IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 23, 2021, accepted March 26, 2021, date of publication March 31, 2021, date of current version April 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069969

Deep Reinforcement Learning Based Dynamic Spectrum Competition in Green Cognitive Virtualized Networks

QUANG VINH DO AND INSOO KOO
Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, South Korea
Corresponding author: Insoo Koo (iskoo@ulsan.ac.kr)

This work was supported in part by the National Research Foundation of Korea through the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2021R1A2B5B01001721.

ABSTRACT This paper examines the optimal spectrum competing strategy for a virtual network operator in cognitive cellular networks with energy-harvesting base stations. In the scenario for this study, multiple cognitive virtual network operators (CVNOs) obtain spectrum resources from a mobile network operator via spectrum sensing and leasing in order to provide data services to their subscribers. Compared to traditional spectrum leasing via long-term contract, spectrum acquired by sensing is usually cheaper but is unreliable due to the stochastic activities of the licensed users. The CVNOs need to determine the optimal sensing and leasing amount to satisfy the needs of subscribers while guaranteeing a low leasing cost. We aim to find an efficient spectrum sensing and leasing scheme for a CVNO in order to maximize its utility in the long run. The problem is first formulated as the framework of a sequential decision process considering the dynamics of users' activities, spectrum prices, and harvested energy. We then develop a deep reinforcement learning algorithm that uses deep neural networks as function approximators so the CVNO can learn the optimal decision policy by interacting with the environment. We analyze the performance of our proposed scheme through extensive simulations. The experiment results show that the proposed mechanism can significantly improve the CVNO's long-term benefit compared to other learning and non-learning methods.

INDEX TERMS Cognitive radio, deep reinforcement learning, energy harvesting, spectrum leasing, spectrum sensing, virtual network.

I. INTRODUCTION

Spectrum resources are becoming more and more scarce due to the tremendous growth in mobile subscribers and wireless communication services. However, most of the spectrum bands allocated via licensing are often under utilized, even in densely populated urban areas, since mobile users do not always require radio resources [1]. To enhance spectrum resource utilization, many methods for dynamic spectrum access have been proposed [2]–[4]. For example, the concepts of cognitive radio [5] and wireless network virtualization (WNV) [6] were introduced to improve spectrum utilization and efficiency. Specifically, cognitive radio technology can help the network operators to tackle spectrum scarcity and inefficient spectrum usage by allowing unlicensed users to utilize spectrum holes in the licensed bands without affecting the primary users. With WNV, a mobile network operator (MNO) dynamically leases temporarily unused spectrum to virtual network operators, which allows the MNO to gain more revenue and improve spectrum utilization.

WNV is a process of abstracting and sharing network infrastructure and radio resources among multiple parties in order to improve resource utilization and reduce operational costs [7]. Therefore, many studies have been conducted to promote WNV in future wireless networks. For example, Nguyen *et al.* [8] developed an optimal trading contract to maximize the total utility at the network operator while satisfying the requirements of service providers. Rawat *et al.* [9] formulated a three-layer game that incorporates the dynamics of wireless infrastructure providers (WIPs), mobile virtual network operators (MVNOs), and internet of thing devices into a sequential decision-making process. The authors derived a unique optimal solution that facilitates a tradeoff between quality of service (QoS) of end users, payoffs of MVNOs, and payoffs of WIPs.

The associate editor coordinating the review of this manuscript and approving it for publication was Hayder Al-Hraishawi.

Competitive spectrum sharing among multiple secondary users in cognitive radio networks can also enhance the utilization of scarce radio spectrum [1]. Our study considers a cognitive virtual network operator (CVNO) that can access licensed spectrum via both spectrum sensing and spectrum leasing. With the cognitive capability, the CVNO can sense the spectrum holes in the licensed spectrum, and then it can decide whether to access those holes or not without violating the current operations of the licensed users. Since the availability of the primary channels depends on the activity of the licensed users, which is not known in advance, the amount of efficient spectrum obtained via sensing is usually uncertain. So far, some research has been conducted to investigate the interactions between the MNO and the CVNOs. For example, Sun *et al.* [10] investigated an oligopoly offloading market, where several MVNOs compete to serve end users using network infrastructures leased from the host MNO. The authors formulated the interactive behaviors of MVNOs and the host MNO as an inventory game, and they proposed two algorithms to achieve the equilibrium. In other work [11], Yi and Cai investigated spectrum sharing with power-constrained multi-radio secondary users (SUs) in cognitive radio networks. In the considered scenario, there exists a primary spectrum owner who runs auctions for leasing her idle channels and multiple SUs bidding for winning the usage of spectrum channels. However, these studies did not include the explicit cost of the primary channels that the CVNOs need to pay when using them.

Furthermore, most work did not consider small-cell networks with energy harvesting for WNV. Recently, small-cell networks have been regarded as one of the key components of future wireless communications to improve spectrum efficiency and energy efficiency. With the increase in the number of small-cell networks, energy-harvesting technology is considered a promising solution to energy conservation in low-power systems [12]. In this paper, we investigate the problem of competitive spectrum leasing in a cognitive virtualized network that is powered by renewable energy. This network consists of one MNO, a set of CVNOs and their subscribed users. The CVNOs compete for spectrum resources owned by the MNO in short-term periods via both spectrum-sensing and spectrum-leasing methods to provide specific services to the users. The cost of accessing spectrum holes in a primary band is usually cheaper than directly leasing the available spectrum from the MNO. However, the spectrum acquired by sensing is unreliable due to the uncertainty of the sensing results. Hence, the CVNOs need to adapt their strategies in terms of requested spectrum sizes in order to provide their users with the best performance while paying the MNO a low leasing cost.

