A MIMO-NOMA Framework With Complex-Valued Power Coefficients

Di Tong, Yuehua Ding, Yonggui Liu, and Yide Wang

***Abstract*—This paper proposes a widely linear processing framework for multiple-input multiple-output non-orthogonal multiple access (MIMO-NOMA) downlink systems. In the framework, a widely linear MIMO-NOMA (WL-MIMO-NOMA) model is derived by assuming that the base station transmits real-valued downlink signals. WL-MIMO-NOMA adopts complex-valued power allocation coefficients to stagger the user signals in phase. The main features of WL-MIMO-NOMA are the following: first, in general case, WL-MIMO-NOMA can remove all the inter-cluster interference and at least half of the intra-cluster interference; and second, in user pairing case, both interferences can be completely eliminated. This is distinct from the existing work where real power coefficients are used, which cannot guarantee the complete separation of the paired user signals because the signals transmitted to the paired users are overlapped in phase. In addition, the closed-form expressions of outage probabilities are derived. The phase difference of the complex power coefficients is optimized to minimize the outage probability. It is proven that, with the optimal phase difference, successive interference cancellation is unnecessary in user pairing case. Finally, the framework is extended to the mixed case of real/complex circular signals. Simulation results show that the proposed framework outperforms the existing work, and the numerical results agree well with the analytical analysis.**

***Index Terms*—Non-orthogonal multiple access (NOMA), multiple-input multiple-output (MIMO), widely linear processing, detection.**

Manuscript received May 6, 2018; revised August 22, 2018 and November 11, 2018; accepted December 25, 2018. Date of publication January 7, 2019; date of current version March 14, 2019. The work of D. Tong and Y. Ding was supported in part by the Guangzhou Science and Technology Program (201807010071), in part by the National Natural Science Foundation of China under Grants 61501192 and 61401157, and in part by the funds of central universities under Grant 2018ZD09. The work of Y. Liu was supported in part by the National Natural Science Foundation of China under Grant 61573153, in part by the Natural Science Foundation of Guangdong Province under Grant 2016A030313510, and in part by the funds of central universities (2018ZD51). The review of this paper was coordinated by Prof. S.-H. Leung. *(D. Tong and Y. Ding are co-first authors and contributed equally to this work.) (Corresponding author: Yonggui Liu.)*

D. Tong and Y. Ding are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China, and also with the Sino-French Research Center in Information and Communication, Guangzhou 510641, China (e-mail: 1466926693@qq.com; eeyhding@scut.edu.cn).

Y. Liu is with the Key Laboratory of Autonomous Systems and Network Control, Ministry of Education, School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: auygliu@scut.edu.cn).

Y. Wang is with the IETR Laboratory, UMR 6164, University of Nantes, Nantes 44300, France, and also with the Sino-French Research Center in Information and Communication, Nantes 44300, France (e-mail: yide.wang@univ-nantes.fr).

Digital Object Identifier 10.1109/TVT.2018.2890546

I. Introduction

With the rapid development of mobile communication, spectrum resources are increasingly scarce, future wireless communication technology needs to further increase system capacity and spectrum efficiency. In recent years, non-orthogonal multiple access (NOMA) has attracted substantial attention due to its high spectral efficiency, and it has become a promising candidate for 5G network [1]–[4]. Unlike orthogonal multiple access (OMA), NOMA allows multiple users in a cell to share the same radio resource in power domain, the user with better channel condition is assigned with less power, and more power is assigned to the user with worse channel condition.

To exploit the potential of NOMA, lots research works appeared. [5] combined NOMA with millimeter wave (mmWave) technology, which significantly reduced the interferences and enhanced sum rates. In [6], the concept of user paring was proposed for NOMA , and the impact of user paring on two different NOMA systems was investigated. Distinct from the usual investigation based on synchronous transmission, [7] proposed an application of asynchronous NOMA in downlink transmission by exploiting symbol asynchronism to reduce interferences. In [8], an optimal NOMA-enabled traffic off-loading was proposed in densely deployed small cells, where the spectrum was reused. In this case, NOMA was quite an efficient approach in traffic off-loading. A NOMA-based device-to-device (D2D) framework was proposed in [9], which allowed multiple users to share the same time-frequency resource. [10] applied NOMA into cooperative multicast cognitive radio networks by proposing a dynamic secondary user scheduling strategy according to channel state information (CSI).

For a smooth development from 4G to 5G, MIMO-NOMA is another hot topic in wireless communications. [11]–[13] show that NOMA could significantly improve the spectral efficiency of multiple-input multiple-output (MIMO) system. [14] applied the MIMO technique in NOMA system, the inter-cluster interference was eliminated by a new detection matrix, and the outage probability was significantly reduced, compared with the conventional OMA. [15] gave a capacity comparison between MIMO-NOMA and MIMO-OMA for a cluster including multiple users. In [16], a multi-cell MIMO-NOMA network was investigated, the proposed coordinated beamforming technique aimed to reduce the cell-edge users' interference which was a remarkable problem in multi-cell network. Layered transmission in MIMO-NOMA systems was investigated in [17], and an optimal power allocation approach was proposed. A novel

MIMO-NOMA scheme was proposed in [18], which was practical in Internet of Things (IoT).

Based on the investigation of the state of the art, the authors note that it is important to control intra-cluster interference and inter-cluster interference to fully exploit the potential of MIMO-NOMA. However, the existing work is mainly focused on inter-cluster interference elimination, and the intra-cluster interference control (rather than elimination) by proper power allocation. For the intra-cluster interference canceling, it remains a problem to be addressed. Motivated by this observation, this paper proposes a widely linear processing (WLP) framework for MIMO-NOMA, which is capable of eliminating both interferences mentioned above. Actually, WLP is a promising technique for interference control. Because WLP can expand the "observation space" by considering both the observation data and their complex conjugates, so that the statistical information contained in both the covariance matrix and conjugate covariance matrix of the received signal can be fully exploited. That is why WLP has been attracting increasing attention in beamforming [19]–[22], direction-of-arrival (DoA) finding [23], wireless communications [24]–[26], [28]–[34], parameter estimation [35], [36], speech enhancement [37], image processing [38] etc. However, to the best of the authors' knowledge, few research considers WLP in NOMA. This paper investigates the WLP in MIMO-NOMA system. The main contributions of this paper are summarized as follows:

Firstly, a widely linear MIMO-NOMA (WL-MIMO-NOMA) model is proposed by assuming that real-valued downlink signals are transmitted by base station. In the proposed model, complex-valued power allocation coefficients are introduced to stagger the users' signals in phase. The advantage is that all the inter-cluster interference and at least half of the intra-cluster interference can be eliminated. This is totally different from the existing work, where the intra-cluster interference can not be removed due to the signal overlapping caused by real-valued power allocation coefficients. More specifically, WL-MIMO-NOMA can completely eliminate both interferences in user pairing case. In general case, WL-MIMO-NOMA can completely remove the inter-cluster interference and reduce by half the intra-cluster interference.

Secondly, performance analysis is carried out to derive the closed form expressions of the users' outage probabilities. The phases of the complex power allocation coefficients are optimized[1] in terms of user's outage probability. In the case of user pairing, the users' outage probabilities take their minimum values when the phase difference of the power coefficients of the two users is $k\pi + \frac{\pi}{2}$. It is proven that, with orthogonal phase difference, SIC is not necessary for the detection and separation of the paired users. Similar conclusion can be obtained for general case.

Finally, the proposed WL-MIMO-NOMA framework is extended to the general mixed case, i.e. the users with poor channel use real signals, and those with good channel adopt complex circular signals, which is quite common in adaptive modulation.[2] Simulation results shows that WL-MIMO-NOMA has superior outage performance to MIMO-NOMA both in real-signal case and mixed-signal case.

The rest of this paper is organized as follows: Section II describes the system model for MIMO-NOMA downlink transmission; Section III proposes a WL-MIMO-NOMA model with complex power coefficients. Section IV shows Performance analysis and system optimization. A further discussion on the mixed case of real/complex circular signals is given in Section V. Numerical simulation results are provided in Section VI. Section VII draws the conclusions.

Notation: Capital letters of boldface are used for denoting matrices. Lowercase letters of boldface denote the column vectors. $(\cdot)^H$ and $(\cdot)^T$ denote the operations of Hermitian (complex conjugate transpose) and transpose, respectively. The operation of complex conjugate is denoted by $(\cdot)^*$. $(\cdot)^\dagger$ represents the left pseudo inverse. $\|\cdot\|$ represents the Frobenius norm. j is used to represent the imaginary part of a complex number. $\mathbf{I}_M$ represents the identity matrix of dimension M. $\mathbf{0}$ represents a zero matrix or vector. $E[\cdot]$ denotes the statistical expectation.

II. System Model

We consider a MIMO-NOMA downlink communication scenario, in which the number of transmitting antennas is M, and the number of receiving antennas is N. We assume that $N \geq M$. This case is justified by the ultra-densely deployed small cells in 5G networks, where small-cell base stations will be used [40]. In this scenario, low-cost small-cell base station with same (or less) number of antennas as user handsets is likely to be used. Cloud radio access networks (C-RANs) can also be another example, where users are served by a small number of low cost remote radio heads (RRHs) in order to reduce the fronthaul overhead [14]. The signals vector transmitted by the BS is represented as:

$$\overline{\mathbf{s}} = \left[\, \overline{s}_1 \; \overline{s}_2 \; \cdots \; \overline{s}_M \,\right]^T \tag{1}$$

where $\overline{s}_m = \boldsymbol{\alpha}_m^H \mathbf{s}_m$ denotes the signals intended for the m-th cluster.

$$\mathbf{s}_m = \left[\, s_{m,1} \cdots s_{m,K} \,\right]^T, \boldsymbol{\alpha}_m = \left[\, \alpha_{m,1} \cdots \alpha_{m,K} \,\right]^T \tag{2}$$

where K is the number of users in the m-th cluster. $s_{m,k}$ denotes the signal intended for the k-th user in the m-th cluster, $k \in \{1, 2, \ldots, K\}$, $\alpha_{m,k}$ is the power allocation coefficient. Conventionally, $\alpha_{m,k}$ is a real positive number and meets $\alpha_{m,1}^2 + \alpha_{m,2}^2 + \cdots + \alpha_{m,K}^2 = 1$. In this model, the signal received by the k-th user in the i-th cluster is given by:

$$\mathbf{y}_{i,k} = \mathbf{H}_{i,k}\mathbf{P}\overline{\mathbf{s}} + \mathbf{n}_{i,k} \tag{3}$$

where $\mathbf{H}_{i,k}$ is a $N \times M$ Rayleigh fading channel from the base station to the k-th user in the i-th cluster, and $\mathbf{n}_{i,k}$ is a Gaussian noise vector with dimension of $N \times 1$. Similar to [14], we focus on the signal detection by setting $\mathbf{P} = \mathbf{I}_M$. In this work, we have the following assumptions:

[1]This paper discusses the impact of the phases of power coefficients on the system performance. The magnitude optimization of power coefficients is beyond the scope of this paper.

[2]We do not discuss the case of circular complex signals, because it is well known that WLP is equivalent to its traditional counterpart in this case [39].

1) $E[\overline{\mathbf{s}}\overline{\mathbf{s}}^H] = \mathbf{I}_M$ (each element has unit power);
2) $E[\mathbf{n}_{i,k}\mathbf{n}_{i,k}^H] = \sigma^2\mathbf{I}_N$;
3) $s_{m,k}$ is real except in Section V, where mixed signals are discussed;

III. WLP FOR MIMO-NOMA

A. WL-MIMO-NOMA With Complex Power Coefficients ($K = 2$)

In this section, a WL-MIMO-NOMA model is proposed by focusing on the user pairing case. Different from the existing work where the power allocation coefficients are real and positive, this paper introduces phase angles to the power coefficients within a cluster. This operation can produce a staggering angles between the users in each cluster. It is useful for the receiving end to decode the signals. Furthermore, the optimal phase difference can be derived through mathematical analysis in the following section. The complex power allocation coefficients can be expressed as follows:

$$\boldsymbol{\alpha}_m = \left[\, \alpha_{m,1}e^{-j\theta_1} \;\; \alpha_{m,2}e^{-j\theta_2} \,\right]^T \tag{4}$$

where $\alpha_{m,k}$ is the strength coefficient of power allocation, it takes the real positive value and $\alpha_{m,1}^2 + \alpha_{m,2}^2 = 1$; θ_1 and θ_2 are the introduced phase angles. We can express $\mathbf{y}_{i,k}$ in more details with $\mathbf{h}_{m,ik}$ denoting the m-th column of $\mathbf{H}_{i,k}$.

$$\mathbf{y}_{i,k} = [\mathbf{h}_{1,ik}, \mathbf{h}_{2,ik}, \ldots, \mathbf{h}_{M,ik}] \begin{bmatrix} \boldsymbol{\alpha}_1^H \mathbf{s}_1 \\ \boldsymbol{\alpha}_2^H \mathbf{s}_2 \\ \vdots \\ \boldsymbol{\alpha}_M^H \mathbf{s}_M \end{bmatrix} + \mathbf{n}_{i,k} \tag{5}$$

The above equation can be further expressed as:

$$\mathbf{y}_{i,k} = \left[\mathbf{h}_{1,ik}e^{j\theta_1}, \mathbf{h}_{1,ik}e^{j\theta_2}, \ldots, \mathbf{h}_{M,ik}e^{j\theta_1}, \mathbf{h}_{M,ik}e^{j\theta_2}\right] \widetilde{\mathbf{s}} + \mathbf{n}_{i,k} \tag{6}$$

where $\widetilde{\mathbf{s}} = [\, \alpha_{1,1}s_{1,1}, \alpha_{1,2}s_{1,2}, \ldots, \alpha_{M,1}s_{M,1}, \alpha_{M,2}s_{M,2} \,]^T$. An extended model is constructed as follows by jointly exploiting the received signal $\mathbf{y}_{i,k}$ and its conjugate version:

$$\widetilde{\mathbf{y}}_{i,k} = \begin{bmatrix} \mathbf{y}_{i,k} \\ \mathbf{y}_{i,k}^* \end{bmatrix} = \widetilde{\mathbf{H}}_{i,k}\widetilde{\mathbf{s}} + \widetilde{\mathbf{n}}_{i,k} \tag{7}$$

where

$$\widetilde{\mathbf{H}}_{i,k} = \begin{bmatrix} \mathbf{h}_{1,ik}e^{j\theta_1}, \mathbf{h}_{1,ik}e^{j\theta_2}, \ldots, \mathbf{h}_{M,ik}e^{j\theta_1}, \mathbf{h}_{M,ik}e^{j\theta_2} \\ \mathbf{h}_{1,ik}^*e^{-j\theta_1}, \mathbf{h}_{1,ik}^*e^{-j\theta_2}, \ldots, \mathbf{h}_{M,ik}^*e^{-j\theta_1}, \mathbf{h}_{M,ik}^*e^{-j\theta_2} \end{bmatrix} \tag{8}$$

