IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 5, October, 2024, pp. 2615-2622
Journal homepage: http://iieta.org/journals/ts

Deep Reinforcement Learning-Based Signal Processing for Cell-Free Massive MIMO Networks in Coal Mine Power Grids

Zhaoxi Zhang*, Dachen Zhai, Shangge Ning, Zong Li, Chuancheng Jiang, Ping Song

Dongtan Coal Mine, Yankuang Energy Group Co., Ltd., Jining 273500, China

Corresponding Author Email: 13355126954@163.com

https://doi.org/10.18280/ts.410534

Received: 5 May 2024
Revised: 27 September 2024
Accepted: 12 October 2024
Available online: 31 October 2024

Keywords:
cell-free massive MIMO, mining area, power grid, H-DDPG, signal processing

ABSTRACT

In coal mine power grids, ensuring reliable, high-capacity, and low-latency communication is critical to maintaining efficient operations. To meet these demands, combining non-orthogonal multiple access (NOMA) with cell-free massive multiple input multiple output (CF-mMIMO) networks presents a powerful solution. This paper focuses on the signal processing and resource allocation challenges inherent in the downlink of CF-mMIMO-NOMA systems, specifically tailored to the complex communication environment of coal mines. We propose a hierarchical deep deterministic policy gradient (H-DDPG) algorithm to optimize system performance, with a focus on user pairing and power allocation. The algorithm addresses signal processing tasks at both system and link levels by leveraging two-layer control networks that operate on distinct time scales. At the system level, the user pairing problem is solved, while power allocation is optimized at the link level, with dedicated DDPG agents guiding both processes. Extensive simulations demonstrate that the proposed H-DDPG method significantly enhances the system sum rate compared to benchmark approaches, making it a robust solution for improving signal processing and resource management in coal mine power grid CF-mMIMO networks.

1. INTRODUCTION

The increasing demand for reliable, high-capacity communication poses a significant challenge to existing cellular network architectures, especially in complex environments such as coal mine power grids. In these power grids, where uninterrupted operations are crucial for both safety and efficiency, the need for advanced communication systems that can handle massive connectivity, low-latency data exchange, and high reliability is paramount. Efficient resource management becomes even more critical in these settings, as power grids in coal mines often operate in harsh conditions with substantial interference and high user density.

Non-orthogonal multiple access (NOMA) is a key technique for enhancing spectral efficiency and has attracted significant attention in both academic and industrial fields. The main idea of NOMA is to overlap multiple users' data signals in the power domain, utilizing differences in their channel gains to serve several users within the same spectrum block [1]. Unlike orthogonal multiple access (OMA), NOMA allows multiple users to share time-frequency resources while minimizing inter-user interference, leading to improved system performance [2]. This is particularly advantageous in environments such as coal mines, where efficient use of spectrum and user fairness are crucial. This paper focuses on power domain NOMA, commonly referred to as NOMA. By using successive interference cancellation (SIC), stronger users can decode and eliminate interference from weaker users, thus boosting both spectral efficiency and fairness. NOMA is also effective at reducing pilot contamination in cell-free massive multiple input multiple output (CF-mMIMO) systems by allocating more power to users with weaker channels.

A large body of research has explored the application of NOMA in wireless communications. For example, Bui et al. [3] investigates the influence of power allocation on downlink NOMA fairness under both perfect and average channel state information feedback. Similarly, Ding et al. [4] proposes pairing users with strong and weak channel gains for collaborative data transmission in NOMA systems. In study [5], user pairing and power allocation are treated as a joint optimization problem to maximize user rates under minimum rate constraints, formulated as mixed-integer programming. Additionally, Ali et al. [6] combines dynamic user clustering with power allocation in typical NOMA systems to maximize throughput, while Katwe et al. [7] proposes a two-stage dynamic user clustering method for UAS-assisted full-duplex NOMA systems to reduce cross-interference.

As traditional cellular networks struggle with issues such as low spectrum utilization and limited system capacity, researchers have proposed combining NOMA with CF-mMIMO systems as an innovative solution to address these limitations. By deploying a large number of antennas at access points and utilizing NOMA for power allocation and channel coding, CF-mMIMO systems can serve multiple users simultaneously. This combination leverages the spatial multiple access capabilities of CF-mMIMO and the multiple access advantages of NOMA, enhancing spectral efficiency and improving service quality. In coal mine power grids,

where communication environments are more challenging due to underground structures and electromagnetic interference, combining these technologies is particularly valuable. NOMA can also save pilot resources in CF-mMIMO systems, leaving more capacity for data transmission, while user pairing and power allocation further enhance the system's overall performance.

Contemporary studies highlight the critical role of combining NOMA and CF-mMIMO technologies in meeting the evolving requirements of next-generation wireless networks, particularly in achieving enhanced spectral efficiency, supporting massive connectivity, ensuring minimal latency, and maintaining equitable user access [8]. A notable contribution by Ding et al. [9] examines the upstream communication capabilities of NOMA-enabled CF-mMIMO systems, proposing an innovative algorithm that optimizes spectral efficiency through careful consideration of service quality metrics, power management, and interference mitigation strategies. Further research efforts, as demonstrated in studies [10-14], have explored various user grouping techniques that leverage channel characteristic similarities, aiming to optimize NOMA system performance while simultaneously reducing system complexity and minimizing inter-user interference effects.

In coal mine power grid environments, traditional communication systems struggle to meet the stringent communication quality requirements for equipment monitoring and safety warning services due to complex underground structures, severe electromagnetic interference, and high user density. CF-mMIMO-NOMA systems can effectively overcome multipath fading and channel attenuation in coal mine power grids through distributed antenna deployment and non-orthogonal multiple access, significantly improving system capacity and reliability. However, existing research has mainly focused on theoretical analysis, lacking system optimization solutions for practical mining applications. Deep reinforcement learning (DRL) has recently become a powerful tool for optimizing complex decisions in wireless communications. It combines deep learning and reinforcement learning to directly learn control strategies from high-dimensional data, offering a smarter approach to network optimization. In study [15], DRL is applied to optimize intelligent reflective surfaces and NOMA-assisted CF-mMIMO systems, focusing on phase shifting, power allocation, and user pairing through the Deep Deterministic Policy Gradient (DDPG) algorithm.

