proposed scheme solves an optimization problem on the basis of the actor–critic algorithm, and the action that brings the highest reward is selected. The optimal policy is determined by updating the optimization problem parameters. The reward is calculated by the critic component through interaction with the environment, and the value function for each state is updated, which then updates the policy function. Simulation results show that the proposed scheme closely follows the exhaustive search scheme and outperforms a myopic scheme in terms of average throughput achieved.

1. Introduction

Wireless sensor networks are pervasive. They are employed in a variety of applications and services ranging from smart grids, Internet of things, to cognitive radios. Wireless sensor networks have sensing at its core. Sensing is required not only to gather information about events, i.e., to record events and behavior of processes, but also to decide the occurrence of a phenomena. When sensor networks are employed to decide the occurrence of an event or a phenomenon, it can be generalized as the basis of diverse set of communication networks. In this way, sensor networks form the basis for cognitive radios when cognitive radios are deployed overlay, i.e., the cognitive radios have to use the licensed spectrum when not in use by the licensed user. Wireless sensor networks form the sensing base for cognitive radios in deciding whether the spectrum is free or not. So, spectrum sensing is an important task in cognitive radio networks (CRNs). The spectrum has to be sensed and the whole spectrum appropriated for the cognitive radio (CR) user if the full promise of the CRN is to be exploited. On the basis of spectrum sensing, the CR user can transmit, provided it vacates as soon as the primary user (PU) appears. This is more complicated if energy-constrained networks are considered. CRNs are ad hoc networks, and the energy for transmission may not always be available. Given the limited energy budget, the CR

user has to take into consideration the long-term operation of the network. So, one of the multiple levels of transmission power can be selected, or given that transmission power is unavailable, no transmission will happen.

In an uncertain but stochastically known environment, behavior can be learned and the policy of the CR user can be adjusted. The CR user can be seen as an agent in terms of reinforcement learning, and through the acknowledgment (ACK) signal, the environment can be known. In each time frame, the optimal decision has to be taken on the basis of the sensing decision and other system parameters, such as historical information about the belief of the presence or absence of the PU. The optimal decision can be learned, but it will take a lot of iterations by the system to reach it. This is the biggest drawback of simple reinforcement learning (RL) [1]. Instead, deep RL algorithms can be considered to make the optimal decision in the current time slot and to design an optimal policy for the agent.

RL and deep RL solutions and algorithms are most effective in situations where there are patterns and models, so they can be learned. The abstractions of the environment are translated into parameters on the basis of which decision can be taken. In the area of communications, the methods of deep RL can be effective in solving many challenges, particularly in emerging fields varying from the CRN to the Internet of things (IoT) to heterogeneous networks. As communications networks become more and more autonomous and decentralized, the need for optimal operation on the basis of both the environment and gained knowledge becomes paramount. The problems of the CRN, such as spectrum access and transmission power level in energy-constrained situations to optimize different network parameters, can be devised as a decision problem. In the stochastic nature of wireless communications, the decision-making problem can be modeled as a Markov decision process (MDP). An MDP can be solved through deep RL techniques. The sensing–transmission problem is an MDP problem, and it is solved in this paper.

After the sensing decision is made, CR users either transmit or stay silent. The decision to transmit is based on the current state and the transition probability to the next state, as well as the remaining energy, if energy constraints of the network are considered. The states for transition are limited by the current state. The set of actions is also determined by the current state, which may be a combination of spectrum-sensing decision, the belief function, and the remaining energy. In Markov decision systems, the state transition should strictly follow the Markov chain. So, all the possible states and possible actions for those states will have to be computed. In this case, the computation becomes complex and expensive.

This paper presents a model-free reinforcement learning algorithm called an actor–critic algorithm [2, 3]. The advantage of the actor–critic algorithm is that it is not as computationally complex as partially observable MDP (POMDP) approaches and it does not need all of the state transition information to make a decision. In the training phase, the critic updates the approximation of state information on the basis of simulation parameters and feeds this information to the actor to update the policy parameters. The actor–critic algorithm may converge to a local optimal policy, but it generates the policy directly from training, so much less computational complexity and formulation are required. The actor–critic algorithm obtains a solution to the nonconvex optimization problem as presented in this paper without complete and accurate information about the wireless network environment.

Deep learning is composed of a set of algorithms and techniques that attempt to find important features of data and that try to model high-level abstractions. The main goal of deep learning is to avoid manual descriptions of a data structure (like handwritten features) by automatically learning from the data. Typically, deep learning structures are any neural network which has two or more hidden layers. The algorithms can be divided into policy optimization and dynamic programming. In policy optimization, the policy is parameterized, and the expected reward is maximized. Methods in this category include policy gradients and derivative-free optimization and evolutionary algorithms, whereas dynamic programming methods can exactly solve some simple control problems (i.e., MDPs) through iteration and subdivision. Dynamic programming includes policy iteration and value iteration, and (for more useful and realistic problems) approximate versions of these algorithms are considered (e.g., Q-learning). The actor–critic methods are policy-gradient methods that use value functions.

The notion of learning from the environment is embedded in the concept of cognitive radio. CR users are meant to monitor the environment and adapt their operating characteristics (operating frequency, transmitting power, etc.,) to the changing conditions. To enable CR users to learn from the environment, several authors have considered machine learning algorithms for spectrum sensing [4–12]. In [4], a dynamic game is formulated where the secondary users incur sensing charges as a result of PU activity and then after a successful completion of a bidding also pay for transmission opportunities. To solve the problem of bidding, a Bayesian nonparametric belief update is used and learning algorithms are employed, so the users can decide whether to participate in the bidding process or not. In [5], to maximize the available spectrum for secondary users' transmission, a select number of CR users sense multiple bands simultaneously through RL. Q-learning is also employed for diverse purposes. In [7], it is employed to mitigate interference in CRNs while in [8], it is employed for detection of primary signal for spectrum sensing. In [9], two kinds of approaches in a no-regret-based machine learning framework with the presence of malicious activity were proposed with two different algorithms, one had perfect observation of the environment and the other had a partial observation, while in [10], pattern classification techniques as SVM and KNN were investigated for spectrum sensing. In [11, 12], a TV white space database was constructed through machine learning and data fusion was carried for global spectrum sensing decision, respectively. In [11], k-nearest neighbors (KNN) were simply used for data recovery in a white space database as a mechanism for majority voting.

