Wireless Networks (2021) 27:1319–1334
https://doi.org/10.1007/s11276-020-02520-y

Uplink NOMA-based long-term throughput maximization scheme for cognitive radio networks: an actor–critic reinforcement learning approach

Hoang Thi Huong Giang[1] · Tran Nhut Khai Hoan[2] · Insoo Koo[1]

Accepted: 8 December 2020 / Published online: 2 January 2021

Abstract
Non-orthogonal multiple access (NOMA) is one of the promising techniques for spectrum efficiency in wireless networks. In this paper, we consider an uplink NOMA cognitive system, where the secondary users (SUs) can jointly transmit data to the cognitive base station (CBS) over the same spectrum resources. Thereafter, successive interference cancellation is applied at the CBS to retrieve signals transmitted by the SUs. In addition, the energy-constrained problem in wireless networks is taken into account. Therefore, we assume that the SUs are powered by a wireless energy harvester to prolong their operations; meanwhile, the CBS is equipped with a traditional electrical supply. Herein, we propose an actor–critic reinforcement learning approach to maximize the long-term throughput of the cognitive network. In particular, by interacting and learning directly from the environment over several time slots, the CBS can optimally assign the amount of transmission energy for each SU according to the remaining energy of the SUs and the availability of the primary channel. As a consequence, the simulation results verify that the proposed scheme outperforms other conventional approaches (such as Myopic NOMA and OMA), so the system reward is always maximized in the current time slot, in terms of overall throughput and energy efficiency.

Keywords Cognitive radio network · NOMA · Energy harvesting · Actor–critic

1 Introduction

Spectrum scarcity is one of the critical issues in fifth-generation (5G) communications systems and for future wireless networks, because the lack of accessible spectrum is hindering the application of novel communications technologies [1–3]. However, in [4], the authors revealed that the licensed spectrum remains under utilized. In order to deal with spectrum inefficiency, the dynamic spectrum access techniques are studied, with cognitive radio (CR), known as the key enabling technology [5]. In a CR network, the unlicensed secondary users (SUs) can access and utilize the unused spectrum of the licensed primary users (PUs) [6–8].

Nowadays, the CR network paradigm can broadly be categorized into three main models [9–11]: underlay, overlay, and interweave. In the underlay CR model, the SUs can perform their operations if and only if the interference caused by all SUs is lower than a given threshold. In the overlay CR model, the SUs assist as relays for the PUs, and jointly transmit their signals using a portion of the licensed spectrum. In the interweave CR model, the SUs can only transmit when the primary channel is not occupied by any PU. With this model, vacant spectrum is temporarily available over certain time instants, such that an SU can opportunistically transmit data when the PU is inactive. In order to reduce collisions with the PUs and ensure energy-efficient utilization, the SUs sense the

✉ Insoo Koo
iskoo@ulsan.ac.kr

Hoang Thi Huong Giang
huonggiangtdt@gmail.com

Tran Nhut Khai Hoan
tnkhoan@ctu.edu.vn

[1] School of Electrical Engineering, University of Ulsan (UOU), Ulsan, Republic of Korea

[2] Can Tho University, Can Tho, Vietnam

surrounding spectrum to verify the availability of the primary channel in order to transmit their data.

As another potential technique for the next generation of wireless networks, non-orthogonal multiple access (NOMA) has lately gotten noticeable attention, enabling multiple users to simultaneously access the spectrum, and it has become an important fundamental to designing radio access techniques for future wireless networks [12–15]. The key with NOMA is allowing multiple users to access the same spectrum resource block together, with the objective being spectral efficiency. NOMA is generally classified into two major approaches: power-domain NOMA [16–18] and code-domain NOMA [19–22]. In power-domain NOMA, different power levels are used to jointly serve multiple users at the same time using the channel frequency under different channel conditions. At the receiver, the signals of the different transmitters are superposed and then decoded via successive interference cancellation (SIC).

Moreover, by introducing the two aforementioned concepts, NOMA can be combined with a CR network in order to improve spectral efficiency. Liu et al. [23] proposed a stochastic geometry model for a large-scale CR network in order to depict the outage performance from the paradigm of integrated NOMA and CR. In [24], spectrum efficiency was enhanced by developing a NOMA-based secure transmission scheme in CR networks. A cooperative NOMA spectrum-sharing network over the Nakagami fading channel was investigated in [25]. Besides, multicast NOMA is also adopted in 5G systems in terms of user scheduling in order to improve network performance [26]. It has been pointed out that higher spectral efficiency can be promised by combining NOMA with CR networks.

In recent years, prolonging the long-term operation of the network is also one of the nearly essential purposes of wireless systems [27]. Using renewable energy sources for wireless users is considered a potential solution for dealing with the energy constraints of wireless devices. In particular, energy for the SUs can be harvested from natural ambient sources (solar [28, 29], wind [30, 31], radio frequency [32], etc.). Hence, the battery of a wireless user can replenish itself without manual recharging. Nevertheless, the harvested energy is still restricted to users. That means finding a way for SUs to effectively utilize the harvested energy needs to be carefully investigated. For that reason, Celik et al. [33] proposed hybrid energy harvesting in a heterogeneous CR network to enhance spectrum efficiency while reducing the energy consumption of the system.

Furthermore, a dynamic power allocation algorithm is carefully being investigated owing to the significant role of power allocation for wireless users in uplink NOMA (i.e. the effect of power allocation on the rate of each user) [34–36]. In [34], the authors studied both power control and beamforming methods in order to maximize the sum rate of the system for millimeter-wave communications. The joint optimization problem for sum-throughput maximization under transmission power constraints, the minimum rate requirements of users, and SIC constraints were formulated in [35] for both uplink and downlink NOMA in a cellular system. In [36], the authors took into account channel assignment and power control to maximize the sum rate for a NOMA-based uplink network. They mathematically derived a more tractable form of the formulated problems as a maximum weighted independent set issue, and then used graph theory to deal with them.

1.1 Related works

For cognitive-based resource control management, the optimization problems are ordinarily formulated as a Markov decision process (MDP) [37–40]. The optimal solution can be obtained by using a value iteration-based dynamic programming method [37]. However, this work focused on joint channel and transmission mode allocation for SUs on multiple non-overlapping channels where each secondary user was assigned to use a channel for interference avoidance. Furthermore, the scheme in [37] requires prior knowledge of energy arrival distribution, which is challenging to achieve in practical wireless networks. To combat with this issue, the authors in [38] provided a model-free reinforcement learning approach such that the agent is able to learn the optimal policy by smartly interacting with dynamic environment. Nevertheless, allowing multiple access without using successive interference cancellation at receiving side can significantly degrade overall system throughput due to severe interference among wireless devices. With the purpose of providing approximating mechanism (e.g. Q-function, action function, and value-state function) to let the system work well in the large scale scenarios, the deep reinforcement learning methods were proposed [39, 40]. Specifically, the authors in [39] considered the underlay cognitive networks where a secondary transmitter and a primary transmitter can simultaneously transmit their data on the licensed channel, and the deep Q learning was proposed for the secondary user to intelligently adjust its transmission power such that data of both secondary and primary transmitters can be successfully transmitted under their own quality-of-service requirements (QoS). Meanwhile, Yang et al. [40] developed an actor–critic deep reinforcement learning to efficiently solve the intelligent transmission scheduling problem in cognitive internet-of-things systems. Nonetheless, the sensing error in terms of licensed user activities at devices was not taken into account. The summary of the introduced solutions in cognitive radio systems can be shown in the Table 1.

Table 1 Table of literature summary in cognitive radio networks

References	Frameworks	Advances/key contributions	Loopholes/drawbacks
[37] Maximized the secrecy rate of the secondary system under the presence of passive eavesdroppers	POMDP-based	Optimal policy can be obtained so far	Requires prior knowledge of environment dynamics and high complexity overhead
[38] Optimized spectrum sensing accuracy and data transmission performance under energy constraints	Reinforcement learning	Nearly optimal policy is achieved by directly interacting from the environment	Limited to large-scale systems and low convergence rate in complex system
[39] Enhanced resource utilization efficiency to overcome the random variations of the received signal strength measurements	Deep Q learning	Action-value function of Q-learning method can be approximated to deal with high dimensional system state	Received signal strength accuracy is not considered
[40] Scheduled the intelligent transmissions in cognitive Internet of Things under high-dimensional variables to improve the system throughput	Deep actor critic learning based on fuzzy normalized radial basis function neural network	Action function of the actor and value function of the critic are approximated to deal with high dimensional system state	Spectrum sensing imperfection is not considered

In addition, reinforcement learning schemes were also implemented in NOMA systems [41, 42]. More particularly, the problem of resource allocation in an uplink NOMA system was investigated to maximize the long-term energy efficiency [41]. On the other hand, the authors in [42] developed the deep reinforcement learning algorithm to reduce the collisions resulting from the uncoordinated and non-orthogonal spectrum access and further to enhance long-term throughput of grant-free NOMA system. Both of these works were designed for licensed frequency band systems where energy harvesting for users was not considered. Besides, the cooperative communication system was considered in which the secondary base station can forward primary information to far primary users and transmit data to secondary users in the underlay mode [43]. In order to explore the optimal sensing policy, Zhong et al. [44] proposed a dynamic multichannel access approach, and the framework was corroborated in various channel switching patterns and probabilities. The authors in [45] investigated the algorithm to optimize the system throughput of the smart grid backscatter communication networks in which the secondary users can select an action, such as energy harvesting, backscatter, or active transmission in each time slot.