However, few works have jointly optimized user pairing and power allocation in CF-mMIMO-NOMA systems, particularly in the context of coal mine power grids, where robust communication is vital. To address this gap, we present a joint optimization framework for user pairing and power allocation using an improved DRL algorithm, specifically designed for CF-mMIMO-NOMA systems in coal mines. The main contributions of this paper are summarized as follows:

- We propose a joint optimization problem for user pairing and power allocation in the downlink of CF-mMIMO-NOMA systems, focusing on maximizing system sum rate while satisfying user pairing constraints and maximum transmit power constraints for each access point.
- To solve this problem, we implement a hierarchical deep deterministic policy gradient (H-DDPG) algorithm, creating a cooperative optimization framework that addresses system-level and link-level tasks on different time scales. DDPG is employed at the system level for user pairing and at the link level for power allocation, ensuring minimal interference and improved transmission performance.
- Numerical simulation results demonstrate that the proposed H-DDPG algorithm significantly improves system sum rate compared to benchmark methods, while also verifying its effectiveness in terms of convergence speed and spectral efficiency.

2. SYSTEM MODEL

As shown in Figure 1, we study the downlink of a CF-mMIMO-NOMA system for a mining area power grid, where L single-antenna APs serve N users within the same time-frequency block. The N users are grouped into M clusters, each consisting of K users, i.e., $N = MK$. Denote the k_{th} user in the m_{th} cluster with $UE_{m,k}$. All APs are linked to the CPU via a high-speed network to facilitate the exchange of collaborative information.

Utilizing a traditional TDD and block fading model, the time-frequency resources are compartmentalized into coherent blocks, allowing for the assumption that the channel coefficients remain constant within each block. Each coherent block is made up of τ_c symbols. This research solely targets the downlink resource allocation for the CF-mMIMO-NOMA system with the objective of maximizing the cumulative system throughput. Consequently, the uplink training period is hypothesized to last for τ_p symbols, and the subsequent coherence interval of τ_c-τ_p symbols is designated for the transmission of downlink data.

The channel vector between the l_{th} AP and $UE_{m,k}$ is given by:

$$\boldsymbol{g}_{l,m,k} = \sqrt{\beta_{l,m,k}}\,\boldsymbol{h}_{l,m,k} \tag{1}$$

where, $\beta_{l,m,k}$ is LSF coefficient, reflecting path loss and shadow fading; $h_{l,m,k} \sim CN(0,1)$ denotes SSF vector and considers Rayleigh fading.

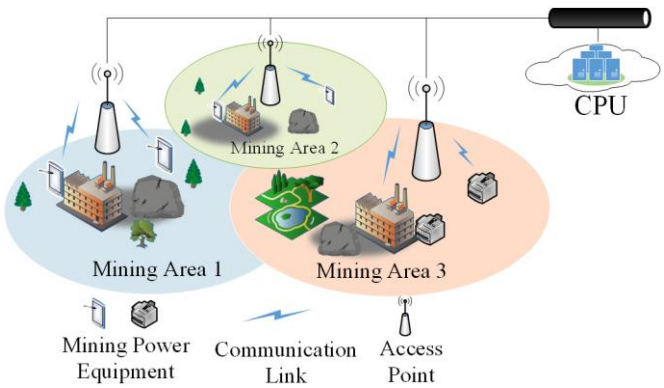

Figure 1. CF-mMIMO-NOMA system model of mining area power grid

A. Channel estimation

The AP estimates the uplink channel from user pilots using TDD reciprocity to obtain the CSI. Users in the same cluster share a pilot sequence to reduce overhead. The M pilot sequences assigned to M clusters is $\phi_m \in \mathbb{C}^{\tau_p \times 1}$, satisfying $\|\phi_m\|^2 = 1$ and for $n \neq i$, $\phi_m^H \phi_i = 0$. The pilot signal received by the l_{th} AP can be written as:

$$y_l^p = \sqrt{\tau p_p}\sum_{m=1}^{M}\sum_{k=1}^{K} g_{l,m,k}\phi_m + n_l \tag{2}$$

where, p_p represents the power allocated for transmitting the pilot signal, and $n_l \sim CN_{\tau_p \times 1}\left(0_{\tau_p \times 1}, I_{\tau_p}\right)$ signifies the vector of additive Gaussian white noise at the l_{th} AP.

In order to estimate $g_{l,m,k}$, projecting y_l^p onto ϕ_m yields:

$$\tilde{\boldsymbol{y}}_{l,m}^p = \phi_m^H \boldsymbol{y}_l^p = \sqrt{\tau p_p}\sum_{k=1}^{K}\boldsymbol{g}_{l,m,k} + \phi_m^H \boldsymbol{n}_l \tag{3}$$

According to the MMSE principle, the channel estimation of $g_{l,m,k}$ is given by

$$\begin{aligned}\hat{\boldsymbol{g}}_{l,m,k} &= \frac{E\left\{\tilde{\boldsymbol{y}}_{l,m}^{p*}\boldsymbol{g}_{l,m,k}\right\}}{E\left\{\left|\tilde{\boldsymbol{y}}_{l,m}^{p}\right|^2\right\}}\tilde{\boldsymbol{y}}_{l,m}^p \\ &= \frac{\sqrt{\tau p_p}\beta_{l,m,k}}{1+\tau p_p\sum_{k'=1}^{K}\beta_{l,m,k'}}\tilde{\boldsymbol{y}}_{l,m}^p\end{aligned} \tag{4}$$

Since $\tilde{y}_{l,m}^p$ is a Gaussian distribution, $\hat{g}_{l,m,k}$ is given by:

$$\hat{\boldsymbol{g}}_{l,m,k} = \sqrt{\delta_{l,m,k}}\boldsymbol{P}_{l,m} \tag{5}$$

where, $P_{l,m} \sim CN(0,1)$ and

$$\delta_{l,m,k} = \frac{\tau p_p \beta_{l,m,k}^2}{1+\tau p_p\sum_{k'=1}^{K}\beta_{l,m,k'}} \tag{6}$$

The corresponding estimation error is $\tilde{g}_{l,m,k} = g_{l,m,k} - \hat{g}_{l,m,k}$.

