Energy Efficient Transmission in Underlay CR-NOMA Networks Enabled by Reinforcement Learning

Wei Liang[1], Soon Xin Ng[2], Jia Shi[3], Lixin Li[1], Dawei Wang[1,*]

[1] School of Information and Electronic, Northwestern Polytechnical University, No.127, YouYiXi Road, Xi'An, China, 710072
[2] School of Electrical and Computer Science, University of Southampton, UK
[3] State Key Lab. of Integrated Services Networks, Xidian University,No.2, TaiBaiNan Road ,Xi'An, China, 710071
*The corresponding author, email: wangdawei@nwpu.edu.cn; Liangwei@nwpu.edu.cn

Abstract: In order to improve the energy efficiency (EE) in the underlay cognitive radio (CR)networks, a power allocation strategy based on an actor-critic reinforcement learning is proposed, where a cluster of cognitive users (CUs) can simultaneously access to the same primary spectrum band under the interference constraints of the primary user (PU), by employing the non-orthogonal multiple access (NOMA) technique. In the proposed scheme, the optimization of the power allocation is formulated as a non-convex optimization problem. Additionally, the power allocation for different CUs is based on the actor-critic reinforcement learning model, in which the weighted data rate is set as the reward function,and the generated action strategy (i.e. the power allocation) is iteratively criticized and updated. Both the CU's spectral efficiency and the PU's interference constrains are considered in the training of the actor-critic reinforcement learning. Furthermore, the first order Taylor approximation as well as other manipulations are adopted to solve the power allocation optimization problem for the sake of considering the conventional channel conditions. According to the simulation results, we find that our scheme could achieve a higher spectral efficiency for the CUs compared to a benchmark scheme without learning process as well as the existing Q-learning based method, while the resultant interference affecting the PU transmission can be maintained at a given tolerated limit.

Keywords: cognitive radio network; non-orthogonal multiple access scheme; power allocation; reinforcement learning

I. INTRODUCTION

Recently, the growing demand for high data rates as well as the increased number of users leads to a large amount of the energy consumption. So that, energy consumption becomes the serious challenges in wireless communication according to the limited battery capacity. Specifically, the energy efficiency (EE) of cognitive radio (CR) networks has attracted the attentions of the researchers [1]–[3] . Therefore, the spectrum sharing strategies for the EE maximization in a CR network system have been investigated in [4]. Underlay CR network is capable of resolving spectrum scarcity problem, where CUs can transmit simultaneously with PUs by using the same frequency spectrum, under the constraints that the interference inflicted by CUs does not degrade PUs' quality of service [5]. On the other hand, non-orthogonal multiple access (NOMA) has emerged as a promising approach to improve

Received: Jul. 21, 2020
Revised: Aug. 21, 2020
Editor: Jie Hu

access efficiency of future wireless networks, and it has fundamentally reshaped the design of future multiple access (MA) techniques. As one of the most popular regimes of NOMA technique, the key idea is to explore the difference in power domain for MA while achieving non-orthogonality in other domains, such as time, frequency, etc. More specifically, in a downlink NOMA scheme, a base station (BS) can serve multiple users within the same time/frequency channel via different power allocation coefficients, where the users with poorer channel conditions are given more transmission power. Hence, NOMA encourages users to share the available spectrum, where MA interference is handled by applying advanced transceiver design, such as the successive interference cancellation as well as the superposition coding [6]. In this trend, by blending the concepts of NOMA and underlay CR, it develops a new novel scheme, namely underlay CR-NOMA, which can significantly improve network spectral efficiency. The benefits of using underlay CR-NOMA networks have been presented in [7]. Moreover, most existing works investigated the EE maximization in a CR network, the relative works related to the CR-NOMA system are less.

Recently, the use of the reinforcement learning (RL) has gained popularity in many fields. Specifically, each agent learns to change its own actions according to its study as well as the environment, where only one reward can be obtained and fed back for each action. Additionally, several RL technologies have also been implemented in wireless communications [8]–[11]. In[8], the authors have used the RL approach for channel selection aiming for reducing the amount of sensing required, which leads to the increased throughput and energy efficiency. Additionally, in [9], the Q-learning RL algorithm was employed to explore the spectrum sensing problem in the CR network. So far as known, very limited research efforts have been devoted to investigate the RL scheme based resource allocation in wireless systems, especially, none works have been done for that in underlay CR-NOMA systems.

In this paper, we focus on studying the resource allocation problem in the underlay CR-NOMA system, and propose the novel RL based power allocation algorithm. In particular, we assume that multiple CUs could access into the PU's spectrum simultaneously, by employing the NOMA technology. For the sake of explication, the main contributions of this paper can be summarized as follows.

- We formulate and analyze the novel power allocation problem for the underlay CR-NOMA system, where aiming to maximize the energy efficiency (EE) of the CUs while constraining the PU's minimum rate requirement, in order to achieve the energy Self-Sustainability (ESS) in the CR-NOMA system. In particular, all the CUs must be served at the same time on the same spectrum, which requires to balance the interference among the CUs themselves, and to control the interference from the CUs to the PU below a certain threshold.
- We propose the novel power allocation method, namely actor critic (AC)-RL based power allocation algorithm, which can efficiently managing the CUs' transmit power with the knowledge of the quantized channel state information (CSI) only. By employing the reinforcement learning framework, we design a novel actor-and-critic scheme with limited environment information, where the state transition probabilities and the rewards are unknown. Furthermore, the policy gradient based method is developed to learn and iteratively update the stochastic policies, which eventually result in a promising power allocation strategy for the cognitive network.
- We carry out the comprehensive performance evaluation for the proposed AC-RL power allocation algorithm. The simulation results show that the proposed algorithm can present a superior performance in terms of the achievable EE for cognitive network, in comparison with the Q-learning algorithm and the case without the learning process. Furthermore, we show that, the

In this paper, we have proposed a novel AC-RL algorithm for coordinating the power allocation among different CUs in underlay CR-NOMA system with the objective of maximizing the EE of the CUs while guaranteeing the PU's minimum data rate requirement.

AC-RL algorithm has a good convergence rate being similar to that of the Q learning algorithm which however demands much higher implementation complexity for acquiring the full knowledge of CSI.

Therefore, we may conclude that, for the underlay CR-NOMA system, our AC-RL based power allocation is a promising and practical solution for efficiently coordinating the transmit power for cognitive network with NOMA scheme employed. The rest of this paper is organized as follows. The system model is introduced in Section II and an energy efficient optimization problem is formulated in Section II. Then in Section III, a power allocation algorithm based on the reinforcement learning method is derived. Additionally, by using the first order Taylor approximation method, another power allocation algorithm is considered in Section IV. Simulation results and conclusions are finally presented in Section V and Section VI.

II. System Model

In this paper, we consider an underlay CR-NOMA system as described in Figure 1, consisting of a PU transmitter-receiver pair as well as a cognitive network. In particular, there are one cognitive transmitter, referred to as cognitive base station (CBS), and K number of CUs, whose indexes are included in the set $\mathcal{K}$. In this work, we focus on the downlink transmission from the CBS to the CUs. By employing NOMA technique, the K CUs are capable of accessing the same spectrum simultaneously. Therefore, different from the existing studies, instead of one CU, multiple CUs are allowed to access the PU's spectrum band as long as their interferences to the PU's transmission are below the pre-defined threshold I_{th}. In that case, the transmit power of CUs should be restricted according to the performance requirement of the primary transmission. Otherwise, the CUs need to remain silent.