Although there has been some excellent work on spectrum investment and pricing in cognitive virtualized networks [13]–[16], there is little research considering online learning-based approaches (e.g., reinforcement learning [RL] algorithms) for spectrum sensing and leasing in cognitive virtualized networks. For example, Li *et al.* [13] proposed an optimization-based approach to solving a spectrum investment problem for a MVNO in cognitive radio networks. In [14], Yu *et al.* modeled the spectrum leasing and sensing decisions of an MVNO as a non-convex optimization problem, and solved the problem by using a backward induction method. In [15], Wu *et al.* used backward induction to characterize a dynamic game for competitive spectrum acquisition and pricing strategies between two MVNOs. Similarly, Li *et al.* [16] proposed cooperative pricing strategies for MVNOs in order to maximize their profits. Specifically, the authors studied the pricing decisions for MVNOs within the uncertainties of spectrum inventories in both cooperative and non-cooperative situations. Furthermore, to the best of our knowledge, none of these works integrated deep neural networks (DNNs) into RL to solve spectrum leasing problems with large state and action spaces.

In a nutshell, we propose a deep reinforcement learning (DRL)-based method for efficient spectrum competition under the uncertainties of harvested energy, spectrum prices, and users' activities in cognitive virtualized networks. We first model the spectrum leasing problem as a sequential decision-making process and then develop a DRL-based framework to solve the problem. The main contributions of this paper are summarized as follows.

- We introduce WNV into small-cell, cognitive radio networks, and propose a novel spectrum-leasing scheme for a CVNO considering the dynamics of the network environment, such as users' activities, spectrum prices, and harvested energy.
- We model the interactions between the MNO and the CVNOs as a stochastic decision-making process. During this process, the CVNOs compete for spectrum resources by announcing their requested spectrum sizes. We aim to find the optimal decision policy to maximize the utility for a CVNO in the long run.
- We develop an RL-based algorithm (i.e., Q-learning) to solve the formulated problem, based on which the CVNO can learn the optimal decision policy through interactions with the network environment.
- We further use DNNs as function approximators to estimate the Q-values of all decisions, given the network state, which can enhance the efficiency of the proposed method in cases of large state and action spaces.

The rest of this paper is organized as follows. In Section II, we introduce the system model and the formulation of the spectrum-leasing problem. In Section III, we present the proposed deep RL-based solution to the formulated problem. In Section IV, we evaluate the performance of the proposed method with various numerical results. We finally conclude this paper in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

As shown in Figure 1, we consider a cognitive virtualized network with one MNO and a set, $\mathcal{V} = \{1, 2, \ldots, V\}$, of CVNOs that acquire spectrum resources from the MNO to provide

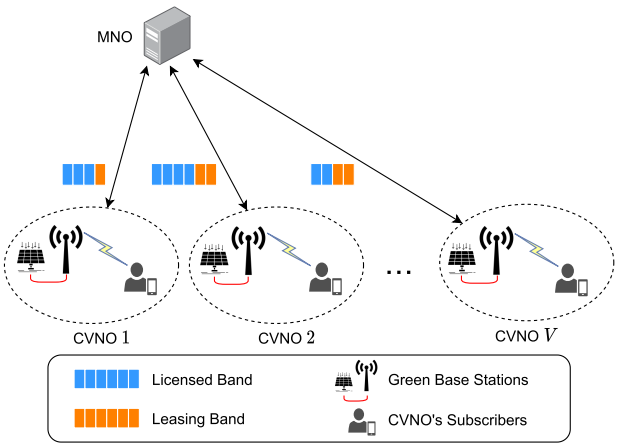

FIGURE 1. The considered model of a cognitive virtualized network with green base stations.

data services to subscribed users. The spectrum resources are divided into the licensed band and the leasing band. The licensed band is primarily reserved for serving licensed users (LUs) of the MNO, and each channel is occupied by an LU. We assume that CVNO $v \in \mathcal{V}$ can access a finite number of channels from this band, denoted as N_v^{li}, through spectrum sensing without explicit communication with the MNO. Meanwhile, the leasing band contains the temporarily unused portions of the spectrum resources that are available for lease, and the CVNO can lease the radio channels from this band by communicating with the MNO. We denote as N^{le} the total number of radio channels in the leasing band, which will be shared among the CVNOs in the network. We assume that the MNO has already reserved a finite number of radio channels from the leasing band for each CVNO according to the contract agreement between the MNO and the CVNO. However, the channels acquired through advance reservation might not be enough for the demands of the users, so the CVNOs might request extra channels. Since the spectrum resources are limited, the CVNOs need to compete for radio channels by announcing their requested spectrum sizes (i.e., the number of channels) to the MNO based on the demands of their users and the prices offered by the MNO.

The system operates in a time-slotted fashion, in which each slot is indexed by notation t and is of equal time duration (in miliseconds). In each time slot, the MNO charges CVNO v for the spectrum of the leasing and the licensed bands at rates of $\pi_v^{le}(t)$ and π_v^{li}, respectively, per unit channel. It is important to note that $\pi_v^{le}(t)$ might change over time according to the demands of the CVNOs while π_v^{li} is a constant, and that the cost of accessing the licensed band is lower than the leasing band, which is upper bounded by π_{max} (i.e., $\pi_v^{li} \leq \pi_v^{le}(t) \leq \pi_{max}$). The benefit that CVNO v gains from utilizing a channel to serve its subscribers is denoted by g_v. Each CVNO owns a green base station (BS) that is equipped with an energy-harvesting device in order to harvest energy from renewable sources (e.g., solar power). The BS stores its harvested energy (packets) in a battery with a finite capacity, E_v^{bat}, and uses these energy packets for data transmissions. We denote as $e_v^h(t)$ the number of energy packets that a BS can harvest in time slot t, which is given as

$$e_v^h(t) \in \{1, 2, \ldots, \xi\} \tag{1}$$

where $0 < \xi \leq E_v^{bat}$, and we assume that $e_v^h(t)$ follows a Poisson point process with mean μ_e.

