Buffer-Aided Relay Selection for Cooperative Hybrid NOMA/OMA Networks With Asynchronous Deep Reinforcement Learning

Chong Huang, Graduate Student Member, IEEE, Gaojie Chen, Senior Member, IEEE, Yu Gong, Member, IEEE, Zhu Han, Fellow, IEEE, and Jonathon A. Chambers, Fellow, IEEE

Abstract—This paper investigates asynchronous reinforcement learning algorithms for joint buffer-aided relay selection and power allocation in the non-orthogonal-multiple-access (NOMA) relay network. With the hybrid NOMA/OMA transmission, we investigate joint relay selection and power allocation to maximize the throughput with the delay constraint. To solve this complicated high-dimensional optimization problem, we propose two asynchronous reinforcement learning-based schemes: the asynchronous deep Q-Learning network (ADQN)-based scheme and the asynchronous advantage actor-critic (A3C)-based scheme, respectively. The A3C-based scheme achieves better performance and robustness when the action space is large, while the ADQN-based scheme converges faster with a small action space. Moreover, a-priori information is exploited to improve the convergence of the proposed schemes. The simulation results show that the proposed asynchronous learning-based schemes can learn from the environment and achieve good convergence.

Index Terms—Buffer-aided relay selection, NOMA, power allocation, asynchronous reinforcement learning.

I. INTRODUCTION

NON-ORTHOGONAL multiple access (NOMA) has attracted much attention in various systems including the Internet of Things (IoT) [1], vehicle-to-everything [2], heterogeneous networks [3], signal detection [4], and edge computing [5]. Of particular interest to this paper is NOMA in cooperative networks. Many works have appeared in this area. For example, in [6], a two-stage max-min relay selection algorithm with fixed power allocation based on users’ quality of service (QoS) and channel quality was proposed to reduce the outage probability. In [7], the outage performance was enhanced by the relay selection scheme with adaptive power allocation in cooperative NOMA networks. In [8], based on the instantaneous channel state, two two-stage schemes with fixed and adaptive power allocation for cooperative NOMA networks were proposed to reduce the outage probability. In [9], the diversity orders of single-state and two-stage relay selection algorithms were obtained with full-duplex (FD)/half-duplex (HD) relays in cooperative NOMA networks.

The buffer-aided relay scheme has been widely applied in cooperative networks to improve the system performance such as the outage probability and throughput [10]–[13]. The buffer-aided relaying technique has been implemented in NOMA transmissions in [14]–[16]. The throughput performance was enhanced by an adaptive transmission mode selection scheme in [14]. In [15], the average transmission rate was improved by using NOMA transmission in a buffer-aided cooperative network. In [16], a buffer-state-based relay selection was proposed to reduce the outage and average packet delay in the NOMA network.

To further improve the throughput, the hybrid NOMA/OMA transmission mode has been investigated in various works. In [17], the buffer-aided relaying system was proposed with adaptive power allocation to improve the outage performance in the hybrid downlink NOMA/OMA network. In [18], based on the instantaneous channel state information (CSI) and the buffer state, the outage probability was reduced by giving high priority to the NOMA transmission mode for uplink buffer-aided networks. In [19], the average sum rate and delay performance were improved by giving higher prioritization to relay-to-user transmissions in the downlink cooperative network with hybrid NOMA/OMA relay-to-user transmissions. Similarly in [20], prioritization-based buffer-aided relay selection was proposed to enhance the throughput and outage performance for downlink hybrid NOMA/OMA cooperative networks. Of particular relevance to this paper are [19] and [20], which also investigate relay selection in the 2-hop
buffer-aided NOMA relay network. In both works, NOMA is only applied over the relay-to-destination hop, while time-division-multiple-access (TDMA) is applied to the source-to-relay hop. By dividing one time slot into two sub-slots, the source can send two data packets in two sub-slots to the selected relay. This, however, doubles the data rate over the corresponding link. In this paper, we assume fixed-rate transmission over every link and TDMA is not applied. Instead, hybrid NOMA/OMA transmission is applied over both the source-to-relay and relay-to-destination hops.

On the other hand, while employing buffers at the relay improves the data throughput, it may increase the packet delay which is a crucial issue in future wireless communications [20]–[22]. In this paper, we will investigate joint buffer-aided relay selection and power allocation in hybrid NOMA/OMA networks to improve the data throughput with the delay constraints. This is a complicated high-dimensional problem. Deep reinforcement learning algorithms such as the deep Q-network (DQN) can be introduced to solve such complicated high-dimensional space problems [23], [24]. Various works have been done in this area. For example, the outage performance was improved by applying a DQN-based relay selection scheme in [25]. In [26], a DQN-based hybrid relay selection algorithm was proposed to improve robustness and speed up learning. Although DQN has high sample efficiency in training, the impact of overestimation is particularly serious in Q-learning algorithms [27], [28]. To solve this problem, we introduce the asynchronous DQN algorithm to overcome the hardships. Moreover, in high-dimensional systems with large action space, the DQN algorithm suffers from the problem of sample complexity, which leads to slow convergence and local maxima [29]. We then investigate the asynchronous advantage actor-critic (A3C) algorithm to further improve the performance of convergence with a large action space [30].

This paper investigates the joint optimization of relay selection and power allocation to maximize the throughput in cooperative NOMA networks, subject to the delay constraint, by using asynchronous reinforcement learning algorithms with a-priori information. The main contributions of this paper are summarized as follows:

- We investigate the hybrid NOMA/OMA transmission mode in buffer-aided cooperative networks. Unlike existing works [19], [20], the hybrid NOMA/OMA transmission is applied over both the source-to-relay and relay-to-user hops, and delay constraint and power allocation are also considered.
- We propose the asynchronous DQN-based and A3C-based schemes to optimize the throughput with the delay constraint. A-priori information is used to reduce the exploration range of the proposed schemes and improve the convergence performance during training.
- The simulation results show that with a-priori information, both the A3C-based scheme (A3C-PI) and asynchronous DQN-based scheme (ADQN-PI) outperform the benchmark. ADQN-PI can achieve high performance with a small action space, while A3C-PI performs significantly with a large action space.

- For comparison, we also show the results of OMA-only and NOMA-only relay selection in the simulation, based on a similar proposed reinforcement learning algorithm. The results show that the proposed hybrid NOMA/OMA scheme takes advantage of both the OMA-only and NOMA-only approaches.

The rest of the paper is organized as follows: Section II introduces the system model and the problem formulation. The tuples of reinforcement learning and a-prieri information are defined in Section III. Section IV investigates the ADQN-PI algorithm. The A3C-PI scheme is proposed in Section V. Simulation results verify the proposed scheme in Section VI. Finally, Section VII concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

As shown in Fig. 1, we consider a two-hop buffer-aided relay network which consists of one source $S$, $K$ HD decode-and-forward (DF) relays $R_k$, where $k \in \{1, \ldots, K\}$, and two users $U_1$ and $U_2$. We assume that there is no direct link between $S$ and the two users as in [31], and each relay node is equipped with two data buffers $B_{k,1}$ and $B_{k,2}$ of size $L$ for $U_1$ and $U_2$, respectively, where $L$ denotes the maximum number of packets stored in a buffer. The buffers follow the First-In-First-Out rule. Moreover, we assume $S$ knows the instantaneous CSI of all channels and the buffer states, and then makes the decision at each time slot as in [32], [33]. All channels are assumed to be quasi-static Rayleigh fading, the channel coefficient $g_{ij} = h_{ij}d_{ij}^{-\beta/2}$ between nodes $i$ and $j$ remains unchanged during one time slot and varies from one time slot to another [13], where $h_{ij}$ is the fading coefficient, $d_{ij}$ is the distance between two nodes, and $\beta$ is the path loss exponent.