1.2 Main contributions and novelty

Our aforementioned survey showed that the application of the NOMA technique to energy harvesting-powered CRN is still an open issue that needs to be thoroughly investigated. Moreover, the sensing error is one of the critical problems in cognitive radio networks, which is not carefully investigated in the literature. Motivated by the foregoing analysis, in this work, we consider a transmission power allocation problem for solar-powered uplink CR networks in the sensing-imperfect scenario. Different from [38] where the interference cancellation among users using the same channel was not investigated, and from [37, 45] that mainly focused on channel assignment using the OMA technique in which each user can be assigned on an orthogonal frequency band to avoid the interference among users, we study an uplink actor–critic learning-based transmission power allocation scheme that allows multiple SUs access on the same channel by adopting NOMA in order to maximize the long-term throughput of the network. Herein, the SUs employ the NOMA technique to simultaneously transmit their information to the cognitive base station (CBS), and then the CBS can exploit SIC to decode the information. The key contributions of this paper can be outlined as follows.

- We consider a CR network with uplink NOMA, where the SUs are allowed to concurrently access the same primary channel when it is not used by the PU. Specifically, by adopting NOMA, the SUs can transmit data on the same channel and in the same time slot when the sensing result indicates the primary channel is free. However, the SUs are equipped with a limited-capacity battery. Therefore, solar energy harvesting is executed by the SUs such that they can externally harvest energy to replenish the battery for use in long-term data transmission. In addition, the energy-constrained problem and sensing error issues of the SUs are also taken into account.
- To do this, we first formulate the problem of throughput maximization based on a Markov decision process (MDP). Afterward, the actor–critic reinforcement learning approach is adopted such that the CBS can adaptively interact with the environment and dynamically assign the optimal amount of energy for each user

in every time slot without prior information about the harvested energy model of the SUs, which is normally needed by some kinds of partially observable MDP schemes.
- Simulation results show that the proposed scheme outperforms other conventional schemes under various network parameters in terms of average throughput and energy efficiency.

The rest of the paper is structured as follows. The system model is outlined in Sect. 2. The proposed power allocation algorithm is discussed in Sect. 3, and simulation results are provided in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 System model

2.1 Network model

In this paper, uplink NOMA in a cognitive radio network (CRN) is considered, as shown in Fig. 1, which comprises a CBS, a pair of PU transceivers and a set of SUs denoted by $\mathbb{N} = \{SU_1, SU_2, \ldots, SU_N\}$. Although the PUs have priority to use the licensed spectrum, the SUs are allowed to simultaneously and opportunistically access the licensed spectrum of the PU when the sensing result indicates that the primary channel is free. In the network, the CBS and SUs are equipped with a single antenna to receive and transmit signals in a time slot on the currently free primary channel. In particular, at the beginning of a time slot, the SUs will share the primary channel to concurrently transmit data to the CBS if the sensing result for the primary channel is free, and then, the CBS will decode all data sent from the SUs by using the SIC technique. In this paper, the SUs are equipped with a finite battery capacity E_{ca}, and they can replenish energy by themselves using an integrated solar energy harvester.

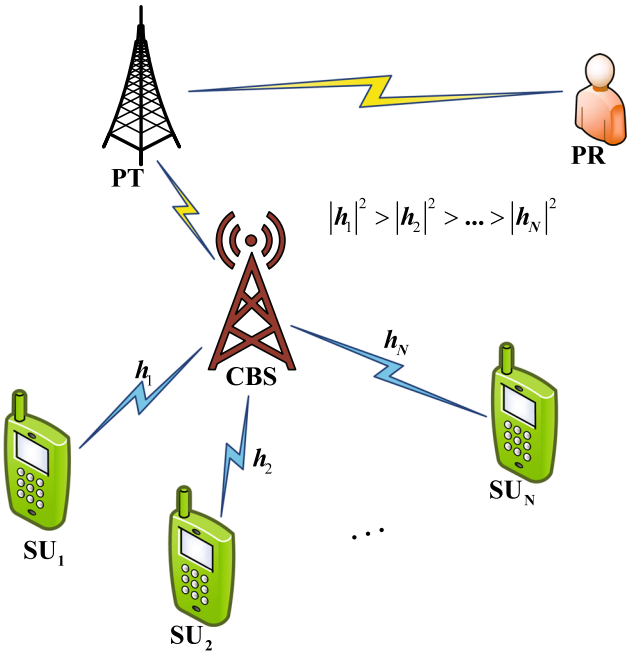


Fig. 1 System model of the proposed scheme

The operation of the considered network consists of three phases: the sensing and decision phase, the data transmission phase, and the energy information update phase, as shown in Fig. 2. In the first phase, with duration τ_{ss}, the SUs perform their individual sensing, and then, they report their local decisions to the CBS; afterward, the CBS gives its global sensing decision (about the state of the primary channel) and its global action decisions on the actions assigned to all SUs. The second phase, with duration τ_{tr}, is the time for the SUs to transmit their data to the CBS. In the last phase, the SUs will send their remaining information to the CBS. Herein, it is assumed that the SUs always have information available to transmit. Furthermore, each transmission session may last several time frames, until all the information is successfully transmitted.

In this paper, we adopt cooperative spectrum sensing for the SUs in the network. More specifically, the SUs perform spectrum sensing at the beginning of a time slot to check whether the licensed spectrum is occupied by the PU or not. There are several major sensing approaches, such as matched filtering, energy detection, and the cyclo-stationary method [46]. Energy detection is one of the most effective methods due to its low computational complexity [47–49]. All sensing results from the SUs are then gathered

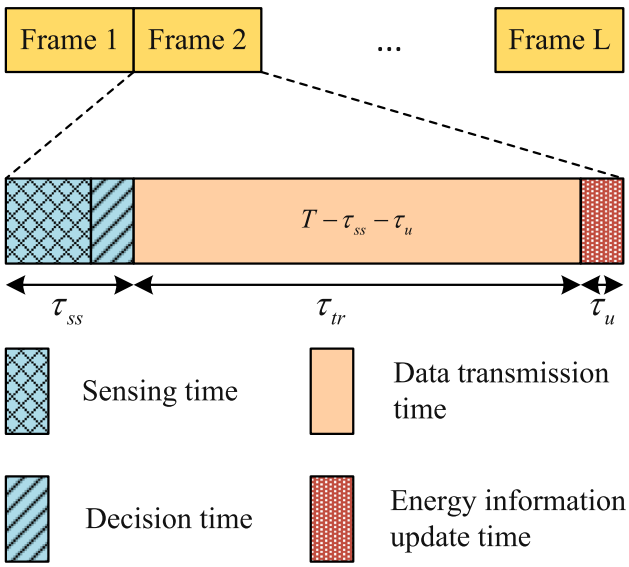


Fig. 2 Time frame of the three phases in the secondary users's operations

and sent to the CBS. After that, the CBS makes a global sensing decision about the activity or silence of the PU on the primary channel, and then decides whether the SUs should transmit data to the CBS or stay silent. Normally, the global sensing decision is done by following a combination rule at the CBS [50–53]. However, in this paper, we do not focus on cooperative spectrum sensing, which has been widely investigated in the literature. Thus, we mainly study a power allocation algorithm for the SUs in order to efficiently use energy to transmit data to the CBS.

When the CBS determines that the PU is absent in the current time slot, all SUs can concurrently transmit their signals to the CBS. The received signal at the CBS is given as follows [34]

$$y(t) = h_1 x_1(t) + h_2 x_2(t) + \cdots + h_N x_N(t) + \omega, \tag{1}$$

where h_i is the channel gain between the CBS and SU_i, $i \in \{1, 2, \ldots, N\}$, $x_i(t) = \sqrt{P_i(t)} s_i(t)$, and $|h_1|^2 > |h_2|^2 > \cdots > |h_N|^2$, when $s_i(t)$ is the signal transmitted by SU_i $\left(\mathbb{E}\{|s_i(t)|^2\} = 1\right)$ with transmission power $P_i(t) = e_i^{tr}(t)$ τ_{tr}, in which $e_i^{tr}(t)$ is the transmission energy assigned to SU_i for the t^{th} time slot; and ω is the additive white Gaussian noise (AWGN) at the CBS with zero mean and variance σ_ω^2.

