Anti-Jamming RIS Communications Using DQN-Based Algorithm

PHAM DUY THANH\textsuperscript{1}, HOANG THI HUONG GIANG\textsuperscript{2}, AND IC-PYO HONG\textsuperscript{1}, (Member, IEEE)

\textsuperscript{1}Department of Information and Communication Engineering, Kongju National University, Cheonan 31080, South Korea
\textsuperscript{2}School of Electrical Engineering, Korea University, Seoul 02841, South Korea
Corresponding author: Ic-Pyo Hong (iphong@kongju.ac.kr)

This work was supported in part by the Basic Science Research Program under Grant 2020R11A3057142, in part by the Priority Research Center Program through the National Research Foundation under Grant 2019R1A6A1A03032988, and in part by the Underground City of the Future Program funded by the Ministry of Science and ICT.

\begin{abstract}
Wireless sensor networks (WSNs) are widely applied in battlefield surveillance, where the data collection (e.g. target tracking or content observation) employed by local sensor nodes needs to send to military bases for tactical decisions. Since data transmissions are susceptible to malicious attacks, jammers of adversaries can successfully block their victim’s communications by transmitting interfering signals to legitimate transmissions. Nowadays, owing to the ability to reconfigure the wireless propagation medium, reconfigurable intelligent surface (RIS) is regarded as an effective tool to enhance transmission performance, especially in the jamming context. This paper considers the anti-jamming communication tactical scenario of a solar-powered RIS network, in which the RIS is used to improve the uplink transmission performance between a wireless device (WD) and a base station (BS). We investigate the long-term anti-jamming communications of the WD powered by a solar energy harvester. Our objective is to jointly assign the optimal amount of transmission energy and the RIS phase shifts to maximize the data rate of the system in the long run. To this end, we formulate an anti-jamming communication optimization problem as a Markov decision process (MDP) framework and then design a deep Q network (DQN)-based algorithm to generate an optimal policy. As a result, the optimal resource allocation is achieved through trial-and-error interactions with environment by observing the predefined rewards and the network state transition. The Python simulation results conducted by averaging $10^4$ time slots show that the proposed algorithm is not only able to learn from environment, but also yields better performance than baseline schemes under network changes. Moreover, the performance of RIS communication schemes is verified to be superior to that of without-RIS communication schemes in the jamming context.
\end{abstract}

\begin{IEEEkeywords}
Reconfigurable intelligent surface, deep reinforcement learning, Markov decision process, solar-powered network, jamming attack, battlefield surveillance.
\end{IEEEkeywords}

\section{I. INTRODUCTION}

With the tremendous growth in data services for wireless communications (e.g., music, video, and games), demands for utilizing wireless resources have been also emerging explosively in recent years. However, due to constrained wireless resources and increasing power consumption, conventional networks might not satisfy these services. Therefore, both the academic and the industrial communities are paying more attention to developing efficient resource management schemes to improve both spectrum utilization and energy conservation for the emerging wireless devices in mobile Internet [1]. To this end, energy harvesting technology plays a crucial role, by which it allows a harvesting device to harvest energy from ambient sources, which is considered a potential solution to attain sustainable operation in dense cellular networks [2].

Nowadays, WSN is regarded as a crucial underlying framework to realize many applications of the internet of things (IoT) scenarios, such as health monitoring [3], traffic monitoring [4], military target tracking [5], and underwater communications [6]. However, WSN is extensively vulnerable to artificial malicious jamming owing to the open nature of wireless transmission medium and simple network
architecture. For instance, the jammers aim to inject interfering electromagnetic signals into wireless channels to suppress legitimate communications of sensor nodes [7]. Moreover, they are capable of tampering with data or masquerading as normal sensor nodes to transmit interfering false data [8]. For the above-mentioned issues, investigating countermeasures against disruptive threats caused by jammers is imperative and challenging in WSNs.

A. CONSIDERED JAMMING CONTEXT

Along with a proliferation of WSNs in civil cases, the military sensor network has been paid more attention to in battlefield surveillance applications, such as attack reconnaissance, battle damage assessment, and so on [5], [9]. In this paper, we mainly focus on the anti-jamming military battlefield context where reliable data transmissions are dominant to defend against interfering attacks. In the military, sensor nodes are deployed heterogeneously on the battlefield to gather data for combat purposes. The sensor nodes collect the data from combat zones and frequently send it to the main fusion center for tactical decisions. In this article, we consider the uplink transmission between a sensor node to a base station (i.e. main fusion controller), where there exist a lot of jamming attacks conducted by the adversary sides. The considered jamming attacks use high-power noise signals to disrupt legitimate communications. Apart from jamming attacks, sensor nodes generally suffer from short lifetimes due to a limited-capacity battery. Consequently, they cannot guarantee Quality of Service (QoS) for their critical transmission missions. Using energy harvesting becomes a promising solution in military sensor networks to prolong the battery lifespan of sensor nodes [10]. Therefore, it is imperative to design the optimal countermeasures against jamming attacks in energy harvesting-powered military sensor networks.

B. RELATED WORKS

To deal with jamming issues, several spectrum spreading technologies have been studied, e.g., direct sequence spectrum spreading (DSSS) [11] or frequency-hopping spectrum spreading (FHSS) [12]. However, on the one hand, they have some limitations such as significantly relying on a local pseudo-random code or a predetermined frequency-hopping pattern. On the other hand, new trends such as the diversity and dynamics of intelligent jammers, the higher requirements are needed for anti-jamming communication strategies. Recently, many studies have been conducted for the cooperative network using trusted relays to ameliorate the physical layer security [13]–[15]. There exist some efforts of designing the optimal strategy for anti-jamming communications [16]–[18]. The authors in [16] studied a dynamic channel allocation mechanism to deal with the spectrum conflicts and reduce the impacts of various jamming attacks. However, the wireless channels are simply modeled without considering the channel fading effect of wireless devices, leading to low feasibility of the proposed approaches. The spectrum assignment and routing scheme was developed to address jamming issues in a multi-hop scenario in [17], while the authors in [18] modeled the jamming attack using the game theory-based method to define the optimal strategies by assuming complete information of the adversary. Nevertheless, these works, [16]–[18], did not consider the energy harvesting context.