A. Transmission Mode

At a given time slot, in terms of $S \rightarrow R_k$ links, the transmission can operate in two modes: OMA and NOMA. For example, if the transmission operates in the OMA mode, one single data packet for user $U_n$, $n \in \{1, 2\}$, can be transmitted from source $S$ to the corresponding buffer $B_{k,n}$ at a single
relay \( R_k \). Therefore, the received signal at relay \( R_k \) is given by
\[
y_{R_k} = \sqrt{P} g_{SR_k} x_S^{(O)} + n_{R_k},
\]
where \( P \) denotes the transmit power at each transmitter, \( x_S^{(O)} \) denotes the packet from source \( S \), and \( n_{R_k} \) denotes the additive-white-Gaussian-noise (AWGN) noise with variance \( \sigma_n^2 \) at relay \( R_k \). Then the channel capacity between source \( S \) and relay \( R_k \) is given by
\[
C_{SR_k} = \log_2 \left( 1 + \frac{P |h_{SR_k}|^2}{d_{SR_k}^2 \sigma_n^2} \right).
\]

Otherwise, if the transmission operates in the NOMA mode, source \( S \) can transmit one packet for user \( U_{n_1} \) to the corresponding buffer \( B_{k_1,n_1} \) of relay \( R_{k_1} \), and one packet for user \( U_{n_2} \) to the corresponding buffer \( B_{k_2,n_2} \) of relay \( R_{k_2} \) simultaneously, where \( k_1, k_2 \in \{1, \ldots, K\} \), \( n_1, n_2 \in \{1,2\} \), \( g_{SR_{k_1}} \leq g_{SR_{k_2}} \). Then the superimposed information symbol at source \( S \) is defined as
\[
x_S^{(N)} = \sqrt{\alpha} x_{SR_{k_1}} + \sqrt{1 - \alpha} x_{SR_{k_2}},
\]
where \( \alpha \in (0,1) \) is the power allocation coefficient, \( x_{SR_{k_1}} \) and \( x_{SR_{k_2}} \) are the signals for relays \( R_{k_1} \) and \( R_{k_2} \), respectively. Then we can obtain the received signals at relays \( R_{k_1} \) and \( R_{k_2} \) as
\[
y_{R_{k_1}} = \sqrt{\alpha} P g_{SR_{k_1}} x_{SR_{k_1}} + \sqrt{1 - \alpha} P g_{SR_{k_1}} x_{SR_{k_2}} + n_{R_{k_1}},
\]
\[
y_{R_{k_2}} = \sqrt{\alpha} P g_{SR_{k_2}} x_{SR_{k_1}} + \sqrt{1 - \alpha} P g_{SR_{k_2}} x_{SR_{k_2}} + n_{R_{k_2}},
\]
respectively, where \( n_{R_{k_1}} \) and \( n_{R_{k_2}} \) denote the AWGN noise with variance \( \sigma_n^2 \) at relays \( R_{k_1} \) and \( R_{k_2} \), respectively. We apply the successive interference cancelation (SIC) scheme in [34]\(^1\) to decode the signal for relay \( R_{k_1} \) first, and then remove \( x_{SR_{k_1}} \) from the received signal at relay \( R_{k_2} \) and decode \( x_{SR_{k_2}} \). Thus, the channel capacity regions for \( S \to R_{k_1} \) and \( S \to R_{k_2} \) are given by
\[
C_{SR_{k_1}} = \log_2 \left( 1 + \frac{\alpha P |h_{SR_{k_1}}|^2 d_{SR_{k_1}}^{-\beta}}{(1 - \alpha) P |h_{SR_{k_1}}|^2 d_{SR_{k_1}}^{-\beta} + \sigma_n^2} \right),
\]
\[
C_{SR_{k_2}} = \log_2 \left( 1 + \frac{(1 - \alpha) P |h_{SR_{k_2}}|^2 d_{SR_{k_2}}^{-\beta}}{\sigma_n^2} \right),
\]
respectively. Notice that because both the signals \( x_{SR_{k_1}} \) and \( x_{SR_{k_2}} \) are decoded at relay \( R_{k_2} \) via the SIC scheme, we consider \( k_1 = k_2 \) is possible for \( S \to R_k \) NOMA transmissions.

On the other hand, the transmissions between relay \( R_k \) and users can also operate in two modes: OMA and NOMA. If the transmission operates in the OMA mode, a single data packet for user \( U_n \), \( n \in \{1,2\} \), can be transmitted from relay \( R_k \) to the corresponding user \( U_n \). The channel capacity between the relay \( R_k \) and user \( U_n \) is given by
\[
C_{R_k U_n} = \log_2 \left( 1 + \frac{P |h_{R_k U_n}|^2 d_{R_k U_n}^{-\beta}}{\sigma_n^2} \right).
\]

Otherwise, in terms of the NOMA mode, the signals for users \( U_1 \) and \( U_2 \) can be superimposed. We assume \( g_{R_k U_{n_1}} \leq g_{R_k U_{n_2}} \), where \( n_1, n_2 \in \{1,2\} \). After applying SIC, we can obtain the channel capacities for \( R_k \to U_{n_1} \) and \( R_k \to U_{n_2} \) as
\[
C_{R_k U_{n_1}} = \log_2 \left( 1 + \frac{\alpha P |h_{R_k U_{n_1}}|^2 d_{R_k U_{n_1}}^{-\beta}}{(1 - \alpha) P |h_{R_k U_{n_1}}|^2 d_{R_k U_{n_1}}^{-\beta} + \sigma_n^2} \right),
\]
\[
C_{R_k U_{n_2}} = \log_2 \left( 1 + \frac{(1 - \alpha) P |h_{R_k U_{n_2}}|^2 d_{R_k U_{n_2}}^{-\beta}}{\sigma_n^2} \right),
\]
respectively.

B. Problem Formulation

For both NOMA and OMA modes in buffer-aided cooperative networks, the link between nodes \( i \) and \( j \) is available for a single data transmission when its channel capacity satisfies
\[
C_{ij} \geq \eta,
\]
where \( \eta \) is the target rate. Moreover, we consider the delay constraint in buffer-aided relay networks. The delay \( \varpi \) for packet \( \varepsilon \) is defined as the period between packet \( \varepsilon \) being successfully transmitted from the source and arriving at the corresponding user. For example, if a packet \( \varepsilon \) is transmitted from source \( S \) to relay \( R_k \) successfully at time slot \( t \), and then successfully arrives at the corresponding user at time slot \( t + 2 \), it takes three time slots \( (t, t+1, t+2) \) for the transmission and its delay \( \varpi = 3 \).

Based on (8), if the link between nodes \( i \) and \( j \) satisfies \( C_{ij} \geq \eta \), the OMA mode can be applied to transmit one single packet from nodes \( i \) to \( j \). On the other hand, if the NOMA mode is applied for a transmit node \( i \) and two selected receiver nodes \( j_1 \) and \( j_2 \), the channel capacities should satisfy \( C_{ij_1} \geq \eta \) and \( C_{ij_2} \geq \eta \) simultaneously.