Figure 3 illustrates the SIC detection process of the received signals at the CBS, where $|h_1|^2 > |h_2|^2 > \cdots > |h_N|^2$. In uplink NOMA, an SU with the strongest channel gain will definitely have the priority for decoding by the CBS, and then, it vanishes from received signals y at the CBS, which continues decoding the other SU signals. Consequently, the attainable throughput of SU_1 is affected by interference from other users $(\mathrm{SU}_2, \mathrm{SU}_3, \ldots, \mathrm{SU}_N)$, and meanwhile, the throughput of the lowest channel gain user (i.e SU_N) is obtained without any interference from the other SUs because interference from stronger signals is eliminated by the SIC technique. Thereby, the throughput for SU_i, $\forall i \in \{1, 2, \ldots, N\}$ in uplink NOMA can be computed as [35]

$$R_i(t) = \frac{T - \tau_{ss} - \tau_u}{T} \log_2 \left(1 + \frac{P_i(t)|h_i|^2}{\sum_{j=i+1}^{N} P_j(t)|h_j|^2 + \sigma_\omega^2} \right), \tag{2}$$

where T, τ_{ss}, and τ_u denote the whole-frame time, the sensing and decision time, and the energy information update time, respectively. The total received throughput at the CBS can be given by

$$R(t) = \sum_{i=1}^{N} R_i(t). \tag{3}$$

2.2 Energy harvesting and primary user models

Herein, we assume that the SUs always harvest energy during the whole of time slot T, and the amount of harvested energy is stored in their finite capacity batteries. Since the SUs perform the energy harvesting process in the same environment, it is also worth noting that they have the same distribution. The amount of harvested energy, $e^{hv,i}$, of SU_i in each time slot follows a Poisson distribution process with mean value e_{mean}^{hv}. The value for $e^{hv,i}$ in time slot t can be expressed as $e^{hv,i}(t) = \{e_1^{hv}, e_2^{hv}, e_3^{hv}, \ldots, e_v^{hv}\}$ where $0 < e_1^{hv} < e_2^{hv} < e_3^{hv} < \cdots < e_v^{hv} < E_{ca}$. The probability mass function for harvested energy can be given as [37]

$$p_{hv}(k) = \Pr\left(e^{hv} = e_k^{hv}\right) = \frac{e^{-e_{mean}^{hv}} \left(e_{mean}^{hv}\right)^k}{k!}, k = 1, 2, \ldots, v. \tag{4}$$

In each time slot, cooperative sensing is performed by the secondary network to predict the state of the PU. At the beginning of each time slot, the PU activity on the licensed

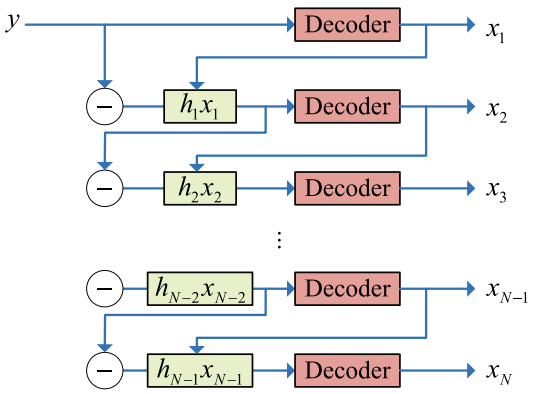


Fig. 3 Illustration of SIC detection of the signals at the CBS

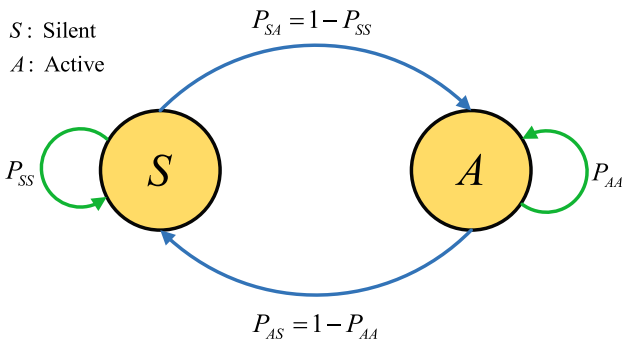


Fig. 4 Markov chain model of the primary user

channel may switch between silent (S) and active (A) states according to a two-state Markov discrete-time process, which is assumed stationary during the entire time slot, T. The state transition probability for two contiguous time slots is given by $P_{ij}|i,j \in \{S,A\}$ as shown in Fig. 4. For example, P_{SA} refers to the probability that the PU transfers from the silent state in the current time slot to the active state in the next time slot.

2.3 Imperfect spectrum sensing

In the network, the SUs need to perform spectrum sensing in every single time slot to determine the activity of the PU on the primary channel, and then, they report these local sensing decisions to the CBS. The global sensing decision is assumed to be obtained by the soft combination scheme from [53]. However, the sensing engine may induce sensing errors in practice, which results in low transmission performance by the users. Accordingly, we consider the imperfect spectrum-sensing model for the CR network. The sensing performance can be evaluated principally by two probabilities: a detection probability P_d and false alarm probability P_f, which are defined as

$$P_d = \Pr(H_A(t) = A|A) \text{ and } P_f = \Pr(H_A(t) = A|S), \quad (5)$$

respectively. P_d represents the probability that the PU is correctly found to be active, whereas P_f is the probability that the PU is found active but is actually silent. $H_A(t)$ denotes the state of the PU (i.e. the global sensing decision at the CBS) in time slot t. As such, the value of P_d is set according to the maximum acceptable probability that collisions between the secondary transmission and the primary transmission can happen [54, 55]. Besides, we further assume in this paper that the value of P_d and P_f are available to the CBS.

2.4 Problem formulation

The objective of this paper is to enhance the long-term throughput of uplink NOMA at the CBS. The power allocation problem for throughput maximization of an uplink NOMA system in time slot t can be formulated as follows:

$$\begin{aligned} &\arg\max_{\boldsymbol{A}(t)\in\mathbb{A}} \sum_{k=t}^{\infty}\sum_{i=1}^{N} R_i(k) \\ &s.t.\ 0 \le e_i^{tr}(t) \le e^{tr,\max} \end{aligned}, \quad (6)$$

where $\boldsymbol{A}(t) = \begin{bmatrix} \boldsymbol{a}(t) \\ \boldsymbol{e^{tr}}(t) \end{bmatrix}$ is the global action that the CBS assigns to the SUs in time slot t, $\boldsymbol{a}(t)$ and $\boldsymbol{e^{tr}}(t)$ represent the assigned action mode vector and the assigned transmission energy vector for the SUs, respectively. The assigned action mode, and the transmission energy for the SUs are described in the row vectors with the same dimension. The index of each element in these vectors represents the index of the corresponding SU. Particularly, $\boldsymbol{a}(t) = [a_1(t), a_2(t), \ldots, a_N(t)]$ includes the assigned action modes of all SUs in time slot t, where $a_i(t) = \{\text{“}SL\text{”}, \text{“}TM\text{”}\}$ denotes the different action modes for SU_i, and SL and TM stand for silent mode and transmission mode, respectively. Meanwhile, $\boldsymbol{e^{tr}}(t) = \left[e_1^{tr}, e_2^{tr}, \ldots, e_N^{tr}\right]$ represents the assigned amount of transmission energy for the SUs in time slot t, where $e_i^{tr}(t) \in \left\{0, e^{tr,1}, e^{tr,2}, \ldots, e^{tr,\psi}\right\}$ denotes the transmission energy of SU_i, in which $e^{tr,j}|j \in \{1,2,\ldots,\psi\}$ represents the transmission energy level. In the next section, we propose an actor–critic reinforcement learning approach to maximizing the overall reward from uplink NOMA. In particular, at the start of time slot t, the CBS will determine the most appropriate action (i.e. silent mode or transmission mode with different transmission energy levels) for each SU based on the remaining energy in each of the SUs and the belief that the PU will be inactive in the current time slot. The actor–critic framework will learn and interact directly with the environment to obtain the optimal solution for the problem in Eq. (6) after a large enough number of time slots.

3 Actor–critic reinforcement learning–based algorithm for uplink NOMA in cognitive radio networks

In this section, the actor–critic reinforcement learning approach is presented with the goal of allocating the optimal action for the SUs such that the maximum long-term throughput of uplink NOMA can be achieved according to information directly collected via practical interactions with the environment. If an SU does not have enough energy for data transmission in a time slot, it has to stay silent to save energy for the next time slot regardless of the active or inactive states of the primary user. If the channel is sensed as active, the SUs have to stay silent; otherwise, they will be assigned to concurrently transmit data with the corresponding amount of transmission energy on the channel, which is described in the following subsection.

3.1 Markov decision process

The actor–critic approach is a type of MDP [56], that can be defined as a quintuple $\langle \mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{R}, \gamma \rangle$, where $\mathbb{S}$ is the state space, $\mathbb{A}$ represents the action space set, $\mathbb{P}$ is the

transition probability set in which the state of the agent changes from the current state to the next state when action $\boldsymbol{A}$ is taken, $\mathbb{R}$ is the reward space, and $\gamma \in [0, 1)$ denotes the discount factor.

- *State space*: The state of the network in time slot t is $s(t) = (\mu(t), \boldsymbol{e^{re}}(t))$, where $\mu(t)$ is the belief representing the probability that the PU is idle in this time slot, and $\boldsymbol{e^{re}}(t)$ is a vector that includes the remaining energy of the SUs

$$\left(\boldsymbol{e^{re}}(t) = \left[e_1^{re}(t), e_2^{re}(t), \ldots, e_N^{re}(t)\right]\right)$$

  at the beginning of time slot t.
- *Action space*: The CBS assigns global action $\boldsymbol{A}(t)$, which comprises two vectors: $\boldsymbol{a}(t) = [a_1(t), a_2(t), \ldots, a_N(t)]$ and $\boldsymbol{e^{tr}}(t) = \left[e_1^{tr}, e_2^{tr}, \ldots, e_N^{tr}\right]$. Note that each element of these vectors is sorted by following the corresponding index of each SU in the network.
- *Reward*: Given the state of the system, $s(t)$, and the action, $\boldsymbol{A}(t)$, each SU performs its own assigned action. As a result, the system can obtain an immediate reward which is defined as the summation of the SUs' throughput in the current time slot: $R(\boldsymbol{A}(t), s(t))$.