Recently, there has been much interest in the efficient use of energy resources for the sake of energy-aware network architectures and to reduce energy costs [13, 14]. To that end, energy resources have to be used according to system performance requirements. Energy harvesting from renewable energy sources (thermal, vibration, solar, acoustic, and ambient radio power) is designed to meet energy requirements, as well as contribute towards reducing harmful emissions [13]. Energy harvesting allows for theoretically perpetual lifetime operation of an ad hoc network without any need for an external power source or battery replacement. This is an attractive model for many kinds of future wireless and cellular networks and is most important for CRNs. Pei et al. [15] studied an energy-efficient design of a spectrum-sensing scheme in cognitive radio networks. Chen et al. [16] maximized throughput given an energy budget, whereas Hoang et al. [17] designed a utility-based objective function for penalizing energy consumption. Studies have been conducted on exploiting energy harvesting in ad hoc networks [18]. For multiple relay channels, energy harvesting-based schemes were proposed [19, 20]. Wang et al. [21] studied energy harvesting in heterogeneous networks, and both they and Anisi et al. [22] studied the effect of energy harvesting in sensing and body-area networks.

Deep learning is a vast area and is now excessively used in all aspects of communications, particularly in intelligent communications such as cognitive radios, so giving an exhaustive list of literature references is out of the scope of this work, but here, the application of deep learning to spectrum sensing and spectrum access in CRN will be discussed briefly. In [23] a deep Q-learning (DQL) algorithm is used to select the channels for sensing and ultimately access to achieve maximum rate. The action is selected on the basis of the channel SINR exceeding a particular QoS requirement. In [24], a DQL is proposed in a heterogeneous network which consists of multiple users and base stations. The base stations for the users to be associated with are selected through DQL. The simulation results in [24] also confirmed that DQL has better performance than a simple Q-learning scheme. In [25], a deep learning scheme was proposed for a joint user association, spectrum access, and content caching problem in an LTE network. The users are controlled by cloud-based servers, and they are able to access both licensed and unlicensed spectrum. Different actions are selected on the basis of the DQL algorithm such as optimal user association, the bandwidth allocated to users, the time slot allowed to the users, and deciding the content popularity for caching. In [26], a dynamic channel access scheme for a sensor-based IoT is proposed which employs deep learning, while in [27], experience replay is used to find the optimal policy. In [28], a joint channel selection and packet forwarding scheme is proposed in a multisensor scenario, while to counter the energy limitation problem in [29], channel access in an energy harvesting-based IoT system is investigated.

In this paper, an energy-constrained CRN is considered. When transmission is considered jointly with the results from sensing, then in an energy-constrained CRN, that becomes a hard problem to solve. In the literature, machine learning was applied extensively (as shown above) to the process of spectrum sensing, but not to taking a combined sensing–transmission decision. The problem of transmission is considered an MDP, and the conventional methods are used for solving the problem [30–32]. The wireless environment is a partially observable MDP, and thus, on the basis of some observations, the next state is transitioned to. The problem of the POMDP is solved by visiting all the next possible states from taking all possible actions in the current state and selecting an action that optimizes certain network parameters. When sensing decisions along with the remaining energy in energy-constrained CRNs are considered to be the state, the state space becomes too large to compute and solve. Another problem with such POMDP solutions is that the environment is not aptly learned. As an alternative to the conventional POMDP, the actor–critic algorithm is emerging as a promising alternative. In an actor–critic algorithm, the agent has two parts. The actor takes an action according to a policy, whereas the critic adjusts the policy through parameters like temporal difference. The policy in the actor is optimized on the basis of optimal value function. The value function can have two definitions: (1) the total accumulated reward while starting in the current state and (2) going to the next states according to the given policy. In the state-action value function, the expected accumulated rewards are calculated while taking an action in the current state, and then, in the next state taking other actions according to the given policy. The value function adjusts the policy function in the actor on the basis of observations from the environment. Convergence in the actor–critic algorithm is possible. It is achieved with less complexity and fewer iterations and computations in the state space. Also, in the actor–critic algorithm, the policy is directly learned from the operational environment.

In this paper, an actor–critic-algorithm-based sensing and transmission framework is proposed. CR users take a local sensing decision and then send it to the fusion center (FC) to take a global decision, and these two form part of the local state. The belief function and the remaining energy form the rest of the state. The action space is formed of either the silent mode or transmission with a level of energy that is able to fulfil the long-term energy requirements of the system. CR users are able to harvest energy, and the transmission energy in the current slot also has to take into consideration the long-term energy requirements. On the basis of the transmission and ACK signal, a reward is assigned to each action. The critic evaluates the reward brought by the action and updates the policy function. At the end of the training phase, the optimal value function and the optimal policy function are obtained.

The rest of this paper is organized as follows. Section 2 presents the system model. In Section 3, the system energy constraints, the energy harvesting process, and the Markov decision process are explained. In Section 4, the actor–critic algorithm is presented, while Section 5 presents simulation results and Section 6 concludes the paper.

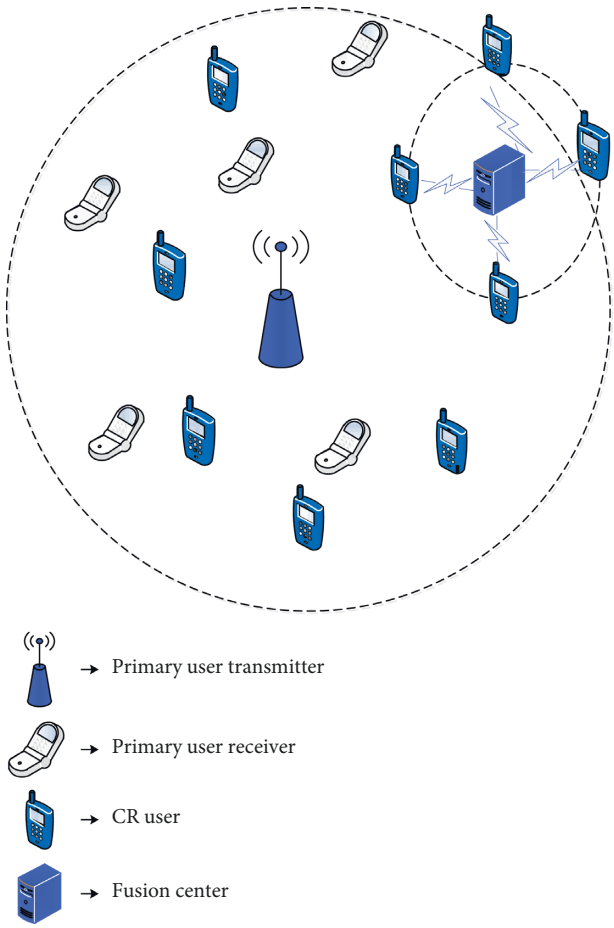

Figure 1: Basic system model.

2. System Model and Methods

We consider that a single PU is active and a CRN that consists of N CR users, as shown in Figure 1, monitors the activity of the PU. Considering multiple PUs complicates the sensing process as well as involves other processes such as scheduling and spectrum-handoff. These particular problems are beyond the scope of this paper. The CR users perform spectrum sensing and report their results to the FC. We assume a slotted time-frame structure where each slot is divided into two slots: a sensing slot for spectrum sensing and a transmission slot, which is used for data transmission. The slotted frame structure is considered in [33–41]. In this method of spectrum sensing, the time frame is divided into two parts. The first part of the time frame is known as sensing slot while the other is transmission slot. Spectrum is sensed in the first while the CR users either transmit or stay silent in the transmission slot on the basis of global decision. The durations of both slots among CR users are synchronized through a common control channel and by the FC.

