Jang and Yoo *EURASIP Journal on Wireless Communications and Networking* (2018) 2018:160
https://doi.org/10.1186/s13638-018-1155-9

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Q-learning-based dynamic joint control of interference and transmission opportunities for cognitive radio

Sung-Jeen Jang and Sang-Jo Yoo*

Abstract

In cognitive radio (CR) system, secondary user (SU) should use available channels opportunistically when the primary user (PU) does not exist. In CR network, SUs have to detect the PU signal with sufficient sensing time to guarantee the detection probability and minimize the interference to the PU, while the CR system should have enough data transmission time to maximize the transmission opportunity of the SU. Therefore, the sensing time and data transmission time of the SU are generally considered as main optimization parameters to maximize the throughput of the CR system. In this paper, a separate sensing node is designated and the sensing is continuously performed using the interference alignment (IA) technique. In this paper, the designated sensing node estimates the interference ratio and transmission opportunity loss ratio. To satisfy the primary user's interference requirement and maximize secondary throughput, we proposed dynamic adjustment mechanism for sensing slot time and sensing report interval using reinforcement learning in time-varying communication environment. The experimental results show that the proposed approach can minimize the interference on PU and enhance the transmission opportunity of SUs.

Keywords: Cognitive radio, Interference alignment, Spectrum sensing, Q-learning

1 Introduction

As the demand for multimedia services explosively increases, the need for bandwidth to meet the requirements of communication systems is also rapidly increasing. Periodic frequency auctions in each country are a major issue of interest to telecom operators due to astronomical costs, and these costs are included in the CAPEX plans of operators, so they are inevitable to be passed on to the consumers for the purpose of operating profit safeguard guarantee [1]. Therefore, it is necessary to use the frequency more efficiently to drastically reduce the cost of the frequency and reduce the communication cost. Cognitive radio (CR) is a technology that can improve the efficiency of frequency spectrum use and has been in continuous research since it was proposed by Mitola [2]. CR should be able to intelligently monitor and adapt to the surrounding environment to share the frequency band with the licensed primary user (PU) in a frequency band not occupied by the PU [3]. In the last few years, several research works have been done in order to apply CR to spectrum sharing and for secondary users (SUs) to coexist with PUs in relation to wireless standards: the IEEE 802.22 Wireless Regional Access Network in TV white spaces (TVWS), IEEE 802.11af for wireless local area network service in TVWS, IEEE 802.19, IEEE 1900.x, and the European Telecommunications Standards Institute's Reconfigurable Radio Systems for coexistence of license-exemption systems. Besides these considerations, CR systems are considered in the complex system problem consisting of heterogeneous systems, like the coexistence problem of a femtocell that can be installed indiscriminately in a macrocell, device-to-device (D2D) coexistence problem within the licensed band, and the coexistence of WiFi and Long Term Evolution in unlicensed spectrum (LTE-U) in the Industiral, Scientific and Medical (ISM) band [4, 5].

CR users are basically required to have spectrum sensing to gain access to the spectrum without interfering with the primary networks. Therefore, various efforts have been made to improve the accuracy of sensing, such as matched filter detection, energy detection, feature detection, and cooperative detection [6]. Since the radio frequency front

* Correspondence: sjyoo@inha.ac.kr
Multimedia Network Lab, Inha University, 253 Yonghyun-Dong, Nam-gu, Incheon 402-751, South Korea

end cannot distinguish between the PU and SU signals, sensing and data transmission must be separated [7]. Although feature detection can identify the modulation types of PU signal, the processing time is considerably longer, and higher computational complexity is required [8]. There is a tradeoff between the interference with PUs and throughput of the SU in the mechanism that separates the sensing and data transmissions. For interference avoidance, the observation time (i.e., sensing time) should be long enough to ensure the accuracy of PU detection. However, a longer observation time reduces the transmission time of the SU, thereby reducing the data throughput of the CR system. In this regard, Liang et al. constructed an optimization function using the ratio of sensing time and transmission time, and detection probability in the SU's frame, and showed that the maximum data throughput of the optimization function can be found by concavity [9]. It is also proposed to specify the operation region for the bounded false alarm assuming the characteristic of the PU activity and to find the optimal sensing time and period in the operation region. Lee and Akyildiz estimated the detection probability, false alarm probability, expected interference ratio, and lost spectrum opportunity by assuming the PDF (probability density function) of the busy/idle times of the PU, and calculated the guaranteed operation interval related to the SU's sensing time and the data frame time through the constraints calculated [10]. In addition to the conventional optimization considering tradeoff between sensing time and period, an effort to further reduce the interference to the PU was proposed. Choi and Yoo defined the PU as an unprotected primary user transmission (UPT) in case the SU cannot detect the PU in the transmission interval of the SU [11, 12]. In addition to the constraints related to the existing time-constrained detection probability, the UPT constraint is additionally used to optimize the sensing schedule. In the multi-input multi-output (MIMO)-based CR system for increasing the channel capacity, spectrum sensing utilizing the characteristics of the MIMO system has also been proposed. In order to solve the tradeoff between sensing and transmission time, Lee and Cho proposed a system in which a MIMO-based SU can perform sensing and data transmission simultaneously using zero forcing (ZF) [13]. Moghimi et al. proposed a system that divides the receiving stream of a MIMO-based SU and operates the data reception stream as a tradeoff for sensing and receiving, while the rest is dedicated to sensing [14].

In wireless communication networks composed of various communication systems, interference is an unavoidable phenomenon. In order to solve this problem, multiple access methods such as time division multiple access, frequency division multiple access, code division multiple access, and space division multiple access are used to make the signals orthogonal to each other in terms of time, frequency, and spatial domain. These methods can avoid interference by dividing the resources in each area, but cannot use enough of the capacity that the channel can provide. In this context, interference alignment (IA) has recently attracted attention as a technique capable of eliminating interference between multiple links and maximizing transmission capacity. In an existing communication system, since each user pair cannot know information about other users in the network, the optimal strategy is to maximize its own transmission rate. Thus, the sum of data rates in the network increases to the same order of a single communication link. However, using IA, the sum rate increases linearly with the number of users at high SNR (signal-to-noise ratio). The IA arranges all interference in a common subspace of a total received signal space in a receiver configured with a multi-antenna system. It separates the interference space from a desired signal space so that a plurality of transceivers can operate at the same time and or same frequency. Applying the IA to the CR was recently studied because of the advantages in eliminating interference with the PU and removing mutual interference between the SUs to increase the transmission capacity of the SUs [15].

The IA is combined with the CR system to divide the signal space and the interference space so that the interference can be avoided between PU and SU signals. Amir et al. analyzed the degree of freedom (DoF) available in IA-based PU and SU networks and maximized the transmission rate of the PU through water-filling [16]. Zhou et al. proposed optimizing the precoding matrix and power allocation to increase the transmission rate of SUs in a network of single PUs and multiple SUs [17]. Men et al. proposed an algorithm that guarantees the transmission performance of the PU using a partial IA algorithm [18]. IA-based CR can be applied to applications considering interference between heterogeneous systems. Chatzinotas and Ottersten applied IA to small cells to mitigate interference in macrocell base stations in small cells [19], and Huang et al. investigated a joint opportunistic interference avoidance scheme using the interweave paradigm-based Gale-Shapley spectral-sharing scheme to mitigate interference between a macrocell network and a femtocell network [20]. Sharma et al. proposed a coordinated and uncoordinated approach in a system consisting of monobeam and multibeam satellites as well as macrocell and small-cell systems [21]. On the other hand, methods of accessing the information of the PU and performing IA were proposed. Chen et al. proposed a system that helps the PU to transfer data while the SU uses a DoF that is not used by the PU [22], and Guler and Yener collected all channel information and proposed an interference technique using successive semi-definite programming (SDP) relaxation [23]. And Perlaza et al. allow the SU to transmit signals without interfering with the PU through the remaining eigenmodes

that are not used by the PU [24]. Hasani-Baferani et al. enabled the SU to perform IA by providing a femtocell with eigenmode information of the PU in a macrocell and femtocell system [25]. Zhao et al. optimized the sum rate of SUs, limiting it to the transmission rate threshold of the PU so that the sum rate of the entire network is optimized according to PU requirements [26].

Previous researches on the optimization of sensing and data transmission time cannot fundamentally block the interference to the PU because it cannot continuously sense the spectrum. Also, assuming the operating characteristics of the PU cannot be a reasonable assumption because the PU cannot be continuously observed or the signal of SUs cannot be transmitted while they process spectrum sensing even if it can continuously observe the spectrum for a while. Meanwhile, even if sensing is continuously performed through the MIMO-based CR system, still secondary systems need the optimization of the sensing and transmission time, and this optimization is very difficult to derive as a closed form specially in dynamic wireless environments in terms of time-varying channel characteristics and primary activities. In a CR system research with IA, if it is not possible to include the PU in IA or perform IA process by providing PU information to the CR, the conventional IA-based CR system cannot achieve the desired interference alignment performance. Therefore, in this paper, we design a designated sensor in an IA-based SU system that uses a conventional IA algorithm to limit transmission signals of SUs to the interference space and to sense the PU through the remaining signal space. Since the sensing and data transmission roles are separated, continuous sensing is possible, and the tradeoff problem of sensing time and transmission time disappears. The problem of not detecting the PU during data transmission also disappears. However, since the sensing role is limited to a specific SU, another problem arises in that the sensing result should be transmitted to other SUs. In this paper, we propose a dynamic adjustable sensing report interval control mechanism using reinforcement learning.

There is a tradeoff regarding the period of sensing interval (or the sensing result transmission). If it is too long, interference time to the PU increases due to the transmission time increases for the SUs; conversely, if it is too short, transmission performance of the SUs decreases. In this regard, this paper proposes an algorithm that dynamically determines the sensing time and reporting interval of sensing result by using Q-learning based on reinforcement learning for interference control in target-level and enhancement of SU's transmission opportunities. Meanwhile, the interference ratio and the transmission opportunity loss ratio are defined as the criteria for selecting the sensing time and the reporting interval according to the performance of the system. From the busy and idle time that are statistical characteristics of the PU, interference ratio and the transmission opportunity loss ratio can be estimated precisely from only the sensing results without any knowledge of operating characteristics about PU. First, in order to minimize the interference to the PU, the sensing is basically set to satisfy the required detection probability preferentially so that the interference to the PU can be ensured in the sensing step. In the reward design of Q-learning, the target interference ratio value is used so that the interference resides in desired range. Also, the target transmission opportunity loss ratio value is used to secure the transmission opportunity of the SU. Based on the designed reward, Q-learning dynamically selects the sensing time and reporting interval time to operate the system in the selected interference ratio and loss ratio range.

Since the Q-learning is a model-free reinforcement learning technique, Q-learning could be very fascinating method for spectrum sensing in time-varying environment. Liwang et al. used Q-learning to minimize mutual interference between SUs according to the sensing order [27]. Jan et al. has assigned a sub-band for spectrum sensing order to obtain high throughput while minimizing the number of sub-bands that the SU should sense [28]. Das et al. has cooperatively calculated the reward for the idle and busy states of the channel and evaluated the priority of the channel [29]. Yang et al. considered hardware reconfiguration energy consumptions and time delays when selecting a sub-band for wide-band sensing [30]. Oliver et al. solved the throughput optimization for the sensing and transmission time with Q-learning [31]. In this paper, we can observe the activity of the PU continuously so that the interference ratio and the transmission loss ratio can be calculated. Therefore, we can use Q-learning to select the appropriate sensing time and reporting interval to meet the primary protection and secondary throughput requirements in a dynamic environment.

In this paper, we describe proposed system model in Section 2 and examine the feasibility of the proposed IA-based sensing system through DoF analysis. Section 3 shows the real-time estimation method for the interference ratio and transmission opportunity loss ratio used as states related with Q-learning. Section 4 compares the proposed system with conventional schemes, and it shows that the proposed method provides stable and desired operation in the dynamic wireless environments. Finally, the conclusion is made in Section 5.

2 Proposed IA-based sensing structure for continuous spectrum sensing

2.1 System model

Typical IA in cognitive radio system is shown in Fig. 1 [18]. There are K s secondary transmitter-receiver pair and one SU with MIMO-CR interface and one primary

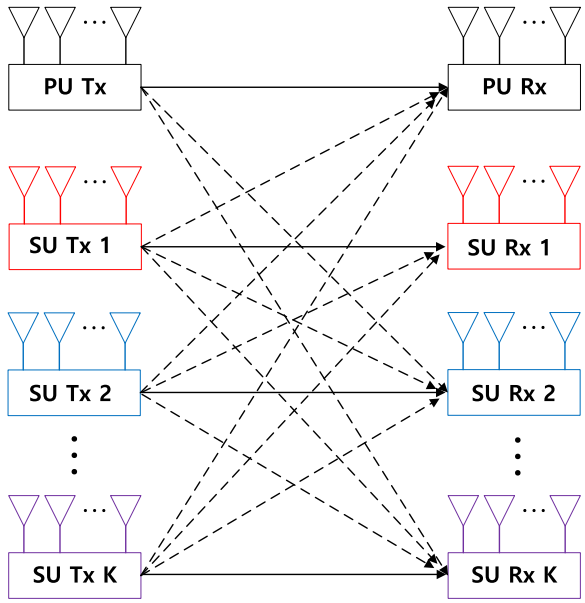

Fig. 1 General interference alignment in CR system. The PU participates in the interference alignment of the transceiver pairs of the CR system and they can transmit data without interfering with each other

transmitter. All the SUs share the transmission resource with the PU at the same time. Each transmitter and receiver has M and N antennas, respectively. $\mathrm{H}^{[ij]} \in \mathbb{C}^{N \times M}$ represents the channel between the jth transmitter and the ith receiver, where $i, j \in \{0, 1, \ldots, K\}$ and 0th user represents the PU. All the elements of $\mathbf{H}^{[ij]}$ are independent and identically distributed (i.i.d.) and follow complex Gaussian distribution with zero mean and unit variance $\mathcal{CN}(0,1)$. Then, the received signal at the ith receiver is expressed as:

$$\mathbf{y}^{[i]} = \mathbf{H}^{[ii]}\mathbf{x}^{[i]} + \sum_{j=0, j\neq i}^{K} \mathbf{H}^{[ij]}\mathbf{x}^{[j]} + \mathbf{z}^{[i]} \tag{1}$$

where $\mathbf{x}^{[i]}$ expresses the transmitted symbols of user i and $\mathbf{z}^{[i]} \in \mathbb{C}^{N \times 1}$ represents the circularly symmetric additive white Gaussian noise vector with $\mathcal{CN}(0, \sigma^2 \mathbf{I}_N)$, in which σ^2 is noise variance and $\mathbf{I}_N$ is an identity matrix.

In the IA system consisting of a pair of MIMO-based transceivers, the transmitter controls the precoding matrix so that the transmitted signal is limited to the interference space at an undesired receiver, and the receiver controls the decoding matrix to remove undesired received signals and to recover the signal. This IA design conditions can be represented as follows:

$$\mathbf{U}^{[i]^*}\mathbf{H}^{[ij]}\mathbf{V}^{[j]} = 0 \tag{2}$$

$$\mathrm{rank}\left(\mathbf{U}^{[i]^*}\mathbf{H}^{[ii]}\mathbf{V}^{[i]}\right) = d^k, \forall i \neq j \tag{3}$$

where d^k is the desired number of streams of user i. $\mathbf{V}^{[j]} \in \mathbb{C}^{M \times d}$ and $\mathbf{U}^{[i]} \in \mathbb{C}^{N \times d}$ denote precoding matrix of jth user and decoding matrix of ith user, respectively. $\mathbf{U}^{[i]^*}$ is the conjugate transpose of $\mathbf{U}^{[i]}$.