One notes that the WL-MIMO-NOMA model is transformed into a classical MIMO model. The advantage is that both the inter-cluster interference and intra-cluster interference can be eliminated completely. Without loss of generality, we focus on the first cluster. In the first cluster, the signal detection for the two users can be performed by adopting zero-forcing (ZF) method. Denote $\mathbf{W}_{1,1} = \widetilde{\mathbf{H}}_{1,1}^{\dagger}$ and $\mathbf{W}_{1,2} = \widetilde{\mathbf{H}}_{1,2}^{\dagger}$ respectively. Let $\mathbf{w}_{1,11}^H$ be the first row of $\mathbf{W}_{1,1}$, and $\mathbf{w}_{2,12}^H$ the second row of $\mathbf{W}_{1,2}$, the two users are detected respectively as:

$$\mathbf{w}_{1,11}^H\widetilde{\mathbf{y}}_{1,1} = \alpha_{1,1}s_{1,1} + \mathbf{w}_{1,11}^H\widetilde{\mathbf{n}}_{1,1} \tag{9}$$

$$\mathbf{w}_{2,12}^H\widetilde{\mathbf{y}}_{1,2} = \alpha_{1,2}s_{1,2} + \mathbf{w}_{2,12}^H\widetilde{\mathbf{n}}_{1,2} \tag{10}$$

where $\|\mathbf{w}_{1,11}\|^2$ and $\|\mathbf{w}_{2,12}\|^2$ stand for the noise amplification coefficients of ZF on the first and the second users, respectively. For ZF, the user's channel gain is fixed to one, so the noise amplification coefficient determines the user's channel condition. Without loss of generality, we assume that:[3]

$$\|\mathbf{w}_{1,11}\|^2 \leq \|\mathbf{w}_{2,12}\|^2 \tag{11}$$

which means that the first user's detection vector brings less noise gain, i.e. the first user's channel condition is better than the second user. According to the NOMA power allocation strategy, the power allocation coefficients should be ordered as follows:

$$\alpha_{1,1} < \alpha_{1,2} \tag{12}$$

Based on the signal model described above, the second user in the first cluster directly demodulates its own information with the following signal to noise ratio (SNR):

$$SNR_{1,2} = \frac{\alpha_{1,2}^2}{\|\mathbf{w}_{2,12}\|^2 \frac{1}{\rho}} \tag{13}$$

where $\rho = \frac{P}{\sigma^2}$. For the first user, the detection performance can be improved by SIC, namely it firstly detects the second user's signal as follows:

$$\mathbf{w}_{2,11}^H\widetilde{\mathbf{y}}_{1,1} = \alpha_{1,2}s_{1,2} + \mathbf{w}_{2,11}^H\widetilde{\mathbf{n}}_{1,1} \tag{14}$$

where $\mathbf{w}_{2,11}^H$ denotes the second row of $\mathbf{W}_{1,1}$. Following the representation style of [14], let $SNR_{1,1}^2$ be the SNR of the second user at the first user, $SNR_{1,1}^2$ is given by:

$$SNR_{1,1}^2 = \frac{\alpha_{1,2}^2}{\|\mathbf{w}_{2,11}\|^2 \frac{1}{\rho}} \tag{15}$$

Once the second user's message is decoded, the received signal is updated as $\overline{\mathbf{y}}_{1,1}$ by removing the second user's signal. Then the first user detects its own signal, the detection model in (9) is rewritten as:

$$\overline{\mathbf{v}}_{1,11}^H\overline{\mathbf{y}}_{1,1} = \alpha_{1,1}s_{1,1} + \overline{\mathbf{v}}_{1,11}^H\widetilde{\mathbf{n}}_{1,1} \tag{16}$$

For simplicity, denote $\widetilde{\mathbf{H}}_{1,1} = [\, \widetilde{\mathbf{h}}_{1,11} \;\; \widetilde{\mathbf{h}}_{2,11} \;\; \overline{\overline{\mathbf{H}}}_{1,1} \,]$, where $\overline{\mathbf{v}}_{1,11}^H$ is given by the first row of $\overline{\mathbf{H}}_{1,1}^{\dagger}$, and $\overline{\mathbf{H}}_{1,1}$ is obtained by removing the second column from $\widetilde{\mathbf{H}}_{1,1}$, namely $\overline{\mathbf{H}}_{1,1} = [\, \widetilde{\mathbf{h}}_{1,11} \;\; \overline{\overline{\mathbf{H}}}_{1,1} \,]$. The SNR in the detection of the first user's signal is then given by:

$$SNR_{1,1} = \frac{\alpha_{1,1}^2}{\|\overline{\mathbf{v}}_{1,11}\|^2 \frac{1}{\rho}} \tag{17}$$

If the fixed power allocation strategy is adopted, the SNR of the two users are determined by the noise amplification effect of

[3]The base station needn't to know the users' channel matrices, it needs only the noise amplification coefficient for each user, which reflects the user's channel condition, this brings a much less demanding requirement than acquiring the global channel condition.

ZF detection vector $\overline{\mathbf{v}}_{1,11}$ and $\mathbf{w}_{2,12}$, which can be optimized by choosing the appropriate phase angles θ_1 and θ_2.

B. Extension to K-User Clusters ($K > 2$)

The above discussion is limited to the case where each cluster has only 2 users (we call 2-user cluster). In this section, the discussion is extended to the case where each cluster has K ($K > 2$) users (we call K-user cluster). Different from 2-user cluster, widely linear processing can not completely separate K users in a cluster, because the available degrees of freedom for a complex-valued power coefficient are only 2. To deal with K users in a cluster, the following partial separation strategy are proposed:

1) The m-th cluster is divided into two subsets, denoted by $\Theta_{m,1}$ and $\Theta_{m,2}$, respectively. Assume that $|\Theta_{m,1}| = K_1$ and $|\Theta_{m,2}| = K_2$, where $K_1 + K_2 = K$. Without loss of generality, $\Theta_{m,1}$ and $\Theta_{m,2}$ can be represented as follows:

$$\Theta_{m,1} = \{s_{m,1}, s_{m,2}, \ldots, s_{m,K_1}\}$$
$$\Theta_{m,2} = \{s_{m,K_1+1}, s_{m,K_1+2}, \ldots, s_{m,K}\}$$

2) The users in subsets $\Theta_{m,1}$ and $\Theta_{m,2}$ use the power coefficient vectors $\boldsymbol{\alpha}_m$, $\boldsymbol{\beta}_m$, respectively. $\boldsymbol{\alpha}_m$, $\boldsymbol{\beta}_m$ are given by

$$\boldsymbol{\alpha}_m = \left[\, \alpha_{m,1} e^{-j\Theta_1} \;\; \alpha_{m,2} e^{-j\theta_1} \cdots \alpha_{m,K_1} e^{-j\theta_1} \,\right]^T$$
$$\boldsymbol{\beta}_m = \left[\, \beta_{m,K_1+1} e^{-j\theta_2} \;\; \beta_{m,K_1+2} e^{-j\theta_2} \cdots \beta_{m,K} e^{-j\theta_2} \,\right]^T$$

where $\alpha_{m,1}^2 + \cdots + \alpha_{m,K_1}^2 + \beta_{m,K_1+1}^2 + \cdots + \beta_{m,K}^2 = 1$. One notes that, complex-valued power coefficients in $\Theta_{m,1}$ have the identical phase θ_1, and those in $\Theta_{m,2}$ share another phase θ_2. Recalling (1), the signal intended for the m-th cluster is given by:

$$\overline{s}_m = \underbrace{\left(\sum\nolimits_{k=1}^{K_1} \alpha_{m,k} s_{m,k}\right) e^{j\theta_1}}_{\Theta_{m,1}} + \underbrace{\left(\sum\nolimits_{k=K_1+1}^{K} \beta_{m,k} s_{m,k}\right) e^{j\theta_2}}_{\Theta_{m,2}} \tag{18}$$

The above signal model shows that the users in $\Theta_{m,1}$ are overlapped in phase, and so are the users in $\Theta_{m,2}$. However, $\Theta_{m,1}$ and $\Theta_{m,2}$ are staggered by θ_1 and θ_2. This is quite similar to the user pairing in the upper section if we consider $\Theta_{m,1}$ and $\Theta_{m,2}$ as two 'big users' in a cluster.

It is proven by Lemma 4 in Section V that the system reaches the optimal performance if $\theta_1 - \theta_2 = k\pi + \pi/2$. Therefore, we consider only the scenario where θ_1 and θ_2 are orthogonal. We should emphasize that the technique of dividing K-user cluster into subsets $\Theta_{m,1}$ and $\Theta_{m,2}$ is beyond the scope of this paper. Following Section III, we can get a general system model as follows:

$$\widetilde{\mathbf{y}}_{i,k} = \widetilde{\mathbf{H}}_{i,k} \widetilde{\mathbf{s}} + \widetilde{\mathbf{n}}_{i,k}$$

The $2M \times 1$ vector $\widetilde{\mathbf{s}}$ is given by:

$$\widetilde{\mathbf{s}} = \begin{bmatrix} \alpha_{1,1} s_{1,1} + \cdots + \alpha_{1,K_1} s_{1,K_1} \\ \beta_{1,K_1+1} s_{1,K_1+1} + \cdots + \beta_{1,K} s_{1,K} \\ \vdots \\ \alpha_{M,1} s_{M,1} + \cdots + \alpha_{M,K_1} s_{M,K_1} \\ \beta_{M,K_1+1} s_{M,K_1+1} + \cdots + \beta_{M,K} s_{M,K} \end{bmatrix} \tag{19}$$

Without loss of generality, we focus on the users in the first cluster. Denote $\mathbf{W}_{1,k} = \widetilde{\mathbf{H}}_{1,k}^{\dagger}$ the ZF matrix at the k-th user. Let $\mathbf{w}_{1,k}$ be the ZF detection vector of the k-th user. For $k \in \{1, 2, \ldots K_1\}$, the k-th user is in subset $\Theta_{1,1}$, the detection vector $\mathbf{w}_{1,k}$ is the first row of $\mathbf{W}_{1,k}$; and for $k \in \{K_1 + 1, \ldots K\}$, the k-th user belongs to $\Theta_{1,2}$, and $\mathbf{w}_{1,k}$ is the second row of $\mathbf{W}_{1,k}$. So the k-th user is detected as:

$$\mathbf{w}_{1,k}^H \widetilde{\mathbf{y}}_{1,k} = \sum_{n=1}^{K_1} \alpha_{1,n} s_{1,n} + \mathbf{w}_{1,k}^H \widetilde{\mathbf{n}}_{1,k}, k \in \{1, \ldots K_1\} \tag{20}$$

$$\mathbf{w}_{1,k}^H \widetilde{\mathbf{y}}_{1,k} = \sum_{n=K_1+1}^{K} \beta_{1,n} s_{1,n} + \mathbf{w}_{1,k}^H \widetilde{\mathbf{n}}_{1,k}, k \in \{K_1 + 1, \ldots K\} \tag{21}$$

According to the above analysis, $\|\mathbf{w}_{1,k}\|^2$ stands for the noise amplification coefficients of ZF, which reflects the user's channel condition. Without loss of generality, we assume that:

$$\|\mathbf{w}_{1,1}\|^2 \leq \|\mathbf{w}_{1,2}\|^2 \leq \cdots \leq \|\mathbf{w}_{1,K_1}\|^2, \text{for } \Theta_{1,1} \tag{22}$$

$$\|\mathbf{w}_{1,K_1+1}\|^2 \leq \|\mathbf{w}_{1,K_1+2}\|^2 \cdots \leq \|\mathbf{w}_{1,K}\|^2, \text{for } \Theta_{1,2} \tag{23}$$

Although the channel conditions are ordered within $\Theta_{1,1}$ (or $\Theta_{1,2}$), one should note that, in general, there is no specific magnitude order for any pair across the subsets, i.e. $\|\mathbf{w}_{1,k_1}\|^2$ in $\Theta_{1,1}$ and $\|\mathbf{w}_{1,K_1+k_2}\|^2$ in $\Theta_{1,2}$, with $1 \leq k_1 \leq K_1$, $1 \leq k_2 \leq K_2$. $\|\mathbf{w}_{1,k_1}\|^2$ can be bigger or smaller than $\|\mathbf{w}_{1,K_1+k_2}\|^2$. Following the principle of NOMA power allocation, the power allocation coefficients should be ordered as follows:

$$\alpha_{1,1} \leq \cdots \leq \alpha_{1,K_1}, \beta_{1,K_1+1} \leq \cdots \leq \beta_{1,K} \tag{24}$$

We consider the detection in subset $\Theta_{1,1}$, as shown by (20). The K_1-th user in subset $\Theta_{1,1}$ can directly decode its messages, it will be detected with the following signal-to-interference-plus-noise ratio (SINR):

$$SINR_{1,K_1} = \frac{\alpha_{1,K_1}^2}{\sum_{m=1}^{K_1-1} \alpha_{1,m}^2 + \|\mathbf{w}_{1,K_1}\|^2 \frac{1}{\rho}} \tag{25}$$

The k-th user, $1 \leq k < K_1$, needs to use SIC technique to decode the signal of the j-th user, $1 + k \leq j \leq K_1$, and then remove all these users' signal before detecting its own. So the SINR for the j-th user at the k-th user is given by:

$$SINR_{1,k}^j = \frac{\alpha_{1,j}^2}{\sum_{m=1}^{j-1} \alpha_{1,m}^2 + \|\mathbf{w}_{1,k}\|^2 \frac{1}{\rho}} \tag{26}$$

The first user in $\Theta_{1,1}$, needs to decode all the other users' messages (from the K_1-th user to the second user) and remove their

contribution, the first user is then detected with the following SNR:

$$SNR_{1,1} = \frac{\alpha_{1,1}^2}{\|\mathbf{w}_{1,1}\|^2 \frac{1}{\rho}} \tag{27}$$

Next, we discuss subset $\Theta_{1,2}$ by considering (21). Similar to the detection in subset $\Theta_{1,1}$, the K-th user in subset $\Theta_{1,2}$ is detected with the following SINR:

$$SINR_{1,K} = \frac{\beta_{1,K}^2}{\sum_{m=K_1+1}^{K-1} \beta_{1,m}^2 + \|\mathbf{w}_{1,K}\|^2 \frac{1}{\rho}} \tag{28}$$

The k-th user, $K_1 + 1 \leq k < K$, needs to perform SIC to decode the signal of the j-th user, $1 + k \leq j \leq K$, and then remove all these users' signals before detecting its own. So the SINR for the j-th user at the k-th user is:

$$SINR_{1,k}^{j} = \frac{\beta_{1,j}^2}{\sum_{m=K_1+1}^{j-1} \beta_{1,m}^2 + \|\mathbf{w}_{1,k}\|^2 \frac{1}{\rho}} \tag{29}$$

For the $K_1 + 1$-th user in $\Theta_{1,2}$, if other users can be detected successfully, its SNR is given by:

$$SNR_{1,K_1+1} = \frac{\beta_{1,K_1+1}^2}{\|\mathbf{w}_{1,K_1+1}\|^2 \frac{1}{\rho}} \tag{30}$$

According to [14], the sum capacity of WL-MIMO-NOMA is given by:

$$C \triangleq \sum_{i=1}^{M} \left[\sum_{k=1}^{K} \log_2 \left(1 + SINR_{i,k}\right) \right] \tag{31}$$

Following Section IV of [14], $E[C]$ for the case $N = M$ is given by (32), where l_k means that the k-th user in the subset is ranked at the l_k-th position in the whole cluster, and $\varphi(l_n, \phi)$ is expressed as follows:

$$\begin{aligned} E[C] = M \Bigg\{ & \sum_{k=1}^{K_1} \varphi \left(l_k, \sum_{m=1}^{k} 2\rho\alpha_{1,m}^2 \right) \\ & + \sum_{k=K_1+1}^{K} \varphi \left(l_k, \sum_{m=K_1+1}^{k} 2\rho\beta_{1,m}^2 \right) \\ & - \sum_{k=2}^{K_1} \varphi \left(l_k, \sum_{m=1}^{k-1} 2\rho\alpha_{1,m}^2 \right) - \sum_{k=K_1+2}^{K} \varphi \left(l_k, \sum_{m=K_1+1}^{k-1} 2\rho\beta_{1,m}^2 \right) \Bigg\} \end{aligned} \tag{32}$$

$$\begin{aligned} & \varphi(l_n, \phi) = \\ & \frac{\phi}{\ln 2} \left(1 - \sum_{p=0}^{l_n - 1} \binom{l_n - 1}{p} \gamma_{l_n} (-1)^p \frac{1}{K - l_n + p + 1} \right) \\ & \int_0^{\infty} \frac{1}{1 + x\phi} dx + \frac{1}{\ln 2} \sum_{p=0}^{l_n - 1} \binom{l_n - 1}{p} \gamma_{l_n} (-1)^p \frac{1}{K - l_n + p + 1} \\ & \sum_{l=1}^{K - l_n + p + 1} \binom{K - l_n + p + 1}{l} (-1)^l e^{\frac{l}{\phi}} \mathrm{E}_i \left(-\frac{l}{\phi} \right) \end{aligned} \tag{33}$$

where $\gamma_{l_n} = \frac{K!}{(K - l_n)!(l_n - 1)!}$, $\mathrm{E}_i(\cdot)$ denotes the exponential integral function, $\mathrm{E}_i(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt$.