Furthermore, the optimal power allocation coefficient \( \alpha \) is a crucial criterion in NOMA transmissions. Therefore, to maximize the throughput with delay constraint for the buffer-aided relay hybrid NOMA/OMA networks, we formulate the optimization problem as follows:
\[
O = \max_{\nu, \alpha, i, j, j_1, j_2} \prod_{t=1}^{N} \left( v(t) \cdot \mu(C_{ij}(t) \geq \eta) \cdot \mu(\varpi_1(t) \leq \Delta) + (1 - v(t)) \cdot \mu(C_{ij_1}(t) \geq \eta \& C_{ij_2}(t) \geq \eta) \cdot \left( \mu(\varpi_1(t) \leq \Delta) + \mu(\varpi_2(t) \leq \Delta) \right) \right) u(t),
\]
\[
\text{s.t. } v(t) = 0 \text{ or } 1, \quad u(t) = 0 \text{ or } 1, \quad l_{k,j}(t) > 0 \text{ if } u(t) \cdot v(t) = 1,
\]
\[
\text{(9a) } \quad \text{(9b) } \quad \text{(9c)}
\]
\(^1\)Notice that the imperfect SIC will affect the system performances. As in the existing literature (e.g. [18]–[20], [34]), the perfect SIC assumption provides a clear benchmark scenario for performance comparison. Detailed discussion of the imperfect SIC is beyond the scope of this work and can be found in [35], [36].
\begin{align}
l_{k,u_1}(t) &> 0 \text{ if } u(t) - v(t) = 1, \quad (9d)
l_{k,u_2}(t) &> 0 \text{ if } u(t) - v(t) = 1, \quad (9e)
l_{k,u_1}(t) &\leq L \text{ and } l_{k,u_2}(t) \leq L, \quad (9f)
0 &< \alpha < 1, \quad (9g)
\end{align}

where \( u(t) = 0 \) denotes \( S \rightarrow R_k \) transmissions and \( u(t) = 1 \) denotes \( R_k \rightarrow U_n \) transmissions, \( N \) is the number of time slots, \( v(t) = 1 \) denotes the OMA mode, \( v(t) = 2 \) denotes the NOMA mode, \( \mu(\cdot) = 1 \) if the enclosing function holds and \( \mu(\cdot) = 0 \) otherwise, \( \Delta \) denotes the target delay, \( \varpi_1(t) \) denotes the delay of the packet which arrives at node \( j \) in the OMA transmission mode at time slot \( t \), \( \varpi_2(t) \) and \( \varpi_3(t) \) denote the delays of the packets which arrive at nodes \( j_1 \) and \( j_2 \) in the NOMA transmission mode at time slot \( t \), respectively. \( l_{k,j}(t) \) denotes the buffer state of the corresponding buffer at relay \( R_k \) for user node \( j \) when an OMA \( R_k \rightarrow U_n \) transmission is selected and \( j \) denotes user \( U_n \). \( l_{k,u_1}(t) \) and \( l_{k,u_2}(t) \) denote the buffer state of the two buffers at relay \( R_k \). (9a) shows the selection of NOMA/OMA for the transmission at time slot \( t \). (9b) shows the selection of \( S \rightarrow R_k \) or \( R_k \rightarrow U_n \) transmission. (9c)–(9e) show that the corresponding buffer should not be empty for \( R_k \rightarrow U_n \) transmissions, while (9f) shows that each buffer can store no more than \( L \) packets. (9g) denotes the range of power allocation coefficient in the NOMA mode. To be specific, if the OMA mode is selected for \( R_k \rightarrow U_n \) transmissions, the buffer at relay \( R_k \) for user \( U_n \) should not be empty. On the other hand, if the NOMA mode is selected for \( R_k \rightarrow U_n \) transmissions, both the two buffers at relay \( R_k \) should not be empty. The optimization problem function ensures that the capacity should satisfy (8) for each selected transmission link, and throughput only includes the packets arriving at the corresponding user with the delay constraint.

According to (9), to maximize the throughput with delay constraint, we need to optimize the selection of transmission mode and the corresponding nodes at each given time slot. When the NOMA transmissions are not available, or available NOMA transmissions are not the optimal selection when considering the delay constraint at a given time slot, the system can switch to the OMA mode to reduce the outage.

For the NOMA transmission mode, the optimal power allocation coefficient \( \alpha \) is required to guarantee the availability of links for data transmission. In terms of the transmit node, if we choose a \( S \rightarrow R_k \) transmission with NOMA, we need to perform relay pairing to select two available relays to receive packets from \( S \). On the other hand, in a successful \( R_k \rightarrow U_n \) NOMA transmission, relay \( R_k \) can transmit one packet to each user from the corresponding buffer. Therefore, the optimal decision should also consider the impact of buffer state to maximize the throughput.

Moreover, the delay constraint is introduced into the optimization function. To be specific, in a time-varying system, the optimal buffer states for maximizing the throughput are affected by the channel states, and the transmission mode selection should satisfy the delay constraint. Due to the relay pairing, power allocation, NOMA/OMA switching, delay constraint, buffer states and time-varying system in the proposed network, the optimization function in (9) is a complicated non-convex high-dimensional optimization problem to achieve the long-term performance.

III. TUPLES OF REINFORCEMENT LEARNING AND A-PRIORI INFORMATION

A. Overview

To solve the complicated non-convex high-dimensional optimization problem in (9), we apply reinforcement learning algorithms in buffer-aided relay hybrid NOMA/OMA networks to achieve the long-term performance. Reinforcement learning generally consists of states, actions and rewards, which need to be defined carefully for the implementation in buffer-aided relay hybrid NOMA/OMA networks. Moreover, a-priori information can help reinforcement learning schemes reduce the action-state space and improve the convergence.

B. State

The system state of the buffer-aided relay hybrid NOMA/OMA network is characterized by the buffer state and the channel states. The buffer state \( l_{k,u_1}(t) \) denotes the number of packets stored in the buffer at relay \( R_k \) for user \( U_n \), where \( n \in \{1, 2\} \) and \( k \in \{1, K\} \). In terms of the channel states, directly applying the CSI in reinforcement learning will cause an infinite number of states due to the continuous values of CSI. Thus, the channel state can be processed by

\[ \tau(t) = \{c_1, ..., c_K, up_{12}, up_{13}, ..., up_{K-1K}\} \]

where \( c_k \) denotes the availabilities for the \( S \rightarrow R_k \) OMA transmission and \( R_k \rightarrow U_n \) NOMA/OMA transmissions at time slot \( t \), \( up_{ki,k2} \) denotes the availability for the NOMA transmission between source \( S \) and the relay pair \( \{R_{ki}, R_{k2}\} \). Therefore, the system state is given by

\[ s(t) = \{l_{1,u_1}(t), l_{1,u_2}(t), ..., l_{K,u_1}(t), l_{K,u_2}(t), \tau(t)\}. \]

In reinforcement learning, the environment can transit from a state \( s(t) \) to the next state \( s(t+1) \) by taking an action, the action set will be defined in the following subsection.

C. Action

In buffer-aided relay hybrid NOMA/OMA networks, an action is not only to select links for data transmission but also to decide the transmission mode. Moreover, an action also includes the value of power allocation coefficient \( \alpha \) for the NOMA transmission mode. Furthermore, in the NOMA mode for \( S \rightarrow R_k \) transmissions, the relay pair selection should be considered in an action. Therefore, we define the action as \( a(t) = a_{r,\alpha,i,j_1,j_2} \), where \( u = 0 \) denotes the NOMA mode and \( u = 1 \) denotes the OMA mode, \( \alpha \in \{0, 1\} \), \( i \) denotes the transmit node, \( j \) denotes the receiver node in the OMA mode, \( j_1 \) and \( j_2 \) denote the receiver nodes in the NOMA mode, respectively. To achieve stable convergence, we consider a discrete action space the same as in [37], and the power levels coefficient is \( \delta \), thus we can obtain

\[ \alpha \in \left\{ \frac{1}{\delta + 1}, \frac{2}{\delta + 1}, ..., \frac{\delta}{\delta + 1} \right\}. \]
Based on the action and state, the reinforcement algorithm can be applied to make decisions in the buffer-aided relay hybrid NOMA/OMA network. To optimize the decisions to achieve the maximum throughput with delay constraint, we introduce the reward to help train the reinforcement algorithm.