B. Data transmission

In the process of downlink data transmission, the access point (AP) utilizes a conjugate beamforming technique that is constructed based on locally estimated CSI. Presuming that the data signal for the k_{th} user in the m_{th} cluster is $\varsigma_{m,k}$, the aggregated signal dispatched by the l_{th} AP to the users within the m_{th} cluster is a composite of the data signals, where each signal is weighted by the corresponding beamforming vector informed by the locally estimated CSI.

$$\boldsymbol{x}_{l,m} = \sqrt{\rho_d}\sum_{k=1}^{K}\sqrt{\eta_{l,m,k}}\boldsymbol{P}_{l,m}^*\varsigma_{m,k} \tag{7}$$

where, ρ_d and $\eta_{l,m,k}$ denotes the transmit power and the power control coefficient. Moreover, $\eta_{l,m,k}$ needs to satisfy:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\eta_{l,m,k} \le 1, \forall l \tag{8}$$

Therefore, the signal received signal is:

$$\begin{aligned}\boldsymbol{y}_{m,k} &= \sum_{l=1}^{L}\sum_{m'=1}^{M}\boldsymbol{g}_{l,m,k}^T\boldsymbol{x}_{l,m'} + \boldsymbol{n}_{m,k} \\ &= \sqrt{\rho_d}\sum_{l=1}^{L}\sum_{m'=1}^{M}\sum_{k'=1}^{K}\sqrt{\eta_{l,m',k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m'}^*\varsigma_{m',k'} + \boldsymbol{n}_{m,k}\end{aligned} \tag{9}$$

where, $n_{m,k} \sim CN(0,1)$ is additive Gaussian white noise.

From Eq. (9), rewrite $y_{m,k}$ as:

$$\begin{aligned}\boldsymbol{y}_{m,k} &= \sqrt{\rho_d}\sum_{l=1}^{L}\sqrt{\eta_{l,m,k}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m}^*\varsigma_{m,k} \\ &+ \sqrt{\rho_d}\sum_{l=1}^{L}\sum_{k'\neq k}^{K}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m}^*\varsigma_{m,k'} \\ &+ \sqrt{\rho_d}\sum_{l=1}^{L}\sum_{m'\neq m}^{M}\sum_{k'=1}^{K}\sqrt{\eta_{l,m',k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m'}^*\varsigma_{m',k'} + \boldsymbol{n}_{m,k}\end{aligned} \tag{10}$$

Based on the effective channel gain, we rank the users in the m_{th} cluster, as:

$$\begin{aligned}&E\left\{\sum_{l=1}^{L}\hat{\boldsymbol{g}}_{l,m,1}^T\boldsymbol{P}_{l,m}^*\right\} \ge E\left\{\sum_{l=1}^{L}\hat{\boldsymbol{g}}_{l,m,2}^T\boldsymbol{P}_{l,m}^*\right\} \\ &\ge \ldots \ge E\left\{\sum_{l=1}^{L}\hat{\boldsymbol{g}}_{l,m,K}^T\boldsymbol{P}_{l,m}^*\right\}, \forall m\end{aligned} \tag{11}$$

In power domain NOMA, higher power is allocated to users with lower channel gain, allowing the k_{th} user in the m_{th} cluster to decode the i_{th} user's signal after decoding its own. For $\forall i \ge k$, the k_{th} user eliminates interference from the ith user's signal via successive interference cancellation (SIC) before decoding its own. Signals from users $\forall i < k$ are treated as interference. Therefore, the above equation can be rewritten as:

$$\begin{aligned}\boldsymbol{y}_{m,k} &= \sqrt{\rho_d}\sum_{l=1}^{L}\sqrt{\eta_{l,m,k}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m}^*\varsigma_{m,k} + \sqrt{\rho_d}\sum_{l=1}^{L}\sum_{k'=1}^{k-1}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m}^*\varsigma_{m,k'} \\ &+ \sqrt{\rho_d}\left(\sum_{l=1}^{L}\sum_{k'=k+1}^{K}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m}^*\varsigma_{m,k'} - E\left\{\sum_{l=1}^{L}\sum_{k'=k+1}^{K}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m}^*\hat{\varsigma}_{m,k'}\right\}\right) \\ &+ \sqrt{\rho_d}\sum_{l=1}^{L}\sum_{m'\neq m}^{M}\sum_{k'=1}^{K}\sqrt{\eta_{l,m',k'}}\boldsymbol{g}_{l,m,k}^T\boldsymbol{P}_{l,m'}^*\varsigma_{m',k'} + \boldsymbol{n}_{m,k}\end{aligned} \tag{12}$$

where, $\hat{\varsigma}_{m,k'}$ is an estimate of $\varsigma_{m,k'}$. There are:

$$\varsigma_{m,k'} = c_{m,k'}\hat{\varsigma}_{m,k'} + b_{m,k'} \tag{13}$$

where, $\hat{\varsigma}_{m,k'} \sim CN(0,1)$, $b_{m,k'} \sim CN\left(\sigma_{b_{m,k'}}^2/\left(1+\sigma_{b_{m,k'}}^2\right)\right)$ and $c_{m,k'} = 1/\left(1+\sigma_{b_{m,k'}}^2\right)^{\frac{1}{2}}$. The parameter $c_{m,k'}$ reflects the

correlation between $\hat{\varsigma}_{m,k'}$ and $\varsigma_{m,k'}$. Note that $c_{m,k'}=1$ does not stand for complete SIC [16, 17]. This is because when $c_{m,k'}=1$, the interference term in the residual cluster is not equal to zero. Assuming that the statistics CSI is known, it can be further written.

$$\begin{aligned}\boldsymbol{y}_{m,k} &= DS_{m,k}\varsigma_{m,k} + BU_{m,k}\varsigma_{m,k} \\ &+ \sum_{k'=1}^{k-1} ICI_{m,k,k'}\varsigma_{m,k'} + \sum_{k'=k+1}^{K} RICI_{m,k,k'} \\ &+ \sum_{m'\neq m}^{M}\sum_{k'=1}^{K} UI_{m,k,m',k'}\varsigma_{m',k'} + \boldsymbol{n}_{m,k}\end{aligned} \tag{14}$$