The state–value function $V(s(t))$ is the cumulative discounted reward from current state $s(t)$. If the CBS uses the policy, $\pi_t(\boldsymbol{A}(t)|s(t))$, the probability that the CBS will assign action $\boldsymbol{A}(t)$ for given state $s(t)$, then the state–value function, $V(s(t))$, can be expressed as follows [57]

$$V(s(t)) = \sum_{k=t}^{\infty} \gamma^{k-t} R(s(k), \boldsymbol{A}(k)). \tag{7}$$

The objective of the actor–critic reinforcement learning algorithm is to find an optimal policy $\pi_t^*(\boldsymbol{A}(t)|s(t))$ that maximizes the state–value function of each state $s(t)$ defined by Eq. (7). The optimal policy can be described by [57]

$$\pi_t^*(\boldsymbol{A}(t)|s(t)) = \arg\max_{\boldsymbol{A}(t)\in\mathbb{A}} \left\{ \sum_{k=t}^{\infty} \gamma^{k-t} R(\boldsymbol{s}(k), \boldsymbol{A}(k)|s(k) = s) \right\}. \tag{8}$$

3.2 Actor–critic reinforcement learning algorithm

In this paper, we present the actor–critic approach as a model-free reinforcement learning framework to solve the MDP problem. The advantage of this algorithm is that it does not require any prior information from the dynamic environment (i.e. the harvested energy distribution). That is, it is worthwhile to utilize the actor–critic scheme in practice from the viewpoint that prior information from the environment is not easy to acquire. On the other hand, the system can directly interact with the environment to learn the information about the harvested energy.

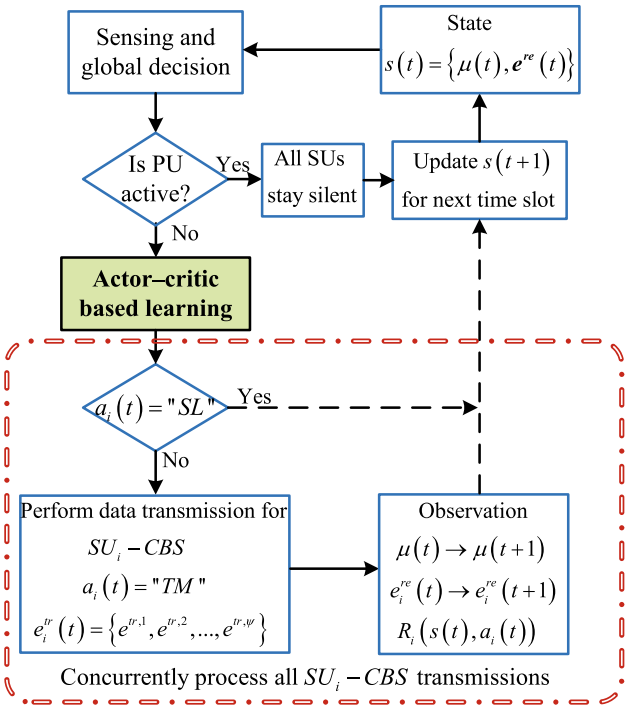


Fig. 5 The flowchart of the proposed scheme

Figure 5 depicts the flowchart of the proposed scheme based on the actor–critic learning method. In a time slot, after cooperative spectrum sensing is executed, the CBS makes the global sensing decision about the existence

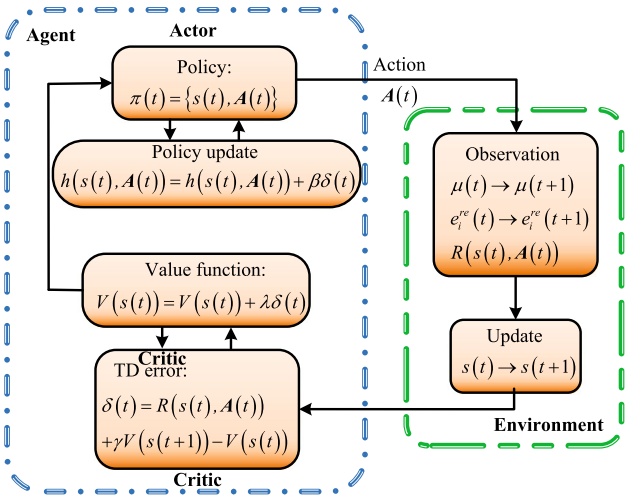


Fig. 6 The actor–critic learning process in the proposed scheme

of the PU on the licensed channel. If the global sensing decision indicates that the PU is active, the SUs accept this result and stay silent. Then, the SUs send to the CBS an update on their remaining energy, and the belief $\mu(t+1)$ can be calculated at the end of time slot t. Otherwise, if the global sensing decision indicates the PU is inactive, the CBS can choose the possible actions from set $\mathbb{A}$ by applying the proposed actor–critic learning approach. The belief can be updated by observing the successful/unsuccessful transmissions of the SUs if they are assigned to transmit data to the CBS.

The actor–critic learning process consists of two components (the *actor* and the *critic*), as shown in Fig. 6. The *actor* is used to define the policy and generate actions based on the observed environment state, while the *critic* learns the state–value function and criticizes the actions selected by the actor. At the start of each time slot, the actor employs an action, $\boldsymbol{A}(t) \in \mathbb{A}$, by following policy $\pi_t(\boldsymbol{A}(t)|s(t))$. The policy is calculated via Gibbs soft-max distribution [57]:

$$\pi_t(\boldsymbol{A}(t)|s(t)) = \frac{e^{h(s(t),\boldsymbol{A}(t))}}{\sum_{\boldsymbol{A}\in\mathbb{A}} e^{h(s(t),\boldsymbol{A})}}, \tag{9}$$

where $h(s(t),\boldsymbol{A}(t))$ indicates the tendency to select action $\boldsymbol{A}(t)$ in state $s(t)$.

At the end of a time slot, the system will update the immediate reward, $R(s(t),\boldsymbol{A}(t))$, and the next state $s(t+1)$. Subsequently, the critic will criticize the selected action and evaluate the policy by using the temporal difference (TD) error, which is computed as [57]

$$\delta(t) = R(s(t),\boldsymbol{A}(t)) + \gamma V(s(t+1)) - V(s(t)), \tag{10}$$

where $\delta(t)$ denotes the difference between state–value function $V(s(t))$ from the preceding state and the state–value function after taking the selected action. Then, the critic uses the TD error to criticize the selected action as follows [57]:

$$V(s(t)) = V(s(t)) + \lambda\delta(t), \tag{11}$$

where λ is the learning step-size of the critic. Thereafter, the TD error is fed back to the actor, and the tendency to select the action is upgraded as [57]

$$h(s(t),\boldsymbol{A}(t)) = h(s(t),\boldsymbol{A}(t)) + \beta\delta(t), \tag{12}$$

where β is the learning step-size of the actor. Ultimately, the policy is updated by Eqs. (9) and (12) for action selection in the subsequent time slots. The training process will be completed when state–value function $V(s(t))$ and policy $\pi_t(\boldsymbol{A}(t)|s(t))$ converge to $V^*(s(t))$ and $\pi_t^*(\boldsymbol{A}(t)|s(t))$ with probability 1 as $t \to \infty$ [58].

Hereafter, when the CBS assigns an action for each SU, one of the following observations may happen.

3.2.1 Silent mode (Ω_1)

If the global sensing decision indicates that the PU is active in the current time slot. The CBS will trust this result and assign action *SL* to all SUs. In this case, no reward is achieved, i.e. $R(s(t),\boldsymbol{A}(t)|\Omega_1) = 0$. The belief in the current time slot can be updated using Bayes' rule [59] as follows:

$$\mu(t) = \frac{\mu(t)P_f}{\mu(t)P_f + (1-\mu(t))P_d}. \tag{13}$$

The updated belief for the next time slot is given as

$$\mu(t+1) = \mu(t)P_{SS} + (1-\mu(t))P_{AS}. \tag{14}$$

For simplicity in this work, we assume that the energy consumed for the information update of the SUs is tiny and can be ignored. Hence, the remaining energy of SU_i for the next time slot can be calculated as follows:

$$e_i^{re}(t+1) = \min\{e_i^{re}(t) + e^{hv,i}(t) - e_{ss}, E_{ca}\}, \tag{15}$$

where e_{ss} denotes the energy consumed for spectrum sensing.

Algorithm 1 Actor–critic reinforcement learning procedure of the transmission power decision policy for the SUs

Input: $\mathbb{S}$, $\mathbb{A}$, γ, λ, β, $\boldsymbol{e^{re}}(t)$, $\mu(t)$, E_{ca}, e^{hv}_{mean}, T. // Initial parameters
Initialize state–value function $V(s(t))$, tendency $h(s(t), \boldsymbol{A}(t))$, and policy $\pi_t(\boldsymbol{A}(t)|s(t))$;
Repeat until convergence
for each time slot,
  Define the current state $s(t) \in \mathbb{S}$
  Choose an action, $\boldsymbol{A}(t) \in \mathbb{A}$, according to policy $\pi_t(\boldsymbol{A}(t)|s(t))$ in Eqn. (9) after considering the sensing result and the remaining energy of the SUs.

  Simultaneously excute the process for all SUs:
  if $a_i(t) =$ "SL" // if SU_i is assigned to stay silent
    SU_i stays silent and only harvests solar energy.
  else// if SU_i is assigned to transmit data with the transmission energy $e_i^{tr}(t)$
    SU_i transmits data to CBS with assigned transmission energy and harvests solar energy.
  end if
  Compute instant reward $R_i(s(t), a_i(t))$; update network state $s(t+1)$.

