IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE FOR PHYSICAL-LAYER WIRELESS COMMUNICATIONS

Received July 2, 2019, accepted July 21, 2019, date of publication July 30, 2019, date of current version August 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2932016

Deep Reinforcement Learning Based Intelligent User Selection in Massive MIMO Underlay Cognitive Radios

ZHAOYUAN SHI[1,2], (Student Member, IEEE), XIANZHONG XIE[1], (Member, IEEE), AND HUABING LU[1], (Student Member, IEEE)

[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Normal University, Anqing 246011, China

Corresponding author: Zhaoyuan Shi (shizy123@126.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61502067, in part by the Key Research Project of Chongqing Education Commission under Grant KJZD-K201800603, in part by the Chongqing Nature Science Foundation under Grant CSTC2018jcyjAX0432 and Grant CSTC2016jcyjA0455, in part by the Project of Anhui Education Department under Grant AQKJ2015B008, in part by the Doctoral High School Talent Training Project under Grant BYJS2016003, and in part by the Chongqing Graduate Scientific Research Innovation Project under Grant CYB17131.

ABSTRACT Cognitive radio (CR) and massive multiple-input multiple-output (MIMO) have attracted much interest recently due to the amazing ability to accommodate more users and improve spectrum utilization. This paper investigates the QoS-aware user selection approach for massive MIMO underlay cognitive radio. Two main CR scenarios are considered: 1) the channel state information (CSI) of the cross channels are available at the secondary base station (SBS), and 2) any CSI of cross-network is unavailable at SBS. For the former, we develop the low-complexity increase-user-with-minimum-power algorithm (IUMP) and decrease-user-with-maximum-power algorithm (DUMP) which both can address the problem of user selection with power allocation. However, the CSI is typically not available in practice. To address the intractable issue, we propose a deep reinforcement learning-based approach, which can enable the SBS to realize efficient and intelligent user selection. The simulation results show that the IUMP and DUMP algorithms have obvious performance advantages over traditional user selection methods. In addition, results also verify that our constructed neural network can efficiently learn the optimal user selection policy in the unknown dynamic environment with fast convergence and high success rate.

INDEX TERMS Cognitive radio, massive MIMO, power allocation, deep reinforcement learning, user selection.

I. INTRODUCTION

With the increasing number of communication devices and the growing demand for the spectrum resource in 5G & B5G network, improving the utilization efficiency of spectrum resource is urgent for the scarce available spectrum resources [1], [2]. Cognitive radio (CR) and massive multiple-input multiple-output (MIMO) have been widely envisaged as the major candidates for future wireless network to accommodate more users with the limited spectrum [3]. As an intelligent wireless communication system [4]–[7], CR allows secondary user (SU) to share the spectrum with licensed primary user (PU) when collisions or harmful interference can be avoided. The massive MIMO system has been employed hotly in cognitive radio network (CRN) recently, since its powerful precoding potential, which can ensure more users achieve reliable downlink transmission concurrently [8]. In addition, large-scale antenna array can provide huge spectral efficiency and energy efficiency gain for CRN [3], [9].

User selection is always an important issue in CRN, which can improve various system performance, such as reducing excessive overhead and computational complexity for cooperative spectrum sensing (CSS) [10]–[13], improving the performance of PU [14]–[16], maximizing the sum rate of the SUs [17], [18]. With the huge number of devices and scarce spectrum resources in 5G communications, some literatures have begun to study the issue of accommodating more SUs to participate in communications by user selection [19]–[21].

The associate editor coordinating the review of this manuscript and approving it for publication was Guan Gui.

For admitting more SUs to participate in communications, [19] studied the SUs selection scheme for the massive MIMO underlay CRN, which consisted of a secondary base station (SBS), K SUs and L primary transmitter-receiver pairs, the authors proposed a QoS aware power allocation and user selection algorithm with available global channel state information (CSI). The algorithm starts by selecting all SUs and then deletes the SU with maximum power per iteration. [20] considered a CRN wherein the SUs want to share a number of frequency bands licensed to PUs. For enhancing the spectrum utilization and maintaining user fairness, the author presented a mixed-integer programming framework and proposed an optimal algorithm based on branch and bound method for the joint resource allocation, user maximization and beamforming problem. In [21], multiple computationally efficient user selection strategies were proposed based on channel correlation, orthogonality, and water-filling for the downlink MIMO CRN.

The algorithms mentioned above can achieve efficient user selection and improve system performance. However, the algorithms in [19], [20] are only applicable to the CRN that each primary transmitter is equipped with a signal antenna. The scheme in [21] can only be applied to the CRN containing one PU, which both lead to the limitation of algorithm application. In addition, the CSI of the cross channels (the channels between SBS and each PU) is always needed [19], [20] which may result in additional feedback overhead to the PUs. Furthermore, the states of PUs in [19]–[21] are fixed to ON, which means that the PUs are always in communication. However, instead of keeping the deterministic states, the PU is idle when there is no communication task in practice [22].

According to the above analysis, we will study the QoS aware user selection algorithms for the massive MIMO underlay CRN wherein the SBS and primary base station (PBS) are both equipped with massive antennas, besides, the active states of the PUs are dynamical. Specifically, we consider two main scenarios: 1) The SBS has the CSI of cross channels, and 2) the SBS has absolutely no CSI of cross channels. Obviously, user selection is particularly difficult to achieve for the second scenario. Besides, with the advent of the 5G & B5G standards, the communication with faster rate, higher QoS and intelligent requirements increase dramatically [3], [23]. Both these lead to a new opportunity to the introduction of deep learning into the study of massive MIMO underlay CRN [7]. Reinforcement learning (RL) [24], [25] is one of the most powerful machine learning tools for intelligent decision making since RL can be invoked to find an optimal action policy for any given Markov decision process, especially when the system model is dynamic [26], [27]. Formulating the selection problem as a Markov Decision Process and using RL tools to solve has been a popular approach [28], [29].

In recent years, RL has been explored for CR systems in some literatures. [28], [30], [31] used ideas from Thompson sampling to propose an algorithm for channel selection in CR. In [32], three route selection schemes were proposed to enhance the networks performance of CRN. A two-stage RL algorithm was proposed in [28], in which a channel was selected from N for SU based on RL algorithm, then Bayesian approach was adopted to shorten the sensing time. Q learning [33], as a classical reinforcement learning algorithm, is also widely used in CRN [29], [34], [35]. A form of real-time multi-agent Q learning RL was proposed to manage the aggregated interference generated by multiple WRAN systems [34]. In [35], a new Q earning-based transmission scheduling mechanism with deep learning for the cognitive radio-based Internet of Things (IoT) was proposed to maximize the system throughput. In [29], a deep Q learning-based method was developed to conduct power control.

It can be seen that the introduction of RL can effectively improve the performance of CRN, by overcoming the uncertainty of the system, and facilitating the system intellectualization. However, the problem of user selection with RL in massive MIMO underlay CRN has not been studied in existing literatures. In this paper, for the scenario without CSI of cross channels at SBS, we develop a user selection algorithm based on deep RL for massive MIMO underlay CRN. The main contributions of our work can be summarized as follow:

(1) When the CSI of cross channels is available at SBS, we propose two low-complexity QoS aware user selection approaches: Decrease-User-with-Maximum-Power (DUMP) algorithm and Increase-user-with-minimum-power (IUMP) algorithm, which address the problem of joint user maximization and power allocation for the massive MIMO underlay CRN. Besides, the two algorithms are different from the algorithms in [19]–[21] which are only applicable to CRN with one PU or the primary transmitter can be only equipped with a single antenna.