Remarks:

1) User grouping in a cluster is an interesting problem, like user clustering in NOMA. The system performance can be enhanced by optimizing the user grouping;
2) Even though user grouping optimization is not in the scope of this paper, we can adopt a simple way as follows: the users are ordered according to their channel conditions; the even-ordered user are gathered in one group, and the remaining users are in the other group. The following simulation shows that, even with this simple grouping method, WL-MIMO-NOMA can achieve better performance than MIMO-NOMA.

IV. Outage Performance Analysis and Optimization

In this section the system performance is analyzed in terms of the outage probability of users, which is closely related to the SNR (or SINR). Before analyzing the general case of K-user cluster, the user pairing case is considered firstly.

A. User Pairing Case

Based on the SNR expression derived in Section III, the outage probability of the first user is given by:

$$\begin{aligned} P_{1,1}^{o} &= P\left(SNR_{1,1} < \zeta_{1,1} \text{ or } SNR_{1,1}^2 < \zeta_{1,2}\right) \\ &= 1 - P\left(\frac{\alpha_{1,1}^2}{\|\overline{\mathbf{v}}_{1,11}\|^2 \frac{1}{\rho}} > \zeta_{1,1}, \frac{\alpha_{1,2}^2}{\|\mathbf{w}_{2,11}\|^2 \frac{1}{\rho}} > \zeta_{1,2} \right) \end{aligned} \tag{34}$$

For the second user, its outage probability is expressed as:

$$P_{1,2}^{\circ} = P\left(SNR_{1,2} < \zeta_{1,2}\right) = P\left(\frac{\alpha_{1,2}^2}{\|\mathbf{w}_{2,12}\|^2 \frac{1}{\rho}} < \zeta_{1,2} \right) \tag{35}$$

where $\zeta_{1,k} = 2^{R_{1,k}} - 1$, with $R_{1,k}$ denoting the targeted data rate of the k-th user in the first cluster. One notes that, for the given power transmission strategy and fixed targeted data rate, the outage probability is determined by the noise amplification effect of the detection vectors, i.e. $\|\overline{\mathbf{v}}_{1,11}\|^2$, $\|\mathbf{w}_{2,11}\|^2$ and $\|\mathbf{w}_{2,12}\|^2$. To minimize the outage probability, we have to minimize the noise amplification effect. However, it's difficult to directly analyze and optimize the problem. Alternatively, we turn to find the relationship between the detection vectors. The analysis begins with the noise amplification effect of $\overline{\mathbf{v}}_{1,11}$.

Lemma 1: Let $\widetilde{\mathbf{H}}_{1,k} = [\widetilde{\mathbf{h}}_{1,1k}\ \widetilde{\mathbf{h}}_{2,1k}\ \overline{\overline{\mathbf{H}}}_{1,k}]$, and $\underline{\mathbf{H}}_{1,k}$ (or $\overline{\overline{\mathbf{H}}}_{1,k}$) be matrix $\widetilde{\mathbf{H}}_{1,k}$ with its first (or second) column removed, therefore $\underline{\mathbf{H}}_{1,k} = [\widetilde{\mathbf{h}}_{2,1k}\ \overline{\overline{\mathbf{H}}}_{1,k}]$, $\overline{\mathbf{H}}_{1,k} = [\widetilde{\mathbf{h}}_{1,1k}\ \overline{\overline{\mathbf{H}}}_{1,k}]$. $\overline{\mathbf{H}}_{1,k}^{\dagger}$ and $\underline{\mathbf{H}}_{1,k}^{\dagger}$ have equal noise amplification impact in their first rows. Namely:

$$\|\overline{\mathbf{v}}_{1,1k}\|^2 = \|\underline{\mathbf{v}}_{1,1k}\|^2 \tag{36}$$

where $\overline{\mathbf{v}}_{1,1k}^H$ (or $\underline{\mathbf{v}}_{1,1k}^H$) denotes the first row of $\overline{\mathbf{H}}_{1,k}^{\dagger}$ (or $\underline{\mathbf{H}}_{1,k}^{\dagger}$).

Proof: Please refer to Appendix A.

Lemma 1 means that the correlation between $\widetilde{\mathbf{h}}_{1,1k}$ and $\overline{\overline{\mathbf{H}}}_{1,k}$ is the same as that between $\widetilde{\mathbf{h}}_{2,1k}$ and $\overline{\overline{\mathbf{H}}}_{1,k}$ in terms of noise amplification. Lemma 1 is obtained based on the absence of either $\widetilde{\mathbf{h}}_{1,1k}$ or $\widetilde{\mathbf{h}}_{2,1k}$, i.e. the absence of either user in a cluster, which can be achieved by SIC technique. Actually, the conclusion in Lemma 1 still holds even if two users coexist in a cluster, i.e. the state before SIC, as illustrated in Lemma 2.

Lemma 2: Let $\mathbf{W}_{1,k} = \widetilde{\mathbf{H}}_{1,k}^{+}$, and $\mathbf{w}_{i,1k}^{H}$ denote the i-th row of $\mathbf{W}_{1,k}$, we have:

$$\|\mathbf{w}_{1,1k}\|^2 = \|\mathbf{w}_{2,1k}\|^2 \tag{37}$$

Proof: Please refer to Appendix B.

Lemma 2 implies that $\widetilde{\mathbf{H}}_{1,k}^{+}$ has equal noise amplification effect on its first two rows, i.e. the two users in a cluster suffer from the same noise amplification effect. Based on Lemma 1 and Lemma 2, we consider the noise amplification reduction brought by SIC. The ratio $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2}$ is employed to denote the improvement, which is given by Lemma 3.

Lemma 3: The ratio $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2}$ depends only on θ_1 and θ_2, and it is given by:

$$\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2} = \frac{\|\mathbf{w}_{2,1k}\|^2}{\|\underline{\mathbf{v}}_{1,1k}\|^2} = \frac{1}{1-\cos^2(\theta_1-\theta_2)} \tag{38}$$

Proof: Please refer to Appendix C.

Base on Lemma 3, the noise amplification is minimized by Lemma 4.

Lemma 4: $\|\overline{\mathbf{v}}_{1,1k}\|^2$ and $\|\underline{\mathbf{v}}_{1,1k}\|^2$ are independent of θ_1 and θ_2. $\|\mathbf{w}_{1,1k}\|^2$ and $\|\mathbf{w}_{2,1k}\|^2$ take the minimum values when θ_1 and θ_2 are orthogonal; which leads to the lowest outage probabilities. In this case, $\|\mathbf{w}_{1,1k}\|^2$ and $\|\mathbf{w}_{2,1k}\|^2$ are given by:

$$\|\mathbf{w}_{1,1k}\|^2 = \|\mathbf{w}_{2,1k}\|^2 = \|\overline{\mathbf{v}}_{1,1k}\|^2 = \|\underline{\mathbf{v}}_{1,1k}\|^2 \tag{39}$$

Proof: Please refer to Appendix D.

Lemma 4 implies that: if θ_1 and θ_2 are orthogonal, the first user's signal can be directly detected by ZF, and SIC is not necessary. In this case, we just use $\mathbf{w}_{1,11}$ to detect the first user's signal, as shown in (9). The outage probability experienced by the first user is given by:

$$P_{1,1}^{\circ} = P\left(SNR_{1,1} < \zeta_{1,1}\right) = P\left(\frac{\alpha_{1,1}^2}{\|\mathbf{w}_{1,11}\|^2 \frac{1}{\rho}} < \zeta_{1,1}\right) \tag{40}$$

Finally, with the above analysis, Theorem 1 is derived to provide an expression of the outage probability.

Theorem 1: In the WL-MIMO-NOMA systems with two users in each cluster, the outage probability of the k-th user in the i-th cluster is given by:

$$P_{i,k}^{\circ} = \sum_{m=0}^{k-1} \binom{k-1}{m} \frac{2!(-1)^m \left[\frac{\Gamma(N-M+1,\varepsilon_{i,k})}{(N-M)!}\right]^{2-k+m+1}}{(2-k)!(k-1)!(2-k+m+1)} \tag{41}$$

- when Θ_1 and θ_2 are orthogonal, $\varepsilon_{i,k} = \frac{\zeta_{i,k}}{2\rho\alpha_{i,k}^2}$.
- when θ_1 and θ_2 are nonorthogonal

$$\varepsilon_{i,1} = \max\left\{\frac{\zeta_{i,1}}{2\rho\alpha_{i,1}^2}, \frac{\zeta_{i,2}}{2\rho\alpha_{i,2}^2\left(1-\cos^2\left(\theta_1-\theta_2\right)\right)}\right\}$$

$$\varepsilon_{i,2} = \frac{\zeta_{i,2}}{2\rho\alpha_{i,2}^2(1-\cos^2(\theta_1-\theta_2))}$$

where $\zeta_{i,k} = 2^{R_{i,k}} - 1$, with $R_{i,k}$ the targeted rate of the k-th user in the i-th cluster, and $\Gamma(\cdot)$ denotes the incomplete gamma function.

Proof: Please refer to Appendix E.

Similar to Lemma 4, we can also conclude, from Theorem 1, that the outage probabilities of the users take the minimum when θ_1 and θ_2 are orthogonal. The optimization problem can be considered as follows:

The minimization of $P_{i,k}^{\circ}$ requires the minimization of $\varepsilon_{i,k}$. To balance $P_{i,1}^{\circ}$ and $P_{i,2}^{\circ}$, a multi-objective problem can be formulated as follows by using the weight $\eta_i > 0$ $(i = 1, 2)$, $\eta_1 + \eta_2 = 1$. The problem can be expressed as follows:

$$\min_{\Delta\theta} : z = \eta_1\varepsilon_{i,1} + \eta_2\varepsilon_{i,2} \quad \text{s.t. } \Delta\theta = \theta_1 - \theta_2 > 0 \tag{42}$$

Due to the expression of $\varepsilon_{i,1}$, we have either $z = \frac{\eta_1\zeta_{i,1}}{2\rho\alpha_{i,1}^2} + \frac{\eta_2\zeta_{i,2}}{2\rho\alpha_{i,2}^2(1-\cos^2\Delta\theta)}$ or $z = \frac{\zeta_{i,2}}{2\rho\alpha_{i,2}^2(1-\cos^2\Delta\theta)}$. Obviously, both cases turn out that we can minimize the outage probabilities of two users when the phases are orthogonal, e.g. $\Delta\theta = \pi/2$.

B. General Cases of K-User Cluster

The case of K-user cluster is more complicated than user pairing case. We can perform an analysis similar to that in [14]. In this case, the users in each cluster are divided into two subsets. For detecting a user in a subset, the interferences from the other $2M-1$ subsets are completely removed by considering 2 subsets in a cluster as 2 'big users', and then SIC is employed to detect the user by considering the other users within the same subset as interferences. The analysis can be summarized as Theorem 2.

Theorem 2: In the WL-MIMO-NOMA systems with K users in each cluster, the outage probability of the k-th user in the i-th cluster is given by:

$$P_{i,k}^{\circ} = \sum_{m=0}^{l_k-1} \binom{l_k-1}{m} \frac{K!(-1)^m \left[\frac{\Gamma(N-M+1,\varepsilon_{i,k})}{(N-M)!}\right]^{K-l_k+m+1}}{(K-l_k)!(l_k-1)!(K-l_k+m+1)} \tag{43}$$

where l_k means that the k-th user in the subset is ranked at the l_k-th position in the whole cluster, it reflects the overall channel condition of the user within the cluster, smaller value of l_k represents better channel condition, detailed explanation is shown in proof.