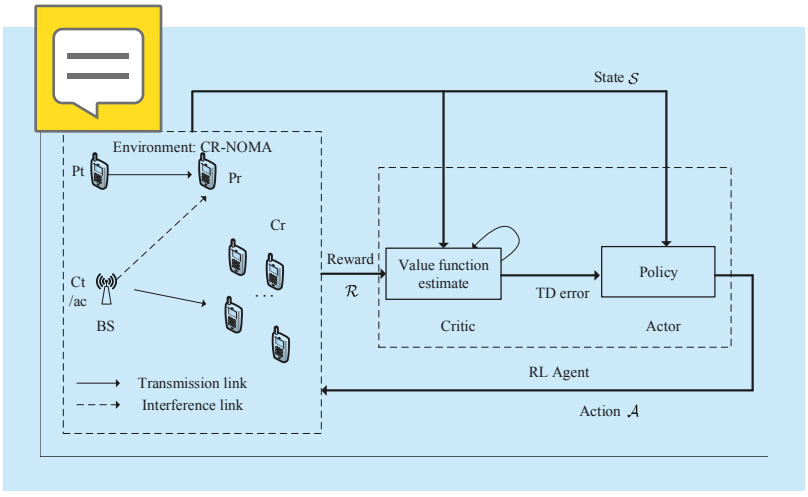


Fig. 1. *The schematic of AC RL based on a underlay CR-NOMA environment conceived.*

In the downlink cognitive transmission, the CBS transmits the superposed information of the K CUs together. For example, the received signal at CU_k is given by

$$Y_k^{CU} = h_k^{CU} \sum_{k=1}^{K} \sqrt{\alpha_k P_C} X_k + h_{PU,k} \sqrt{P_{PU}} X^{PU} + n_C, \tag{1}$$

where $k = 1, \ldots, K$ and α_k denotes the power allocation coefficient for CU_k.Note that,h_k^{CU} is the channel gain between the CBS and the CU_k,which is assumed to follow independent Rayleigh fading. Further, the noise term at CU_k is denoted by n_C, which follows Gaussian distribution with a zero mean and a noise variance of $N_0/2$ per dimension. Let us assume that, the transmitted symbol for CU_k satisfies $X_k \sim \mathcal{CN}(0,1)$, and that for PU also follows $X^{PU} \sim \mathcal{CN}(0,1)$. In the context of the underlay CR-NOMA system, the PU should satisfy its own requirement before allowing a group of CUs to access its spectrum, namely $R^{PU} = \log_2(1 + \frac{P_{PU} |h_{PU}|^2}{|h_{CU,PU}|^2 + \sigma^2}) \geq \eta R_{req}^{PU}$. Note that,$\eta$ is a positive scale factor, defined by the PU and can be adjusted according to different demands [1]. Note that, we assume the requirement R_{req}^{PU} or PU is set as the data rate without interference from cognitive transmission. Accordingly, by considering the interference constraints of the PU, for the downlink transmission, the transmit power at the CBS

should satisfy that

$$P_C \le \frac{1}{|h_{CU,PU}|^2}\left(\frac{P_{PU}|h_{PU}|^2}{2^{\eta R_{req}^{PU}}-1}-\sigma^2\right), \quad (2)$$

which ensures the unwanted CUs' interference at the PU is controlled by imposing a limit on the maximum transmitting power at the CBS. In (2),h_{PU} is the channel gain between the PU transmitter and receiver, and P_{PU} denotes the transmit power of the PU. Note that,$h_{CU,PU}$ is the equivalent channel gain between the interfering CUs and the PU. Note that Eq. (2) denotes the maximum permissible interference power at the PU's receiver, which ensures the unwanted CU's interference at the PU is controlled by imposing a limit on the maximum transmitting power for the CU's transmitter. Furthermore, based on Eq. (2) as well as considering the other system setups, the total transmit power at the CBS should be constrained as follow [12]:

$$P_C = \min\left[\frac{1}{|h_{CU,PU}|^2}\left(\frac{P_{PU}|h_{PU}|^2}{2^{\eta R_{req}^{PU}}-1}-\sigma^2\right), P_S\right], \quad (3)$$

where P_S is the maximum transmission power at the CBS.

Without loss of generality, let us assume that the CBS has the information of the ordering for the CUs' channel qualities. For example, we have that $|h_1|^2 \le |h_2|^2 \le \cdots \le |h_K|^2$. When employing the NOMA technique, the power allocation coefficients for the CUs should satisfy $\sum_k^K \alpha_k = 1$. Therefore, the successive interference cancellation (SIC) technique can be carried out at the CU receivers. Furthermore, let us assume that, the decoding order of the SIC at each receiver always follows the ascending order of CUs' channel conditions, and each CU's minimum SINR requirement (i.e. minimum decoding threshold) can be met [13]. Then, according to the principle of NOMA scheme, the spectral efficiency of each CU, such as CU_k, can be given by

$$SE_k^{CU} = \log_2\left(1+\frac{|h_k|^2 P_C\alpha_k}{|h_k|^2\sum_{k'\ne k,|h_{k'}|<|h_k|}\alpha_{k'}P_C + P_{PU}|h_{PU,k}|^2+\sigma^2}\right), \quad (4)$$

where the interference from the PU to the CUs $\{P_{PU}|h_{PU,k}|^2, \forall k\}$ is assumed to be known by the CBS. On the other hand, the spectral efficiency of CU_K, which has the best channel condition,

can be expressed as

$$SE_K^{CU} = \log_2\left(1+\frac{|h_K|^2 P_C}{P_{PU}|h_{PU,K}|+\sigma^2}\right). \quad (5)$$

In this work, it focuses on the power allocation in the underlay CR-NOMA system considered,

and aims to maximize the EE of the CUs. Therefore, energy efficiency is defined as the ratio of

the rate over the power [14], is shown as follows

$$EE^{CU} = \frac{SE^{CU}}{P_C}. \quad (6)$$

By observe the Eq. (6), maximize the energy efficiency of the system is identical to maximum

the spectral efficiency. When given the PU's predefined QoS requirement, the objective function

of maximizing the cognitive system's spectrum efficiency are formulated as:

$$\begin{aligned} &\max_{\mathbf{A}} \sum_k^K SE_k^{CU}, \\ &s.t.\ (a)\sum_k^K \alpha_k = 1, \\ &\quad (b) 0 \le \alpha_k \le 1, \forall k \in \mathcal{K}, \\ &\quad (c) Eq.(3). \end{aligned} \quad (7)$$