Among them,

$$DS_{m,k} = \sqrt{\rho_d}E\left\{\sum_{l=1}^{L}\sqrt{\eta_{l,m,k}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m}^{*}\right\} \tag{15}$$

$$BU_{m,k} = \sqrt{\rho_d}\left(\begin{array}{l}\sum_{l=1}^{L}\sqrt{\eta_{l,m,k}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m}^{*} \\ -E\left\{\sum_{l=1}^{L}\sqrt{\eta_{l,m,k}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m}^{*}\right\}\end{array}\right) \tag{16}$$

$$ICI_{m,k,k'} = \sqrt{\rho_d}\sum_{l=1}^{L}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m}^{*} \tag{17}$$

$$RICI_{m,k,k'} = \sqrt{\rho_d}\left(\begin{array}{l}\sum_{l=1}^{L}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m}^{*}\varsigma_{m,k'} \\ -E\left\{\sum_{l=1}^{L}\sqrt{\eta_{l,m,k'}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m}^{*}\hat{\varsigma}_{m,k'}\right\}\end{array}\right) \tag{18}$$

$$UI_{m,k,m',k'} = \sqrt{\rho_d}\sum_{l=1}^{L}\sqrt{\eta_{l,m',k'}}\boldsymbol{g}_{l,m,k}^{T}\boldsymbol{P}_{l,m'}^{*} \tag{19}$$

They correspond to the coherent beamforming gain, variations in beamforming gain, interference within the cluster, residual interference due to imperfect SIC, and interference between clusters, all of which collectively influence the achievable rate as:

$$R_{m,k} = \frac{\tau_c-\tau_p}{\tau_c}\log_2\left(1+SINR_{m,k}\right) \tag{20}$$

Among them,

$$SINR_{m,k} = \frac{\left|DS_{m,k}\right|^2}{E\left\{\left|BU_{m,k}\right|^2\right\}+\sum_{k'=1}^{k-1}E\left\{\left|ICI_{m,k,k'}\right|^2\right\}+\sum_{k'=k+1}^{K}E\left\{\left|RICI_{m,k,k'}\right|^2\right\}+\sum_{m'\neq m}^{M}\sum_{k'=1}^{K}E\left\{\left|UI_{m,k,m',k'}\right|^2\right\}+1} \tag{21}$$

3. PROBLEM FORMULATION

This chapter covers the joint optimization of user pairing and power allocation to maximize the sum rate in CF-mMIMO-NOMA systems. User pairing variables $\alpha_{n,n'}$ are introduced, and the optimization problem is defined as follows:

$$\begin{aligned}&\underset{\alpha_{n,n'},\eta_{l,m,k}}{\text{Maximize}} \quad \sum_{m=1}^{M}\sum_{k=1}^{K}R_{m,k} \\ &s.t. \quad \text{C1: } \alpha_{n,n'}\in\{0,1\},\ \forall n,n' \\ &\text{C2: } \alpha_{n,n}=0,\ \forall n \\ &\text{C3: } \sum_{n'=1}^{N}\alpha_{n,n'}\le 1,\ \forall n \\ &\text{C4: } \sum_{n=1}^{N}\alpha_{n,n'}\le 1,\ \forall n' \\ &\text{C5: } \alpha_{n,n'}+\sum_{i=1}^{N}\alpha_{n',i}\le 1,\ \forall n,n' \\ &\text{C6: } \alpha_{n,n'}+\sum_{i=1}^{N}\alpha_{n,i}\le 1,\ \forall n,n' \\ &\text{C7: } \sum_{m=1}^{M}\sum_{k=1}^{K}\eta_{l,m,k}\le 1,\ \forall l \\ &\text{C8: } \eta_{l,m,k}\ge 0,\ \forall l,m,k\end{aligned} \tag{22}$$

The problem is under the corresponding user pairing constraints and maximum transmit power constraints for each AP to maximize the sum rate. Constraints C1-C2 and C8 are necessary for user pairing and power control factors, respectively. Constraints C3-C6 guarantee that each user is involved in a pairing with no more than a single other user, and that every user is part of precisely one pairing. Constraint C7 verifies that the transmit power of the AP remains within the maximum allowable power limit. The non-convex optimization problem and large number of APs and users in CF-mMIMO-NOMA systems make joint user pairing and power distribution complex. To tackle this, we apply the HDRL algorithm in the next section for an effective solution.

4. OPTIMIZATION SCHEME BASED ON H-DDPG ALGORITHM

Joint optimization in the CF-mMIMO-NOMA system of a mining area power grid involves optimizing multiple variables, making the problem highly complex. Additionally, user pairing is a network-wide decision, while power allocation is specific to individual links after connections are established. As a result, these two sub-problems must be addressed at different timescales and scopes, due to their distinct impacts and optimization objectives. To tackle these challenges, we introduce an H-DDPG framework to simultaneously optimize user pairing and power allocation within CF-mMIMO-NOMA systems, utilizing a "divide and conquer" strategy. This framework employs a two-tier control network architecture operating on different timescales to improve the system's overall data rate through hierarchical coordinated optimization.

A. HDRL Framework

User pairing and power allocation joint optimization problems usually have a lot of nonlinear constraints. It is difficult for traditional optimization methods to solve these

nonlinear problems effectively, and the computational complexity of convex optimization algorithm is higher. On the other hand, the DRL algorithm possesses certain advantages when it comes to addressing complex nonlinear issues. As a DRL framework, HDRL is comprised of two distinct tiers of DRL. This stratified methodology is designed to handle intricate decision-making challenges more effectively by breaking them down into higher-level and lower-level tasks, thereby enhancing the efficiency and effectiveness of the solutions derived.