To cope with the uncertainty of jamming behaviors and power levels, some reinforcement learning (RL) algorithms were proposed to acquire the optimal jamming resistance strategy in wireless networks [19], [20]. Although the effectiveness of the foregoing schemes was validated [13]–[15], [19], [20], deploying many active relays may increase the hardware cost as well as complexity for the networks. Nowadays, thanks to the ability to reconfigure the wireless transmission medium for enhancing the system capacity, reconfigurable intelligent surface (RIS) has received increasing attention [21]. RIS is a planar array that consists of passive reflecting elements for adaptively controlling the electromagnetic wave. Especially, it can enhance or weaken the signals received by various wireless users. Consequently, the RIS has been broadly investigated for wireless security performance optimization [22]–[28]. However, in [22]–[25], the authors only pay attention to the security enhancement for RIS-assisted communications systems in presence of eavesdroppers by jointly determining the optimal beamforming and phase shifts of the RIS.

To the best of the authors’ knowledge, only a few works focus on anti-jamming strategies for RIS-assisted communications [26]–[28]. The RIS carried by an unmanned aerial vehicle (UAV) is used to simultaneously mitigate jamming signals and enhance legitimate signals [26]. However, employing the assisting UAV may impose a high cost, and a trade-off between the achievable throughput and additional overhead was not considered. The authors in [27] designed an RL-based anti-jamming scheme by jointly deriving optimal power allocation and reflecting beamforming, while the RIS-assisted approach against both jamming and eavesdropping attacks was proposed to maximize the achievable rate [28]. Nevertheless, the studies in [26]–[28] assumed the knowledge of the jamming channel states is known, which is not realistic because the jammer is considered an external user whose channel information is hard to achieve. Furthermore, none of these existing works considers the sustainable operation of the wireless devices using energy harvesting and the impact of jamming sensing errors on the network performance. Thus, this paper studies a solar-powered RIS communications under jamming attacks, where a wireless device capable of harvesting solar energy aims at achieving efficient data transmissions against external jamming. On the one hand, we consider the impact of jamming sensing errors when the legitimate user senses the presence of jamming signals in its communication region. On the other hand, we employ the deep reinforcement learning (DRL) framework to solve the issues of large state and action spaces of the considered optimization problem, which is challenging for traditional RL
approaches. We summarize the main contribution of this work as follows:

- We first investigate the RIS-assisted military wireless communication system in the presence of jamming attacks. In contrast to [26]–[28], where the illegitimate channel state information is assumed to be known or predicted, we adopt the spectrum sensing using energy detection to detect the state active/inactive of the jammer. Although this sensing method does not require illegitimate channel knowledge of the third party, the system may experience the degradation of system capacity due to false alarm and misdetection circumstances. Therefore, we propose a scheme to obtain intelligent resource allocation under dynamic jamming and imperfect sensing.

- We further apply the solar energy harvesting for the wireless device to attain sustainable operation in the battlefield deployment. The constraints of limited harvesting energy, battery capacity, and environment dynamics are taken into account to develop a secure RIS transmission strategy against jamming attacks.

- Next, the joint transmission energy allocation and RIS phase-shift configuration problem for long-term data rate maximization against jamming attacks is formulated as a MDP framework. Subsequently, the DRL-based resource allocation scheme is proposed to achieve an optimal solution to the MDP problem. Thereby, the legitimate user can interact directly with the environment to learn the environment knowledge and dynamic behavior of the jammer, thus obtaining the optimal transmission energy at the wireless device and phase shifts of the RIS via trial-and-error training.

- Lastly, the proposed scheme is verified to efficiently work with the dynamic changes of environment and the imperfect sensing through simulation results. The effectiveness of our algorithm is assessed through the simulation in comparison with the conventional existing approaches, in which long-term and short-term optimization are taken into consideration. Moreover, the algorithms with non-RIS-assisted communication are also evaluated to validate the advances of employing the RIS on improving the performance of wireless communications under the jamming attack scenario.

The remainder of the work is organized as follows. Section II elaborates the system description. Next, Section III presents the DRL-based resource allocation scheme. We discuss the numerical simulation results in Section IV and then conclude the study in Section V.

II. SYSTEM DESCRIPTION

We consider a RIS-assisted battlefield uplink transmission network, which is composed of a wireless device (WD), a base station (BS), a reconfigurable intelligent surface, and a malicious jammer (JM). The WD attempts to transmit data collected in the observed battlefield zone to the BS, while the JM aims to transmit artificial noise signals to block the transmissions between the WD and the BS, as shown in Fig. 1. This jamming scenario usually occurs when one side attempts to combat with its opposing side in warfare. In the network, the WD has a single antenna while the BS is equipped with $M$ antennas. On the other hand, the RIS has $K$ reflecting elements, thus it can adjust the phase shifts to reconfigure wireless channel conditions between the WD and the BS. The system operation time is divided into $T$ time slots with time slot duration $\tau$, and each slot is indexed by $t \in \mathcal{T} = \{1, 2, \ldots, T\}$.

A. TRANSMISSION MODEL

By leveraging RIS’s aid, the WD transmits the signal to the BS through two different wireless links: direct link and reflected RIS link. The direct path between the WD and the BS is assumed to be a non-LOS (NLOS) channel. The channel of the direct link is given by $\mathbf{h}_{\text{DB}} = \mathbf{g}_{\text{DB}}[h_{1}^\text{DB}, \ldots, h_{K}^\text{DB}]^T \in \mathbb{C}^{M \times 1}$, where $\mathbf{g}_{\text{DB}} = 1/\sqrt{d_{\text{DB}}^2}$ is the path loss between the WD and the BS at the distance $d_{\text{DB}}$ with a path loss exponent $\alpha_{\text{L}}$; and $h_{mk}^\text{D}$ denotes the small-scale fading between the WD and the BS. The reflected RIS path consists of two links: an NLOS link between the WD and the RIS and an LOS link between the RIS and the BS. The NLOS channel between the WD and the RIS is denoted by $\mathbf{h}_{\text{DR}} = \mathbf{g}_{\text{DR}}[h_{1}^\text{R}, \ldots, h_{K}^\text{R}]^T \in \mathbb{C}^{K \times 1}$, where $\mathbf{g}_{\text{DR}}$ is the path loss between the WD and the RIS and $h_{mk}^\text{R}$ is the complex Gaussian random variable, $\mathcal{CN}(0, 1)$. Let $\mathbf{G}_{\text{RB}} = \mathbf{g}_{\text{RB}}[h_{mk}]_{m \in \{1, M\}, k \in \{1, K\}} \in \mathbb{C}^{M \times K}$ represent the channel between the RIS and the BS, where $\mathbf{g}_{\text{RB}} = 1/\sqrt{4\pi d_{\text{RB}}^2}$ represents a free space path loss attenuation between the RIS and the BS at the distance $d_{\text{RB}}$, and $h_{mk}$ is the LOS channel state between antenna $m$ of the BS and element $k$ in the RIS. We assume that $\mathbf{h}_{\text{DB}}, \mathbf{h}_{\text{DR}},$ and $\mathbf{G}_{\text{RB}}$ are quasi-static and perfectly known at the WD by using the pilot signal transmissions and feedback. Hence, the WD can obtain these information at the beginning of each time slot to make
the transmission decision. When the RIS receives the signal transmitted by the WD, it will reflect the signal to the BS with the vector of phase shifts $\Phi = \text{diag} ( e^{\phi_1}, e^{\phi_2}, \ldots, e^{\phi_K} )$.