Above, let us define the variable vector of power allocation by $\mathbf{A} = [\alpha_1, \alpha_2, \ldots, \alpha_K]^T$.In the problem, constraint (a) ensures that all the transmit power available at the CBS will be allocated

to the CUs. Further, condition (b) means

that the transmit power for each CU is lower-bounded by 0, and is upper-bounded by full power. By contrast, condition (c) indicates that the transmit power of the CUs should satisfy the interference constraint of the PU. Furthermore, observed from the objective function, problem (7) is a non-convex problem, which is difficult to solve and requires to know the information of all CUs' CSI. However, in practical scenarios, it is hard to obtain the perfect knowledge of users' CSI due to the limited ability of feedback channels and dynamic communication environments. In our CR-NOMA system, we assume frequency division duplex (FDD) is used, in which case the channel information at the CBS can only be obtained by the feedback from CUs. With practical consideration of expensive feedback resources, limited amount of quantized channel gains are allowed in our system. Without loss of generality, each CU is assumed to have the knowledge of its channel gain, such that, CU k knows $|h_k|$.By employing Lloyds vector quantization (LVQ) approach [15], each CU feedbacks an index of B bits to the CBS. Note that, let us assume the feedback links are delay less and error-free [16].Hence, the CBS can obtain the quantized channel gains $\{\hat{h}_k, \forall k\}$ when receiving the indexes.Therefore, in contrast to the existing literatures requiring the knowledge of full CSI between the CBS and CUs, in this paper we assume that, the CBS has the information about *quantized channel gains only*, *such that* $\{|\hat{h}_k| = [|h_k|], \forall k\}$(note that $[x]$ is the nearest integer of x),which can significantly reduce the burden on feedback channel [7]. In that case, conventional approaches, such as convex optimization, will be impossible to solve the problem (7) in the absence of full CSI. By contrast, deep learning approach is able to self-adapt wireless systems with inaccurate channel information, and can be employed to find promising power allocation strategy, which will be descried in Section III. Apparently, it is extremely difficult to find a promising solution for problem 7 in the absence of full CSI. Hence, so far as we know, there is no published studies that have applied the conventional approaches, such as convex optimization, to solving resource allocation problems in NOMA systems upon having the quantized CSI only. By contrast, reinforcement learning (RL) approach is able to self-adapt wireless systems with inaccurate channel information, and can be employed to find promising power allocation strategy. In particular, by using the actor-critic (AC) RL approach, the actor process generates continuous policy and the critic process evaluates the actor, forming a trial-and-error learning process, which can dynamically adjust the strategy toward a promising result as long as the objective is properly set up. Therefore, the above features guarantee that the AC based RL approach perfectly match our power allocation problem especially for lack of full CSI.

III. Reinforcement Learning Based Power Allocation Algorithm

In this section, we propose the AC-RL based power allocation algorithm, which aims to solve problem (7). Let us first analyze our optimization problem, and then discuss how to apply the AC-RL approach.

3.1 Problem analysis & general theory of AC-RL

In the underlay CR-NOMA system, our power allocation problem in (7) can be regarded as a discrete-time Markov decision process (MDP) with continuous states and actions. To address the practical implementation, traditional convex optimization approach cannot find the optimal solution, even a sub-optimal solution, to problem (7) with the knowledge of quantized channel gains. On the contrary, The model-free RL framework can be applied to our problem, since it only needs to know partial information about wireless environment to derive the state (i.e. power allocation strategy) transition probability and expected rewards of the states.

By exploiting the features of general RL framework, we adopt the AC-RL approach for

our problem, where the schematic has been shown in Figure 1. In particular, the agent learns the optimum policy as well as the corresponding value function via the interactions with the environment. Note that, there are two parts in the agent, namely the actor (i.e. policy) and the critic (i.e. estimated value function). Specifically, the aim of the actor is to define the parameterized policy and also to generate the actions based on the observed environment state. On the other hand, the critic will evaluate and criticize the current policy by processing the rewards received from the environment. Then, the actor is able to use the output from the critic so that its policy parameters are updated, after the approximation of the value function. Hence, the output of the critic is proportional to the temporal difference (TD) error which indicates whether the results get better or worse than the expectation. Additionally, it can also be used to adjust both the actor and the critic for the sake of reducing the TD error. As proved in [9], the actor-critic method has a good property of quick convergence when learning continuous-valued stochastic policies. In order to maximize the final reward, in the AC-RL approach, it maps the states to achieve the optimal behaviors by using the trial and error learning experiences. Moreover, for the sake of accelerating the learning process, the AC-RL needs to learn the value and strategy functions separately. Additionally, the final obtained strategy aims to maximize the objective function as described in Eq.(7).

3.2 Principles of AC-RL based power allocation algorithm

Let us first introduce the following definitions. When considering the AC-RL for the underlay CR-NOMA system, the factor matrix is given by

$$M \triangleq \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}, \tag{8}$$

where $\mathcal{S}$ denotes the state space,$\mathcal{A}$ indicates the action space,$\mathcal{P}$ represents the probability of transition state space, and $\mathcal{R}$ denotes the reward function. The key idea of the AC-RL approach is to learn how to map $\mathcal{S}$ to optimal states via the trial-and-error learning experience, aiming to maximize the accumulated sum of the rewards for the CUs in time.

In the AC-RL approach, the actions are obtained at the beginning of each step by learning the existing environment state. For our CR-NOMA system, the state is determined by the Signalto-Interference-and-Noise Ratio (SINR) of CU, such as, the state set S_t during step t can be defined as:

$$S_t = \{\hat{\gamma}_1^{CU}(t), \hat{\gamma}_2^{CU}(t), \cdots, \hat{\gamma}_K^{CU}(t)\}, \tag{9}$$

where $t = \{1, 2, \cdots, T_{max}\}$ with T_{max} being the maximum number of the steps, and $S_t \in \mathcal{S}$.Let us define that

$$\hat{\gamma}_k^{CU} = \frac{|\hat{h}_k|^2 P_C \alpha_k}{|\hat{h}_k|^2 \sum_{k'=1, k' \neq k, |\hat{h}_{k'}| < |\hat{h}_k|}^{K} \alpha_{k'} P_C + P_{PU} |\hat{h}_{PU,k}|^2 + \sigma^2}, \tag{10}$$

where $k = 1, 2, \ldots, K$.Note that, the first term of denominator in (10) vanishes for $k = K$.

In our problem, the agent decides how much power could be allocated to each CU during every step. Additionally, the action set A_t for step t can be defined as:

$$A_t = \{\alpha_1(t), \alpha_2(t), \cdots, \alpha_K(t)\}. \tag{11}$$

Again, $\alpha_k(t)$ is the power allocation factor for CU k.Specifically, when an action A_t is executed and the system is at state s_t the corresponding reward R_t of the system can be obtained by computing Eq. (4) or Eq. (5), namely $R(A_t) = \sum_{k=1}^{K} SE_k^{CU}(\alpha_k(t))$. Since the action space $\mathcal{A}$ is continuous, assume that, the action A_t is dependent on the stochastic policy $\pi(a|s) = \Pr(A_t = a | S_t = s)$ which is a mapping from the network situations to the probability of taking actions.More specifically, an agent can evaluate and change its policy according to the value function which is defined as the expected value of cumulative discounted rewards received over the entire process. In our system, it employs the state-action value

function which is the expected value of accumulated rewards starting from the current state and action, and then, uses the given policy to select the next action. Thus, the state-action value function can be computed as:

$$\mathcal{V}\pi(s \mid a) = \mathbf{E}\left(\sum_{t=1} \beta_t R_t \mid S_t, A_t, \pi \right). \quad (12)$$

Above,$\mathbf{E}(\cdot)$ indicates the expectation, and β_t is (belonging to (0, 1)) the discount factor which can weigh the myopic or foresighted decisions. As described by Eq. (12), the state-value function is used to quantify the value of executing the action set A_t at the state set S_t and to make a decision according to the given policy π.Moreover, our application has infinite number of states and action spaces, since the transmit power of CUs is continuous variable. On the other hand, the perfect channel information of CUs is not available. Hence, it is impractical to compute all the value functions for every possible state-action pair.

