A Deep Reinforcement Approach for Energy-Efficient Resource Assignment in Cooperative NOMA-Enhanced Cellular Networks

Yan-Yan Guo, Xiao-Long Tan, Yun Gao, Jing Yang, and Zhi-Chao Rui

***Abstract*—In this article, an energy efficiency (EE) maximization problem of cooperative nonorthogonal multiple access (CNOMA) network is proposed to jointly determine the user pairing, subchannel assignment, and power control scheme. We decompose it into two steps: in the first step, the optimal closed-form expressions of the power control problem are derived. Based on these, the EE optimization of whole system is formulated as a self-play Go game with the maximum EE as the winner, by constructing a virtual Go board with rows and columns representing indices of users and subchannels, respectively. Each move is to select a position on the Go board (i.e., select a user that has not been assigned to any channel and then assign a channel to it), until all users are assigned on the subchannels. Then, a deep Monte Carlo tree search (MCTS) model is proposed, where an MCTS guided by a neural network simulates multiple possible trajectories to search each move by evaluating its achievable EE reward, while the neural network is trained by the training data generated from the searching of MCTS to predict move selections and also the winner of games. The simulation results show that the proposed method is superior to a variety of conventional schemes in terms of EE in negligible computational time.**

***Index Terms*—Cooperative nonorthogonal multiple access (CNOMA), decode-and-forward (DF), deep neural network, deep reinforcement learning (DRL), energy efficiency (EE), full duplex (FD), half duplex (HD), Monte Carlo tree search (MCTS).**

I. Introduction

A. Motivation

NONORTHOGONAL multiple access (NOMA) has recently received widespread attention for its promising application in next-generation wireless networks [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. The key feature of NOMA is splitting multiple users into power domain to simultaneously serving them on same radio resources block (i.e., frequency and time) to improve spectral efficiency (SE). In the NOMA system, users with good channel conditions can extract their own information from other users' information by using successive interference canceler (SIC). However, this leads to an unavoidable limitation of NOMA that the high power needs to be assigned to a user with poor channel condition for successfully decoding the superimposed signal, which will reduce the SE of the system [4]. On the other hand, cooperative communication is a key technology for future mobile communications due to its ability to provide diversity gain through multiple-path user cooperative transmission, thereby reducing the impact of fading [12]. Thus, cooperative NOMA (CNOMA), which allows the users with good channel conditions to relay the removed information through the SIC to the users with weak channel conditions for enhancing their performance gains, has sparked a great deal of research interests [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23].

For a practical CNOMA scenario where there exists a direct link from the transmitter to each user, a user with good channel, referenced as "strong user," assists user with weak channel, referenced as "weak user," to communicate with the transmitter as a half-duplex (HD) or full-duplex (FD) relay. In the HD mode, direct and relay links occur over two consecutive time slots, whereas in the FD mode, they can take place simultaneously [5]. Due to the enhanced system throughput and ability to reduce the effect of fading, CNOMA has been an effective solution to improve the SE and coverage of wireless networks [1], [4]. Apart from the indicator of SE, energy efficiency (EE) is also one of the key performance requirements for 5G system [24]. However, on the one hand, in CNOMA systems, the strong user needs extra transmit power plus extra circuit power to relay the weak's signal [25]. On the other hand, HD CNOMA requires an additional transmission time slot for implementing cooperation and an FD CNOMA applies self-interference (SI) mitigation at the receiver of strong user. Those will result in complicated EE optimization formula to capture the characteristics of NOMA as well as the HD/FD decode-and-forward (DF) procedures [4]. Therefore, achieving a high EE in CNOMA systems remains challenging.

On the other hand, deep reinforcement learning (DRL) algorithms, such as deep Q-network (DQN), deep deterministic policy gradient (DDPG), etc., have attracted wide attention [26], [27]. For these DRL algorithms with a neural network trained by past experience transitions, the learned optimal policy is usually represented to be the maximal expectation of long-term cumulative reward. However, in the dynamic network environment, some actions become unavailable or obsolete by the reinforcement learning rules when suddenly adding or removing devices. Recently, AlphaGo Zero,

Manuscript received 16 January 2023; accepted 28 February 2023. Date of publication 7 March 2023; date of current version 7 July 2023. *(Corresponding author: Yan-Yan Guo.)*

The authors are with the College of Physics and Electronic Engineering, Shanxi University, Taiyuan 030004, China (e-mail: guoyanyan@sxu.edu.cn; 1350355014@qq.com; 812113623@qq.com; 1209183561@qq.com; 825179663@qq.com).

Digital Object Identifier 10.1109/JIOT.2023.3253129

based solely on a deep Monte Carlo tree search (MCTS) combined with a tree search and a neural network, without human data or guidance, has achieved superhuman performance in game of Go [28]. In addition to exploit the past experience like the existing DRL algorithms, the deep MCTS can simulate multiple possible trajectories in the future and need not determine the reward of action taken at each state until the game is over. Therefore, it is necessary to apply deep MCTS method to solve the single time slot decision-based resource allocation issues in CNOMA systems.

B. Literature Review

There have been extensive researches on CNOMA performance analysis [6], [7], [8], [9] and efficient performance optimization schemes [6], [10], [11], [12], [13], [14], [20]. As a major aspect of CNOMA, power allocation (PA) problems for the links from base station (BS) to strong user and weak user are of great importance and have been studied in [1], [6], [10], [11], and [12]. Kara and Kaya [6] discussed the PA coefficient of BS and threshold (TS) selection and proposed a joint PA-TS optimization to maximize sum rate under the constraint of required bit error rate (BER). Alharbi et al. [12] applied proportional rate constraints to the PA of the FD-CNOMA system to maximize the sum rate. However, these researches assumed that relaying power is a given positive constant. In fact, apart from the links from BS to user for signal transmission, the relay links between users play a vital role in improving the performance of CNOMA-enhanced networks. PA problems for two links of BS–user and user–user have been studied in [4], [5], [13], [14], and [20]. In [5], [13], and [14], maximizing the minimum achievable user rate in HD CNOMA and an FD CNOMA systems was formulated as a nonconcave power optimization problem for a NOMA user pair. In [4], the power control closed-form solutions to a given user pair for the HD and an FD cases were derived to maximize the sum rate.

The aforementioned researches on CNOMA systems focused error probability [2], [7], [8], achievable rate [4], [12], [13], [14], ergodic capacity, and outage probability [3], [6], [9], [16], [17]. Recently, the EE optimization strategies for CNOMA with simultaneous wireless information and power transfer (SWIPT)-aided transmission have been investigated [15], [16], [17], [18], [19], [21], [22], [23], [24]. For example, in [29], the near nodes harvest energy from radio frequency (RF) signals and use only their harvested energy stored in energy buffers to relay symbols to the far nodes. The optimization problem becomes parameter optimization regarding the PA coefficients and the power splitting (PS) factor. Although the ability of harvesting energy from RF signals can achieve low energy consumption, the practical implementation of SWIPT-aided CNOMA has been considered unfeasible due to performing information demodulation and energy harvesting at the same time, especially, for small and light devices, such as Internet of Things (IoT) and biomedical sensors. Motivated by this, Lamba et al. [1] employed a Stackelberg game (SG) for price-based PA based on charges per unit power in HD CNOMA networks. However, in this research, the strong user uses a constant power to relay message to the weak user. Wei et al. [25] transformed the EE maximization problem for an FD CNOMA system into a standard semidefinite programming (SDP) problem, which was implemented by MATLAB's CVX package. Besides, Ning et al. [30] formulated a stochastic-based EE optimization problem in HD CNOMA-based networks by considering the long-term queue stability. In this system, it was assumed that the weak user cannot directly communicate with the BS. As opposed to other goal optimization schemes, the EE research on the power optimization of CNOMA is still far from mature. One important issue is that the optimal power control closed-form expression for case of HD or an FD cannot directly inherit from conventional CNOMA.

To further improve throughput, EE, and the fairness of the CNOMA system, the user pairing and channel assignment schemes, which determine which link should serve each user, are critical [1], [4], [15], [20], [23], [31], [32], [33], [34], [35], [36]. Huu et al. [4], Dinh et al. [20], Liu et al. [31], Salehi et al. [32], Ma et al. [33], Dinh et al. [34], Lima et al. [35], and Obeed et al. [36] investigated the user pairing problem of multiple users in a subcarrier. For instance, in [33], it was converted to a critical ratio-based scheme and in [4], [20], [34], and [36], it was solved by using Hungarian method. Besides, the scenario with multiuser multisubcarrier CNOMA has been investigated in [1], [35], and [37]. In [1] and [35], the user pairing process depends on all available users' channel gains on each available subchannel [1]. The disadvantage of this method is that the order of available subchannels used for pairing users will influence the final performance of system. Thus, Cheng et al. [37] designed a two step pairing strategy that jointly considers all users' channel gains on all subchannel. In fact, the joint user pairing and channel assignment problem is clearly a combinatorial problem so that the traditional method to solve this problem is impractical, especially, for the network with large number of subchannels and users [1].

Recently, DRL algorithms have been extensively explored for the complicated scenarios of communication systems. Huang et al. [26] solved the relay selection and PA for cooperative hybrid NOMA/OMA networks by the asynchronous deep Q-Learning network (ADQN) and the asynchronous advantage actor–critic (A3C) network, respectively. In [38], a DRL-based model is used to decide the relay power from the near user (NU) in the CNOMA system. Moreover, the deep MCTS algorithm has been applied to the collaborative resource optimization in the mobile-edge computing (MEC) network [39] and the topology optimization in self-organized EE wireless sensor networks (WSNs) [40].

C. Contributions

In this article, the joint optimization problem of user pairing, subchannel assignment, and power control is investigated for a CNOMA-enhanced downlink cellular system, where users with NOMA capability can assist the transmissions for users with poor channel quality through DF cooperative communication in the HD or FD mode. We propose a new optimization problem, whose goal is to maximize the EE of the whole CNOMA system under the constraint of a certain QoS for all users. The main contributions of this article are given as follows.