where $\phi_k \in \mathcal{F}_{\text{ris}} = \left\{ \frac{2\pi l}{2^K} \right\}_{l=0}^{2^K-1}$ is the phase shift of element $k$, and $b$ represents the resolution of the phase shifter of the RIS elements [29]. The signal received at the BS is expressed as follows

$$y = \sqrt{p_t} ( G_{RB} \Phi h_{DR} + h_{DB} ) x + n,$$

where $p_t = \frac{e_t}{\tau f_0}$ is the transmission power of the WD, $e_t$ represents the transmission energy, and $x$ denotes the information of the WD; $n = [n_1, n_2, \ldots, n_M]^T$ represents noise vector where each element is the zero-mean complex white Gaussian noise with variance $\sigma^2$, i.e., $n_m \sim CN (0, \sigma^2)$. The signal-to-noise ratio (SNR) received at the BS can be given as follows

$$\gamma = \frac{p_t \| G_{RB} \Phi h_{DR} + h_{DB} \|^2}{\sigma^2}.$$  

(2)

In this paper, we model the jamming behavior of the JM on the channel as a 2-state discrete-Time Markov Chain process. Particularly, the state of the JM in a time slot is denoted as “A” (active) or “I” (inactive), and the state transition probability between two adjacent time slots of the JM is represented as $P_{ij} \forall i, j \in \{ A, I \}$. We further assume that the JM has enough energy for its jamming operation, such that data received at the BS can not be successfully decoded due to the high-power interference injected by the JM once it performs jamming action on the channel (i.e., the JM is active). Thus, the achievable data rate of the system in (b/s/Hz) is calculated as

$$r = \frac{\tau - t_s}{\kappa \log_2 (1 + \gamma)},$$

(3)

where $\kappa = \begin{cases} 1 & \text{if the JM is inactive} \\ 0 & \text{otherwise} \end{cases}$ is the jamming indicator, and $t_s$ represents the sensing duration. In this article, the WD is assumed to have a battery with the limited capacity, $E_{\text{bat}}$, and it is equipped with an energy harvester to collect solar energy from environment for long-term operation. We assume that the harvested energy of the WD during time slot $t$, $e_{h}(t)$, is limited, where $0 < e_{h}(t) < E_{\text{bat}}$, and that it follows a Poisson distribution with mean harvested energy, $\epsilon$ [30].

The operation in a time slot of the system can be expressed as follows. At the start of a time slot, the WD performs spectrum sensing to detect the presence of the channel. Subsequently, if the sensing result indicates the state “inactive” of the JM, the WD will transmit data to the BS; otherwise, it has to stay silent to avoid the interference of the JM. In this paper, we take into account the sensing errors of the WD. The sensing result, $\theta (t) \in \{ A, I \}$, indicates the state (i.e., active or inactive) of the JM in time slot $t$. Nevertheless, the sensing error is indeed inevitable owing to the imperfection of the sensing engine. Two crucial metrics generally considered are false alarm probability, $P_f = \Pr ( \theta(t) = A \mid I )$, and misdetection probability, $P_m = \Pr ( \theta(t) = I \mid A )$. The former represents the probability that the state of the jammer is sensed as “active”, but it is actually “inactive”. In contrast, the latter indicates the probability that the state of the JM is sensed as “inactive” but it is actually “active”. The transmission performance can be significantly degraded due to these sensing errors. Hence, designing a jamming-resistance communication strategy of the RIS-assisted network under imperfect spectrum sensing is essential.

### B. PROBLEM FORMULATION

Although maximizing the transmission power at the WD can optimize the immediate reward, it may lead to the energy shortage for the upcoming time slots due to limited-capacity battery and stochastic harvested energy. Furthermore, the accumulative data rate in the long run can also be affected significantly by channel fading, misdetections, and false alarms. We study the problem of jointly determining transmission energy at the WD and phase shifts of the RIS for uplink transmissions. Our objective is to maximize the long-term system rate. The problem formulation is expressed as

$$\max_{e_t(t), \Phi(t)} \sum_{t=1}^{\infty} r(t)$$

subject to

$$0 \leq e_t(t) \leq e_{\max}$$

(b)\hspace{1cm} r(t) \geq r_{\min}$$

(c)$$

$$\Phi_k(t) \in \mathcal{F}_{\text{ris}}$$

(4)

where $e_t(t)$ denotes the transmission energy at the WD and $e_{\max}$ is the maximum transmission energy at the WD. $\Phi(t)$ is the vector of phase shifts in the RIS. Constraint (a) ensures the transmission energy assigned for the WD does not exceed the maximum level, while the received instant rate will meet the minimum requirement of data rate, $r_{\min}$, with constraint (b). Constraint (c) is imposed to satisfy that the phase shift of each RIS element should be chosen in the set of phase shift candidates, $\mathcal{F}_{\text{ris}}$.

It is challenging to obtain directly the solution for non-trivial optimization problem (4) due to network constraints of the limited battery capacity, time-varying energy resource, and jammers’ dynamic behavior. The exhaustive search may have to be adopted to acquire the optimal solution by utilizing classical mathematical tools and sophisticated formulations. However, it is impossible for such a large-dimension optimization problem. The partially observable Markov decision process, considered as one of the feasible solutions, can be adopted to achieve the optimal solution using value iteration-based programming. However, the input state values are generally discretized and the environment knowledge is assumed to know in advance, leading to the impractical implementation. On the other hand, the conventional RL algorithms, such as Q-learning and actor-critic can provide the sub-optimal solution to address the unknown prior knowledge of the network by optimizing the Q-value function through trial-and-error training. Unfortunately, these
RL approaches can not effectively work with the enormous dimension problem.

Unlike most of the above-mentioned schemes which generally face with the difficulties of prior knowledge, continuous state spaces, and enormous dimension optimization problems, DRL is considered one of the potential tools. The key idea is to use a deep neural network where the value function is approximated by continuously modifying parameters through training. Therefore, in this paper, rather than directly solving the challenging optimization problem mathematically, we reformulate the long-term data rate optimization problem (4) to the MDP framework and then obtain the optimally feasible $e_r$ and $\Phi$ by using the DRL-based method. By leveraging the RL principle, the agent (i.e., the WD) can continuously interact with environment to learn the channel knowledge, energy harvesting pattern, and jammer behavior, resulting in the optimal strategy over time without prior knowledge from environment. The formulation of the MDP framework and proposed DRL scheme are respectively elaborated in the following section.