**D. Reward**

Reinforcement learning algorithms require a set of rewards to evaluate the action-state space. In the proposed networks, the reward is designed to help the reinforcement learning algorithms maximize the throughput with delay constraint. We consider giving a positive bonus as the reward if there is a packet arriving at the corresponding user within the target delay. Furthermore, negative rewards are considered when unavailable links are selected for data transmissions. Therefore, the reinforcement learning algorithm can learn the experience from the total reward during training and evaluate the action-state space.

**E. A-Priori Information**

Reinforcement learning needs to explore the state-action space and evaluate the rewards during training. However, based on the states in (11) and actions in Section III-C, we can obtain a high-dimensional action-state space. Therefore, the range of exploration is quite large in the proposed network, and a high-dimensional action-state space. Therefore, the range of exploration and improve the convergence efficiency, a-priori information is introduced for the proposed schemes.

For a given state, the action space is still large. However, we can remove the invalid actions for a given state to reduce the exploration range. We assume that an action for the OMA transmission between nodes $i$ and $j$ is valid at time slot $t$ when it satisfies the requirements as follows:

$$C_{ij}(t) \geq \eta,$$

$$l_{k,U_n}(t) < L \text{ for } S \rightarrow R_k, \ k \in \{1, \ldots, K\},$$

$$l_{k,U_n}(t) > 0 \text{ for } R_k \rightarrow U_n, \ k \in \{1, \ldots, K\},$$

$$n \in \{1, 2\}.$$  

Therefore, the OMA transmission requires an available link for single packet transmission. Moreover, an empty buffer is invalid for $R_k \rightarrow U_n$ transmissions, whilst a full buffer is invalid for $S \rightarrow R_k$ transmissions. On the other hand, we assume that an action for the NOMA transmission from node $i$ to nodes $j_1$ and $j_2$ is valid at time slot $t$ when it satisfies the following requirements:

$$C_{ij_1}(t) \geq \eta,$$

$$C_{ij_2}(t) \geq \eta,$$

$$l_{R_k,U_n}(t) < L \text{ for } S \rightarrow R_k, \ k \in \{1, \ldots, K\},$$

$$l_{R_k,U_n}(t) > 0 \text{ for } R_k \rightarrow U_n, \ k \in \{1, \ldots, K\},$$

$$n \in \{1, 2\}.$$  

where $U_{n_1}$ and $U_{n_2}$ denote the corresponding user for the selected buffer in NOMA transmissions. A valid NOMA transmission requires two available links for the transmit node. Moreover, the corresponding buffer is not empty for $R_k \rightarrow D_m$ transmissions, and not full for $S \rightarrow R_k$ transmissions.

Furthermore, we consider that the buffer state can also be a priori information. By considering the target delay, a trade-off between different buffer states is required to ensure the delay constraint. Moreover, the channel gains also have an impact on the delay. Strong links may be always available, while weak links have low probability to be available.

Therefore, we introduce the target buffer length $\xi_{k,n}$ for user $U_n$ at relay $R_k$ to improve the convergence performance of reinforcement learning, where $n \in \{1, 2\}$. $\xi_{k,n}$ can be expressed as

$$\xi_{k,n} = \min \left( \frac{\Delta \log_2(1 + Pd_{SD}^\beta_R)}{\sum_{k=1}^K \log_2(1 + Pd_{SD}^\beta_R)}, L \right).$$  

If the buffer state $l_{k,n}(t) > \xi_{k,n}$ at time slot $t$, we assume the corresponding actions, which transmit a packet to buffer $n$ at relay $R_k$, are invalid.

Based on the priori information in (13), (14) and (15), we can remove invalid actions from the actions set at a given time slot. Therefore, the exploration efficiency is improved by reducing the action-state space. With high exploration efficiency, the proposed algorithm can improve the performance and converge faster, which will be shown in the simulation section.

**IV. ASYNCHRONOUS DQN-BASED SCHEME**

In this section, we propose the asynchronous DQN-based scheme (ADQN) for joint buffer-aided relay selection and power allocation. Moreover, we further develop the ADQN scheme with the priori information (ADQN-PI). In the DQN, there is an agent that takes the action and obtains the next state for the system. With applying the $\epsilon$-greedy strategy, the agent can explore the system randomly in the exploration mode. If in the exploitation mode, the agent takes the action estimated from the neural network. A DQN consists of two neural networks: the prediction and target deep neural networks. The prediction network is used to estimate the current action for the current state in the exploitation mode, and the target network is used to estimate the target value for updating the prediction network. The target value function in the DQN is given by

$$Q = r_{s(t),a(t)} + \rho \left( \max_a Q^\text{Tar}(s(t+1), a) \right),$$  

where $r_{s(t),a(t)}$ denotes the reward for $s(t)$ and $a(t)$, $\rho$ denotes the discount coefficient, $Q^\text{Tar}(s(t+1), a)$ denotes the estimated Q-value from the target network. Therefore, the prediction network can update its weights by minimizing the loss between the true value and the target value. The loss function of the DQN-based scheme is given by

$$\zeta_{DQN}(t) = \left( r_{s(t),a(t)} + \rho \left( \max_a Q^\text{Tar}(s(t+1), a) \right) - Q^\text{Pre}(s(t), a(t)) \right)^2,$$  

where $Q^\text{Pre}(s(t), a(t))$ denotes the true value estimated from the prediction network. Then, we apply a widely used scheme
Adam [38] to minimize the loss and obtain the gradient to update the weights of the prediction neural network. To update the target network, we copy the parameters from the prediction network to the target network periodically.

To solve the problem of overestimation in the conventional DQN, we apply the multi-threaded asynchronous DQN scheme to achieve better convergence [30]. We design multi agents for the asynchronous scheme, each agent has its own copy of the environment. The agents can use the same networks to explore the same environment in parallel. However, because we apply the $\epsilon$-greedy strategy and the channel coefficients vary independently from one time slot to another, the agents can explore different parts of the environment in their own threads. The framework of the asynchronous reinforcement learning is shown in Fig. 2. To be specific, in the proposed asynchronous DQN scheme, the two neural networks are shared with all agents. At time slot $t$, a local agent applies the $\epsilon$-greedy strategy with the priori information to explore the environment, and then obtains the current state $s(t)$, the current action $a(t)$, the corresponding reward $r(t)$, and the next state $s(t+1)$. Then the agent forms these elements as a sample $(s(t), a(t), r(t), s(t+1))$. The loss from this sample can be calculated based on (17), and the Adam method is used to obtain the gradient based on the loss. Then after exploring the environment for $W_P$ time slots, the total gradient from this agent can be used to update the prediction neural networks. Due to the multi agents in the asynchronous DQN algorithm, the algorithm can obtain a different gradient to update the network. After updating the prediction network $W_T$ times, we copy the parameters from the prediction network to the target network. The pseudo code of the proposed asynchronous DQN based scheme with priori information is shown in Algorithm 1.