Each CR user employs the energy detection scheme for spectrum sensing. CR users receive energy and on the basis of received energy take a decision. Energy detection is the simplest technique, given the limited resources (e.g., energy and computational power) of most CR users. Common spectrum sensing problems such as multipath fading and shadowing can be overcome by exploiting spatial diversity using cooperative spectrum sensing, thereby ensuring that PU constraints are met [33]. CR users can either report the actual value of energy received or take a decision locally or report it to the FC. The first one is called soft decision combination and results in optimal detection performance but theoretically requires infinite bandwidth [34], while the latter is hard decision combination which saves bandwidth but produces inferior results as compared to soft reporting. To balance performance and bandwidth efficiency, a combination of both soft and hard decisions can be used where the energy range can be quantized, as in references [34, 37]. In [34], the authors used a so-called softened hard combination scheme, in which the observed energy is quantized into four regions using two bits, where each region is represented by a label. This achieves an acceptable trade-off between the improved performance resulting from soft reporting and information loss during quantization process [41]. In this paper, quantization is considered where the received energy is quantized into four quantization zones.

The signal received by the i-th CR user, during the sensing slot, when the PU is absent, i.e., H_0, and when the PU is present, i.e., H_1, is given as

$$y_i = \begin{cases} w(n), & H_0, \\ s(n) + w(n), & H_1, \end{cases} \tag{1}$$

where $w(n)$ is additive white Gaussian noise and $s(n)$ is the energy received from the PU's signal. The received energy is quantized as

$$l_i = \begin{cases} H_0 \begin{cases} Z_1, & Y_i \le \lambda_{Z_1}, \\ Z_2, & Y_i \le \lambda_{Z_2}, \end{cases} \\ H_1 \begin{cases} Z_3, & Y_i \le \lambda_{Z_3}, \\ Z_4, & Y_i > \lambda_{Z_3}, \end{cases} \end{cases} \tag{2}$$

where λ_{Z_1}, λ_{Z_2}, and λ_{Z_3} are the quantization thresholds and $\{Z_1, Z_2, Z_3, Z_4\}$ represent different quantization zones in which the received energies are quantized. A global decision is taken on the basis of majority rule, i.e., the majority of the reported symbols become the global decision, denoted by D_t, where t represents the time index. The combination of local and global decisions determines the state of the CR in the current slot.

3. System Constraints and Definitions

In the section below, the system constraints and processes are explained in detail.

3.1. Energy Harvesting Process. The CR users are able to harvest energy. If the energy arrival process, $e_h^t \subset R^+$, is assumed to be independent and identically distributed (i.i.d.) sequences of variables, then $E\{e_h^t\} = e_h$ [13]. It is also assumed that the energy harvested in time slot t is immediately available in slot $t+1$.

The total energy spent in the current slot, t, if the CR user transmits, is given as

$$e_c(t) = e_s + \alpha_t e_r, \tag{3}$$

where e_s is the sensing energy, e_r is the transmission energy, and α_t is given as

$$\alpha_t = \begin{cases} 1, & \text{if the CR user transmits,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The residual energy for the next time slot, if the CR user transmits in the current time slot, is

$$e_{rem}(t+1) = \min\left\{e_{cap}, e_{rem}(t) - e_c(t) + e_h\right\}, \tag{5}$$

where e_{cap} is the maximum battery capacity and $e_{rem}(t)$ is the residual energy at time t.

To ensure long-term operation of the network, the energy remaining in the current time slot has to satisfy the energy requirements for some future time slots. Because the transmission energy is dependent upon the sensing decision, its future value cannot be estimated. There is assumed to be a number of future time slots for which energy is required. To maintain energy conservation and long-term operation of the network, it is necessary to save energy in the current time slot. The sensing energy for future time slots remains fixed as sensing happens in each time slot in the sensing slot. The consumption by transmission energy is dependent on the sensing decision, so it cannot be determined in advance. Thus, on the basis of sensing energy, a constraint for energy preservation to ensure long-term operation of the network can be formulated. Let us suppose that we want the CRN to be functional for next N future slots; then, the constraint for long-term energy availability can be formulated as

$$e_{rem} \geq N(e_s - e_h). \tag{6}$$

3.2. The Markov Process. Let the PU activity follow a two-state Markov process, as shown in Figure 2. Figure 2 illustrates the Markov process where the CR user either transitions to another state or remains in the same state. On the edges, the transition probabilities are given. For the sake of simplicity, H is not written as the subscript of P.

The sensing and transmission framework is formulated as a Markov decision process. The Markov decision process tuple is $\langle S, A, P, R\rangle$, where S represents the state, A is the action taken, P is the transition probability, and R is the reward received on taking an action given a state.

The state is composed of the local and global sensing decisions, the remaining energy, and the transition probability, denoted by $\mu(t)$. For simplicity, let us denote the combination of local and sensing decisions as Q_{ld}. The state at time t is given as

$$s(t) = \{Q_{ld}, e_{rem}(t), \mu(t)\}. \tag{7}$$

The transition probabilities are dependent upon the current local and global sensing decisions. They will be presented in detail later.

The CR user, after sensing, can either be silent or transmit at different levels of transmission energy to meet long-term energy requirements. Two transmission energy levels are considered. There can be many levels, but the formulation and the solution will become untenable. The action space is defined as

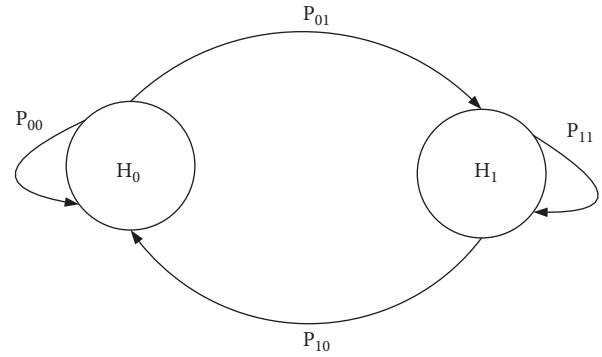

Figure 2: The Markov process.

$$A = \left\{\text{SIL}, e_r^1(t), e_r^2(t)\right\}, \tag{8}$$

where $e_r^1(t)$ represents transmitting with transmission energy e^1 and $e_r^2(t)$ denotes transmitting with energy level e^2, while SIL indicates no transmission.