We can represent the received signal recovered by decoding matrix and adjusted by precoding matrix from the IA design conditions as follows.

$$\tilde{\mathbf{y}}^{[i]} = \underbrace{\mathbf{U}^{[i]^*}\mathbf{H}^{[ii]}\mathbf{V}^{[i]}\mathbf{s}^{[i]}}_{\text{desired signal}} + \underbrace{\sum_{j=0, j\neq i}^{K} \mathbf{U}^{[i]^*}\mathbf{H}^{[ij]}\mathbf{V}^{[j]}\mathbf{s}^{[j]}}_{\text{interference signals}} + \underbrace{\mathbf{U}^{[i]^*}\mathbf{z}^{[i]}}_{\text{noise}} \tag{4}$$

where $\mathbf{s}^{[i]}$ is the transmission signal of the ith transmitter. From the IA condition, interference signals term of (4) is eliminated.

In this paper, we designate one of the SUs as a sensor node only responsible for spectrum sensing, as shown in Fig. 2, in order to eliminate dependence on the PU information in the CR system for IA.

The designated sensor node should sense the primary signal continuously so that it cannot transmit or receive data during the sensing process. It may also consume

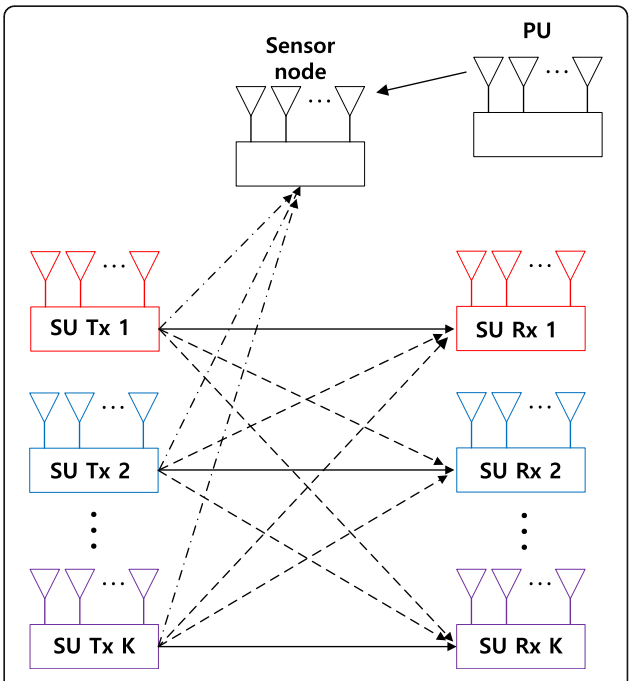

Fig. 2 Proposed system with interference alignment in CR system. The sensor node can sense the PU by using interference alignment even while the SUs are transmitting data use, and it will notify whether or not to transmit to the SUs depending on whether or not the PU is present

more energy than other secondary nodes. Therefore, a new sensor node needs to be selected after certain given time. In wireless ad hoc networks or wireless sensor networks, there have been similar studies on selecting the cluster head (CH) [27, 32–34]. In order to select the CH, various parameters can be considered, such as the number of member nodes covered by the CH, the current residual energy level, and the history of CH nodes. In this paper, the sensor is selected by considering the residual energy and energy draining rate of every node as in [34]. In this paper, for the selection of spectrum sensing node, the node with the largest ratio of residual energy to draining rate of energy is selected.

$$\text{sensor}(t + T_{\text{sel}}) = \underset{i}{\text{argmax}} \left(\frac{E_i}{D_i} \right) \tag{5}$$

where T_{sel} is a cycle for selecting sensor node. E_i and D_i are the residual energy and the draining rate of energy of node i, respectively.

The use of the designated sensor node can reduce the sensing overhead of other secondary nodes, but it may bring sensing accuracy degradation in some wireless scenarios. When a cooperative sensing method is used in the conventional CR system, by combining each SU's sensing result, it can increase the sensing accuracy and detect hide primary transmitters. Even though the sensor node in the proposed method is the only node that senses the primary signal, by performing consecutive spectrum sensing as a form of sequential chaining of a fixed sensing time, in which the fixed sensing time is determined to satisfy the required minimum primary detection probability, our proposed method also can combine time-domain multiple sensing results. It can compensate the lack of physical cooperative sensing.

As with the usual IA scheme, the sensor performs the IA process with other SUs to limit signals from other SUs to interference space and remove them through the decoding matrix. The remaining signal space can be used to sense the PU. As the sensing role is dedicated to a particular sensor, other SUs do not need to participate on spectrum sensing and also can transmit data without wasting of time for sensing.

To satisfy the required primary detection probability, a fixed sensing time slot (t_s) is determined as in (27). The sensor node performs spectrum sensing every consecutive sensing time slot. In this paper, to notify the sensing result by the sensor node to all CR SUs, we propose two mechanisms.

i. Periodic notification (default mode): the sensor node broadcasts sensing report which includes primary detection information at every predetermined sensing reporting interval. When SUs receive the primary detection notification, they should not transmit data until the next report broadcasting time. The sensing reporting interval is represented as t_r.
ii. Notification using dedicated control channel transceiver: every secondary nodes including sensor node have dual transceivers, in which one is for data transmission (or spectrum sensing) and the other one is control signal exchange. When the sensor node detects the primary signal, it transmits detection notification signal on the dedicated narrow band channel, and other SUs seize their data transmission. If the data channel returns to idle, then the sensor also send channel idle notification and then SUs can again utilize the data channel.

The spectrum sensing and primary detection comparison for the conventional CR system and proposed IA-based spectrum sensing system is represented in Fig. 3. Figure 3a represents the primary system activities as a form of busy with times. Figure 3b shows the conventional CR system which uses a fixed spectrum sensing time and interval. The conventional CR system can only sense the primary signal only during the short sensing time so that if the primary appears during the secondary data transmission time (i.e., between two consecutive sensing times), then the secondary system will give harmful interference to the primary system. As shown in Fig. 3c, the designated sensor node can continuously sense the primary signal. At time (1), the sensor node broadcasts the primary non-detection report to SUs so that SUs can utilize the data channel using interference alignment. At time (2) and time (3), the sensor node detects the primary signal so that it sends the primary detection report at (4). As we can see in Fig. 3c, SUs can seize their transmission until the primary signal is not detected. In conventional CR in Fig. 3b, since SUs are not able to detect the primary signal during the short sensing time, they send data and cause strong interference to the primary user. Figure 3d shows the case that dual transceivers are used. At time (5) when the sensor node detects the primary signal, it can immediately send the detection notification using the dedicated control channel. And when the data channel returns to idle at time (6), the sensor node notifies the non-detection notification to SUs and SUs utilize the data channel again. Therefore, secondary node's data throughput is enhanced.

The sensing time t_s and sensing reporting interval t_r impact on not only primary protection performance but also secondary system data transmission opportunity. The longer the sensing time, the lower miss detection and false alarm probabilities are obtained. On the other hand, the shorter sensing time results in the higher miss

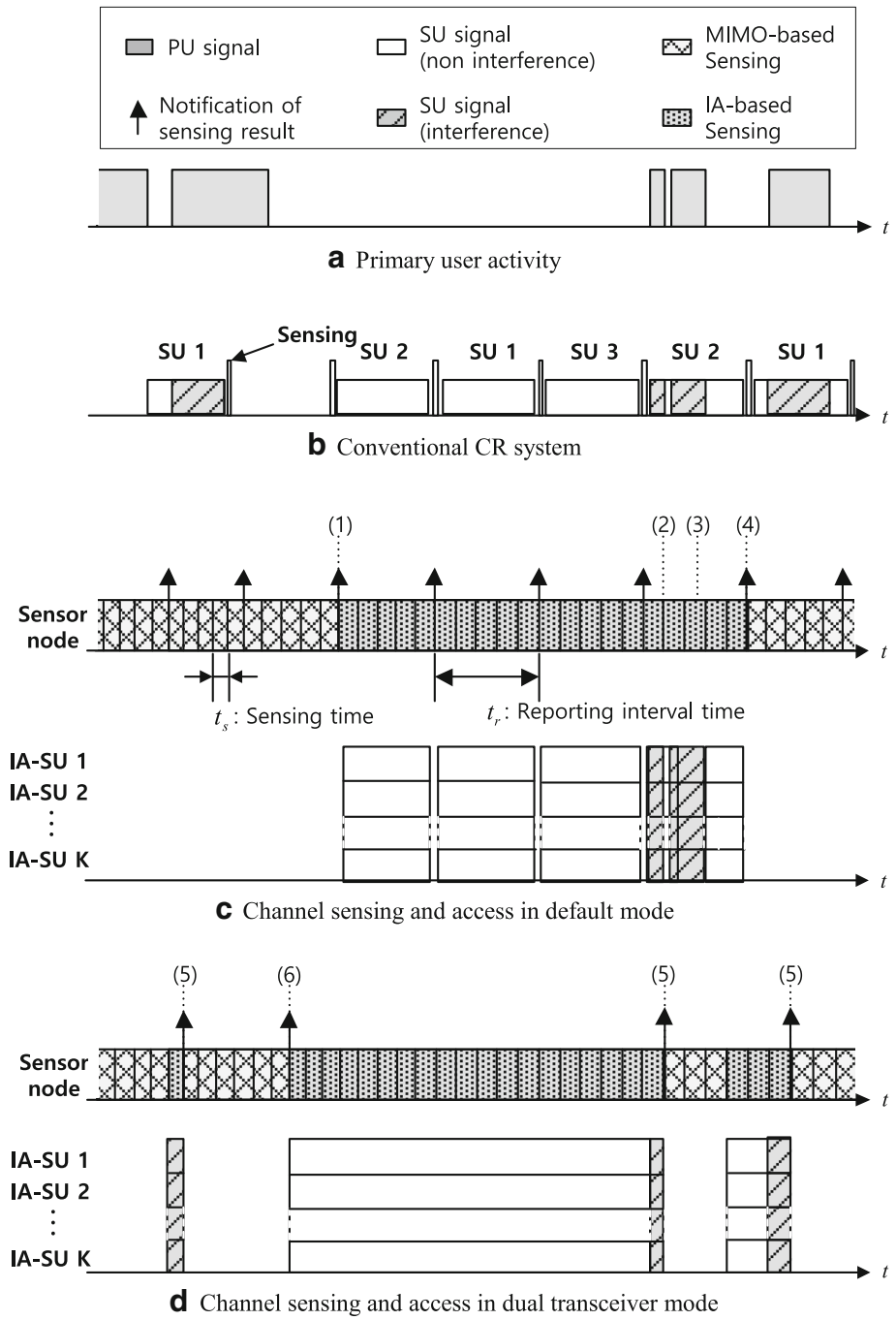

Fig. 3 Conventional CR system and proposed IA-based spectrum sensing. According to the busy/idle state of PU in **a**, this figure represents the process of conventional CR system in **b**, default mode in **c**, and dual transceiver mode in **d**. Since the conventional CR system can only sense the primary signal only during the short sensing time, the secondary system gives harmful interference to the primary system during the data transmission as shown in **b**. As shown in **c**, the sensor node can continuously sense the primary signal. At time (1), the sensor node broadcasts the primary non-detection report to SUs so that SUs can utilize the data channel using interference alignment. At time (2) and time (3), the sensor node detects the primary signal so that it sends the primary detection report at (4). As we can see in **c**, SUs can seize their transmission until the primary signal is not detected. **d** The case that dual transceivers are used. At time (5) when the sensor node detects the primary signal, it can immediately send the detection notification using the dedicated control channel

detection and false alarm probabilities especially at low signal-to-noise ratio (SNR) of the longer sensing reporting interval makes the more transmission opportunity for secondary users; however, it generates the higher possible interference to primary users. In CR wireless network, the primary activity and wireless channel condition vary dynamically so that it is very difficult to derive the optimal sensing time and reporting interval. Therefore, in this paper, we propose a new dynamic optimal parameter control using reinforcement learning. The multi-objective function of the secondary system is given as in (6), in which the multi-objective function consists of three reward functions: interference ratio reward, transmission opportunity loss ratio, and overhead for sensing. The reward functions will be explained in detail in Section 3.3. Therefore, the proposed method derives the optimal (t_s^*, t_r^*) value that maximizes the multi-objective function with subject to primary protection and secondary throughput requirements.

$$\begin{aligned} &\text{Maximize} : f_{\text{intf}}(R_I) + f_{\text{loss}}(R_L) + f_{\text{overhead}}(t_s, t_r) \\ &\text{Find} : t_s^*, t_r^* \\ &\text{Subject to} : P_d \geq P_d^{\text{th}}, R_I \leq R_I^{\text{th}}, R_L \leq R_L^{\text{th}} \end{aligned} \tag{6}$$

where $f_{\text{intf}}(R_I)$ and $f_{\text{loss}}(R_L)$ are the functions of interference and transmission opportunity loss ratio; $f_{\text{overhead}}(t_s, t_r)$ is the function of the overhead related to sensing time t_s and reporting interval (integer multiple of t_s); t_s^*, t_r^* are the optimal spectrum sensing time and reporting interval, respectively; P_d^{th} is the required primary detection probability P_d; R_I^{th} and R_L^{th} are the tolerable interference ratio and secondary transmission opportunity loss ratio, respectively.

The main novel features of the proposed system architecture are as follows:

1. The dedicated sensor is responsible for the sensing function and can operate spectrum sensing by IA process when SUs transmit the signal so that the operation of the PU can be continuously observed.
2. We specify the target detection probability to basically satisfy the detection probability and operate in the range that satisfies the interference ratio and the secondary transmission opportunity loss ratio.
3. We use the Q-learning to determine sensing time and reporting interval dynamically and design the suitable reward function.

2.2 Interference alignment and degree of freedom in the proposed system

A minimum DoF must be ensured for each transceiver pair to communicate using IA process. We derive the condition of DoF that the proposed system can obtain. In addition, this section provides a theoretical basis for the sensor node to perform sensing while the SU is transmitting. Suppose there is a MIMO-CR interference network with K SUs, one sensor and one PU in Fig. 4. It is assumed that SU's IA network is consist of symmetric (i.e., same transmission antennas and receive antennas). The 0th SU is a sensor, and each transmitter and receiver of the SU has M and N antennas. Then, received signals at the sensor and the ith SU receiver are as shown in (7) and (8):

$$\mathbf{y}^{[0]} = \mathbf{H}^{[0p]}\mathbf{x}^{[p]} + \sum_{j=1}^{K} \mathbf{H}^{[0j]}\mathbf{x}^{[j]} + \mathbf{z}^{[0]} \tag{7}$$

$$\mathbf{y}^{[i]} = \mathbf{H}^{[ii]}\mathbf{x}^{[i]} + \sum_{j=0, j\neq i}^{K} \mathbf{H}^{[ij]}\mathbf{x}^{[j]} + \mathbf{z}^{[i]} \tag{8}$$

where $\mathbf{x}^{[p]}$ and $\mathbf{x}^{[i]}$ are the transmission symbol of PU and SU i, $\mathbf{z}^{[0]}$, $\mathbf{z}^{[i]}$ are circularly symmetric additive white Gaussian noise vectors, with $\mathcal{CN}(0, \sigma^2\mathbf{I}_N)$. $\mathbf{H}^{[ij]} \in \mathbb{C}^{N\times M}$ represents the channel between the jth transmitter and the ith receiver, where $i, j \in \{0, 1, \ldots, K\}$, K represents the number of SUs, and the index p represents the PU. All elements of $\mathbf{H}^{[ij]}$ are i.i.d. distributed and follow $\mathcal{CN}(0, 1)$. We assumed a quasi-static channel, i.e., the channel realization remains fixed throughout the duration of transmission.