  Critic Process:
  Calculate TD error $\delta(t)$ with Eqn. (10).
  Update state–value function $V(s(t))$ with Eqn. (11).

  Actor Process:
  Update tendency to select an action, $\boldsymbol{A}(t)$, $h(s(t), \boldsymbol{A}(t))$, with Eqn. (12).
  Update policy to choose action $\boldsymbol{A}(t)$ under the given state, $\pi_t(\boldsymbol{A}(t)|s(t))$, with Eqn. (9).
end for
Output: Final policy $\pi_t^*(\boldsymbol{A}(t)|s(t))$. // Optimal action at given state.

3.2.2 Transmission mode

If the global sensing decision indicates that the PU is silent, then the CBS allows all SUs to transmit data with $a_i(t) = TM$ and the corresponding transmission energy level $e_i^{tr}(t)$. In this case, there are two observations: Ω_2 and Ω_3.

Observation 2 (Ω_2): The CBS can successfully decode the signals transmitted by the SUs at the end of the time slot. In this case, the system recognizes that the PU was actually silent in the time slot. The reward can be computed as

$$R(s(t), \boldsymbol{A}(t)|\Omega_2) = \sum_{i=1}^{N} R_i(t), \tag{16}$$

where the throughput of the SU_i, $R_i(t)$ can be calculated with Eq. (2). Belief $\mu(t+1)$ for the next time slot can be updated as

$$\mu(t+1) = P_{SS}. \tag{17}$$

The remaining energy of SU_i can be updated as follows:

$$e_i^{re}(t+1) = \min\left\{e_i^{re}(t) + e^{hv,i}(t) - e_{ss} - e_i^{tr}(t), E_{ca}\right\}, \tag{18}$$

where $e_i^{tr}(t)$ denotes the transmission energy of SU_i in time slot t.

Observation 3 (Ω_3): The CBS can not successfully decode the signals transmitted by the SUs due to collisions between the SUs and PUs transmissions. The system can infer that misdetection occurred in this case. There is no reward: $R(s(t), \boldsymbol{A}(t)|\Omega_3) = 0$.

Belief $\mu(t+1)$ for the next time slot is updated as

$$\mu(t+1) = P_{AS}. \tag{19}$$

The remaining energy of SU_i for the next time slot is updated by:

$$e_i^{re}(t+1) = \min\left\{e_i^{re}(t) + e^{hv,i}(t) - e_{ss} - e_i^{tr}(t), E_{ca}\right\}. \tag{20}$$

In the actor–critic algorithm, the state–value function and the policy parameters are sequentially and concurrently updated based on the action of the CBS over the time slots. The policy of the system can be dynamically obtained from a practical learning process, such that the local optimal policy can converge over a large number of time slots [60]. Finally, we summarize the learning process of the proposed actor–critic scheme in **Algorithm 1**.

4 Simulation results

In this section, we analyze the performances of the proposed scheme under various conditions of the network by using simulation results based on MATLAB R2019a. In

Table 2 Simulation parameters

Parameter	Description	Value
N	Number of SUs	2
T	Time slot duration	200 ms
τ_{ss}	Sensing duration	2 ms
τ_u	Update duration	1 ms
E_{ca}	Battery capacity	20 μJ
e_{ss}	Sensing cost	1 μJ
e^{tr}	Transmission energy	5, 10, 15 μJ
e^{hv}_{mean}	Mean value of harvested energy	6 μJ
μ	Initial belief that the PU is absent	0.5
P_{SS}	Transition probability of the PU from state S to itself	0.8
P_{AS}	Transition probability of the PU from state A to state S	0.2
P_d	Probability of detection	0.9
P_f	Probability of false alarm	0.1
h_1	Channel gain between SU_1 and the CBS	− 20 dB
h_2	Channel gain between SU_2 and the CBS	−35 dB
σ^2_ω	Noise variance	−80 dB
γ	Discount factor	0.95
λ, β	Learning step-sizes of critic, actor	0.1, 0.1

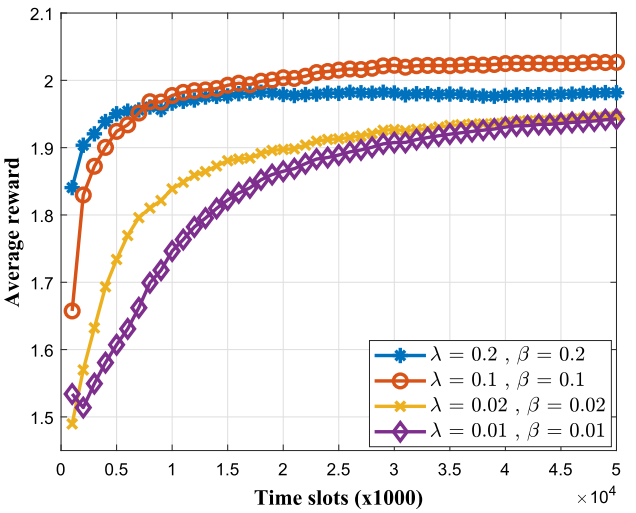


Fig. 7 The convergence process of the actor–critic according to different values of learning step-size when $e^{hv}_{mean} = 6\,\mu\text{J}, E_{ca} = 20\,\mu\text{J}, h_1 = -20\,\text{dB}, h_2 = -35\,\text{dB}$, and $\sigma^2_\omega = -80$ dB

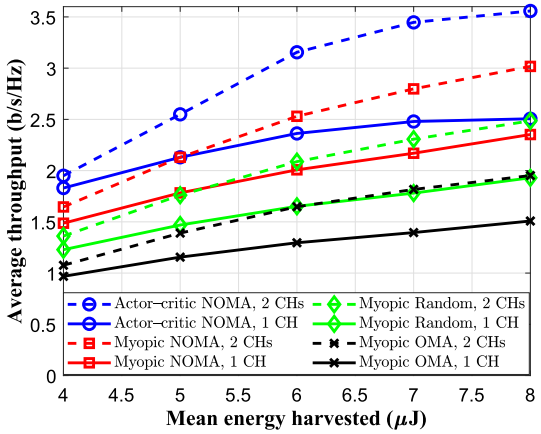


Fig. 8 Average throughput of the secondary system under various values of harvested energy in two cases (1 CH and 2 CHs) when $E_{ca} = 20\,\mu\text{J}, h_1 = -20\,\text{dB}, h_2 = -35\,\text{dB}$, and $\sigma^2_\omega = -80$ dB

addition, we also compare the proposed scheme with the other conventional schemes, such as the Myopic NOMA scheme, the Myopic OMA scheme, and the Myopic Random scheme, where the term "Myopic" refers to the policy that only maximizes the instant reward of the system [61]. In the Myopic NOMA scheme, if the sensing result indicates the absence of the PU, the SUs will simultaneously transmit data to the CBS at the highest transmission energy level, and then, similar to the proposed scheme the received signals will be decoded at the CBS by applying the SIC technique. For the Myopic OMA scheme, the decision is made by combining the myopic approach in [61] and TDMA-based technique in [62]. More specifically, the entire data transmission phase in a time slot is equally divided into sub-phases according to the number of SUs. After that, the SUs transmit their information in rapid succession during their respective sub-phases, one after the other. For the simulation $(N = 2)$, we assumed each SU transmits data in half of the time for the transmission phase

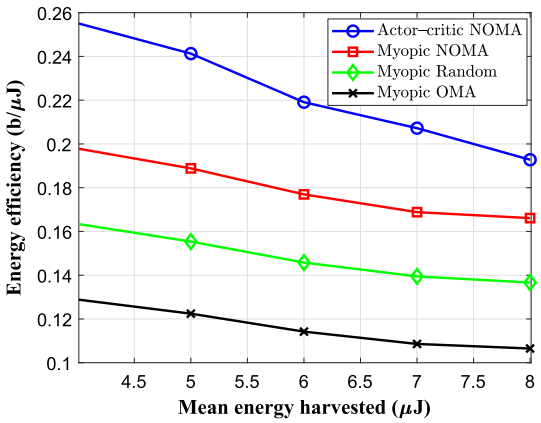


Fig. 9 Energy efficiency of the secondary system for various values of harvested energy when $E_{ca} = 20\,\mu\text{J}, h_1 = -20\,\text{dB}, h_2 = -35\,\text{dB}$, and $\sigma_\omega^2 = -80$ dB

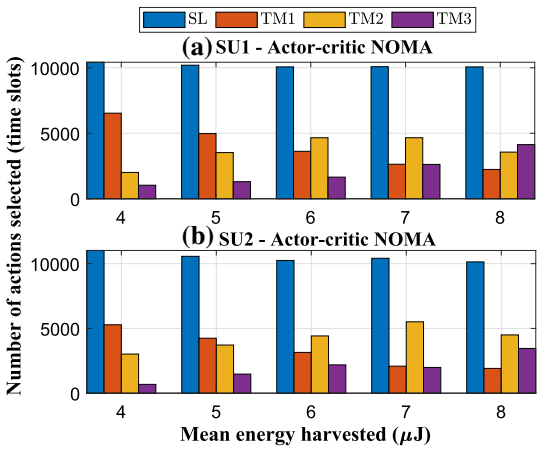


Fig. 10 The selected action statistics of each secondary user according to various values of harvested energy in the case of the actor–critic NOMA approach when $E_{ca} = 20\,\mu\text{J}, h_1 = -20\,\text{dB}, h_2 = -35$ dB, and $\sigma_\omega^2 = -80$ dB