1) The optimal power control closed-form expressions are derived for cases of the HD and FD CNOMA, respectively, which maximize the EE of a given pair under constraints of users' required QoS.
2) Inspired by AlphaGo, for the first time, the user pairing and subchannel assignment for the CNOMA system is formulated as a self-play game of Go by constructing a vitual Go board whose rows and columns denote indices of users and subchannels in the network, respectively. Then, a deep MCTS model combined with an MCTS module and a neural network is proposed to achieve a near optimal solution to the EE optimization problem for CNOMA systems.
3) The proposed resource assignment method can intelligently learn the network environment and make decisions in a dynamic environment to maximize the EE performance of the CNOMA systems while satisfying the QoS requirements of users. The action could be taken under the guidance of a specific optimization goals, such as sum rate of system, outage probability, and so on.
4) Numerical results also illustrate that our proposed method achieves better EE performance on CNOMA systems compared with existing schemes in negligible computational time.

The remainder of this article is organized as follows. Section II demonstrates the system model and problem formulation. The power control scheme for a given user pair is presented in Section III. The user pairing and subchannel assignment solution is shown in Section IV. Simulation results are given in Section V. Finally, Section VI gives the conclusion of this article.

II. System Model and Problem Formulation

A. System Model

We consider a downlink cellular system with one BS and a set of $2M$ users, which has K subchannels of equal bandwidth B_K. Each user is equipped with a transmit antenna and a receive antenna. Denoted $\mathbf{U}$ as the set of all users and $\mathbf{K}$ as the set of all subchannels, respectively. In each subchannel, a CNOMA pair communicates with the BS, in which a strong user relays a weak user's signal through DF in either the HD mode or the FD mode. Additionally, the channels between users on the same subchannel are assumed be to be asymmetric, e.g., $h_{i\to j}^k \neq h_{j\to i}^k$, where $h_{i\to j}^k$ and $h_{j\to i}^k$ denote the channel coefficients of links from user i to j and from user j to i on the kth subchannel, respectively. For analytical simplicity, we assume that the communications on the different subchannels are based on the orthogonal frequency-division multiplexing (OFDM) mode. The channel coefficients on each subchannel remain constant in each time slot and unaffected over each transmission phase, but may change independently between different time slots [41].

Consider a given pair of users, denoted as (m, n), where m and n, respectively, represent the strong user and the weak user within the pair. We define the channel coefficients of links from the BS to the strong user m and the weak user n on the kth subchannel, $|h_{B\to m}^k| > |h_{B\to n}^k|$. The communication between the users and the BS includes direct transmission (first) phase and cooperative transmission (second) phase. The BS with the fixed transmit power p_B transmits the superimposed signals $x^k = \sqrt{(1-\alpha_{m,n}^k)p_B}x_m + \sqrt{\alpha_{m,n}^k p_B}x_n$ on the kth subchannel, where x_m and x_n represent the messages sent by the BS to the strong user m and the weak user n, respectively, and $\alpha_{m,n}^k \in [0, 1]$ is the PA coefficient of user pair (m, n) on the kth subchannel.

1) HD CNOMA Mode: In the first phase, the received signals at the strong user m and the weak user n are expressed as $y_m^k = h_{B\to m}^k(\sqrt{(1-\alpha_{m,n}^k)p_B}\,x_m + \sqrt{\alpha_{m,n}^k p_B}\,x_n) + \omega_m^k$ and $y_n^k = h_{B\to n}^k(\sqrt{(1-\alpha_{m,n}^k)p_B}\,x_m + \sqrt{\alpha_{m,n}^k p_B}\,x_n) + \omega_n^k$, respectively, where ω_m^k and ω_n^k denote the additive white Gaussian noise (AWGN) at the strong user m and the weak user n on the kth subchannel, respectively. Since more power from the BS is allocated to the users with lower channel gain [1], the strong user m first decodes the message x_n of the weak user n from its received signal, and then decodes its own message x_m from its rest interference free signal. The achievable data rates of weak user n and the strong user m decoded by the strong user m on the kth subchannel are, respectively, expressed as

$$\left(R_n^k\right)^{H-1} = \frac{B_K}{2}\log_2\left(1+\frac{\left|h_{B\to m}^k\right|^2\alpha_{m,n}^k p_B}{\sigma^2+\left|h_{B\to m}^k\right|^2(1-\alpha_{m,n}^k)p_B}\right) \quad (1)$$

$$\left(R_m^k\right)^{H} = \frac{B_K}{2}\log_2\left(1+\frac{\left|h_{B\to m}^k\right|^2(1-\alpha_{m,n}^k)p_B}{\sigma^2}\right) \quad (2)$$

where σ^2 is the variance of AWGN. Since each phase in the HD mode utilizes the half time slot, the prelog factors in (1) and (2) are set to be (1/2) [10]. In the second phase, the strong user m assigns the power p_m^k to relay the decoded message x_n of the weak user n. The weak user n merges the two signals forwarded from the BS and the strong user m and then, decodes its own message, expressed as

$$\left(R_n^k\right)^{H-2} = \frac{B_K}{2}\log_2\left(1+\frac{\left|h_{m\to n}^k\right|^2 p_m^k}{\sigma^2} + \frac{\left|h_{B\to n}^k\right|^2\alpha_{m,n}^k p_B}{\sigma^2+\left|h_{B\to n}^k\right|^2(1-\alpha_{m,n}^k)p_B}\right) \quad (3)$$

where $h_{m\to n}^k$ is the channel coefficient of link from the strong user m to the weak user n on the kth subchannel. Therefore, according to the two-phase transmission, the achievable data rate of the weak user n on the kth subchannel is expressed as [10]

$$\left(R_n^k\right)^{H} = \min\left\{\left(R_n^k\right)^{H-1}, \left(R_n^k\right)^{H-2}\right\}. \quad (4)$$

2) FD CNOMA Mode: The two communication phases are simultaneously executed [10]. In the first phase, the strong user m and the weak user n receive the superposition signal of the BS. In the second phase, the strong user m relays the decoded message x_n to the weak user n, simultaneously. As a result, the strong user m will suffer the residual SI from its own input to output [10]. The received signals at the strong user m and the weak user n on the kth subchannel are expressed as $y_m^k = h_{B\to m}^k(\sqrt{(1-\alpha_{m,n}^k)p_B}\,x_m + \sqrt{\alpha_{m,n}^k p_B}\,x_n) + h_{m\to m}^k\sqrt{p_m^k}\,x_n + \omega_m^k$ and $y_n^k = h_{B\to n}^k(\sqrt{(1-\alpha_{m,n}^k)p_B}\,x_m + \sqrt{\alpha_{m,n}^k p_B}\,x_n) + h_{m\to n}^k\sqrt{p_m^k}\,x_n + \omega_n^k$, respectively, where $h_{m\to m}^k$ is SI channel coefficient of the strong user m on the kth subchannel. The achievable data rates of the weak user n and the strong user m decoded by the strong user m on the kth subchannel are, respectively, expressed as

$$\left(R_n^k\right)^{F-1} = B_K\log_2\left(1+\frac{\left|h_{B\to m}^k\right|^2\alpha_{m,n}^k p_B}{\sigma^2+\left|h_{B\to m}^k\right|^2(1-\alpha_{m,n}^k)p_B+\left|h_{m\to m}^k\right|^2 p_m^k}\right) \tag{5}$$

$$\left(R_m^k\right)^{F} = B_K\log_2\left(1+\frac{\left|h_{B\to m}^k\right|^2(1-\alpha_{m,n}^k)p_B}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k}\right). \tag{6}$$

By the same way as the HD CNOMA mode, the weak user n merges the two signals forwarded from BS and the strong user m [10] on the kth subchannel and, hence, its own decoded rate is achieved as

$$\left(R_n^k\right)^{F-2} = B_K\log_2\left(1+\frac{\left|h_{m\to n}^k\right|^2 p_m^k}{\sigma^2} + \frac{\left|h_{B\to n}^k\right|^2\alpha_{m,n}^k p_B}{\sigma^2+\left|h_{B\to n}^k\right|^2(1-\alpha_{m,n}^k)p_B}\right). \tag{7}$$

Finally, the achievable data rate of the weak user n on the kth subchannel is expressed as [10]

$$\left(R_n^k\right)^{F} = \min\left\{\left(R_n^k\right)^{F-1}, \left(R_n^k\right)^{F-2}\right\}. \tag{8}$$

B. Problem Formulation

The joint optimization of user pairing, subchannel assignment, and power control for the EE problem of the CNOMA system can be given by

$$\max_{\alpha,\partial,\mathbf{P}} \eta = \sum_{k\in\mathbf{K}}\sum_{m,n\in\mathbf{U}}\varphi_{m,n}^k\left(\frac{R_{m,n}^k}{p_m^k+p_B}\right) \tag{9}$$

$$\text{s.t. } 0\le\alpha_{m,n}^k\le 1 \quad \forall k\in\mathbf{K} \quad \forall m,n\in\mathbf{U} \tag{9a}$$

$$0\le p_m^k\le p^{\max} \quad \forall k\in\mathbf{K} \quad \forall m\in\mathbf{U} \tag{9b}$$

$$\varphi_{m,n}^k=\{0,1\} \quad \forall k\in\mathbf{K} \quad \forall m,n\in\mathbf{U} \tag{9c}$$

$$R_m^k\ge R\text{th}, R_n^k\ge R\text{th} \quad \forall k\in\mathbf{K} \quad \forall m,n\in\mathbf{U} \tag{9d}$$

$$\sum_{k\in\mathbf{K}}\varphi_{m,n}^k = 1 \quad \forall m,n\in\mathbf{U} \tag{9e}$$

$$\sum_{m,n\in\mathbf{U}}\varphi_{m,n}^k = 1 \quad \forall k\in\mathbf{K} \tag{9f}$$

where $R_{m,n}^k = (R_m^k)^\zeta + (R_n^k)^\zeta$ is the rate sum of the user pair (m,n) on the kth subchannel, $\zeta = \text{F}$ and $\zeta = \text{H}$ denote FD and HD modes, respectively, $p^{\max}$ is the maximum relaying power of user, Rth denotes the user's minimum rate requirement, $\alpha = \{\alpha_{m,n}^k\}$ denotes the PA vector of the BS for user pairs on all subchannels, $\varphi_{m,n}^k$ is a factor for channel assignment and user pairing, i.e., $\varphi_{m,n}^k = 1$ indicates that the strong user m and the weak user n form a CNOMA pair on the kth subchannel and $\varphi_{m,n}^k = 0$; otherwise, $\partial = \{\varphi_{m,n}^k\}$ indicates the vector of user pairing and channel assignment in the network, and $\mathbf{P} = \{p_m^k\}$ denotes the relaying power vector of strong users on all subchannels. Equation (9b) denotes the user's relaying power threshold. Equation (9d) denotes the QoS requirement for a user to successfully decode its intended message. Equations (9e) and (9f), respectively, indicate that one pair of users is allocated only on one subchannel and one subchannel is allocated only to one pair of users.