With the above discussions, let us now introduce the principles of the proposed AC-RL based power allocation algorithm, which is given by Algorithm 1. At the start of the algorithm, the agent observes the environment and generates an action based on the Gaussian policy and also the immediate reward. In the proposed AC-RL approach, the actions are obtained at the beginning of each step by learning the existing environment state, in which the quantized channel gains can be fedback to the CBS. Note that, we employ the equal assignment of transmit power for initializing the actor set A_0, which in turn gives the initialized state set S_0.Note that, the critic part can estimate the value function and calculate the TD errors. However, the critic updates its parameters in proportion to TD error and eligibility trace. The actor part uses the outcome from the critic part to compute the advantage function and estimate the policy. After that its parameters are updated towards the final policy.

Algorithm 1. Principles of AC-RL algorithm.

Initialization:

Set the learning rate for the actor space η_a and that for the critic space η_c;Set the discount factor β;

Initialize $A_0 = \{\alpha_k(0) = 1/K, \forall k\}$,and compute $S_0 = \{\hat{\gamma}_k(0), \forall k \,|\, A_0\}$;

Set the initial value function as $Q(S_0, A_0) = 0$;

Input and set the the basic elements.

Set the action set A_t as $A_t = \{\alpha_1(t), \alpha_2(t), ..., \alpha_K(t)\}$;

Set the state set S_t as $S_t = \{\hat{\gamma}_k(t), \forall k \,|\, A_t\}$;

While $t \le T_{max}$

Calculate immediate reward:$R_t = \sum_{k=1}^{K} SE_k^{CU}(\alpha_k(t))$;

Calculate TD error: *using* $\delta_t = R_t + \beta Q(S_t, A_t) - Q(S_{t-1}, A_{t-1})$;

Select the action based on the value function: *update as* $Q(S_{t+1}, A_{t+1}) \leftarrow Q(S_t, A_t) + \eta_c \delta_t$;

Update the strategy function:$\pi(S_{t+1}, A_{t+1}) \leftarrow \pi(S_t, A_t) - \eta_a \delta_t$;

Update $t \leftarrow t+1$;

End

IV. Optimal Power Allocation

Different to the AC-RL power allocation method employed in the Section III. In this section, we use the optimization method to solve the objection function of Eq. (7) in the perfect channelcondition. As described in Section II, the objective function of Eq. (7) is a monotonically and non-decreasing log(·) function, which cannot affect the convexity of this problem. Apparently,Eq. (7) is not a convex problem. Then we obtain the following propositions.

4.1 The proposed optimal power allocation algorithm

Proposition 1: The optimization problem in (7) for power allocation is quasi-concave problem.

Proof :A maximization optimization problem is quasi-concave when the objective function is quasi-concave and the constraints are convex [17]. The conditions (a) and (b) in Problem (7) are convex, since both of them are linear.

It is observed that Problem (7) is NP-hard even though the existing of the convex constraints of (7)(a) and (7)(b) based on the the-

ory in [18]. Hence, Problem (7) can be transformed into the sequence of linear programs and develop a customized low-complexity polynomial algorithm, in order to get its local optimal solution.

Let us first transform the origin problem of (7) into

$$\max_{\mathcal{A}} \Big(\prod_{k\in\mathcal{K}} R_k^{CU}\Big)^{\frac{k_l}{|K'|}} \tag{13}$$

where R_k^{CU} has been defined in Eq. (4) and $|K'|$ s the number of CUs within a specific group. Based on the objective function of Eq. (7), we introduce two slack vectors $\mathbf{t}\in\mathbb{R}^{|K'|}$ and $\mathbf{b}\in\mathbb{R}_+^{|K'|-1}$ with the elements t_k and b_k,respectively. Then Eq. (13) can be rewritten as

$$\max_{\mathcal{A},b,t} \Big(\prod_{k\in\mathcal{K}} t_k\Big)^{\frac{1}{|K'|}}, \tag{14}$$

$$\begin{aligned}
s.t.\quad & (a)\,|h_{CU}|^2 P_C\alpha_k \ge (t_{k'}-1)b_{k'}, \forall k, \forall k' \in K' \\
& (b)\,|h_{CU}|^2 P_C^2 \sum_{p=k+1}^{K}\alpha_p + P_p |h_{PU,k}|^2 \\
& \quad + \sigma^2 \le b_{k'}, \forall k, \forall k' \in K' \\
& (c) 1 + \frac{|h_{CU}|^2 P_C\alpha_K}{P_p |h_{PU,k}|^2 + \sigma^2} \ge t_{k'}, \forall k, \forall k' \in K' \\
& (d)\alpha_1 \ge \cdots \ge \alpha_k \ge \cdots \ge \alpha_K, \\
& (e) \sim 6(a)\ 6(b)
\end{aligned}$$

Additionally, the objective function of Eq. (14) is a geometric mean function, which is concave and increasing, respect to vector **t**. Constraints (b) and (c) of problem (14) are derived from Eq. (5) and Eq. (4). The original power allocation problem of (7) is equivalent to the transformation problem of (14). However, the objective function of Eq. (14) is still non-convex and intractable due to the bi-linear term in constraint (a). Therefore, we need to implement the further constraint (a) of (14) can be rewritten as $|h_{CU}|^2 P_C\alpha_k \ge t_{k'}b_{k'} - b_{k'}$. Hence, the transformation of this bi-linear term $t_{k'}b_{k'}$ can be computed as

$$t_{k'}b_{k'} = \frac{1}{4}[(t_{k'}+b_{k'})^2 - (t_{k'}-b_{k'})^2]. \tag{15}$$

It is observed that Eq. (15) is a non-convex function, and the term $(t_{k'}+b_{k'})^2 - (t_{k'}-b_{k'})^2$ of Eq. (15) is a difference between these two convex functions. Moreover, the second quadratic term on the right side of Eq. (15) can be approximated by its low bound, which is linear and is obtained by using Taylor series expansion, based on the convexity of itself. Then, we obtain the following first order Taylor approximation around the points $(t_{k'}^{(n)}, b_{k'}^{(n)})$ and is given by

$$\begin{aligned}
f(t_{k'}, b_{k'}) &= (t_{k'} - b_{k'})^2 \\
&\ge (t_{k'}^{(n)} - b_{k'}^{(n)})^2 + 2(t_{k'}^{(n)} - b_{k'}^{(n)}) \\
&\quad \times (t_{k'} - t_{k'}^{(n)} - b_{k'} + b_{k'}^{(n)}) \\
&\triangleq g(t_{k'}, t_{k'}^{(n)}, b_{k'}, b_{k'}^{(n)})
\end{aligned} \tag{16}$$