**Algorithm 1 ADQN-PI:**

1. Initialize the variables.
2. Initialize the shared prediction network $\theta_p$ and target network $\theta_t$.
3. Initialize the thread-specific prediction network $\theta_p'$ and target network $\theta_t'$.
4. repeat for each agent thread:
   5. for $w = 1, \ldots, W_T$ do
   6.   Synchronize the thread-specific networks $\theta_p' = \theta_p$ and $\theta_t' = \theta_t$.
   7. for $t = 1, \ldots, W_P$ do
   8.   Use the $\epsilon$-greedy strategy to decide the exploration/exploitation mode, and then select $a(t)$ based on $\theta_p'$ or randomly.
   9.   Get the reward $r_s(t,a(t))$ and the next state $s(t+1)$.
   10. Generate sample $(s(t), a(t), r_{s(t),a(t)}, s(t+1))$.
   11. Calculate the loss $\zeta(t)$ of sample, based on the loss function (17) with $\theta_p'$ and $\theta_t'$.
   12. Obtain the gradient $\nu(t)$ based on $\zeta(t)$ by using the Adam method.
   13. Accumulate gradients $\nu = \nu + \nu(t)$.
   14. end for
   15. Asynchronous update $\theta_p$ with $\nu$.
   16. $\nu = 0$.
   17. end for
18. Update $\theta_t$ by using $\theta_p$.
19. until final convergence.

Thus, compared with the conventional DQN, the asynchronous DQN-based scheme can not only explore more parts of the environment within one training iteration, but also solve the problem of correlation between successive experiences and overestimation. Notice that DQN is a value-based reinforcement learning algorithm which has high sample efficiency [27]. Therefore, DQN can converge well in environments with small action space. However, it propagates the impact of rewards to the related action-state space by using the target value. Though we apply the asynchronous method to the DQN, the robustness and stability are still open problems for the proposed DQN-based scheme in buffer-aided cooperative NOMA networks with a large action space. Therefore, we will consider the A3C-based scheme to achieve high performance with a large action space in the next section.

**V. A3C-BASED SCHEME**

Because policy-based reinforcement learning algorithms are effective and stable for function approximation, we consider to combine the advantages of the value-based scheme and the policy-based scheme to improve the convergence performance and robustness of training. Therefore, we introduce the actor-critic scheme to solve the optimization problem in (9). In the actor-critic algorithm, there are two networks: the actor network and the critic network. Unlike using the prediction and target networks in the DQN algorithm, the actor network uses the actor network to evaluate the probabilities for the action-state space based on the policy, and the critic network...
is used to evaluate the advantage for each action-state pair. To be specific, the input of both networks is the current state $s(t)$ at time slot $t$, the outputs of the actor network are the probabilities of all actions for $s(t)$, and the output of the critic network is the state value $Q$ which is used to determine the average value of $s(t)$. Thus, the estimated value function in A3C is given by

$$V(s(t)) = r(s(t), a(t)) + \gamma V(s(t+1), a(t+1)) + \cdots + \gamma^{N-1} V(s(t+N-1), a(t+N-1)) + \gamma^{N} Q(s(t+N); \theta_c),$$

where $\gamma$ is the discount coefficient, and $\theta_c$ is the weights matrix of the critic network.

Therefore, we can use the average value and the estimated value to evaluate the advantage of taking action $a(t)$ for state $s(t)$ at time slot $t$. The advantage function is given by

$$A(s(t)) = V(s(t)) - Q(s(t); \theta_c),$$

which can help the agent understand the direction of updating networks. To be specific, the advantage function is used to evaluate the advantage or disadvantage of actions for the policy from the actor network.

The actor network is a deep neural network with the weights matrix $\theta$ and policy $\pi$. The actor network aims to optimize the policy $\pi$, which is the action probability sets for each state, to achieve the maximum throughput with the delay constraint. In the actor-critic algorithm, the action with the maximum probability is selected for the corresponding state. The loss function in the actor network is given by

$$\zeta_A(t) = \log \pi(a(t), s(t); \theta) \left( V(s(t)) - Q(s(t); \theta_c) \right),$$

The critic network can provide the value function to obtain the loss of the actor network in (20). It aims to evaluate the policy $\pi$ from the actor network. The loss function in the critic network is given by

$$\zeta_C(t) = \left( V(s(t)) - Q(s(t); \theta_c) \right)^2.$$

Moreover, we apply the RMSProp method [41], which is a widely used optimization method for deep neural networks, to obtain the gradients based on the loss in the actor-critic algorithm [30]. The framework of the actor-critic algorithm is shown in Fig. 3.

To improve the convergence efficiency and performance for the actor-critic algorithm in training, the asynchronous advantage actor-critic (A3C) method was introduced [30]. Similar with ADQN, multi agents are designed for the actor-critic algorithm, where each agent has its own copy of the environment and can use its own copy of the shared networks to explore its environment in parallel.

To be specific, each agent performs actor-critic learning asynchronously in its thread. At a given time slot $t$, a local agent can use its actor network to estimate the action $a(t)$ with policy $\pi$. Notice that the priori information is introduced to help the agent avoid invalid actions. After receiving the action $a(t)$, the environment takes $a(t)$ to change to the next state $s(t+1)$, and then obtains the corresponding reward $r_{s(t), a(t)}$, and feeds back the reward to the agent. Finally, the agent generates a sample as $\{s(t), a(t), r_{s(t), a(t)}\}$ and stores it to the memory.

After repeating $W_A$ time slots, the local agent can obtain the state value $Q(s(W_A))$ from the critic network, and then the estimated value for each state within $W_A$ time slots can be calculated based on (18). Therefore, the losses of the action network and critic networks can be obtained based on (20) and (21). Then, the local agent applies the RMSProp method to calculate the gradients and accumulate them. The shared networks are updated with the accumulated gradient from each agent asynchronously, and then the updated weights are sent to each agent. The Pseudo code of the proposed A3C-based scheme with priori information is shown in Algorithm 2. We build the two deep neural networks, which consist of three fully-connected layers with 256, 128 and 128 neurons, respectively. The computational complexity of the proposed A3C-based algorithm for training is $W_A$ as the number of iterations of loops in [39], [40].
Remark 1: From (20) and (21), it is clear that the actor network and critic network are trained separately. With training iteration increase, the actor network learns to optimize its policy to find the optimal estimation of the action-state space, while the critic network learns to evaluate the actions estimated from the actor network efficiently. Therefore, the critic network can help the actor network converge more stably. Moreover, compared with conventional reinforcement learning algorithms, the asynchronous method can help them explore the environment more efficiently and converge faster [30]. However, it is difficult to let both the actor network and critic network converge well, when the action space is not large to generate a sufficient number of samples for evaluating the action-state pairs. This analysis is verified in the simulation section.

The computational complexity of the proposed reinforcement learning at the training stage depends on the layer structure of the neural networks and the amount of the training data. After the training stage, the neural network can immediately output the relay selection decision given the input states, making it an attractive scheme in communications.

VI. SIMULATION RESULTS

The results of the proposed asynchronous learning-based schemes are shown in this section. The simulation parameters include system parameters and learning parameters for the buffer-aided cooperative NOMA systems and the deep reinforcement learning algorithm. Unless otherwise stated, the system parameters are set as follows: the number of relays $K = 4$, the buffer size $L = 10$, the transmit power to noise ratio for all transmitters $P/\sigma_n^2 = 20$ dB, the number of power levels $\delta = 10$, the path loss exponent $\beta = 3$, the target rate $\eta = 1.1$ bps/Hz, and the target delay $\Delta = 10$ time slots. The locations of $S, U_1, U_2, R_1, R_2, R_3$ and $R_4$ are $(0, 0)$ m, $(4.8, 0.1)$ m, $(4.9, -0.3)$ m, $(2.4, 0.1)$ m, $(2.5, -1.1)$ m, $(2.4, 0.7)$ m and $(2.5, -1.2)$ m, respectively. On the other hand, the learning parameters are set as follows: the discount coefficients $\rho = 0.9$, the training sample number $W_p = W_A = 100$, and the iteration number of updating the target network $W_T = 100$.