In order to eliminate interference in a sensor and each SU, we use a decoding matrix and the received signal with d data streams of the ith user is recovered as follows:

$$\begin{aligned} \tilde{\mathbf{y}}^{[0]} = \mathbf{U}^{[0]^*}\mathbf{H}^{[0p]}\mathbf{V}^{[p]}\mathbf{s}^{[p]} + \sum_{j=1}^{K} \mathbf{U}^{[0]^*}\mathbf{H}^{[0j]}\mathbf{V}^{[j]}\mathbf{s}^{[j]} \\ + \mathbf{U}^{[0]^*}\mathbf{z}^{[0]} \end{aligned} \tag{9}$$

$$\tilde{\mathbf{y}}^{[i]} = \mathbf{U}^{[i]^*}\mathbf{H}^{[ii]}\mathbf{V}^{[i]}\mathbf{s}^{[i]} + \sum_{j=0, j\neq i}^{K} \mathbf{U}^{[i]^*}\mathbf{H}^{[ij]}\mathbf{V}^{[j]}\mathbf{s}^{[j]} + \mathbf{U}^{[i]^*}\mathbf{z}^{[i]} \tag{10}$$

where $\mathbf{s}^{[p]}$, $\mathbf{s}^{[j]}$ are transmission signals of the PU and the jth SU. $\mathbf{V}^{[p]}$ and $\mathbf{V}^{[j]}$ are the precoding matrix of the PU and jth SU, and $\mathbf{U}^{[0]}$, $\mathbf{U}^{[i]} \in \mathbb{C}^{N\times d}$ are the decoding matrix of the sensor and the ith user. To completely remove interference from the SUs to the sensor or between the SUs, $\mathbf{V}^{[j]}$, $\mathbf{U}^{[0]}$, and $\mathbf{U}^{[i]}$ must satisfy the following conditions:

$$\mathbf{U}^{[0]^*}\mathbf{H}^{[0j]}\mathbf{V}^{[j]} = \mathbf{0}_d \tag{11}$$

$$\mathbf{U}^{[i]^*}\mathbf{H}^{[ij]}\mathbf{V}^{[j]} = \mathbf{0}_d \tag{12}$$

$$\text{rank}\left(\mathbf{U}^{[i]^*}\mathbf{H}^{[ii]}\mathbf{V}^{[i]}\right) = d^{[i]} \tag{13}$$

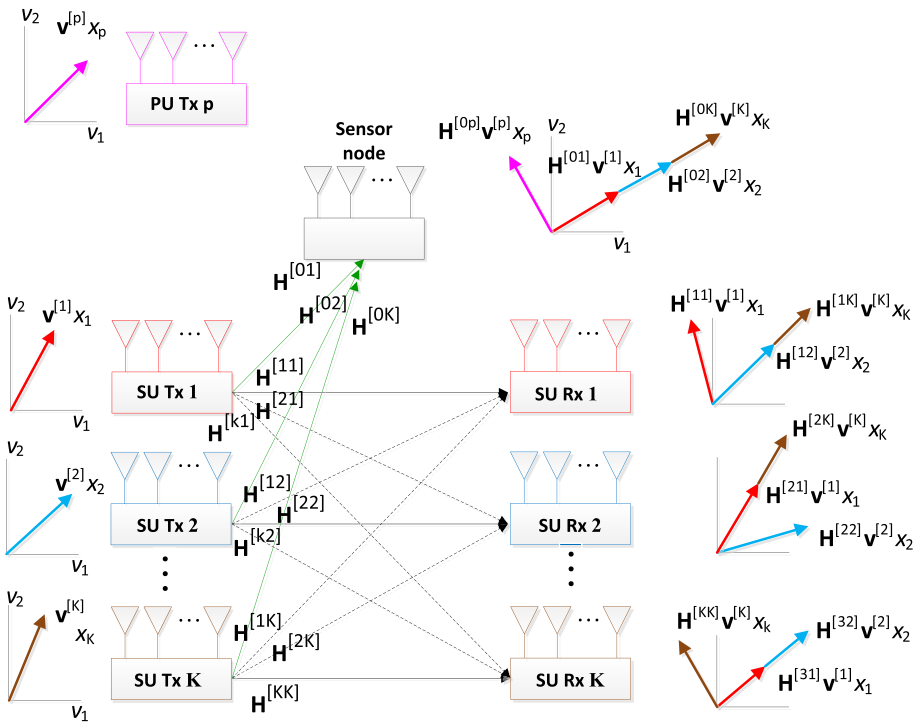

Fig. 4 Interference alignment in sensor. In the proposed system, the SU transceiver pair behaves like a normal interference alignment, but added sensor node removes the transmission signals of other SUs in the interference space. Therefore, the signal of the PU can be sensed in the remaining signal space of the sensor node

$$\operatorname{rank}\left(\mathbf{U}^{[0]^*}\mathbf{H}^{[0p]}\mathbf{V}^{[p]}\right) = d^{[0]}, \forall i \neq j, i \neq 0, j \neq 0, \forall i, j \in \{1, 2, \ldots, K\} \tag{14}$$

where $d^{[i]}$ is the desired number of stream of user i.

Equations (11) and (12) show that interference in receiving signal dimension of sensor and SU receivers should be zero. Equations (13) and (14) represent the number of signal stream that each SU transceiver pair and sensor node can acquire. From these constraints, the DoF condition is expressed by (15):

$$d \leq \frac{N}{K+1} + \frac{KM}{(P+K)(K+1)} \tag{15}$$

Proof. See the Appendix.

Therefore, in the network with satisfying the DoF condition from (11), the sensor node can remove the interference from the signals of other SU and sense the PU signal.

To fulfill the requirements in (11) and (12), the iterative IA algorithms in [35, 36] can be adopted with some modifications. The sensor should minimize the total leakage interference that remains after canceling the interference by decoding. Other SUs can obtain the precoding and decoding matrix by the maximum SINR algorithm considering the total leakage interference of the sensor. By fixing all $\mathbf{V}^{[i]}$, we can solve $\mathbf{U}^{[j]}$ as

$$\mathbf{U}^{[j]} = \nu_{\max}\left(\frac{\mathbf{H}^{[ji]}\mathbf{V}^{[j]}\mathbf{V}^{[j]^*}\mathbf{H}^{[ji]^*}}{\sum_{i\neq j}\mathbf{H}^{[ji]}\mathbf{V}^{[i]}\mathbf{V}^{[i]^*}\mathbf{H}^{[ji]^*} + \sigma^2\mathbf{I}_N}\right) \tag{16}$$

where $\nu_{\max}(\cdot)$ denotes the dominant eigenvector when the eigenvalues are real.

Reversely, by fixing all $\mathbf{U}^{[j]}$, we can solve $\mathbf{V}^{[i]}$ as

$$\mathbf{V}^{[i]} = \nu_{\max}\left(\frac{\mathbf{H}^{[ii]^*}\mathbf{U}^{[i]^*}\mathbf{U}^{[i]}\mathbf{H}^{[ii]}}{\sum_{j\neq i}\mathbf{H}^{[ji]^*}\mathbf{U}^{[j]^*}\mathbf{U}^{[j]}\mathbf{H}^{[ji]} + \mathbf{Q}_0}\right) \tag{17}$$

where $\mathbf{Q}_0$ is the interference covariance matrix at the sensor.

The interference covariance matrix at the sensor is

$$\mathbf{Q}_0 = \sum_{i\neq 0}^{K}\mathbf{U}^{[0]^*}\mathbf{H}^{[0i]}\mathbf{H}^{[0i]^*}\mathbf{U}^{[0]} \tag{18}$$

The decoder minimizing the total leakage interference at the sensor is

$$\mathbf{U}_0 = \nu_{\min}(\mathbf{Q}_0) \tag{19}$$

where $\nu_{\min}(\cdot)$ is the least dominant eigenvector.

Summarizing the process, the transmitters choose the initial precoders randomly and receivers choose the decoders maximizing SINR. The sensor node calculates the interference covariance matrix and chooses the decoding matrix. The transmitters choose the precoders by maximizing SINR by considering the total interference leakage at the sensor node. Then, the choices of decoding matrix of the receivers are followed and this sequence of processes continues to convergence.

2.3 Energy detection with/without interference alignment

Spectrum sensing in the proposed system stops transmission of SUs when PU is detected and performs general MIMO-based spectrum sensing. If the sensor determines that the PU is in an idle state, the SUs send the signal through IA and the sensor performs IA-based spectrum sensing that allows sensing during communication of the SUs. Therefore, in the proposed system, MIMO-based or IA-based spectrum sensing is selected according to the detection result of the PU signal. Since the parameters to be used to set thresholds for energy detection depend on the choice of sensing method, this section focuses on MIMO-based sensing and IA-based sensing.

If an SU does not transmit because the PU state is determined as busy state, the hypothesis from the received signal y_i of the sensor is expressed in (20):

$$\begin{aligned} &H_0 : y_i(n) = z_i(n) \\ &H_1 : y_i(n) = h_i s(n) + z_i(n), \text{where } 1 \le i \le N \end{aligned} \tag{20}$$

where H_0 represents the hypothesis corresponding to "no signal transmitted," and H_1 represents "signal transmitted." $s(n)$ is the signal waveform, and $z_i(n)$ is a zero mean additive white Gaussian noise (AWGN). The PU is assumed to phase shift keying (PSK) modulated signal. The channel coefficient h_i follows $\mathcal{CN}(0, \sigma_h^2)$, and z_i follows $\mathcal{CN}(0, \sigma_n^2)$; σ_h^2 and σ_n^2 are the variance in channel gain and Gaussian noise. N is the number of receiving antennas.

The test statistic for the energy detector is given by

$$Y = \sum_{i=1}^{N} \sum_{n=0}^{n_s - 1} |y_i(n)|^2 \tag{21}$$

where n_s is the samples of spectrum sensing.

We assumed the test statistic follows a Gaussian distribution under the central limit theorem. Therefore, each pdf of (21) under H_0 and H_1 is given by

$$\begin{aligned} Y|H_0 &\sim \mathcal{N}\left(\mu_{0,non-IA}, \sigma^2_{0,non-IA}\right), Y|H_1 \\ &\sim \mathcal{N}\left(\mu_1, non\text{–}IA, \sigma^2_{1,non-IA}\right) \end{aligned} \tag{22}$$

where $\mu_{0,\text{non–IA}} = N n_s \sigma_n^2$, $\sigma^2_{0,\text{non–IA}} = N n_s \sigma_n^4$, $\mu_{1,\text{non–IA}} = N n_s (P\sigma_h^2 \lambda_m + \sigma_n^2)$, $\sigma_1^2 = N n_s (P\sigma_h^2 \lambda_m + \sigma_n^2)^2$. λ_m is eigenvalue of the correlation matrix, and P is transmission power of the PU.

False alarm and detection probability for the non-IA case are given by

$$\begin{aligned} P_f^{non-IA} &= Pr\{Y > \varepsilon_{non-IA} | H_0\} = \mathcal{Q}\left(\frac{\varepsilon_{non-IA} - \mu_{0,non-IA}}{\sigma_{0,non-IA}}\right) \\ P_d^{non-IA} &= Pr\{Y > \varepsilon_{non-IA} | H_1\} = \mathcal{Q}\left(\frac{\varepsilon_{non-IA} - \mu_{1,non-IA}}{\sigma_{1,non-IA}}\right) \end{aligned} \tag{23}$$

The hypothesis for the received signal of the spectrum sensor produced from the decoding matrix when a SU transmits because the PU state is determined as idle is expressed with (24):

$$\begin{aligned} &H_0 : \tilde{y}_i(n) = \tilde{z}_i(n) \\ &H_1 : \tilde{y}_i(n) = \sum_{j=1}^{M_p} \tilde{\mathbf{G}}^{[ij]} s_j(n) + \tilde{z}_i(n), \text{where } 1 \le i \le N \end{aligned} \tag{24}$$

where $s_j(n)$ is the signal waveform from jth antenna of PU, and $\tilde{z}_i(n)$ is an AWGN. $\tilde{\mathbf{G}}^{[ij]}$ is the compound channel gain between the PU transmitter and the sensor, and M_p is the number of PU's transmit antenna. We assumed that the gain does not change for multiple CR frames and can be estimated blindly while the PU is known to be present. Each statistical pdf of (24) is given by

$$\begin{aligned} Y|H_0, \tilde{\mathbf{H}}_p &\sim \mathcal{N}\left(\mu_{0,IA}, \sigma^2_{0,IA}\right), Y|H_1, \tilde{\mathbf{H}}_p \\ &\sim \mathcal{N}\left(\mu_{1,IA}, \sigma^2_{1,IA}\right) \end{aligned} \tag{25}$$

where $\mu_{0,\text{IA}} = N n_s \sigma_n^2$, $\sigma^2_{0,\text{IA}} = N n_s \sigma_n^4$, $\mu_{1,\text{IA}} = N n_s (P\tilde{g}_m^2 \sigma_h^2 \lambda_m + \sigma_n^2)$, $\sigma^2_{1,\text{IA}} = N n_s (P\tilde{g}_m^2 \sigma_h^2 \lambda_m + \sigma_n^2)^2$. $\tilde{g}_m^2$ is the sum of $\tilde{\mathbf{G}}^{[ij]}$ on j indexes.

False alarm and detection probability for the IA process case are given by

$$\begin{aligned} P_f^{IA} &= Pr\{Y > \varepsilon_{IA} | H_0\} = \mathcal{Q}\left(\frac{\varepsilon_{IA} - \mu_{0,IA}}{\sigma_{0,IA}}\right) \\ P_d^{IA} &= Pr\{Y > \varepsilon_{IA} | H_1\} = \mathcal{Q}\left(\frac{\varepsilon_{IA} - \mu_{1,IA}}{\sigma_{1,IA}}\right) \end{aligned} \tag{26}$$

For a given pair of target probabilities $(P_d^{\text{th}}, P_f^{\text{th}})$, the number of required samples can be determined by

$$n_s = \left(\frac{Q^{-1}\left(P_f^{\text{th}}\right) - Q^{-1}\left(P_d^{\text{th}}\right)\left(\zeta \lambda_m \tilde{g}_m^2 + 1\right)^2}{\zeta \lambda_m \tilde{g}_m^2} \right)^2 / N \tag{27}$$

where ζ is SNR and $\tilde{g}_m^2 = 1$ when this represents about the non-IA case.