III. DEEP Q NETWORK FRAMEWORK
In this section, we reformulate problem (4) into a MDP framework. Subsequently, we investigate a DQN-based resource allocation scheme to obtain the solution to the MDP problem where the energy harvesting distribution is unknown in advance. By leveraging the advancement of the reinforcement learning, the proposed scheme can directly learn the optimal policy by interacting with the environment. Furthermore, it can also cope with the issue of the large dimensional space of the optimization problem, which is hard to tackle for traditional RL approaches.

A. FRAMEWORK OF MARKOV DECISION PROCESS
We first define the state, action spaces, and reward of the MDP problem as follows. Let $\mathcal{S}$ be the state space, in which each state includes the jamming probability of the jammer on the channel (i.e., the probability that the JM is active), $\rho(t)$, the energy level in the battery of the WD, $e_r(t)$, and the channel information, $\mathbf{h}(t)$ as the following:

$$s(t) = \left(\rho(t), e_r(t), \mathbf{h}(t)\right) \in \mathcal{S},$$

where $\mathbf{h}(t) = \mathbf{G}_{\text{RB}}(t)\mathbf{\Phi}_{\text{DR}}(t) + \mathbf{h}_{\text{DB}}(t)$. The action with respect to state $s(t)$ is denoted by $a(t)$. We define the action space of the network by

$$A = \left\{a_1, a_2, \ldots, a_{|A|}\right\},$$

where $a_i(t) = (e_p(t), \Phi(t))$ is the action including the amount of assigned transmission energy of the WD and the phase shifts of the RIS. The system receives an immediate reward after executing action $a(t)$ in state $s(t)$, which is calculated by equation (3), i.e., $R(t) = r(t)$. The operation of the WD can be described as follows. At the beginning of a time slot, the WD recognizes the current state of the network, $s(t)$. Then it selects and executes an action, $a(t)$, by which the WD will transmit data to the BS with an assigned amount of transmission energy and phase shifts. Subsequently, the BS receives an immediate reward and then feedbacks ACK/NACK messages regarding whether the data is successfully decoded or not to update the state $s(t+1)$ at the end of the time slot. It is worth noting that data received by the BS is unsuccessfully decoded if the JM is active due to the assumption of the high noise power infected by the JM to the legitimate transmission. In the following, we will elaborate the process of updating the network state based on possible observations.

Observation 1 ($O_1$): The sensing result indicates the state “A” of the JM, the WD will trust the result and stay silent. Thus, there is no reward in this case, i.e., $R(s(t), a(t) | O_1) = 0$. The jamming probability for the next time slot is updated as

$$\rho(t+1) = \frac{\rho(t)P_H + (1 - \rho(t))(1 - P_m)P_{AI}}{\rho(t)P_F + (1 - \rho(t))(1 - P_m)},$$

where $P_H$ and $P_{AI}$ respectively represent the transition probability from state I to state I, and from state A to state I of the JM between two adjacent time slots. The energy level of the WD can be calculated by

$$e_r(t+1) = e_r(t) - e_s(t) + e_h(t),$$

where $e_s(t)$ denotes the sensing energy of the WD.

Observation 2 ($O_2$): The WD obtains the state “I” from the sensing engine and then transmits data to the BS, and it finally receives the ACK message sent by the BS at the end of the time slot. In this case, the reward is given by

$$R(s(t), a(t) | O_2) = r(t).$$

The jamming probability in the next time slot is

$$\rho(t+1) = P_H.$$

We update the energy level of the WD as follows

$$e_r(t+1) = e_r(t) - e_s(t) - e_{tr}(t) + e_h(t).$$

Observation 3 ($O_3$): The sensing result shows the state “I” of the JM and then the WD transmits data to the BS, and it finally receives the NACK message sent by the BS at the end of the time slot. Thus, the misdetection is confirmed in this case; thus, there is no reward obtained, i.e., $R(s(t), a(t) | O_3) = 0$. The jamming probability is updated by

$$\rho(t+1) = P_{AI}.$$

The energy level of the WD is calculated same as Eq. (11).

In this work, our target is to maximize the long-term data rate of the solar-powered network under the jamming attack from the current state. That means the impact of the current action on the future reward is taken into consideration owing to the uncertainty of environment. We define the state-action value function (or Q-value function) as the expected value
of accumulative system reward with state $s$ and action $a$ as follows:

$$Q_\psi(s, a) = \mathbb{E}_\psi \left[ \sum_{i=0}^{\infty} \alpha^{i-t} R(t) | s(t) = s, a(t) = a \right],$$  \hspace{1cm} (13)

where $\mathbb{E} [.]$ refers to the expectation operator, and $\alpha$ is the discount rate. The Q-value function is a metric to assess the effect of the action choice on the expected future reward. Based on Q-learning principle, the optimal Q-value function associated with the optimal policy $\psi^*$ is expressed as

$$Q^*(s, a) = R(s, a) + \alpha \sum_{s' \in S} P^a_{ss'} \mathbb{E}_{\psi} \left[ \sum_{a' \in A} \psi(s', a') Q_\psi(s', a') \right].$$  \hspace{1cm} (14)

where $s'$ and $a'$ are respectively the next state and next action; $R(s, a)$ is the instant reward received by performing the action $a$ at state $s$; $P^a_{ss'}$ represents the state transition probability of the network from state $s$ to state $s'$ with action $a$ being performed. Based on Q-learning principle, the optimal Q-value function associated with the optimal policy $\psi^*$ is obtained by the training process under policy $\psi$. The Q-value function satisfies the following Bellman equation:

$$Q_\psi(s, a) = R(s, a) + \alpha \sum_{s' \in S} P^a_{ss'} \psi(s', a') \left[ \sum_{a' \in A} Q_\psi(s', a') \right].$$  \hspace{1cm} (13)

where $\mathbb{E} [.]$ refers to the expectation operator, and $\alpha$ is the discount rate. The Q-value function is a metric to assess the effect of the action choice on the expected future reward. Based on Q-learning principle, the optimal Q-value function associated with the optimal policy $\psi^*$ is expressed as

$$Q^*(s, a) = R(s, a) + \alpha \sum_{s' \in S} P^a_{ss'} \max_{a' \in A} Q^*(s', a').$$  \hspace{1cm} (15)