The activity of an LU is described by a two-state discrete-time Markov chain process, as shown in Figure 2. In a given time slot, an LU might be in one of two states: active (1) or inactive (0) with state-transition probabilities P_{10} and P_{01}. To access the licensed channels, the CVNO needs to perform spectrum sensing. We assume that the CVNO can collect the sensing information from a sensor network and combine the local sensing data using a specific rule (e.g., a soft combination approach [17], [18]) to decide the states of LUs. The sensing performance can be evaluated by the probabilities of detection (P_d) and false alarm (P_f). The former metric refers to the probability that the active state of the LU is detected correctly, whereas the latter metric is the probability that the sensing result indicates the presence of the LU signal on the corresponding channel, when actually there is no signal. For simplicity, we call the CVNOs' subscribers the SUs, when no ambiguity arises.

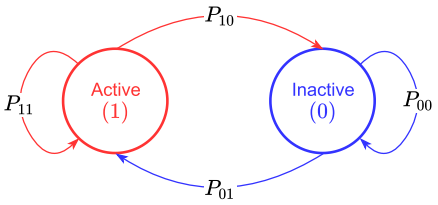

FIGURE 2. The two-state Markov chain process used in this paper to model the activities of the licensed users.

B. PROBLEM FORMULATION

The operation of the system is illustrated in Fig. 3. At the beginning of time step t, the MNO announces the unit prices for radio resources in both bands, π_v^{li} and $\pi_v^{le}(t)$. CVNO v first determines the amount of licensed channels that it is going to sense, $x_v^{li}(t)$. We denote as e^s the energy consumption to perform spectrum sensing on one channel, hence, the total energy consumption for spectrum sensing is $e^s x_v^{li}(t)$. Let $y_v^{li}(t)$ denote the number of channels that the CVNO obtains after spectrum sensing, which can be calculated as

$$y_v^{li}(t) = \lambda x_v^{li}(t) \tag{2}$$

where $\lambda \in [0, 1]$ denotes the fraction of sensed spectrum that is temporarily available for use. The CVNO then announces its leasing amount, $x_v^{le}(t)$, to the MNO by considering the sensing results, the demands of their subscribers, and the resource prices. Since the spectrum resources are limited, the MNO fairly allocates the channels to the CVNOs

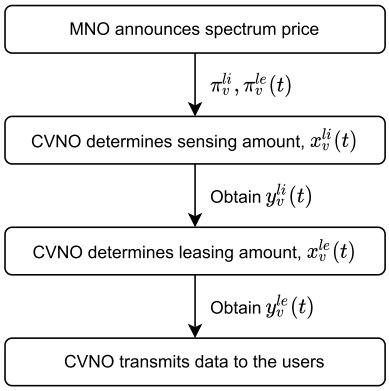

FIGURE 3. The operation of the considered system.

as follows:

$$y_v^{le}(t) = \min\left(x_v^{le}(t), \frac{N^{le}}{\sum_{v\in\mathcal{V}} x_v^{le}(t)} x_v^{le}(t)\right) \tag{3}$$

where $y_v^{le}(t) \leq x_v^{le}(t)$ is the number of channels allocated to CVNO v. Finally, the CVNOs use the obtained channels, $y_v^{li}(t) + y_v^{le}(t)$, to transmit data to the SUs. We denote as e^{tr} the number of energy packets required for transmitting data to the SUs (per unit channel). We can see from Eq. (3) that if we increase $x_v^{le}(t)$, the value of $y_v^{le}(t)$ will increase, which means that if the channel requirements increase, the energy consumption for data transmission might also increase. Therefore, the current energy capacity of the BS will also affect the sensing and leasing decision of the CVNO.

A CVNO needs to satisfy the requirements of its subscribers while guaranteeing a low leasing cost by competing for radio resources using its own sensing and leasing strategy. At a given time step, subscriber m_v of CVNO v might request $b_{m_v}(t) \in \{0, 1, \cdots, b_{max}\}$ wireless channels for their data service, where b_{max} denotes the maximum number of channels that can be used to provide a data service to an SU; and $b_{m_v}(t) = 0$ means that the SU does not require data from the CVNO. Thus, the total demand from the SUs can be calculated as

$$b_v(t) = \sum_{m_v} b_{m_v}(t) \tag{4}$$

We assume that a user (i.e., an SU) can receive better service quality if the CVNO assigns more channels to that user. We aim to develop a learning-based framework for the CVNO in order to maximize its utility in the long run.

We define the utility function for a CVNO, which is the difference between the benefit from, and the total cost of, leasing spectrum as follows:

$$U_v(t) = g_v\left(y_v^{li}(t) + y_v^{le}(t)\right) - \left(\pi_v^{li} y_v^{li}(t) + \pi_v^{le}(t) y_v^{le}(t)\right) \tag{5}$$

In Eq. (5), the first term on the right side is the benefit that the CVNO can obtain by using the allocated channels to serve its users. The last two terms represent the cost of purchasing resources from the MNO.

The CVNO aims to maximize its utility considering the stochastic properties of harvested energy, the LUs' activities, the SUs' demands, and the spectrum prices. The utility maximization problem at CVNO v for time horizon $T > 0$ can be formulated as follows:

$$\begin{aligned}
&\max_{\{x_v^{li}(t), x_v^{le}(t)\}} \sum_{t=0}^{T} \gamma^t U_v(t) \\
&s.t.\ C1:\ (2),\ (3),\ (4),\ (5) \\
&\quad C2:\ x_v^{le}(t) + y_v^{li}(t) \leq b_v(t) \\
&\quad C3:\ x_v^{le}(t) + y_v^{li}(t) \leq \frac{e_v^r(t) - e^s x_v^{li}(t)}{e^{tr}} \\
&\quad C4:\ x_v^{li}(t) \in \left[0, N_v^{li}\right], \quad x_v^{le}(t) \in \left[0, N^{le}\right]
\end{aligned} \tag{6}$$

where $\gamma \leq 1$ is a non-negative coefficient (i.e., a discount rate) that prioritizes immediate utilities over future utilities; $\sum_{t=0}^{T} \gamma^t U_v(t)$ is a cumulative discounted utility from the current time slot to the future. (C2) ensures that the number of wireless channels acquired from sensing and leasing can be fully utilized. (C3) makes sure that the total energy consumption for spectrum sensing and for data transmission using the requested channels can not exceed the current energy level in the battery of the base station, $e_v^r(t)$. To solve this utility maximization problem, we develop a deep RL-based framework so the CVNO can learn the dynamics of the environment, and thus, make better decisions through interactions with the environment.