3 Control of interference and transmission opportunity loss through Q-learning

3.1 Probabilistic estimation of interference ratio and transmission opportunity loss ratio

In the proposed system, the sensing time and reporting interval time must be determined to protect the PU and to guarantee the SU transmission rate. To determine these sensing parameters appropriately according to the loss of PU and gain of CR system, we must be able to predict interference with the PU and the transmission opportunity access of the CR user as a result of the selected sensing parameters. Most of conventional sensing time and sensing period optimization algorithms have assumed that the statistics of the PU activation (busy and idle). In the proposed system, we estimate the required probabilistic parameters without any prior knowledge of PU operations by sensor node's observation using IA process. The performance of interference with the PU can be predicted by the interference ratio, which means the predicted rate at which the PU's busy state interval is interrupted by transmissions by CR users. The transmission performance of the CR users can be estimated by the transmission opportunity loss ratio, which is the idle state ratio of the PU that is not detected by the SU, compared to the transmission-possible interval.

Figure 5 shows the interference ratio. Figure 5a indicates the operation of the PU, and Fig. 5c indicates the operation of the SU according to the sensor node of Fig. 5b. As shown in Fig. 5b, the sensor node operates as IA-based sensing until the second sensing period, detects PU in the second sensing period, and switches to non-IA-based sensing. Finally, it confirms that the PU is in the idle state in the fifth sensing period and switches to IA-based sensing. Since the sensor node instructs the SU to stop transmission after the second sensing period by the k out of N rule, the transmission stream of the SU overlapping with the transmission interval of the PU in the second sensing interval interferes with the PU.

The interference ratio is estimated by the following equation:

$$R_I = \frac{\sum_{\text{window}}\{R_{I1}\}}{\sum_{\text{window}}\{R_{I1} + R_{I2}\}} \tag{28}$$

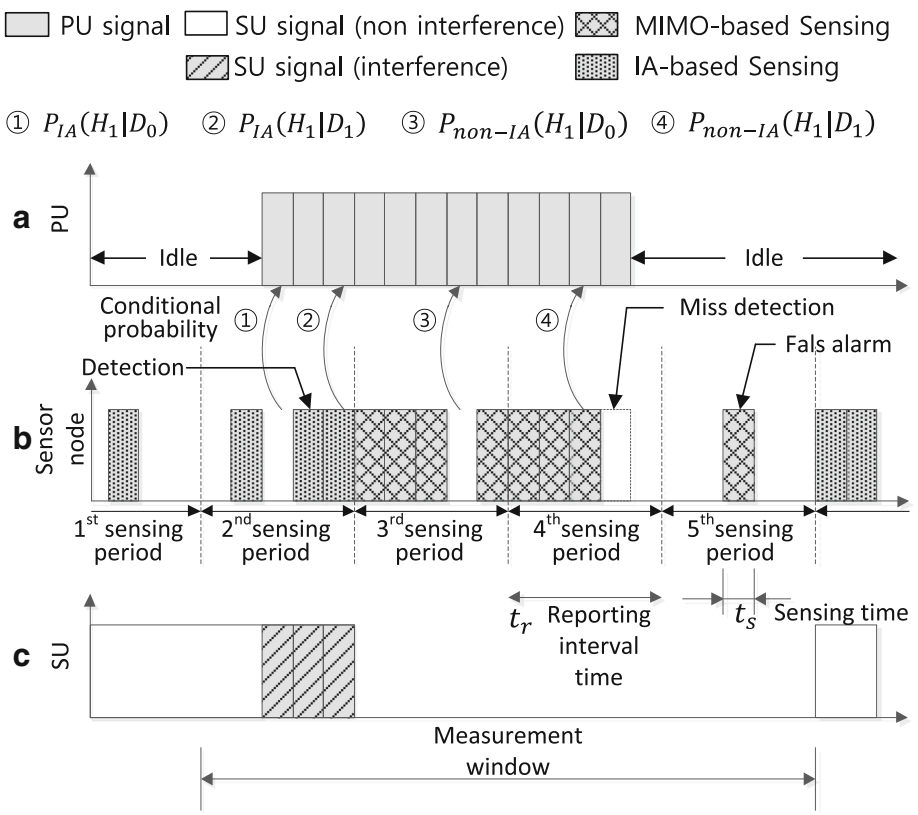

Fig. 5 Interference ratio. The interference ratio is the ratio between the conditional probability that the SU interferes with the PU and the PU exists within the measurement interval

$$R_{I1} = \sum_{SU_{on},D_0} t_s P_{IA}(H_1|D_0) + \sum_{SU_{on},D_0} t_s P_{IA}(H_1|D_1) \quad (29)$$

$$R_{I2} = \sum_{SU_{\text{off}},D_0} t_s P_{non-IA}(H_1|D_0) + \sum_{SU_{off},D_1} t_s P_{non-IA}(H_1|D_1) \quad (30)$$

where t_s denotes the sensing time, and D_0 indicates that PU is not detected, and D_1 indicates the PU is detected. H_0 and H_1 represent the hypothesis of the PU's presence.

The interference ratio is the ratio of the transmission interval of the SU overlap with the transmission interval of the PU over the transmission interval of the PU within the measurement window. In (28), the numerator is the interference probability and denominator is the probability of PU existence. In order to estimate the interference probability, we calculate the conditional probabilities and sum the results of that PU is busy from each sensing result (D_0, D_1) in the interval (SU_{on}) where the SU operates from the decision the PU is in idle state in the previous sensing time as show in (29). In the interval in which the SU operates, the sensor node performs the IA-based sensing. In order to estimate the probability of the PU busy state (H_1) in the measurement window, we compute the conditional probabilities and sum the results of that PU is busy for each sensing result (D_0, D_1) in non-IA-based sensing intervals in which the SU operates (SU_{on}) and the interval in which the SU is idle (SU_{off}). Therefore, the probability of PU existence in

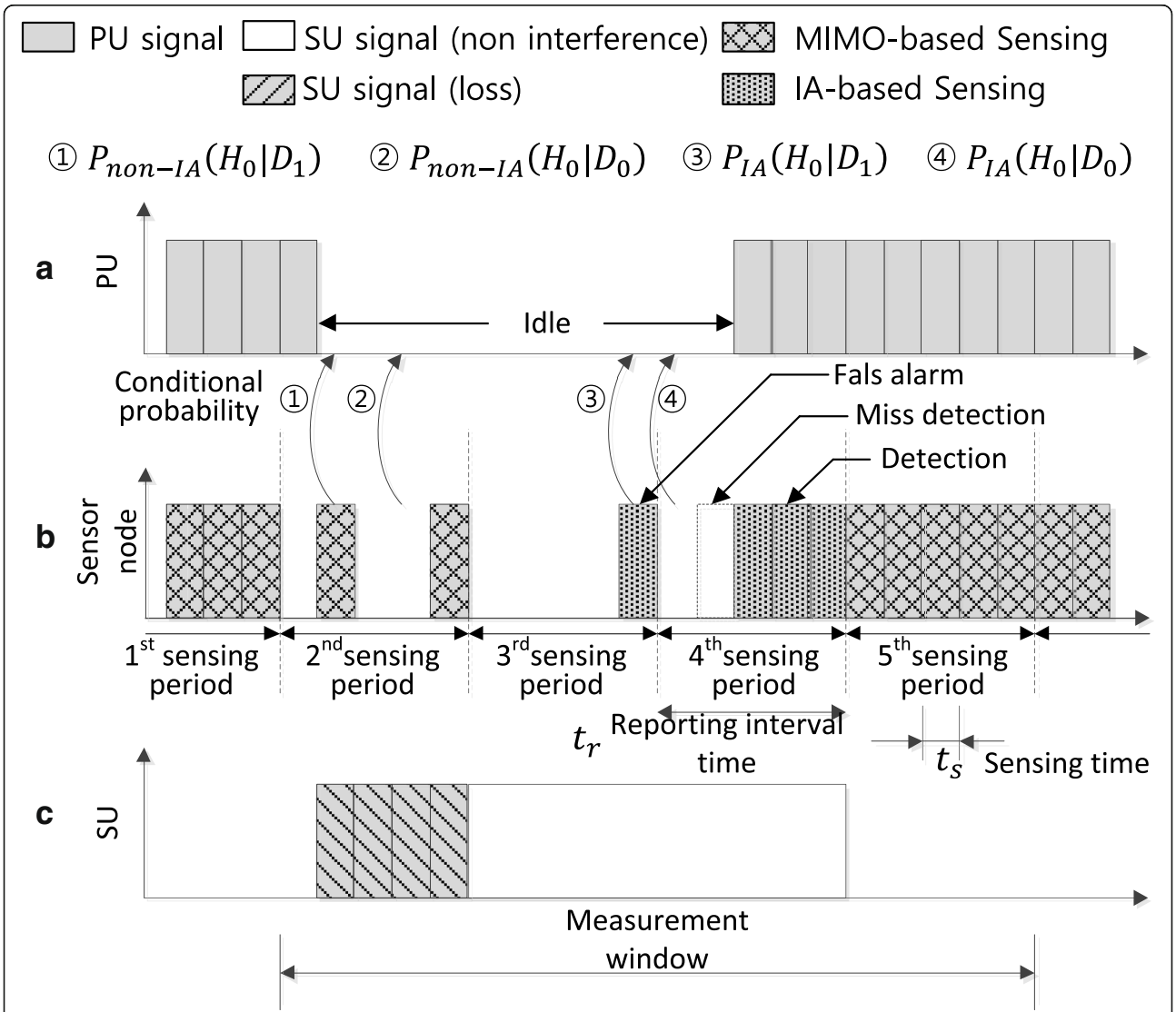

Fig. 6 Transmission opportunity loss ratio. The transmission opportunity loss ratio is the ratio between the conditional probability that the SU misses a transmission opportunity and the PU is idle within the measurement interval

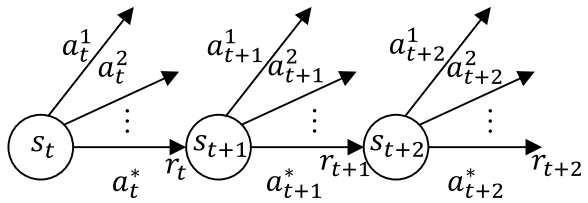

Fig. 7 General Q-learning mechanism. In general Q-learning mechanism, system moves to the next state using the optimal action in the corresponding state and selects the optimal action in the changed state and transitions to the next state

measurement window is expressed in the denominator of (28) as the sum of (29) and (30).

Figure 6 shows the transmission opportunity loss ratio. The PU switches to idle in the second sensing interval, but the sensor node instructs the SUs to operate from the third sensing interval by k out of N rule. Therefore, a disabled term in the second sensing period is the loss of transmission opportunity.

The transmission opportunity loss ratio of the SU is expressed in (31):

$$R_L = \frac{\sum_{\text{window}}\{R_{L1}\}}{\sum_{\text{window}}\{R_{L1}+R_{L2}\}} \tag{31}$$

$$R_{L1} = \sum_{SU_{off},D_1} t_s P_{non-IA}(H_0|D_1) + \sum_{SU_{off},D_0} t_s P_{non-IA}(H_0|D_0) \tag{32}$$

$$R_{L2} = \sum_{SU_{on},D_1} t_s P_{IA}(H_0|D_1) + \sum_{SU_{on},D_0} t_s P_{IA}(H_0|D_0) \tag{33}$$

The transmission opportunity loss ratio is the ratio of the interval that the SU does not transmit and for the interval in which the PU is idle within the measurement interval when the sensor node cannot detect the idle state of PU. In (31), the numerator is the probability where PU is idle when SU do not use the vacant time and the denominator is the probability where PU is idle in the measurement window. First, we calculate the conditional probabilities and sum of them that the PU is idle for each sensing results (D_0, D_1) in the idle interval of SU (SU_{off}) to estimate the probability of the interval in which the SU does not transmit despite the idle PU as shown in (32). In order to estimate the probability of the interval in which the PU is idle, we sum up the conditional probabilities of the PUs idle (H_0) for each of the sensing results (D_0, D_1) in intervals in which the SU operates and IA-based sensing is operated and in which the SU is idle and performs non-IA-based sensing. The estimate for the probability in the measurement window in which the PU is idle is expressed in the denominator of (31) as the sum of (32) and (33).

Each conditional probability from (28) to (33) can be expressed by Baye's rule as follows.

$$\begin{aligned} &P_{non-IA\,or\,IA}(H_0D_0) \\ &= \frac{P_{non-IA\,or\,IA}(D_0|H_0)P(H_0)}{P_{non-IA\,or\,IA}(D_0|H_0)P(H_0)+P_{non-IA\,or\,IA}(D_0|H_1)P(H_1)} \\ &= \frac{(1-P_{f,non-IA\,or\,IA})P_{off}}{(1-P_{f,non-IA\,or\,IA})P_{off}+P_{m,non-IA\,or\,IA}P_{on}} \end{aligned} \tag{34}$$

$$\begin{aligned} &P_{non-IA\,or\,IA}(H_0|D_1) \\ &= \frac{P_{non-IA\,or\,IA}(D_1|H_0)P(H_0)}{P_{non-IA\,or\,IA}(D_1|H_0)P(H_0)+P_{non-IA\,or\,IA}(D_1|H_1)P(H_1)} \\ &= \frac{P_{f,non-IA\,or\,IA}P_{off}}{P_{f,non-IA\,or\,IA}P_{off}+P_{d,non-IA\,or\,IA}P_{on}} \end{aligned} \tag{35}$$

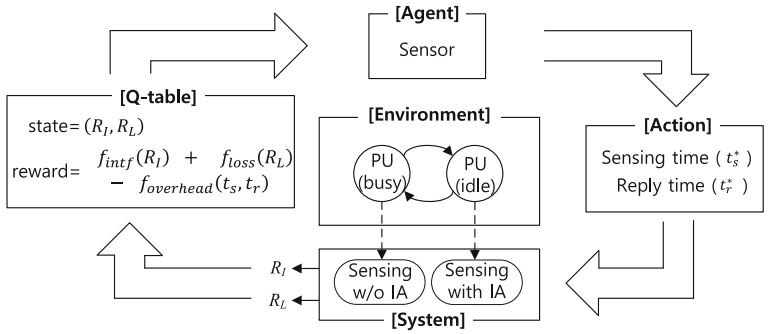

Fig. 8 Q-learning mechanism of the proposed scheme. In the proposed Q-learning system, the sensor acts as an agent and uses sensing time and reply time as an action. In an environment where the PU operates in an alternating busy/idle state, it obtains the interference ratio and the transmission opportunity loss ratio for the action, and uses it to change the state and calculate the reward