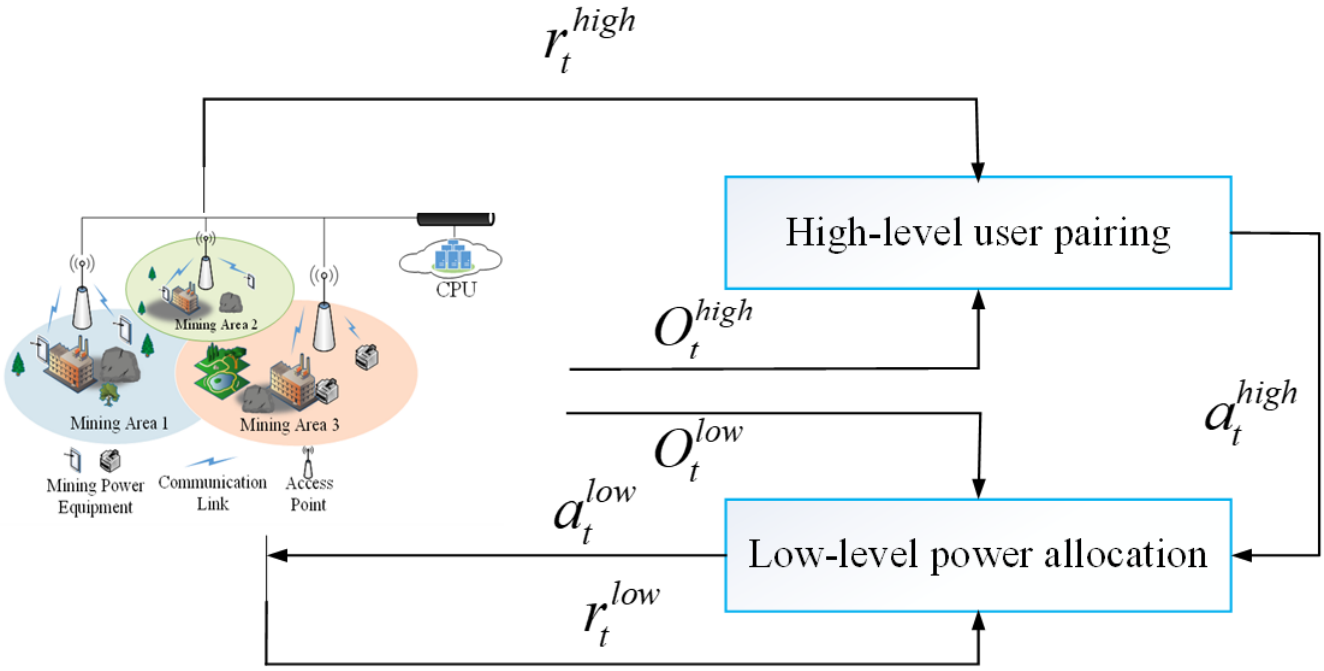

Figure 2. HDRL algorithm framework

A hierarchical DRL framework, shown in Figure 2, is proposed, using high- and low-level control strategies for network optimization through cooperative hierarchy. The system requires a large action space due to the many users and APs. To address this, the DDPG algorithm is employed for fast and stable learning. The H-DDPG algorithm contains two levels of DDPG agents: the high-level agent is responsible for user pairing decisions, while the low-level agent executes power allocation. The two agents collaborate through shared state and reward information. The high-level DDPG agent updates user pairing strategy each training cycle. The low-level DDPG agent optimizes power allocation based on current pairing results. The two-level agents alternate training until convergence

B. Optimization Schemes

When H-DDPG algorithm is used to optimize user pairing and power allocation, two DDPG agents are also designed, namely agent $UE(s_t^{UE}, a_t^{UE}, r_t^{UE})$ and agent $PA(s_t^{PA}, a_t^{PA}, r_t^{PA})$. Agent *UE* aims to satisfy the corresponding constraints and obtain a better user pairing scheme. The agent *PA* allocates transmission power to the user to minimize inter-user interference, thereby improving signal quality. In general, the agent *UE* relies on the agent *PA* to group users, and the return of the agent *UE*'s actions depends on the agent *PA*'s actions. On the other hand, the input s_t^{PA} of the agent *PA* depends on the action and environmental state of the agent *UE*. The states, actions, and reward functions associated with these two agents are defined as follows.

1. User Pairing

User pairing constitutes a holistic decision that encompasses the entire network, whereas power allocation is considered a localized decision once a specific connection is set up. These two issues possess distinct realms of impact and optimization goals, and they necessitate resolution at varying time scales and levels of abstraction. Thus, user pairing is handled as a high-level DDPG task, with the algorithm defined as follows.

1) State s_t

The state is given as follows:

$$s_t = \left\{\alpha_{1,1}^{t-1}, \ldots, \alpha_{N,N}^{t-1}, \boldsymbol{g}_{1,1,1}, \ldots, \boldsymbol{g}_{L,M,K}, R_{1,1}, \ldots, R_{M,K}\right\} \tag{23}$$

2) Action a_t

The action is given as:

$$a_t = \left\{\alpha_{1,1}, \alpha_{1,2}, \ldots, \alpha_{N,N}\right\} \tag{24}$$

3) Reward $r(s_t, a_t)$

The reward function is given by

$$r\left(s_t, a_t\right) = \sum_{m=1}^{M}\sum_{k=1}^{K} R_{m,k} - R_{penalty} \tag{25}$$

where, $R_{penalty}$ is a penalty term that gives a penalty when the user pairing constraint is not satisfied. Algorithm 1 presents the high-level DDPG algorithm.

2. Power Allocation

The DDPG algorithm for power allocation, which acts as the lower tier within the H-DDPG framework, is constructed in the following manner.

1) State s_t

The state space consists of the user clusters α^t, the channel

coefficient vector $\boldsymbol{g}_{l,m,k}$, the transmit power η^{t-1}, and the rate $R_{m,k}$, and is given by:

$$s_t = \begin{Bmatrix} \alpha_{1,1}^t,\ldots,\alpha_{N,N}^t,\boldsymbol{g}_{1,1,1},\ldots,\boldsymbol{g}_{L,M,K}, \\ \eta_{1,1,1}^{t-1},\ldots,\eta_{L,M,K}^{t-1},R_{1,1},\ldots,R_{M,K} \end{Bmatrix} \tag{26}$$