III. DYNAMIC SPECTRUM COMPETITION WITH DRL

In this section, we present a learning-based method for spectrum sensing and leasing by a CVNO in cognitive virtualized networks, upon which the CVNO can adapt to the variations in the environment during a decision-making process. In this network, the environment dynamics are unknown in advance, and hence, the agent needs to learn the changes in the users' activities, the harvested energy, and the spectrum prices in order to make sensing and leasing decisions. In particular, we employ a DRL algorithm in which DNNs are used as function approximators to estimate the value function to solve the formulated problem. The main purpose of this algorithm is to maximize the utility of the CVNO in the long run.

A. TWO-STEP SEQUENTIAL DECISION PROCESS

By using RL algorithms, the agent can estimate the system dynamics through a sequential decision-making process. During this process, the agent gradually learns how to map the observed state of the environment to a suitable action in order to maximize utility for the CVNO. Among those well-known RL algorithms, Q-learning is the most widely used training technique, and is a model-free approach that can help the agent to learn the optimal policy without having prior information about the environment [19]. Therefore, we employ Q-learning to optimize the sensing and leasing policy for a CVNO in the considered virtualized network.

The problem is reformulated as a two-step sequential decision process, where the CVNO first determines the sensing amount, and subsequently determines the leasing amount. Regarding the learning algorithm, we first define the state space, the action space, and the reward function. The system state obtained by CVNO v at the beginning of time slot t is denoted by

$$s_v(t) = \left\{e_v^r(t), b_v(t), \pi_v^{le}(t)\right\} \tag{7}$$

where $e_v^r(t)$ represents the current energy level of the BS; $\pi_v^{le}(t)$ is the unit price of the leasing channels; $b_v(t)$ is the total demand from the SUs. We employ two Q-learning agents to learn the optimal policy, where each agent takes the system state as input to make sensing and leasing decisions. In the proposed method, agent 1 will make sensing decisions, $x_v^{li}(t)$, after observing the system state. Meanwhile, agent 2 will make leasing decisions, $x_v^{le}(t)$, after observing the system state and the sensing result. Therefore, the action spaces for agent 1 and agent 2 can be defined as follows:

$$\mathcal{A}_1 = \left\{0, 1, \ldots, N_v^{li}\right\} \tag{8}$$

and

$$\mathcal{A}_2 = \left\{0, 1, \ldots, N^{le}\right\} \tag{9}$$

respectively. For the reward function, we use the utility function $U_v(t)$ defined in Eq. (5) to represent the immediate reward that the agents can receive after taking actions. At the end of each time slot, the agents automatically adjust their behaviors based on the returned rewards.

The long-term utility of the CVNO can be estimated by using the Q-value function, denoted by $Q(s, a)$, which represents the expected sum of discounted utilities when the agent is in state s and is taking action a:

$$Q(s, a) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t U(t) | s(0) = s, a(0) = a\right] \tag{10}$$

During the decision process, the algorithm updates the Q-value at each time t using a temporal-difference (TD) update rule, as follows:

$$Q^{(t+1)}(s, a) = Q^{(t)}(s, a) + \eta \delta^{(t)} \tag{11}$$

where $\eta \in (0, 1)$ is the learning rate, and $\delta^{(t)}$ is the TD error, which is the difference between the target value, $U(t) + \gamma \max_{a'} Q^{(t)}(s', a')$, and the current one, $Q^{(t)}(s, a)$, as follows:

$$\delta^{(t)} = U(t) + \gamma \max_{a'} Q^{(t)}(s', a') - Q^{(t)}(s, a) \tag{12}$$

where s' is the next state of the CVNO after taking the action, and a' is the action taken to maximize the Q-value at state s'. The Q-value update process aims at minimizing the TD error, and then gradually converges to the optimal value function, from which the agents can select the optimal decisions at each time step. The algorithm repeats this decision-making process until convergence. With classic RL algorithms, Q-values are simply stored in a lookup table (namely, the Q table), which might be ineffective for large-scale models since iteratively updating the Q-values might take time and effort, and thus, significantly reduce the performance of the algorithm. Therefore, we combine deep learning with Q-learning and exploit DNNs to represent the Q function, referred to as a deep Q-network (DQN) algorithm.

B. DQN-BASED FRAMEWORK

We denote as $Q(s, a; \boldsymbol{\theta_i})$ $\forall i \in \{1, 2\}$ the Q-network that agent i uses to represent the Q-value of state-action pair (s, a), where parameter $\boldsymbol{\theta_i}$ stands for the weight of the DNN. The Q-network is used to map the input states into appropriate actions, and it is composed of three main parts: one input layer, several hidden layers, and one output layer. The input layer stores state $s_v(t)$ using a finite number of cells that is equal to the number of elements of the state. The hidden layers contain finite cells and uses a rectified linear unit function as the activation function to perform a threshold operation on each input element, as follows:

$$f(s) = \max(\boldsymbol{\theta_i} s + \boldsymbol{b}, 0) \tag{13}$$

where $\boldsymbol{b}$ is a bias vector. The output layer uses a linear activation function to produce the estimated Q-values for each action, given state s, and thus, its size is equal to the size of the action space, $|\mathcal{A}_i|$. To improve the performance of the DQN, we also employ two other techniques (i.e., *experience replay* and *fixed target network*) in the design of the DQN algorithm. With the experience replay technique, the agent needs to store any new experience, $\psi_i^{(t)} = \{s, a, U, s'\}_i$, at each training step into a replay memory, $\mathcal{M}_i^{(t)} = \left\{\psi_i^{(0)}, \psi_i^{(1)} \ldots, \psi_i^{(t)}\right\}$, where U and s' are the instant utility and the next system state. The agent then uniformly selects mini-batches from $\mathcal{M}_i$ to train the Q-network. With the fixed target network technique, the agent needs to build a second DNN that has the same structure as the current Q-network (namely, a target network). We denote by $\boldsymbol{\theta_i'}$ the weight of the target network, and we periodically update these parameters during the training process. In the traditional DQN, the Q-network is iteratively optimized to minimize the loss function, which is given as

$$L(\boldsymbol{\theta_i}) = \mathbb{E}_{\mathcal{M}_i}\left[(R - Q(s, a; \boldsymbol{\theta_i}))^2\right] \tag{14}$$

where R is the target value, and is given as

$$R = U + \gamma \max_{a'} Q\left(s', a'; \boldsymbol{\theta_i'}\right) \tag{15}$$

However, the max operation in Eq. (15) might result in overly optimistic value estimates, since it is used for both choosing and evaluating an action. To reduce over-estimation, we employ a double DQN (DDQN) algorithm [20] and rewrite the target value as follows:

$$R = U + \gamma Q\left(s', \arg\max_{a'} Q\left(s', a'; \boldsymbol{\theta_i'}\right); \boldsymbol{\theta_i}\right) \tag{16}$$

With this approach, actions are chosen based on the online Q-network, $\boldsymbol{\theta_i}$, while both the Q-network and the target network are used to evaluate the values of the chosen actions.

Weight $\boldsymbol{\theta_i}$ is gradually optimized by using a stochastic gradient descent with back propagation algorithm, as follows:

$$\Delta\boldsymbol{\theta_i} = \eta\delta\nabla_{\boldsymbol{\theta}_i} Q(s, a; \boldsymbol{\theta_i}) \tag{17}$$

where the TD error, δ, is rewritten as follows:

$$\delta = U + \gamma Q\left(s', \arg\max_{a'} Q(s', a'; \boldsymbol{\theta_i}); \boldsymbol{\theta_i'}\right) - Q(s, a; \boldsymbol{\theta_i}) \tag{18}$$

And the target network parameters, $\boldsymbol{\theta_i'}$, are updated after every finite time steps. The proposed DDQN framework is illustrated in Figure 4.

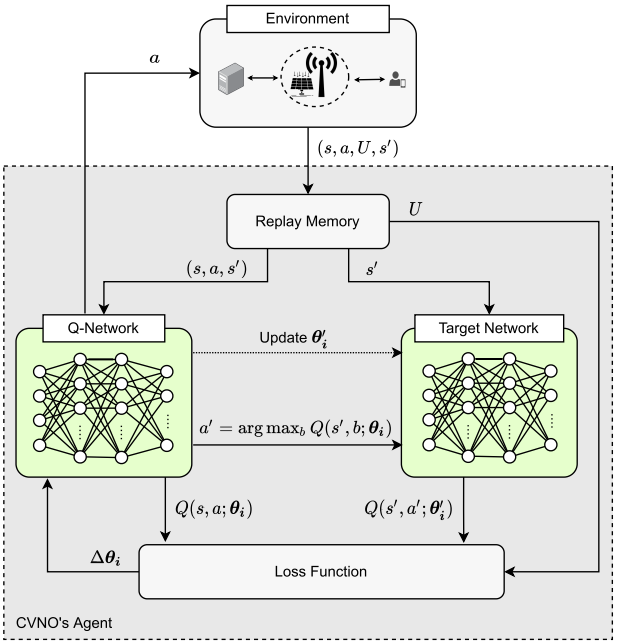

FIGURE 4. A flowchart of the proposed DDQN framework.

C. ACTION SELECTION PROCEDURE

During the sequential decision process, the agents interact with the environment by executing actions according to an ϵ-greedy policy, where $\epsilon \in (0, 1)$ is the exploration rate. It is important to note that the chosen actions must satisfy the constraints in Eq. (6). We denote as $a_v^i(t)$ the action that agent i can select at time t. In this paper, agent 1 makes sensing decisions after observing the system state. Hence, we can derive from (C3) and (C4) that $a_v^1(t)$ can be limited in $\left[0, a_{max}^1(t)\right]$, where $a_{max}^1(t) = \min\left(\frac{e_v^r(t)}{e^s + e^{tr}}, N_v^{li}\right)$. Similarly, agent 2 makes leasing decisions after observing the sensing result and the system state, thus action $a_v^2(t)$ can be limited in $[0, a_{max}^2(t)]$, where $a_{max}^2(t) = \min\left(\min\left(b_v(t), \frac{e_v^r(t) - e^s x_v^{li}(t)}{e^{tr}}\right) - y_v^{li}(t), N^{le}\right)$.

Let $\mathcal{A}_i^f(t) \subset \mathcal{A}_i$ denote the feasible action space for agent i at time t, which can be given as

$$A_i^f(t) = \left[0, 1, \ldots, a_{max}^i(t)\right] \ \forall i \in \{1, 2\} \tag{19}$$

At a given time step, an agent can either choose a random action in the feasible action space with probability ϵ or choose an action that maximizes the Q-value of the state-action pair in the current time slot with probability $1 - \epsilon$, as follows:

$$a_v^i(t) = \arg\max_{a \in \mathcal{A}_i^f(t)} Q(s_v(t), a; \boldsymbol{\theta_i}) \tag{20}$$

The overall learning process using DDQN algorithm is described in Algorithm 1.