Moreover, as there is no existing proper benchmark, we design a Max-Min-SNR joint relay selection and power allocation scheme as the benchmark. The Max-Min-SNR scheme uses the discrete power level coefficient $\delta$ for power allocation, and can switch the NOMA/OMA mode for transmissions in both two hops. The decision from the Max-Min-SNR scheme is carried out as follows.

1) First, the scheme checks all valid NOMA mode actions. In each NOMA mode action, two links are selected for transmission, and we define the one with smaller SNR as a “weak link”. Then the priority is given to the NOMA mode action which has the strongest “weak link” among all “weak links”.
2) If all NOMA transmission actions are invalid, the priority is given to the valid OMA transmission action which uses the link with the strongest SNR.
3) Otherwise, no actions can be taken.

Fig. 4 shows the throughput vs. training iterations for the proposed asynchronous learning based schemes with the delay constraint, where the number of power levels $\delta = 10$ and relay number $K = 4$. It is clearly shown that the A3C-PI scheme converges to 0.85 packet/time slot after 3,000 training iterations, while the A3C-based scheme only achieves about 0.7. Without the priori information, the reinforcement learning based scheme needs to explore a large action-state space to achieve the optimal solution, and it is much more difficult to avoid the local optima problem. Furthermore, a large numbers of relays and power levels lead to large action space, and both the A3C-PI and the A3C-based scheme can outperform the DQN-PI and ADQN schemes in this case. These results shows that with a large action space, the A3C algorithm improves the convergence performance by combining the advantages of the value-based scheme and the policy-based scheme.

As we can see in Fig. 5, the ADQN-PI scheme converges to 0.48 packets/time slot after 2,000 training iterations, while the A3C-PI scheme only achieves about 0.43. Compared with Fig. 4, we reduce the action space by decreasing the number of relays and power levels to improve the convergence efficiency, and the DQN-based scheme can achieve better results than the A3C-based scheme due to its high sample efficiency. It can be seen that both two proposed algorithms converge unstably in a small action space, because of the impact of the small
number of power levels. Moreover, both the ADQN-PI and the A3C-PI schemes can achieve higher throughput than the other two schemes without the a-priori information, this result confirms that reducing the range of exploration can still help reinforcement learning algorithms converge with a small action space.

The results in Fig. 6 show the comparison of throughput with delay constraint vs. different target rates between the proposed schemes and benchmark. All proposed schemes outperform Max-Min-SNR significantly, the A3C-PI and ADQN-PI achieve the throughput of 0.89 and 0.6 with the target rate $\eta = 1$ bps/Hz, respectively, while Max-Min-SNR only achieves 0.09. This clearly indicates that the proposed schemes can optimize the throughput with the delay constraint. Moreover, the A3C-PI scheme with fixed $\alpha = 0.7$ only achieves the throughput of 0.5 when the target rate $\eta = 1$ bps/Hz, while the ADQN-PI scheme with fixed $\alpha = 0.7$ achieves only 0.38. The results show that the power allocation can improve the throughput in NOMA transmissions.

Fig. 7 indicates the impact of target delay for the proposed schemes and Max-Min-SNR scheme. The A3C-PI and ADQN-PI achieve throughputs of 0.85 and 0.57 when the target delay $\Delta = 10$ time slots, while Max-Min-SNR only obtains 0.09. One of the advantages of the proposed scheme is that the learning based algorithm can search the solution in environments with different requirements. Therefore, with different target delay, the proposed schemes can achieve high throughput with delay constraint. Moreover, both ADQN-PI and A3C-PI can perform better than learning-based schemes with fixed $\alpha$. It shows that the power allocation is an efficient way to improve the throughput. Notice that all schemes can only achieve throughput with delay constraint when $\Delta \geq 2$, because a packet takes at least two time slots to arrive at the corresponding user.

Fig. 8 shows the throughput vs. transmit power to noise ratio $P/\sigma_n^2$ for the proposed schemes and Max-Min-SNR scheme with the delay constraint. It is clearly shown that A3C-PI and ADQN-PI achieve throughputs of 0.88 and 0.6 when $P/\sigma_n^2 = 25$ dB, while Max-Min-SNR only achieves the throughput of 0.1. Due to the requirement of the target rate, the low value of $P/\sigma_n^2$ leads to a large number of outages in transmissions. With low transmit power, many links are not available for transmission at a given time slot. Although the proposed scheme can learn to optimize the action to avoid selecting unavailable links for transmission at a given time slot, a packet requires many time slots to arrive at the corresponding user, and the delay is difficult to be minimized.

Fig. 9 shows the throughput with the delay constraint vs. power level coefficient $\delta$ for the proposed schemes and Max-Min-SNR scheme. It is clear that the A3C-PI and ADQN-PI schemes achieve throughputs of 0.7 and 0.5 when $\delta = 4$, while Max-Min-SNR only achieves the throughput of 0.06.
The results indicate that higher throughput is achieved with a larger power level coefficient, because the power allocation can perform more efficient in a discrete action space with a large $\delta$. We can observe the unstable uptrends for all schemes, because the higher discrete power allocation coefficient may not provide a corresponding larger action space frequently.

Fig. 10 compares the delay-constrained throughput between the hybrid NOMA/OMA, NOMA-only and OMA-only schemes. The NOMA-only and OMA-only schemes only apply the NOMA and OMA transmission, respectively, based on using ADQN similar to that in the hybrid NOMA/OMA scheme. It is clearly shown that, at high $P/\sigma_n^2$, the hybrid scheme performs similar to the NOMA-only approach. This is because when the channel SNR is high, NOMA will be selected by the learning to ensure high data throughput. On the other hand, at low $P/\sigma_n^2$, the hybrid scheme performs similar to the OMA-only approach. This also matches the intuition that the channels with low SNR will not support the NOMA transmission, and the learning process has successfully captured this. Therefore, the proposed hybrid NOMA/OMA scheme takes advantages of both the OMA-only and NOMA-only approaches.

VII. CONCLUSION

This paper proposed two asynchronous learning algorithms for joint hybrid NOMA/OMA relay selection and power allocation in buffer-aided delay-constrained networks. We compared the asynchronous DQN-based scheme and the A3C-based scheme with different action spaces. A-priori information is exploited to further improve the learning process. The comparison shows that the A3C-PI scheme outperforms ADQN-PI for large action spaces, and the ADQN-PI outperforms A3C-PI for small action spaces. The simulation results also showed the advantages of exploiting the priori information in the learning. Finally, in our future work, we will investigate the scalability of the proposed learning scheme in dynamic scenarios with varying parameters, including relays numbers, buffer lengths and others.

REFERENCES

[1] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, “A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends,” IEEE J. Sel. Areas Commun., vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[2] B. Di, L. Song, Y. Li, and Z. Han, “V2X meets NOMA: Non-orthogonal multiple access for 5G-enabled vehicular networks,” IEEE Wireless Commun., vol. 24, no. 6, pp. 14–21, Dec. 2017.

[3] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, “Spectrum allocation and power control for non-orthogonal multiple access in HetNets,” IEEE Trans. Wireless Commun., vol. 16, no. 9, pp. 5825–5837, Sep. 2017.

[4] T. Wang, S. Liu, F. Yang, J. Wang, J. Song, and Z. Han, “Generalized spatial modulation-based multi-user and signal detection scheme for terrestrial return channel with NOMA,” IEEE Trans. Broadcast., vol. 64, no. 2, pp. 211–219, Jun. 2018.

[5] Q.-V. Pham, H. T. Nguyen, Z. Han, and W.-J. Hwang, “Coalitional games for computation offloading in NOMA-enabled multi-access edge computing,” IEEE Trans. Veh. Technol., vol. 69, no. 2, pp. 1982–1993, Feb. 2020.