Algorithm 1: The high-level DDPG algorithm

1: **Initialize:**High level network parameter
2: **for** episode = 1 to EP **do**
3: **for** t = 1 to T_1 **do**
4: Initialize the exploration noise $\mathcal{N}_t^H$ at time t
5: Add noise to the action, $a_t^H = \mu\left(s_t^H \mid \theta_H^\mu\right) + \mathcal{N}_t^H$
6: Apply tanh to each element of a_t^H, projected to a finite range [-1, 1]
7: If a_t^H is greater than zero, consider action $a_t^H = 1$, otherwise set $a_t^H = 0$
8: Evaluate action a_t^H taken by the actor
9: Obtain the reward r_t^H and next state s_{t+1}^H
10: Store $b_t^H = \left(s_t^H, a_t^H, r_t^H, s_{t+1}^H\right)$ in replay buffer D_H
11: Sample a mini-batch B from D_H randomly
12: Evaluate target $y_t^H = r\left(s_t^H, a_t^H\right) + \gamma Q'\left(s_{t+1}^H, \mu'\left(s_{t+1}^H \mid \theta^{\mu'}\right) \mid \theta^{Q'}\right)$
13: Determine the gradient:

$$\nabla_{\theta_H^\mu} J\left(\theta_H^\mu\right) = \nabla_{\theta_H^\mu} \frac{1}{|\mathrm{B}|} \sum_{b_t \in \mathrm{B}} Q_H\left(s_t^H, \mu\left(s_t^H \mid \theta_H^\mu\right) \mid \theta_H^Q\right)$$

$$\nabla_{\theta_H^Q} L\left(\theta_H^Q\right) = \nabla_{\theta_H^Q} \frac{1}{|\mathrm{B}|} \sum_{b_t \in \mathrm{B}} \left(Q_H\left(s_t^H, a_t^H \mid \theta_H^Q\right) - y_t^H\right)^2$$

14: Refine the parameters within the assessment network.

$$\theta_H^\mu \leftarrow \theta_H^\mu + \alpha^\mu \nabla_{\theta_H^\mu} J\left(\theta_H^\mu\right), \theta_H^Q \leftarrow \theta_H^Q - \alpha^Q \nabla_{\theta_H^Q} L\left(\theta_H^Q\right)$$

15: Synchronize the weights in the target network.

$$\theta_H^{\mu'} \leftarrow \tau\theta_H^{\mu'} + (1-\tau)\theta_H^\mu, \theta_H^{Q'} \leftarrow \tau\theta_H^{Q'} + (1-\tau)\theta_H^Q$$

16: **end**
17: **end**

2) Action a_t

The agent's action, defined as the transmission power allocated to the user, is:

$$a_t = \left\{\eta_{1,1,1}, \eta_{1,1,2}, \eta_{1,2,2}, \ldots, \eta_{L,M,K}\right\} \tag{27}$$

3) Reward $r(s_t, a_t)$

The reward function, based on the problem's objective, is defined as:

$$r\left(s_t, a_t\right) = \sum_{m=1}^{M}\sum_{k=1}^{K} R_{m,k} - \lambda\left(\sum_{l=1}^{L} u\left(\sum_{m=1}^{M}\sum_{k=1}^{K} \eta_{l,m,k} - 1\right)\right) \tag{28}$$

where, $\lambda = 10^4$, $u(x)$ is step functions that are 0 when $x \leq 0$ and 1 otherwise. A penalty applies if the maximum transmit power constraint is violated.

As can be seen, although the proposed strategies have different objectives, by coordinating their optimization processes, they are able to promote monotonic improvement of the system and the rate. Algorithm 1 presents the low-level DDPG algorithm.

The H-DDPG algorithm updates high- and low-level networks sequentially, with agents refining decisions to improve system transmission performance. The full steps are in Algorithm 3.

Algorithm 2: The main steps of low-level DDPG algorithm

1: **Initialize:** Low-level network parameters
2: **for** episode = 1 to EP **do**
3: **if** t can be divided by ΔT **or (** $T_1 < t < T_2$ **) do**
4: Initialize the exploration noise $\mathcal{N}_t^L$ at time t
5: Add noise to the action, $a_t^L = \mu\left(s_t^L \mid \theta_L^\mu\right) + \mathcal{N}_t^L$
6: Apply sigmoid to each element of a_t^L, projected to a finite range [0,1].
7: Evaluate action a_t^L taken by the low-level actor and get a_t^H based on the high-level actor
8: Obtain the reward r_t^L and next state s_{t+1}^L
9: Store $b_t^L = \left(s_t^L, a_t^L, r_t^L, s_{t+1}^L\right)$ in replay buffer D_L
10: Sample a mini-batch B from D_L randomly
11: Evaluate target $y_t^L = r\left(s_t^L, a_t^L\right) + \gamma Q'\left(s_{t+1}^L, \mu'\left(s_{t+1}^L \mid \theta^{\mu'}\right) \mid \theta^{Q'}\right)$
12: Determine the gradient:

$$\nabla_{\theta_L^\mu} J\left(\theta_L^\mu\right) = \nabla_{\theta_L^\mu} \frac{1}{|\mathrm{B}|} \sum_{b_t \in \mathrm{B}} Q_L\left(s_t^L, \mu\left(s_t^L \mid \theta_L^\mu\right) \mid \theta_L^Q\right)$$

$$\nabla_{\theta_L^Q} L\left(\theta_L^Q\right) = \nabla_{\theta_L^Q} \frac{1}{|\mathrm{B}|} \sum_{b_t \in \mathrm{B}} \left(Q_L\left(s_t^L, a_t^L \mid \theta_L^Q\right) - y_t^L\right)^2$$

13: Update the weights in the evaluation network

$$\theta_L^\mu \leftarrow \theta_L^\mu + \alpha^\mu \nabla_{\theta_L^\mu} J\left(\theta_L^\mu\right), \theta_L^Q \leftarrow \theta_L^Q - \alpha^Q \nabla_{\theta_L^Q} L\left(\theta_L^Q\right)$$

14: Update the weights in the target network

$$\theta_L^{\mu'} \leftarrow \tau\theta_L^{\mu'} + (1-\tau)\theta_L^\mu, \theta_L^{Q'} \leftarrow \tau\theta_L^{Q'} + (1-\tau)\theta_L^Q$$

15: **end**
16: **end**

Algorithm 3: The main steps of H-DDPG algorithm

1: **Initialize:**High-level and low-level network parameters;
2: **for** episode = 1 to EP **do**
3: **for** t = 1 to T_1 **do**
4: **Algorithm 1:** The high-level DDPG algorithm's main step
5: **if** t can be divided by ΔT **or (** $T_1 < t < T_2$ **) do**
6: **Algorithm 2:** The low-level DDPG algorithm consists of the following main steps
7: **end**
8: **end**

5. NUMERICAL RESULTS

In this section, we delineate the implementation specifics of the HDRL approach, succeeded by an assessment of its efficacy and a comparative analysis with other prevailing techniques.