Algorithm 1 Learning Procedure With DDQN

Initialize $\boldsymbol{\theta_i}$, $\boldsymbol{\theta_i'}$, and $\mathcal{M}_i$ $\forall i \in \{1, 2\}$.
for episode $ep = 1, 2, \ldots$ **do**
  for step $t = 0, 1, \ldots, T$ **do**
    Agent 1 observes state $s_v(t)$ and specifies $\mathcal{A}_1^f(t)$, then executes action $a_v^1(t)$.
    Agent 2 observes state $s_v(t)$, sensing result and specifies $\mathcal{A}_2^f(t)$, then executes action $a_v^2(t)$.
    Obtain reward $U_v(t)$ and observe state $s_v(t+1)$.
    for Agent $i = 1, 2$ **do**
      Store tuple $\{s_v(t), a_v^i(t), U_v(t), s_v(t+1)\}$ in $\mathcal{M}_i$.
      Select K random tuples from the memory as the training samples.
      for $k = 1, 2, \ldots, K$ **do**
        if $t < T$ **then**
          Compute target value by using Eq. (16)
        else
          $R = U_v(t)$
        end if
        Compute TD error by using Eq. (18).
      end for
      Update $\boldsymbol{\theta_i}$ by using Eq. (17).
    end for
  end for
  Replace $\boldsymbol{\theta_i'}$ with $\boldsymbol{\theta_i}$.
end for

IV. PERFORMANCE ANALYSIS

In this section, we provide simulation results for the proposed spectrum sensing and leasing scheme in a cognitive virtualized network under different system configurations. The simulations were implemented by using Python-integrated software with Keras and TensorFlow deep learning libraries (Python 3.7, Anaconda 2020 distribution, The Anaconda Inc., Austin, Texas, USA, 2020).

A. SIMULATION PARAMETERS

We conducted simulations with the following parameters. The number of CVNOs in the network is $V = 3$, and each CVNO provides data services to 5 subscribers. There are 30 orthogonal channels in total, in which $N_v^{li} = 15$ and $N^{le} = 15$. The state transition probabilities of each LU in the discrete-time Markov process were set to $P_{10} = P_{01} = 0.2$. We set the value of the desired probability of detection at

$P_d = 0.9$, and the probability of false alarm at $P_f = 0.1$. The channel requirements of the SUs were uniformly generated from the set $\{0, 1, 2, 3\}$, which means the SU might request channels from this set with the same probability. We assume the MNO sets prices for the resources in the current time slot based on the demands of the CVNOs in the previous time slot, which can be defined by using elastic pricing functions as follows [21]:

$$\pi_v^{le}(t+1) = \alpha + \frac{\beta}{N^{le}} \left(\sum_{v \in \mathcal{V}} y_v^{le}(t) \right)^{\tau} \tag{21}$$

where α, β, and τ are positive coefficients, and $\tau \geq 1$. This equation indicates that the MNO intends to set a higher price for a unit wireless channel when the demand increases. For the pricing function of the leasing band, we use $(\alpha, \beta, \tau) = (1, 0.5, 1.5)$. The price range for the leasing band was set to $1 \leq \pi_v^{le}(t) \leq 3$. The MNO charges the CVNOs much lower price for the channels in the licensed band than those in the leasing band, that is $\pi_v^{li} = 1$. The benefit for serving the SUs by using one channel was set at $g_v = 4$. We assume that each CVNO does not know the pricing strategy of the MNO, or the leasing strategies of other CVNOs in the network. Regarding energy harvesting, the average harvested energy at a base station was set at $\mu_e = 5$ energy packets. The total energy capacity of a base station is $E_v^{bat} = 16$ energy packets. Energy consumption for spectrum sensing and data transmission were set at $e^s = 1$ and $e^{tr} = 2$, respectively. As for the proposed DDQN algorithm, the Q-network includes two hidden layers, each of which contains 100 cells. The discount factor was set to $\gamma = 0.99$, and the learning rate of the algorithm was set to $\eta = 0.01$. We set $\epsilon = 1$ at the beginning of the training process and gradually decreased it to 0 at a rate of 0.01 per time slot. We trained the network over 100 episodes of $T = 2000$ time steps each. In addition, the convergence condition was defined as $|U_{ep} - U_{ep-1}| < 0.005$, where U_{ep} denotes the average utility that the CVNO obtains at episode ep. All the results were obtained by averaging over a large number of independent runs.

We compared the performance of the proposed spectrum sensing and leasing scheme with the following schemes:

- Single-agent scheme: the CVNO employs only one DRL agent to learn the optimal policy. The agent will make sensing and leasing decisions simultaneously at the beginning of each time slot after observing the system state.
- Myopic scheme: the CVNO aims to maximize its utility in the current time slot. This scheme is equivalent to our scheme when the discount factor is set to zero (i.e., $\gamma = 0$).
- Random scheme: the CVNO makes decisions about the sensing amount and the leasing amount randomly.

In the considered system, the first CVNO might employ the proposed DDQN or the single-agent scheme algorithm, whereas the second and the third CVNOs use the myopic and the random schemes, respectively.

B. SIMULATION RESULTS

We first verify the convergence performance of the proposed algorithm during the training process, as shown in Figure 5. We can observe from the figure that the average utility for the CVNO using the learning method increases with the increment in the number of training episodes, and then gradually converges at the 80th episodes. We can also see that the proposed method can converge to the optimal policy faster than the single-agent algorithm. This is because by using the proposed method, the CVNO makes leasing decisions after knowing the available amount of sensed spectrum. This can guarantee that the CVNO will not lease too many expensive channels from the leasing band and thus can improve its utility. Meanwhile, utility for the CVNOs using non-learning methods was unchanged when the number of episodes increased. Furthermore, the DDQN agent can provide the CVNO with the best performance, since it can learn the environment's dynamics during the training process.