where $(t_{k'}^{(n)}, b_{k'}^{(n)})$ can be updated by $(t_{k'}^{(n+1)}, b_{k'}^{(n+1)})$). However, in every iteration the approximated inequality satisfies the following conditions [19], [20], given by

$$f(t_{k'}, b_{k'}) \ge g(t_{k'}, t_{k'}^n bk', b_{k'}^n), \tag{17}$$

$$f(t_{k'}^n, b_{k'}^n) = g(t_{k'}^n, t_{k'}^n, b_{k'}^n, b_{k'}^n), \tag{18}$$

$$f(t_{k'}, bk')\big|_{(t_{k'}^n, \beta_{k'}^n)} = \nabla g(t_{k'}, t_{k'}^n, b_{k'}, b_{k'}^n)\big|_{(t_{k'}^n, b_{k'}^n)} \tag{19}$$

Above, ∇f denotes the gradient of f.Further, the auxiliary variables $(t_{k'}, b_{k'})$ can be updated by $t_{k'}^{(n+1)} = t_{k'}^{(n)}$ and $b_{k'}^{(n+1)} = b_{k'}^{(n)}$,when condition (19) is satisfied in each iteration.

Hence, the problem of (14) can be converted to convex problem at the n th iteration by replacing $(t_{k'} - b_{k'})^2$ with $g(t_{k'}, t_{k'}^{(n)}, b_{k'}, b_{k'}^{(n)})$,and it is given by

$$\max_{\mathbf{A},\mathbf{b},\mathbf{t}} \Big(\prod_{k\in\mathcal{K}} t_k\Big)^{\frac{1}{|K'|}}. \tag{20}$$

$$\begin{aligned}
s.t.\quad & (a) 4(|h_{CU}|^2 P_C\alpha_k + b_k) + g(t_k, t_k^{(n)}, b_k, b_k^{(n)}) \\
& \ge (t_k + b_k)^2, s_i = 1, \ldots, K-1, \\
& (b)\ 6(a),\ 6(b),\ 14(a),\ 14(b),\ 14(c).
\end{aligned}$$

Known from [21], Problem (20) is convex at the nth iteration round points $(t_k^{(n)}, b_k^{(n)})$. Therefore, we propose a successive iterative algorithm (SIA) to locally solve the approximation of Problem(20), which is related to the sequential convex approximation method [19].

As a result, the proposed optimal power allocation algorithm can be described in Algo-

rithm 2.

4.2 Convergence analysis

In this section, we analyze the convergence of the proposed SIA algorithm as shown in Algorithm 2. Let us define the feasible set as $\mathcal{Z}^{(n)}$ and the sequence of variables set as $\mathcal{B}^{(n)}=[t_k^{(n)},b_k^{(n)}]$,for all s_i in the n th iteration of the problem in Eq. (20). Furthermore, we assume that feasible set of the problem in Eq. (14) as $\mathcal{Z}$. Then, we readily obtain the Proposition 3.

Proposition 2: The sequence of variables $\mathcal{B}^{(n)}$ belongs to $\mathcal{Z}$.By defining the objective value of Eq. (20) as $\mathcal{F}^{(n)}$,we have $\mathcal{F}^{(n+1)}\geq\mathcal{F}^{(n)}$,and the algorithm converges.

Proof : In order to prove $\mathcal{B}^{(n)}$ belongs to $\mathcal{Z}$,we only need to prove $\mathcal{Z}^{(n)}$ is included in $\mathcal{Z}$.Since $\mathcal{B}^{(n)}$ is the optimal solution for Problem (20), the sequence of variables $\mathcal{B}^{(n)}$ belongs to $\mathcal{Z}^{(n)}$, such that $\mathcal{B}^{(n)}\in\mathcal{Z}^{(n)}$.We assume that the variables $t_k^{(n)^*}$ and $b_k^{(n)^*}$ are the optimal solutions for Problem (20) at the n th iteration. Then $t_l^{(n)^*}$ and $b_l^{(n)^*}$ belong to the feasible set $\mathcal{Z}^{(n)}$ and they satisfy the all the constraints of Problem (20). We can clearly prove that $t_k^{(n)^*}$ and $b_k^{(n)^*}$ also satisfy all constraints in Problem (14). That is because we employ the convex approximation $g(t_k,t_k^{(n)},b_k,b_k^{(n)})$ for $(t_k-b_k)^2$ and the conditions in Eq.(19). Hence, $\mathcal{Z}^{(n)}$ is included in $\mathcal{Z}$ and the sequence of variables $\mathcal{B}^{(n)}$ belongs to $\mathcal{Z}$.Since the point $(t_k^{(n+1)},b_k^{(n+1)})$ at the $n+1$ th iteration is updated by the sequence of variables $\mathcal{B}^{(n)}$ at the n th iteration, the point $(t_k^{(n+1)},b_k^{(n+1)})$ belongs to the feasible set of Problem (20) at the $n+1$ th iteration.

Algorithm 2. Proposed successive iterative algorithm (SIA) for solving the problem of Eq. (20).

1 *Initial state*: Set $(t_k^{(n)},b_k^{(n)})$ of Problem (20) while $n=0$.

2 *Repeat*

1) Solve Problem (20) with $(t_{k'}^{(n)},b_{k'}^{(n)})$ to get the optimal values (t_k^*,b_k^*).

2) Set $n\rightarrow n+1$, update $(t_{k'}^{(n+1)},b_{k'}^{(n+1)})=(b_k^*,t_k^*)$.

until Converge.

Output the optimal solution of Problem (20).

Furthermore, Problem (20) have been solved by using the proposed SIA of Algorithm 2, which generates a non-decreasing sequence of objective values, i.e.$\mathcal{F}^{(n+1)}\geq\mathcal{F}^{(n)}$. According to the proposition 2, the proposed iterative algorithm converges to a finite value. However, we cannot prove the algorithm converges to the global optimum solution, due to the original problem of (14) is non-convex.

In the following proposition, we will prove that our proposed algorithm satisfies the Karush-Kuhn-Tucker (KKT) conditions.

Proposition 3: Algorithm 2 converges to the KKT point of Problem (14) when it achieves the pre-defined accuracy.