The optimal Q-value function in Eq. (15) can be obtained by recursive rule without the exact knowledge of environment dynamics and state transition models which are highly dependent on each action. To this end, the state-action value function is frequently updated with the learning rate $\lambda \in (0, 1)$ as follows:

$$Q_\psi(s, a) = (1 - \lambda) Q_\psi(s, a) + \lambda \left( R + \max_{a' \in A} Q_\psi(s', a') \right).$$  \hspace{1cm} (16)

$Q(s, a)$ will converge to the optimal value through the training process if it is continuously updated at every time slot under a proper configuration. Nevertheless, the original Q-learning algorithm might yield wide variances in function approximation, resulting in the locally optimal policy. Especially, the problem becomes more challenging when the size of the problem increases [31]. To address the large dimensional issues, we adopt the deep reinforcement learning mechanism, in which a neural network is used to approximate the Q-value function. More particularly, we proposed the DRL scheme using a neural network with weight $w$ to approximate the Q-value function, denoted by $Q(s, a, w)$, by which the Q-value function is determined by $w$ as $Q(s, a, w) = Q(s, a, w)$. Consequently, the proposed scheme can work efficiently in the large dimension networks. The optimal policy is explored by using $\varepsilon-$greedy policy regarding the action choice in each step, where an action is randomly selected with probability $\varepsilon$, or extracted from the policy with probability $1 - \varepsilon$. The proposed DQN-based scheme is presented in more details in the following subsection.

B. THE PROPOSED DQN SOLUTION

In this subsection, we present the proposed DQN scheme to obtain the solution of the MDP problem, in which the feed-forward neural network (FNN) is adopted to approximate the Q-value function according to each state, named a Q-network. The FFN network is composed of an input layer, multiple hidden layers, and an output layer, as illustrated in Fig. 2, where the input is the system state $s$ while the output represents a set of the Q-values of state-action pairs. The input layer contains $M + 2$ neuron units that refer to the number of elements in each state. Each hidden layer is a fully connected layer, and the rectified linear unit function is utilized as a
nonlinear activation function. The output of the hidden layers can be expressed by
\[ z = \max (w \cdot s + u, 0), \quad (17) \]
where \( w \) and \( u \) denote the weight and bias, respectively. The output of the FNN is designed to connect with the last hidden layer to estimate the Q-value of each state-action pair. It has the size of \(|A|\) and uses the linear activation function. Let us define the loss function as the mean square error between the current value and the target Q-value as follows:
\[ L(w) = \mathbb{E} \left[ \left( R + \alpha \max_{a'} Q(s', a', w) - Q(s, a, w) \right)^2 \right], \quad (18) \]
where \( R + \alpha \max_{a'} Q(s', a', w) \) refers to the target value.

By training the FNN, the network parameters are adjusted to minimize the loss function. However, the crucial issue in Q-value function approximation using neural network is that, the states are remarkably correlated in time domain and cause the reduction of randomness of states because they are all extracted from the same episode, resulting in the high oscillation as well as low convergent rate during the training. To improve the network performance, we adopt two well-known methods: experience replay [32] and fixed target network [31]. By using the experience replay, the sample \( \{s, a, R, s'\} \) is restored in a data storage, namely, a replay buffer, \( D \), from which a random mini-batch is chosen for training the Q-network in each time slot. By doing this, rather than updating the parameters from the last state, the DNN can update from a batch of randomly sampled states to the experience replay.

On the other hand, with fixed target network method, we can compute the target value by using another neural network with weight \( \hat{w} \), and the network parameters are kept unchanged within a finite number of time slots. Consequently, by using fixed target neural network, the system reward is optimized by using the following equation:
\[ L(w) = \mathbb{E}_D \left[ \left( R + \alpha \max_{a'} Q(s', a', \hat{w}) - Q(s, a, w) \right)^2 \right]. \quad (19) \]

The target network is designed with the same structure of the Q-network, and its parameters are frequently replaced by those of the Q-network. We update the weight vector \( w \), such that the loss function is minimized by using stochastic gradient descent as follows
\[ w = w + \lambda \delta \nabla_w Q(s, a, w), \quad (20) \]
where \( \lambda \in (0, 1) \) is the learning rate, and the temporal difference (TD) error, \( \delta \), between the current Q-value and the target value, is computed by
\[ \delta = R + \alpha \max_{a'} Q(s', a', \hat{w}) - Q(s, a, w). \quad (21) \]

Fig. 2 illustrates the flowchart of the proposed DQN-based scheme. With this method, the agent selects an action at the beginning of each time slot based on \( \varepsilon \)-greedy policy, where \( \varepsilon \) denotes the exploration rate. After the WD transmits data to the BS, the system receives a reward and turns to the next state. The data sample, \( \{s, a, R, s'\} \), will be stored in the replay buffer. Subsequently, the mini-batch is selected among the dataset in the replay buffer to update the parameters of the target network and deep Q network. Through each step, the exploration rate decays with the decay rate \( \eta \), and the training process is repeated until convergence. Algorithm 1 outlines the training procedure of the proposed DQN-based scheme.

Algorithm 1 DQN-Based Training Process

1. Initialize the parameters \( w \) and \( \hat{w} \).
2. Initialize exploration rate \( \varepsilon \).
3. Initialize replay memory \( D \).
4. for each training episode do
5. Select an initial state \( s \in S \).
6. for each step \( t = 1, 2, \ldots, T \) do
7. Determine \( \varepsilon = \max (\varepsilon_{\eta}, \varepsilon_{\min}) \).
8. Observe network state, \( s(t) \).
9. Sense environment to detect the JM’s state.
10. if \( \theta (t) = "A" \) then
11. Stay silent.
12. else
13. Select action \( a(t) \) using \( \varepsilon \)-greedy policy.
14. end if
15. Perform action \( a(t) \).
16. Receive reward \( R(t) \) using Eq. (3).
17. Observe next state \( s(t + 1) \).
18. Store the sample of the transition to memory \( D \).
19. Randomly select a mini-batch from memory \( D \).
20. for \( i \) in mini-batch’s size do
21. Compute the estimated value of \( Q(s_i, a_i, w) \).
22. Compute \( R_i + \alpha \max_{a'} Q(s_{i+1}, a', \hat{w}) \).
23. end for
24. Update next state \( s(t + 1) \).
25. Update Q-network parameter \( w \) using Eq. (20).
26. end for
27. Update target network parameter, \( \hat{w} \leftarrow w \).
28. end for

IV. NUMERICAL RESULTS AND DISCUSSIONS

In this section, we present numerical simulation results to assess the effectiveness of the proposed scheme, compared to other baseline schemes. The results are obtained by employing Python 3.8 with Anaconda 2021 distribution. The comparisons were made by considering three baseline schemes: classical DQN scheme [33], Myopic scheme [34], and Random scheme. In the classical DQN scheme, the agent adopts the deep Q-learning method without considering the sensing error metrics such as false alarm and misdetection. The term “myopic” in the Myopic scheme implies that the system cares...
more about the immediate reward than the future reward. In particular, the agent in Myopic scheme selects the actions by optimizing the immediate reward. On the other hand, the policy in the random scheme was made based on random fashion.