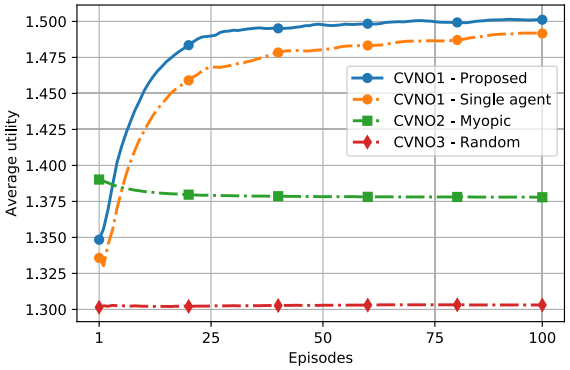

FIGURE 5. Average utility from the different schemes as a function of the number of episodes.

Figure 6 shows the impact of harvested energy on the average utility for the CVNOs in the network. The average harvested energy at each BS varied from 2 to 7 (energy packets). As depicted from the figure, the average utility obtained by the CVNOs increased significantly with growth in the number of energy packets that the BSs harvested from the

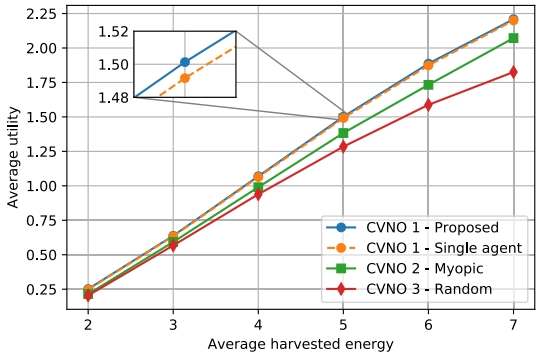

FIGURE 6. Average utility according to the average amount of harvested energy.

environment in each time slot. The reason is that the BSs can store more energy in their batteries, so the CVNOs can lease more resources from the MNO to serve their subscribers. Thus, the CVNOs can earn more revenue, since they can sell more resources to users. Moreover, the DDQN agent can learn about the arrival of harvested energy through interaction with the environment. Therefore, it can choose the appropriate action in each time slot in order to maximize the utility for the corresponding CVNO. As a consequence, the CVNO that uses the proposed algorithm for spectrum sensing and leasing can achieve the best performance.

Figure 7 shows the effect of the battery capacity of the BSs on the performance of the proposed method in terms of average utility. We can observe from the figure that a larger battery capacity allows a BS to store more harvested energy for further use. Hence, the CVNOs can request more resources from the MNO in order to provide more services to users, which results in an increase in the utility for each CVNO. Compared with the myopic and the random schemes, the utility achieved by the first CVNO dominates the other schemes. To explain this, the CVNO with the myopic scheme aims to maximize its utility in the current time slot by requesting as many resources as possible. However, due to the limitation on harvested energy, this kind of action might cause the corresponding BS to lack energy for future use, which leads to lower utility. The CVNO with the random scheme will request random spectrum sizes from the MNO based on the remaining energy, the demands of its users, and the available spectrum announced by the MNO. However, traffic is time-varying, and the harvested energy and system bandwidth have a stochastic feature, so the CVNO might not request enough resources to serve the users in the current time slot. Therefore, the final utility is low.

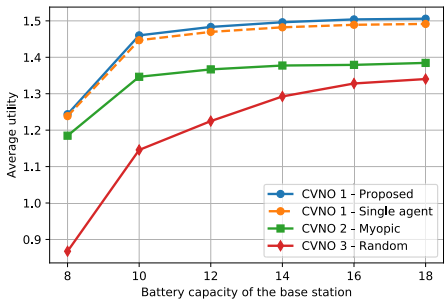

FIGURE 7. Average utility based on battery capacity in each BS.

We further examined the effect of the number of licensed channels on the performance of the proposed algorithm. In this case, the number of licensed channels was set at $N_v^{li} \in [8, 16]$, as shown in Figure 8. We can see from the figure that growth in the number of licensed channels in the system can provide the CVNOs with better performance. In particular, utility for the learning CVNO rises quickly as N^{li} increases. Meanwhile, utility for those using non-learning methods also increases, but at a very low rate. The reason is

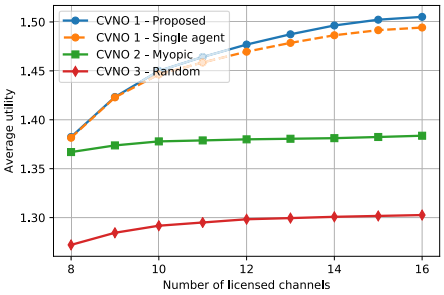

FIGURE 8. Average utility based on the number of licensed channels.

that the DDQN agent can learn the activities of the LUs during the training process, and thus, it can guide the CVNO to select better actions after interacting with the environment. As a consequence, the proposed learning algorithm can provide the CVNO with the best performance.

Figures 9 and 10 present the average utility for each CVNO under the effect of false alarm probability P_f and correct detection probability P_d. In this scenario, we change the values of P_f and P_d to verify the performance of the resource-leasing methods. As can be seen from Fig. 9, there are significant decreases in utility when the probability of false alarm rises from 0.1 to 0.5. Since the CVNOs select actions based partly on the sensing results, a high false alarm rate might cause CVNOs to select bad actions, which thus results in lower utility. For example, the CVNOs might take a risk and lease busy channels, which means the CVNOs do not have enough resources for data transmissions. As a result, the revenue gained from serving the users will be reduced

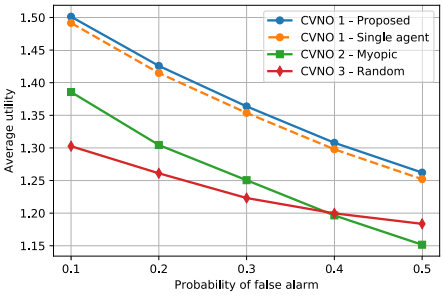

FIGURE 9. Average utility based on the probability of false alarm.