Table 1 Q-learning algorithm for the proposed scheme

Algorithm 1 Q-learning algorithm for the proposed scheme

1: **while** required packet exists
2: **if** time index > sensing window size
3: initialize the sensing results and obtained results
4: **else**
5: time index = time index +1
6: **while** time index < $n_{sensing}n_{reply}$
7: Determine sensing tool (with IA or without IA)
8: from the *k out of N* rule with $N = n_{sensing}n_{reply}$ sensing results
9: **end**
10: Obtain $P(D_0)$, $P(D_1)$ from the sensing results
11: Calculate $P(H_0)$, $P(H_1)$ from the simultaneous equations of $P(D_0)$,$P(D_1)$
12: Calculate R_I,R_L from the conditional probabilities and determine the state
13: Select an action a_t based on the optimal policy from the current state s_t
14: Obtain the immediate payoff R from action a_t
15: Observe the next state s_{t+1} by R_I,R_L
16: Update $\mathcal{Q}(s_t, a_t)$ based upon this experience as
17: $\mathcal{Q}(s_t, a_t) \leftarrow (1-\alpha)\mathcal{Q}(s_t, a_t) + \alpha\{r_t + \gamma \max_{a_{t+1}} \mathcal{Q}(s_{t+1}, a_{t+1})\}$
18: **end**
19: **end**

$$P_{non-IA\ or\ IA}(H_1|D_0) = \frac{P_{non-IA\ or\ IA}(D_0|H_1)P(H_1)}{P_{non-IA\ or\ IA}(D_0|H_1)P(H_1) + P_{non-IA\ or\ IA}(D_0|H_0)P(H_0)} = \frac{P_{m,\ non-IA\ or\ IA}P_{on}}{P_{m,\ non-IA\ or\ IA}P_{on} + \left(1-P_{f,\ non-IA\ or\ IA}\right)P_{off}} \tag{36}$$

$$P_{non-IA\ or\ IA}(H_1|D_1) = \frac{P_{non-IA\ or\ IA}(D_1|H_1)P(H_1)}{P_{non-IA\ or\ IA}(D_1|H_1)P(H_1) + P_{non-IA\ or\ IA}(D_1|H_0)P(H_0)} = \frac{P_{d,\ non-IA\ or\ IA}P_{on}}{P_{d,\ non-IA\ or\ IA}P_{on} + P_{f,\ non-IA\ or\ IA}P_{off}} \tag{37}$$

where each conditional probability can be the case of non-IA or IA-based sensing. P_{on} and P_{off} represent $P(H_1)$ and $P(H_0)$. They can be estimated from measured $P(D_0)$ and $P(D_1)$, as follows:

$$P(D_0) = P_{\text{non-IA or IA}}(D_0|H_0)P(H_0) + P_{\text{non-IA or IA}}(D_0|H_1)P(H_1) \tag{38}$$

$$P(D_1) = P_{\text{non-IA or IA}}(D_1|H_0)P(H_0) + P_{\text{non-IA or IA}}(D_1|H_1)P(H_1) \tag{39}$$

Through the simultaneous equations, $P(H_0)$ and $P(H_1)$ are calculated as follows.

$$P(H_0) = 1\text{–}P(H_1) \tag{41}$$

As a result, we can obtain $P(D_0)$ and $P(D_1)$ from the sensing result in the set window and obtain $P(H_0)$ and $P(H_1)$ from these simultaneous equations and then estimate the each conditional probability. Again, the conditional probability can be used to track the state of the PU and the CR system by interference ratio and transmission opportunity loss ratio for the selected sensing time and reporting interval time. In the next section, we propose a system dynamically operates by Q-learning. This system uses sensing time and reporting interval as an action. The response of an action is state which can be represented as interference ratio and transmission opportunity loss ratio.

3.2 Dynamic sensing parameter control using Q-learning

Q-learning is one of the off-policy techniques of reinforcement learning based on a Markov decision rule. It operates adaptively to the experienced environment and allows the system to operate dynamically to suit the desired purpose [37]. First, the object that recognizes and learns the surrounding environment is called an agent. The agent obtains a response from the environment after determining the action and recognizes where the agent belongs in the defined states. The mechanism in which general Q-learning operates is shown in Fig. 7. s_t, r_t, and a_t^i represent state, reward, and ith action at time t, respectively. The agent recognizes the state s_t and selects the action a_t^* which gives the maximum reward r_t among the selectable actions. The operation of Q-learning is performed by repeating this series of processes.

As shown in Fig. 8, there is an agent to perform decision-making and learning in the environment of the current state. The agent performs an action, evaluates the effect of the environment on the taken action, obtains the reward, and acts as a series of processes in which the state changes. In the proposed method, the agent is the sensor, and the sensor operates by using

$$P(H_1) = \frac{P_{\text{non-IA or IA}}(D_1|H_0)P(D_0) - P_{\text{non-IA or IA}}(D_0|H_0)P(D_1)}{P_{\text{non-IA or IA}}(D_1|H_0)P_{\text{non-IA or IA}}(D_0|H_1) - P_{\text{non-IA or IA}}(D_0|H_0)P_{\text{non-IA or IA}}(D_0|H_1)} \tag{40}$$

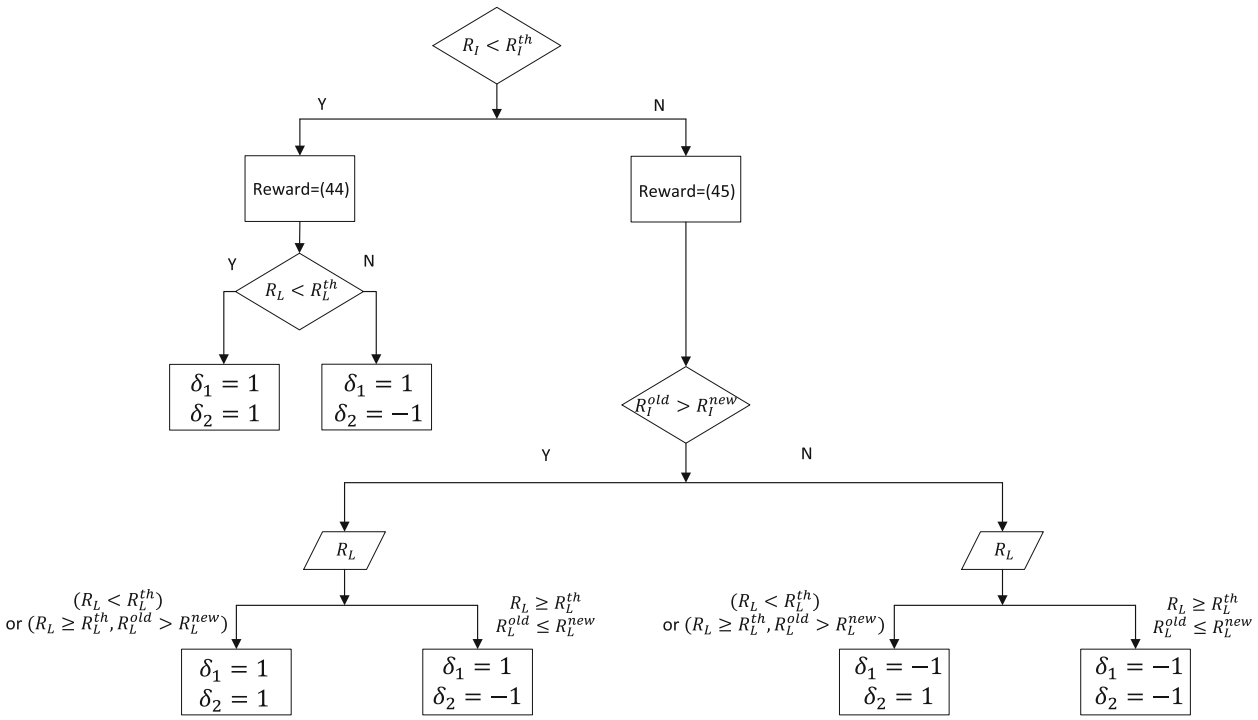

Fig. 9 Flow chart for reward function. We use (27) or (26) for the reward function depending on whether the interference ratio meets the constraint. If (27) is used, the sign of the parameter changes depending on whether or not the transmission opportunity loss ratio satisfies the restriction condition. If (26) is used, the sign of the parameter is changed according to the tendency of whether the interference ratio is improved or deteriorated. After that, the sign of the parameter is changed according to the tendency of the transmission opportunity loss ratio is improved or not for each divided case

sensing time and reporting interval as an action. When the sensor performs a specific action in the environment of repeated PU busy/idle state, the sensor can estimate R_I and R_L from the sensing results. After that, the sensor obtains the reward by using the gain or loss function related to R_I and R_L, and the overhead function of the action. The sensor recognizes the state change composed of the observed R_I and R_L and performs another action based on it. When the PU is busy, the SUs do not transmit, so the sensor performs basic MIMO-based sensing, and when it is idle, the sensor performs IA-based sensing because SUs transmit using IA.

In the Q-learning, the Q-table is used as a data base in which the agent selects action in a given environment and records information with the reward and state changes obtained from the selected actions. The Q-table records this information for (state, action) pairs. When the system starts to operate in a given environment, there is no information in the Q-table. The Q-table stores information on how to maximize the designed reward in the given environment in the Q-table as a series of process that obtain the reward through the selected action and change the state is repeated. A Q-table consists of rows representing the states and columns representing the actions. An action is a set of products of sensing time and reporting interval, which can be expressed as $\mathcal{A} = \{a_1, a_2, ..., a_t, ..., a_M\}$ where a_t is $\{a_t^s, a_t^r\}$, and t represents the serial number of the action; a_t^s is sensing time, and a_t^r is a multiple of the sensing time to express the reporting interval. The state is the product set of R_I and R_L, which can be expressed as $\mathcal{S} = \{s_1, s_2, ..., s_t, .., s_N\}$ where s_t is $\{s_t^I, s_t^L\}$, and t means serial number of the state; s_j^I is interference ratio, s_j^L represents loss ratio. Each action and state is quantized as some steps because Q-learning implementation requires the input and environment to be modeled as a finite-state system.

Table 2 Parameters of PU activity and spectrum sensing

E[T_{busy}] of PU [ms]	E[T_{idle}] of PU [ms]	Sensing bandwidth [MHz]	Measurement window [ms]
2.5	2.5	1.5	500

Table 3 Target value and experimental parameter

Target interference ratio (R_I^{th})	Target loss ratio (R_L^{th})	Parameters for reward
0.1	0.2	ρ_1=3.15, ρ_2=1.35, ω_1=3, ω_2=2, φ=25, K=40, L=30

The Q-value is updated according to (42):

$$\mathcal{Q}(s_t, a_t) \leftarrow (1-\alpha)\mathcal{Q}(s_t, a_t) + \alpha\left\{ r_t + \gamma \max_{a_{t+1}} \mathcal{Q}(s_{t+1}, a_{t+1}) \right\} \tag{42}$$

where $\alpha \in [0, 1]$ is the learning rate. If α has high value, the system consider more for present and future experience. If α has low value, it takes longer time to learn the environment as the stored Q-values have more weight. On the right side of (42), $r_t + \gamma \max_{a_{t+1}} \mathcal{Q}(s_{t+1}, a_{t+1})$ is the actual value for the action and future value. Q-learning finds the Q-value by iteratively approximating the Q-function using the difference between the predicted value and the actual value as the estimation error [38]. $\gamma \in [0, 1]$ is the discount factor and if γ is high, the system gives a higher weight to the Q-value of the new state by the action than the reward of the past action. On the other hand, if γ is low, the immediate reward is weighted and is more influenced by the current action.

If an agent chooses an action by only the maximum value of the Q-value, a local optimization problem occurs. Therefore, we used the action in consideration of the ε-greedy policy, as follows:

$$a = \begin{cases} \underset{\tilde{a} \in \mathcal{A}}{\text{argmax}}\, \mathcal{Q}(s, \tilde{a}), & \text{with probability } 1-\varepsilon \\ \text{random } a \in \mathcal{A}, & \text{with probability } \varepsilon \end{cases} \tag{43}$$

where $\varepsilon \in [0, 1]$ is the probability of choosing a random action. When the random number is less than ε, the action is selected randomly, and in the opposite case, the highest Q-value is chosen. Table 1 shows the sequence of steps for Q-learning in the proposed algorithm.

3.3 Reward function design

In this section, we describe the proposed reward function of the Q-learning-based dynamic sensing time and reporting interval selection algorithm for the sensor. A false alarm depends on the sensing time from fixing the required detection probability to protect the PU preferentially. The false alarm is a parameter that seriously affects the capture of the transmission opportunity for the SU. A long reporting interval can increase interference with the PU due to the sudden appearance of the PU, while ensuring the continuity of the transmission of the SU and saving sensor power by not sending the sensing results frequently. Therefore, in Q-learning, action is defined as a combination of sensing time and reporting interval, and state is defined as a combination of interference ratio and loss ratio that change through action. In addition, we designed the reward as expressed in the flow chart of Fig. 9. It should be noted that because the wireless environment conditions can change dynamically and we do not assume any prior environmental statistics, the proposed learning-based mechanism may have a difficulty to meet the required constraints in terms of the interference ratio threshold and secondary transmission opportunity loss ratio at every time instance. Therefore, we have relaxed the constraints in (6) not for every time instance but long-term average. From this point of view, the proposed reward function composed of the multi-objective in (6) dynamically controls the sensing and reporting parameters and satisfies the constraint condition.

Table 4 Parameters for Q-learning

Learning rate (α)	Discount factor (γ)	Random selection probability(ε)
0.5	0.5	0.3→0.1

Reward is divided into two types. Considering the interference ratio R_I as a priority, (44) is used when R_I is smaller than the threshold of R_I ($R_I < R_I^{\text{th}}$), and (45, 46) is used in the opposite case.

$$r = \delta_1 \rho_1 \exp\left(\varphi \left| R_I - R_I^{\text{th}} \right|\right) + \delta_2 \rho_2 \exp\left(\varphi \left| R_L - R_L^{\text{th}} \right|\right) - K \frac{E[N_s N_r]}{N_s N_r} + L \tag{44}$$

where φ, ρ_1, ρ_2, and L are the constant value. ρ_1, and ρ_2 should be carefully selected to reduce or increase the reward appropriately if R_1 and R_L exceed the threshold or not; δ_1 and δ_2 are the signs for the first and the second term; R_I^{th} and R_L^{th} are the threshold of interference ratio and loss ratio; N_s and N_r are the sample numbers of sensing and a multiple number of the sensing time to represent the reporting interval. These are obtained by sampling the sensing and reporting intervals at twice the sensing frequency. The first, second, and third term of (44) is related to $f_{\text{intf}}(R_I)$, $f_{\text{loss}}(R_L)$, and $f_{\text{overhead}}(t_s, t_r)$ of (6), respectively.

The first term gives a positive value ($\delta_1 = +1$) for R_I satisfying the condition ($R_I < R_I^{\text{th}}$), and the second term gives a positive ($\delta_2 = +1$) or negative ($\delta_2 = -1$) value according to whether R_L is satisfied ($R_L < R_L^{\text{th}}$). The value of ρ_1 is greater than ρ_2 in order to consider R_I prior to R_L. The exponential function is used for more dramatic Q-value changes in the reward terms for R_I and R_L when the interference ratio and loss ratio become significantly worse. The third term is designed to indicate that the overhead on the system increases, as the response time (the product of the sensing length and the multiple for the response length) is shorter than the average response time.