Proof : Let us define an accumulation point $\mathcal{B}^*$ of the sequence of variables $\mathcal{B}^{(n)}$ at convergence. We first need to simplify the constraints in the original problem Eq. (14) and the approximation problem Eq. (20). Since the only difference in the constraints of both problems is the first constraint, we set the remain constraints as a function $G_k(\mathcal{B})$.The approximation problem (20) satisfies the Slater's conditions since it is strictly convex program [21]. Then, the KKT conditions of Eq. (20) at the accumulation point $\mathcal{B}^*$ can be expressed as following with the optimal dual variables λ_k^*,μ_k^*

$$\nabla f_0(t_k^*)-\sum_{k=1}^{K-1}\lambda_k^*\left(g(t_k^*,t_k^{(*)},\beta_k^*,\beta_k^{(*)})-A_k^*\right)+\sum_{k=1}^{K}\mu_k^*G_k(\mathcal{B}^*)=0 \tag{21}$$

$$\lambda_k^*\left(g(t_k^*,t_k^{(*)},\beta_k^*,\beta_k^{(*)})-A_k^*\right)=0 \tag{22}$$

$$\mu_k^*G_k(\mathcal{B}^*)=0. \tag{23}$$

where $A_k^*=(t_k^*+\beta_k^*)^2-4(|h_{Ct,Cr_k}|^2P_c\alpha_k^*+\beta_k^*)$. According to the equality condition Eq. (16), we know $(t_k^*-\beta_k^*)^2=f(t_k^*,\beta_k^*)=g(t_k^*,t_k^{(*)},\beta_k^*,\beta_k^{(*)})$. Then replacing $g(t_k^*,t_k^{(*)},\beta_k^*,\beta_k^{(*)})$ by $(t_k^*-\beta_k^*)^2$, the above KKT conditions can be reformulated as

$$\nabla f_0(t_k^*)-\sum_{k=1}^{K-1}\lambda_k^*\left((t_k^*-\beta_k^*)^2-A_k^*\right)+\sum_{k=1}^{K}\mu_k^*G_k(\mathcal{B}^*)=0 \tag{24}$$

$$\lambda_k^*\left((t_k^*-\beta_k^*)^2-A_k^*\right)=0 \quad (25)$$

$$\mu_k^* G_k(\mathcal{B}^*)=0. \quad (26)$$

It is clearly shown that the reformulated KKT conditions above are also the KKT conditions of the original problem of Eq. (14). Therefore, the convergent point $\mathcal{B}^*$ generated by the proposed algorithm is a KKT point of Eq. (14).

V. Simulations Results and Analysis

In our considered underlay CR network, the CBS is located at the origin of a disc $\mathcal{D}_i$ with radius one. Assume that, the CUs are randomly distributed in the disc $\mathcal{D}_c$.The downlink transmission from the CBS to the K CUs is based on NOMA technique. The PU stays in the margin of a disc $\mathcal{D}_o$ with radius two, and is outside of the $\mathcal{D}_o$.Moreover, we assume that Pt and Pr are located at the opposite sides of the square with the distance of four. By considering the tradeoff between complexity and performance, in our simulation we set $B=3$ bits for acquiring relative precise channel quantization results. Furthermore, let us set up: 1) the number of CUs is $K=6$, 2) the transmit power for PU is $P_{PU}=15$ W, 3) the discount factor is $\beta=0.9$,4) the scale factor is given as $\eta=0.8$, 5) the learning rate of the actor is $\eta_a=0.01$, and 6) the learning rate of the critic as $\eta_c=0.001$.

Figure 2 shows the performance of the SE of the CUs versus the time index. Note that, we use the fixed actor's learning rate as well as the critic's learning rate. As known, Q-learning is a value-based RL method which has been widely used to solve the stochastic optimization problems[8]. Hence, in the figure we also evaluate the Q-learning based power allocation scheme. By using the Q-learning, the continuous-valued states and actions have to be quantized and the real value is replaced by an approximation from a finite number of discrete-values. In contrast to our AC-RL algorithm, the Q-learning based power allocation algorithm requires to know the full knowledge of instantaneous CSI of the CUs. From Figure 2, it confirms that our power allocation based on the AC-RL algorithm successfully finds an optimal tracking controller for underlay CR system.

Figure 3 compares the SE performance of the proposed AC-RL with the other algorithms. In the figure, we have employed a benchmark strategy, referred to as the case without learning process[3], in which a fixed fractional power allocation method is used. In particular, the transmit power of CU k is allocated by: $\alpha_k=\frac{P_C\,|\,h_k\,|^{-\beta_F}}{\sum_k^K |h_k\,|^{-\beta_F}}$, where β_F is the decay factor, namely $0\le\beta_F\le 1$.Without loss of generality, we set $\beta_F=0.8$ in our paper, and this setup retains the same during the entire transmission period considered. Note that, the case without learning process requires the CBS to have the full knowledge of CSI for the CUs' links. Observed from the figure, the SE of the CUs increases in majority cases as the transmit power available for cognitive network gets bigger, i.e. the ratio P_S/P_{PU} gets bigger. The reason behind is that, the SE of the CUs is positively related to their transmit power available, which can be known from (4) and (5). However, we should aware that, if keep increasing the transmit power for the cognitive

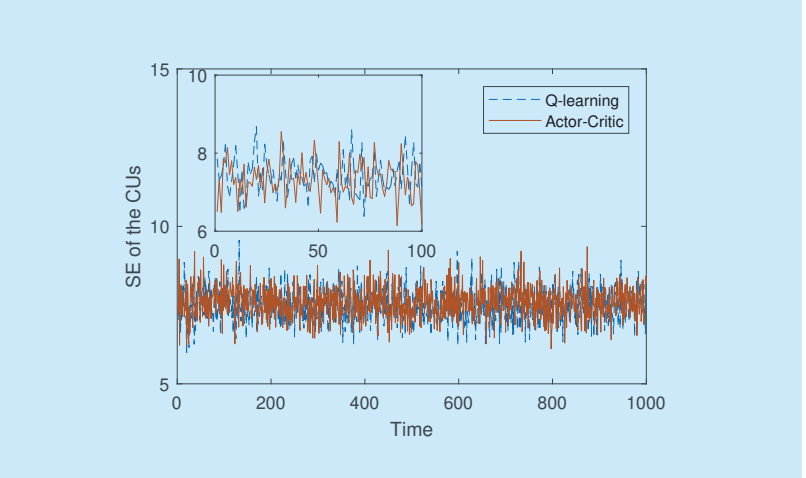


Fig. 2. *The performance of the SE of the CUs versus the time index.*

network, it will inevitably lead to the increase of interference from the cognitive transmissions to the primary transmission. In that case, this may violate the constraint of interference threshold at the PU, i.e. the constraint in Eq. (2), which in turn causes the result of a reduced number of CUs accessing to the PU's spectrum. Therefore, we may conclude that, in practical systems one should consider a trade-off between the maximization of the SE of cognitive network and maximization of the number of accessing CUs.

Furthermore, as shown in Figure 3, we observe that, the SE achieved by our AC-RL algorithm may decrease and/or keep constant as the ratio P_S / P_{PU} keep increasing, especially when $P_S / P_{PU} \geq 1$,i.e. the total transmit power for cognitive network is greater than that for primary network. This observation implies that, our AC-RL may result in a power allocation strategy that assigns too much transmit power to the CUs with strong channel conditions, which causes heavy interference among the CUs. Moreover, in the figure, an overlap between the Q-learning approach and our AC-RL approach can be observed for the scenario of $P_S / P_{PU} \geq 1$.In this case, the actor part of AC-RL at that situation can be treated as a fixed strategy and cannot predict the errors, which results in the performance similar to that of the Q learning. Note that, the performance saturation for our AC-RL is also due to the lack of knowledge about instantaneous CSI. As seen from Figure 3, a much higher SE performance can be obtained by employing the AC-RL algorithm, in comparison to that by the case without learning process, although our AC-RL approach has the quantized CSI of the CUs only.