The reward is based on the ACK signal. The rewards are assigned as follows, where $T^{\{\cdot\}}$ represents the throughput achieved with the given state and transmission energy:

$$\begin{aligned} R\left(\frac{s(t), e_r^1(t)}{\text{ACK}}\right) &= T^{\{e_r^1(t)\}}, \\ R\left(\frac{s(t), e_r^2(t)}{\text{ACK}}\right) &= T^{\{e_r^2(t)\}}, \\ R\left(\frac{s(t), e_r^i(t)}{\text{ACK}}\right) &= 0, \quad \text{and } i = 1, 2, \end{aligned} \tag{9}$$

where $T = \log_2(1 + e_r \ell_i)$ and ℓ_i is the SNR received by i-th CR user.

4. Actor–Critic Algorithm

The CR user can take an action, given a particular state, and transition to another state in the current time slot, determined as follows:

$$P\left(\frac{s'}{s(t), a(t)}\right) = \begin{cases} 1, & \text{if } s' = s(t+1), \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The reward associated with each state is given in (9).

The total discounted reward in the t-th time slot is given by a value function when the current state is $s(t)$, computed as follows [3]:

$$V(s(t)) = \sum_{k=t}^{\infty} \gamma^k R(s(t), a(t)). \tag{11}$$

The policy function is given as follows [3]:

$$\pi\left(\frac{a(t)}{s(t)}\right) = P\left(\frac{a(t) \in A}{s(t)}\right) = \frac{e^{\phi(a(t),s(t))}}{\sum_{a\in A} e^{\phi(a(t),s(t))}}, \quad (12)$$

where $e^{\phi(a(t),s(t))}$ is the tendency to select action $a(t)$ in state $s(t)$.

After the action is taken, the reward will be calculated. After calculating the reward, the temporal difference is determined as follows:

$$\delta(t) = [R(s(t), a(t)) + \gamma V(s(t+1)) - V(s(t))], \quad (13)$$

where γ determines the effect of the state transition from the current state to the next state.

On the basis of the temporal difference, the value function is updated by the critic as

$$V(s(t)) = V(s(t) + \beta \cdot \delta(t)), \quad (14)$$

where β is the positive parameter of the critic. The tendency to select an action, given a state, is updated as

$$\varphi(s(t), a(t)) = \varphi(s(t), a(t)) + \chi \cdot \delta(t), \quad (15)$$

where χ is a positive step-size parameter, which determines the number of states to be visited.

The decision in the current time slot is based on the sum of the reward in the current time slot and the expected future reward in the next time slot. If the global decision is Z_1, calculating the reward from the current and future time slots on the basis of the status of the ACK signal is

$$B^i_{Z_1} = P_{\text{ACK}} \times \left\{R\left(Q_{l1}, e_{\text{rem}}(t), \mu(t), e^i_r(t)\right)\right\} + P_{\text{ACK}} \times \sum_{\left\{\begin{array}{c} t=t+1 \\ e_{\text{rem}}(t+1) \end{array}\right\}} P[*]V(s(t)) + P_{\text{ACK}} \times \sum_{\left\{\begin{array}{c} t=t+1 \\ e_{\text{rem}}(t+1) \end{array}\right\}} P[*]V(s(t)). \quad (16)$$

For Z_2, it is

$$B^i_{Z_2} = P_{\text{ACK}} \times \left\{R\left(Q_{l2}, e_{\text{rem}}(t), \mu(t), e^i_r(t)\right)\right\} + P_{\text{ACK}} \times \sum_{\left\{\begin{array}{c} t=t+1 \\ e_{\text{rem}}(t+1) \end{array}\right\}} P[*]V(s(t)) + P_{\text{ACK}} \times \sum_{\left\{\begin{array}{c} t=t+1 \\ e_{\text{rem}}(t+1) \end{array}\right\}} P[*]V(s(t)), \quad (17)$$

where $P[*]$ gives the probability that (6) will be satisfied and $i \in (1, 2)$. The decision function for the current time slot can then be formulated as

$$B_0\left(e_{\text{rem}}(t+1), \mu(t), Q_{ld}\right) = \text{Arg} \max_A \left\{B^i_{Z_1}, B^i_{Z_2}\right\}, \quad (18)$$

where $l, d \in (Z_1, Z_2)$, and A is given by (8).

The training process is meant to find the set comprising the policy and the optimal value function corresponding to each state. The CR users take a local decision and send the quantized energy zone to the FC. The FC takes a global decision and sends it to the CR users. Based on the local decision and the global decision, the CR users can stay silent or transmit at one of two levels of transmission energy. At the beginning of each time slot, a CR user takes action $a(t) \in A$ according to policy $\pi(a(t)/s(t))$ in a given state. There will be a transition to another state or the current state will be retained, and the next state will be $s(t+1)$ based on the residual energy and the feedback. The rewards will be calculated according to (9). The temporal difference is calculated according to (12) after calculating the reward on the basis of temporal differences, updating the value function in (13). The tendency to select action $a(t)$ in state $s(t)$ is updated in (13). After the convergence is achieved, there will be an optimal value function, $V(s)$, and an optimal set of policies, π. The following are the possible cases for CR users, on the basis of which the value function is updated and the policy function found. These cases are determined by the system model and the level of transmission energy. They are run till the optimal value function and optimal policy function are obtained.

Case 1. If $D_t = Z_1$ or Z_2 && $l_1 = Z_1$, then stay silent. The belief that the PU is absent in the current time slot is updated using Bayes' rule [2] as

$$\mu^*(t) = \frac{\mu(t) P_{\text{f}}^{Z_1}}{\mu(t) P_{\text{f}}^{Z_1} + (1 - \mu(t)) P_{\text{d}}^{Z_1}}, \quad (19)$$

where $P_{\text{f}}^{Z_i}$ is the local probability of false alarm for zone Z_i and $P_{\text{d}}^{Z_i}$ is the local probability of detection for zone Z_i, with $i \in (1, 2, 3, 4)$. The belief for the next time slot is given as

$$\mu(t+1) = \mu^*(t) P_{11} + \left(1 - \mu^*(t)\right) P_{01}. \quad (20)$$

The residual energy for the next time slot is updated as

$$e_{\text{rem}}(t+1) = \min\left\{e_{\text{cap}}, e_{\text{rem}}(t) - e_{\text{s}} + e_{\text{h}}\right\}. \quad (21)$$

Case 2. If $D_t = Z_1$ or Z_2 && $l_1 = Z_2$, then stay silent. The belief that the PU is absent in the current time slot is updated using Bayes' rule as

$$\mu^*(t) = \frac{\mu(t)P_{\mathrm{f}}^{Z_2}}{\mu(t)P_{\mathrm{f}}^{Z_2} + (1-\mu(t))P_{\mathrm{d}}^{Z_2}}. \tag{22}$$

The belief for next time slot and the residual energy for next time slot for cases 2 to 5 are given in (20) and (21), respectively.