It is obviously a nonconvex mixed-integer nonlinear program (MINLP) problem [4]. We can observe from (9) that the EE expressions for given user pairs do not include the subchannel assignment and user pairing variable, ∂. Thus, the optimization problem (9) can be transformed into two steps. In the first step, the power control to maximize the EE of a given user pair on a subchannel is the optimization problem regarding the BS's PA coefficient and the strong user's relaying power within this pair. Then, based on the achieved optimal PA, maximizing EE of the whole system is converted into the optimization strategy with respect to the channel assignment and user pairing.

III. Power Control for CNOMA Pair

In the section, we tackle the power control problem to maximize EE for the case when a user pair is assigned on a given subchannel, expressed as

$$\max_{\alpha_{m,n}^k,\, p_m^k} \eta_{m,n}^k = \frac{R_{m,n}^k}{p_m^k+p_B}$$
$$\text{s.t.} \quad \text{(9a), (9b), (9d)} \tag{10}$$

where $\eta_{m,n}^k$ denotes EE of a given user pair (m,n) on the kth subchannel. Obviously, the problem (10) is not convex. According to (4) and (8), the lower decoded rate of weak user during the two phases determines its final achieved rate in cases of both HD and FD CNOMA. Therefore, if the PA strategy for $\alpha_{m,n}^k$ and p_m^k makes the achieved rates of weak user during the two phases unequal, i.e., $(R_n^k)^{\zeta-1} \ne (R_n^k)^{\zeta-2}$, the higher achieved rate during the two phases will waste the extra energy of system. Therefore, the approach for solving the problem (10) is feasible if and only if the weak user in a CNOMA pair during the two phases achieves equal transmission rate, i.e., $(R_n^k)^{\zeta-1} = (R_n^k)^{\zeta-2}$. Based on this premise, we will derive the optimal power control solutions of each CNOMA pair for HD and FD modes in the following sections.

A. HD CNOMA Mode

1) Condition A: When the weak user's decoded rate during the first phase is lower than during the second phase, according to (1) and (2), the EE optimization problem of the user pair (m, n) on the kth subchannel in (10) is rewritten by

$$\max_{p_m^k} \eta_{m,n}^k = \frac{\frac{B_K}{2}\log_2\left(1+\gamma_B\left|h_{B\to m}^k\right|^2\right)}{p_m^k+p_B} \quad \text{s.t.} \quad \text{(9b), (9d)} \tag{11}$$

where $\gamma_B = (p_B/\sigma^2)$. As illustrated in (11), it does not include $\alpha_{m,n}^k$. Obviously, to maximize $\eta_{m,n}^k$ in (11), the optimal relaying power of the strong user, $(p_m^k)^*$, should be set to be the minimum value in the feasible solutions, which is determined by $(R_n^k)^{H-2}$. It will be solved after Condition B.

2) Condition B: When the weak user's decoded rate during the first phase is higher than during the second phase, according to (2) and (3), the EE optimization problem of the user pair (m, n) on the kth subchannel in (10) is rewritten by

$$\begin{aligned}\max_{\alpha_{m,n}^k, p_m^k} \eta_{m,n}^k = \frac{B_k}{2(p_m^k+p_B)}&\left(\log_2\left(1+\frac{\left|h_{B\to m}^k\right|^2(1-\alpha_{m,n}^k)p_B}{\sigma^2}\right)\right.\\ &+\log_2\left(1+\frac{\left|h_{B\to n}^k\right|^2\alpha_{m,n}^k p_B}{\sigma^2}\right.\\ &\left.\left.+\left|h_{B\to n}^k\right|^2\left(1-\alpha_{m,n}^k\right)p_B+\frac{\left|h_{m\to n}^k\right|^2 p_m^k}{\sigma^2}\right)\right)\\ \text{s.t.} \quad &\text{(9a), (9b), (9d).}\end{aligned} \tag{12}$$

To solve the optimal solution to $\alpha_{m,n}^k$, the first-order derivatives of (12) with respect to $\alpha_{m,n}^k$ is given by

$$\begin{aligned}\frac{\partial\eta_{m,n}^k}{\partial\alpha_{m,n}^k} = \frac{B_K}{\ln(2)\left(p_m^k+p_B\right)}&\left(G\frac{\gamma_B}{\frac{1}{\left|h_{B\to n}^k\right|^2}+\gamma_B\left(1-\alpha_{m,n}^k\right)}\right.\\ &\left.-\frac{\gamma_B}{\frac{1}{\left|h_{B\to m}^k\right|^2}+\gamma_B\left(1-\alpha_{m,n}^k\right)}\right)\end{aligned} \tag{13}$$

where

$$G = 1-\frac{\frac{\left|h_{m\to n}^k\right|^2 P_m^k}{\sigma^2}}{1+\frac{\left|h_{B\to n}^k\right|^2\alpha_{m,n}^k p_B}{\sigma^2+\left|h_{B\to n}^k\right|^2(1-\alpha_{m,n}^k)p_B}+\frac{\left|h_{m\to n}^k\right|^2 p_m^k}{\sigma^2}}. \tag{14}$$

As illustrated in (13), because of $G < 1$ and $|h_{B\to n}^k|^2 < |h_{B\to m}^k|^2$, $(\partial\eta_{m,n}^k/\partial\alpha_{m,n}^k) < 0$ and, thus, $\eta_{m,n}^k$ in (12) is a monotonically decreasing function with respect to $\alpha_{m,n}^k$. Consequently, the optimal value of $\alpha_{m,n}^k$ should be set to be minimum in the feasible solutions, which is determined by $(R_n^k)^{H-1}$. We set $(R_n^k)^{H-1} = R\text{th}$ and according to (1), the optimal value of $\alpha_{m,n}^k$ is obtained as

$$\left(\alpha_{m,n}^k\right)^* = \frac{C}{1+C}\left(1+\frac{1}{\gamma_B\left|h_{B\to m}^k\right|^2}\right) \tag{15}$$

where $C = 2^{2R\text{th}/B_K} - 1$. Similarly, the achieved rate of strong user m should satisfy $(R_m^k)^H \geq R\text{th}$. According to (2), we have $(\alpha_{m,n}^k)^* \leq 1 - (C/[\gamma_B|h_{B\to m}^k|^2])$. If $(\alpha_{m,n}^k)^* > 1 - (C/[\gamma_B|h_{B\to m}^k|^2])$, the strong user m and the weak user n cannot form an HD CNOMA pair to communicate with the BS on the kth subchannel.

According to Condition A, the optimal relaying power $(p_m^k)^*$ of the strong user m, which is determined by $(R_n^k)^{H-2}$, is the minimum value in the feasible solutions. Consequently, by setting $(R_n^k)^{H-2} = R\text{th}$ and putting $(\alpha_{m,n}^k)^*$ of (15) into (3), $(p_m^k)^*$ is given by

$$\begin{aligned}\left(p_m^k\right)^* = &\frac{\sigma^2 C(1+C)}{\left|h_{m\to n}^k\right|^2}\\ &\times\frac{\left(\left|h_{B\to m}^k\right|^2-\left|h_{B\to n}^k\right|^2\right)}{\gamma_B\left|h_{B\to m}^k\right|^2\left|h_{B\to n}^k\right|^2+(C+1)\left|h_{B\to m}^k\right|^2-C\left|h_{B\to n}^k\right|^2}.\end{aligned} \tag{16}$$

If $(p_m^k)^* > p^{\max}$, the strong user m and the weak user n cannot form an HD CNOMA pair to communicate with the BS on the kth subchannel.

B. FD CNOMA Mode

1) Condition A: When the weak user's decoded rate during the first phase is lower than during the second phase, according to (5) and (6), we obtain the EE optimization problem of the user pair (m, n) on the kth subchannel in (10), written by

$$\begin{aligned}\max_{p_m^k} \eta_{m,n}^k = \frac{B_K}{p_m^k+p_B}&\left(\log_2\left(\sigma^2+\left|h_{B\to m}^k\right|^2 p_B+\left|h_{m\to m}^k\right|^2 p_m^k\right)\right.\\ &\left.-\log_2\left(\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k\right)\right)\\ \text{s.t.} \quad &\text{(9b), (9d).}\end{aligned} \tag{17}$$

As illustrated in (17), it does not include $\alpha_{m,n}^k$. By differentiating the objective function of (17) with respect to p_m^k, we have

$$\begin{aligned}\frac{d\eta_{m,n}^k}{dp_m^k} = -\frac{B_K}{\left(p_m^k+p_B\right)^2}&\left(\frac{\left|h_{m\to m}^k\right|^2\left(p_B+p_m^k\right)}{\ln(2)}\right.\\ &\times\left(\frac{1}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k}-\frac{1}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k+\left|h_{B\to m}^k\right|^2 p_B}\right)\\ &\left.+\log_2\left(\frac{\sigma^2+\left|h_{B\to m}^k\right|^2 p_B+\left|h_{m\to m}^k\right|^2 p_m^k}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k}\right)\right)\end{aligned} \tag{18}$$

where

$$\frac{1}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k}-\frac{1}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k+\left|h_{B\to m}^k\right|^2 p_B} > 0$$

$$\text{and} \quad \log_2\left(\frac{\sigma^2+\left|h_{B\to m}^k\right|^2 p_B+\left|h_{m\to m}^k\right|^2 p_m^k}{\sigma^2+\left|h_{m\to m}^k\right|^2 p_m^k}\right) > 0.$$

Obviously, $(d\eta_{m,n}^k/dp_m^k) < 0$ and $\eta_{m,n}^k$ in (17) is a monotonically decreasing function with respect to p_m^k. Therefore, the optimal relaying power $(p_m^k)^*$ of the strong user m should be set to be minimum value in the feasible solutions. We will solve it after Condition B.