We set the distance between the WD and the BS, the WD and the RIS, and the RIS and the BS at 28, 21, and 15 meters, respectively. The number of antennas of the BS is $M = 8$ and the number of elements in RIS is $K = 4$. The battery of the WD has a capacity of $E_{bat} = 30 \mu J$. The mean value of harvested energy is $\epsilon = 7 \mu J$. The maximum transmission energy is set to 14 $\mu J$. The state transition probabilities of the JM between two adjacent time slots are $P_{H} = 0.8$ and $P_{AL} = 0.2$. The false alarm and misdetection probabilities are $P_{f} = 0.1$, and $P_{m} = 0.9$, respectively. We set the discount rate to $\alpha = 0.9$. The learning rate is $\lambda = 0.001$. Four layers are designed in the neural network: an input layer, two hidden layers with 24 nodes each, and an output layer. We use the ReLU function and the linear function for the activation function of the hidden layers and the output layer. We train the Q-network over 200 episodes, each of which contains $10^4$ time slots. The simulation results were realized by averaging $10^4$ time slots.

**TABLE 1. Simulation parameters.**

| Description                  | Parameter | Value      |
|------------------------------|-----------|------------|
| Time slot duration           | $\tau$   | 0.2 $s$    |
| Sensing time                 | $t_s$     | $2 \times 10^{-3}$ $s$ |
| Resolution of phase shifter  | $b$       | 2 bits     |
| Path loss exponent           | $\alpha_L$ | 3.6        |
| Battery capacity             | $E_{bat}$ | 30 $\mu J$ |
| Noise variance               | $\sigma^2$ | $-65$ dBm |
| Minimum requirement of data rate | $r_{min}$ | 1 b/s/Hz |
| Mean harvested energy        | $\epsilon$ | 7 $\mu J$ |
| Discount rate                | $\alpha$  | 0.9        |
| Learning rate                | $\lambda$ | 0.001      |

The selection of network parameters determines the learning convergence rate and system efficiency. To demonstrate the importance of network parameters selection, we examine the convergence rate of the proposed algorithm with the increasing number of training episodes and different learning rates, i.e., $\lambda = \{0.0001, 0.001, 0.01, 0.1\}$, in Fig. 3. In the simulation, we regularly compute the average value of the cumulative rewards received in the number of time slots through each episode to visualize a curve with less oscillation during the training. In this paper, the Q-network keeps training until it meets the convergence condition ($\leq 0.001$) or reaches a predefined number of training episodes. We can see that learning rates yield significant impacts on the reward performance of the proposed DQN algorithm. Although the figure shows that the proposed algorithm is able to learn the environment knowledge by observing the immediate rewards and improves the effectiveness of the policy step by step, the reward may converge to the locally optimal values. If we set too-small learning rate of $\lambda = 0.0001$, the system requires more time to acquire the convergence, resulting in a low reward. On the other hand, there exists oscillation in behavior and worse convergent property for too-large learning rate of $\lambda = 0.1$. With the case of learning rate of $\lambda = 0.001$, the system can obtain the best reward compared to the others, although it has a slightly lower convergence rate than in the case of $\lambda = 0.01$. Thus, we should select the proper learning rate which is neither too large nor too small, and the value is chosen at 0.001 for the rest of the simulation.

Fig. 4 compares the performance of the proposed algorithm with the other conventional methods. The increase in the average rewards of the proposed scheme and classical DQN scheme shows that the policies are improved as time goes on. During the first 20 episodes, these policies bring similar rewards because the system requires more time to explore the environment dynamics by frequently executing random actions. As depicted from the figure, the classical DQN converges much faster than the proposed scheme because it does not consider the jamming probability in the state, which lowers the state space but can not obtain the optimal policy due to the sensing errors. In contrast to the classical DQN scheme, the proposed scheme can estimate the jamming probability of the JM to obtain the proper action for the WD against the JMs’ attack in each time slot. As the result, the reward in the proposed algorithm rapidly increases within about the first 75 episodes and then gradually converges to the optimal value during the training time of 200 episodes. On the other hand, the rewards of the non-learning methods: Myopic...
Figure 4. Convergence property of the proposed algorithm and other baseline schemes.

Figure 5. Reward with respect to the maximum transmission powers.

Figure 6. Data rate versus the harvested energy.

Figure 7. Energy efficiency according to the harvested energy.

In order to get better understanding of our proposed scheme, we further inspect the impact of the maximum transmission power $e_{\text{max}}$ on the performance in Fig. 5, in which two settings $e_{\text{max}} = 11 \mu J$ and $e_{\text{max}} = 14 \mu J$ with episodic average reward and episodic instant reward are considered as the function of the number of episodes. In the following simulation, the episodic average rewards at episode $j$ are calculated by

$$ R_{\text{avg}}^j = \frac{1}{j} \sum_{z=1}^{j} R_{z}^{\text{inst}}, $$

where $R_{z}^{\text{inst}} = \frac{1}{T} \sum_{t=1}^{T} R(t)$ represents the episodic instant reward at episode $z$. We can see that, the instant reward converges faster at the low SNR (i.e., $e_{\text{max}} = 11 \mu J$) than high SNR (i.e., $e_{\text{max}} = 14 \mu J$). It is because the dynamic range of the transmission energy amount is smaller, leading to the good convergence. In contrast, with the higher SNR, the transmission energy range is large, the system can cause more fluctuations and worse convergence during the training process.

To assess the effectiveness of the proposed scheme, we plot the data rate of the proposed scheme and other baseline schemes versus increasing mean values of harvested energy in both cases of with-RIS and without-RIS scenarios in Fig. 6. It is observed that the system can obtain higher throughput as the harvested energy increases with all schemes. The reason is that the WD has more energy to transmit data to the BS with the increment of energy harvested from environment. Furthermore, the proposed scheme brings the best performance as compared to other traditional schemes in both cases of with-RIS and without-RIS scenarios. In this article, the energy efficiency of the system is defined as the number of achievable bits over the amount of energy consumption at the WD. We further see that the energy efficiency of the proposed scheme is superior to that of the classical DQN, Myopic, and Random schemes, as illustrated in Fig. 7.