[6] Z. Ding, H. Dai, and H. V. Poor, “Relay selection for cooperative NOMA,” IEEE Wireless Commun. Lett., vol. 5, no. 4, pp. 416–419, Aug. 2016.

[7] Z. Yang, Z. Ding, Y. Wu, and P. Fan, “Novel relay selection strategies for cooperative NOMA,” IEEE Trans. Veh. Technol., vol. 66, no. 11, pp. 10114–10123, Nov. 2017.

[8] P. Xu, Z. Yang, Z. Ding, and Z. Zhang, “Optimal relay selection schemes for cooperative NOMA,” IEEE Trans. Veh. Technol., vol. 67, no. 8, pp. 7851–7855, Aug. 2018.

[9] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, “Spatially random relay selection for full/half-duplex cooperative NOMA networks,” IEEE Trans. Commun., vol. 66, no. 8, pp. 3294–3308, Aug. 2018.

[10] Z. Tian, G. Chen, Y. Gong, Z. Chen, and J. A. Chambers, “Buffer-aided max-link relay selection in amplify-and-forward cooperative networks,” IEEE Trans. Veh. Technol., vol. 64, no. 2, pp. 553–565, Feb. 2015.

[11] G. Chen, Z. Tian, Y. Gong, and J. Chambers, “Decode-and-forward buffer-aided relay selection in cognitive relay networks,” IEEE Trans. Veh. Technol., vol. 63, no. 9, pp. 4723–4728, Nov. 2014.

[12] Z. Tian, Y. Gong, G. Chen, Z. Chen, and J. Chambers, “Buffer-aided link selection with network coding in multihop networks,” IEEE Trans. Veh. Technol., vol. 65, no. 9, pp. 7195–7206, Sep. 2016.

[13] Z. Tian, Y. Gong, G. Chen, and J. A. Chambers, “Buffer-aided relay selection with reduced packet delay in cooperative networks,” IEEE Trans. Veh. Technol., vol. 66, no. 3, pp. 2567–2575, Mar. 2017.

[14] S. Luo and K. C. Teh, “Adaptive transmission for cooperative NOMA system with buffer-aided relaying,” IEEE Commun. Lett., vol. 21, no. 4, pp. 937–940, Apr. 2017.

[15] H. Cao, J. Cai, S. Huang, and Y. Lu, “Online adaptive transmission strategy for buffer-aided cooperative NOMA systems,” IEEE Trans. Mobile Comput., vol. 18, no. 5, pp. 1133–1144, May 2019.

[16] J. Kochi, R. Nakai, and S. Sugiyama, “Performance evaluation of generalized buffer-state-based relay selection in NOMA-aided downlink,” IEEE Access, vol. 7, pp. 173320–173328, 2019.

[17] Q. Zhang, Z. Liang, Q. Li, and J. Qin, “Buffer-aided non-orthogonal multiple access relaying systems in Rayleigh fading channels,” IEEE Trans. Commun., vol. 65, no. 1, pp. 95–106, Jan. 2017.

[18] P. Xu, J. Quan, Z. Yang, G. Chen, and Z. Ding, “Performance analysis of buffer-aided hybrid NOMA/OMA in cooperative uplink system,” IEEE Access, vol. 7, pp. 168759–168773, 2019.

[19] N. Nomikos, T. Charalambous, D. Vouyloukas, G. K. Karagiannidis, and R. Wichman, “Hybrid NOMA/OMA with buffer-aided relay selection in cooperative networks,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 3, pp. 524–537, Jun. 2019.

[20] M. Alkhwawatrah, Y. Gong, G. Chen, S. Lambotharan, and J. A. Chambers, “Buffer-aided relay selection for cooperative NOMA in the Internet of Things,” IEEE Internet Things J., vol. 6, no. 3, pp. 5722–5731, Jun. 2019.

[21] N. Zlatanov, A. Ihleif, T. Islam, and R. Schober, “Buffer-aided cooperative communications: Opportunities and challenges,” IEEE Commun. Mag., vol. 52, no. 4, pp. 146–153, Apr. 2014.

[22] Y. Gong, G. Chen, and T. Xie, “Using buffers in trust-aware relay selection networks with spatially random relays,” IEEE Trans. Wireless Commun., vol. 17, no. 9, pp. 5818–5826, Sep. 2018.

[23] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, pp. 529–533, 2015.

[24] C. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.

[25] Y. Su, X. Lu, Y. Zhao, L. Huang, and X. Du, “Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks,” IEEE Sensors J., vol. 19, no. 20, pp. 9561–9569, Oct. 2019.
Y. Zou, Y. Xie, C. Zhang, S. Gong, D. T. Hoang, and D. Niyato, “Optimization-driven hierarchical deep reinforcement learning for hybrid relaying communications,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Seoul, South Korea, May 2020, pp. 1–6.

O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, “Bridging the gap between value and policy based reinforcement learning,” in Proc. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, vol. 30, Dec. 2017, pp. 2775–2785.

H. V. Hasselt, “Double Q-learning,” in Proc. Neural Inf. Process. Syst. (NIPS), Vancouver, BC, Canada, vol. 23, Dec. 2010, pp. 2613–2621.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in Proc. ICML, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 1861–1870.

V. Mnih et al., “Asynchronous methods for deep reinforcement learning,” in Proc. 33rd Int. Conf. Mach. Learn., New York, NY, USA, Jun. 2016, pp. 1928–1937.

R. Nakai and S. Sugiura, “Physical layer security in buffer-state-based max-ratio relay selection exploiting broadcasting with cooperative beamforming and jamming,” IEEE Trans. Inf. Forensics Security, vol. 14, no. 2, pp. 431–444, Feb. 2019.

M. Darabi, V. Jamali, B. Maham, and R. Schober, “Adaptive link selection for cognitive buffer-aided relay networks,” IEEE Commun. Lett., vol. 19, no. 4, pp. 693–696, Apr. 2015.

R. Zhang, R. Nakai, K. Sezaki, and S. Sugiura, “Generalized buffer-state-based relay selection in cooperative cognitive radio networks,” IEEE Access, vol. 8, pp. 1164–11657, 2020.

Z. Ding, Z. Yang, P. Fan, and H. V. Poor, “On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users,” IEEE Signal Process. Lett., vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

J.-M. Kang, I.-M. Kim, and C.-I. Chun, “Deep learning-based MIMO-NOMA with imperfect SIC decoding,” IEEE Syst. J., vol. 14, no. 3, pp. 3414–3417, Sep. 2020.

Y. Zou, Y. Xie, C. Zhang, S. Gong, D. T. Hoang, and D. Niyato, “Optimization-driven hierarchical deep reinforcement learning for hybrid relaying communications,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Seoul, South Korea, May 2020, pp. 1–6.

O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, “Bridging the gap between value and policy based reinforcement learning,” in Proc. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, vol. 30, Dec. 2017, pp. 2775–2785.

H. V. Hasselt, “Double Q-learning,” in Proc. Neural Inf. Process. Syst. (NIPS), Vancouver, BC, Canada, vol. 23, Dec. 2010, pp. 2613–2621.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in Proc. ICML, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 1861–1870.

V. Mnih et al., “Asynchronous methods for deep reinforcement learning,” in Proc. 33rd Int. Conf. Mach. Learn., New York, NY, USA, Jun. 2016, pp. 1928–1937.

R. Nakai and S. Sugiura, “Physical layer security in buffer-state-based max-ratio relay selection exploiting broadcasting with cooperative beamforming and jamming,” IEEE Trans. Inf. Forensics Security, vol. 14, no. 2, pp. 431–444, Feb. 2019.