If R_I does not satisfy the condition ($R_I < R_I^{\text{th}}$), we do not consider the system overhead in order to focus on the interference ratio satisfaction as follows:

$$r = \delta_1 \omega_1 \exp\left(\phi \left| R_I^{\text{new}} - R_I^{\text{old}} \right|\right) + \delta_2 \omega_2 \exp\left(\phi \left| R_L^{\text{new}} - R_L^{\text{old}} \right|\right) \tag{45}$$

where ϕ, ω_1, and ω_2 are the constant values. R_I^{new} and

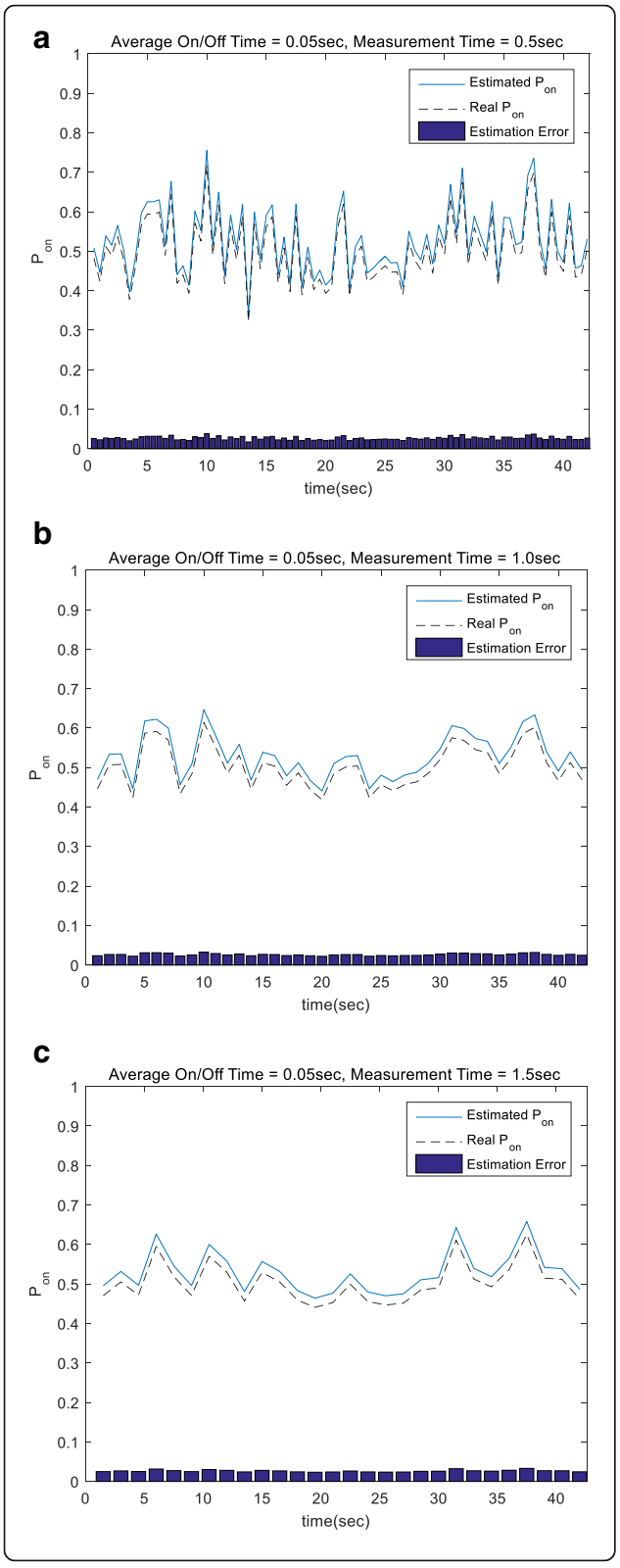

Fig. 10 Comparison between estimated and real P_{on}. **a** Measurement time = 0.5 s. **b** Measurement time = 1 s. **c** Measurement time = 1.5 s. We must calculate the P_{on} to accurately calculate the interference ratio and the transmission opportunity loss ratio. Therefore, P_{on}'s estimation must also be accurate. The average on/off cycle of the PU is 0.05 s. The measurement time is 0.5 s in **a**, 1 s in **b**, and 1.5 s in **c**. For each case, the solid line represents the estimated P_{on}, the dashed line represents the actual P_{on}, and bar represents the estimation error. **a**, **b**, and **c** indicate that the proposed method accurately estimate the real P_{on}. As the measurement time becomes longer, the change of the value decreases by the nomalization

R_L^{new} are the present values of interference ratio and loss ratio. R_I^{old} and R_L^{old} are the previous value of interference ratio and loss ratio. Since the interference ratio and loss ratio could not be changed by the action, we try to choose the action which shows a good tendency when the interference ratio does not satisfy the constraint. The first and second term of (44) is related to $f_{intf}(R_I)$ and $f_{loss}(R_L)$, respectively. The constant values should be carefully selected to increase or reduce the reward appropriately if tendency of interference ratio and loss ratio is good or not.

The expression in (45, 46) is again divided based on whether the previous value (R_I^{old}) of R_I of measurement window is larger than the current value (R_I^{new}) of the measurement window or not. The first term of (45, 46) takes a positive value ($\delta_1 = +1$) when R_I is improved for $R_I^{old} > R_I^{new}$ and has a negative value ($\delta_1 = -1$) when R_I deteriorates for $R_I^{old} \le R_I^{new}$. Similarly with respect to R_L, ω_2 has a positive value ($\delta_2 = +1$) when R_L is lower than the threshold or is improved ($R_L \ge R_L^{th}$, $R_L^{old} > R_L^{new}$), and has a negative value ($\delta_2 = -1$) when R_L is exacerbated $R_L \ge R_L^{th}$, $R_L^{old} \le R_L^{new}$). Using the algorithm designed in this way, we will see that we select the sensing time and reporting interval dynamically according to the surrounding environment through the results of Section 4.

4 Simulation results

In this section, we first show an accurate estimation of P_{on}, P_{off}, interference ratio, and transmission opportunity loss ratio. Second, we compare the simulation results of interference ratio and transmission opportunity loss ratio at each SNR about Q-learning and the case in which the sensing and reporting interval are fixed. We also show the simulation results of interference ratio and transmission opportunity loss ratio of Q-learning according to SNR change with time. Finally, we represent the advantage from the continuous sensing.

The simulation was performed using MATLAB. The alternate sequence about busy and idle states of PU follows exponential distribution. More details could be found in [39] and for the MIMO-based sensing, antenna correlation

is 0.5 which is referred to [40]. The parameters of PU activity and spectrum sensing are shown in Table 2. Table 3 represents the experimental parameter for (44), (45) and (46) in Section 3.3 and the target value for interference ratio and loss ratio. The parameters used in Q-learning are shown in Table 4. The parameters for reward are selected experimentally. And we assume the compound channel gain is 0.9.

As shown in Section 3.1, we can estimate P_{on} and P_{off} without any assumption about the statistics of the PU as [10, 39]. We can successfully estimate R_I and R_L as (27) and (30). Figure 10 shows the results of P_{on} estimation for each measurement time. We set SNR = − 5 dB, sensing bandwidth as 1.5 MHz, the sum of the average busy/idle time of the PU 5 ms, and the ratio of ON time is 0.5. The average estimation error is only 0.0264, 0.0265, and 0.0264 for each of (a), (b), and (c) in Fig. 10 according to measurement time. They are similar to each other. If the measurement interval is short, it is possible to estimate the ON state of the PU that is dynamically fluctuating, and it can be confirmed that the variation of the value decreases as the measurement interval becomes longer.

Figure 11 shows the estimated R_I and R_L using the P_{on} estimation of about 0.5 s of the measurement window. Similar to Fig. 10, we can see that the estimated R_I and R_L are close to the actual values, and the average estimation errors are 0.0395 and 0.0427, respectively. From this result, it is possible to guarantee the reliability of the effect estimation from the selection of each action because the interference ratio and transmission opportunity loss ratio can be estimated approximately which are the response of the actions selected by the Q-learning.

In the simulation, we evaluated the performance of the proposal compared with fixed action (i.e., fixed sensing time and reporting interval). The actions of Q-learning are combinations of sensing time in the form of number of sensing samples (ss = 200, 600, 1000, 1400) and the reporting interval as multiples of sensing time (sr = 3, 6, 9). For the state, we divide R_I into four states (0~ 0.07, 0.07~ 0.1, 0.1~ 0.2, and 0.2~), and we split R_L into three states (0~ 0.1, 0.1~ 0.2, 0.2~). The states are the combinations of R_I and R_L. We set the parameters for Q-learning at $\alpha = 0.5$, $\gamma = 0.5$, and the random action choice parameter is $0.1 \le \varepsilon \le 0.3$ (starts from 0.3, and the lower limit is 0.1) using ε-greedy exploration. The value of the low limit of ε is necessary to allow for a flexible adaptation of the Q-table when the environment changes. The fixed case is the combination of the sensing sample (ss = 200, 600, 1000, and 1400) and the reporting interval as multiples of sensing time (sr = 3 and 9).

Figures 12, 13, and 14 show boxplots of R_I and R_L according to fixed cases and Q-learning at a SNR of − 6, −9, and −12 dB, respectively. If the response time is short, like ss200sr3 (ss = 200, sr = 3, reply length = 200×3) and ss200sr9 (ss = 200, sr = 9), a high false alarm probability arises because there is not enough sensing time. Although the SU system will abandon the transmission for this reason and has very low interference, the loss of transmission opportunity increases at a low SNR (−9 and −12 dB). For ss600sr3 and ss600sr9, the performance is good in terms of interference ratio and loss ratio up to −9 dB, but transmission loss increases at −12 dB for the same reason. Both ss1000sr9 and ss1400sr9 have high values in interference ratio and loss ratio because the reporting interval itself is long. For ss1400sr3, the interference ratio and loss ratio are stable at all SNRs. However, it has low performance compared to Q-learning on the reward side, since Q-learning that dynamically selects actions in all environments has better performance in terms of system load

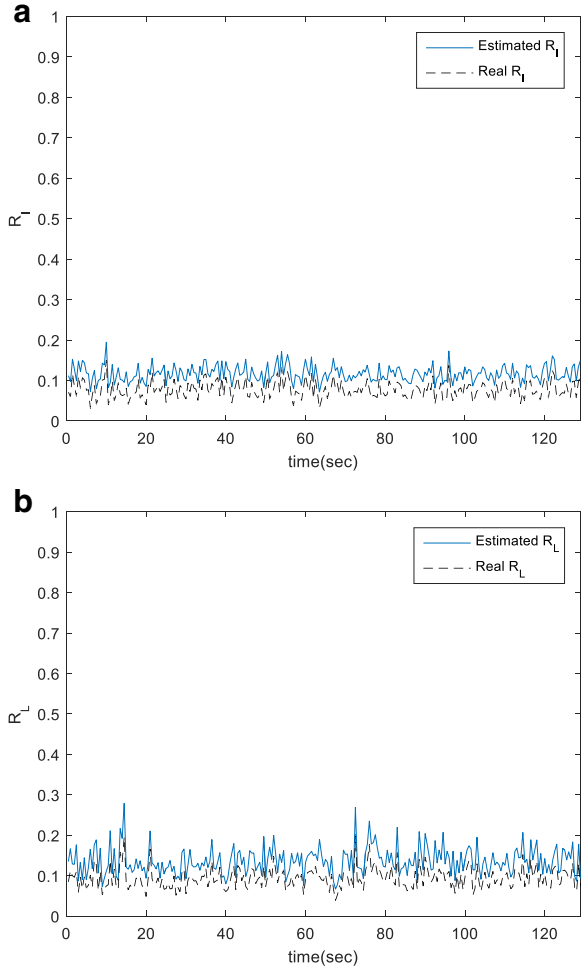

Fig. 11 Comparison between estimated and real R_I, R_L **a** interference ratio and **b** transmission opportunity loss ratio, **a** shows the interference ratio and **b** shows the transmission opportunity loss ratio. For each case, the solid line represents the estimated value and the broken line represents the real value. **a** and **b** are approximate to the real values

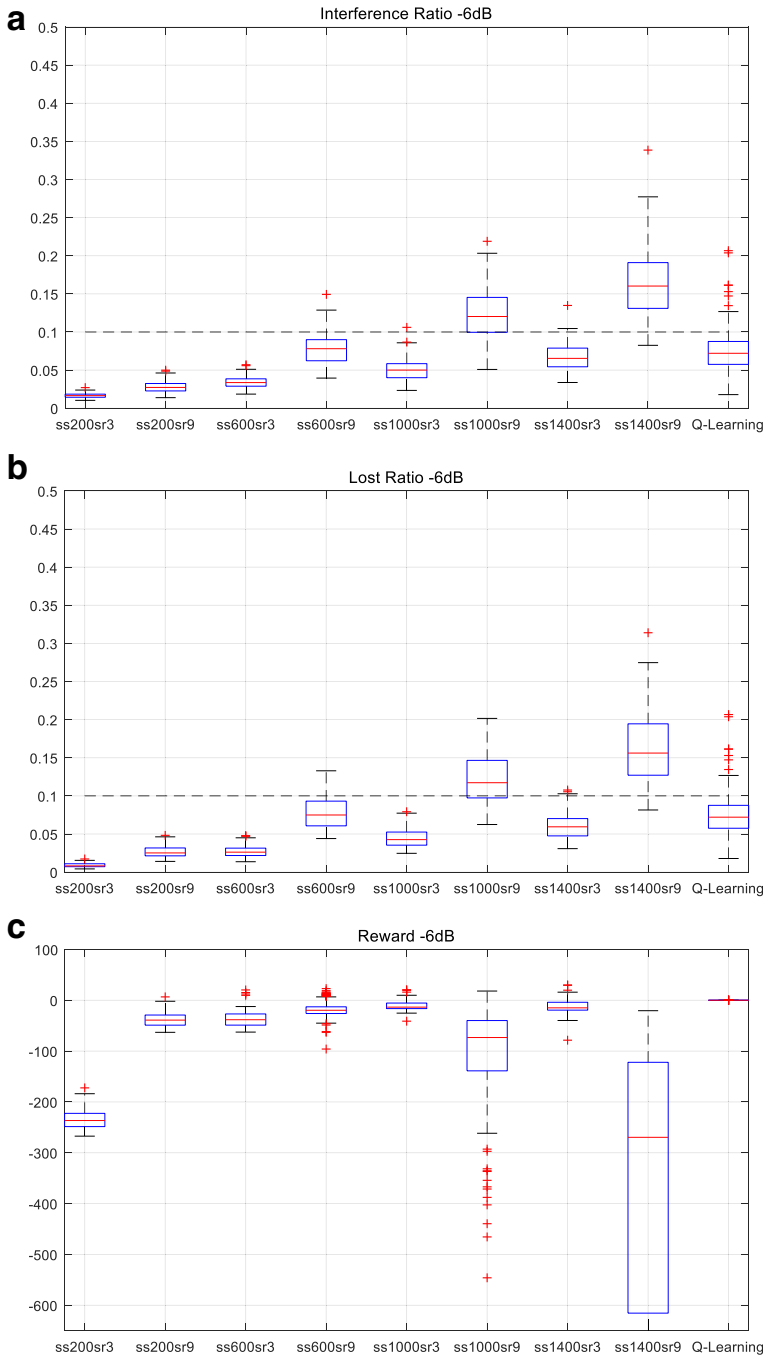

Fig. 12 Interference ratio, transmission opportunity loss ratio, and reward of the fixed case and Q-learning @ − 6 dB. **a** Interference radio. **b** Transmission opportunity loss ratio. **c** Reward. The performance of the interference ratio, transmission opportunity loss ratio, and reward is compared for cases where the sensing time, and the reply time are fixed and the proposed Q-learning is used for SNR of − 6 dB. In **a** representing the interference ratio and **b** representing the transmission loss ratio, the Q-learning operates within a stable range, but other cases for fixed sensing time and reply time also operate within a stable range. However, in **c**, which indicates reward, Q-learning shows overwhelming performance difference compared to other cases. Thus, it can be shown that the system load can be reduced by using Q-learning