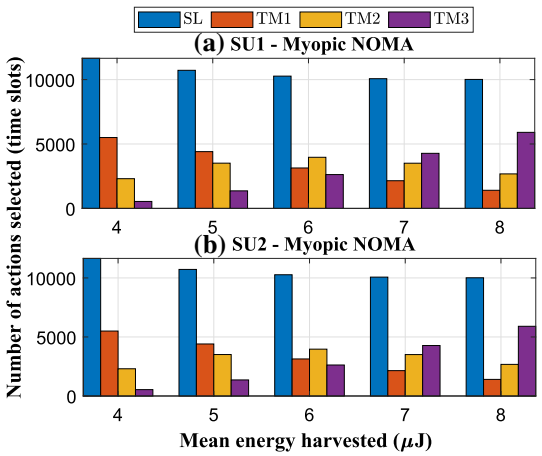


Fig. 11 The selected action statistics of each secondary user according to various values of harvested energy in the case of the Myopic NOMA approach when $E_{ca} = 20\,\mu\text{J}, h_1 = -20\,\text{dB}, h_2 = -35$ dB, and $\sigma_\omega^2 = -80$ dB

after sensing, and then, the achievable throughput of SU_i at the CBS can be computed as [63]

$$R_i^{OMA}(t) = \frac{\tau_{tr}}{2T}\log_2\left(1 + \frac{P_i(t)|h_i|^2}{\sigma_\omega^2}\right), \tag{21}$$

In the Myopic Random scheme, the CBS randomly assigns NOMA/OMA to the SUs when the global sensing decision indicates that the PU is silent in the current time slot. For simplicity, the channel gain for each SU is fixed, and there are three levels for the transmission energy of the SUs: $\text{TM1} = 5\,\mu\text{J}, \text{TM2} = 10\,\mu\text{J}, \text{TM3} = 15\,\mu\text{J}$. The span of each belief is 0.1. Furthermore, the performance of the proposed scheme was verified over 30,000 time slots, and the results were acquired by averaging 10 separate loops. Table 2 shows the simulation parameters for the scheme proposed in this paper.

In Fig. 7, we examine the convergence of the proposed scheme's algorithm over time slots with various step-size parameters, λ and β, based on the reward (throughput) of the system. In this paper, the convergence condition for the proposed scheme was set at 10^{-3}. We can see that the system reward greatly increases during the first 15,000 time slots, and then gradually converges to the optimal value, which depends on the different values of λ and β. Obviously, if larger values of λ and β are used, the faster the convergence and the higher the throughput. However, from the figure, we can see that increasing the value of λ and β does not always guarantee a higher reward for the network due to underfitting, whereas the system might be prone to overfitting if we reduce the learning parameters. As a result, we set the value of actor and critic learning step-sizes as $\lambda = 0.1$, and $\beta = 0.1$, respectively, for the proposed scheme in the upcoming simulation.

In Fig. 8, we show the effect of the amount of harvested energy of the SUs and the number of primary channels on average throughput of the system. We can see that as e_{mean}^{hv} increases, the SUs can collect more energy from the solar source, and transmit at a higher transmission energy level, which leads to higher achievable throughput at the CBS. In addition, the performance of the proposed scheme outperforms the conventional schemes, since the conventional schemes disregard the effect of the current decision on future rewards. For that reason, whenever the PU is sensed as silent on the primary channel, then these conventional schemes will allow SUs to use as much energy as possible to maximize the immediate throughput at the CBS. However, this action causes the SUs to stay silent longer than under the proposed actor–critic NOMA scheme owing to the limitations on battery capacity and harvested energy. It is also observed that average throughput achieved at the

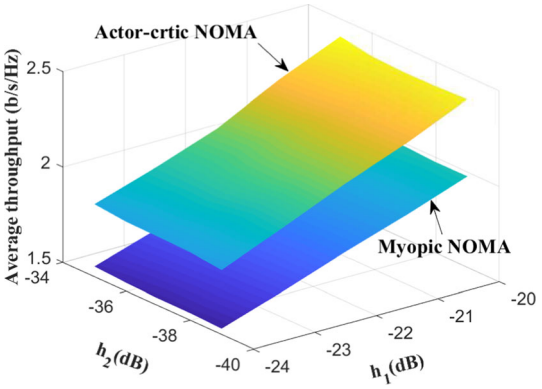


Fig. 12 Average throughput for different values of h_1 and h_2 when $e^{hv}_{mean} = 6\,\mu\text{J}, E_{ca} = 20\,\mu\text{J}$, and $\sigma^2_\omega = -80$ dB

CBS in the case of two primary channels (2 CHs) is larger than that of a single channel case (1 CH). Obviously, with more primary channels, the SUs have more chances to transmit their data. As a consequence, the performance of the cognitive radio system is enhanced in the case of multiple channels. However, the proposed scheme provides the highest throughput in both cases of single and multiple channels.

Figure 9 shows the energy efficiency of the system under various mean values of energy harvesting. In this paper, we define energy efficiency as the average long-term throughput over the total energy-harvesting amount during the operations spanning M $(M = 20,000)$ time slots $\left(EE = \sum_{t=1}^{M}\sum_{i=1}^{N} R_i(t) \Big/ \sum_{t=1}^{M}\sum_{i=1}^{N} e^{hv,i}(t)\right)$. In order to enhance the energy efficiency of the proposed scheme, we set the maximum transmission energy level for the SU if its battery is likely to overflow in each time slot [64]. From Fig. 9, we can see that the energy efficiency drops with increased levels of energy harvesting, e^{hv}_{mean}. The reason is

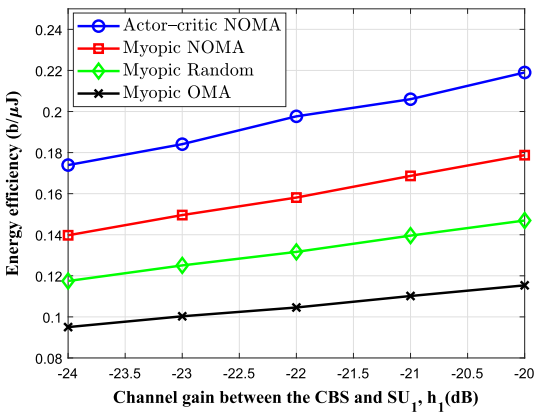


Fig. 13 Energy efficiency according to the channel gain between the CBS and SU_1 when $e^{hv}_{mean} = 6\,\mu\text{J}, E_{ca} = 20\,\mu\text{J}, h_2 = -35$ dB, and $\sigma^2_\omega = -80$ dB

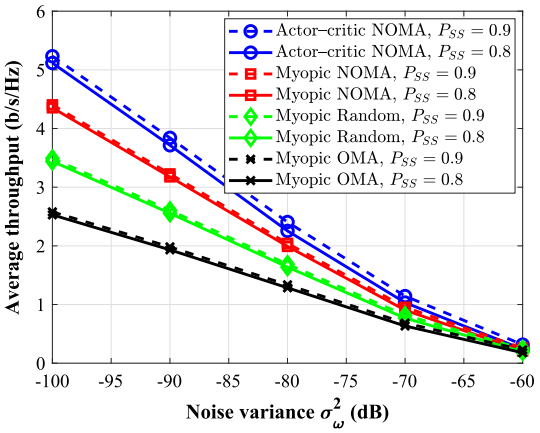


Fig. 14 Average throughput according to the noise variance in two cases of $P_{SS} = 0.9$ and $P_{SS} = 0.8$ when $e^{hv}_{mean} = 6\,\mu\text{J}, E_{ca} = 20\,\mu\text{J}, h_1 = -20$ dB, and $h_2 = -35$ dB

that when e^{hv}_{mean} increases, the SUs can gather more energy for their operations but the total amount of overflow energy in the SUs increases concurrently. As a consequence, the figure shows that the proposed scheme is still superior to other conventional schemes under different amounts of harvested energy. For instance, when $e^{hv}_{mean} = 6\,\mu\text{J}$, the energy efficiency of the proposed scheme can provide 23.8%, 50.3%, and 91.8% in energy utilization improvement for the Myopic NOMA, Myopic Random, and Myopic OMA schemes, respectively. Myopic OMA brings the lowest result, because the SUs transmit data in turn during each half-phase of the data transmission duration, while other NOMA schemes allow the SUs to simultaneously transmit data over the entire data transmission phase.

Specific information about the number of actions selected for each SU under a change in e^{hv}_{mean} for the proposed scheme and the Myopic NOMA scheme is illustrated in Figs. 10 and 11, respectively. We can see that the proposed scheme normally assigns the proper amount of

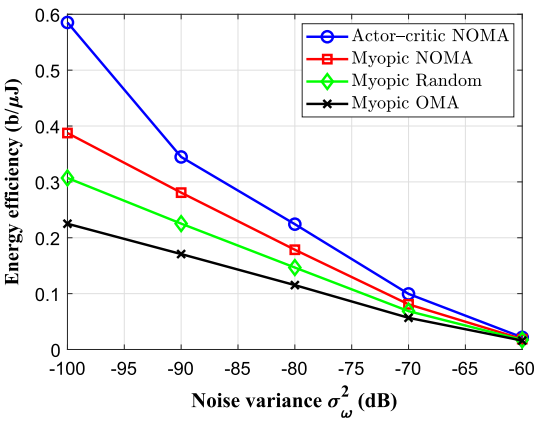


Fig. 15 Energy efficiency according to the noise variance when $e^{hv}_{mean} = 6\,\mu\text{J}, E_{ca} = 20\,\mu\text{J}, h_1 = -20$ dB, and $h_2 = -35$ dB

transmission energy for the SUs at low values for the harvested energy mean, e^{hv}_{mean}. Meanwhile Myopic NOMA scheme assigns the SUs the highest possible transmission energy at all values of e^{hv}_{mean}. This creates inefficiency in terms of both energy and throughput metrics, as presented in Figs. 8 and 9. It is obvious that although harvested energy may vary over time slots, the SUs in the proposed scheme usually utilize TM3 to obtain the highest throughput, provided that the PU is most likely absent, or energy overflow might happen.