2) Condition B: When the weak user's decoded rate during the first phase is higher than during the second phase, according to (6) and (7), the EE optimization problem of the user pair (m, n) on the kth subchannel in (10) is rewritten by

$$\begin{aligned}\max_{\alpha_{m,n}^k, p_m^k} \ & \eta_{m,n}^k \\ & = \frac{B_K}{p_m^k + p_B}\left(\log_2\left(1 + \frac{|h_{B\to m}^k|^2(1-\alpha_{m,n}^k)p_B}{\sigma^2 + |h_{m\to m}^k|^2 p_m^k}\right)\right. \\ & \quad \left. + \log_2\left(1 + \frac{|h_{B\to n}^k|^2 \alpha_{m,n}^k p_B}{\sigma^2 + |h_{B\to n}^k|^2(1-\alpha_{m,n}^k)p_B} + \frac{|h_{m\to n}^k|^2 p_m^k}{\sigma^2}\right)\right) \\ \text{s.t.} \quad & (9a), (9b), (9d). \end{aligned} \tag{19}$$

To solve the optimal value of $\alpha_{m,n}^k$, the first-order derivatives of (19) with respect to $\alpha_{m,n}^k$ are given by

$$\begin{aligned}\frac{\partial \eta_{m,n}^k}{\partial \alpha_{m,n}^k} = \frac{B_K}{\ln(2)(p_m^k + p_B)}&\left(G\frac{\gamma_B}{\frac{1}{|h_{B\to n}^k|^2} + \gamma_B(1-\alpha_{m,n}^k)}\right. \\ & \left. - \frac{\gamma_B}{\frac{|h_{m\to m}^k|^2 p_m^k + 1}{|h_{B\to m}^k|^2} + \gamma_B(1-\alpha_{m,n}^k)}\right). \end{aligned} \tag{20}$$

In (20), obviously, $|h_{m\to m}^k|^2 \cdot p_m^k \ll 1$ and $|h_{B\to n}^k|^2 < |h_{B\to m}^k|^2$. Consequently, by omitting $|h_{m\to m}^k|^2 \cdot p_m^k$, we can rewrite (20) as

$$\begin{aligned}\frac{\partial \eta_{m,n}^k}{\partial \alpha_{m,n}^k} \approx \frac{B_K}{\ln(2)(p_m^k + p_B)}&\left(G\frac{\gamma_B}{\frac{1}{|h_{B\to n}^k|^2} + \gamma_B(1-\alpha_{m,n}^k)}\right. \\ & \left. - \frac{\gamma_B}{\frac{1}{|h_{B\to m}^k|^2} + \gamma_B(1-\alpha_{m,n}^k)}\right). \end{aligned} \tag{21}$$

It is exactly the same as (13) and obviously, $(\partial\eta_{m,n}^k/\partial\alpha_{m,n}^k) < 0$, so that $\eta_{m,n}^k$ in (19) is a monotonically decreasing function with respect to $\alpha_{m,n}^k$. Therefore, the optimal solution of $\alpha_{m,n}^k$ is the minimum value in the feasible solutions.

Considering Conditions A and B, both of $\alpha_{m,n}^k$ and p_m^k should be set to be minimum in the feasible solutions, which are determined jointly by the two-phase transmission. Therefore, we set $(R_n^k)^{F-1} = (R_n^k)^{F-2} = R\text{th}$. According to (5) and (7), we have the following equations:

$$\begin{cases} \frac{|h_{B\to m}^k|^2 \alpha_{m,n}^k p_B}{\sigma^2 + |h_{B\to m}^k|^2(1-\alpha_{m,n}^k)p_B + |h_{m\to m}^k|^2 p_m^k} = c_1 \\ \frac{|h_{B\to n}^k|^2 \alpha_{m,n}^k p_B}{\sigma^2 + |h_{B\to n}^k|^2(1-\alpha_{m,n}^k)p_B} + \frac{|h_{m\to n}^k|^2 p_m^k}{\sigma^2} = c_1 \end{cases} \tag{22}$$

where $c_1 = 2^{R\text{th}/B_K} - 1$. By solving (22), we omit the value of $\alpha_{m,n}^k > 1$ and obtain the optimal value of $\alpha_{m,n}^k$ and p_m^k written as (23) and (24) on bottom of the next page, where $a_1 = |h_{B\to m}|^2 p_B$, $a_2 = |h_{B\to n}|^2 p_B$, $b_1 = \sigma^2 + |h_{B\to m}|^2 p_B$, $b_2 = \sigma^2 + |h_{B\to n}|^2 p_B$, $d_1 = |h_{m\to m}|^2$, and $d_2 = (|h_{m\to n}|^2/\sigma^2)$.

Note that $(p_m^k)^* \le p^{\max}$ and otherwise, the strong user m and the weak user n cannot form an FD CNOMA pair to communicate with the BS on the kth subchannel.

IV. Joint Resource Assignment of EE CNOMA System

After deriving the optimal EE of given CNOMA pair, $(\eta_{m,n}^k)^*$, the EE optimization problem of the CNOMA system in (9) is transformed into the user pairing and channel assignment ∂, written as

$$\max_{\partial} \sum_{k\in\mathbf{K}} \sum_{m,n\in\mathbf{U}} \varphi_{m,n}^k \left(\eta_{m,n}^k\right)^* \text{s.t.} \quad (9a), \ (9e), (9f). \tag{25}$$

It is clearly a combinatorial problem which requires an exhaustive search [1]. To solve this obstacle, we construct a $2M \times K$-order user–subchannel matching matrix $[x_{i,k}]_{i\in\mathbf{U},k\in\mathbf{K}}$, where row i and column k correspond to user i and subchannel k in the network, respectively. $x_{i,k} = 1$ represents that user i is assigned on the kth subchannel and $x_{i,k} = 0$, otherwise. Therefore, if user pair (m, n) is assigned on the kth subchannel, then $\varphi_{m,n}^k = x_{m,k} \cdot x_{n,k} = 1$, and $\varphi_{m,n}^k = x_{m,k} \cdot x_{n,k} = 0$, otherwise.

A. Formation of Channel Assignment and User Pairing as Self-Play Game of Go

We treat the user–subchannel matching matrix $[x_{i,k}]_{i\in\mathbf{U},k\in\mathbf{K}}$ as a Go board of size $2M \times K$. The optimization problem for (25) is formulated as a self-play game of Go as shown in Fig. 1.

1) State s is the user–subchannel matching matrix $[x_{i,k}]_{i\in\mathbf{U},k\in\mathbf{K}}$.
2) Action a taken at each step is $x_{i,k} = 1 \ \forall i \in \mathbf{U}, \ k \in \mathbf{K}$, representing that user i is selected and then, assigned on the kth subchannel. First, we start the game at a root state, i.e., set every element of the matching matrix to be zero, $s_0 = [0]_{2M\times K}$. Then, each state $s_t \in \mathbb{S}$, is transferred to the next state s_{t+1} by taking an action $a_t \in \mathbb{A}$ according to a policy $\pi(s_t, a_t)$, with a probability matrix $P(s_{t+1}|s_t)$ and a reward r_t.
3) The reward is not determined until the terminal state s_{2M} is reached (i.e., all users are assigned on the subchannels), and then, a final reward according to optimization goal in (25) defined as

$$r_{2M} = \begin{cases} \sum\limits_{k\in\mathbf{K}} \sum\limits_{m,n\in\mathbf{U}} \left(x_{m,k} \cdot x_{n,k}\right)\left(\eta_{m,n}^k\right)^*, & \text{if } \forall R_m^k \ge R\text{th} \\ & \text{and } \forall R_n^k \ge R\text{th} \\ -1, & \text{otherwise} \end{cases} \tag{26}$$

is propagated back along the state trajectory, as the reward for every action on the path.

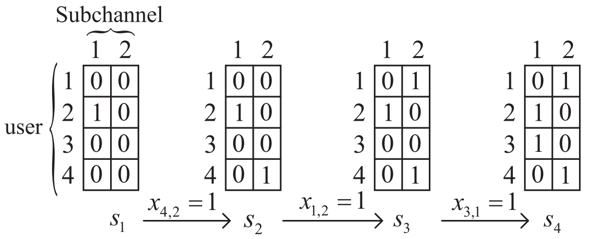
We can describe this procedure by a fully observable finite-horizon Markov decision process (MDP) of a 4-tuple $(\mathbb{S}, \mathbb{A}, r, P)$ [42].