Case 3. If $D_t = Z_1$ or Z_2 && $l_1 = Z_3$, then stay silent. The belief that the PU is absent in the current time slot is updated using Bayes' rule as

$$\mu^*(t) = \frac{\mu(t)P_{\mathrm{f}}^{Z_3}}{\mu(t)P_{\mathrm{f}}^{Z_3} + (1-\mu(t))P_{\mathrm{d}}^{Z_3}}. \tag{23}$$

Case 4. If $D_t = Z_1$ or Z_2 && $l_1 = Z_4$, then stay silent. The belief that the PU is absent in the current time slot is updated using Bayes' rule as

$$\mu^*(t) = \frac{\mu(t)P_{\mathrm{f}}^{Z_4}}{\mu(t)P_{\mathrm{f}}^{Z_4} + (1-\mu(t))P_{\mathrm{d}}^{Z_4}}. \tag{24}$$

Case 5. If $D_t = Z_1$ or Z_2 && $l_1 = Z_1$, then take a decision according to (18). The belief that the PU is truly absent in the current time slot is given by Bayes' rule (if transmission happens and ACK is received) as follows:

$$\mu^*(t) = \frac{\mu(t)P_{\mathrm{f}}^{Z_1}}{\mu(t)P_{\mathrm{f}}^{Z_1} + (1-\mu(t))P_{\mathrm{d}}^{Z_1}}. \tag{25}$$

The residual energy at the CR user for the next time slot is given as

$$e_{\mathrm{rem}}(t+1) = \min\left\{e_{\mathrm{cap}}, e_{\mathrm{rem}}(t) - e_{\mathrm{r}}^{j} - e_{\mathrm{s}} + e_{\mathrm{h}}\right\}, \quad \text{where } j \in (1,2). \tag{26}$$

The belief that the PU will be absent in the next time slot is given as

$$\mu(t+1) = P_{01}. \tag{27}$$

Case 6. If $D_t = Z_2$ or Z_2 && $l_1 = Z_1$, then take a decision according to (18). The belief that the PU is truly absent in the current time slot is given by Bayes' rule (if transmission happens and ACK is received) as follows:

$$\mu^*(t) = \frac{\mu(t)\left(1-P_{\mathrm{f}}^{Z_2}\right)}{\mu(t)\left(1-P_{\mathrm{f}}^{Z_2}\right) + (1-\mu(t))\left(1-P_{\mathrm{d}}^{Z_2}\right)}. \tag{28}$$

The residual energy at the CR user for the next time slot and the belief are given in (23) and (24), respectively. In the absence of an ACK signal, in both Case 5 and Case 6, the belief probability for the next time slot is updated according to (20).

Based on the ACK signal, the rewards are assigned if Case 5 and Case 6 occur on the basis of (9).

In Figure 3, the basic flow chart of the proposed scheme is presented. First, the local and global sensing decisions are made. Combined with the remaining energy and the belief function, a decision is taken according to the cases explained above. The decision is taken by the actor component of the actor–critic algorithm. On the basis of the action taken, there is interaction with the environment. On the basis of observations, the belief function (along with remaining energy) is updated and the reward calculated. The parameters of the optimization problem given in (18) are updated, and the value function is calculated by the critic. On the basis of the updated value function, the temporal difference as given in (13) is determined. Ultimately, the policy is updated implicitly by updating the tendency to select an action, given a state, according to (15).

5. Simulation Results

In this section, the performance of the proposed actor–critic-algorithm-based sensing and transmission framework is assessed through simulation. In the simulation, the proposed scheme is compared with an exhaustive search scheme where the reward brought by each action in each state is calculated and the best selected, rather than finding the transition probabilities for each state to another state. This scheme can be considered an upper bound for the proposed scheme. A myopic scheme is also considered, where energy is harvested, but the long-term energy constraint is not considered, and the maximum available power is used for transmission.

The exhaustive search scheme has precedent in the literature, where in [42], an offline scheme was proposed as an upper bound for a deep Q-network based scheme. The offline scheme assumed that the base station has perfect knowledge of all the random process and thus it can take the optimal decision. The proposed scheme does not have noncausal knowledge of the processes involved, i.e., the future state of battery and the PU activity. The number of mathematical operations involved for the exhaustive search scheme depends on the state space and action space but with the added advantage of having knowledge of all the random processes. The maximum number of mathematical operations, i.e., computational complexity, for the exhaustive search scheme is $O(S_t \times A)$, while the computational complexity of the proposed scheme is given by $O(S_t \times A \times \gamma \times \chi)$. As both γ and χ are positive values below 1, the computational complexity of the proposed scheme is less than the exhaustive search scheme.

When simulating the proposed scheme, the initial value of residual energy is assumed to be $0.6(e_{\mathrm{c}} + e_{\mathrm{r}}^1)$; e_{r}^1 is 150 mW, and e_{r}^2 is 110 mW. The circuit power consumption was kept fixed at 210 [13–15]. These energy settings parameters are considered because of the energy harvesting model. The energy harvesting model considered in this paper was investigated in detail by the authors in [13] and the values for the parameters are based on simulation results as obtained there. The two different values of transmission energies are considered to be realistic in comparison with

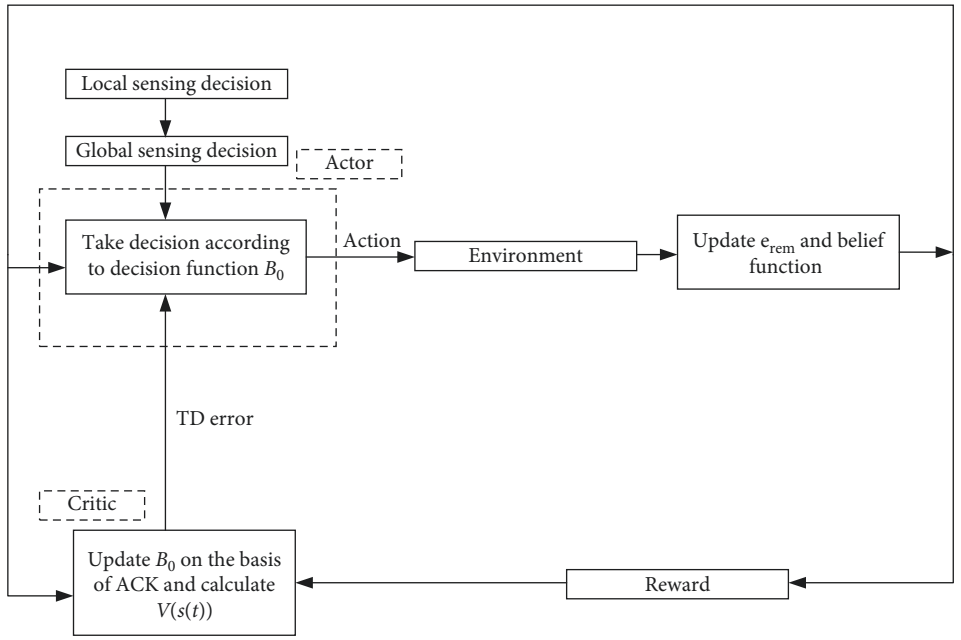

Figure 3: Flow chart of the proposed scheme.

sensing energy and overall circuit consumption. The transmit power was varied for the different simulation results and is given for each. The time slot duration was 200 ms, and the sensing duration was eight of the total slot durations. The noise spectrum density was taken to be 4.4×10^{-21} W/Hz [43]. The number of CR users considered was three. The probability of detection is 0.9, and the false alarm probability is 0.1, while the initial belief about the PU is 0. The state transition probabilities of the channel are 0.2. The value of γ was kept at 0.4, while χ was 0.3. These values were taken experimentally for achieving convergence, and different values will have different convergence behavior.