(2) When the CSI of cross channels is unavailable at SBS, we introduce RL into massive MIMO underlay CRN to address the problem of user selection. In addition, we develop a deep Q learning network (DQN)-based user selection algorithm by which the SBS can learn an efficient policy to select as many appropriate SUs as possible.

(3) We evaluate and analyze the performance of our proposed SUs selection algorithms, the algorithm IUMP and DUMP are compared with some traditional classical user selection schemes. In addition, the DQN-based approach is evaluated from the perspectives of the loss function of the deep neural network (DNN), success rate, average transition step and average number of selected SUs. Simulation results prove that our proposed algorithms can efficiently deploy QoS aware user selection for the massive MIMO underlay CRN regardless of whether CSI of cross channels is available or not.

The following notations are used in this paper. We use the upper case boldface letters for matrices and lower case boldface letters for vectors. $[\mathbf{A}]_n$ stands for the n-th column of matrix $\mathbf{A}$, $\mathbf{A}^T$ and $\mathbf{A}^H$ respectively represent the transpose and conjugate transpose of $\mathbf{A}$, $\mathcal{F} \leftarrow \{\mathcal{F} - \{n\}\}$ denotes

removing $\{n\}$ from $\{\mathcal{F}\}$, $\mathcal{F} \leftarrow \{\mathcal{F}+\{n\}\}$ stands for adding $\{n\}$ to $\{\mathcal{F}\}$, $\emptyset$ is empty set, $|\mathcal{F}|$ stands for the cardinality of set $\mathcal{F}$, % represent the remainder operation.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

In our formulation, we consider a generic massive MIMO downlink underlay CRN, wherein the primary network consists of a PBS and K PUs and the secondary network consists of one SBS and N SUs. SBS and PBS are both equipped with a large-scale antenna, and the number of antennas is M. All PUs and SUs are configured with a single antenna. Let $\mathcal{N} = \{1, 2, \cdots, N\}$ be the set of SUs and $\mathcal{K} = \{1, 2, \cdots, K\}$ be the set of PUs. All users are randomly distributed around own base station, the distance between the PBS and SBS is d_0. SUs aim to share the spectrum resource belongs to PUs. The system model is depicted in Fig. 1.

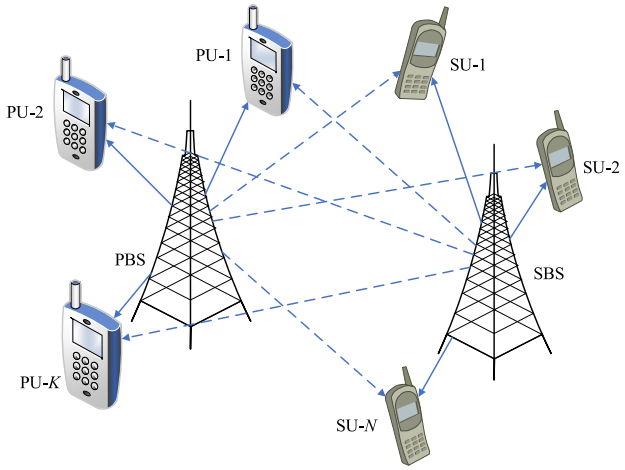


FIGURE 1. System model.

In our setup, similar to the existed works [36], [37], we assume that the activity states of each PU are modeled by Discrete-Time Markov Chain (DTMC), i.e., each PU decides whether to communicate based on DTMC model, where the transfer probability is shown in Fig. 2. There are generally two states, idle shows OFF state, i.e., spectrum is not occupied by the PU, and busy shows ON state, i.e., spectrum is occupied by the PU.

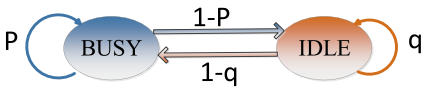


FIGURE 2. The Markov chain model for PU activity.

We set $\mathbf{h}_{Sn}^{S} = \sqrt{\beta_{Sn}^{S}}\widetilde{\mathbf{h}}_{Sn}^{S} \in \mathbb{C}^{1\times M}$ as the channel gain from SBS to the n-th SU, where $\widetilde{\mathbf{h}}_{Sn}^{S} \sim \mathcal{CN}(0, I)$, $\beta_{Sn}^{S} = \left(\frac{\lambda}{4\pi d}\right)^2$ is the path loss of the $\mathbf{h}_{Sn}^{S}$, where λ is the signal wavelength, d presents the distance between transmitter and receiver. $\mathbf{h}_{Sn}^{P} = \sqrt{\beta_{Sn}^{P}}\widetilde{\mathbf{h}}_{Sn}^{P} \in \mathbb{C}^{1\times M}$ stands for the channel gain between PBS to the n-th SU and $\widetilde{\mathbf{h}}_{Sn}^{P} \sim \mathcal{CN}(0, I)$. Similarly, $\mathbf{h}_{Pk}^{P} = \sqrt{\beta_{Pk}^{P}}\widetilde{\mathbf{h}}_{Pk}^{P} \in \mathbb{C}^{1\times M}$ and $\mathbf{h}_{Pk}^{S} = \sqrt{\beta_{Pk}^{S}}\widetilde{\mathbf{h}}_{Pk}^{S} \in \mathbb{C}^{1\times M}$ respectively denote the channel gain from PBS and SBS to the k-th PU. As precoding vector can improve system rates while reducing user interference [38], [39]. we design the unit-norm Zero-forcing (ZF) precoding vectors $\mathbf{v}_n \in \mathbb{C}^{M\times 1}$, $(n \in \mathcal{N})$ and $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$, $(k \in \mathcal{K})$ for the n-th SU and k-th PU. In addition, binary variable $U_{Pk}(i)$, $(k \in \mathcal{K})$ and U_{Sn}, $(n \in \mathcal{N})$ are used to represent the state indicator of the k-th PU and the n-th SU at time frame i.

B. PROBLEM OPTIMIZATION

Our objective is to maximize the number of selected SUs while guaranteeing the QoS requirements of all communication users. In this paper, QoS is measured in terms of the specific rate and interference. In particular, the instantaneous rate of the each selected SU has to be greater than R_0 and the interference towards each communication PU must below I_0. The instantaneous rate of the n-th SU and the interference received by the k-th PU at time frame i can be respectively expressed as follow

$$R_{Sn}(i) = log_2\left(1 + \frac{\left|\sqrt{p_{Sn}}\mathbf{h}_{Sn}^{S}\mathbf{v}_n\right|^2}{I_{Sn}^{P}(i) + I_{Sn}^{S}(i) + N_{Sn}(i)}\right), \quad (1)$$

$$I_{Pk}(i) = I_{Pk}^{P}(i) + I_{Pk}^{S}(i), \quad (2)$$

where $I_{Sn}^{P}(i)$ and $I_{Sn}^{S}(i)$ represent the interference from primary network and secondary network to the n-th SU respectively. Similarly, $I_{Pk}^{S}(i)$ and $I_{Pk}^{P}(i)$ stand for the interference from secondary network and primary network to the k-th PU respectively. p_{Sn} stands for the transmit power of the n-th SU, which must satisfy the power constraint $\sum_{n=1}^{N} U_{Sn}(i)p_{Sn} \leq P^S$, and $N_{Sn}(i)$, a Gaussian random variable with zero mean and variance σ_w^2, is used to characterize the random variation caused by shadow effects and noise. Then the QoS aware user selection optimization problem can be formulated as

$$\max_{U_{S1}(i),\cdots,U_{Sn}(i)} \sum_{n=1}^{N} U_{Sn}(i), \quad (3)$$

$$\text{S. t. : } U_{Sn}(i), \quad U_{Pk}(i) \in \{0, 1\}, \ (n \in \mathcal{N}, k \in \mathcal{K}), \quad (4)$$

$$\sum_{n=1}^{N} U_{Sn}(i)p_{Sn} \leq P^S, \quad (n \in \mathcal{N}), \quad (5)$$

$$R_{Sn}(i) \geq R_0, \quad (if\ U_{Sn}(i) = 1, n \in \mathcal{N}), \quad (6)$$

$$I_{Pk}(i) \leq I_0, \quad (if\ U_{Pk}(i) = 1, k \in \mathcal{K}). \quad (7)$$

III. USER SELECTION FOR UNDERLAY MASSIVE MIMO CRN

Obviously, the optimization problem of user selection in (3)-(7) is difficult to solve as it is a non-convex combination and NP-hard problem. Here, the problem of user selection is studied for the following two cases: 1) CSI of cross channels is available at SBS; and 2) CSI of cross channels is unavailable at SBS.