Case I: The users are in subset $\Theta_{i,1}$

- if $k = 1$

$$\varepsilon_{i,1} = \max\left\{\frac{\zeta_{i,K_1}}{2\rho(\alpha_{i,K_1}^2 - \zeta_{i,K_1}\sum_{m=1}^{K_1-1}\alpha_{i,m}^2)}, \ldots, \frac{\zeta_{i,2}}{2\rho(\alpha_{i,2}^2 - \zeta_{i,2}\alpha_{i,1}^2)}, \frac{\zeta_{i,1}}{2\rho\alpha_{i,1}^2}\right\}$$

- if $2 \leq k < K_1$

$$\varepsilon_{i,k} = \max \left\{ \frac{\zeta_{i,K_1}}{2\rho(\alpha_{i,K_1}^2 - \zeta_{i,K_1} \sum_{m=1}^{K_1-1} \alpha_{i,m}^2)}, \ldots, \frac{\zeta_{i,k}}{2\rho(\alpha_{i,k}^2 - \zeta_{i,k} \sum_{m=1}^{k-1} \alpha_{i,m}^2)} \right\}$$

- if $k = K_1$

$$\varepsilon_{i,K_1} = \frac{\zeta_{i,K_1}}{2\rho(\alpha_{i,K_1}^2 - \zeta_{i,K_1} \sum_{m=1}^{K_1-1} \alpha_{i,m}^2)}$$

Case II: The users are in subset $\Theta_{i,2}$

- if $k = K_1 + 1$

$$\varepsilon_{i,K_1+1} = \max \left\{ \frac{\zeta_{i,K}}{2\rho(\beta_{i,K}^2 - \zeta_{i,K} \sum_{m=K_1+1}^{K-1} \beta_{i,m}^2)}, \ldots, \frac{\zeta_{i,K_1+2}}{2\rho(\beta_{i,K_1+2}^2 - \zeta_{i,K_1+2}\beta_{i,K_1+1}^2)}, \frac{\zeta_{i,K_1+1}}{2\rho\beta_{i,K_1+1}^2} \right\}$$

- if $K_1 + 2 \leq k < K$

$$\varepsilon_{i,k} = \max \left\{ \frac{\zeta_{i,K}}{2\rho(\beta_{i,K}^2 - \zeta_{i,K} \sum_{m=K_1+1}^{K-1} \beta_{i,m}^2)}, \ldots, \frac{\zeta_{i,k}}{2\rho(\beta_{i,k}^2 - \zeta_{i,k} \sum_{m=K_1+1}^{k-1} \beta_{i,m}^2)} \right\}$$

- if $k = K$

$$\varepsilon_{i,K} = \frac{\zeta_{i,K}}{2\rho(\beta_{i,K}^2 - \zeta_{i,K} \sum_{m=K_1+1}^{K-1} \beta_{i,m}^2)}$$

It is important to clarify that Theorem 2 in this paper is similar to the Theorem 1 in [14], the proof of Theorem 2 in this paper can be readily obtained from the proof in [14]. For completeness of the work, we provide the proof in Appendix F.

V. Further Discussion: Coexistence of Real Signals and Complex Circular Signals

In the above sections, the discussion on WL-MIMO-NOMA is for the real signals. In practice, users are likely to suffer from various channel fading. To deal with the problem, communication system usually adopts an adaptive strategy by using real signals for users with poor channels and complex circular signals for those experiencing good channels. Therefore, the coexistence of real and complex circular signals is quite usual. In this section, we extend the model to the mixed case where real and complex circular signals coexist in a cluster.

We consider (18). $\Theta_{m,1}$ and $\Theta_{m,2}$ are supposed to be the sets for the complex circular and real signals, respectively. An extended mixed signal model can be expressed as follows:

$$\widetilde{\mathbf{y}}_{i,k} = \begin{bmatrix} \mathbf{y}_{i,k} \\ \mathbf{y}_{i.k}^* \end{bmatrix} = \begin{bmatrix} e^{j\theta}\mathbf{H}_{i,k}\mathbf{s}_c \\ e^{-j\theta}\mathbf{H}_{i.k}^* s_c^* \end{bmatrix} + \begin{bmatrix} \mathbf{H}_{i,k} \\ \mathbf{H}_{i.k}^* \end{bmatrix} \mathbf{s}_n + \begin{bmatrix} \mathbf{n}_{i,k} \\ \mathbf{n}_{i.k}^* \end{bmatrix} \tag{44}$$

where θ is the difference of two phase angles (θ_1 and θ_2 for $\Theta_{m,1}$ and $\Theta_{m,2}$, respectively), which can simplify the expression. $\mathbf{s}_n$ and $\mathbf{s}_c$ denote the real signal and complex circular signals, respectively. They are expressed as follows:

$$\mathbf{s}_c = \begin{bmatrix} \alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K_1}s_{1,K_1} \\ \vdots \\ \alpha_{M,1}s_{M,1} + \cdots + \alpha_{M,K_1}s_{M,K_1} \end{bmatrix} \tag{45}$$

$$\mathbf{s}_n = \begin{bmatrix} \beta_{1,K_1+1}s_{1,K_1+1} + \cdots + \beta_{1,K}s_{1,K} \\ \vdots \\ \beta_{M,K_1+1}s_{M,K_1+1} + \cdots + \beta_{M,K}s_{M,K} \end{bmatrix} \tag{46}$$

Without loss of generality, we focus on the signals in the first cluster. Let $\mathbf{c}_{1,k}$ be the detection vector at the k-th user, so the detected signal is described as:

$$\widetilde{s}_{1,k} = \mathbf{c}_{1,k}^H \widetilde{\mathbf{y}}_{1,k} \tag{47}$$

The detection vector $\mathbf{c}_{1,k}$ is obtained by minimizing the following criterion:

$$\min_{\mathbf{c}_{1,k}} L_{1,k} = E\left[|\widetilde{s}_{1,k} - s_{1,k}|^2\right] \tag{48}$$

Let $\frac{\partial L_{1,k}}{\partial \mathbf{c}_{1,k}} = 0$, we can obtain:

$$\mathbf{c}_{1,k} = \left(E\left[s_{1,k}\widetilde{\mathbf{y}}_{1,k}^H\right]\left(E\left[\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^H\right]\right)^{-1}\right)^H$$

$$L_{1,k} = E\left[s_{1,k}s_{1,k}^H\right] - E\left[s_{1,k}\widetilde{\mathbf{y}}_{1,k}^H\right] \times \left(E\left[\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^H\right]\right)^{-1}\left(E\left[s_{1,k}\widetilde{\mathbf{y}}_{1,k}^H\right]\right)^H$$

The SINR of k-th user is given by:

$$SINR_{1,k} = \frac{|s_{1,k}|^2}{L_{1,k}} \tag{49}$$

We assume that the channel conditions of complex circular signals are better than those of real signals, and the channel condition in each subsets is ordered as (22) and (23). The K-th user in real signals subset $\Theta_{1,2}$ can directly decode its messages. The outage probability of the K-th user (real signal) is given by:

$$P_{1,K}^o = P\left(SINR_{1,K} < \zeta_{1,K}\right) = P\left(\frac{\beta_{1,K}^2}{L_{1,K}} < \zeta_{1,K}\right) \tag{50}$$

The k-th user, $k < K$, needs to use SIC technique to decode the signal of j-th user, $1 + k \leq j \leq K$, and then remove all these users' signal before detecting its own. The SIC process is similar to those in the above sections. And the outage probability experienced by the k-th user is shown as follows:

$$P_{1,k}^o = 1 - P\left(SINR_{1,k}^j > \zeta_{1,j}, j \in \{k, \ldots, K\}\right) \tag{51}$$

One should note that, for the case of circular complex signals, it is well known that WLP is equivalent to its traditional counterpart [39].

VI. Numerical Simulation Results

In this section, the performance of WL-MIMO-NOMA system is assessed by numerical simulation. The simulation includes five parts. The first part is devoted to examine the analytical results in Section IV. The second part is to compare

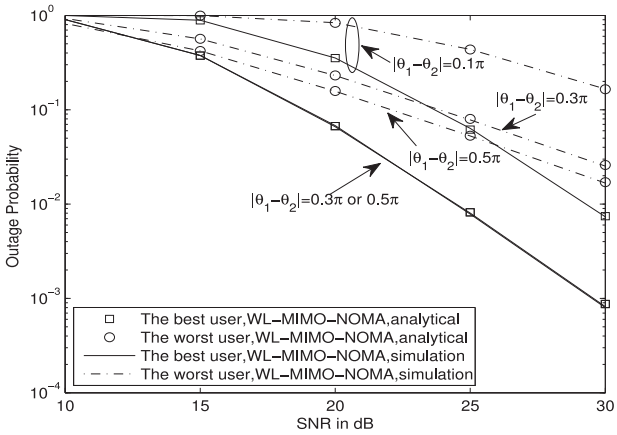

Fig. 1. WL-MIMO-NOMA with different $|\theta_1 - \theta_2|$, $N = M = 3$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 4$ BPCU, $R_2 = 3.8$ BPCU (BPCU means bits per channel use).

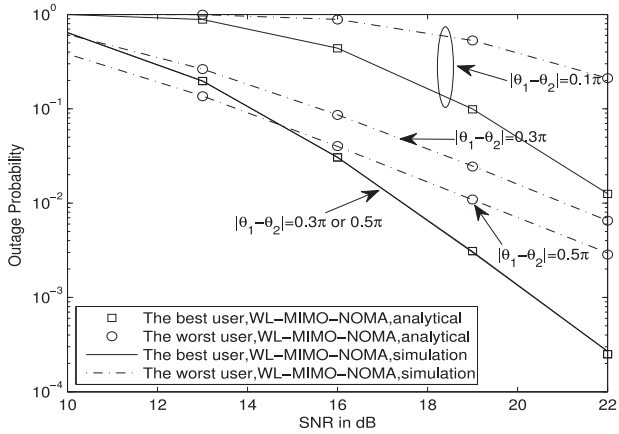

Fig. 2. WL-MIMO-NOMA with different $|\theta_1 - \theta_2|$, $N = 3$, $M = 2$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 4$ BPCU, $R_2 = 3.8$ BPCU.

the performance of WL-MIMO-NOMA with that of conventional MIMO-NOMA in terms of user pairing. The third part simulates the WL-MIMO-NOMA system with K users in each cluster. The fourth part compare WL-MIMO-NOMA with WL-MIMO-OMA, and the final part simulates the mixed case of real and complex circular signals.

A. Analytical Results and Monte-Carlo Simulation

Fig. 1 simulates a WL-MIMO-NOMA downlink communication system with fixed power allocation. In the simulation, the base station is equipped with 3 transmission antennas serving 3 clusters, each of which includes 2 users. Each user is equipped with 3 receiving antennas. In each cluster, the users have different data rate requirement according to the users' channel condition, higher rate means that the user has better channel condition. Fig. 1 shows the outage probability of the system under various phase differences in $|\theta_1 - \theta_2|$. One can observe that simulation results agree well with the analytical results in Theorem 1. Besides, the outage probabilities of the users decrease with the increasing phase difference from 0 to $\pi/2$, because larger phase difference means less noise amplification in eliminating intra-cluster interferences. In particular, the outage probabilities of two users reach their minimum values when θ_1 and θ_2 are orthogonal, namely $|\theta_1 - \theta_2| = k\pi + \frac{\pi}{2}$. One can also observe that, when $|\theta_1 - \theta_2| = \frac{\pi}{3}$ or $|\theta_1 - \theta_2| = \frac{\pi}{2}$, the two curves for the first user are exactly overlapped by each other. It's because the curves are determined by $\max\{\frac{\zeta_{i,1}}{2\rho\alpha_{i,1}^2}, \frac{\zeta_{i,2}}{2\rho\alpha_{i,2}^2(1-\cos^2(\theta_1-\theta_2))}\}$. When $|\theta_1 - \theta_2| = \frac{\pi}{3}$ or $|\theta_1 - \theta_2| = \frac{\pi}{2}$, the above maximization function returns the same value $\frac{\zeta_{i,1}}{2\rho\alpha_{i,1}^2}$, which is independent of $|\theta_1 - \theta_2|$. Similarly, the analytical results are also validated by Fig. 2, where the system has a different antenna configuration with $N = 3$, $M = 2$.

B. WL-MIMO-NOMA Versus MIMO-NOMA: $K = 2$

In this subsection, the performance of WL-MIMO-NOMA is compared with that of conventional MIMO-NOMA when

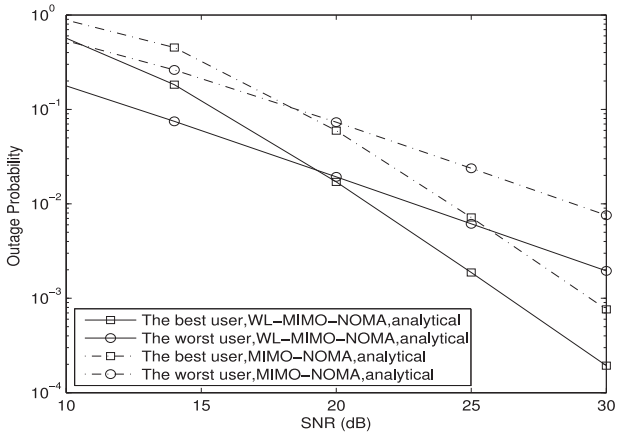

Fig. 3. WL-MIMO-NOMA v.s. MIMO-NOMA, $N = 3$, $M = 3$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 3$ BPCU, $R_2 = 1.3$ BPCU.

$K = 2$. Fig. 3 provides a comparison between the proposed WL-MIMO-NOMA system with real signals and the classical MIMO-NOMA system in [14]. The simulation is based on a system with $M = N = 3, K = 2$. One can note that the outage probabilities of the users in WL-MIMO-NOMA are lower than those in MIMO-NOMA. Fig. 4 further shows the superiority of WL-MIMO-NOMA to MIMO-NOMA by simulating a WL-MIMO-NOMA system whose antenna configuration is $N = 3$, $M = 2$. We observe that, in Fig. 3-4, the 'best user' sometimes has a worse performance than the 'worst user'. The reason for this is that two users have different targeted data rates, and different channel conditions, which leads to an equivalent SNR offset on the curves in the figures. Actually, the curve of 'best user' has faster decreasing rate (steeper slope) than the worst user, which determines that the former will surpass the latter in terms of outage performance.

C. WL-MIMO-NOMA Versus MIMO-NOMA: $K \geq 2$

Fig. 5 gives a simulation on a WL-MIMO-NOMA system with $M = N = 3, K = 4$. In this system, users in one

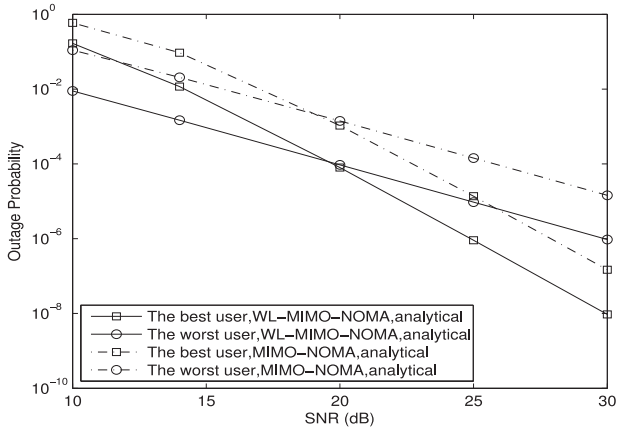

Fig. 4. WL-MIMO-NOMA v.s. MIMO-NOMA, $N = 3$, $M = 2$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 3$ BPCU, $R_2 = 1.3$ BPCU.