As we mentioned in Section II, the expression of EE is shown in Eq.(6). Figure 4 compares the convergence performance of the proposed AC-RL algorithm with the Q learning approach, which is a value-based RL method for solving the stochastic optimization problems Eq. (7). By using the Q learning algorithm, it requires the CBS to know the full knowledge of CUs' CSI. Seen from the figure, both the algorithms can converge within a relatively small time steps, while our proposed AC-RL requires slightly more time steps. Note that, during each step of the AC-RL algorithm, the state and action spaces contains six elements corresponding the six CUs. Further, the state space is computed based on the knowledge of quantized CSI, which can be acquired by feedback links. Moreover, in comparison with the Q learning, the curve for

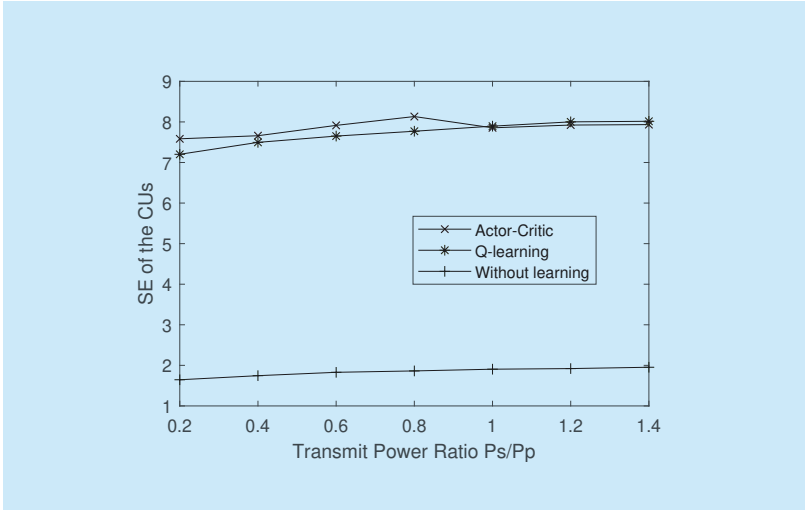


Fig. 3. *Comparision of the CU's SE perofrmance when varying the transmit power ratio between the maximum transmit power at CBs an PU,P_S / P_{PU}.*

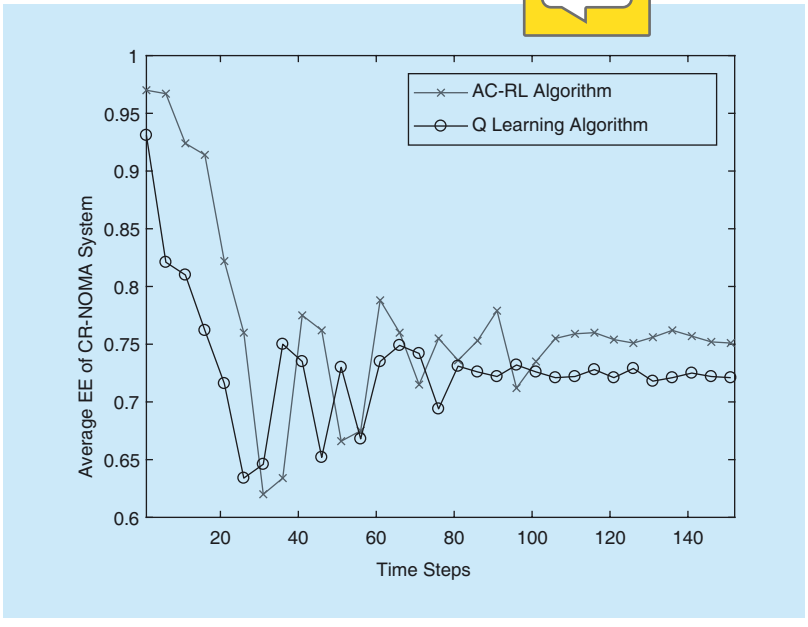


Fig. 4. *The average EE performance of the cognitive network versus time steps.*

our AC-RL algorithm clearly appears more fluctuating during the whole learning process. The main reason behind is that, our AC-RL algorithm only uses the quantized CSI, while the Q learning has the full knowledge of CSI. This can certainly confirms the effectiveness of our proposed AC-RL algorithm in terms of convergence rate. Therefore, we may conclude that, the convergence speed of a RL approach for resource allocation heavily determined by the amount of information about wireless environment observed.

According to the action space definition, each CU needs to update its transmission power level. Therefore, for each state, the action of the CU is a function of the transmission power level. By contrast, the complexity of the Q learning is $\mathcal{O}(K^K)$,where requiring to search all possible combinations of the actions from the Q table. Moreover, based on the actor part, AC-RL algorithm occurs less error compared to the Q-learning algorithm during the training process and with less complexity. In general, we may conclude that, for the underlay CR-NOMA system, our AC-RL based power allocation is a promising and practical solution for efficiently coordinating the transmit power for the CUs to maximize the SE according to adaptively training from wireless environment, in the absence of perfect channel information. At last, referring to [22], the computational complexity of the training AC-RL algorithm is $\mathcal{O}(K^2)$,and that of the Q learning is $\mathcal{O}(K^K)$,where requiring to search all possible combinations of the actions from the Q table. Moreover, the AC-RL algorithm occurs less error compared to the Q learning algorithm during the training process. Hence, we may conclude that, for our CR-NOMA system, the ACRL is a promising and practical solution for efficiently coordinating the transmit power for the cognitive network in the absence of perfect channel information, where the size of the state function of the CUs as well as their action must be considered. Additionally, by using the SIA algorithm which is to exhaustive search of the optimal solution and the relative complexity is $\mathcal{O}(K!2^{2K})$[23]. Thus, by using the AC-RL algorithm the computational complexity is less by comparing with the SIA algorithm.

VI. Conclusions

In this paper, we have proposed a novel AC-RL algorithm for coordinating the power allocation among different CUs in underlay CR-NOMA system with the objective of maximizing the EE of the CUs while guaranteeing the PU's minimum data rate requirement. Note that, our ACRL algorithm can be operated upon knowing the quantized channel gains only. With the aid of setting up the reward function as the weighted data rate, the proposed AC-RL approach iteratively result in the action of the power allocation strategy criticized by the TD error. Numerical results have proved that our proposed scheme significantly outperforms the benchmark scheme without considering learning process, and can achieve similar or superior performance compared to the Q learning approach which however demanding much higher implementation complexity. By comparing with the traditional optimization solution, our AC-RL algorithm has less complexity and can also be employed in the quantized channel condition scenario. Overall, we may conclude that, for the underlay CR-NOMA system, our AC-RL algorithm is a promising and practical solution for efficiently coordinating the transmit power for the cognitive network.

References

[1] L. Hu, R. Shi, M. Mao, Z. Chen, H. Zhou, and W. Li, "Optimal energy-efficient transmission for hybrid spectrum sharing in cooperative cognitive radio networks," *China Communications*, vol. 16, no. 6, pp. 150–161, 2019.