A. USER SELECTION WITH CROSS CHANNELS CSI

We first assume that the CSI of the cross channels is available at SBS. For this scenario, we design unit-norm ZF precoding vectors $\mathbf{v}_n$ and $\mathbf{w}_k$ as follow

$$\mathbf{v}_n = \frac{\left[\Psi^H\left(\Psi\Psi^H\right)^{-1}\right]_n}{\left\|\left[\Psi^H\left(\Psi\Psi^H\right)^{-1}\right]_n\right\|}, \quad \mathbf{w}_k = \frac{\left[\Phi^H\left(\Phi\Phi^H\right)^{-1}\right]_k}{\left\|\left[\Phi^H\left(\Phi\Phi^H\right)^{-1}\right]_k\right\|}, \tag{8}$$

where $\Psi = \left[\left(\mathbf{h}_{S1}^S\right)^T, \cdots, \left(\mathbf{h}_{SN}^S\right)^T, \left(\mathbf{h}_{P1}^S\right)^T, \cdots, \left(\mathbf{h}_{PK}^S\right)^T\right]^T$, and $\Phi = \left[\left(\mathbf{h}_{P1}^P\right)^T, \cdots, \left(\mathbf{h}_{PK}^P\right)^T\right]^T$. Obviously, the precoding vectors $\mathbf{v}_n$, $(n \in \mathcal{N})$ can eliminate the interference of the n-th SU on primary network and secondary network. However, $\mathbf{w}_k$, $(k \in \mathcal{K})$ can only eliminate the interference among the PUs, since in practical applications, the PU does not actively acquire the CSI of cross channels from PBS to SUs. Hence, for this case, various interference in equation (1)-(2) can be calculated as follow

$$I_{Sn}^P(i) = \sum_{k=1}^{K} p^P |U_{Pk}(i)\mathbf{h}_{Sn}^P \mathbf{w}_k|^2, \tag{9}$$

$$I_{Sn}^S(i) = 0, \quad I_{Pk}^P(i) = I_{Pk}^S(i) = 0, \tag{10}$$

where p^P denotes the transmit power of each PU. According to equation (2) and (10), we can see that the PUs will not receive any interference due to the design of the ZF precoding vectors. In addition, instantaneous rate of the n-th SU $(U_{Sn}(i) = 1)$ can be expressed as

$$R_{Sn}(i) = log_2\left(1 + \frac{p_{Sn}|\mathbf{h}_{Sn}^S \mathbf{v}_n|^2}{\sum_{k=1}^{K} p^P |U_{Pk}(i)\mathbf{h}_{Sn}^P \mathbf{w}_k|^2 + N_{Sn}(i)}\right). \tag{11}$$

Then the optimization problem (3)-(7) can be simplified as

$$\max_{U_{S1}(i),\cdots,U_{SN}(i)} \sum_{n=1}^{N} U_{Sn}(i), \tag{12}$$

$$\text{S.t.: } U_{Sn}(i), U_{Pk}(i) \in \{0, 1\}, (n \in \mathcal{N}, k \in \mathcal{K}), \tag{13}$$

$$\sum_{n=1}^{N} U_{Sn}(i) p_{Sn} \le P^S, \tag{14}$$

$$log_2\left(1 + \frac{p_{Sn}|\mathbf{h}_{Sn}^S \mathbf{v}_n|^2}{\sum_{k=1}^{K} p^P |U_{Pk}(i)\mathbf{h}_{Sn}^P \mathbf{w}_k|^2 + N_{Sn}(i)}\right) \ge R_0. \tag{15}$$

Obviously, in order to achieve the specific threshold rate R_0, the power assigned to the n-th SU must satisfy the following inequality

$$p_{Sn} \ge \frac{(2^{R_0} - 1)\left(\sum_{k=1}^{K} p^P |U_{Pk}(i)\mathbf{h}_{Sn}^P \mathbf{w}_k|^2 + N_{Sn}(i)\right)}{|\mathbf{h}_{Sn}^S \mathbf{v}_n|^2}. \tag{16}$$

In order to select more SUs that meet the rate requirements (6) and power constrain $\sum_{n=1}^{N} U_{Sn}(i) p_{Sn} \le P^S$, we perform the following power allocation

$$p_{Sn} = \frac{(2^{R_0} - 1)\left(\sum_{k=1}^{K} p^P |U_{Pk}(i)\mathbf{h}_{Sn}^P \mathbf{w}_k|^2 + N_{Sn}(i)\right)}{|\mathbf{h}_{Sn}^S \mathbf{v}_n|^2}. \tag{17}$$

Let $\mathcal{F}$ denotes the set of selected users, $\bar{\mathcal{F}}$ stands for the set of unselected users, and $\mathcal{F}^*$ is one of the optimal sets. The easiest way to find the optimal set $\mathcal{F}^*$ is to list all possible sets in incrementing or descending order. For the incrementing order, we need one-by-one list all possible sets of cardinalities $1, 2 \cdots, |\mathcal{F}^*|$. For each set $\mathcal{F}$, we need to check whether the constraints (13)-(15) are satisfied, and then find the one of the optimal sets with biggest cardinality. However, the total number of all sets is $\sum_{n=1}^{|\mathcal{F}^*|} C_N^n$, which increases exponentially with N. Therefore, it is necessary to design low-complexity user selection schemes.

Firstly, we design a low-complexity user selection algorithm called Increase-User-with-Minimum-Power (IUMP). The algorithm is initialized by $\mathcal{F} = \emptyset$, i.e., none of the SUs are selected. Then the ZF precoding vector and power allocation are carried out for all SUs by (8) and (17), respectively. The user with minimum power allocation will be selected per iteration if the power constraint (14) is satisfied. For clarity, the IUMP-based user selection scheme is summarized in Algorithm 1.

Algorithm 1 IUMP-Based SUs Selection Algorithm

Initialize: All SUs are not selected,
i.e., $U_{Sn}(i) = 0$, $(\forall n \in \mathcal{N})$, $\mathcal{F} = \varnothing$ and $\bar{\mathcal{F}} = \mathcal{N}$;
Compute the unit-norm ZF precoding vectors $\mathbf{v}_n$ and $\mathbf{w}_k$ by formula (8);
Perform power allocation for all SUs according (17).
while $\bar{\mathcal{F}} \neq \emptyset$ **do**
  find the SU with minimum power in $\bar{\mathcal{F}}$, i.e., $n' = \arg\min_{n \in \bar{\mathcal{F}}} p_{Sn}$;
  if $\sum_{n=1}^{N} U_{Sn}(i) p_{Sn} + p_{Sn'} < P^S$ **then**
    Increase the SU with minimum power to $\mathcal{F}$, i.e., set $U_{Sn'}(i) = 1$, $\bar{\mathcal{F}} \leftarrow \left\{\bar{\mathcal{F}} - \{n'\}\right\}$, $\mathcal{F} \leftarrow \left\{\mathcal{F} + \{n'\}\right\}$;
  else
    $\mathcal{F}^* = \mathcal{F}$
    Stop
  end if
end while

Contrary to the principle of user selection in the IUMP algorithm, Decrease-User-with-Maximum-Power (DUMP) based user selection algorithm is to find the user with the largest allocated power in each iteration. Furthermore, the algorithm is different form the user selection scheme in [19] which is not applicable to the CRN in this paper, because it assumes that PBS is configured with a single antenna, and the interference among PUs is not taken into account.