We further investigated the joint effect of channel gains between the CBS and SUs' on throughput of the system, as shown in Fig. 12. It is evident that the performance of the system improves with an increase in channel gain h_1 and h_2. The reason is that the throughput of system throughput is dependent on the channel gain as shown in Eqs. (2) and (3). Thus, the larger channel gains are, the higher throughput the system obtains. Specifically, when h_1 becomes larger, the average throughput of the system significantly increases. On the other hand, the average throughput of the system only slightly increases when h_2 increases. That is because the value of h_2 is quite a bit smaller than the value of h_1, and subsequently, it has less influence on the signal of SU_1 and further on total throughput of the system. Thus, increasing h_2 does not much influence on the overall throughput at the CBS due to its small channel gain.

In Fig. 13, we investigate the energy efficiency of the proposed scheme versus the various values for channel gain between the CBS and SU_1. The curves show that the energy efficiency of the system benefits from larger values of h_1 because with the same transmission power for SUs, the higher channel gain will bring more throughput at the CBS. Consequently, the proposed scheme is verified to be superior to other conventional schemes in terms of efficient energy utilization under the variation of channel gain.

In Fig. 14, we jointly study average system throughput of the schemes under the impact of various noise variances σ^2_ω and primary user activity which is expressed as transition probability of PUs from state *silent* to state *silent*. Figure 14 shows that a large amount of noise variance can significantly degrade the obtained throughput. It can be explained as following: when the noise variance goes up, it will severely interfere with the received signals at the CBS. According to Eqs. (2) and (21), the noise variance, which is an interfering component in the denominator of the signal-to-interference-plus-noise ratio (SINR), will lower the system throughput as it increases, and vice versa. In addition, we can see that the performances of all schemes can get better as P_{SS} increases. The reason is that when the transition probability of PUs from state *silent* to itself rises, the probability that the primary channel is free also goes up, which results in more opportunities for the SUs to transmit data on the primary channel.

Finally, in Fig. 15, we examine the effect of noise variance on the energy efficiency of the schemes.It is observed that the energy efficiency of all schemes deteriorates as the noise variance increases. The reason for this is as following: the energy efficiency is calculated by the achieved long-term throughput over the total energy-harvesting during operational time. Hence, for the large value of the noise variance, the average long-term throughput tends to be reduced, which results in the low energy efficiency in the system. The figure shows that the energy efficiency of the proposed schemes outperforms the other conventional schemes. Furthermore, the figure also shows that the proposed scheme is more robust to noise variation at the CBS than other schemes.

5 Conclusion

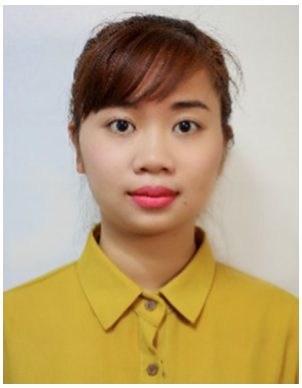
In this paper, we propose an actor–critic reinforcement learning approach using uplink NOMA in a cognitive radio network. In the network, the solar energy–powered SUs can simultaneously transmit data to a cognitive base station. The energy-constrained and the imperfect-sensing problems are also taken into account. Consequently, the optimal policy can be obtained by using the proposed scheme, where the SUs can be assigned the proper action mode (i.e. stay silent or transmit data) to maximize the long-term throughput of the secondary system. Simulation results demonstrate that the proposed scheme can improve both long-term throughput and the energy efficiency of the network, compared with conventional schemes.

References

1. Khan, F. A., Ratnarajah, T., & Sellathurai, M. (2010). Multiuser diversity analysis in spectrum sharing cognitive radio networks. In *2010 Proceedings of the fifth international conference on cognitive radio oriented wireless networks and communications, Cannes* (pp. 1–5).
2. Wang, C., et al. (2014). Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Communications Magazine*, *52*(2), 122–130.
3. Hong, X., Wang, J., Wang, C., & Shi, J. (2014). Cognitive radio in 5G: A perspective on energy-spectral efficiency trade-off. *IEEE Communications Magazine*, *52*(7), 46–53.
4. Wang, B., & Liu, K. J. R. (2011). Advances in cognitive radio networks: A survey. *IEEE Journal of Selected Topics in Signal Processing*, *5*(1), 5–23.
5. Akyildiz, I. F., Lee, W., Vuran, M. C., & Mohanty, S. (2008). A survey on spectrum management in cognitive radio networks. *IEEE Communications Magazine*, *46*(4), 40–48.

6. Mitola, J, I. I. I., & Maguire, G. Q, Jr. (1999). Cognitive radio: Making software radios more personal. *IEEE Personal Communications*, *6*(4), 13–18.
7. Haykin, S. (2005). Cognitive radio: Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, *23*(2), 201–220.
8. Hossain, E., Niyato, D., & Han, Z. (2009). *Dynamic spectrum access and management in cognitive radio networks*. Cambridge: Cambridge University Press.
9. Lv, L., Chen, J., Ni, A., Ding, Q. Z., & Jiang, H. (2018). Cognitive non-orthogonal multiple access with cooperative relaying: A new wireless frontier for 5G spectrum sharing. *IEEE Communications Magazine*, *56*(4), 188–195.
10. Goldsmith, A., Jafar, S. A., Maric, I., & Srinivasa, S. (2009). Breaking spectrum gridlock with cognitive radios: An information theoretic perspective. *Proceedings of the IEEE*, *97*(5), 894–914.
11. Giang, H. T. H., Hoan, T. N. K., Thanh, P. D., & Koo, I. (2019). A POMDP-based long-term transmission rate maximization for cognitive radio networks with wireless-powered ambient backscatter. *International Journal of Communication Systems*,. https://doi.org/10.1002/dac.3993.
12. Ding, Z., et al. (2017). Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Communications Magazine*, *55*(2), 185–191.
13. Dai, L., Wang, B., Yuan, Y., Han, S., Chih-Lin, I., & Wang, Z. (2015). Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends. *IEEE Communications Magazine*, *53*(9), 74–81.
14. Wei, Z., Yuan, J., Ng, D. W. K., Elkashlan, M., & Ding, Z. (2016). A survey of downlink non-orthogonal multiple access for 5G wireless communication networks. *ZTE Communications*, *14*(4), 17–26.
15. Saito, Y., Kishiyama,Y., Benjebbour, A., Nakamura, T., Li, A., & Higuchi, K. (2013). Non-orthogonal multiple access (NOMA) for cellular future radio access. In *2013 IEEE 77th vehicular technology conference (VTC Spring), Dresden* (pp. 1–5).
16. Ding, Z., Peng, M., & Poor, H. V. (2015). Cooperative non-orthogonal multiple access in 5G systems. *IEEE Communications Letters*, *19*(8), 1462–1465.
17. Wan, D., Wen, M., Ji, F., Yu, H., & Chen, F. (2018). Non-orthogonal multiple access for cooperative communications: Challenges, opportunities, and trends. *IEEE Wireless Communications*, *25*(2), 109–117.
18. Tabassum, H., Hossain, E., & Hossain, J. (2017). Modeling and analysis of up-link non-orthogonal multiple access in large-scale cellular networks using Poisson cluster processes. *IEEE Transactions on Communications*, *65*(8), 3555–3570.
19. Razavi, R., Hoshyar, R., Imran, M. A., & Wang, Y. (2011). Information theoretic analysis of LDS scheme. *IEEE Communications Letters*, *15*(8), 798–800.
20. AL-Imari, M., Imran, M. A. C., & Tafazolli, R. (2012). Low Density Spreading for next generation multicarrier cellular systems. In *2012 International conference on future communication networks, Baghdad* (pp. 52–57).
21. Du, Y., Dong, B., Chen, Z., Fang, J. C., & Wang, X. (2016). A fast convergence multiuser detection scheme for uplink SCMA systems. *IEEE Wireless Communications Letters*, *5*(4), 388–391.
22. Nikopour, H., et al. (2014). SCMA for downlink multiple access of 5G wireless networks. In *2014 IEEE global communications conference, Austin, TX* (pp. 3940–3945).
23. Liu, Y., Ding, Z., Elkashlan, M., & Yuan, J. (2016). Nonorthogonal multiple access in large-scale underlay cognitive radio networks. *IEEE Transactions on Vehicular Technology*, *65*(12), 10152–10157.
24. Wang, D., & Men, S. (2018). Secure energy efficiency for NOMA based cognitive radio networks with nonlinear energy harvesting. *IEEE Access*, *6*, 62707–62716.
25. Lv, L., Ni, Q., Ding, Z., & Chen, J. (2017). Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over Nakagami-*m* fading channels. *IEEE Transactions on Vehicular Technology*, *66*(6), 5506–5511.
26. Lv, L., Chen, J., Ni, Q., & Ding, Z. (2017). Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis. *IEEE Transactions on Communications*, *65*(6), 2641–2656.
27. Simjee, F. I., & Chou, P. H. (2008). Efficient charging of supercapacitors for extended lifetime of wireless sensor nodes. *IEEE Transactions on Power Electronics*, *23*(3), 1526–1536.
28. Chen, Z., Law, M., Mak, P., & Martins, R. P. (2017). A single-chip solar energy harvesting IC using integrated photodiodes for biomedical implant applications. *IEEE Transactions on Biomedical Circuits and Systems*, *11*(1), 44–53.
29. Wang, C., Li, J., Yang, Y., & Ye, F. (2018). Combining solar energy harvesting with wireless charging for hybrid wireless sensor networks. *IEEE Transactions on Mobile Computing*, *17*(3), 560–576.
30. Stuyts, J., Horn, G., Vandermeulen, W., Driesen, J., & Diehl, M. (2015). Effect of the electrical energy conversion on optimal cycles for pumping airborne wind energy. *IEEE Transactions on Sustainable Energy*, *6*(1), 2–10.
31. Zhao, L., Tang, L., Liang, J., & Yang, Y. (2017). Synergy of wind energy harvesting and synchronized switch harvesting interface circuit. *IEEE/ASME Transactions on Mechatronics*, *22*(2), 1093–1103.
32. Zou, Z., Gidmark, A., Charalambous, T., & Johansson, M. (2016). Optimal radio frequency energy harvesting with limited energy arrival knowledge. *IEEE Journal on Selected Areas in Communications*, *34*(12), 3528–3539.
33. Celik, A., Alsharoa, A., & Kamal, A. E. (2017). Hybrid energy harvesting-based cooperative spectrum sensing and access in heterogeneous cognitive radio networks. *IEEE Transactions on Cognitive Communications and Networking*, *3*(1), 37–48.
34. Zhu, L., Zhang, J., Xiao, Z., Cao, X., Wu, D. O., & Xia, X. (2018). Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications. *IEEE Transactions on Wireless Communications*, *17*(9), 6177–6189.
35. Ali, M. S., Tabassum, H., & Hossain, E. (2016). Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems. *IEEE Access*, *4*, 6325–6343.
36. Zhai, D., & Du, J. (2018). Spectrum efficient resource management for multi-carrier-based NOMA networks: A graph-based method. *IEEE Wireless Communications Letters*, *7*(3), 388–391.
37. Thanh, P. D., Hoan, T. N. K., & Koo, I. (2020). Joint resource allocation and transmission mode selection using a POMDP-based hybrid half-duplex/full-duplex scheme for secrecy rate maximization in multi-channel cognitive radio networks. *IEEE Sensors Journal*, *20*(7), 3930–3945.
38. Shah, H. A., Koo, I., & Kyung, S. K. (2019). Actor–critic-algorithm-based accurate spectrum sensing and transmission framework and energy conservation in energy-constrained wireless sensor network-based cognitive radios. *Wireless Communications and Mobile Computing*,. https://doi.org/10.1155/2019/6051201.
39. Li, X., Fang, J., Cheng, W., Duan, H., Chen, Z., & Li, H. (2018). Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach. *IEEE Access*, *6*, 25463–25473.