(transmission power loss, continuous transmission possibility). Q-learning has superior mean and a low variance over the fixed case for all SNRs on the reward side. From these results, it can be seen that the loss of transmission opportunity increases in order to minimize the interference to the PU in almost fixed cases, and both

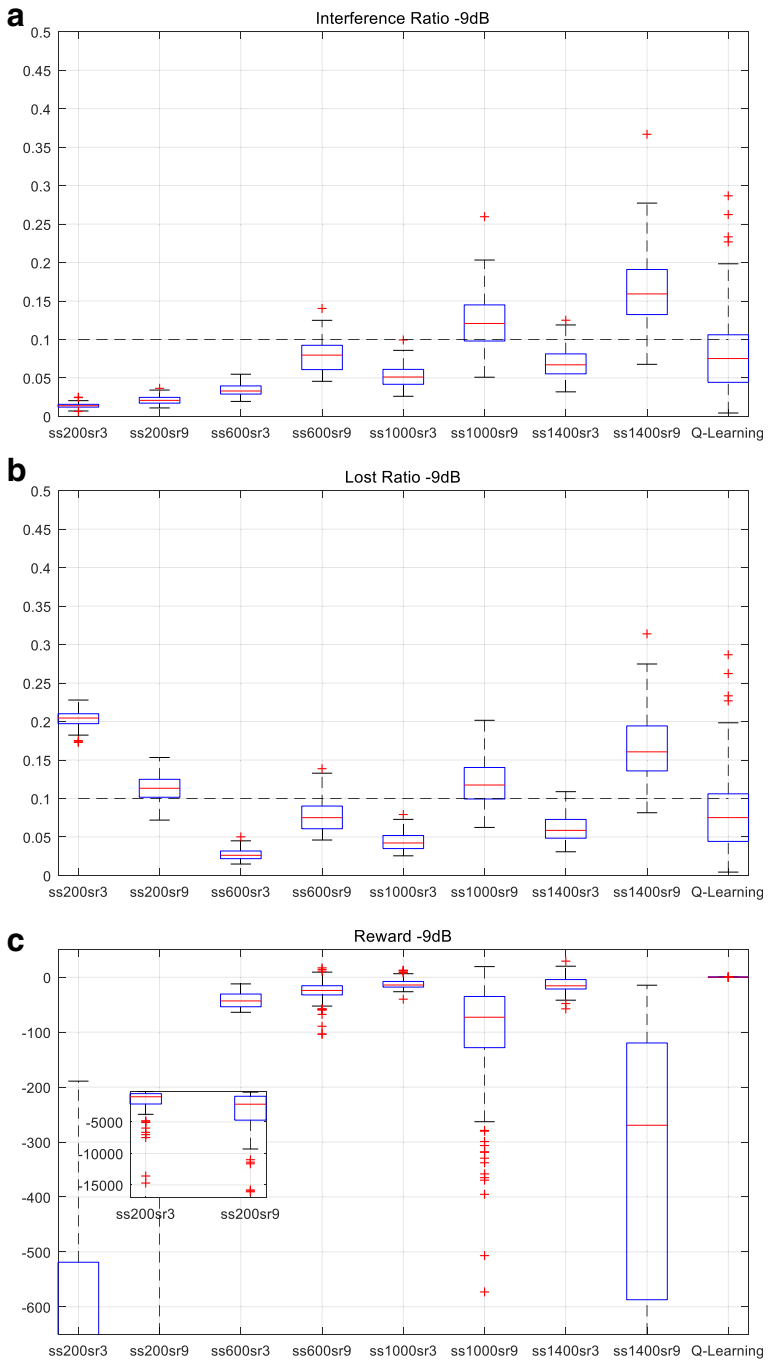

Fig. 13 Interference ratio, transmission opportunity loss ratio, and reward of the fixed case and Q-learning @ − 9 dB. **a** Interference radio. **b** Transmission opportunity loss ratio. **c** Reward. The performance of the interference ratio, transmission opportunity loss ratio, and reward is compared for cases where the sensing time and the reply time are fixed and the proposed Q-learning is used for SNR of − 9 dB. In **a** representing the interference ratio and **b** representing the transmission loss ratio, the Q-learning operates within a stable range. For the case where the sensing time and the reapply time are fixed, the transmission opportunity loss ratio increases for ss200sr3 and ss200sr9, and the intereference ratio and transmission opportunity loss ratio for ss100sr9 and ss1400sr9 increase. In **c** indicating a reward, the Q-learning shows overwhelming performance difference compared to other cases. Thus, it can be shown that the system load can be reduced by using Q-learning

the interference to the PU and the loss of the CR system are unsatisfactory in certain cases. On the other hand, the Q-learning dynamically tracks the busy/idle of the PU which are frequently change when the SNR is fixed, satisfying the interference ratio and the transmission opportunity ratio within the selected range. In the reward

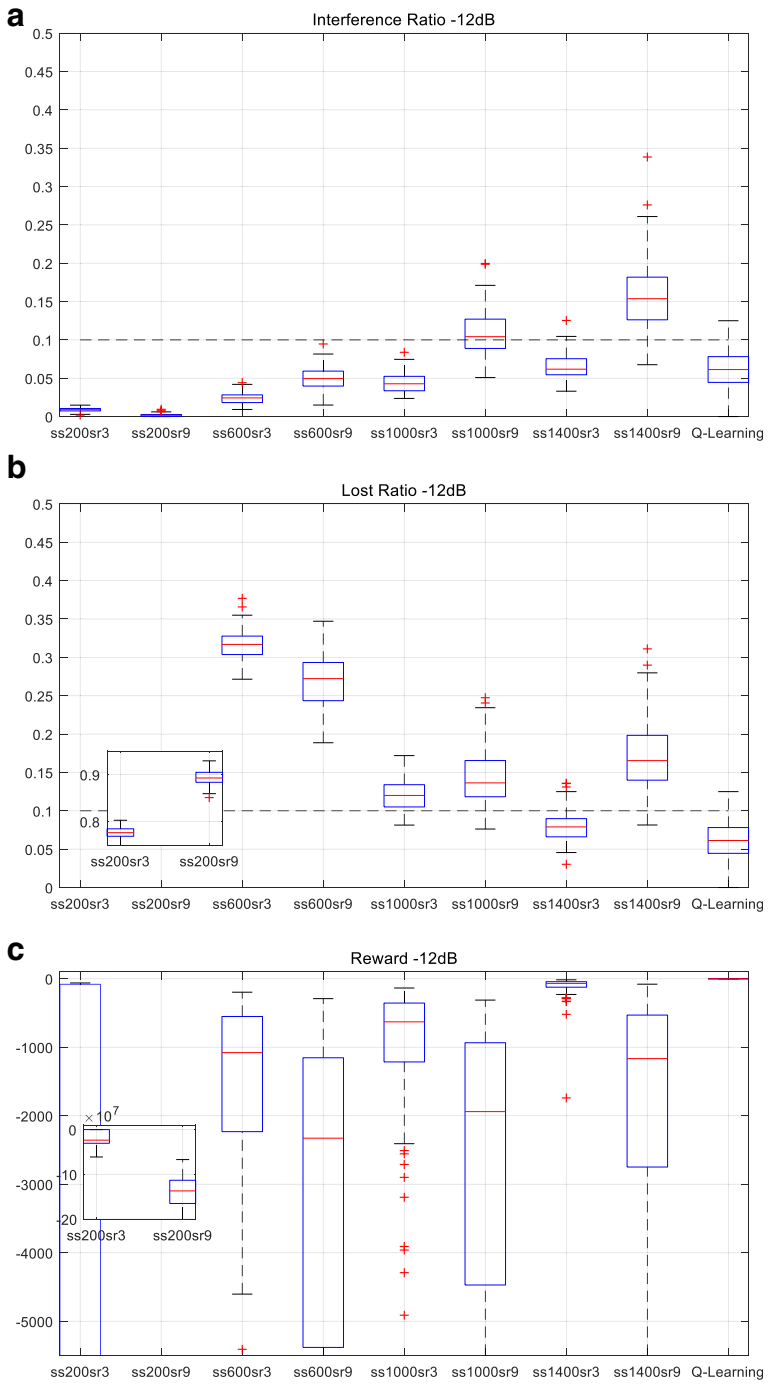

Fig. 14 Interference ratio, transmission opportunity loss ratio, and reward of the fixed case and Q-learning @ − 12 dB. **a** Interference radio. **b** Transmission opportunity loss ratio. **c** Reward. The performance of the interference ratio, transmission opportunity loss ratio, and reward is compared for cases where the sensing time and the reply time are fixed and the proposed Q-learning is used for SNR of − 12 dB. In **a** representing the interference ratio and **b** representing the transmission loss ratio, the Q-learning operates within a stable range. For the case where the sensing time and the reapply time are fixed, the interference ratio is low, but the transmission opportunity loss ratio increases significantly in most cases. In **c** indicating a reward, the Q-learning shows overwhelming performance difference compared to other cases. Thus, it can be shown that the system load can be reduced by using Q-learning

aspect, it can be seen that the overhead of sensor is considerably reduced because the reward is higher than the fixed case and low variance.

Figures 15 and 16 show that Q-learning operates adaptively according to SNR changes. In the overall region, we can see that the interference ratio stays around the threshold, and for the loss ratio, we can see the phenomenon of returning to around the threshold although it sometimes has a value of bouncing at −12 dB. Therefore, Q-learning can select the sensing time and the reporting interval dynamically even in the environment where the SNR varies, so that the system can operate within the desired range of interference ratio and transmission opportunity loss ratio.

Figure 17 shows the actual interference ratio according to the average primary ON time E[on] changes, and the proposed method is compared with the general sensing time optimization method for throughput maximization [9] for which the data frame length is 17.5 ms (52,500 samples). For the proposed method, we implemented two primary detection notification models (default periodic reporting and dual transceiver). Since the conventional method cannot detect the PU in the data transmission period, the interference ratio is always larger than that of the proposed method. For the smaller average primary ON time (i.e., the shorter primary system activation time), the more conventional method cannot detect primary signal and gives the more harmful interference because the primary may appear only between consecutive sensing times. In the case of two proposed notification models, the dual transceiver method shows little better performance than that of periodic reporting because it can make immediate stop of secondary data transmission using the dedicate narrow band control channel.

Figure 18 shows the primary appearance detection ratio for different average primary ON time E[on]. The primary appearance detection ratio indicates whenever primary turns on how accurately the secondary system detects the primary appearance. In the conventional method, for shorter E[on] case, the primary activation time can be smaller than the sensing interval so that secondary system cannot sense the primary appearance. The primary only can be detected when primary activation time is overlapped with the secondary sensing time. The proposed methods always show very high (> 0.98) primary appearance detection ratio.

5 Conclusions

In this paper, we proposed an algorithm to dynamically select the sensing time and reporting interval in order to adapt to the surrounding environment using Q-learning in an IA-based CR network. We change the system to eliminate the dependence on the PU information unlike the conventional IA-based CR system. Therefore, we can continuously monitor the PU by designating a sensor dedicated to sensing. However, the remaining SUs need to periodically receive the sensing results from the sensor. In this mechanism, the optimization issue turns into how often to receive the sensing results. We define this as the sensing time and its multiple for reporting interval, and solve this problem using Q-learning, a typical

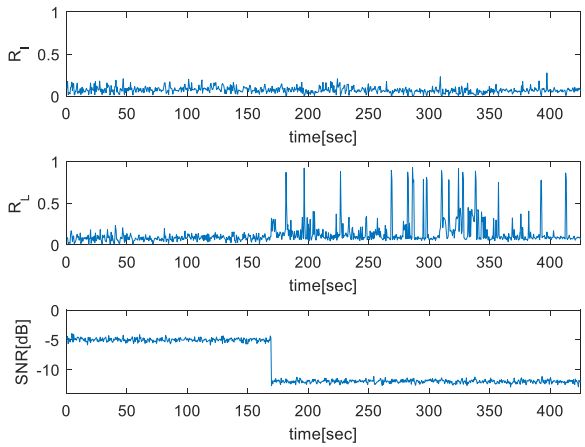

Fig. 15 R_I, R_L according to SNR change. Figure 15 shows the change in interference ratio and transmission opportunity loss ratio when SNR changes over time. Even when the environment changes, Q-learning uses a Q-table to find the state of the interference ratio and the transmission opportunity loss ratio and uses the appropriate action to maintain a stable system. At low SNRs, the transmission opportunity loss ratio increases, but it can be confirmed that it returns to a stable range

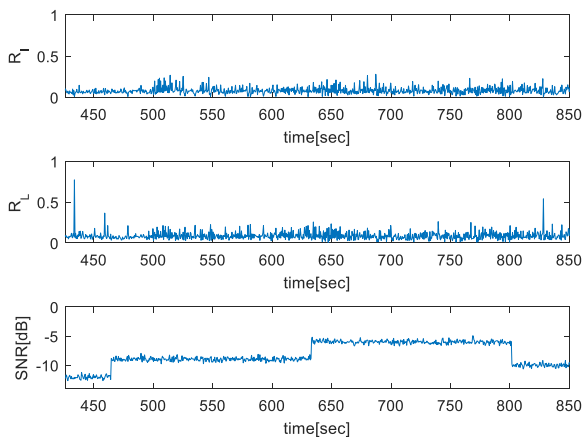

Fig. 16 R_I, R_L according to SNR change. Figure 16 shows the change in interference ratio and transmission opportunity loss ratio when SNR changes over time. Even when the environment changes, Q-learning uses a Q-table to find the state of the interference ratio and the transmission opportunity loss ratio and uses the appropriate action to maintain a stable system. At low SNRs, the transmission opportunity loss ratio increases, but it can be confirmed that it returns to a stable range

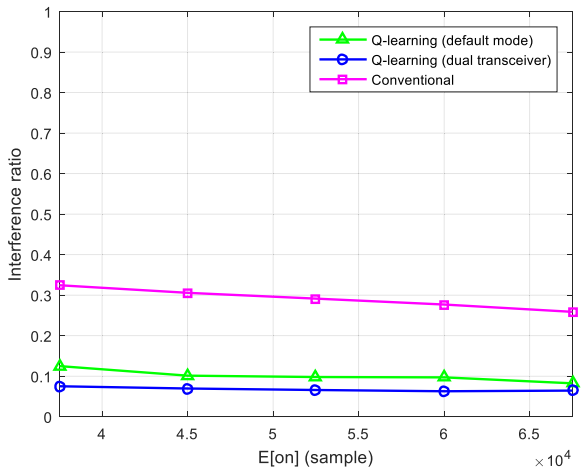

Fig. 17 Interference ratio according to *E*[on] ratio. Compares the actual interference ratio according to the ratio of *E*[on] with the proposed method and the general sensing optimization method. If the sensing and transmission intervals are alternating, there is no way to recognize the PU during data transmission. Therefore, even if the idle is detected in the sensing interval, the interference ratio is high because the PU occurs in the data transmission interval. As the *E*[on] increases, the probability of the PU of the sensing interval continuing to the data transmission interval is high and the interference ratio decreases. Both proposed methods considerably lower the interference ratio than the conventional method, and the second method shows a lower interference ratio because the data transmission is immediately stopped using the narrow band as soon as the PU is detected