M. Darabi, V. Jamali, B. Maham, and R. Schober, “Adaptive link selection for cognitive buffer-aided relay networks,” IEEE Commun. Lett., vol. 19, no. 4, pp. 693–696, Apr. 2015.

R. Zhang, R. Nakai, K. Sezaki, and S. Sugiura, “Generalized buffer-state-based relay selection in cooperative cognitive radio networks,” IEEE Access, vol. 8, pp. 1164–11657, 2020.

Z. Ding, Z. Yang, P. Fan, and H. V. Poor, “On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users,” IEEE Signal Process. Lett., vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

J.-M. Kang, I.-M. Kim, and C.-I. Chun, “Deep learning-based MIMO-NOMA with imperfect SIC decoding,” IEEE Syst. J., vol. 14, no. 3, pp. 3414–3417, Sep. 2020.

A. S. de Sena et al., “Massive MIMO-NOMA networks with imperfect SIC: Design and fairness enhancement,” IEEE Trans. Wireless Commun., vol. 19, no. 9, pp. 6100–6115, Sep. 2020.

C. He, Y. Hu, Y. Chen, and B. Zeng, “Joint power allocation and channel assignment for NOMA with deep reinforcement learning,” IEEE J. Sel. Areas Commun., vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.

F. Hussain, R. Hussain, A. Anpalagan, and A. Benslimane, “A new block-based reinforcement learning approach for distributed resource allocation in clustered IoT networks,” IEEE Trans. Veh. Technol., vol. 69, no. 3, pp. 2891–2904, Mar. 2020.

I. Budhiraja, N. Kumar, and S. Tyagi, “Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication,” IEEE Internet Things J., vol. 8, no. 5, pp. 3143–3156, Mar. 2021.

J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Netw., vol. 61, pp. 85–117, Jan. 2015.

Chong Huang (Graduate Student Member, IEEE) received the B.Eng. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2011, and the M.Sc. degree in electrical and electronic engineering from Loughborough University, Loughborough, U.K., in 2015. He is currently pursuing the Ph.D. degree in signal processing with the University of Leicester, Leicester, U.K. His research interests include deep learning, reinforcement learning, physical layer security, cognitive radio, non-orthogonal multiple access (NOMA), intelligent reflecting surface (IRS), and the Internet of Things (IoT).

Gaojie Chen (Senior Member, IEEE) received the B.Eng. and B.Ec. degrees in electrical information engineering and international economics and trade from Northwest University, China, in 2006, and the M.Sc. (Hons.) and Ph.D. degrees in electrical and electronic engineering from Loughborough University, Loughborough, U.K., in 2008 and 2012, respectively. From 2008 to 2009, he was a Software Engineer with DT Mobile, Beijing, China. From 2012 to 2013, he was a Research Associate with the School of Electronic, Electrical and Systems Engineering, Loughborough University. From 2014 to 2015, he was a Research Fellow with 5GIC, Faculty of Engineering and Physical Sciences, University of Surrey, U.K. From 2015 to 2018, he was a Research Associate with the Department of Engineering Science, University of Oxford, U.K. He is currently a Lecturer with the School of Engineering, University of Leicester, U.K. His current research interests include information theory, wireless communications, cooperative communications, cognitive radio, the Internet of Things, secrecy communications, and random geometric networks. He received the Exemplary Reviewer Certificates of the IEEE WIRELESS COMMUNICATIONS LETTERS in 2018 and the IEEE TRANSACTIONS ON COMMUNICATIONS in 2019. He serves as an Associate Editor for the IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, Machine Learning in Communications and Networks, and Electronics Letters (IEL).

Yu Gong received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Electronics and Science Technology of China in 1992 and 1995, respectively, and the Ph.D. degree in communications from the National University of Singapore in 2002. Since 2012, he has been with the School of Electronic, Electrical and Systems Engineering, Loughborough University, U.K. After Ph.D. graduation, he took several research positions in the Institute for Infocomm Research, Singapore, and Queen’s University Belfast, U.K., respectively. From 2006 to 2012, he was a Lecturer with the School of Systems Engineering, University of Reading, U.K. His research interests include signal processing and communications, including wireless communications, cooperative networks, non-linear and non-stationary system identification, and adaptive filters.

Peng Xu (Member, IEEE) received the B.Eng. and Ph.D. degrees in electronic and information engineering from the University of Science and Technology of China, Hefei, China, in 2009 and 2014, respectively. From June 2014 to July 2016, he was working as a Post-Doctoral Researcher with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He is currently an Associate Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China. His current research interests include cooperative communications, information-theoretic secrecy, NOMA techniques, and reconfigurable intelligent surface. He received IEEE WIRELESS COMMUNICATIONS LETTERS Exemplary Reviewer 2015 and Excellent Paper of Chongqing Association for Science and Technology 2018.
Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 1999 and 2003, respectively.

From 2000 to 2002, he was a Research and Development Engineer with JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate with the University of Maryland. From 2006 to 2008, he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department as well as the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (Best Paper Award in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in IEEE conferences. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018, AAAS Fellow since 2019, and an ACM Distinguished Member since 2019. He has been 1% Highly Cited Researcher through 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: “for contributions to game theory and distributed management of autonomous communication networks.”

Jonathon A. Chambers (Fellow, IEEE) received the Ph.D. and D.Sc. degrees in signal processing from the Imperial College of Science, Technology and Medicine (Imperial College London), London, U.K., in 1990 and 2014, respectively. From 1991 to 1994, he was a Research Scientist with the Schlumberger Cambridge Research Center, Cambridge, U.K. He returned to Imperial College London as a Lecturer in signal processing in 1994 and was promoted to a Reader (an Associate Professor) in 1998. From 2001 to 2004, he was the Director of the Center for Digital Signal Processing and a Professor of signal processing with the Division of Engineering, King’s College London. From 2004 to 2007, he was a Cardiff Professorial Research Fellow with the School of Engineering, Cardiff University, Cardiff, U.K. From 2007 to 2014, he led the Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering, Loughborough University, where he is currently a Visiting Professor. He joined the School of Electrical and Electronic Engineering, Newcastle University in 2015, where he was a Professor of signal and information processing and led the ComS2IP Group and is also a Visiting Professor. He became the Head of the School of Engineering, University of Leicester, in 2017. He is also an International Honorary Dean and a Guest Professor with Harbin Engineering University, China, with support from the 1000 Talents Scheme. He is the coauthor of the books: Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability (New York, NY, USA: Wiley, 2001) and EEG Signal Processing (New York, NY, USA: Wiley, 2007). He has advised more than 90 researchers through to Ph.D. graduation and published more than 500 conference papers and journal articles, many of them are in the IEEE journals. His research interests include adaptive signal processing, and machine learning and their applications. He is also a Fellow of the Royal Academy of Engineering, U.K., and the Institution of Electrical Engineers. In 2007, he received the first QinetiQ Visiting Fellowship for his outstanding contributions to adaptive signal processing and his contributions to QinetiQ, as a result of his successful industrial collaboration with the international defense systems company QinetiQ. He was the Technical Program Chair of the 15th International Conference on Digital Signal Processing and the 2009 IEEE Workshop on Statistical Signal Processing, both held at Cardiff, U.K., and a Technical Program Co-Chair of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic. He has served on the IEEE Signal Processing Theory and Methods Technical Committee for a period of six years and the IEEE Signal Processing Society Awards Board for a period of three years, together with the Jack Kilby Award Committee. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING for two terms over the periods 1997–1999 and 2004–2007 and as a Senior Area Editor from 2011 to 2014.