40. Yang, H., & Xie, X. (2020). An actor–critic deep reinforcement learning approach for transmission scheduling in cognitive internet of things systems. *IEEE Systems Journal*, *14*(1), 51–60.
41. Zhang, Y., Wang, X. & Xu, Y. (2019). Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning. In *2019 11th International conference on wireless communications and signal processing (WCSP), Xi'an, China* (pp. 1–6).
42. Zhang, J., Tao, X., Wu, H., Zhang, N., & Zhang, X. (2020). Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system. *IEEE Internet of Things Journal*, *7*(7), 6369–6379.
43. Manimekalai, T., Joan, S. R., & Laxmikandan, T. (2020). Throughput maximization for underlay CR multicarrier NOMA network with cooperative communication. *ETRI Journal, 42*(6), 846–858.
44. Zhong, C., Lu, Z., Gursoy, M. C., & Velipasalar, S. (2018). Actor–critic deep reinforcement learning for dynamic multichannel access. In *2018 IEEE global conference on signal and information processing (GlobalSIP), Anaheim, CA, USA* (pp. 599–603).
45. Yang, Z., Feng, L., Chang, Z., Lu, J., Liu, R., Kadoch, M., & Cheriet, M. (2020). Prioritized uplink resource allocation in smart grid backscatter communication networks via deep reinforcement learning. *Electronics, 9*(4), 1–16.
46. Yucek, T., & Arslan, H. (2009). A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Communications Surveys & Tutorials*, *11*(1), 116–130.
47. Cordeiro, C., & Challapali, K. (2005). Spectrum agile radios: Utilization and sensing architectures. In *First IEEE international symposium on new frontiers in dynamic spectrum access networks (DySPAN 2005), Baltimore, MD, USA* (pp. 160–169).
48. Pawelczak, P., Janssen, G. J. M., & Prasad, R. V. (2006). WLC10-4: Performance measures of dynamic spectrum access networks. In *IEEE Globecom 2006, San Francisco, CA* (pp. 1–6).
49. Liu, X., & Shankar, S. (2006). Sensing-based opportunistic channel access. *Mobile Networks and Applications*, *11*(4), 577–591.
50. Cichoń, K., Kliks, A., & Bogucka, H. (2016). Energy-efficient cooperative spectrum sensing: A survey. *IEEE Communications Surveys & Tutorials*, *18*(3), 1861–1886.
51. Quan, Z., Ma, W.-K., Cui, S., & Sayed, A. (2010). Optimal linear fusion for distributed detection via semidefinite programming. *IEEE Transaction Signal Processing*, *58*(4), 2431–2436.
52. Ribeiro, F., de Campos, M., & Werner, S. (2012). Distributed cooperative spectrum sensing with adaptive combining. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), Kyoto, Japan* (pp. 3557–3560).
53. Han, W., Li, J., Li, Z., Si, J., & Zhang, Y. (2013). Efficient soft decision fusion rule in cooperative spectrum sensing. *IEEE Transaction Signal Processing*, *61*(8), 1931–1943.
54. Stevenson, C. R., Chouinard, G., Lei, Z., Hu, W., Shellhammer, S. J., & Caldwell, W. (2009). IEEE 802.22 The first cognitive radio wireless regional area network standard. *IEEE Communications Magazine*, *47*(1), 130–138.
55. Liang, Y., Zeng, Y., Peh, E. C. Y., & Hoang, A. T. (2008). Sensing-throughput tradeoff for cognitive radio networks. *IEEE Transactions on Wireless Communications*, *7*(4), 1326–1337.
56. Crites, R. H., & Barto, A. G. (1995). An actor/critic algorithm that is equivalent to Q-learning. In *Advances in neural information processing systems, Denver, CO* (Vol. 7, pp. 401–408).
57. Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.
58. Singh, S., Jaakkola, T., Littman, M., & Szepesvri, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, *38*(3), 287–308.
59. Stone, J. V. (2013). *Bayes' rule: A tutorial introduction to Bayesian analysis*. Sheffield: Sebtel Press.
60. Konda, V. R., & Tsitsiklis, J. N. (2000). Actor–critic algorithms. In *Advances in neural information processing systems, CO* (Vol. 12, pp. 1008–1014).
61. Wang, K. (2018). Optimally myopic scheduling policy for downlink channels with imperfect state observation. *IEEE Transactions on Vehicular Technology*, *67*(7), 5856–5867.
62. Nguyen, T., Nguyen, V., Lee, J., & Kim, Y. (2019). Sum rate maximization for multi-user wireless powered IoT network with non-linear energy harvester: Time and power allocation. *IEEE Access*, *7*, 149698–149710.
63. Islam, S. M. R., Avazov, N., Dobre, O. A., & Kwak, K. (2017). Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Communications Surveys & Tutorials*, *19*(2), 721–742.
64. Thanh, P. D., Hoan, T. N. K., Vu-Van, H., & Koo, I. (2019). Efficient attack strategy for legitimate energy-powered eavesdropping in tactical cognitive radio networks. *Wireless Networks*, *25*(6), 3605–3622.



Hoang Thi Huong Giang received the B.E. degree in Electronics and Telecommunications Engineering from Ton Duc Thang University, Vietnam, in 2013, and the M.S. degree from the Graduate Institute of Digital Mechatronic Technology, College of Engineering, in Chinese Culture University, Taiwan, in 2015. She is currently pursuing her Ph.D. degree in Electrical and Electronic Engineering at University of Ulsan, Korea. Her current research interests include NOMA, reinforcement learning, and deep learning in wireless communications.

Tran Nhut Khai Hoan received the B.E. degree in electronics engineering from Can Tho University, Can Tho, Vietnam, in 2002, and the M.E. degree in Electronics Engineering from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2008. He received Ph.D. degree in Electrical Engineering from the University of Ulsan (UOU), Korea, in 2018. His research interests include cognitive radio and next generation wireless communication networks.

Insoo Koo received the B.E. degree from Kon-Kuk 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 the Ultrafast Fiber-Optic Networks Research Center, GIST, as a Research Professor. In 2003, he was a Visiting Scholar with the Royal Institute of Science and Technology, Stockholm, Sweden. In 2005, he joined the University of Ulsan, South Korea, where he is currently a Full Professor. His current research interests include next generation wireless communication systems and wireless sensor networks.