A. Simulation Parameters

All APs and users are assumed to be randomly distributed within a 0.5-kilometer radius. Once a random topology is generated, the user and AP locations are fixed during the evaluation phase. Each AP has a maximum downlink transmit power $p_{\max} = 1\,W$, and each user transmits with uplink power

$p_i = 100\,mW$ during the pilot phase. The path loss model used to generate the large-scale fading coefficients is

$$\beta_{l,m,k} = d_{l,m,k}^{-\mu} \tag{29}$$

where, μ represents the path loss and $d_{l,m,k}$ denotes the distance between the l_{th} AP and the k_{th} user in the m_{th} cluster. Table 1 summarizes the simulation parameters. The simulation parameters are chosen based on actual coal mine communication environments: 1W maximum transmission power meets safety requirements, and the path loss model references underground tunnel propagation characteristics.

Table 1. Environmental parameters

Regional Area	0.5 km × 0.5 km
Bandwidth	20 MHz
AP number	$L = 12$
Number of UEs	$N = 8$
Maximum transmit power of a single AP	$P_{\max}^{\text{dl}} = 1W$
Coherent block length	$\tau_c = 200$
Pilot length	$\tau_p = N$

In the H-DDPG algorithm, the neural network architecture is consistent across both the high-level and low-level tiers. The actor network comprises a fully connected feedforward neural network with two layers, featuring a single hidden layer that incorporates both batch normalization and layer normalization processes. The critic neural network's overall structure is a fully connected network with three layers, including two hidden layers. The hidden layers undergo batch normalization and layer normalization.

B. Discussion

To evaluate the H-DDPG algorithm's performance in CF-mMIMO-NOMA resource allocation, it is compared with three algorithms: "S-DDPG," a single-layer DDPG for joint user pairing and power allocation; "UE-DDPG," which uses DDPG for user pairing and a traditional method for power allocation; and "Tradition," a conventional joint optimization approach.

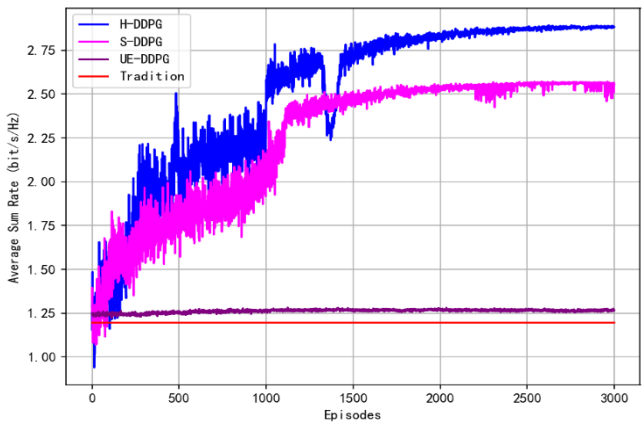

Figure 3. Comparison of average sum rate under different schemes at $L = 18$ and $N = 10$

The results in Figure 3 show that the H-DDPG algorithm achieves the highest sum rate in the CF-mMIMO-NOMA system. While the S-DDPG algorithm performs slightly worse, the AP-DDPG algorithm only marginally outperforms the conventional optimization scheme. Despite the lower computational complexity of S-DDPG and AP-DDPG, H-DDPG significantly excels in optimization performance. This is because H-DDPG tackles user pairing and power allocation separately, with each assigned to a distinct policy network, allowing more flexible and efficient decision-making. As the number of APs and users increases, the performance advantages of H-DDPG become even more pronounced.

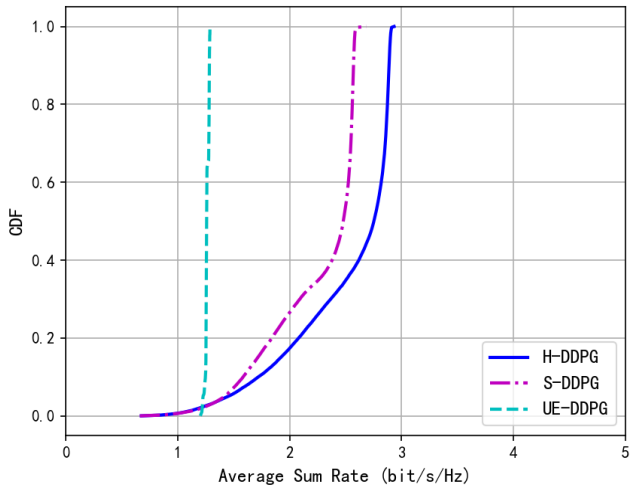

Figure 4. CDF of sum rate under different schemes for $L = 18$ and $N = 10$

Figure 4 shows the CDF of sum rate across different schemes. The H-DDPG algorithm outperforms the others, followed by S-DDPG, which is slightly behind but still better than AP-DDPG. The H-DDPG algorithm leverages its hierarchical structure for more accurate decision-making and system optimization, allowing it to acquire strategies more efficiently and achieve faster convergence.

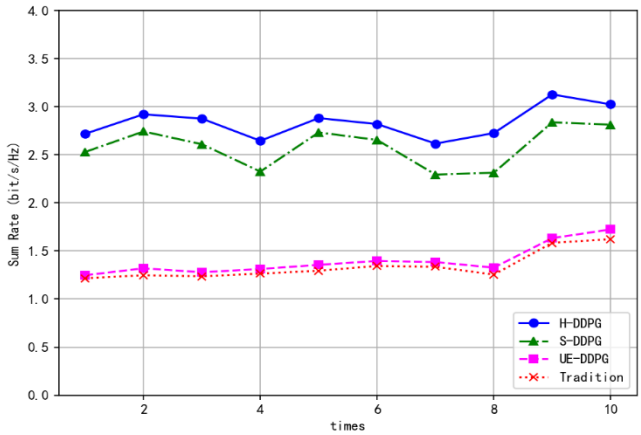

Figure 5. Variation of sum rate with relative position between AP and user

Figure 5 illustrates the effect of changing relative positions between APs and users on the system's sum rate over time. The H-DDPG algorithm consistently outperforms the other methods, while S-DDPG is only slightly behind and performs well. The AP-DDPG algorithm shows only a slight improvement over the conventional method. H-DDPG's superior performance stems from its hierarchical control strategy, which efficiently handles multi-objective problems, improving task management and execution. Additionally, H-DDPG balances exploration and execution, enabling finer strategy adjustments throughout the learning process, allowing

the agent to explore the environment more effectively.