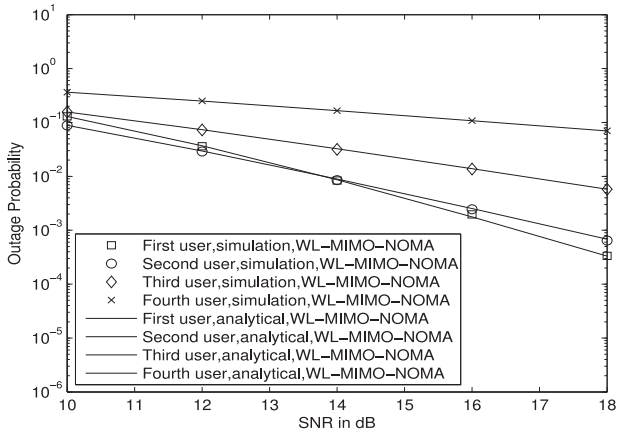

Fig. 5. WL-MIMO-NOMA, $N = M = 3$, $K = 4$, $\alpha_1^2 = \frac{1}{10}$, $\alpha_2^2 = \frac{1}{5}$, $\alpha_3^2 = \frac{3}{10}$, $\alpha_4^2 = \frac{2}{5}$. $R_1 = 1.5$ BPCU, $R_2 = 1.3$, $R_3 = 0.9$, $R_4 = 0.7$ BPCU.

cluster are divided into 2 subsets. For simplicity, the first user and the third user are grouped in $\Theta_{m,1}$, the second user and the fourth user are grouped in $\Theta_{m,2}$, which seems more equitable. Based on the simulation results, we can note the consistence of the analytical results with those of Monte-Carlo simulation. In addition, Fig. 6 compares WL-MIMO-NOMA with the conventional MIMO-NOMA in terms of outage probability, the performance of our proposed scheme is superior to the conventional MIMO-NOMA. Besides, Fig. 7 shows the sum capacity comparison between WL-MIMO-NOMA and MIMO-NOMA, we can observe that WL-MIMO-NOMA outperforms MIMO-NOMA in sum capacity, and the numerical results agree well with the analytical analysis.

D. WL-MIMO-NOMA Versus WL-MIMO-OMA

Fig. 8 compares WL-MIMO-NOMA and WL-MIMO-OMA when the real signals are transmitted by BS. In the simulation, WL-MIMO-NOMA is allocated with complex power coefficients, and the difference of two phase angles is $\pi/2$. This simulation is based on a system with $M = N = 3$ and $K = 2$. From

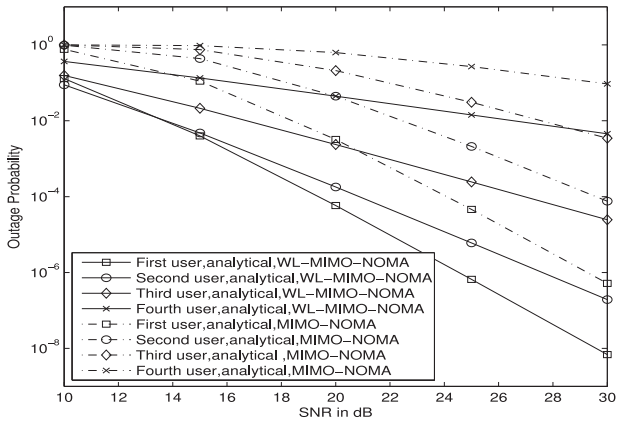

Fig. 6. WL-MIMO-NOMA v.s. MIMO-NOMA, $N = M = 3, K = 4, \alpha_1^2 = \frac{1}{10}, \alpha_2^2 = \frac{1}{5}, \alpha_3^2 = \frac{3}{10}, \alpha_4^2 = \frac{2}{5}$. $R_1 = 1.5$ BPCU, $R_2 = 1.3$, $R_3 = 0.9$, $R_4 = 0.7$ BPCU.

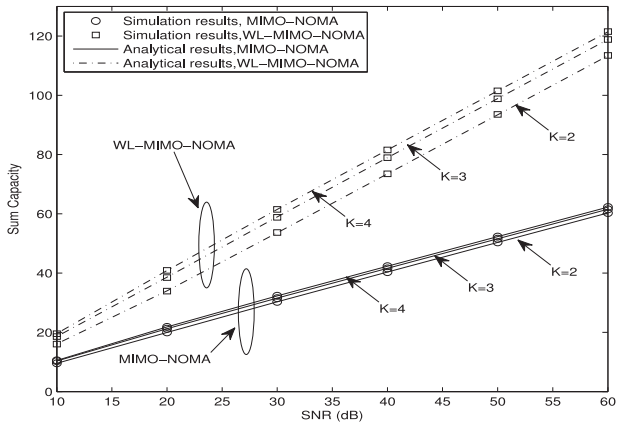

Fig. 7. Sum capacity comparison between WL-MIMO-NOMA and MIMO-NOMA, $N = 3$, $M = 3$.

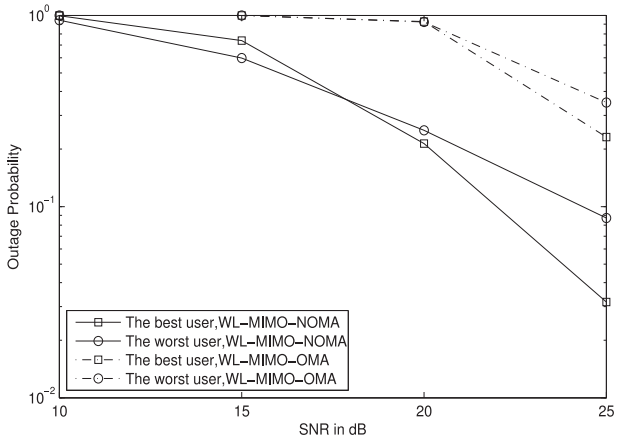

Fig. 8. WL-MIMO-NOMA v.s. WL-MIMO-OMA with real signals, $N = 3$, $M = 3$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 5$ BPCU, $R_2 = 4.5$ BPCU.

the result we can see that the WL-MIMO-NOMA is superior to WL-MIMO-OMA.

E. Coexistence of Real Signals and Complex Circular Signals

The mixed case of real and complex circular signals is simulated to compare WL-MIMO-NOMA with MIMO-NOMA and

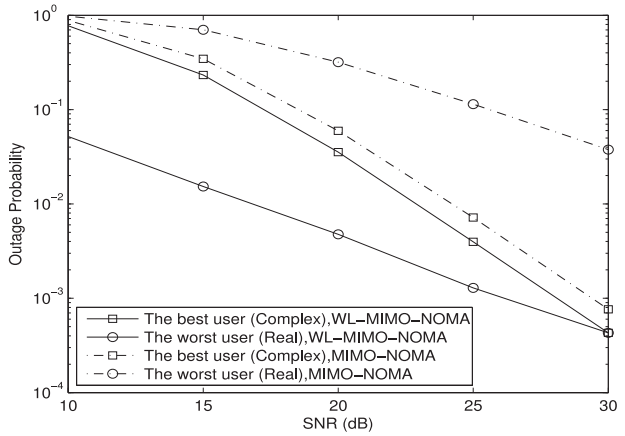

Fig. 9. WL-MIMO-NOMA v.s. MIMO-NOMA with mixed signals, $N = 3$, $M = 3$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 3$ BPCU, $R_2 = 1.8$ BPCU.

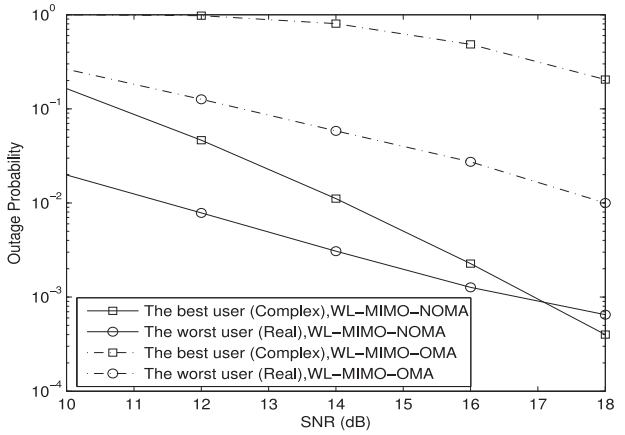

Fig. 10. WL-MIMO-NOMA v.s. WL-MIMO-OMA with mixed signals, $N = 3$, $M = 2$, $K = 2$, $\alpha_1^2 = \frac{1}{4}$, $\alpha_2^2 = \frac{3}{4}$, $R_1 = 2.5$ BPCU, $R_2 = 2$ BPCU.

WL-MIMO-OMA. The difference of two phase angles in WL-MIMO-NOMA system is $\pi/2$. Fig. 9 is obtained by the simulation using $M = N = 3$. It shows the advantage of the proposed WL-MIMO-NOMA over MIMO-NOMA. WL-MIMO-NOMA and WL-MIMO-OMA are also compared when the mixture of real and complex circular signals are transmitted by BS. The simulation is based on a system with $N = 3, M = 2$ and $K = 2$. In WL-MIMO-OMA system, one real signal user and one complex circular signal user are served during one time slot. The result is shown in Fig. 10, which illustrates the superiority of WL-MIMO-NOMA to WL-MIMO-OMA.

VII. Conclusion

In this paper, a WLP framework is proposed for MIMO-NOMA downlink systems, where the signals transmitted by the base station are real-valued. Under the WLP framework, complex power coefficients are used to stagger the user signals. In user pairing case, a conventional MIMO-NOMA model can be transformed into a classical MIMO model with expanded dimension, both the intra-cluster and inter-cluster interferences can be eliminated completely. The WLP framework can be extended to the general case of K-user cluster, where all the inter-cluster interference and at least half of the intra-cluster interference can be removed. Performance analysis is done in terms of the outage probability; the close-form expressions for outage probability are derived. Performance analysis shows that the user outage probability reaches the minimum value when the phases of complex power coefficients are orthogonal, which is validated by simulation results. The simulation also shows that WL-MIMO-NOMA system outperforms conventional MIMO-NOMA system. The framework is finally extended to the general mixed case including both real signals and complex circular signals.

Appendix A
Proof for Lemma 1

Without loss of generality, we can firstly consider the matrix $\overline{\mathbf{H}}_{1,k} = [\widetilde{\mathbf{h}}_{1,1k} \quad \overline{\overline{\mathbf{H}}}_{1,k}]$, obtained by removing the second column from the channel matrix $\widetilde{\mathbf{H}}_{1,k}$. Denote by $\overline{\mathbf{v}}_{1,1k}^H$ the first row of $\overline{\mathbf{H}}_{1,k}^{\dagger}$. It is easy to follow that $\overline{\mathbf{v}}_{1,1k}^H \widetilde{\mathbf{h}}_{1,1k} = 1, \overline{\mathbf{v}}_{1,1k}^H \overline{\overline{\mathbf{H}}}_{1,k} = \mathbf{0}$. Hence, $\overline{\mathbf{v}}_{1,1k}^H$ belongs to the null space of $\overline{\overline{\mathbf{H}}}_{1,k}$. Assume $\mathbf{U}$ is an orthonormal basis of the null space of $\overline{\overline{\mathbf{H}}}_{1,k}$, with $\mathbf{U}^H \overline{\overline{\mathbf{H}}}_{1,k} = \mathbf{0}$. $\overline{\mathbf{v}}_{1,1k}^H$ is represented by a combination of orthonormal basis as follows:

$$\overline{\mathbf{v}}_{1,1k} = \mathbf{U}\mathbf{x}_1 \tag{52}$$

where $\mathbf{x}_1$ is a coefficient vector to be optimized in the following form:

$$\min \mathbf{x}_1^H \mathbf{x}_1 \quad \text{s.t. } \mathbf{x}_1^H \mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} = 1 \tag{53}$$

The problem can be easily solved by the Lagrangian multiplier method, an auxiliary equation is constructed as follows [42]:

$$L = \mathbf{x}_1^H \mathbf{x}_1 + \lambda \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{x}_1 + (\lambda \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{x}_1)^*$$

To solve this problem, we take the partial derivative of L with respect to $\mathbf{x}_1$, and set $\frac{\partial L}{\partial \mathbf{x}_1} = \mathbf{0}$, $\mathbf{x}_1$ is given by:

$$\mathbf{x}_1^H = \frac{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}}{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}} \tag{54}$$

Because $\overline{\mathbf{v}}_{1,1k}^H \overline{\mathbf{v}}_{1,1k} = \mathbf{x}_1^H \mathbf{U}^H \mathbf{U}\mathbf{x}_1 = \mathbf{x}_1^H \mathbf{x}_1$, hence, $\mathbf{x}_1^H \mathbf{x}_1$ stands for the noise amplification effect of $\overline{\mathbf{v}}_{1,1k}$.

$$\mathbf{x}_1^H \mathbf{x}_1 = \frac{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}}{(\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k})^2} = \frac{1}{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}} \tag{55}$$

For computing $\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}$, we can exploit the special structure of $\widetilde{\mathbf{h}}_{1,1k}$ that the lower part of which is the conjugation of the upper part. Accordingly, $\mathbf{U}$ is partitioned into two parts as follows:

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \tag{56}$$

where $\mathbf{U}_1$ is the upper half of $\mathbf{U}$, $\mathbf{U}_2$ is the lower half of $\mathbf{U}$, and $\widetilde{\mathbf{h}}_{1,1k} = [\begin{smallmatrix}\mathbf{h}_{1,1k}e^{j\theta_1}\\ \mathbf{h}^*_{1,1k}e^{-j\theta_1}\end{smallmatrix}]$. Consequently, we have:

$$\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k} = \mathbf{h}^H_{1,1k}\mathbf{U}_1\mathbf{U}_1^H\mathbf{h}_{1,1k} + \mathbf{h}^T_{1,1k}\mathbf{U}_2\mathbf{U}_2^H\mathbf{h}^*_{1,1k} + \mathbf{h}^H_{1,1k}e^{-j\theta_1}\mathbf{U}_1\mathbf{U}_2^H\mathbf{h}^*_{1,1k}e^{-j\theta_1} + \mathbf{h}^T_{1,1k}e^{j\theta_1}\mathbf{U}_2{\mathbf{U}_1}^H\mathbf{h}_{1,1k}e^{j\theta_1} \tag{57}$$

The dimension of $\overline{\overline{\mathbf{H}}}_{1,k}$ is $2N \times (2M-2)$, it's rank is $(2M-2)$, so the dimension of $\mathbf{U}$ is $2N \times (2N-2M+2)$. Assuming the upper half of $\overline{\overline{\mathbf{H}}}_{1,k}$ is $\overline{\overline{\mathbf{H}}}_1$, thus the lower half of $\overline{\overline{\mathbf{H}}}_{1,k}$ is $\overline{\overline{\mathbf{H}}}^*_1$, which is given as follows:

$$\overline{\overline{\mathbf{H}}}_{1,k} = \begin{bmatrix}\overline{\overline{\mathbf{H}}}_1\\ \overline{\overline{\mathbf{H}}}^*_1\end{bmatrix} \tag{58}$$

The dimension of $\overline{\overline{\mathbf{H}}}_1$ and $\overline{\overline{\mathbf{H}}}^*_1$ is $N \times (2M-2)$, their rank is $(M-1)$, so the dimension of an orthonormal basis for the null space of $\overline{\overline{\mathbf{H}}}_1$ and $\overline{\overline{\mathbf{H}}}^*_1$ is $N \times (N-M+1)$. Assuming $\mathbf{V}$ is an orthonormal basis of the null space of $\overline{\overline{\mathbf{H}}}_1$; accordingly, $\mathbf{V}^*$ is an orthonormal basis of the null space of $\overline{\overline{\mathbf{H}}}^*_1$. $\mathbf{U}$ can be constructed in the following form:

$$\mathbf{U} = \begin{bmatrix}\mathbf{V} & \mathbf{0}_{N\times(N-M+1)}\\ \mathbf{0}_{N\times(N-M+1)} & \mathbf{V}^*\end{bmatrix} \tag{59}$$