Specifically, algorithm DUMP is initialized by selecting all SUs, i,e., $\mathcal{F} = \mathcal{N}$ and $\bar{\mathcal{F}} = \emptyset$. The ZF precoding vectors and power allocation are also carried out for all SUs by (8)

and (17), respectively. Then the user with maximum power allocation will be found and decreased from $\mathcal{F}$ if the power constraint (14) cannot be satisfied. The algorithm steps are summarized in Algorithm 2.

Algorithm 2 DUMP-Based SUs Selection Algorithm

Initialize: Selected all SUs,
i.e., $U_{Sn}(i) = 1, (\forall n \in \mathcal{N})$, $\mathcal{F} = \mathcal{N}$ and $\bar{\mathcal{F}} = \emptyset$;
Compute the unit-norm ZF precoding vectors $\mathbf{v}_n$ and $\mathbf{w}_k$ by formula (8);
Perform power allocation for all SUs according (17).
while $\mathcal{F} \neq \emptyset$ **do**
  if $\sum_{n=1}^{N} U_{Sn}(i) p_{Sn} > P^S$ **then**
    find the SU with maximum power in $\mathcal{F}$, i.e., $n' = \arg\max_{n\in\mathcal{F}} p_{Sn}$;
    Decrease the SU with maximum power from $\mathcal{F}$, i.e., set $U_{Sn'}(i) = 0$, $\mathcal{F} \leftarrow \{\mathcal{F} - \{n'\}\}$, $\bar{\mathcal{F}} \leftarrow \{\bar{\mathcal{F}} + \{n'\}\}$;
  else
    $\mathcal{F}^* = \mathcal{F}$
    Stop
  end if
end while

Algorithm 1 and algorithm 2 achieve joint power allocation and user selection with low complexity. Those SUs who meet the system QoS requirements are selected as many as possible. However, both of the algorithms are based on the premise that the CSI of cross channels is available at SBS.

B. USER SELECTION WITHOUT CROSS CHANNELS CSI

The worst case is that the CSI of cross channels is absolutely unknown, i.e., the primary network and the secondary network work in non-cooperative mode, we cannot get any informations about $\mathbf{h}_{Pk}^S, (k \in \mathcal{K})$ at SBS. However, this case is more common in practical applications. Because the primary network usually does not actively open the interface to exchange with the secondary network, nor does it actively provide own CSI to the secondary network in reality. For this case, the ZF precoding vectors $\mathbf{v}_n$ and $\mathbf{w}_k$ can only be designed to eliminate the interference of internal network, which can be expressed as

$$\mathbf{v}_n = \frac{\left[\Psi^H\left(\Psi\Psi^H\right)^{-1}\right]_n}{\left\|\left[\Psi^H\left(\Psi\Psi^H\right)^{-1}\right]_n\right\|}, \quad \mathbf{w}_k = \frac{\left[\Phi^H\left(\Phi\Phi^H\right)^{-1}\right]_k}{\left\|\left[\Phi^H\left(\Phi\Phi^H\right)^{-1}\right]_k\right\|}, \tag{18}$$

where $\Psi = \left[\left(\mathbf{h}_{S1}^S\right)^T, \cdots, \left(\mathbf{h}_{SN}^S\right)^T\right]^T$, and the $\Phi = \left[\left(\mathbf{h}_{P1}^P\right)^T, \cdots, \left(\mathbf{h}_{PK}^P\right)^T\right]^T$. Obviously, each user will receive external interference. In addition, due to the lack of the CSI of cross channels, the interference of external network cannot be calculated. So we cannot get the instantaneous rate of SUs and the interference of PUs. This makes it impossible to solve the QoS aware user selection problem in conventional methods.

In addition, as an intelligent wireless communication system, the massive MIMO underlay CRN needs an intelligent user selection algorithm to enable the SBS to intelligently make decision. Hence, a DQN-based user selection algorithm is proposed in this section, where the specific rate requirement of selected SUs and interference requirement of PUs can both be directly addressed. Before presenting the proposed algorithm, the main parts of the RL based Markov Decision Processes (MDP) [24] are given with a new proposed reward function, and a Q learning framework is adopted to address the SUs selection problem.

1) MARKOV DECISION PROCESS FOR SUS SELECTION

RL is an important branch of machine learning, which can be used in the CRN to search the optimal policy. Similar to the existing literatures [29], [34], [35], we apply the MDP to model the user selection process in this paper.

In this paper, we model the user selection problem as $(\mathcal{S}, \mathcal{A}, r(i), \gamma)$, where $\mathcal{S}$ is the environment state space set, $\mathcal{A}$ denotes the action space set, $r(i)$ represents the immediate reward at the current time frame i, and γ is a discount factor. When the agent takes action $a(i) \in \mathcal{A}$, the current environment state $S(i) \in \mathcal{S}$ will be transformed into state $S(i+1) \in \mathcal{S}$, while the corresponding reward $r(i)$ will be obtained. The interaction between the agent (SBS) and the CR environment is depicted in Fig. 3.

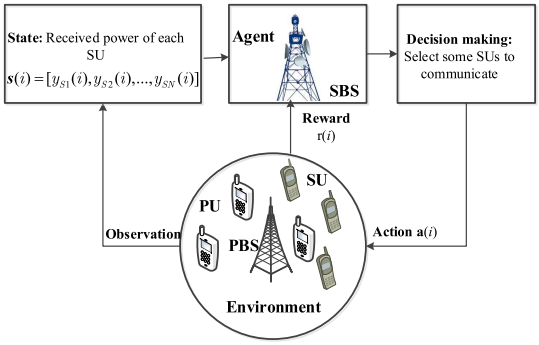


FIGURE 3. The interaction between the agent and the environment.

Agent: SBS

State space $\mathcal{S}$: The environment state information can be depicted by all SUs. We use

$$\mathbf{s}(i) = [y_{S1}(i), y_{S2}(i) \cdots y_{SN}(i)] \tag{19}$$

to denote the environment state in time frame i, where the $y_{Sn}(i), n \in \mathcal{N}$ is the received power at the n-th SU in time frame i, which can be expressed as

$$y_{Sn}(i) = \sum_{n=1}^{N} p^S \left|U_{Sn}(i)\mathbf{h}_{Sn}^S \mathbf{v}_n\right|^2 + \sum_{k=1}^{K} p^P \left|U_{Pk}(i)\mathbf{h}_{Sn}^P \mathbf{w}_k\right|^2 + N_{Sn}(i). \tag{20}$$

To simplify the algorithm, we assume that each SU has the same transmit power p^S. Obviously, the environment state

$\mathbf{s}(i+1)$ is only related to state $\mathbf{s}(i)$, and completely irrelevant to other previous states.

Action space $\mathcal{A}$: The action is considered for SBS to select the SUs intelligently in RL framework. The size of $\mathcal{A}$ increases exponentially with the number of SUs. Note that each action selection of the agent must satisfy the constraints (5)-(7).

Reward function $r(i)$: The search process of the optimal user selection strategy is driven by the reward function since the action is selected with the maximum reward. Considering the purpose of learning is to maximize the number of selected SUs, we propose a new reward function for the RL algorithm, when the QoS of communicating users are satisfied, the reward function can be expressed as

$$r(i) = r_0 + \sum_{n=1}^{N} U_{Sn}(i) * \mu, \tag{21}$$

where $\sum_{n=1}^{N} U_{Sn}(i)$ represents the number of selected SUs at time frame i, r_0 is the basic reward, μ means reward multiplier. Obviously, the reward function enables the agent to select as many SUs as possible to participate in communication.