To implement the MDP, we propose a deep MCTS-based model as shown in Fig. 2, consisting of two parts: 1) a neural network and 2) an MCTS module. The neural network $\Pr = f_\theta(s, \hat{H})$ with parameters θ takes the current state s as well as the channel gains $\hat{H}$ of the corresponding slot as its

Algorithm 1 Implementation Procedure of the MCTS

```
Input: root node $v_0$, slot_id, ResNet $f_\theta$
Output: A trajectory of states and policies $E$: $s_0$, $\pi_0$, $s_1$, $\pi_1$, ..., $s_{2M-1}$, $\pi_{2M-1}$, $s_{2M}$
Function MCTS($v_0$, slot_id)
  while within $N_t$:
    $v_{\text{end}}, r$ = TreeSearch($v_0$, slot_id)
    Save $v_{\text{end}}$ with the best reward so far
    BackUp($v_{\text{end}}$, $r$)    //backup
  end while
  Propagated along the search path with the best reward back to the root node, and get $E$
  return $E$
Function TreeSearch($v$, slot_id)
  if $v$ is $v_L$    //has not been expanded
    Read $s_L$ from $v$ and the normalized channel gains $\hat{H}$ of slot_id
    $Pr$ of node $v \leftarrow$ ResNet($s_L, \hat{H}$)
    Initialize all child nodes of node $v$ and added them to tree
    TreeSearch($v$, slot_id) // Continue search until end node
  else if $v$ is end node
    Read $s$ from $v$ and channel gains of slot_id
    Get reward $r$ according to formula (26)
    return $r$ and $v$
  else    //select a best child
    Get best child $v'$ by formula (28)
    TreeSearch($v'$, slot_id) // Continue search until end node
  end if
Function BackUp($v$, $r$)
  while $v \neq$ the root node
    $N(v) \leftarrow N(v) + 1$
    $Q(v) \leftarrow \frac{Q(v)N(v)+r}{N(v)+1}$
    $v \leftarrow$ the parent node of $v$
  end while
```

input and outputs a vector, $\text{Pr} = \{\Pr(s, a), a \in \mathbb{A}(s)\}$, representing a probability distribution of actions, where $\Pr(s, a)$ is a prior probability taking action a from state s. The MCTS module contains the value function α_θ based on the latest neural network f_θ for actions search to generate self-play data by running simulations.



$$
\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 0 & 0 \\ 4 & 0 & 0 \end{array}
\quad
\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 0 & 0 \\ 4 & 0 & 1 \end{array}
\quad
\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 0 & 1 \\ 2 & 1 & 0 \\ 3 & 0 & 0 \\ 4 & 0 & 1 \end{array}
\quad
\begin{array}{c|cc} & 1 & 2 \\ \hline 1 & 0 & 1 \\ 2 & 1 & 0 \\ 3 & 1 & 0 \\ 4 & 0 & 1 \end{array}
$$

$$
s_1 \xrightarrow{x_{4,2}=1} s_2 \xrightarrow{x_{1,2}=1} s_3 \xrightarrow{x_{3,1}=1} s_4
$$

Fig. 1. MDP of constructing a 4 × 2 user–subchannel matching matrix starting at state 1 and completed after three steps.

B. Collecting Training Data Sets by Using MCTS

Fig. 2(a) illustrates a search tree of MCTS. Each node represents the observed state s of the environment. Each edge (s, a) in the search tree denotes an action, $a \in \mathbb{A}(s)$, taken from state s. Each edge (s, a) or node represents a 5-tuple data $\big(s, a, N(s, a), \Pr(s, a), Q(s, a)\big)$, where $N(s, a)$ is a visit number of node and $Q(s, a)$ is a state–action value, representing the expected reward taking action a from state s, defined as [42]

$$
Q(s, a) \triangleq \mathbb{E}\left[\sum_{\tau=t}^{2M} r_\tau | s_t = s, a_t = a\right]. \tag{27}
$$

A self-play game (corresponding to performing resource allocation for a time slot) uses N_t simulations of MCTS for selecting each node in the search tree. As shown in Fig. 2(a), each simulation of MCTS repeatedly performs three sequential steps [40]: 1) selection; 2) expansion and evaluation; and 3) backup. The procedure of the MCTS subroutine is illustrated in Algorithm 1.

Selection: Each simulation guided by the neural network f_θ starts at the root node with state s_0 and iteratively chooses the best child node s' from the node s that have been fully expanded according to maximizing the upper confidence bound (UCB) [28], i.e.,

$$
s' = \arg\max_{a \in \mathbb{A}(s)} (Q(s, a) + U(s, a)). \tag{28}
$$

In (28), $U(s, a) = c\,\mathbb{P}(s, a)(\sqrt{\sum_{a \in \mathbb{A}(s)} N(s, a)}/[1 + N(s, a)])$ where c is a controlling hyperparameter for the exploration level. To ensure that all nodes of the searching tree may be tried, the Dirichlet noise is added to the prior probability $\Pr(s, a)$ to achieve $\mathbb{P}(s, a) = (1 - \varepsilon)\Pr(s, a) + \varepsilon\mu$, where $\mu \sim \text{Dir}(0.03)$ and $\varepsilon = 0.25$ [28].

$$
\begin{aligned}
\left(a_{m,n}\right)^* = {} & \frac{1}{2a_1a_2d_2(1+c_1)}\big(a_1b_2d_2(1+c_1) + a_2c_1(d_1 + c_1d_1 + b_1d_2) \\
& - \sqrt{(a_1b_2d_2(1+c_1) + a_2c_1(d_1 + c_1d_1 + b_1d_2))^2 - 4a_1a_2b_2c_1d_2(1+c_1)(c_1d_1 + b_1d_2)}\Big)
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
\left(p_m^k\right)^* = {} & \frac{1}{2a_2c_1d_1d_2}\big(a_1b_2d_2(1+c_1) + a_2c_1(d_1 + c_1d_1 - b_1d_2) \\
& - \sqrt{(a_1b_2d_2(1+c_1) + a_2c_1(d_1 + c_1d_1 + b_1d_2))^2 - 4a_1a_2b_2c_1d_2(1+c_1)(c_1d_1 + b_1d_2)}\Big)
\end{aligned} \tag{24}
$$

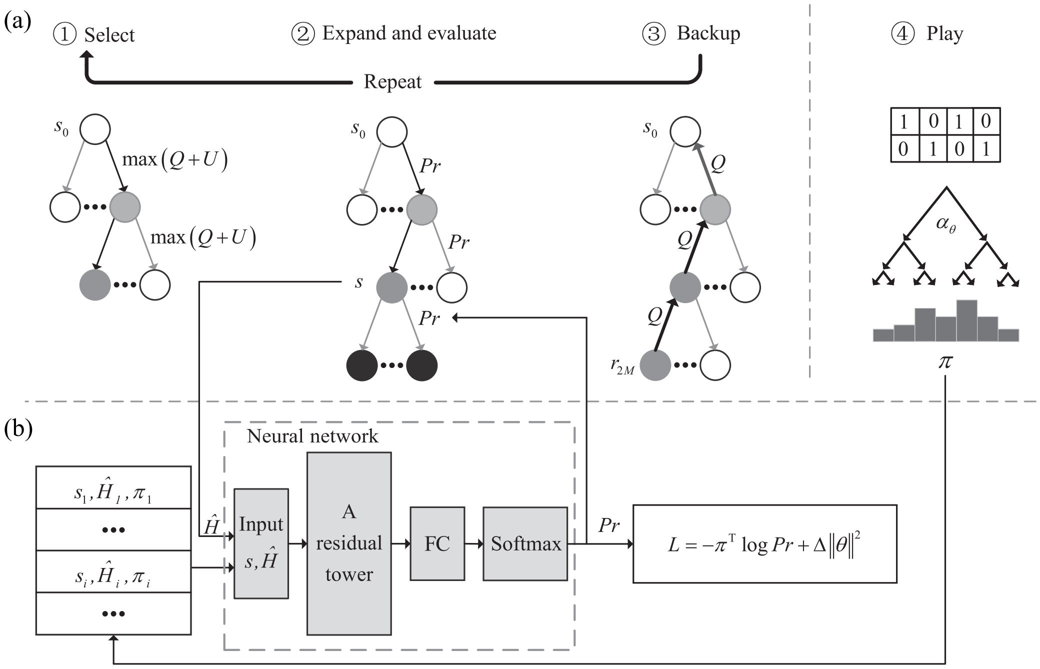


Fig. 2. Proposed deep MCTS-based model. (a) Self-play. (b) Neural network training.

Expansion and Evaluation: Once the selection procedure in the searching tree reaches a leaf node with state s_L, we will expand and evaluate this node according to prior probabilities, $\Pr = f_\theta(s_L, \hat{H})$, generated only once by the neural network. And its each child node is initialized by $(s, a, N(s, a) = 0, \Pr(s, a), Q(s, a) = 0)$. These child nodes are added to the searching tree (see lines 13–15 in Algorithm 1).

Backup: The two above steps will be iterative until the termination state or $t = 2M$ is reached. Then, the final reward in (26) is propagated along the trajectory back to the root state. Each node (s, a) traversed is performed by increasing its visiting number $N(s, a)$, and updating its state–action value $Q(s, a)$ (see lines 27–29 in Algorithm 1).

C. Approximating Policy Functions Using Neural Network

A deep residual network (ResNet) architecture is adopted in order to improve the representational capacity of the neural network while maintaining its training feasibility.

Input Layer: The neural network takes the current matching matrix s as well as the channel gains $\hat{H}$ as inputs. $\hat{H}$ is a 1-D vector with the length of $(4M^2 + 2M) \times K$, which is achieved by reshaping all links' channel gains from the BS to users and between users in the corresponding time slot. Its calculation formula can refer to (4) in [43]. The current matching matrix s is a $2M \times K$ image consisting of binary values.

Hidden Layers: The input features s and $\hat{H}$ are fed into a convolutional block followed by W_1 residual blocks. The convolutional block consists of a convolution of 32 filters of kernel size 3, no padding, batch normalization, and a rectified linear unit (ReLU) activation. Each residual block adopts the following modules [28]: 1) a convolution of 16 filters of kernel size 3, no padding; 2) batch normalization; 3) a ReLU activation; 4) a convolution of 16 filters of kernel size 3, two paddings; 5) batch normalization; 6) a skip connection that adds the input to the block; and 7) a ReLU activation.

Output Layer: The outputs of the hidden layers are processed by a fully connected (FC) layer with 256 neurons followed by a softmax activation for the probabilities distribution, Pr.

Loss Function: The neural network's parameters are updated in a self-supervised learning manner to make the action probabilities, $\Pr = f_\theta(s, \hat{H})$, more closely approximate the enhanced search policies π from the MCTS by Adam gradient descent [44] with a loss function

$$L = -\pi^{\mathrm{T}} \log \Pr + \Delta\|\theta\|^2 \tag{29}$$

where Δ is a weight regularization to prevent overfitting [28].

Algorithm 2 illustrates the self-training process of the proposed model. We set the minimum size of training data set to be N_b. Once the size of training data set in memory $\mathbf{D}$ is more than N_b, the neural network is continually optimized every N_c time slots by randomly extracting N_g data samples from $\mathbf{D}$ (see lines 6–10). Based on new neural network parameters θ, the MCTS will start subsequent games of self-play and return a trajectory of states and policies E, which will be stored as $\{(s_i, \hat{H}, \pi_i)\}, i = 0, \ldots, 2M - 1$ for one time slot, sampled uniformly among all time steps in memory (see lines 11–13).