It is easy to verify that $\mathbf{U}^H\overline{\overline{\mathbf{H}}}_{1,k} = \mathbf{0}$, so $\mathbf{U}$ is an orthonormal basis for the null space of $\overline{\overline{\mathbf{H}}}_{1,k}$. Based on (56) and (59), we can conclude that $\mathbf{U}_1\mathbf{U}_2^H = \mathbf{0}$ and $\mathbf{U}_2\mathbf{U}_1^H = \mathbf{0}$, (55) can be simplified as:

$$\mathbf{x}_1^H\mathbf{x}_1 = \frac{1}{\mathbf{h}^H_{1,1k}\mathbf{U}_1\mathbf{U}_1^H\mathbf{h}_{1,1k} + \mathbf{h}^T_{1,1k}\mathbf{U}_2\mathbf{U}_2^H\mathbf{h}^*_{1,1k}} \tag{60}$$

Next, we consider another scenario that the first column of the channel matrix $\widetilde{\mathbf{H}}_{1,k}$ is removed, which is the case of $\underline{\mathbf{H}}_{1,k} = [\,\widetilde{\mathbf{h}}_{2,1k}\ \overline{\overline{\mathbf{H}}}_{1,k}\,]$. Denote $\underline{\mathbf{v}}^H_{1,1k}$ the first row of $\underline{\mathbf{H}}^\dagger_{1,k}$. Similarly, $\underline{\mathbf{v}}^H_{1,1k}$ can be represented as:

$$\underline{\mathbf{v}}_{1,1k} = \mathbf{U}\mathbf{x}_2 \tag{61}$$

The coefficient vector $\mathbf{x}_2$ is given by the solution to the following problem:

$$\min \mathbf{x}_2^H\mathbf{x}_2 \ \text{ s.t. } \mathbf{x}_2^H\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k} = 1 \tag{62}$$

Similarly, $\mathbf{x}_2^H\mathbf{x}_2$ is given by:

$$\mathbf{x}_2^H\mathbf{x}_2 = \frac{1}{\mathbf{h}^H_{1,1k}\mathbf{U}_1\mathbf{U}_1^H\mathbf{h}_{1,1k} + \mathbf{h}^T_{1,1k}\mathbf{U}_2\mathbf{U}_2^H\mathbf{h}^*_{1,1k}} \tag{63}$$

It means $\mathbf{x}_1^H\mathbf{x}_1 = \mathbf{x}_2^H\mathbf{x}_2$, so is $\|\overline{\mathbf{v}}_{1,1k}\|^2 = \|\underline{\mathbf{v}}_{1,1k}\|^2$. The proof is done.

Appendix B
Proof for Lemma 2

We consider the first row of $\widetilde{\mathbf{H}}^\dagger_{1,k}$; its first row is denoted by $\mathbf{w}^H_{1,1k}$. Then we have $\mathbf{w}^H_{1,1k}\widetilde{\mathbf{h}}_{1,1k} = 1, \mathbf{w}^H_{1,1k}\widetilde{\mathbf{h}}_{2,1k} = 0, \mathbf{w}^H_{1,1k}\overline{\overline{\mathbf{H}}}_{1,k} = \mathbf{0}$. Assuming $\mathbf{U}$ is an orthonormal basis for the null space of $\overline{\overline{\mathbf{H}}}_{1,k}$, as a result, $\mathbf{w}_{1,1k}$ can be expressed as follows:

$$\mathbf{w}_{1,1k} = \mathbf{U}\mathbf{x} \tag{64}$$

where $\mathbf{x}$ is a coefficient vector to be optimized as follows:

$$\min \mathbf{x}^H\mathbf{x} \quad \text{s.t. } \mathbf{x}^H\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k} = 1, \mathbf{x}^H\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k} = 0 \tag{65}$$

The problem can be solved by the Lagrangian multiplier method, an auxiliary equation is constructed as follows [42]:

$$L = \mathbf{x}^H\mathbf{x} + \lambda\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{x} + \gamma\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{x} + (\lambda\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{x} + \gamma\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{x})^*$$

we take the partial derivative of L with respect to $\mathbf{x}$, and set $\frac{\partial L}{\partial \mathbf{x}} = \mathbf{0}$, $\mathbf{x}$ is given by:

$$\mathbf{x}^H = \frac{\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k} - \widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}}{\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k} - \widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}} \tag{66}$$

Based on Lemma 1, we have $\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k} = \widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}$. Thus the final expression of $\mathbf{x}^H\mathbf{x}$ can be written as:

$$\mathbf{x}^H\mathbf{x} = \frac{c}{c^2 - \widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}} \tag{67}$$

where $c = \widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k} = \widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}$.

Next, we consider the second row of $\widetilde{\mathbf{H}}^\dagger_{1,k}$; its second row is denoted by $\mathbf{w}^H_{2,1k}$. Then we have $\mathbf{w}^H_{2,1k}\widetilde{\mathbf{h}}_{1,1k} = 0, \mathbf{w}^H_{2,1k}\widetilde{\mathbf{h}}_{2,1k} = 1, \mathbf{w}^H_{2,1k}\overline{\overline{\mathbf{H}}}_{1,k} = \mathbf{0}$, $\mathbf{w}_{2,1k}$ can be represented as:

$$\mathbf{w}_{2,1k} = \mathbf{U}\widetilde{\mathbf{x}} \tag{68}$$

Similarly, $\widetilde{\mathbf{x}}^H\widetilde{\mathbf{x}}$ is given by:

$$\widetilde{\mathbf{x}}^H\widetilde{\mathbf{x}} = \frac{c}{c^2 - \widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}} \tag{69}$$

It means $\mathbf{x}^H\mathbf{x} = \widetilde{\mathbf{x}}^H\widetilde{\mathbf{x}}$. Because $\mathbf{w}^H_{1,1k}\mathbf{w}_{1,1k} = \mathbf{x}^H\mathbf{U}^H\mathbf{U}\mathbf{x} = \mathbf{x}^H\mathbf{x}$, $\mathbf{w}^H_{2,1k}\mathbf{w}_{2,1k} = \widetilde{\mathbf{x}}^H\mathbf{U}^H\mathbf{U}\widetilde{\mathbf{x}} = \widetilde{\mathbf{x}}^H\widetilde{\mathbf{x}}$, so we have $\|\mathbf{w}_{1,1k}\|^2 = \|\mathbf{w}_{2,1k}\|^2$. The proof is done.

Appendix C
Proof for Lemma 3

According to Lemma 1 and Lemma 2, we can find that $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2} = \frac{\|\mathbf{w}_{2,1k}\|^2}{\|\underline{\mathbf{v}}_{1,1k}\|^2}$, and we focus on the ratio $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2}$. Recall the results, (60) and (67), $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2}$ is given by:

$$\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2} = \frac{1}{1 - \frac{\widetilde{\mathbf{h}}^H_{2,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{2,1k}}{\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}\widetilde{\mathbf{h}}^H_{1,1k}\mathbf{U}\mathbf{U}^H\widetilde{\mathbf{h}}_{1,1k}}} \tag{70}$$

According to Appendix A, $\mathbf{U}$ can be expressed as:

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{bmatrix} = \begin{bmatrix}\mathbf{V} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}^*\end{bmatrix}$$

And $\widetilde{\mathbf{h}}_{1,1k}$ and $\widetilde{\mathbf{h}}_{2,1k}$ can be written as:

$$\widetilde{\mathbf{h}}_{1,1k} = \begin{bmatrix} \mathbf{h}_{1,1k} e^{j\theta_1} \\ \mathbf{h}_{1,1k}^* e^{-j\theta_1} \end{bmatrix}, \widetilde{\mathbf{h}}_{2,1k} = \begin{bmatrix} \mathbf{h}_{1,1k} e^{j\theta_2} \\ \mathbf{h}_{1,1k}^* e^{-j\theta_2} \end{bmatrix}$$

In (70), expression $\frac{\widetilde{\mathbf{h}}_{2,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{2,1k}}{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}}$ can be rewritten as (71), shown at the bottom of this page. Because $\mathbf{U}_1\mathbf{U}_1^H = \mathbf{V}\mathbf{V}^H, \mathbf{U}_2\mathbf{U}_2^H = \mathbf{V}^*\mathbf{V}^T$, so $\mathbf{U}_1\mathbf{U}_1^H = (\mathbf{U}_2\mathbf{U}_2^H)^*$. The value of $\mathbf{h}_{1,1k}^H \mathbf{U}_1\mathbf{U}_1^H \mathbf{h}_{1,1k}$ is real, which can be extracted from both numerator and denominator. So, (71) can be further expressed as:

$$\frac{\widetilde{\mathbf{h}}_{2,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{2,1k}}{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}} = \cos^2(\theta_1 - \theta_2) \tag{72}$$

Combining (70) and (72), the ratio $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2}$ is given by:

$$\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2} = \frac{1}{1 - \cos^2(\theta_1 - \theta_2)} \tag{73}$$

Therefore, $\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2} = \frac{\|\mathbf{w}_{2,1k}\|^2}{\|\underline{\mathbf{v}}_{1,1k}\|^2}$ is only related to angles θ_1 and θ_2. The proof is done.

Appendix D
Proof for Lemma 4

We can prove that the orthonormal basis $\mathbf{U}$ for the null space of $\overline{\overline{\mathbf{H}}}_{1,k}$ is independent of angles θ_1 and θ_2. According to Appendix A, $\mathbf{U}$ can be expressed as:

$$\mathbf{U} = \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^* \end{bmatrix}$$

According to (58) and (59), $\mathbf{V}$ is an orthonormal basis for the null space of $\overline{\overline{\mathbf{H}}}_1$, $\overline{\overline{\mathbf{H}}}_1$ is the upper half of $\overline{\overline{\mathbf{H}}}$, hence, $\mathbf{V}^H \overline{\overline{\mathbf{H}}}_1 = \mathbf{0}$, and we also have $\mathbf{V}^H \overline{\overline{\mathbf{H}}}_1 \overline{\overline{\mathbf{H}}}_1^H = \mathbf{0}$. Therefore, $\mathbf{V}$ is also an orthonormal basis for the null space of $\overline{\overline{\mathbf{H}}}_1 \overline{\overline{\mathbf{H}}}_1^H$. Because the value of $\overline{\overline{\mathbf{H}}}_1 \overline{\overline{\mathbf{H}}}_1^H$ is independent of angles θ_1 and θ_2, so $\mathbf{V}$ and $\mathbf{U}$ are independent of θ_1 and θ_2. Recalling that

$$\|\overline{\mathbf{v}}_{1,1k}\|^2 = \|\underline{\mathbf{v}}_{1,1k}\|^2 = \frac{1}{\mathbf{h}_{1,1k}{}^H \mathbf{U}_1\mathbf{U}_1{}^H \mathbf{h}_{1,1k} + \mathbf{h}_{1,1k}^T \mathbf{U}_2\mathbf{U}_2{}^H \mathbf{h}_{1,1k}^*}$$

Because $\mathbf{U}$ is independent of θ_1 and θ_2, so $\|\overline{\mathbf{v}}_{1,1k}\|^2$ and $\|\underline{\mathbf{v}}_{1,1k}\|^2$ are independent of θ_1 and θ_2. With the aid of Lemma 3, we have:

$$\frac{\|\mathbf{w}_{1,1k}\|^2}{\|\overline{\mathbf{v}}_{1,1k}\|^2} = \frac{\|\mathbf{w}_{2,1k}\|^2}{\|\underline{\mathbf{v}}_{1,1k}\|^2} = \frac{1}{1 - \cos^2(\theta_1 - \theta_2)}$$

Consequently, $\|\mathbf{w}_{1,1k}\|^2$ and $\|\mathbf{w}_{2,1k}\|^2$ take the minimum value when θ_1 and θ_2 are orthogonal. In this case, $\|\mathbf{w}_{1,1k}\|^2$ and $\|\mathbf{w}_{2,1k}\|^2$ are given by:

$$\|\mathbf{w}_{1,1k}\|^2 = \|\mathbf{w}_{2,1k}\|^2 = \|\overline{\mathbf{v}}_{1,1k}\|^2 = \|\underline{\mathbf{v}}_{1,1k}\|^2$$

Recall the outage probability analysis about the two users earlier, when $\|\mathbf{w}_{2,12}\|^2$ takes the minimum value, the outage probability achieved by the second user reaches the minimum value; the outage probability experienced by the first user is determined by $\|\overline{\mathbf{v}}_{1,11}\|^2$ and $\|\mathbf{w}_{2,11}\|^2$, the value of $\|\overline{\mathbf{v}}_{1,11}\|^2$ is independent of θ_1 and θ_2, and $\|\mathbf{w}_{2,11}\|^2$ reaches the minimum value when θ_1 and θ_2 are orthogonal, so the outage probability achieved by the two users reaches the minimum value when θ_1 and θ_2 are orthogonal.

Appendix E
Proof for Theorem 1

The detection strategy for the first user in the first cluster depends on that whether the phases are orthogonal. So the analysis of the outage probability of the first user needs to be discussed in two cases.

Case 1: θ_1 and θ_2 are nonorthogonal

The outage probability experienced by the first user in the first cluster is given in (34), which is re-written as follows:

$$P_{1,1}^o = 1 - P\left(\frac{{\alpha_{1,1}}^2}{\|\overline{\mathbf{v}}_{1,11}\|^2 \frac{1}{\rho}} > \zeta_{1,1}, \frac{{\alpha_{1,2}}^2}{\|\mathbf{w}_{2,11}\|^2 \frac{1}{\rho}} > \zeta_{1,2} \right)$$

According to (36) and (38), the above outage probability can be further expressed as follows:

$$\begin{aligned} P_{1,1}^o &= 1 - P\left(\frac{{\alpha_{1,1}}^2}{\|\underline{\mathbf{v}}_{1,11}\|^2 \frac{1}{\rho}} > \zeta_{1,1}, \frac{{\alpha_{1,2}}^2}{\|\mathbf{w}_{2,11}\|^2 \frac{1}{\rho}} > \zeta_{1,2} \right) \\ &= 1 - P\left(\frac{1}{\|\underline{\mathbf{v}}_{1,11}\|^2} > \varepsilon_{1,1}^* \right) \end{aligned} \tag{74}$$

where $\varepsilon_{1,1}^* = \max\{\frac{\zeta_{1,1}}{\rho\alpha_{1,1}^2}, \frac{\zeta_{1,2}}{\rho\alpha_{1,2}^2(1-\cos^2(\theta_1-\theta_2))}\}$.