The main objective of MDP is to learn the optimal user selection policy for agent: Let $\mathbf{v}^{\pi}(\mathbf{s}(i), a(i))$ denote the state action function, which is the discount cumulative reward for the action $a(i)$ at current state $\mathbf{s}(i)$ with the policy π, which can be expressed as

$$\mathbf{v}^{\pi}(\mathbf{s}(i), a(i)) = \sum_{t=i}^{T} \gamma^{t-i} r(t), \tag{22}$$

where T presents the number of time frame required to reach the goal state. The goal state of this paper is that some appropriate SUs are selected which can meet the QoS requirements of the system. Then the task becomes learning an optimal policy π^* that maximizes $\mathbf{v}^{\pi}$, i.e.,

$$\pi^* = \arg\max_{\pi} \mathbf{v}^{\pi}(\mathbf{s}(i), a(i)). \tag{23}$$

Obviously, it is not straightforward to address the problem.

2) THE DQN-BASED SUS SELECTION ALGORITHM

Instead of computing the problem (23) directly, we adopt the RL tools to learn the optimal policy, which contain Q learning, policy gradient, actor critic and so on [40]. Unlike the one-step update in the Q learning approach, the parameters in policy gradient scheme are rounded up after each exploration, which results in lower learning efficiency. In addition, the policy gradient and actor critic algorithms are more suitable for systems with continuous action. Hence, for the system with continuous-value states and discontinuous-value actions in this paper, deep Q learning algorithm is applied to obtain the optimal policy for the intelligent user selection.

To achieve a better understanding, the classic Q learning algorithm is briefly introduced firstly. In Q learning, a Q value function is invoked to evaluate the discount cumulative reward of taking action $\mathbf{a}(i)$ at state $\mathbf{s}(i)$. The Q value can be iteratively updated by Bellman equation [40].

$$Q(\mathbf{s}(i), a(i)) = r(i) + \gamma \max_{a} \mathrm{Q}(\mathbf{s}(i+1), a), \tag{24}$$

where the $\mathbf{s}(i+1)$ is the next state led by taking action $a(i)$ at the state $\mathbf{s}(i)$, γ is the discount factor. It has been proved that the Q value will be updated to converge [40]. All the convergent values will form a final Q table, that is, for any state, each action has a corresponding Q value in the table. Hence, after the Q table converges, we merely need to search the Q table and select the action with largest Q value for any given states.

Obviously, with the number of states and actions increasing, the scale of Q table will be large, which results in a long search time. Unfortunately, in this paper, the value of states is continuous due to the random variation of the environment. To overcome this difficulty, we introduce DQN [41], in which each Q value can be calculated through a DNN, i.e., $Q(\mathbf{s}, a, \theta)$, where θ indicates the network parameter. For a given input (a certain state), the DNN will output the Q values of all actions in the state. In this paper, the input of DNN is an N-dimensional vector represented by signals power received by the N SUs, and the network output is a 2^N-dimensional vector which includes all Q values for each SUs selection strategy.

For training the DNN with initialized parameter, we need an experience replay memory with capacity D to store sufficient transitions $(\mathbf{s}(l), a(l), r(l), \mathbf{s}(l+1))$, where $a(l)$ is the action taken in state $s(l)$, $r(l)$ is the immediate reward obtained, and $\mathbf{s}(l+1)$ is the next state, where the action in the l-th iteration is selected by

$$a(l) = \begin{cases} \arg\max_a Q(\mathbf{s}(l), a; \theta_l); & \text{with probability } \varepsilon_l, \\ \text{Randomly action}; & \text{otherwise}, \end{cases} \tag{25}$$

where ε-greedy policy [40] is introduced to fully explore the environment, we set it to $\varepsilon_l = 0.8(1 - l/I)$, where I is the total number of iterations. When the number of iteration l is greater than sampling threshold τ, a minibatch of transitions set Ω_l from D will be randomly selected for network training in iteration l. Fig. 4 shows the specific training principle and DNN network structure. The loss function of the neural network in iteration l can be defined as

$$L(\theta_l) = \sum_{l \in \Omega_l} \left(\hat{Q}\left(\mathbf{s}(l), a(l); \theta_l^-\right) - Q\left(\mathbf{s}(l), a(l); \theta_l\right) \right)^2, \tag{26}$$

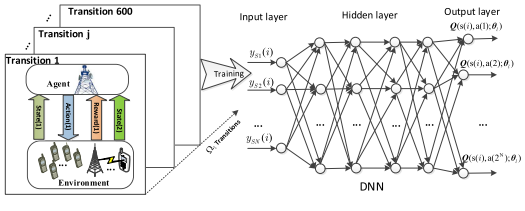


FIGURE 4. Training principle and DNN network structure.

where the $Q(\mathbf{s}(l), a(l); \theta_l)$ denotes the output of the DNN, Ω_l is the training transitions set used at iteration l. It should be noted that, in this paper, we use another network to compute the target value $\hat{Q}\left(\mathbf{s}(l), a(l); \theta_l^-\right)$, which means we create two networks: the online network with parameter θ and target network with parameter θ^-. The target network is the same as the online network except that its parameter is copied from online network every C step. The target network is designed to avoid iterative volatility due to differences and correlations between sample data [42], which makes it difficult to stabilize the network parameter. Hence, we update parameter θ in each step to calculate $Q(\mathbf{s}(l), a(l); \theta_l)$ and update target network parameter every C step. The target value can be defined as

$$\hat{Q}\left(\mathbf{s}(l), a(l); \theta_l^-\right) = r(l) + \gamma \max_a \hat{Q}\left(\mathbf{s}(l+1), a; \theta_l^-\right). \quad (27)$$

Differentiating the loss function, we can get the gradient as follows

$$\nabla_{\theta_l} L(\theta_l) = \sum_{i \in \Omega_l} \Big(r(l) + \gamma \max_a \hat{Q}(\mathbf{s}(l+1), a; \theta_l^-) - Q(\mathbf{s}(l), a(l); \theta_l) \Big) \nabla_{\theta_l} Q(\mathbf{s}(l), a(l); \theta_l). \quad (28)$$

In each iteration, we update the online network parameter at learning rate β

$$\theta_{l+1} = \theta_l + \beta \nabla_{\theta_l} L(\theta_l). \quad (29)$$

For clarity, the proposed DQN-based user selection scheme is summarized in Algorithm 3, in which the agent (SBS) is assumed to know whether the SUs and the PUs transmitted successfully (QoS requirements were met). In practice, this can be achieved by listening the acknowledgement signal of all communication users. Note that the DQN-based algorithm can also applied to the situation where all available CSI is perfect.

After training, the SBS can select the action which yields the largest estimated value $Q(\mathbf{S}(i), a(i); \theta)$. In other words, we realize the SUs intelligent selection in the CRN under the non-cooperative mode by using a deep reinforcement learning algorithm.