[2] B. Han, M. Zeng, Q. Guo, H. Jiang, Q. Zhang, and L. Feng, "Energy-efficient sensing and transmission for multi-hop relay cognitive radio sensor networks," *China Communications*, vol. 15, no. 9, pp. 106–117, 2018.

[3] X. Zhou, M. Sun, G. Y. Li, and B. Fred Juang, "Intelligent wireless communications enabled by cognitive radio and machine learning," *China*

Communications, vol. 15, no. 12, pp. 16–48, 2018.

[4] N. Zhao, F. R. Yu, H. Sun, and M. Li, "Adaptive power allocation schemes for spectrum sharing in interference-alignmentbased cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3700–3714, 2016.

[5] W. Liang, S. X. Ng, and L. Hanzo, "Cooperative overlay spectrum access in cognitive radio networks," *IEEE Communications Surveys Tutorials*, vol. 19, pp. 1924–1944, thirdquarter 2017.

[6] Z. Ding, P. Fan, and H. Poor, "Impact of User Pairing on 5G Non-Orthogonal Multiple Access Downlink Transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 6010–6023, Aug 2016.

[7] S. Arzykulov, T. A. Tsiftsis, G. Nauryzbayev, and M. Abdallah, "Outage performance of cooperative underlay cr-noma with imperfect csi," *IEEE Communications Letters*, vol. 23, pp. 176–179, Jan 2018.

[8] V. Raj, I. Dias, T. Tholeti, and S. Kalyani, "Spectrum access in cognitive radio using a two-stage reinforcement learning approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, pp. 20–34, Feb 2018.

[9] K. A. M, F. Hu, and S. Kumar, "Intelligent spectrum management based on transfer actor-critic learning for rateless transmissions in cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 17, pp. 1204–1215, May 2018.

[10] W. Lee, "Resource allocation for multi-channel underlay cognitive radio network based on deep neural network," *IEEE Communications Letters*, vol. 22, pp. 1942–1945, Sep. 2018.

[11] X. Qiu, L. Liu, W. Chen, Z. Hong, and Z. Zheng, "Online deep reinforcement learning for computation offloading in blockchain-empowered mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 8050–8062, Aug 2019.

[12] J. Zou, H. Xiong, D. Wang, and C. W. Chen, "Optimal power allocation for hybrid overlay/underlay spectrum sharing in multiband cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 62, pp. 1827–1837, May 2013.

[13] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 2852–2857, March 2017.

[14] L. Sboui, Z. Rezki, and M. Alouini, "Energy-efficient power allocation for underlay cognitive radio systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 3, pp. 273–283, 2015.

[15] V. Lau, Y. Liu, and T. A. Chen, "On the design of mimo block-fading channels with feedback-link capacity constraint," *IEEE Transactions on Communications*, vol. 52, pp. 62–70, Jan. 2004.

[16] W. Xu, X. Dong, and W. S. Liu, "Mimo relaying broadcast channels with linear precoding and quantized channel state information feedback," *IEEE Transactions on Signal Processing*, vol. 58, pp. 5233–5245, Oct. 2010.

[17] S. Boyd and L. Vandenberghe, *Convec Optimization*. Cambridge University Press, 2004.

[18] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process*, vol. 2, pp. 57–73, Jan 2008.

[19] A. B. T. A. Beck and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Opt.*, vol. 47, no. 1, pp. 29–51, 2010.

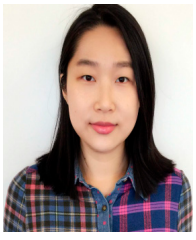
[20] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in nonorthogonal multiple access systems," *IEEE Trans. Signal Process*, vol. 64, pp. 76–88, Jan 2016.

[21] S. Boyd and L. Vandenberghe, "Covex optimization," *Cambridge Univ. Press*, 2004.

[22] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected uavs: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2019.

[23] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Transactions on Communications*, vol. 65, pp. 5319–5332, Dec 2017.

Biographies



Wei Liang, received her M.Sc. and Ph.D. degree in wireless communication at University of Southampton, Southampton, U.K in 2010 and 2015, respectively. She was a Postdoctoral Research Fellow with Lancaster University, UK, during 2015-2018. She is currently an Associate Professor in the School of Electronics and Information, NPU. Her research interests include adaptive coded modulation, network coding, matching theory, game theory, cooperative communication, cognitive radio network, Non-orthogonal multiple access scheme and Mobile edge computing etc.

Soon Xin Ng (S'99-M'03-SM'08), received the B.Eng. degree (First class) in electronic engineering and the Ph.D. degree in telecommunications from the University of Southampton, Southampton, U.K., in 1999 and 2002, respectively.

From 2003 to 2006, he was a postdoctoral research fellow working on collaborative European research projects known as SCOUT, NEWCOM and PHOENIX. Since August 2006, he has been a member of academic staff in the School of Electronics and Computer Science, University of Southampton. He is involved in the OPTIMIX and CONCERTO European projects as well as the IU-ATC and UC4G projects. He is currently an Associate Professor in telecommunications at the University of Southampton. His research interests include adaptive coded modulation, coded modulation, channel coding, space-time coding, joint source and channel coding, iterative detection, OFDM, MIMO, cooperative communications, distributed coding, quantum error correction codes and joint wireless-and-optical-fibre communications. He has published over 180 papers and co-authored two John Wiley/IEEE Press books in this field. He is a Senior Member of the IEEE, a Chartered Engineer and a Fellow of the Higher Education Academy in the UK.

Jia Shi, received both his MSc. and Ph.D degrees from University of Southampton, UK, in 2010 and 2015, respectively. He was a research associate with Lancaster University, UK, during 2015-2017. Then, he became a research fellow with 5GIC, University of Surrey, UK, from 2017 to 2018. Since 2018, he joined Xidian University, China, and now is an Associate Professor in the State Key Lab. of Integrated Services Networks (ISN). His current research interests include mmWave communications, artificial intelligence, resource allocation in wireless systems, covert communications, physical layer security, cooperative communication, etc. He is now an associate editor of Electronics Letters, and is also serving as a guest editor of China Communications.

Lixin Li, received the B.Sc. and M.Sc. degrees in communication engineering, and the Ph.D. degree in control theory and its applications from Northwestern Polytechnical University (NPU), Xian, China, in 2001, 2004, and 2008, respectively. He was a Post-Doctoral Fellow with NPU from 2008 to 2010. In 2017, He was a visiting scholar at the University of Houston, Texas. He is currently an Associate Professor in the School of Electronics and Information, NPU. He has authored or co-authored over 80 technical papers in journals and international conferences, and he holds 10 patents. His current research interests include wireless communications, game theory, and machine learning. He has reviewed papers for many international journals. He received the 2016 NPU Outstanding Young Teacher Award, which is the highest research and education honors for young faculties in NPU.

Dawei Wang (S'14–M'18), received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2018. From 2016 to 2017, he was a Visiting Student with the School of Engineering, The University of British Columbia. He is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an. His current research interests include cognitive radio, green communications, UAV communications, resource allocation, physical layer security, cooperative communication, and non-convex and robust optimization.