D. Robustness and Convergence Analysis of the Neural Network Framework

To distinguish between the contributions of network architecture and algorithm, three neural networks were created as benchmarks, using a convolutional network (CNN) architecture, an FC network (DNN) architecture, and a ResNet architecture in which we use the FC blocks with 256 neurons to replace the convolutional blocks of the proposed ResNet.

Algorithm 2 Self-Training Process of the Proposed Model

```
Input: Number of slots $N_{slot}$
Output: ResNet $f_\theta$
Training dataset $\mathbf{D}=\emptyset$; slot_id=0
while slot_id within $N_{slot}$:
    Read $\hat{H}$ of slot_id
    if size of $\mathbf{D} \geq N_b$ and slot_id mod$N_c = 0$
        Randomly extract $N_g$ data samples from $\mathbf{D}$
        Train ResNet
        $\theta \leftarrow$ formula (29)
    end if
    Create a root node $v_0$
    $E \leftarrow$ MCTS($v_0$, slot_id)
    $\mathbf{D} \cup \left\{(s_i, \hat{H}, \pi_i)\right\}$    // a slot includes $2M$ data samples
    slot_id++
end while
```

TABLE I
PARAMETER SETTINGS OF DEEP MCTS ALGORITHM

Parameter	Value	
Minimum size of training dataset (N_b)	100*2M	
Simulations of MCTS (N_t)	M=5	500
	M=4	300
	M=3	100
	M=2	50
Minibatch size of data (N_g)	128	
Number of interval slots for training (N_c)	100	
Learning rate (β)	0.0001	
Weight regularization (Δ)	0.0001	
c	2	

The CNN and DNN configured with W_2 hidden layers have the same input and output layers as the proposed ResNet. Each hidden layer of the CNN has a convolution of 16 kernels of size 3, no padding and a ReLU activation. Each FC hidden layer of the DNN contains 256 neurons followed by a ReLU activation. Moreover, we will verify whether the change of the number of the hidden layers (or residual blocks in the ResNet architecture) would impact on the performance of the different schemes, respectively. The parameters of the deep MCTS algorithm are illustrated in Table I. In our experiment, we introduce 10^4 independent channel generations. To evaluate the learning error of neural networks, we define the test accuracy as the ratio of the number of correct predictions to total number of predictions [45].

Fig. 3 shows the convergence comparison of the proposed ResNet architecture with $W_1 = 2$, the CNN and DNN architectures with $W_2 = 2$ during self-play reinforcement learning for $M = 3$. From Fig. 3, the probability loss of proposed ResNet architecture very quickly converges to almost zero and enters a stable status once the count of training iterations reaches about 6000, while both of CNN and DNN architectures converge to about 1.5 and then fluctuate in the range of about 0.5. The performance of CNN is slightly better than that of DNN.

Moreover, after 10^4 training iterations of all neural networks, we test 6000 data samples, and the test accuracy is demonstrated in Table II. It is observed that using a residual network achieves lower error and better performance. Also, we can see that a small number of the residual blocks does not make the test accuracy being much degraded, which implies that the ResNet architecture is robust and efficient. Especially, it is witnessed that using the convolutional blocks can improve the test accuracy of the ResNet architecture.

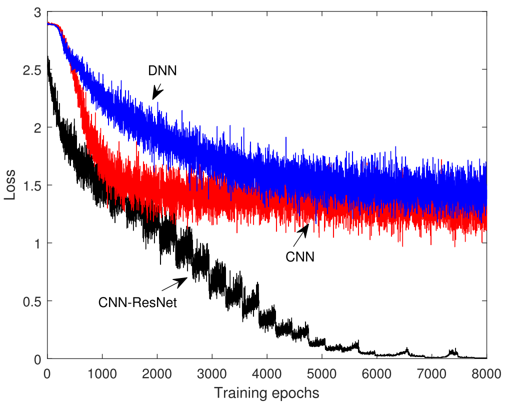


Fig. 3. Convergence of the proposed neural network.

TABLE II
TEST ACCURACY FOR DIFFERENT NEURAL NETWORKS

Number of Hidden Layers (Residual Blocks)	Method	Test Accuracy(%)
8	CNN-ResNet	99.9
4	CNN-ResNet	99.7
2	CNN-ResNet	99.6
8	DNN-ResNet	99.5
4	DNN-ResNet	99.5
2	DNN-ResNet	99.4
3	DNN	45
2	DNN	37.5
3	CNN	48.5
2	CNN	38.4

E. Complexity Analysis and Application of the Deep MCTS Approach for Practical Network

Over the course of self-play reinforcement learning, to balance performance and complexity, we adjust the simulations of MCTS according to the size of the network as shown in Table I. For example, completing one simulation requires six steps for $M = 3$, and, hence, each game needs 600 steps when simulations of MCTS are set to 100. 10^4 games (10^4 independent channel generations) require 6×10^6 steps, corresponding to approximately 0.248 s per step (run in the computer configured with 4 CPU and 8192-MB RAM).

This self-play reinforcement learning will converge to a solution with a sufficient number of training iterations of the neural network (about 6000 for $M = 3$), which corresponds to approximately 8.23 s per iteration. Once the neural network $\Pr = f_\theta(s, \hat{H})$ is trained, to obtain an optimal policy, we start with the root state s_0 and then, at each state s_t, select an action $a_t \sim \pi_t$ predicted by the neural network and sequentially update the state $s_{t+1} = \mathscr{T}(s_t, a_t)$ until all users are paired on each subchannel. Due to the offline training, the training complexity of the deep MCTS model can be neglected and, thus, the optimal policy can be obtained within milliseconds.

V. Numerical Results

In this section, we will validate the performance of the proposed HD and FD CNOMA schemes. We assume that $2M$ users are randomly distributed in a cell with a radius of 250 m and the BS is located at the center of this cell. For simplicity, we set $K = M$ in the simulation. The channel gains of all links follow independent Rayleigh fading with $\mathcal{CN}(0, 1)$ distribution. The path-loss exponent is $\chi = 2$. $B_K = 2$ Hz, $\sigma^2 = -80$ dBm, and $p^{\max} = 27$ dBm and all strong users are assumed to have the same SI channel coefficient, $|h_{m\to m}|^2 = \lambda_{\mathrm{SI}}\ \forall m \in \mathbf{U}$. Furthermore, the parameters of the proposed deep MCTS model are set as shown in Table I and $W_1 = 8$. In the simulation, 1000 independent slot channel gains are generated and as long as one user in a pair cannot satisfy its minimum rate requirement, Rth, the EE of this pair is set to be 0.

A. Performance Comparison of Different Power Control Schemes

Throughout this section, we present the analytical and numerical average EE of the proposed power control schemes for one CNOMA pair. We compared them with the two existing CNOMA power control schemes. One is the price-based PA for case of HD CNOMA in [1], termed as "HD CNOMA with full relaying power," which assigns the full relaying power, $p_m^k = p^{\max}$, to the strong users and fixes the total transmit power budget of the BS on one subchannel, so as to optimize the PA coefficient of BS–user link. The other is the CVX-based scheme for case of an FD CNOMA in [25], termed as "FD CNOMA CVX unfixed," in which the total power budget of the BS on one subchannel is not fixed to optimize the PA of both BS–user link and user–user link. Besides, based on the CVX algorithm proposed in [25], we fixed the total power budget of the BS on each subchannel and presented a term "FD CNOMA CVX fixed" scheme for case of an FD CNOMA. Moreover, similar to the two FD CNOMA schemes based on CVX, two CVX-based schemes for case of HD CNOMA, termed as "HD CNOMA CVX fixed" and "HD CNOMA CVX unfixed," were also illustrated as benchmarks, respectively. Note that for all schemes of fixing the total power budget of the BS on each subchannel, we assign the full transmit power of the BS, p_B, whereas, for all schemes that do not fix the total transmit power budget of the BS, the total power of the BS is set to be lower than or equal to p_B.

We set Rth $= 1.5$ bit/s and $p_B = 20$ dBm. Fig. 4 shows the average EE achieved by the power control schemes of FD CNOMA versus λ_{SI}. It is shown from Fig. 4 that the average EE decreases when the SI channel coefficient increases. This observation indicates, when λ_{SI} increases, decreasing the relaying power is the only way to relieve the effect of SI, which is the same as the conclusion of [4].

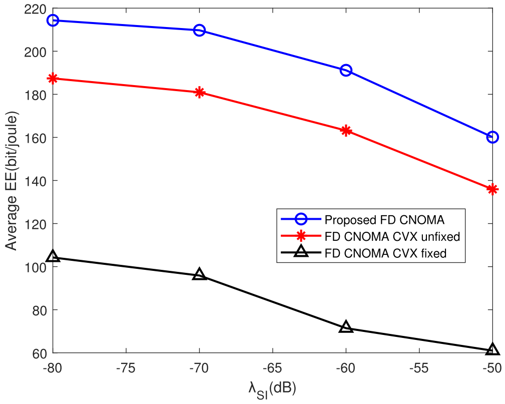


Fig. 4. Average EE achieved by the power control schemes of an FD CNOMA versus λ_{SI} for Rth $= 1.5$ bit/s and $p_B = 20$ dBm.

Fig. 5 depicts the average EE with respect to p_B/σ^2 for Rth $= 1.5$ bit/s and $\lambda_{\mathrm{SI}} = -110$ dB. First, when the value of p_B/σ^2 is small, it is shown that with increase in p_B/σ^2, the average EE for all schemes increases. This can be attributed to the fact that as p_B/σ^2 increases, more available power is allocated to the users, thereby, improving their respective EE. Another important observation is that the HD CNOMA scheme proposed in [1] achieves the highest average EE compared with the other three HD CNOMA schemes. The observation shows that the strong user with a full relaying power can enhance cooperative gain of weak user when the BS power budget p_B is lower. Then, the average EE does not always increase as p_B/σ^2 increases, and there exists an upper limit on EE for all schemes. After reaching the upper limit, the EE curves of the schemes that fix the total power of the BS on each subchannel begin to fall, while the EE values of the schemes that do not fix the total power of the BS on each subchannel no longer change. This shows that there exists an optimized total transmit power of the BS for maximizing the EE of user pair on each subchannel.