Algorithm 3 DQN-Based SUs Selection Algorithm

Initialize: Initialize an experience replay memory with capacity D; Initialize the online network and target network with random weights $\theta = \theta^- = \theta_0$;
Initialize the activity state of all SUs;
for episode $l = 1, I$ **do**
  Update each PU activity state based on Markov chain model;
  Obtain $\mathbf{s}(l)$ versus the random observation model (19);
  Choose an action $a(l)$ by the formula (25), where $\varepsilon_l = 0.8(1 - l/I)$;
  Obtain the next state $\mathbf{s}(l+1)$ and observe reward $r(l)$ by function (21);
  Store transition $d_l = \{\mathbf{s}(l), a(l), r(l), \mathbf{s}(l+1)\}$ in the replay memory;
  if $l \geq \tau$ **then**
    Sample a random minibatch Ω_l from replay memory;
    Update θ by equation (29) where the gradient $\nabla_{\theta_l} L(\theta_l)$ and the loss function $L(\theta_l)$ are given by equation (28) and equation (26) ;
  end if
  if $l\%C = 0$ **then**
    Update the weights of target network by $\theta^- = \theta$;
  end if
  if $\mathbf{s}(l)$ is the goal state **then**
    Initialize the activity state of the SUs and PUs; obtain $\mathbf{s}(l+1)$.
  end if
end for

IV. SIMULATION EXPERIMENT AND ANALYSIS

In this section, simulation results are conducted to evaluate the performance of our proposed IUMP, DUMP and DQN-based algorithms in the underlay massive MIMO CRN. First, the performance of the DNN we built is verified. Then, we examine the performance of three algorithms for selecting users. Several conventional user selection schemes are compared.

A. SIMULATION SETUP

In this paper, we evaluate the constructed DNN via three metrics, namely:

1) Loss function of the DNN, which can be calculated by equation (26).

2) Success rate: it is computed as the ratio of the number of successful trials to the total number of independent runs. A trial is considered successful if the current state can move to the goal state (reward >10) within 5 transition steps.

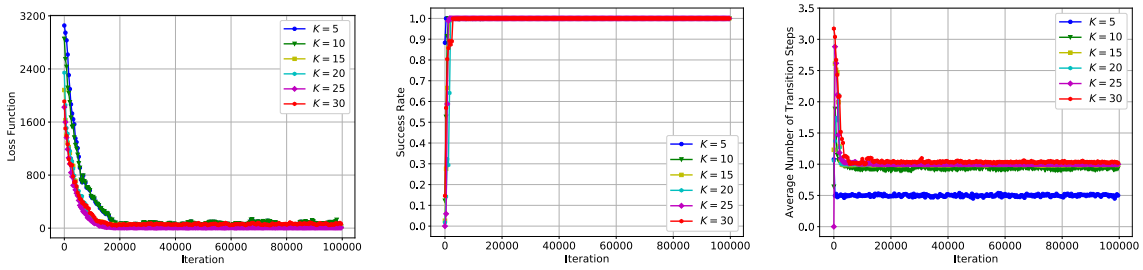
3) Average transition step: the average time frames required to achieve the goal state if the exploration is successful.

The Loss function is used to characterize the convergence of neural networks; Success rate and average transition step can be used to evaluate how well the networks is trained.

We create the DNN model with four fully-connected feed-forward hidden layers, and the number of neurons in each hidden layer is 256, 256, 512, and 512 respectively. The DQN with many hidden layers cannot be fully trained when the data quantity is small. The number of hidden layers is selected according to the simulation comparison. Rectified linear units (ReLUs) are employed as the activation function for the first three hidden layers, and the tanh function for the last hidden layer. Unless otherwise specified, the simulation parameters are considered as table 1.

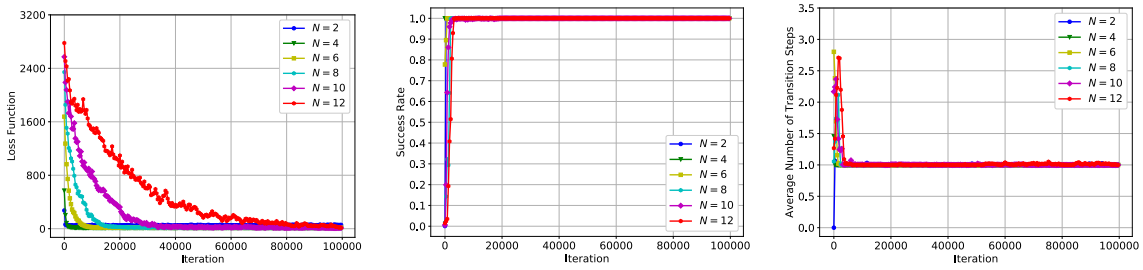
B. PERFORMANCE VERIFICATION

Specifically, in this section we verify the neural networks performance with different PUs number K, SUs number N,



(a) Loss function vs iteration. (b) Success rate vs iteration. (c) Average steps vs iteration.

FIGURE 5. The performance of neural networks with different K.



(a) Loss function vs iteration. (b) Success rate vs iteration. (c) Average steps vs iteration.

FIGURE 6. The performance of neural networks with different N.

TABLE 1. Simulation parameter setting.

Parameters	Value(Description)
The probability q	1/5
The probability p	4/5
Transmit power of PU p^P (W)	1
The signal wavelengt λ (m)	0.1
Distance d_0 (m)	200
The user distribution range (m)	$(50-150)$
Variances of the noise δ_w^2 (dBW)	-97
The capacity of replay memory D	1000
Sampling threshold τ	800
size of Ω_l	600
Basic reward r_0	10
Reward multiplier μ	5
Discount factor γ	0.8
Update time C	200
Learning rate of DQN β	10^{-5}
Exploring probability ε_i	$0.8(1-i/I)$
The total number of iterations I	10×10^4
Total number of PUs K	$5,10,\cdots,30$
Total number of SUs N	$2,4,\cdots,12$
rate threshold R_0 (bps)	$1.5,2,\cdots,4$
rate threshold I_0 (dBW)	$-70,-72,\cdots,-80$

rate threshold R_0 and interference threshold I_0. Next, we will analyze the simulation results in detail.

1) IMPACT OF TOTAL NUMBER OF PUS

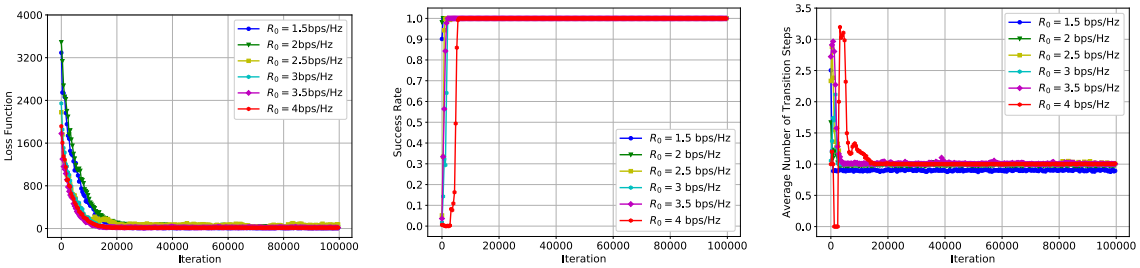
Firstly, we examine the neural networks performance versus different K with $M = 64, N = 8, I_0 = -72$ dBW and $R_0 = 3$ bps/Hz. Fig. 5(a) show that, the loss function converges quickly for different K. In addition, as can be observed in Fig. 5(b) and Fig. 5(c), the smaller the K, the faster the network converges, all the success rate and the average transition step converges after 3×10^3 iterations. Especially, the average transition step will converge to 0.5 when $K = 5$, that is because some original states are the goal state when the number of PUs is small. Furthermore, even if $K = 30$, after 2×10^3 iterations, the average transition step converges to one with 100% success rate, which means that the agent can select the appropriate SUs efficiently in one step for different K.