Figure 4 presents the comparison of the proposed scheme with the exhaustive search scheme. We can see from the figure that the proposed scheme closely follows the exhaustive search scheme, which acts as the upper bound for the proposed scheme. The average SNR of all the CR users was taken to be the value given on the x-axis, and it was changed as shown in the figure. The number of iterations was taken to be 5000. The average rate was calculated for all the slots according to (9). SNR is important both for sensing and for success of transmission. When the SNR is low, the CR users will not be able to sense properly, and even when the CR users decide to transmit, the rate achieved will be less because of the higher ratio of the noise signal. So, checking the performance at different levels of SNR is important. At a very low SNR of −9 dB, the proposed scheme starts to converge, and the behavior is the same as the upper bound. The exhaustive search scheme, rather than taking an optimized decision, searches the subspace of available actions, and thus, the one selected is the global optimum. The proposed scheme, on the other hand, may at best converge to a locally optimal policy, and thus, there is always a gap, even after the training converges to an optimal value function and policy. Because the subspace of all available actions is

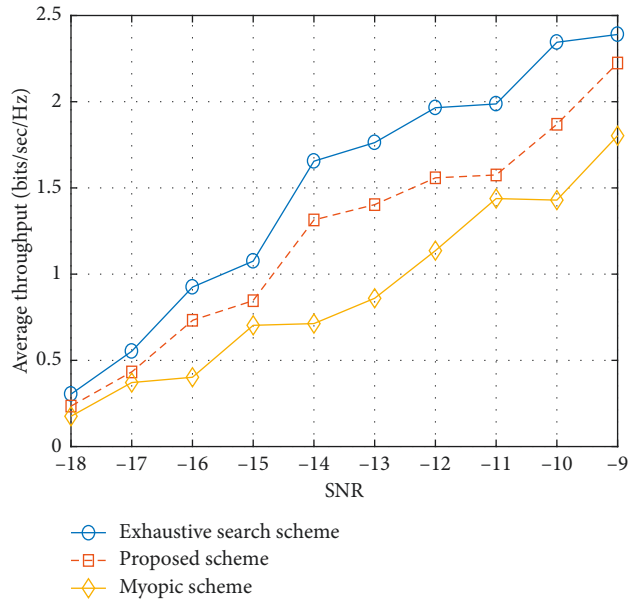

Figure 4: Average throughput of the system.

checked thoroughly in the exhaustive search scheme and all the rewards brought by all possible next states are calculated, it is computationally expensive. The proposed scheme, on the other hand, has less computational complexity but gives performance closely following the upper bound. On the other hand, the proposed scheme outperforms the myopic scheme, which transmits with all the transmission power available. It may achieve high throughput in some slots, but it eventually runs out of transmission energy. The state space is continuous as the energy harvested and the remaining energy are continuous functions. The size of the state space cannot be predetermined; however, the positive step

parameters given in (14) and (15) determine the rate of convergence at the expense of not visiting all the states. In simulation for exhaustive search schemes, all states are visited unless there is convergence, and the reward returned does not change significantly. So, the positive step parameters in (14) and (15) determine the complexity of the proposed scheme while for exhaustive search scheme, all the states will be visited unless there is a convergence to a stable reward value.

Figure 5 presents the false decision rate, which is represented by P_{fd}. The false decision rate measures the probability that the PU is absent and the CR user does not transmit or that the PU is absent and the CR user does transmit. The sensing part is the same for both schemes, but in the exhaustive search scheme, all the states and actions are visited, and the best is selected; in the proposed scheme, only the optimization problem is solved, and that may not be accurate. Though both schemes follow the same quantization-based sensing scheme, the error performance from the exhaustive search scheme is better than with the proposed scheme because the proposed scheme is based on estimating the next state, whereas the exhaustive search scheme checks the reward brought by all the next possible states and selects the best one. In that scenario, the exhaustive search scheme assumes correct information for all the possible states.

Figure 6 shows the convergence of the proposed scheme. The x-axis is the number of iterations. We can see that as the number of iterations increases, system performance improves. The system performance is dependent upon the information obtained from the environment, and the optimization problem presented here learns both the probability that the CR user will have enough energy to transmit and the reward brought by each action taken in a given state. With the increase in the system run, the critic component of the proposed scheme is able to calculate and thus to limit the temporal difference. On the basis of the temporal difference, the policy parameters are optimized. As many times as the proposed scheme is invoked, there is the same number of updates to the temporal difference error, and hence, the best action (given a state) can be selected. When the number of iterations reaches a certain point, we can see that system performance reaches a stable position, and despite further increases in the number of iterations, the performance improves no more. Thus, there is also a limit to performance improvement in terms of the number of iterations, and performance improvement beyond that point would need a new model of energy harvesting or changing the other system's parameters and constraints.

In Figure 7, the effect of the energy harvesting rate is shown. The effect is determined by r which is the energy harvesting rate. The energy harvesting rate affects the harvested energy which is given by $e_{\text{h}} = r\,(e_{\text{c}} + e_{\text{r}}^{1})$. The x-axis in Figure 7 shows different values of r. The proposed scheme closely matches the exhaustive search scheme when the harvested energy is below a certain limit. This is because when there is limited energy available, a transmission cannot be made, despite having the best information about the operating environment, and thus, the exhaustive search scheme cannot outperform the proposed scheme by a big

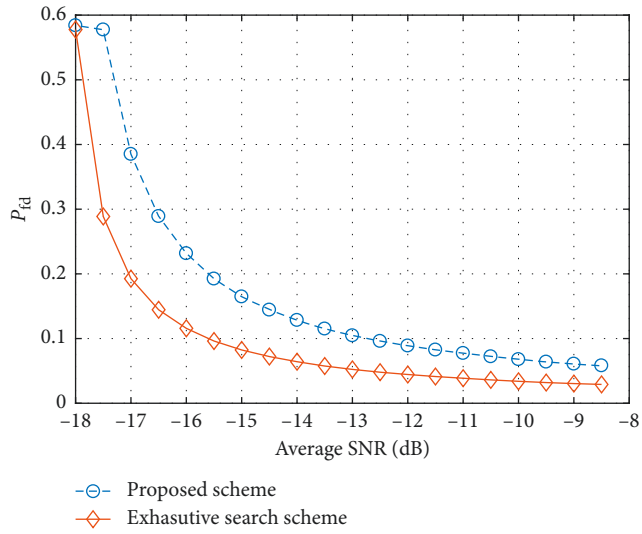

Figure 5: Comparison of the probability of false alarm.