Fig. 8 shows the impact of the battery capacity of the WD on the performance of the schemes. We see that the...
improvement of data rate can be achieved with the increased battery capacity. This can be explained as follows. For the larger battery capacity, the WD stores more harvested energy for its long-term operation, which can enhance the performance by frequently using higher transmission power for data transmissions. It is worth noting that the data rate of our algorithms can dominate other approaches. To explain this, the classical DQN did not consider the sensing error impact, even though it adopts the long-term data rate optimization, and the WD in the short-term scheme (i.e., Myopic scheme) aims to optimize the system performance by only maximizing the immediate reward without considering the impact of the current action on the future utilization. As a consequence, the system using these conventional schemes can not achieve the optimal policy due to the dynamic environment of the wireless network. In addition, we can see from the figure that the schemes using RIS can achieve up to 3% gain in terms of data rate, compared to those without using the RIS.

To evaluate the effect of the number of RIS elements on system rate, the performance comparison of different algorithms is shown in Fig. 9. The curves show that increasing the number of RIS elements will help the system obtain a higher data rate. That is because the gain of reflecting channel in the system goes up with the number of RIS elements. In Fig. 10, we further investigate the system performance according to various changes of false alarm probabilities of the sensing engine, i.e., $P_f \in [0.1, 0.2]$, and state transition probabilities of the JM, i.e., $P_H = [0.6, 0.8]$. It can be seen that the system is significantly degraded as the probability of false alarm becomes large. It can be explained as follows. When the sensing engine of the WD causes more false alarms, the WD will miss more opportunities to transmit data to the BS, resulting in lower throughput. On the other hand, the system achieves more rewards when $P_H$ increases. It reflects the fact that the WD will frequently perform data transmissions. As a result, the proposed scheme yields an optimal strategy for the WD because it can estimate the jamming probability of the JM for selecting the optimal action in each step. This leads to energy-efficient anti-jamming communications under the network changes.
Finally, in Fig. 11, we plot the data rate of the schemes versus the number of antennas of the base station and the varying misdetection probabilities caused by the sensing engine. The curves show that the data rate goes up with the increase in the number of BS antennas. In contrast, the performance decreases with the increasing values of $P_m$, because the system will frequently experience the jamming interference with the higher misdetection probability. The results show that the proposed scheme is robust with the environment dynamics by not only considering the energy conservation at the energy-constrained WD but also predicting the JM activity to acquire the optimal strategy for the anti-jamming optimization problem in the long run.

V. CONCLUSION

In this article, we investigated the joint transmission energy allocation and RIS phase shifts optimization problem for the solar-powered RIS-assisted communication against jamming attacks. The DQN-based algorithm was proposed, in which the agent can adapt its policy to the dynamic variations of environment to attain the optimal strategy under sensing errors, limited battery capacity, and energy harvesting. Specifically, by considering the jamming prediction of the jammer and learning the energy evolution of the wireless device, the proposed scheme can gradually select the optimal strategy to maximize long-term data rate. As a consequence, numerical simulation results validated the superiority of the proposed scheme to other existing benchmarks with respect to the changes in the network situation. Although the effectiveness of the proposed scheme was verified, the system only addressed the considered network problem with discrete actions. Therefore, our future work aims at developing intelligent resource allocation schemes to solve secure transmission problems with continuous action spaces by using other artificial intelligence methods such as deep deterministic policy gradient algorithms.

REFERENCES

[1] M. Masoudi, “Green mobile networks for 5G and beyond,” IEEE Access, vol. 7, pp. 107270–107299, 2019.

[2] G. Piro, M. Miozzo, G. Forte, N. Baldo, L. A. Grieco, G. Boggia, and P. Dini, “HetNets powered by renewable energy sources: Sustainable next-generation cellular networks,” IEEE Internet Comput., vol. 17, no. 1, pp. 32–39, Jan. 2013.

[3] Z. Pengfei, T. K. Boon, and W. Yixin, “Sensor selection in wireless sensor networks for structural health monitoring,” in Proc. IEEE SENSORS, Oct. 2019, pp. 1–4.

[4] B. Rashid and M. H. Rehmani, “Applications of wireless sensor networks for urban areas: A survey,” J. Netw. Comput. Appl., vol. 60, pp. 192–219, Jan. 2016.

[5] A. Ullah, N. Ishaq, M. Azeem, H. Ashraf, N. Z. Jianhui, M. Humayun, T. A. Tabbakh, and Z. A. Almusaylim, “A survey on continuous object tracking and boundary detection schemes in IoT-assisted wireless sensor networks,” IEEE Access, vol. 9, pp. 126324–126336, 2021.

[6] S. Ashraf, M. Gao, Z. Chen, H. Naeem, A. Ahmad, and T. Ahmed, “Underwater pragmatic routing approach through packet reverberation mechanism,” IEEE Access, vol. 8, pp. 163091–163114, 2020.

[7] K. Pelechrinis, M. Iliofotou, and S. Krishnamurthy, “Denial of service attacks in wireless networks: The case of jammers,” IEEE Commun. Surveys Tuts., vol. 13, no. 2, pp. 245–257, May 2010.

[8] M. Young and R. Boutaba, “Overcoming adversaries in sensor networks: A survey of theoretical models and algorithmic approaches for tolerating malicious interference,” IEEE Commun. Surveys Tuts., vol. 13, no. 4, pp. 617–641, 4th Quart., 2011.

[9] D. Thomas, R. Shankaran, M. Orgun, M. Hitchens, and W. Ni, “Energy-efficient military surveillance: Coverage meets connectivity,” IEEE Sensors J., vol. 19, no. 10, pp. 3902–3911, May 2019.

[10] A. Boukerche, Q. Wu, and P. Sun, “Efficient green protocols for sustainable wireless sensor networks,” IEEE Trans. Sustain. Comput., vol. 5, no. 1, pp. 61–80, Jan. 2020.

[11] M. Spuhler, D. Giustiniano, V. Lenders, M. Wilhelm, and J. B. Schmitt, “Detection of reactive jamming in DSSS-based wireless communications,” IEEE Trans. Wireless Commun., vol. 13, no. 3, pp. 1593–1603, Mar. 2014.

[12] A. Ephremides, J. E. Wieselthier, and D. J. Baker, “A design concept for reliable mobile radio networks with frequency hopping signaling,” Proc. IEEE, vol. 75, no. 1, pp. 56–73, Jan. 1987.