2) IMPACT OF TOTAL NUMBER OF SUS

We further demonstrate the DNN performance at different N with $M = 64, K = 20, I_0 = -72$ dBW and $R_0 = 3$ bps/Hz. As depicted in Fig. 6(a), after about 800 iterations, the loss function decrease to zero under $N = 2$, however, the loss function becomes larger and converges more slowly when we increase the number of SUs. This is due to the fact that the loss function increase with the number of output node of DNN. In addition, we can see from Fig. 6(b) and Fig. 6(c) that, the success rate and the average transition step converge quickly for different N. In addition, the number of iterations required for convergence increases slowly with increasing N. Nevertheless, even if $N = 12$, an efficient SUs selection policy can be learned in 4×10^3 iterations with one transition step with 100% success rate.

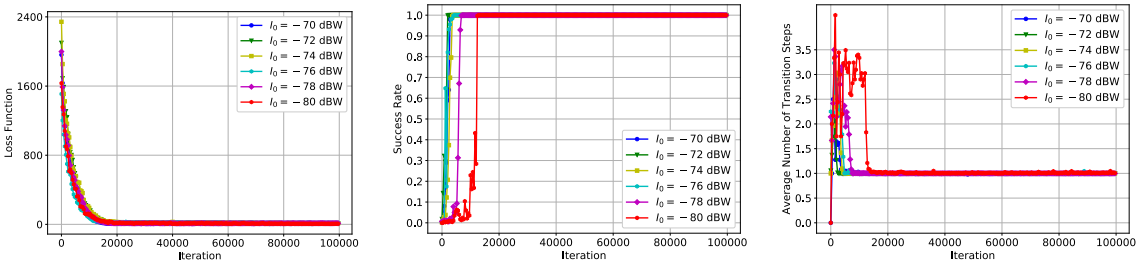
3) IMPACT OF R_0

In order to study the influence of rate threshold, we conduct experiments with different R_0 when $M = 64, K = 20, N = 8$ and $I_0 = -72$ dBW. As Fig. 7 present, all loss functions



(a) Loss function vs iteration. (b) Success rate vs iteration. (c) Average steps vs iteration.

FIGURE 7. The performance of neural networks with different R_0.



(a) Loss function vs iteration. (b) Success rate vs iteration. (c) Average steps vs iteration.

FIGURE 8. The performance of neural networks with different I_0.

fall at a similar rate. In addition, as depicted in Fig. 7(b) and Fig. 7(c), the smaller R_0 is, the less iterations are required. Nevertheless, even $R_0 = 4$ bps/Hz, after about 7×10^3 iterations, 100% success rate is achieved, and the average transition step converges to one after about 1.5×10^4, i.e., the efficient selection policy can also be learned.

4) IMPACT OF I_0

We further explore the impact of the interference threshold of PUs with $M = 64$, $K = 20$, $N = 8$ and $R_0 = 3$ bit/Hz. As shown in Fig. 8(a), the loss function can be reduced to zero before 1.5×10^4 iterations. Fig. 8(b) and Fig. 8(c) present that, even $I_0 = -80$ dBW, after about 1.5×10^4 iterations, the agent can select the appropriate SUs within one step with 100% success rate. In addition, we can notice that the lower the interference threshold, the longer training time it takes.

Through numerous simulation experiments, it is apparent that our constructed neural networks can be quickly trained under various environmental factors. 100% success rate and small transfer step can be quickly obtained, which means that the agent can quickly and efficiently learn the appropriate user selection strategy by our constructed neural networks. In addition, these experimental results prove that our proposed DQN-based user selection scheme can make intelligent decisions efficiently and find the optimal strategy quickly in the face of dynamic changing environment and various system configurations.

C. PERFORMANCE COMPARISON

In this section, we examine the ability of selecting users for the proposed SUs selection algorithms, operating with different system parameters in the massive MIMO underlay CRN.

In Fig. 9 and Fig. 10, the channel similarity-based user selection (CSUS) scheme and precoder-based group user selection (PGUS) algorithm in [21] are compared, in which

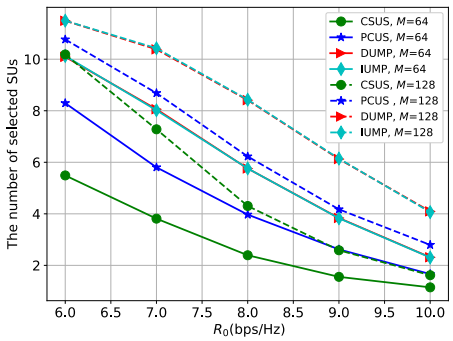


FIGURE 9. Algorithm comparison via R_0.

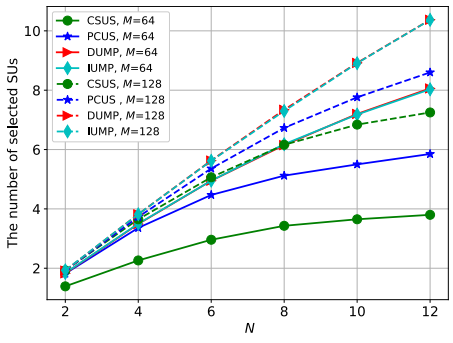
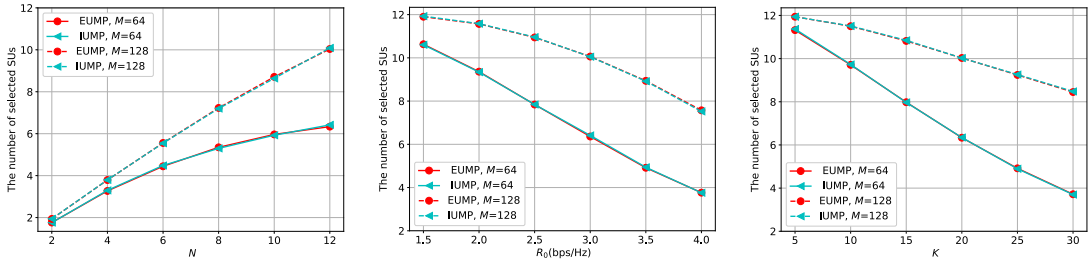
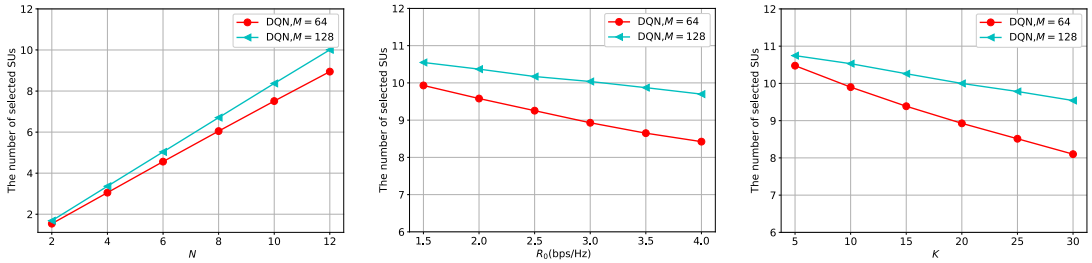


FIGURE 10. Algorithm comparison via N.



(a) Performance via N with $R_0 = 3$(bps/Hz), (b) Performance via R_0 with K=20, $N = 12$. (c) Performance via K with $R_0 = 3$(bps/Hz), K=20. $N = 12$.

FIGURE 11. The performance of IUMP and DUMP SUs selection algorithm with multiple PUs.



(a) Performance via N with $R_0 = 3$(bps/Hz), (b) Performance via R_0 with K=20, $N = 12$. (c) Performance via K with $R_0 = 3$(bps/Hz), K=20. $N = 12$.