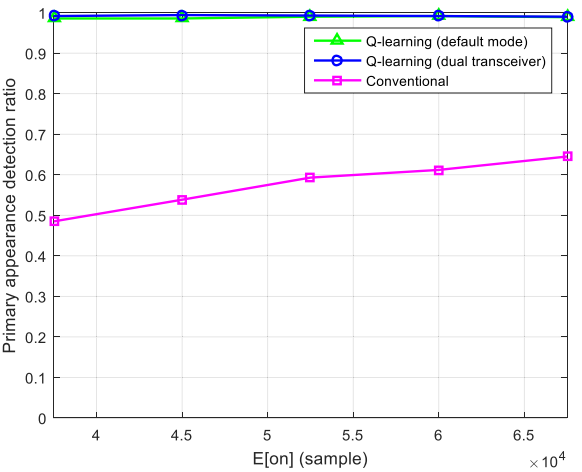

Fig. 18 PU existence count ratio according to *E*[on]. Shows the detection ratio of the number of PU appearances according to the ratio of *E*[on]. The proposed method has a high value in all regions because it continuously sense the channel. However, the conventional method cannot detect if the PU appears only in the data transmission interval because the sensing and transmission intervals are alternately performed. The conventional method increases the PU existence count ratio because *E*[on] increases the probability that the sensing interval overlaps with the PU busy interval

reinforcement learning algorithm. We assigned the action of Q-learning as the set of products from sensing time and multiple of that. We designate state as the set of products of the interference ratio and transmission opportunity loss ratio. We propose a method to predict the interference ratio and loss ratio without any assumptions about the operation of the PU and confirm that it is close to the actual interference ratio and loss ratio. We designed the reward considering the interference ratio, the loss ratio, and the load on the system and compared it with the fixed case for each SNR through simulation. In addition, as the SNR changes, we can confirm that the system operates dynamically and operates stably. Furthermore, we also assure that benefit of the proposal since this system can sense the channel continuously.

6 Methods/experimental

The purpose of this study is to minimize the interference to the primary user and to keep the monitoring of primary users by continuously sensing without alternating between sensing and data transmission. For this purpose, the IA-based cognitive radio system performs the spectrum sensing by assigning a secondary user as a sensor. In this paper, we propose a precoding and decoding method for this system, and propose a sensing method for each case when the secondary user transmit the data and does not transmit according to the existence of primary user. Since the role of the spectrum sensing is limited to the sensor node, it is necessary to determine the period for other secondary user to receive the sensing result from the sensor and the sensing time. Those are selected by the Q-learning. Q-learning is a representative learning algorithm that allows the agent to identify the state of the agent and to take appropriate action by identifying the surrounding information. In the proposed system, the state of the agent (CH) is defined as the combination of the interference ratio of the PU and transmission opportunity loss ratio of the SU. In this paper, we propose a method to calculate the interference ratio and the transmission opportunity loss ratio using the statistical characteristics of the PU obtained by continuous sensing. The action of the agent is the combination of a sensing time and period for reporting the sensing result. The time-dependent mechanical relationship is stored in the Q-table when the interference ratio and the transmission opportunity loss ratio are determined according to the selected action (sensing time and sensing result). The Q-learning uses the information stored in the Q-table to select the most appropriate action for a given state at each time.

Experimental results in this paper had performed using MATLAB R2015b on Intel® Core i7 3.4 GHz system. The exponential random function to generate the

PU over time and Q-table matrix for Q-learning can be made by constructing appropriate MATLAB code.

7 Appendix

According to Bezout's theorem, $N_e \le N_v$ must be satisfied, where N_e is the total number of equations, and N_v is the total number of variables. Considering the conditions in (11), (12), (13), and (14), N_e and N_v can be obtained as follows:

$$\begin{aligned} N_e &= \sum_{j\neq 0, j\in K} d^{[0]}d^{[j]} + \sum_{i,j\neq 0, i,j\in K} d^{[i]}d^{[j]} \\ &= [K + K(K-1)]d^2 \end{aligned} \tag{46}$$

$$\begin{aligned} N_v &= d^{[0]}\left(N-d^{[0]}\right) + \sum_{k=1}^{K}\left[d^{[k]}\left(M-d^{[k]}\right) + d^{[k]}\left(N-d^{[k]}\right)\right] \\ &= d(N-d) + Kd(M+N-2d) \end{aligned} \tag{47}$$

where the number of desired streams is assumed to be the same for simple representation.

Then, the DoF condition is expressed by (48):

$$d \le \frac{N}{K+1} + \frac{KM}{(K+1)^2} \tag{48}$$

If we generalize this to P sensors, we can obtain (49) for N_e and N_v.

$$\begin{aligned} N_e &= [PK + K(K-1)]d^2, N_v \\ &= Pd(N-d) + Kd(M+N-2d) \end{aligned} \tag{49}$$

The DoF condition is expressed by (50):

$$d \le \frac{N}{K+1} + \frac{KM}{(P+K)(K+1)} \tag{50}$$

Therefore, in the network environment in which (48) or (50) are guaranteed, the sensor node can remove the interference to the signals of other SU and sense the PU signal.

Abbreviations

AWGN: Additive white Gaussian noise; CR: Cognitive radio; D2D: Device-to-device; DoF: Degree of freedom; IA: Interference alignment; ISM: Industrial, Scientific and Medical; LTE-U: Long Term Evolution in unlicensed spectrum; MIMO: Multi-input multi-output; PDF: Probability density function; PSK: Phase shift keying; PU: Primary user; SDP: Semi-definite programming; SU: Secondary user; TVWS: TV white spaces; UPT: Unprotected primary user transmission; ZF: Zero forcing

Dataset of simulations

The simulation was performed using MATLAB in Intel Core i7 (32 bit).
The alternate sequence about busy and idle states of PU follows exponential distribution. More details could be found in [39].
For the MIMO-based sensing, antenna correlation is 0.5, which is referred to [40].
The Q-table is made up of tables as defined in the paper and it works according to the Q-table update equation in conjunction with sensing.

Funding

This work was supported by the Inha University research grant.

Authors' contributions

Both authors contribute to the concept, the design and developments of the theory analysis and algorithm, and the simulation results in this manuscript. Both authors read and approved the final manuscript.

Authors' information

- Prof. Sang-Jo Yoo, PhD (corresponding author)
Sang-Jo Yoo received the B.S. degree in electronic communication engineering from Hanyang University, Seoul, South Korea, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, in 1990 and 2000, respectively. From 1990 to 2001, he was a Member of Technical Staff with the Korea Telecom Research and Development Group, where he was involved in communication protocol conformance testing and network design fields. From 1994 to 1995 and from 2007 to 2008, he was a Guest Researcher with the National Institute Standards and Technology, USA. Since 2001, he has been with Inha University, where he is currently a Professor with the Information and Communication Engineering Department. His current research interests include cognitive radio network protocols, ad hoc wireless network, MAC and routing protocol design, wireless network QoS, and wireless sensor networks.
- Mr. Sung-Jeen Jang
Sung-Jeen Jang received a B.S degree in electrical engineering from Inha University Incheon, Korea, 2007. He received his M.S. degree in Graduate School of Information Technology and Telecommunication, Inha University, Incheon Korea, 2009. Since March 2009, he has been pursuing a Ph.D degree at the Graduate School of Information Technology and Telecommunication, Inha University, Incheon Korea. His current research interests include cognitive radio network protocols and machine learning applied wireless communications.

Competing interests

I confirm that I have read Springer Open's guidance on competing interests and none of the authors have any competing interests in the manuscript.

Publisher's Note

Received: 25 October 2017 Accepted: 30 May 2018
Published online: 20 June 2018

References

1. MR Kelley, The spectrum auction: big money and lots of unanswered questions. IEEE Internet Comput. **12**(1), 66–70 (2008)
2. J. Mitola, III, Cognitive radio, licentiate thesis, KTH, Royal Inst. of Technol., Stockholm, Sweden, (1999).
3. S Haykin, Cognitive radio: Brain-empowered wireless communications. IEEE Journal on Selected Areas in Communications **23**(2), 201–220 (2005)
4. YY Liu, SJ Yoo, Dynamic resource allocation using reinforcement learning for LTE-U and WiFi in the unlicensed spectrum. IEEE ICUFN, 471–475 (2017)
5. E Almeida, AM Cavalcante, RCD Paiva, et al., Enabling LTE/WiFi coexistence by LTE blank subframe allocation. IEEE ICC, 5083–5088 (2013)
6. D Cabric, SM Mishra, RW Brodersen, Implementation issues in spectrum sensing for cognitive radios. IEEE, Pacific Grove, 772–776 (2004)
7. N Sai Shankar, C Cordeiro, K Challapali, Spectrum agile radios: utilization and sensing architectures. IEEE DySPAN, 160–169 (2005)
8. Y Hur et al., A cognitive radio (CR) system employing a dual-stage Spectrum sensing technique: a multi-resolution Spectrum sensing (MRSS) and a temporal signature detection (TSD) technique. IEEE Globecom, 1–5 (2006)
9. YC Liang, Y Zeng, ECY Peh, AT Hoang, Sensing-throughput tradeoff for cognitive radio networks. IEEE Trans. Commun. **7**(4), 1326–1337 (2008)
10. W y Lee, IF Akyildiz, Optimal spectrum sensing framework for cognitive radio networks. IEEE Trans. on Wireless Commun. **7**(10), 3845–3857 (2008)
11. JK Choi, SJ Yoo, Undetectable primary user transmissions in cognitive radio networks. IEEE Commun. Letters **17**(2), 277–280 (2013)
12. JK Choi, SJ Yoo, Optimal sensing interval considering per-primary transmission protection in cognitive radio networks. Wireless Personal Commun. **78**, 1891~1903 (2014)

13. W Lee, DH Cho, Enhanced spectrum sensing scheme in cognitive radio systems with MIMO antennae. IEEE Trans. on Veh. Tech. **60**(3), 1072–1085 (2011)
14. F Moghimi, RK Mallik, R Schober, Sensing time and power optimization in MIMO cognitive radio networks. IEEE Trans. on Wireless Commun. **11**(9), 3398–3408 (2012)
15. X Li, N Zhao, Y Sun, FR Yu, Interference alignment based on antenna selection with Imperfect channel state information in cognitive radio networks. IEEE Trans. on Veh. Tech. **65**(7), 5497–5511 (2016)
16. M Amir, A El-Keyi, M Nafie, Constrained interference alignment and the spatial degrees of freedom of MIMO cognitive networks. IEEE Trans. on Infor. Theory **57**(5), 2994–3004 (2011)
17. H Zhou, T Ratnarajah, YC Liang, On secondary network interference alignment in cognitive radio. IEEE DySPAN, 637–641 (2011)
18. H Men, N Zhao, M Jin, JM Kim, Optimal transceiver design for interference alignment based cognitive radio networks. IEEE Commun. Letters **19**(8), 1442–1445 (2015)
19. S Chatzinotas, B Ottersten, Cognitive interference alignment between small cells and a macrocell. IEEE ICT, 1–6 (2012)
20. L Huang, G Zhu, X Du, Cognitive femtocell networks: an opportunistic spectrum access for future indoor wireless coverage. IEEE Wirel. Commun. **20**(2), 44–51 (2013)
21. SK Sharma, S Chatzinotas, B Ottersten, Interference alignment for spectral coexistence of heterogeneous networks. EURASIP Journal on Wireless Commun. and Networking, 1–14 (2013)
22. G Chen, Z Xiang, C Xu, M Tao, On degrees of freedom of cognitive networks with user cooperation. IEEE Wireless Commun. Letters **1**(6), 617–620 (2012)
23. B Guler, A Yener, Interference alignment for cooperative MIMO femtocell networks. IEEE, GLOBECOM, 1–5 (2011)
24. SM Perlaza, N Fawaz, S Lasaulce, M Debbah, From spectrum pooling to space pooling: opportunistic interference alignment in MIMO cognitive networks. IEEE Trans. on Signal Processing **58**(7), 3728–3741 (2010)
25. M Hasani-Baferani, J Abouei, Z Zeinalpour-Yazdi, Interference alignment in overlay cognitive radio femtocell networks. IET Commun. **10**(11), 1401–1410 (2016)
26. X Li, N Zhao, Y Sun, FR Yu, Interference alignment based on antenna selection with imperfect channel state information in cognitive radio networks. IEEE Trans. on Veh. Tech. **65**(7), 5497–5511 (2016)
27. L Li, T Li, J Ge, L Kong, J Liu, Channel sensing order for distributed cognitive networks with multi-user and multi-channel. IEEE ICCSN, 44–50 (2017)
28. J Oksanen, J Lunden, V Koivunen, Reinforcement learning based sensing policy optimization for energy efficient cognitive radio networks. Neurocomputing **80**, 102–110 (2012)
29. A Das, SC Ghosh, N Das, AD Barman, Q-learning based co-operative spectrum mobility in cognitive radio networks. IEEE LCN **2017**, 502–505 (2017)
30. Y Li, SK Jayaweera, M Bkassiny, C Ghosh, Learning-aided sub-band selection algorithms for spectrum sensing in wide-band cognitive radios. IEEE Trans. on Wireless Commun. **13**(4), 2012–2024 (2014)
31. O. van den Biggelaar, J.M. Dricot, P.D. Doncker, F. Horlin, Sensing time and power allocation for cognitive radios using distributed q-learning, EURASIP J. Wirel. Commun. Netw., 2012, 1–40, (2012)
32. SH Kang, T Nguyen, Distance based thresholds for cluster head selection in wireless sensor networks. IEEE Commun. Letters **16**(9), 1396–1399 (2012)
33. D Jia, H Zhu, S Zou, P Hu, Dynamic cluster head selection method for wireless sensor network. IEEE Sensors J. **16**(8), 2746–2754 (2016)
34. B Gangwar, JD Bhosale, N Gangwar, An energy optimized path selection and dynamic cluster head selection for wireless mesh network. ICEI, 272–277 (2017)
35. O El Ayach, SW Peters, RW Heath, The feasibility of interference alignment over measured MIMO-OFDM channels. IEEE Trans. on Veh. Tech. **59**(9), 4309–4321 (2010)
36. SW Peters, RW Heath, Cooperative algorithms for MIMO interference channels. IEEE Trans. on Veh. Tech. **60**(1), 206–218 (2011)
37. RS Sutton, AG Barto, *Reinforcement learning: an introduction* (Cambridge, (MA, MIT Press, 1998)
38. Watkins and Dayan, Q-learning, Machine learning, 8(3–4), pp.279–292(1992)
39. H Kim, KG Shin, Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks. IEEE Trans. on Mobile Computing **7**(5), 533–545 (2008)
40. S Kim, J Lee, H Wang, D Hong, Sensing performance of energy detector with correlated multiple antennas. IEEE Signal Proc. Letters **16**(8), 671–674 (2009)