[13] P. Gu, C. Hua, W. Xu, R. Khatoon, Y. Wu, and A. Serhrouchni, “Control channel anti-jamming in vehicular networks via cooperative relay beamforming,” IEEE Internet Things J., vol. 7, no. 6, pp. 5064–5077, Jun. 2020.

[14] P.-D. Thanh, H. Vu-Van, and I. Koo, “Efficient channel selection and routing algorithm for multihop, multichannel cognitive radio networks with energy harvesting under jamming attacks,” Secur. Commun. Netw., vol. 2018, pp. 1–12, Mar. 2018.

[15] Z. Feng, G. Ren, J. Chen, X. Zhang, Y. Luo, M. Wang, and Y. Xu, “Power control in relay-assisted anti-jamming systems: A Bayesian three-layer Stackelberg game approach,” IEEE Access, vol. 7, pp. 14623–14636, 2019.

[16] Y. Xu, Y. Xu, G. Ren, J. Chen, C. Yao, L. Jia, D. Liu, and X. Wang, “Play it by ear: Context-aware distributed coordinated anti-jamming channel access,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 5279–5293, 2021.

[17] Y. Z. Jembre and Y.-J. Choi, “Distributed and jamming-resistant channel assignment and routing for multi-hop wireless networks,” IEEE Access, vol. 6, pp. 76402–76415, 2018.

[18] A. Signori, F. Chiarotti, F. Campagnaro, and M. Zorzi, “A game-theoretic and experimental analysis of energy-depleting underwater jamming attacks,” IEEE Internet Things J., vol. 7, no. 10, pp. 9793–9804, Oct. 2020.

[19] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhang, “UV relay in VANETs against smart jamming with reinforcement learning,” IEEE Trans. Veh. Technol., vol. 67, no. 5, pp. 4087–4097, May 2018.

[20] Z. Xiao, B. Gao, S. Liu, and L. Xiao, “Learning based power control for mmWave massive MIMO against jamming,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[21] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhaher, “Reconfigurable intelligent surfaces: Principles and opportunities,” IEEE Commun. Surveys Tuts., vol. 23, no. 3, pp. 1546–1577, 3rd Quart., 2021.

[22] Z. Chu, W. Hao, P. Xiao, and J. Shi, “Intelligent reflecting surface aided multi-antenna secure transmission,” IEEE Wireless Commun. Lett., vol. 9, no. 1, pp. 108–112, Jan. 2020.

[23] L. Dong and H.-M. Wang, “Secure MIMO transmission via intelligent reflecting surface,” IEEE Wireless Commun. Lett., vol. 9, no. 6, pp. 787–790, Jun. 2020.

[24] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, “Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications,” IEEE Commun. Lett., vol. 23, no. 9, pp. 1488–1492, Sep. 2019.

[25] H. Yang, Z. Xiong, D. Niyato, L. Xiao, and Q. Wu, “Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications,” IEEE Trans. Wireless Commun., vol. 20, no. 1, pp. 375–388, Jan. 2021, doi: 10.1109/TWC.2020.3024860.

[26] X. Tang, D. Wang, R. Zhang, Z. Chu, and Z. Han, “Jamming mitigation via aerial reconfigurable intelligent surface: Passive beamforming and deployment optimization,” IEEE Trans. Veh. Technol., vol. 70, no. 6, pp. 6201–6237, Jun. 2021.

[27] H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, H. V. Poor, and M. Tornatore, “Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach,” IEEE Trans. Wireless Commun., vol. 20, no. 3, pp. 1963–1974, Mar. 2021.

[28] Y. Sun, K. An, J. Luo, Y. Zhu, G. Zheng, and S. Chatzinotas, “Intelligent reflecting surface enhanced secure transmission against both jamming and eavesdropping attacks,” IEEE Trans. Veh. Technol., vol. 70, no. 10, pp. 11017–11022, Oct. 2021.
[29] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconfigurable intelligent surfaces for energy efficiency in wireless communication,” IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[30] P. Lee, Z. A. Ei, M. Han, and H.-P. Tan, “Empirical modeling of a solar-powered energy harvesting wireless sensor node for time-slotted operation,” in Proc. IEEE Wireless Commun. Netw. Conf., Mar. 2011, pp. 179–184.

[31] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, and A. Graves, “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.

[32] R. Liu and J. Zou, “The effects of memory replay in reinforcement learning,” in Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton), Oct. 2018, pp. 478–485.

[33] F. Meng, P. Chen, L. Wu, and J. Cheng, “Power allocation in multi-user cellular networks: Deep reinforcement learning approaches,” IEEE Trans. Wireless Commun., vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[34] K. Wang, L. Chen, K. A. Agha, and Q. Liu, “On optimality of myopic policy in opportunistic spectrum access: The case of sensing multiple channels and accessing one channel,” IEEE Wireless Commun. Lett., vol. 1, no. 5, pp. 452–455, Oct. 2012.

PHAM DUY THANH received the B.E. degree in electronics and telecommunications engineering from Ton Duc Thang University, Vietnam, in 2013, the M.S. degree from the College of Engineering, Graduate Institute of Digital Mechatronic Technology, Chinese Culture University, Taiwan, in 2015, and the Ph.D. degree in electrical, electronic and computer engineering from the University of Ulsan, South Korea, in 2021. He is currently working as a Postdoctoral Researcher with the Department of Information and Communication Engineering, Kongju National University, Cheonan, South Korea. His current research interests include reinforcement learning, reconfigurable intelligent surfaces, cognitive radio networks, wireless communication security, and next generation wireless communication systems.

HOANG THI HUONG GIANG received the B.E. degree in electronics and telecommunications engineering from Ton Duc Thang University, Vietnam, in 2013, the M.S. degree from the College of Engineering, Graduate Institute of Digital Mechatronic Technology, Chinese Culture University, Taiwan, in 2015, and the Ph.D. degree in electrical, electronic and computer engineering from the University of Ulsan, South Korea, in 2021. She is currently working as a Research Professor with the Institute of Computer-Information and Communication, Korea University, Seoul, South Korea. Her current research interests include NOMA, RIS, reinforcement learning, and deep learning in wireless communications.

IC-PYO HONG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics engineering from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. From 2000 to 2003, he was with the Information and Communication Division, Samsung Electronics Company, Suwon, South Korea, where he was a Senior Engineer with CDMA Mobile Research. Since 2003, he has been with the Department of Information and Communication Engineering, Kongju National University, Cheonan, South Korea, where he is currently a Professor. In 2006 and 2012, he was a Visiting Scholar with Texas A&M University, College Station, TX, USA, and Syracuse University, Syracuse, NY, USA, respectively. His research interests include intelligent reflecting surface, numerical techniques in electromagnetics, and periodic electromagnetic structures.