FIGURE 12. The performance of DQN-based SUs algorithm.

the system model can only contain one PU. In particular, Fig. 9 depicts the number of selected SUs of four algorithms versus R_0 when $P^S = 4$W, $N = 12$ and $K = 1$. Fig. 10 displays the number of selected SUs versus N when $P^S = 4$W, $R_0 = 7$ bit/Hz and $K = 1$. In addition, the performance of four algorithms is demonstrated with two different antenna configurations. As can be observed that the number of selected SUs of all algorithms decrease as R_0 increase, and increase as N increase. The algorithm IUMP and DUMP have similar performance and obvious performance advantages over other algorithms. Since the same precoding design and power allocation are carried out in the two algorithms. Furthermore, the number of selected users increases with the number of antennas, which verifies the theory that large-scale antenna can admit more users to participate in communication.

Fig. 11 presents the impact of N, R_0, K on the algorithm IUMP and DUMP with multiple PUs. We can observe that the number of selected users increases as N increases, and decreases as R_0 and K increase, since a larger N means more alternative users, whereas a larger R_0 or K means the QoS requirement is harder to satisfy. Besides, as mentioned above, the two algorithms have the same performance, and the increase of the number of antennas has obvious performance improvement, compared with $M = 64$, the number of selected users increased faster with the increase of N, and decreased slower with the decrease of K and R_0 when $M = 128$, which further demonstrates the necessity of applying large-scale antenna in the underlay CRN.

Finally, the performance of proposed DQN-based SUs algorithm was verified for the case that SBS could not acquire the CSI of cross channels. In Fig. 12, the performance of number of selected users was studied versus N, R_0, K with $p_{Sn} = 1$W and $I_0 = -72$dBW. As illustrated by Fig. 12, similar to algorithms IUMP and DUMP, the number of SUs selected by algorithm DQN increases as N increases, decreases as K and R_0 increases. Furthermore, even $K = 30$ or $R_0 = 4$bit/Hz, more than average 8 SUs can be selected with $M = 64$. In other words, DQN-base user selection algorithm can efficiently select user for the CRN with different system parameters.

V. CONCLUSION

In this paper, we studied the user selection strategy for massive MIMO underlay CR system, three selection algorithms were presented for two scenarios, i.e., the CSI of cross channels is available and unavailable at SBS. The proposed algorithm IUMP and DUMP for the perfect cross channels CSI scenario are based on ZF precoding and power allocation that satisfies specific interference requirement of PUs and rate requirement of SUs. For the scenario with unavailable CSI of cross channel, we developed a deep reinforcement learning-based algorithm for the SBS to learn how to intelligently select suitable SUs such that both the PUs and SUs are able to transmit their respective data successful with required QoS. Furthermore, sufficient experiments show that the algorithms proposed in this paper can effectively select as many SUs as possible, regardless of whether the CSI of the cross channel is

available at SBS or not. For future work, power allocation will be considered to further improve the performance of DQN algorithm.

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[2] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.

[3] A. Gupta and E. R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, Jul. 2015.

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

[5] A. He, K. K. Bae, T. R. Newman, J. Gaeddert, K. Kim, R. Menon, L. Morales-Tirado, and J. J. Neel, "A survey of artificial intelligence for cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1578–1592, May 2010.

[6] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1136–1159, Jul. 2013.

[7] M. Liu, J. Yang, T. Song, J. Hu, and G. Gui, "Deep learning-inspired message passing algorithm for efficient resource allocation in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 641–653, Jan. 2019.

[8] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive MIMO linear precoding: A survey," *IEEE Syst. J.*, vol. 12, no. 1, pp. 3920–3931, Dec. 2017.

[9] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 172–179, Feb. 2013.

[10] A. S. Cacciapuoti, I. F. Akyildiz, and L. Paura, "Correlation-aware user selection for cooperative spectrum sensing in cognitive radio ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 297–306, Feb. 2012.

[11] X. Li, W. Li, and Y. Hei, "Joint spectrum sensing and user selection strategy for cognitive radio networks," in *Proc. IEEE WCSP*, Oct. 2012, pp. 1–6.

[12] M. Monemian and M. Mahdavi, "Sensing user selection based on energy constraints in cognitive radio networks," in *Proc. IEEE WCNC*, Istanbul, Turkey, Apr. 2014, pp. 3379–3384.

[13] Q.-T. Vien, H. X. Nguyen, and A. Nallanathan, "Cooperative spectrum sensing with secondary user selection for cognitive radio networks over Nakagami-m fading channels," *IET Commun.*, vol. 10, no. 1, pp. 91–97, Jan. 2016.

[14] M. Qin, S. Yang, H. Deng, and M. H. Lee, "Enhancing security of primary user in underlay cognitive radio networks with secondary user selection," *IEEE Access*, vol. 6, pp. 32624–32636, 2018.

[15] C. Zhai, J. Liu, L. Zheng, and X. Wang, "Wireless power transfer based spectrum leasing with user selection in cognitive radio networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.

[16] M. Zhang, P. Si, and Y. Zhang, "Optimal secondary user selection scheme for primary users in cognitive radio networks," in *Proc. 2nd Int. Conf. Consum. Electron. Commun. Netw. (CECNet)*, Apr. 2012, pp. 1166–1170.

[17] S. Dadallage, C. Yi, and J. Cai, "Joint beamforming, power, and channel allocation in multiuser and multichannel underlay MISO cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3349–3359, May 2016.

[18] R. Xie, F. R. Yu, and H. Ji, "Joint power allocation and beamforming with users selection for cognitive radio networks via discrete stochastic optimization," *Wireless Netw.*, vol. 18, no. 5, pp. 481–493, Jul. 2012.

[19] S. Chaudhari and D. Cabric, "QoS aware power allocation and user selection in massive MIMO underlay cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 220–231, Jun. 2018.

[20] K. Cumanan, R. Krishna, L. Musavian, and S. Lambotharan, "Joint beamforming and user maximization techniques for cognitive radio networks based on branch and bound method," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 3082–3092, Oct. 2010.

[21] W. Xiong, A. Mukherjee, and H. M. Kwon, "MIMO cognitive radio user selection with and without primary channel state information," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 985–991, Feb. 2016.

[22] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Phys. Commun.*, vol. 4, no. 1, pp. 40–62, Mar. 2011.

[23] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari, and F. Adachi, "Deep learning for physical-layer 5G wireless techniques: Opportunities, challenges and solutions," 2019, *arXiv:1904.09673*. [Online]. Available: https://arxiv.org/abs/1904.09673

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[25] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Aug. 2009.

[26] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[27] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[28] V. Raj, I. Dias, T. Tholeti, and S. Kalyani, "Spectrum access in cognitive radio using a two-stage reinforcement learning approach," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 20–34, Feb. 2018.

[29] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach," *IEEE Access*, vol. 6, pp. 25463–25473, Apr. 2018.

[30] Y. Gwon, S. Dastangoo, and H. T. Kung, "Optimizing media access strategy for competing cognitive radio networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 1215–1220.

[31] X. Zhang, L. Jiao, O.-C. Granmo, and B. J. Oommen, "Channel selection in cognitive radio networks: A switchable Bayesian learning automata approach," in *Proc. IEEE 24th Annu. Int. Symp. Pers. Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 2362–2367.

[32] A. R. Syed, K.-L. A. Yau, J. Qadir, H. Mohamad, N. Ramli, and S. L. Keoh, "Route selection for multi-hop cognitive radio networks using reinforcement learning: An experimental study," *IEEE Access*, vol. 4, pp. 6304–6324, 2016.

[33] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[34] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1823–1834, May 2010.

[35] J. Zhu, Y. Song, D. Jiang, and H. Song, "A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2375–2385, Aug. 2018.

[36] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[37] S. Filippi, O. Cappe, and A. Garivier, "Optimally sensing a single channel without prior information: The tiling algorithm and regret bounds," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 1, pp. 68–76, Feb. 2011.

[38] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, Dec. 2018.

[39] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019.

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[42] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.