Fig. 6 presents the average EE versus the minimum rate requirement Rth for $p_B = 27$ dBm and $\lambda_{\mathrm{SI}} = -110$ dB. From Fig. 6, wc can see that when Rth increases, the average EE of all schemes decreases. An important observation is that when the minimum rate requirement is low, e.g., Rth $= 1$ bit/s, the FD CNOMA CVX unfixed scheme obtains the best performance. The reason is that in this case, the FD CNOMA schemes require very low transmit power of the BS for strong and weak users so that not fixing total power of the BS may not cause energy waste. Besides, we observe that with increase in Rth, the average EE of the HD CNOMA scheme proposed in [1] becomes closer and surpasses that of our proposed HD CNOMA scheme. The reason is that, for case of HD CNOMA, the increase in Rth requires more power budget of the BS than p_B and, thus, the full relaying power of strong user can enhance cooperative gain of CNOMA, which confirms the conclusion of Fig. 5.

It is clearly observed from the above figures that with the proper total power budget of the BS, p_B, the proposed HD CNOMA power control scheme generally outperforms the HD CNOMA scheme proposed in [1]. We also witness that the proposed HD and FD CNOMA schemes outperform the optimal CVX-based HD and FD CNOMA methods,

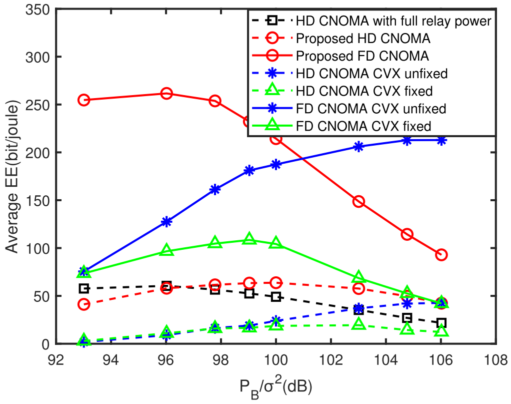


Fig. 5. Average EE versus p_B/σ^2 with different power control schemes (Rth $= 1.5$ bit/s and $\lambda_{\mathrm{SI}} = -110$ dB).

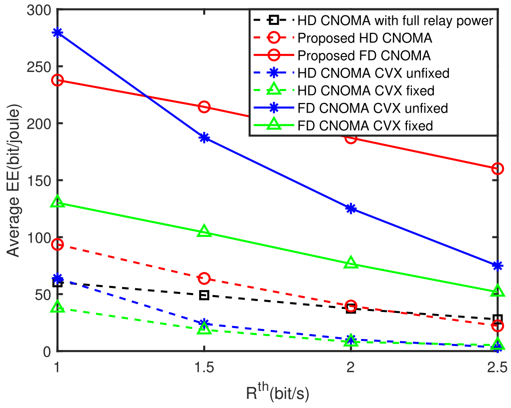


Fig. 6. Average EE versus Rth with different power control schemes ($p_B = 27$ dBm and $\lambda_{\mathrm{SI}} = -110$ dB).

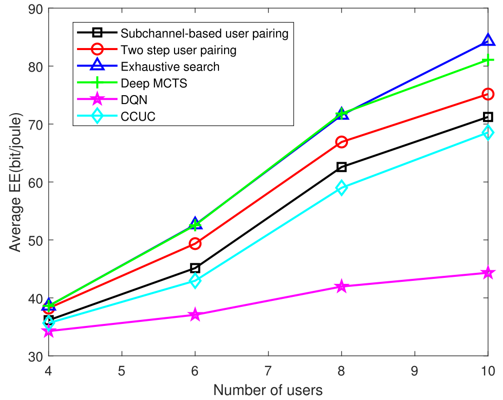


Fig. 7. Average EE versus the number of users, $2M$, with different pairing schemes (Rth $= 1.5$ bit/s).

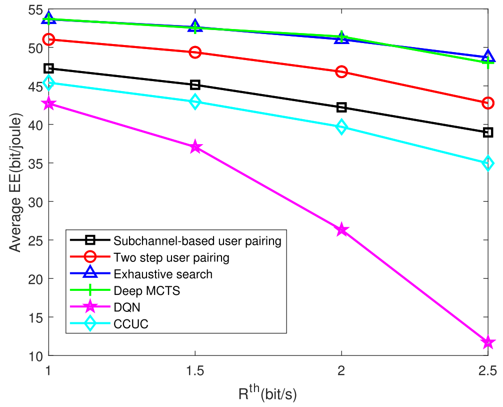


Fig. 8. Average EE versus Rth with different pairing schemes ($M = 3$).

respectively. This is because the CVX-based method in [25] relaxes the constraint of EE maximization problem so that it is formulated as a standard SDP problem. Besides, for the CVX-based schemes, the cases of not fixing the total power budget of the BS outperform the cases of fixing the total power budget of the BS.

B. Performance Comparison of Different User Pairing and Subchannel Assignment Schemes

In this section, we set $p_B = 27$ dBm and validate the performance of the proposed deep MCTS-based user pairing and subchannel assignment policy. It was compared with the three existing pairing schemes termed as: "subchannel-based user pairing" [1], [35]; "two step user pairing" [37]; and "CCUC" [46], respectively, and the DQN algorithm. Moreover, we illustrated the exhaustive search scheme as an optimal solution, where strong and weak users are grouped by their distances from the BS. Note that for all schemes, the EE of each CNOMA pair is computed according to the proposed HD CNOMA power control algorithm.

Figs. 7 and 8, respectively, illustrate the average EE of system versus the number of users for Rth $= 1.5$ bit/s and the minimum rate requirement for $M = 3$. Fig. 7 shows that increasing the number of users improves the average EE of system, while Fig. 8 shows that the increasing of the minimum rate requirement decreases the average EE of the system. We observe from Figs. 7 and 8 that the average EE of the system achieved by the proposed deep MCTS scheme is very close to the EE curve of the exhaustive search scheme, which is much higher than those of the proposed baseline schemes. It is noted that the worst performance is obtained by the DQN algorithm. The reason is that the DQN scheme to maximize the long-term cumulative reward over time slots is not suitable in the dynamic network environment when the channel states changing independently from one time slot to another. From Fig. 7, when the number of users becomes 10, a small gap exists between the proposed scheme and the exhaustive search scheme. The reason is that 500 simulations of MCTS for $M = 5$ is not enough to search the optimal policy. Particularly, the exhaustive search scheme, since grouping strong and weak users according to their distances from the BS, may result in some small gap from the proposed method in some cases, e.g., Rth $= 2$ bit/s in Fig. 8.

C. Performance Analysis of Proposed Joint Optimization Schemes

Since the joint EE optimization scheme of user pairing, subchannel assignment, and power control for the FD CNOMA system has not been involved in the existing literature, we verified the performance of the proposed joint optimization schemes only compared with the existing joint optimization

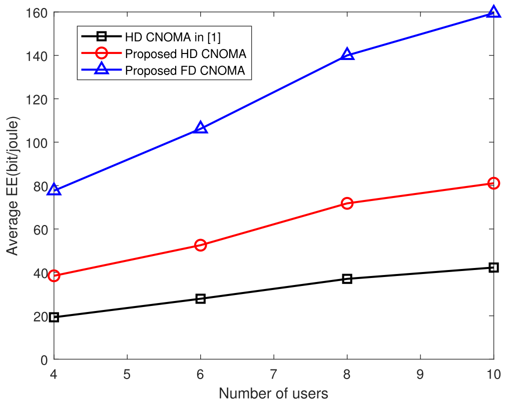


Fig. 9. Performance of the joint optimization CNOMA schemes versus the number of users, $2M$, for Rth = 1.5 bit/s.

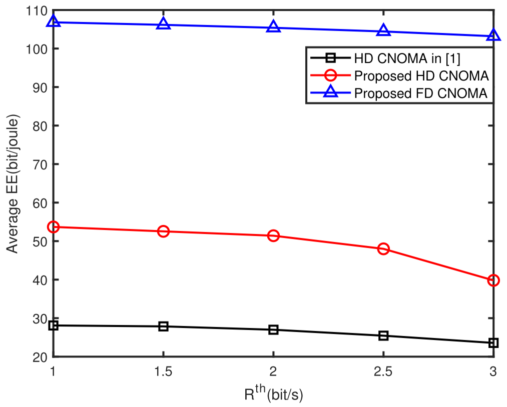


Fig. 10. Performance of the joint optimization CNOMA schemes versus Rth for $M = 3$.

scheme for case of HD CNOMA in [1]. We set $p_B = 27$ dBm and $\lambda_{SI} = -110$ dB.

Figs. 9 and 10 plot the average EE of system against the number of users for Rth $= 1.5$ bit/s and the minimum rate requirement for $M = 3$, respectively. It is shown from Figs. 9 and 10, for all schemes, the increasing number of users improves the average EE of the system, while with the increasing of the minimum rate requirement Rth, the average EE of the system decreases. We can clearly observe from Figs. 9 and 10 the better performance of the proposed two joint optimization schemes in comparison with that of the joint optimization HD CNOMA scheme proposed in [1]. For example, the average EE of system achieved by the proposed HD and FD CNOMA schemes improves roughly by 88.17% and 280.64% of the HD CNOMA scheme proposed in [1] when $M = 3$ and Rth $= 1.5$ bit/s, respectively.

VI. Conclusion

In this article, the joint optimization problem of user pairing, subchannel assignment, and power control for CNOMA was formulated and solved to maximize the achievable EE of the whole system as well as guarantee a certain required QoS of each user. First, the optimal power control closed-form expressions that maximize the EE of a given pair of users were derived for FD and HD CNOMA, respectively. Then, a joint resource optimization model based on the deep MCTS that combines an MCTS and a neural network was proposed. Simulation results have demonstrated the efficacy of the proposed CNOMA scheme over the existing NOMA schemes proposed in [1], [25], [35], [37], and [46] in terms of average EE of the system.

References

[1] A. K. Lamba, R. Kumar, and S. Sharma, "Joint user pairing, subchannel assignment and power allocation in cooperative non-orthogonal multiple access networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11790–11799, Oct. 2020.