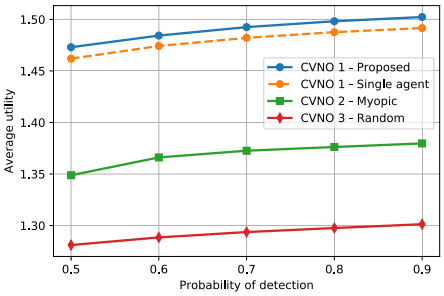

FIGURE 10. Average utility based on the probability of correct detection.

significantly. Conversely, an increment in the probability of correct detection can provide better utility to all CVNOs in the network. This is because the MNO can allocate more unused channels from the licensed band to the CVNOs for their services. Therefore, the CVNOs might gain more revenue by transmitting data to their subscribers. Furthermore, with higher detection probability, the CVNO does not need to request more channels from the leasing band, which might cause the resource price to increase quickly, and hence, profit will increase.

V. CONCLUSION

In this paper, we investigate the spectrum competition problem in cognitive virtualized networks where multiple CVNOs lease spectrum resources from an MNO in order to provide data services to their subscribers. The CVNOs compete for limited spectrum resources by announcing the values of the spectrum sizes they are going to lease from the MNO. The problem is formulated as a sequential decision-making process, during which the CVNOs make their sensing and leasing decisions despite uncertainties in the users' activities, in the amount of harvested energy, and in the resource prices. We propose a DDQN-based framework for efficient spectrum sensing and leasing so a CVNO can maximize its long-term utility. With this method, neural networks are used as function approximators to estimate the value functions, which is useful for solving large-scale problems. By using our proposed approach, the CVNO can learn the optimal decision policy through interaction with the network environment without knowing the system's dynamics in advance. The simulation results show that our proposed method outperforms the other strategies in terms of average utility.

REFERENCES

[1] F. Hu, B. Chen, and K. Zhu, "Full spectrum sharing in cognitive radio networks toward 5G: A survey," *IEEE Access*, vol. 6, pp. 15754–15776, 2018.

[2] X. Liu, M. Jia, X. Zhang, and W. Lu, "A novel multichannel Internet of Things based on dynamic spectrum sharing in 5G communication," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5962–5970, Aug. 2019.

[3] S. K. Sharma, T. E. Bogale, L. B. Le, S. Chatzinotas, X. Wang, and B. Ottersten, "Dynamic spectrum sharing in 5G wireless networks with full-duplex technology: Recent advances and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 674–707, 1st Quart., 2018.

[4] C. Xin, P. Paul, M. Song, and Q. Gu, "On dynamic spectrum allocation in geo-location spectrum sharing systems," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 923–933, Apr. 2019.

[5] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[6] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2015.

[7] N. Zhang, P. Yang, S. Zhang, D. Chen, W. Zhuang, B. Liang, and X. S. Shen, "Software defined networking enabled wireless network virtualization: Challenges and solutions," *IEEE Netw.*, vol. 31, no. 5, pp. 42–49, May 2017.

[8] D. H. N. Nguyen, Y. Zhang, and Z. Han, "A contract-theoretic approach to spectrum resource allocation in wireless virtualization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[9] D. B. Rawat, A. Alshaikhi, A. Alshammari, C. Bajracharya, and M. Song, "Payoff optimization through wireless network virtualization for IoT applications: A three layer game approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2797–2805, Apr. 2019.

[10] F. Sun, F. Hou, H. Zhou, B. Liu, J. Chen, and L. Gui, "Equilibriums in the mobile-virtual-network-operator-oriented data offloading," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1622–1634, Feb. 2018.

[11] C. Yi and J. Cai, "Ascending-price progressive spectrum auction for cognitive radio networks with power-constrained multiradio secondary users," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 781–794, Jan. 2018.

[12] 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.

[13] S. Li, J. Huang, and S.-Y.-R. Li, "Dynamic profit maximization of cognitive mobile virtual network operator," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 526–540, Mar. 2014.

[14] J. Yu, M. H. Cheung, and J. Huang, "Spectrum investment under uncertainty: A behavioral economics perspective," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2667–2677, Oct. 2016.

[15] C. Wu, R. Wang, P. Wang, Y. Cao, L. Liu, K. Zhu, and B. Chen, "On the profit maximization of spectrum investment under uncertainties in cognitive radio networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[16] C. Li, J. Li, Y. Li, and Z. Han, "Pricing game with complete or incomplete information about spectrum inventories for mobile virtual network operators," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11118–11131, Nov. 2019.

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

[18] W. Han, J. Li, Z. Li, J. Si, and Y. Zhang, "Efficient soft decision fusion rule in cooperative spectrum sensing," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 1931–1943, Apr. 2013.

[19] H.-H. Chang, H. Song, Y. Yi, J. Zhang, H. He, and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938–1948, Apr. 2019.

[20] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with Double Q-learning," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.

[21] D. Niyato and E. Hossain, "Competitive spectrum sharing in cognitive radio networks: A dynamic game approach," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2651–2660, Jul. 2008.

QUANG VINH DO received the B.E. degree in electrical and electronic engineering from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2009, the M.E. degree in electronic and computer engineering from RMIT University, Melbourne, Australia, in 2012, and the Ph.D. degree in electrical and computer engineering from the University of Ulsan, Ulsan, South Korea, in 2020.

Since March 2021, he has been a Postdoctoral Researcher with Pusan National University, South Korea. His research interests include optimization theory, deep learning, and deep reinforcement learning for resource management in wireless communications.

INSOO KOO received the B.E. degree from Konkuk University, Seoul, South Korea, in 1996, and the M.S. and Ph.D. degrees from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 1998 and 2002, respectively.

From 2002 to 2004, he was with Ultrafast Fiber-Optic Networks Research Center, GIST, as a Research Professor. In 2003, he was a Visiting Scholar with the KTH Royal Institute of Technology, Stockholm, Sweden. In 2005, he joined the University of Ulsan, Ulsan, South Korea, where he is currently a Full Professor. His current research interests include next-generation wireless communication systems and wireless sensor networks.