The advantages of the H-DDPG algorithm include: 1) hierarchical design reduces problem complexity; 2) two-level agent coordination improves convergence speed; 3) multi-timescale decision-making enhances system performance. Limitations include: 1) relatively high training overhead; 2) high requirements for channel estimation accuracy; 3) computational complexity considerations for practical deployment.

6. CONCLUSION

In this paper, we proposed an HDRL algorithm to address the joint optimization of user pairing and power allocation in the CF-mMIMO-NOMA system tailored for coal mine power grids. These grids demand high reliability, low-latency communication, and efficient resource allocation to ensure safe and continuous operation. Our approach is designed to meet user pairing and maximum transmit power constraints for each AP, aiming to enhance the system's sum rate. To solve this non-convex optimization problem, we developed a hierarchical optimization framework using the H-DDPG algorithm, which divides user pairing and power allocation into two sub-problems. Each is managed by separate DDPG agents trained to optimize performance through interaction with the environment. Simulation results demonstrate that the H-DDPG algorithm significantly improves the system sum rate compared to benchmark methods, highlighting its effectiveness in meeting the demanding requirements of coal mine power grids.

REFERENCES

[1] Timotheou, S., Krikidis, I. (2015). Fairness for non-orthogonal multiple access in 5G systems. IEEE Signal Processing Letters, 22(10): 1647-1651. https://doi.org/10.1109/LSP.2015.2417119

[2] Chen, X., Gong, F.K., Li, G., Zhang, H., Song, P. (2017). User pairing and pair scheduling in massive MIMO-NOMA systems. IEEE Communications Letters, 22(4): 788-791. https://doi.org/10.1109/LCOMM.2017.2776206

[3] Bui, V.P., Nguyen, P.X., Nguyen, H.V., Nguyen, V.D., Shin, O.S. (2019). Optimal user pairing for achieving rate fairness in downlink NOMA networks. In 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan, pp. 575-578. https://doi.org/10.1109/ICAIIC.2019.8669061

[4] Ding, Z., Peng, M., Poor, H.V. (2015). Cooperative non-orthogonal multiple access in 5G systems. IEEE Communications Letters, 19(8): 1462-1465. https://doi.org/10.1109/LCOMM.2015.2441064

[5] Zhu, L., Zhang, J., Xiao, Z., Cao, X., Wu, D.O. (2018). Optimal user pairing for downlink non-orthogonal multiple access (NOMA). IEEE Wireless Communications Letters, 8(2): 328-331. https://doi.org/10.1109/LWC.2018.2853741

[6] Ali, M.S., Tabassum, H., Hossain, E. (2016). Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems. IEEE Access, 4: 6325-6343. https://doi.org/10.1109/ACCESS.2016.2604821

[7] Katwe, M., Singh, K., Sharma, P.K., Li, C.P., Ding, Z. (2021). Dynamic user clustering and optimal power allocation in UAV-assisted full-duplex hybrid NOMA system. IEEE Transactions on Wireless Communications, 21(4): 2573-2590. https://doi.org/10.1109/TWC.2021.3113640

[8] Chen, X., Ng, D.W.K., Yu, W., Larsson, E.G., Al-Dhahir, N., Schober, R. (2020). Massive access for 5G and beyond. IEEE Journal on Selected Areas in Communications, 39(3): 615-637. https://doi.org/10.1109/JSAC.2020.3019724

[9] Ding, Z., Lei, X., Karagiannidis, G.K., Schober, R., Yuan, J., Bhargava, V.K. (2017). A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. IEEE Journal on Selected Areas in Communications, 35(10): 2181-2195. https://doi.org/10.1109/JSAC.2017.2725519

[10] Zhang, L.M., Cong, Y., Meng, F.Z., Wang, Z.Q., Zhang, P., Gao, S. (2021). Energy evolution analysis and failure criteria for rock under different stress paths. Acta Geotechnica, 16(2), 569-580. https://doi.org/10.1007/s11440-020-01028-1

[11] Zhao, W., Wang, H., Song, R. (2021). Two user clustering schemes for cell-free massive MIMO-NOMA system. In 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), Changsha, China, pp. 1-5. https://doi.org/10.1109/WCSP52459.2021.9613341

[12] Nguyen, T.K., Nguyen, H.H., Tuan, H.D. (2020). Max-min QoS power control in generalized cell-free massive MIMO-NOMA with optimal backhaul combining. IEEE Transactions on Vehicular Technology, 69(10): 10949-10964. https://doi.org/10.1109/TVT.2020.3006054

[13] Rezaei, F., Heidarpour, A.R., Tellambura, C., Tadaion, A. (2020). Underlaid spectrum sharing for cell-free massive MIMO-NOMA. IEEE Communications Letters, 24(4): 907-911. https://doi.org/10.1109/LCOMM.2020.2966195

[14] Zhang L.M., Chao W.W., Liu Z.Y., Cong Y., Wang Z.Q. (2022). Crack propagation characteristics during progressive failure of circular tunnels and the early warning thereof based on multi-sensor data fusion. Geomechanics and Geophysics for Geo-Energy and Geo-Resources, 8: 172. https://doi.org/10.1007/s40948-022-00482-3

[15] Dang, X.T., Nguyen, H.V., Shin, O.S. (2023). Optimization of IRS-NOMA-assisted cell-free massive MIMO systems using deep reinforcement learning. IEEE Access, 11: 94402-94414. https://doi.org/10.1109/ACCESS.2023.3310283

[16] Li, Y., Baduge, G.A.A. (2018). NOMA-aided cell-free massive MIMO systems. IEEE Wireless Communications Letters, 7(6): 950-953. https://doi.org/10.1109/LWC.2018.2841375

[17] Zhang, J., Fan, J., Ai, B., Ng, D.W.K. (2020). NOMA-based cell-free massive MIMO over spatially correlated Rician fading channels. In ICC 2020-2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, pp. 1-6. https://doi.org/10.1109/ICC40277.2020.9148861