[2] F. Kara and H. Kaya, "On the error performance of cooperative-NOMA with statistical CSIT," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 128–131, Jan. 2019.

[3] X. Li et al., "Cooperative wireless-powered NOMA relaying for B5G IoT networks with hardware impairments and channel estimation errors," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5453–5467, Apr. 2021.

[4] P. Huu, M. A. Arfaoui, S. Sharafeddine, C. M. Assi, and A. Ghrayeb, "A low-complexity framework for joint user pairing and power control for cooperative NOMA in 5G and beyond cellular networks," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6737–6749, Nov. 2020.

[5] P. Dinh, M.-A. Arfaoui, C. Assi, and A. Ghrayeb, "Exploiting antenna diversity to enhance hybrid cooperative non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2936–2940, Dec. 2020.

[6] F. Kara and H. Kaya, "Threshold-based selective cooperative NOMA: Capacity/outage analysis and a joint power allocation-threshold selection optimization," *IEEE Commun. Lett.*, vol. 24, no. 9, pp. 1929–1933, Sep. 2020.

[7] X. Lai, Q. Zhang, and J. Qin, "Cooperative NOMA short-packet communications in flat rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6182–6186, Jun. 2019.

[8] Q. Li, M. Wen, E. Basar, H. V. Poor, and F. Chen, "Spatial modulation-aided cooperative NOMA: Performance analysis and comparative study," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 715–728, Jun. 2019.

[9] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-Duplex device-to-device aided cooperative non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4467–4471, May 2017.

[10] X. Chen, G. Liu, and Z. Ma, "Statistical QoS provisioning for half/full-duplex cooperative non-orthogonal multiple access," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Toronto, ON, Canada, Sep. 2017, pp. 1–5.

[11] S. Liu, Z. Chen, J. Xie, L. Liang, M. Wang, and Y. Jia, "Optimal power allocation of cooperative NOMA with finite blocklength codes," in *Proc. 12nd Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nanjing, China, Oct. 2020, pp. 1200–1205.

[12] T. E. A. Alharbi, K. Z. Shen, and D. K. C. So, "Full-duplex cooperative non-orthogonal multiple access system with feasible successive interference cancellation," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, Antwerp, Belgium, May 2020, pp. 1–6.

[13] G. Liu, X. Chen, Z. Ding, Z. Ma, and F. R. Yu, "Hybrid half-duplex/full-duplex cooperative non-orthogonal multiple access with transmit power adaptation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 506–519, Jan. 2018.

[14] S. Al-Ahmadi, "On the achievable max–min rates of cooperative power-domain NOMA systems," *IEEE Access*, vol. 8, pp. 173112–173122, 2020.

[15] S.-L. Wang and T.-M. Wu, "Stochastic geometric performance analyses for the cooperative NOMA with the full-duplex energy harvesting relaying," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4894–4905, May 2019.

[16] Y. Liu, H. Ding, J. Shen, R. Xiao, and H. Yang, "Outage performance analysis for SWIPT-based cooperative non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1501–1505, Sep. 2019.

[17] O. Omarov, G. Nauryzbayev, S. Arzykulov, A. M. Eltawil, and M. S. Hashmi, "Outage analysis of EH-based cooperative NOMA networks over generalized statistical models," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–6.

[18] J. Li, C. W. Sung, and C. S. Chen, "Performance study of cooperative non-orthogonal multiple access with energy harvesting," in *Proc. 2nd Int. Conf. Commun. Eng. Technol. (ICCET)*, Nagoya, Japan, Apr. 2019, pp. 30–34.

[19] Y. Liu, Y. Ye, H. Ding, F. Gao, and H. Yang, "Outage performance analysis for SWIPT-based incremental cooperative NOMA networks with non-linear harvester," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 287–291, Feb. 2020.

[20] P. Dinh, M. A. Arfaoui, S. Sharafeddine, C. Assi, and A. Ghrayeb, "A low-complexity approach for sum-rate maximization in cooperative NOMA enhanced cellular networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Dublin, Ireland, Jun. 2020, pp. 1–7.

[21] Z. Zhang, H. Qu, J. Zhao, and W. Wang, "Energy efficient transmission design of cooperative NOMA with SWIPT network," in *Proc. IEEE 4th Int. Conf. Signal Image Process. (ICSIP)*, Wuxi, China, Jul. 2019, pp. 566–572.

[22] Y. Yuan, Y. Xu, Z. Yang, P. Xu, and Z. Ding, "Energy efficiency optimization in full-duplex user-aided cooperative SWIPT NOMA systems," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5753–5767, Aug. 2019.

[23] Y. Alsaba, C. Y. Leow, and S. K. A. Rahim, "Full-duplex cooperative non-orthogonal multiple access with beamforming and energy harvesting," *IEEE Access*, vol. 6, pp. 19726–19738, 2018.

[24] Y. Xu et al., "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4874–4886, Sep. 2017.

[25] Z. Wei, X. Zhu, S. Sun, J. Wang, and L. Hanzo, "Energy-efficient full-duplex cooperative nonorthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10123–10128, Oct. 2018.

[26] C. Huang, G. Chen, Y. Gong, P. Xu, Z. Han, and J. A. Chambers, "Buffer-aided relay selection for cooperative hybrid NOMA/OMA networks with asynchronous deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2514–2525, Aug. 2021.

[27] Y.-H. Xu, C.-C. Yang, M. Hua, and W. Zhou, "Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications," *IEEE Access*, vol. 8, pp. 18797–18807, 2020.

[28] D. Silver et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[29] D. Bapatla and S. Prakriya, "Adaptive multiuser cooperative NOMA scheme with energy buffer-aided near-users for high spectral and energy efficiency," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16643–16662, Sep. 2022.

[30] B. Ning, W. Hao, A. Zhang, J. Zhang, and G. Gui, "Energy efficiency-delay tradeoff for a cooperative NOMA system," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 732–735, Apr. 2019.

[31] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[32] F. Salehi, N. Neda, M. -H. Majidi, and H. Ahmadi, "Cooperative NOMA-based user pairing for URLLC: A max–min fairness approach," *IEEE Syst. J.* vol. 16, no. 3, pp. 3833–3843, Sep. 2022.

[33] Y. Ma, T. Lv, X. Li, S. Zhang, and W. Guo, "User pairing schemes in cooperative downlink NOMA system with SWIPT," in *Proc. Comput. Commun. IoT Appl. (ComComAp)*, Shenzhen, China, Oct. 2019, pp. 82–87.

[34] P. Dinh, M. A. Arfaoui, S. Sharafeddine, C. Assi, and A. Ghrayeb, "Joint user pairing and power control for C-NOMA with full-duplex device-to-device relaying," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[35] B. K. S. Lima, D. B. da Costa, R. Oliveira, R. Dinis, M. Beko, and U. S. Dias, "Power allocation, relay selection, and user pairing for cooperative NOMA systems with rate fairness," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–5.

[36] M. Obeed, H. Dahrouj, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "User pairing, link selection, and power allocation for cooperative NOMA hybrid VLC/RF systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1785–1800, Mar. 2021.

[37] Y. Cheng, K. H. Li, K. C. Teh, S. Luo, and W. Wang, "Two-step user pairing for OFDM-based cooperative NOMA systems," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 903–906, Apr. 2020.

[38] W. J. Ryu, J. W. Kim, and D.-S. Kim, "Deep reinforcement learning based cooperative retransmission in downlink NOMA systems," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju Island, South Korea, Oct. 2022, pp. 883–885.

[39] J. Chen, S. Chen, Q. Wang, B. Cao, G. Feng, and J. Hu, "iRAF: A deep reinforcement learning approach for collaborative mobile edge computing IoT networks," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7011–7024, Aug. 2019.

[40] X. Meng, H. Inaltekin, and B. Krongold, "Deep reinforcement learning-based topology optimization for self-organized wireless sensor networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[41] X. Lan, Y. Zhang, Q. Chen, and L. Cai, "Energy efficient buffer-aided transmission scheme in wireless powered cooperative NOMA relay network," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1432–1447, Mar. 2020.

[42] A. Liu, J. Chen, M. Yu, Y. Zhai, X. Zhou, and J. Liu. "Watch the Unobserved: A Simple Approach to Parallelizing Monte Carlo Tree Search." Feb. 2020. [Online]. Available: http://arxiv.org/abs/1810.11755

[43] W. Lee, "Resource allocation for multi-channel underlay cognitive radio network based on deep neural network," *IEEE Commun. Lett.*, vol. 22, no. 9, pp. 1942–1945, Sep. 2018.

[44] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." Dec. 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[45] H. Huang, Y. Yang, Z. Ding, H. Wang, H. Sari and F. Adachi, "Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 19, no.8, pp. 5373–5388, Aug. 2020.

[46] Y. Fu, M. Zhang, L. Salaun, C. W. Sung, and C. S. Chen, "Zero-forcing oriented power minimization for multi-cell MISO-NOMA systems: A joint user grouping, beamforming, and power control perspective," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1925–1940, Aug. 2020.

Yan-Yan Guo received the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2010.

She is currently an Associate Professor with the School of Physics and Electronics Engineering, Shanxi University, Taiyuan, China. Her research interests include resource allocation optimization, Internet of Things, machine learning, and UAV networks.

Xiao-Long Tan received the master's degree from Shanxi University, Taiyuan, China, in 2022.

He is currently an Algorithm Engineer with Beijing Samsung Telecommunication Research and Development Center, Beijing, China. His research interests include resource allocation optimization, machine learning, and communications algorithm.

Yun Gao received the master's degree from Shanxi University, Taiyuan, China, in 2021.

Her research interests include resource allocation optimization, Internet of Things, and machine learning.

Jing Yang received the master's degree from Shanxi University, Taiyuan, China, in 2022.

She is currently an Algorithm Engineer in Xi'an. Her research interests include resource allocation optimization, machine learning, and communications algorithm.

Zhi-Chao Rui is currently pursuing the master's degree in communication engineering with the School of Physics and Electronics Engineering, Shanxi University, Taiyuan, China.

His research interests include machine learning, artificial intelligence, and image processing.