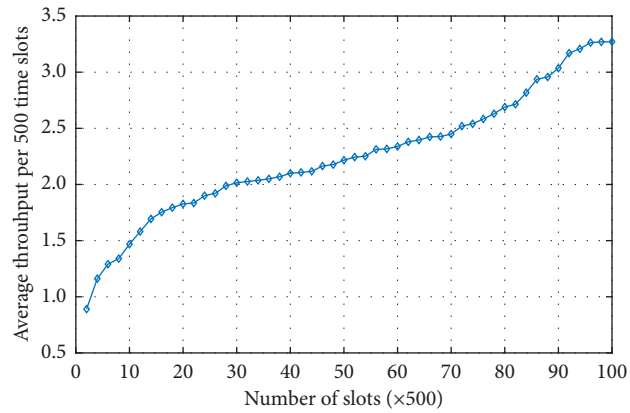

Figure 6: Convergence rate of the proposed scheme.

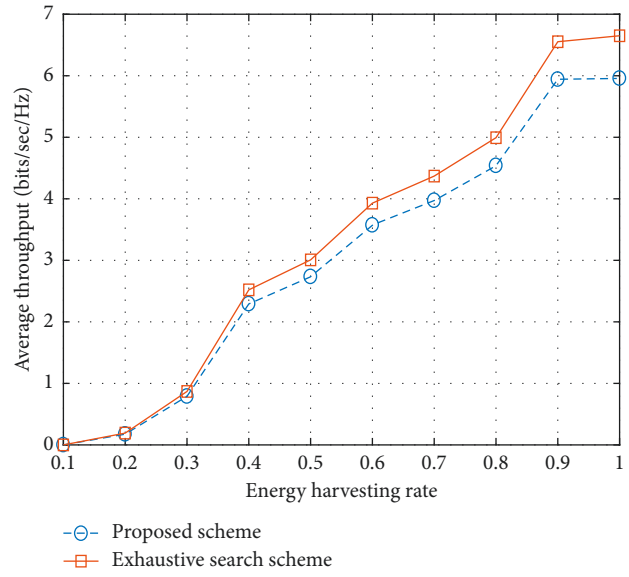

Figure 7: Effect of harvested energy on average throughput.

margin. When there is sufficient transmission energy available and when the energy harvesting rate improves, there is sufficient energy available for transmission, and thus, the decision taken on the basis of the exhaustive search scheme outperforms the solution to the optimization problem, which does not check all the possible solutions. The exhaustive search scheme calculates the next possible states and calculates the actions that can be taken in all future states and thus can better know the future energy state. The proposed scheme, on the other hand, makes a decision on the basis of the probability of receiving the ACK signal and on the energy constraint. Because it solves the optimization problem, rather than visiting all possible states and calculating the reward brought by each action in every possible state, the proposed scheme settles for low transmission energy to satisfy the system's energy constraints. But despite not visiting all possible states, the proposed scheme closely follows the exhaustive search scheme, and under practical conditions (when information about all possible states is not available), the proposed scheme based on the actor–critic algorithm comes across as an acceptable alternative.

As the harvested energy remains low, there is a slight chance that transmission with higher level of power will be selected. Thus, the performance of both the proposed scheme and the exhaustive search scheme remains the same in terms of average throughput despite the exhaustive search scheme visiting more states as ultimately the transmission is carried out with low transmission power. But as the harvested energy increases, the exhaustive search scheme performs better because it can predict well the future state of energy, and so transmission with higher level of power can be carried out. On the other hand, because of having inexact knowledge of the energy harvesting process and thus the future state of the remaining energy, the proposed scheme opts to transmit with low level of transmission power because of the constrain given in (6). Thus, the exhaustive search scheme gives better average throughput than the proposed scheme when the energy harvesting rate increases.

6. Conclusion

In this paper, a joint sensing and transmission framework was considered. The transition probabilities from one state to another and the set of available actions are determined from the sensing result and the amount of residual energy. This allows for a robust framework where the CRN ensures there is energy available for future time slots while achieving throughput in the current slot. The actor–critic algorithm is formulated to decide the next state and the amount of transmission energy, if there is a transmission. The value function takes care to come up with an optimal policy, which associates an optimal action with each state. After the training is complete, there is an optimal policy function as the environment is learned through the interplay of the actor and critic functions. The proposed scheme avoids computing all the state and action space and rather finds an action which optimizes the reward in a given state. The optimal policy is updated in each time slot, and the critic acts to reduce the deviation from the optimal path. The proposed scheme which is based on reinforcement learning-based actor–critic algorithm is less computationally expensive and less exhaustive while solves the optimization problem to find the optimal action in a given state. The simulation results show that the proposed scheme closely follows the exhaustive search scheme despite having less computations and solving an optimal solution.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Research Foundation (NRF) grant through the Korean Government (MSIT) under grant NFR-2018R1AB6001714.

References

[1] N. C. Luong, N. C. Luong, D. T. Hoang, S. Gong et al., "Applications of deep reinforcement learning in communications and networking: a survey," 2018, http://arxiv.org/abs/1810.07862.

[2] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds., Vol. 12, MIT Press, Cambridge, MA, USA, 2000.

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

[4] Z. Han, R. Zheng, and H. V. Poor, "Repeated auctions with Bayesian nonparametric learning for spectrum access in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 3, pp. 890–900, 2011.

[5] J. Lundén, V. Koivunen, S. R. Kulkarni, and H. V. Poor, "Reinforcement learning based distributed multiagent sensing policy for cognitive radio networks," in *Proceedings of the 2011 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, IEEE Aachen, Germany, May 2011.

[6] M. Bkassiny, K. J. Sudharman, and K. A. Avery, "Distributed reinforcement learning based MAC protocols for autonomous cognitive secondary users," in *Proceedings of the 2011 20th Annual Wireless and Optical Communications Conference (WOCC)*, IEEE Newark, NJ, USA, April 2011.

[7] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823–1834, 2010.

[8] B. Y. Reddy, "Detecting primary signals for efficient utilization of spectrum using Q-learning," in *Proceedings of the Fifth International Conference on Information Technology: New Generations (ITNG 2008)*, IEEE Las Vegas, NV, USA, April 2008.

[9] Q. Zhu, Z. Han, and T. Başar, "No-regret learning in collaborative spectrum sensing with malicious nodes," in *Proceedings of the 2010 IEEE International Conference on Communications*, IEEE, Cape Town, South Africa, May 2010.