Case 2: θ_1 and θ_2 are orthogonal According to Lemma 4, the SIC process can be skipped in the detection of the first user. The first user can decode its information directly by using the detection vector $\mathbf{w}_{1,11}$, the outage

$$\frac{\widetilde{\mathbf{h}}_{2,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{2,1k}}{\widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k} \widetilde{\mathbf{h}}_{1,1k}^H \mathbf{U}\mathbf{U}^H \widetilde{\mathbf{h}}_{1,1k}} = \frac{(e^{j\theta_1}e^{-j\theta_2}\mathbf{h}_{1,1k}^H \mathbf{U}_1\mathbf{U}_1^H \mathbf{h}_{1,1k} + e^{-j\theta_1}e^{j\theta_2}\mathbf{h}_{1,1k}^T \mathbf{U}_2\mathbf{U}_2^H \mathbf{h}_{1,1k}^*)(e^{-j\theta_1}e^{j\theta_2}\mathbf{h}_{1,1k}^H \mathbf{U}_1\mathbf{U}_1^H \mathbf{h}_{1,1k} + e^{j\theta_1}e^{-j\theta_2}\mathbf{h}_{1,1k}^T \mathbf{U}_2\mathbf{U}_2^H \mathbf{h}_{1,1k}^*)}{(\mathbf{h}_{1,1k}^H \mathbf{U}_1\mathbf{U}_1^H \mathbf{h}_{1,1k} + \mathbf{h}_{1,1k}^T \mathbf{U}_2\mathbf{U}_2^H \mathbf{h}_{1,1k}^*)^2} \tag{71}$$

probability is given by:

$$P_{1,1}^{\circ} = P\left(\frac{{\alpha_{1,1}}^2}{\|\mathbf{w}_{1,11}\|^2 \frac{1}{\rho}} < \zeta_{1,1}\right) = P\left(\frac{1}{\left\|\underline{\mathbf{v}}_{1,11}\right\|^2} < \frac{\zeta_{1,1}}{\rho\alpha_{1,1}^2}\right) \tag{75}$$

Similarly, the outage probability experienced by the second user in the first cluster can be rewritten as:

$$P_{1,2}^{\circ} = P\left(\frac{{\alpha_{1,2}}^2}{\|\mathbf{w}_{2,12}\|^2 \frac{1}{\rho}} < \zeta_{1,2}\right) = P\left(\frac{1}{\left\|\underline{\mathbf{v}}_{1,12}\right\|^2} < \frac{\zeta_{1,2}}{\rho\alpha_{1,2}^2(1-\cos^2(\theta_1-\theta_2))}\right) \tag{76}$$

Next, we focus on analyzing the property of $\|\underline{\mathbf{v}}_{1,1k}\|^2$ to give the proof of Theorem 1. Recall that $\underline{\mathbf{v}}_{1,1k}$ the first row of $\underline{\mathbf{H}}_{1,k}^{\dagger}$ with $\underline{\mathbf{H}}_{1,k} = [\,\widetilde{\mathbf{h}}_{2,1k}\ \overline{\overline{\mathbf{H}}}_{1,k}\,]$. The $\|\underline{\mathbf{v}}_{1,1k}\|^2$ can be expressed as follows [41]:

$$\|\underline{\mathbf{v}}_{1,1k}\|^2 = \frac{\det(\overline{\overline{\mathbf{H}}}_{1,k}^{H}\overline{\overline{\mathbf{H}}}_{1,k})}{\det\left(\underline{\mathbf{H}}_{1,k}^{H}\underline{\mathbf{H}}_{1,k}\right)} \tag{77}$$

The computation in (77) is based on the complex-valued model. To simplify the analysis, the analysis can be converted into a real-valued model by multiplying the following unitary matrix $\mathbf{P}$:

$$\mathbf{P} = \sqrt{2}\begin{pmatrix} \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} \\ \frac{1}{2j}\mathbf{I} & \frac{-1}{2j}\mathbf{I} \end{pmatrix} \tag{78}$$

One can easily verify that $\overline{\overline{\mathbf{H}}}_{1,k}^{r} = \frac{1}{\sqrt{2}}\mathbf{P}\overline{\overline{\mathbf{H}}}_{1,k}$ is real. Similarly, we can obtain the real matrix by $\underline{\mathbf{H}}_{1,k}^{r} = \frac{1}{\sqrt{2}}\mathbf{P}\underline{\mathbf{H}}_{1,k}$. Therefore, we have:

$$\frac{\det\left(\left(\overline{\overline{\mathbf{H}}}_{1,k}^{r}\right)^{H}\overline{\overline{\mathbf{H}}}_{1,k}^{r}\right)}{\det\left(\left(\underline{\mathbf{H}}_{1,k}^{r}\right)^{H}\underline{\mathbf{H}}_{1,k}^{r}\right)} = \frac{\det\left(\left(\frac{1}{\sqrt{2}}\mathbf{P}\overline{\overline{\mathbf{H}}}_{1,k}\right)^{H}\frac{1}{\sqrt{2}}\mathbf{P}\overline{\overline{\mathbf{H}}}_{1,k}\right)}{\det\left(\left(\frac{1}{\sqrt{2}}\mathbf{P}\underline{\mathbf{H}}_{1,k}\right)^{H}\frac{1}{\sqrt{2}}\mathbf{P}\underline{\mathbf{H}}_{1,k}\right)} \tag{79}$$

Denoting the above equation as $\|\underline{\mathbf{r}}_{1,1k}\|^2$, it can be further expressed as:

$$\left\|\underline{\mathbf{r}}_{1,1k}\right\|^2 = \frac{\det\left(\frac{1}{2}\overline{\overline{\mathbf{H}}}_{1,k}^{H}\overline{\overline{\mathbf{H}}}_{1,k}\right)}{\det\left(\frac{1}{2}\underline{\mathbf{H}}_{1,k}^{H}\underline{\mathbf{H}}_{1,k}\right)} = \frac{\frac{1}{2^{2M-2}}\det\left(\overline{\overline{\mathbf{H}}}_{1,k}^{H}\overline{\overline{\mathbf{H}}}_{1,k}\right)}{\frac{1}{2^{2M-1}}\det\left(\underline{\mathbf{H}}_{1,k}^{H}\underline{\mathbf{H}}_{1,k}\right)} \tag{80}$$

We can find that $\|\underline{\mathbf{r}}_{1,1k}\|^2 = 2\|\underline{\mathbf{v}}_{1,1k}\|^2$. So the outage probability achieved by the users in the first cluster can be rewritten as follows:

$$P_{1,1}^{\circ} = P\left(\frac{1}{\|\underline{\mathbf{r}}_{1,11}\|^2} < \varepsilon_{1,1}\right) \tag{81}$$

$$P_{1,2}^{\circ} = P\left(\frac{1}{\|\underline{\mathbf{r}}_{1,12}\|^2} < \varepsilon_{1,2}\right) \tag{82}$$

When θ_1 and θ_2 are orthogonal, $\varepsilon_{1,1} = \frac{\zeta_{1,1}}{2\rho\alpha_{1,1}^2}$, $\varepsilon_{1,2} = \frac{\zeta_{1,2}}{2\rho\alpha_{1,2}^2}$. When θ_1 and θ_2 are nonorthogonal, $\varepsilon_{1,1} = \max\{\frac{\zeta_{1,1}}{2\rho\alpha_{1,1}^2}, \frac{\zeta_{1,2}}{2\rho\alpha_{1,2}^2(1-\cos^2(\theta_1-\theta_2))}\}$, $\varepsilon_{1,2} = \frac{\zeta_{1,2}}{2\rho\alpha_{1,2}^2(1-\cos^2(\theta_1-\theta_2))}$. Actually, if the first user's targeted data rate is greater than the second user's, we can observe that the former is a special case of the latter.

According to [43], $\frac{1}{\|\underline{\mathbf{r}}_{1,1k}\|^2}$ follows the chi-square distribution with $(2N-2M+2)$ degrees of freedom. Recall that the users are ordered as $\|\mathbf{w}_{1,11}\|^2 \le \|\mathbf{w}_{2,12}\|^2$, it means $\frac{1}{\|\underline{\mathbf{r}}_{1,11}\|^2} \ge \frac{1}{\|\underline{\mathbf{r}}_{1,12}\|^2}$. Define $\widetilde{x}_k$ as the unordered variables following the chi-square distribution with $(2N-2M+2)$ degrees of freedom, where $\widetilde{x}_k = \frac{1}{\|\underline{\mathbf{r}}_{1,1k}\|^2}$. Thus the probability density function (pdf) of $\widetilde{x}_k$ is given by:

$$f_{\widetilde{x}_k}(x) = \frac{e^{-x}}{(N-M)!}x^{N-M} \tag{83}$$

And its cumulative distribution function (CDF) is $F_{\widetilde{x}_l}(x) = \int_0^x f_{\widetilde{x}_l}(y)dy$. Define x_k the ordered copy of $\widetilde{x}_k$, according to $\frac{1}{\|\underline{\mathbf{r}}_{1,11}\|^2} \ge \frac{1}{\|\underline{\mathbf{r}}_{1,12}\|^2}$. Therefore the pdf of ordered x_k is given by:

$$f_{x_k}(x) = \frac{2!f_{\widetilde{x}_k}(x)[F_{\widetilde{x}_k}(x)]^{2-k}[1-F_{\widetilde{x}_k}(x)]^{k-1}}{(2-k)!\,(k-1)!} \tag{84}$$

The outage probability experienced by the k-th user in the first cluster is $P_{1,k}^{\circ} = \int_0^{\varepsilon_{1,k}} f_{x_k}(x)dx$, by applying the PDF of the ordered x_k, it can be expressed as follows:

$$\begin{aligned} P_{1,k} &= \int_0^{\varepsilon_{1,k}} \frac{2!f_{\widetilde{x}_k}(x)[F_{\widetilde{x}_k}(x)]^{2-k}[1-F_{\widetilde{x}_k}(x)]^{k-1}}{(2-k)!(k-1)!}dx \\ &= \sum_{m=0}^{k-1}\binom{k-1}{m}\frac{2!(-1)^m}{(2-k)!(k-1)!} \\ &\quad \times \int_0^{\varepsilon_{1,k}} f_{\widetilde{x}_k}(x)[F_{\widetilde{x}_k}(x)]^{2-k}F_{\widetilde{x}_k}(x)^m\,dx \\ &= \sum_{m=0}^{k-1}\binom{k-1}{m}\frac{2!(-1)^m\,[F_{\widetilde{x}_k}(\varepsilon_{1,k})]^{2-k+m+1}}{(2-k)!(k-1)!(2-k+m+1)} \end{aligned} \tag{85}$$

By applying the CDF of the unordered variable $\widetilde{x}_k$, and then applying the incomplete gamma function, we can get the overall outage probability shown in the Theorem 1‘. The proof is done.

Appendix F
Proof for Theorem 2

The case of K-user cluster is considered from the following perspective. The users in each cluster are divided into two subsets. For detecting a user signal in a subset, the interferences from the other $2M-1$ subsets are completely removed

by considering 2 subsets in a cluster as 2 'big users', then SIC is employed to detect the user signal by considering the other users within the subset as interferences. Without loss of generality, the analysis can be performed by focusing on the first cluster.

1) The analysis begins with subset $\Theta_{1,1} = \{s_{1,1}, s_{1,2}, \ldots, s_{1,K_1}\}$. The user with worst channel condition can decode its own information with regarding other users' messages as noises, so the outage probability for the K_1-th user in subset $\Theta_{1,1}$ is given by:

$$P\left(SINR_{1,K_1} < \zeta_{1,K_1}\right) = P\left(\frac{\alpha_{1,K_1}^2}{\sum_{m=1}^{K_1-1}\alpha_{1,m}^2 + \left\|\mathbf{w}_{1,K_1}\right\|^2 \frac{1}{\rho}} < \zeta_{1,K_1}\right) \quad (86)$$

The outage probability can be further expressed as follows:

$$\begin{cases} P\left(\frac{1}{\left\|\mathbf{w}_{1,K_1}\right\|^2} < \frac{\zeta_{1,K_1}}{\rho(\alpha_{1,K_1}^2 - \zeta_{1,K_1}\sum_{m=1}^{K_1-1}\alpha_{1,m}^2)}\right) & \text{if } C_1 \\ 1 & \text{otherwise} \end{cases} \quad (87)$$

where C_1 denotes $\alpha_{1,K_1}^2 > \zeta_{1,K_1}\sum_{m=1}^{K_1-1}\alpha_{1,m}^2$.
According to Lemma 4 and the formula (80), the above outage probability can be rewritten as:

$$P_{1,K_1}^{\circ} = \begin{cases} P\left(\frac{1}{\left\|\underline{\mathbf{r}}_{1,K_1}\right\|^2} < \varepsilon_{1,K_1}\right) & \text{if } C_1 \\ 1 & \text{otherwise} \end{cases} \quad (88)$$

where $\varepsilon_{1,K_1} = \frac{\zeta_{1,K_1}}{2\rho(\alpha_{1,K_1}^2 - \zeta_{1,K_1}\sum_{m=1}^{K_1-1}\alpha_{1,m}^2)}$, and $\frac{1}{\left\|\underline{\mathbf{r}}_{1,K_1}\right\|^2}$ follows the chi-square distribution with $(2N-2M+2)$ degrees of freedom.

2)) The k-th user in $\Theta_{1,1}, 2 \le k < K_1$, needs to use SIC strategy to decode the information of j-th user, $1+k \le j \le K_1$, and then remove all these users' information before detecting its own. The k-th user can only decode its information successfully when all these users' messages are decoded, so the overall outage probability of the k-th user is shown as follows:

$$\begin{aligned} P_{1,k}^{\circ} &= 1 - P\left(SINR_{1,k}^{j} > \zeta_{1,j}, j \in \{k, \ldots, K_1\}\right) \\ &= 1 - P\left(\frac{\alpha_{1,j}^2}{\sum_{m=1}^{j-1}\alpha_{1,m}^2 + \left\|\mathbf{w}_{1,k}\right\|^2 \frac{1}{\rho}} \right. \\ &\qquad \left. > \zeta_{1,j}, j \in \{k, \ldots, K_1\}\right) \end{aligned} \quad (89)$$

Similarly, the above outage probability can be rewritten as:

$$P_{1,k}^{\circ} = \begin{cases} P\left(\frac{1}{\left\|\underline{\mathbf{r}}_{1,k}\right\|^2} < \varepsilon_{1,k}\right) & \text{if } C_2 \\ 1 & \text{otherwise} \end{cases} \quad (90)$$

where the condition C_2 denotes $\alpha_{1,j}^2 > \zeta_{1,j}\sum_{m=1}^{j-1}\alpha_{1,m}^2$, for all $j \in \{k, \ldots, K_1\}$, and

$$\varepsilon_{1,k} = \max\left\{\frac{\zeta_{1,K_1}}{\rho(\alpha_{1,K_1}^2 - \zeta_{1,K_1}\sum_{m=1}^{K_1-1}\alpha_{1,m}^2)}, \ldots, \frac{\zeta_{1,k}}{\rho(\alpha_{1,k}^2 - \zeta_{1,k}\sum_{m=1}^{k-1}\alpha_{1,m}^2)}\right\}$$

3) The first user in $\Theta_{1,1}$ needs to use SIC strategy to decode its information. Consequently, the outage probability achieved by the first user is given by:

$$P_{1,1}^{\circ} = \begin{cases} P\left(\frac{1}{\left\|\underline{\mathbf{r}}_{1,1}\right\|^2} < \varepsilon_{1,1}\right) & \text{if } C_3 \\ 1 & \text{otherwise} \end{cases} \quad (91)$$

where C_3 denotes $\alpha_{1,j}^2 > \zeta_{1,j}\sum_{m=1}^{j-1}\alpha_{1,m}^2$, for all $j \in \{2, \ldots, K_1\}$, and

$$\varepsilon_{1,1} = \max\left\{\frac{\zeta_{1,K_1}}{2\rho(\alpha_{1,K_1}^2 - \zeta_{1,K_1}\sum_{m=1}^{K_1-1}\alpha_{1,m}^2)}, \ldots, \frac{\zeta_{1,2}}{2\rho(\alpha_{1,2}^2 - \zeta_{1,2}\alpha_{1,1}^2)}, \frac{\zeta_{1,1}}{2\rho\alpha_{1,1}^2}\right\}$$

Following the analysis on $\Theta_{1,1}$, the outage probability experienced by users in $\Theta_{1,2} = \{s_{1,K_1+1}, s_{1,K_1+2}, \ldots, s_{1,K}\}$ can be directly given in the following.