[10] K. M. Thilina, K. W. Choi, N. Saquib, and E. Hossain, "Pattern classification techniques for cooperative spectrum sensing in cognitive radio networks: SVM and W-KNN approaches," in *Proceedings of the 2012 IEEE Global Communications Conference (GLOBECOM)*, pp. 1260–1265, IEEE, Anaheim, CA, USA, December 2012.

[11] M. Tang, Z. Zheng, G. Ding, and Z. Xue, "Efficient TV white space database construction via spectrum sensing and spatial inference," in *Proceedings of the 2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC)*, pp. 1–5, IEEE, Nanjing, China, December 2015.

[12] A. M. Mikaeil, B. Guo, and Z. Wang, "Machine learning to data fusion approach for cooperative spectrum sensing," in *Proceedings of the 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 429–434, IEEE, Shanghai, China, October 2014.

[13] S. Park, H. Kim, and D. Hong, "Cognitive radio networks with energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1386–1397, 2013.

[14] L. Cai, H. Poor, Y. Liu, T. Luan, X. Shen, and J. Mark, "Dimensioning network deployment and resource management in green mesh networks," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 58–65, 2011.

[15] Y. Pei, Y.-C. Liang, K. C. Teh, and K. H. Li, "Energy-efficient design of sequential channel sensing in cognitive radio networks: optimal sensing strategy, power allocation, and sensing order," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1648–1659, 2011.

[16] Y. Chen, Q. Zhao, and A. Swami, "Distributed spectrum sensing and access in cognitive radio networks with energy constraint," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 783–797, 2009.

[17] A. T. Hoang, Y.-C. Liang, D. T. C. Wong, Y. Zeng, and R. Zhang, "Opportunistic spectrum access for energy-constrained cognitive radios," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1206–1211, 2009.

[18] J. Yang, O. Ozel, and S. Ulukus, "Broadcasting with an energy harvesting rechargeable transmitter," *IEEE Transactions on Wireless Communications*, vol. 11, no. 2, pp. 571–583, 2012.

[19] H. Li, N. Jaggi, and B. Sikdar, "Relay scheduling for cooperative communications in sensor networks with energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 2918–2928, 2011.

[20] I. Krikidis, T. Charalambous, and J. S. Thompson, "Stability analysis and power optimization for energy harvesting cooperative networks," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 20–23, 2012.

[21] L. Wang, K.-K. Wong, S. Jin, G. Zheng, and R. W. Heath, "A new look at physical layer security, caching, and wireless energy harvesting for heterogeneous ultra-dense networks," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 49–55, 2018.

[22] M. H. Anisi, G. Abdul-Salaam, M. Y. I. Idris, A. W. A. Wahab, and I. Ahmedy, "Energy harvesting and battery power based routing in wireless sensor networks," *Wireless Networks*, vol. 23, no. 1, pp. 249–266, 2017.

[23] Y. Lin, X. Dai, L. Li, and F.-Y. Wang, "An efficient deep reinforcement learning model for urban traffic control," 2018, http://arxiv.org/abs/1808.01876.

[24] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with synchronous off-policy updates," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3389–3396, IEEE, Singapore, June 2017.

[25] X. Di, K. Xiong, P. Fan, H.-C. Yang, and K. B. Letaief, "Optimal resource allocation in wireless powered communication networks with user cooperation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7936–7949, 2017.

[26] D. Zhao, H. Wang, K. Shao, and Y. Zhu, "Deep reinforcement learning with experience replay based on SARSA," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6, IEEE, Athens, Greece, December 2016.

[27] W. Wang, J. Hao, Y. Wang, and M. Taylor, "Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach," 2018, http://arxiv.org/abs/1803.00162.

[28] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access," in *Proceedings of the International Conference on Computing, Networking and Communications (ICNC)*, IEEE, Guilin, China, July 2017.

[29] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431–5440, 2008.

[30] B. C. Jung, J. Park, T.-W. Ban, W. Lee, and J. M. Kim, "Full-duplex generalized spatial modulation: a compressed sensing-based signal detection," in *Proceedings of the 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, IEEE, Milan, Italy, July 2017.

[31] Y. Cui, W. Xu, and J. Lin, "A novel compressed data transmission scheme in slowly time-varying channel," in *Proceedings of the 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, IEEE, Valencia, Spain, September 2016.

[32] M. Hirzallah, W. Afifi, and M. Krunz, "Full-duplex-based rate/mode adaptation strategies for Wi-Fi/LTE-U coexistence: a POMDP approach," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 1, pp. 20–29, 2017.

[33] F. Rongfei and H. Jiang, "Optimal multi-channel cooperative sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 3, pp. 1128–1138, 2010.

[34] J. Ma, G. Zhao, and Y. Li, "Soft combination and detection for cooperative spectrum sensing in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4502–4507, 2008.

[35] S. Kyperountas, N. Correal, and Q. Shi, *A Comparison of Fusion Rules for Cooperative Spectrum Sensing in Fading Channels*, EMS Research, Motorola, Libertyville, IL, USA, 2010.

[36] H. Guo, W. Jiang, and W. Luo, "Linear soft combination for cooperative spectrum sensing in cognitive radio networks," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1573–1576, 2017.

[37] H. Sakran and M. Shokair, "Hard and softened combination for cooperative spectrum sensing over imperfect channels in cognitive radio networks," *Telecommunication Systems*, vol. 52, no. 1, pp. 61–71, 2013.

[38] H. A. Shah, M. Usman, and I. Koo, "Bioinformatics-inspired quantized hard combination-based abnormality detection for cooperative spectrum sensing in cognitive radio networks," *IEEE Sensors Journal*, vol. 15, no. 4, pp. 2324–2334, 2015.

[39] P. Kaligineedi and V. K. Bhargava, "Sensor allocation and quantization schemes for multi-band cognitive radio cooperative sensing system," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 284–293, 2011.

[40] R. Chen, J. M. Park, and K. Bian, "Robust distributed spectrum sensing in cognitive radio networks," in *Proceedings of the 2008 IEEE INFOCOM-The 27th Conference on Computer Communications*, pp. 31–35, IEEE, Phoenix, AZ, USA, April 2008.

[41] H. A. Shah and I. Koo, "Reliable machine learning based spectrum sensing in cognitive radio," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 5906097, 17 pages, 2018.

[42] M. Chu, H. Li, X. Liao, and S. Cui, "Reinforcement learning-based Multiaccess control and battery prediction with energy harvesting in IoT systems," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2009–2020, 2019.

[43] H. Mu, M. Tao, W. Dang, and Y. Xiao, "Joint subcarrier-relay assignment and power allocation for decode-and-forward multi-relay OFDM systems," in *Proceedings of the 2009 Fourth International Conference on Communications and Networking in China*, pp. 1–6, IEEE, Xian, China, August 2009.