4) The outage probability for the K-th user in $\Theta_{1,2}$ is given by:

$$P_{1,K}^{o} = \begin{cases} P\left(\frac{1}{\left\|\underline{\mathbf{r}}_{1,K}\right\|^2} < \varepsilon_{1,K}\right) & \text{if } C_4 \\ 1 & \text{otherwise} \end{cases} \quad (92)$$

where C_4 denotes $\beta_{1,K}^2 > \zeta_{1,K}\sum_{m=K_1+1}^{K-1}\beta_{1,m}^2$, and $\varepsilon_{1,K} = \frac{\zeta_{1,K}}{2\rho(\beta_{1,K}^2 - \zeta_{1,K}\sum_{m=K_1+1}^{K-1}\beta_{1,m}^2)}$.

5) The outage probability experienced by the k-th user, $K_1 + 1 \le k < K$, is shown as follows:

$$P_{1,k}^{\circ} = \begin{cases} P\left(\frac{1}{\left\|\underline{\mathbf{r}}_{1,k}\right\|^2} < \varepsilon_{1,k}\right) & \text{if } C_5 \\ 1 & \text{otherwise} \end{cases} \quad (93)$$

where C_5 denotes $\beta_{1,j}^2 > \zeta_{1,j}\sum_{m=K_1+1}^{j-1}\beta_{1,m}^2$, for all $j \in \{k, \ldots, K\}$, and

$$\varepsilon_{1,k} = \max\left\{\frac{\zeta_{1,K}}{2\rho(\beta_{1,K}^2 - \zeta_{1,K}\sum_{m=K_1+1}^{K-1}\beta_{1,m}^2)}, \ldots, \frac{\zeta_{1,k}}{2\rho(\beta_{1,k}^2 - \zeta_{1,k}\sum_{m=K_1+1}^{k-1}\beta_{1,m}^2)}\right\}$$

6) The outage probability achieved by the $K_1 + 1$ user is given by:

$$P_{1,K_1+1}^{\circ} = \begin{cases} P\left(\frac{1}{\left\|\underline{\mathbf{r}}_{1,K_1+1}\right\|^2} < \varepsilon_{1,K_1+1}\right) & \text{if } C_6 \\ 1 & \text{otherwise} \end{cases} \quad (94)$$

where C_6 denotes $\alpha_{1,j}^2 > \zeta_{1,j}\sum_{m=1}^{j-1}\alpha_{1,m}^2$, for all $j \in \{K_1+2, \ldots, K\}$. And

$$\varepsilon_{1,K_1+1} = \max\left\{\frac{\zeta_{1,K}}{2\rho(\beta_{1,K}^2 - \zeta_{1,K}\sum_{m=K_1+1}^{K-1}\beta_{1,m}^2)}, \ldots, \frac{\zeta_{1,K_1+2}}{2\rho(\beta_{1,K_1+2}^2 - \zeta_{1,K_1+2}\beta_{1,K_1+1}^2)}, \frac{\zeta_{1,K_1+1}}{2\rho\beta_{1,K_1+1}^2}\right\}$$

Despite the channel condition ordering in (22) and (23), we have underlined that there is no specific magnitude order for any pair across the subsets, i.e. $\|\mathbf{w}_{1,k_1}\|^2$ can be bigger or smaller than $\|\mathbf{w}_{1,K_1+k_2}\|^2$ for $1 \le k_1 \le K_1$ and $1 \le k_2 \le K_2$.

When considering the user's channel conditions throughout the cluster, we give an ordering overall channel conditions as follows:

$$\|\mathbf{w}_1\|^2 \le \|\mathbf{w}_2\|^2 \le \cdots \le \|\mathbf{w}_K\|^2 \tag{95}$$

where $\|\mathbf{w}_{l_k}\|^2 = \|\mathbf{w}_{1,k}\|^2$, $l_k \in \{1, 2, \ldots, K\}$, it means that $\|\mathbf{w}_{1,k}\|^2$ is ranked as the l_k-th position in (95). Accordingly, denote the $\frac{1}{\|\underline{\mathbf{r}}_{1,k}\|^2}$ by $\frac{1}{\|\underline{\mathbf{r}}_{l_k}\|^2}$, the order of $\frac{1}{\|\underline{\mathbf{r}}_{l_k}\|^2}$ is given by:

$$\frac{1}{\|\underline{\mathbf{r}}_1\|^2} \ge \frac{1}{\|\underline{\mathbf{r}}_2\|^2} \ge \cdots \ge \frac{1}{\|\underline{\mathbf{r}}_K\|^2} \tag{96}$$

Define $\widetilde{x}_{l_k}$ as the unordered variables following the chi-square distribution with $(2N - 2M + 2)$ degrees of freedom, where $\widetilde{x}_{l_k} = \frac{1}{\|\underline{\mathbf{r}}_{l_k}\|^2}$. Thus the probability density function (pdf) of $\widetilde{x}_{l_k}$ is given by:

$$f_{\widetilde{x}_{l_k}}(x) = \frac{e^{-x}}{(N-M)!} x^{N-M} \tag{97}$$

And its cumulative distribution function (CDF) is $F_{\widetilde{x}_{l_k}}(x) = \int_0^x f_{\widetilde{x}_{l_k}}(y)dy$. Define x_{l_k} the ordered copy of $\widetilde{x}_{l_k}$, therefore the pdf of x_{l_k} is given as follows:

$$f_{x_{l_k}}(x) = \frac{K! f_{\widetilde{x}_{l_k}}(x)[F_{\widetilde{x}_{l_k}}(x)]^{K-l_k}[1 - F_{\widetilde{x}_{l_k}}(x)]^{l_k-1}}{(K-l_k)!(l_k-1)!} \tag{98}$$

The outage probability experienced by the *k*-th user in the first cluster is equivalent to the outage probability experienced by the l_k-th user in the first cluster, it can be expressed as $P_{1,k}^{\circ} = \int_0^{\varepsilon_{1,k}} f_{x_{l_k}}(x)dx$, by applying the PDF of the ordered x_{l_k}, it can be expressed as follows:

$$\begin{aligned} P_{1,k} &= \int_0^{\varepsilon_{1,k}} \frac{K! f_{\widetilde{x}_{l_k}}(x)[F_{\widetilde{x}_{l_k}}(x)]^{K-l_k}[1 - F_{\widetilde{x}_{l_k}}(x)]^{l_k-1}}{(K-l_k)!(l_k-1)!} dx \\ &= \sum_{m=0}^{l_k-1} \binom{l_k-1}{m} \frac{K!(-1)^m}{(K-l_k)!(l_k-1)!} \\ &\quad \times \int_0^{\varepsilon_{1,k}} f_{\widetilde{x}_{l_k}}(x)[F_{\widetilde{x}_{l_k}}(x)]^{K-l_k} F_{\widetilde{x}_{l_k}}(x)^m dx \\ &= \sum_{m=0}^{l_k-1} \binom{l_k-1}{m} \frac{K!(-1)^m [F_{\widetilde{x}_{l_k}}(\varepsilon_{1,k})]^{K-l_k+m+1}}{(K-l_k)!(l_k-1)!(K-l_k+m+1)} \end{aligned} \tag{99}$$

By applying the CDF of the unordered variable $\widetilde{x}_{l_k}$, and then applying the incomplete gamma function, we can get the overall outage probability shown in the Theorem 2.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf.*, pp. 1–5, 2013.

[2] J. Chen, H. Chen, H. Zhang, and F. Zhao, "Spectral-energy efficiency tradeoff in relay-aided massive MIMO cellular networks with pilot contamination," *IEEE Access*, vol. 4, pp. 5234–5242, 2016.

[3] S. M. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," 2017, arXiv preprint arXiv:1706.08215.

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

[5] S. A. R. Naqvi and S. A. Hassan, "Combining NOMA and mmWave technology for cellular communication," in *Proc. IEEE 84th Veh. Technol. Conf.*, 2016, pp. 1–5.

[6] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[7] J. Cui, G. Dong, S. Zhang, H. Li, and G. Feng, "Asynchronous NOMA for downlink transmissions," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 402–405, Feb. 2017.

[8] Y. Wu and L. P. Qian, "Energy-efficient NOMA-enabled traffic offloading via dual-connectivity in small-cell networks," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1605–1608, Jul. 2017.

[9] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, M. Elkashlan, and J. Alonso-Zarate, "NOMA-based D2D communications: Towards 5G," in *Proc. IEEE Global Commun. Conf.*, pp. 1–6, 2016.

[10] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[11] X. Chen, A. Benjebboui, Y. Lan, A. Li, and H. Jiang, "Evaluations of downlink non-orthogonal multiple access (NOMA) combined with SU-MIMO," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun.*, 2014, pp. 1887–1891.

[12] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, "System-level performance of downlink NOMA combined with SU-MIMO for future LTE enhancements," in *Proc. IEEE Globecom Workshops*, 2014, pp. 706–710.

[13] Y. Lan, A. Benjebboiu, X. Chen, A. Li, and H. Jiang, "Considerations on downlink non-orthogonal multiple access (NOMA) combined with closed-loop SU-MIMO," in *Proc. 8th Int. Conf. Signal Process. Commun. Syst.*, Dec. 2014, pp. 1–5.

[14] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[15] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas. Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[16] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84–87, Jan. 2017.

[17] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016

[18] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the internet of things," *IEEE Access*, vol. 4, pp. 1393–1405, 2016.

[19] P. Chevalier and A. Blin, "Widely linear MVDR beamformers for the reception of an unknown signal corrupted by noncircular interferences," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5323–5336, Nov. 2007.

[20] P. Chevalier, J. P. Delmas, and A. Oukaci, "Properties, performance and practical interest of the widely linear MMSE beamformer for nonrectilinear signals," *Signal Process.*, vol. 97, no. 4, pp. 269–281, 2014.

[21] N. Song, W. U. Alokozai, R. C. de Lamare, and M. Haardt, "Adaptive widely linear reduced-rank beamforming based on joint iterative optimization," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 265–269, Mar. 2014.

[22] D. Xu, L. Huang, X. Xu, and Z. Ye, "Widely linear MVDR beamformers for noncircular signals based on time-averaged second-order noncircularity coefficient estimation," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3219–3227, Sep. 2013.

[23] P. Chargé, Y. Wang, and J. Saillard, "A non-circular sources direction finding method using polynomial rooting," *Signal Process.*, vol. 81, no. 8, pp. 1765–1770, 2001.

[24] P. Chevalier and F. Dupuy, "Widely linear Alamouti receiver for the reception of real-valued constellations corrupted by interferences—The Alamouti-SAIC/MAIC concept," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3339–3354, Jul. 2011.

[25] Y. Ding, N. Li, Y. Wang, and S. Feng, "Widely linear sphere decoding by exploiting the hidden properties of PSK signals," in *Proc. IEEE Global Commun. Conf.*, 2014, pp. 3209–3214.

[26] R. Rodríguez-Ávila, G. Núñez-Vega, R. Parra-Michel, M. E. Guzman, and D. L. Torres-Roman, "A frequency-selective I/Q imbalance analysis technique," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1854–1861, Apr. 2014.

[27] Z. Ahmadian and L. Lampe, "Robust design of widely linear pre-equalization filters for pre-Rake UWB systems," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4206–4217, Oct. 2013.

[28] D. Korpi, L. Anttila, V. Syrjala, and M. Valkama, "Widely-linear digital self-interference cancellation in direct-conversion full-duplex transceiver," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1674–1687, Sep. 2014.

[29] C. Hellings, M. Joham, and W. Utschick, "QoS feasibility in MIMO broadcast channels with widely linear transceivers," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1134–1137, Nov. 2013.

[30] P. Chevalier and F. Pipon, "New insights into optimal widely linear array receivers for the demodulation of BPSK, MSK, and GMSK signals corrupted by noncircular interferences-application to SAIC," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 870–883, Mar. 2006.

[31] G. Gelli, L. Paura, and A. R. P. Ragozini, "Blind widely linear multiuser detection," *IEEE Commun. Lett.*, vol. 4, no. 6, pp. 187–189, Jun. 2000.

[32] P. W. Wolniansky *et al.*, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proc. URSI Int. Symp. Signals, Syst., Electron. Conf. Proc.*, 1998, pp. 295–300.

[33] R. F. H. Fischer and C. Windpassinger, "Real versus complex-valued equalisation in V-BLAST systems," *Electron. Lett.*, vol. 39, no. 5, pp. 470–471, 2003.

[34] Y. Ding, Y. Wang, and J. F. Diouris, "Efficient detection algorithms for multiple-input/multiple-output systems by exploiting the non circularity of transmitted signal source," *IET Signal Process.*, vol. 5, no. 2, pp. 180–186, 2011.

[35] R. Arablouei, S. Werner, and K. Dogancay, "Estimating frequency of three-phase power systems via widely-linear modeling and total least-squares," in *Proc. 5th IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, 2013, pp. 464–467.

[36] N.-M. Jesús, F.-A. Rosa María, and R.-M. Juan Carlos, "A quaternion widely linear model for nonlinear Gaussian estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6414–6424, Dec. 2014.

[37] C. Paleologu, J. Benesty, and S. Ciochin, "Widely linear general Kalman filter for stereophonic acoustic echo cancellation," *Signal Process.*, vol. 94, no. 1, pp. 570–575, 2014.

[38] A. Papaioannou and S. Zafeiriou, "Principal component analysis with complex Kernel: The widely linear model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1719–1726, Sep. 2014.

[39] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 2030–2033, Aug. 1995.

[40] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov. 2014.

[41] A. Hedayat and A. Nosratinia, "Outage and diversity of linear receivers in flat-fading MIMO channels," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5868–5873, Dec. 2007.

[42] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc.*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[43] M. Rupp, C. Mecklenbrauker, and G. Gritsch, "High diversity with simple space time block-codes and linear receivers," in *Proc. IEEE Global Telecommun. Conf.*, 2003, vol. 1, pp. 302–306.

Authors' photographs and biographies not available at the time of publication.