

Received February 24, 2016, accepted March 12, 2016, date of publication April 5, 2016, date of current version April 21, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2551040

# MIMO-NOMA Design for Small Packet Transmission in the Internet of Things

ZHIGUO DING<sup>1,2</sup>, (Senior Member, IEEE), LINGLONG DAI<sup>3</sup>, (Senior Member, IEEE), AND H. VINCENT POOR<sup>1</sup>, (Fellow, IEEE)

<sup>1</sup>Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

<sup>2</sup>School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K.

<sup>3</sup>Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Corresponding author: Z. Ding (z.ding@lancaster.ac.uk)

The work of Z. Ding was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/N005597/1 and Grant H2020-MSCA-690750. The work of L. Dai was supported by the National Natural Science Foundation of China under Grant 61571270. The work of H. V. Poor was supported by the U.S. National Science Foundation under Grant CCF-1420575.

**ABSTRACT** A feature of the Internet of Things (IoT) is that some users in the system need to be served quickly for small packet transmission. To address this requirement, a new multiple-input multiple-output non-orthogonal multiple access (MIMO-NOMA) scheme is designed in this paper, where one user is served with its quality of service requirement strictly met, and the other user is served opportunistically by using the NOMA concept. The novelty of this new scheme is that it confronts the challenge that the existing MIMO-NOMA schemes rely on the assumption that users' channel conditions are different, a strong assumption which may not be valid in practice. The developed precoding and detection strategies can effectively create a significant difference between the users' effective channel gains, and therefore, the potential of NOMA can be realized even if the users' original channel conditions are similar. Analytical and numerical results are provided to demonstrate the performance of the proposed MIMO-NOMA scheme.

**INDEX TERMS** Non-orthogonal multiple access (NOMA), multiple-input multiple-output (MIMO), QR decomposition, MIMO precoding, and power allocation.

## I. INTRODUCTION

The fifth generation (5G) of mobile communications has been envisioned to enable the future Internet of Things (IoT); however, supporting the IoT functionality in 5G networks is challenging since connecting billions of smart IoT devices with diversified quality of service (QoS) requirements is not a trivial task, given the constraint of scarce bandwidth [1]. Non-orthogonal multiple access (NOMA) provides a promising solution to provide massive connectivity by efficiently using the available bandwidth resources [2], and has consequently been included in 3GPP long term evolution (LTE) [3].

The key idea of NOMA is to ask the users to share the same resources, such as frequency channels, time slots, and spreading codes, whereas the power domain is used for multiple access. The performance of NOMA in scenarios with single-antenna users has been studied in [4] and [5]. Achieving user fairness with different channel state information (CSI) in NOMA systems has been addressed in [6], and the impact of user pairing on NOMA has been investigated in [7]. The application of simultaneous wireless information and power transfer to NOMA has been considered in [8].

Since the use of multiple-input multiple-output (MIMO) techniques brings an extra dimension for further performance improvements, the study of the combination of MIMO and NOMA has received considerable attention recently. In [9] and [10], the scenario in which each user has a single antenna has been considered, and various algorithmic frameworks for optimizing the design of beamforming in the NOMA transmission system have been proposed. The sum rate has been used as an objective function in [11] and [12] to formulate various optimization problems in MIMO-NOMA scenarios. In [13] a zero-forcing based MIMO-NOMA transmission scheme was proposed without requiring the full CSI at the transmitter. A signal alignment based precoding scheme was developed in [14], and it requires fewer antennas at the users compared to the scheme proposed in [13]. A more detailed literature review can be found in [15].

In this paper, we consider a MIMO-NOMA downlink transmission scenario with one transmitter sending data to two users, e.g., an access point is serving two IoT devices. The feature of the IoT that users have diversified QoS requirements is used for the design of the

MIMO-NOMA transmission. Particularly, we consider that user 1 needs to be served quickly for small packet transmission, i.e., with a low targeted data rate, and user 2 is to be served with the best effort. Take intelligent transportation as an example, where user 1 can be a vehicle receiving incident warning information which is contained within a few bytes only, but is time-critical. User 2 can be another vehicle that wishes to perform some background tasks, such as downloading multimedia files. The use of NOMA prevents the drawback of conventional orthogonal multiple access (OMA) that user 1 whose targeted data rate is small is served with a dedicated orthogonal channel use. With NOMA, the users are served using the same time/frequency/code resources, which means that the bandwidth resources which are solely occupied by user 1 in the case of OMA can be released to user 2 in NOMA.

Most existing NOMA schemes rely on a key assumption that the users' channel conditions are very different. Take the MIMO-NOMA schemes proposed in [11] and [13] as examples. Within the NOMA user pair, one user is assumed to be deployed close to the base station, and the other is far away from the base station. As shown in [7], this difference in users' channel conditions is crucial for realizing the potential of NOMA. However, in practice, it is very likely that the users who want to participate in NOMA have similar channel conditions. Take our considered IoT scenario as an example, where the two users are categorized by their QoS requirements, not by their channel conditions. It is important to point out that the situation in which users have similar channel conditions can make the benefits of implementing many existing NOMA schemes marginal.

The main contribution of this paper is to design two sets of system parameters, precoding and power allocation coefficients, in order to ensure that the potential of NOMA can be realized even if the users' channel conditions are similar. *Firstly*, the precoding matrix at the base station is designed to degrade the effective channel gains at user 1 and improve the effective channel gains at user 2, at the same time. As a result, the users' effective channel conditions become very different, an ideal situation for the application of NOMA. The reason to degrade user 1's channel condition is that this user is regarded as an IoT user to be served with a low data rate, and therefore, a weaker channel condition could still accommodate this low data rate. *Secondly*, the power allocation coefficients are carefully designed to ensure that user 1's QoS requirements can still be met with its degraded channel conditions. Two types of power allocation policies are developed in this paper. One is to meet user 1's QoS requirements in the long term, e.g., its targeted outage probability can be satisfied. The other is in which to realize user 1's QoS requirements instantaneously, e.g., the power allocation coefficients are designed to realize its targeted data rate for each channel realization.

Analytical results are developed to better demonstrate the performance of the proposed MIMO-NOMA scheme. With the long term power allocation policy, the developed analytical results show that user 1's targeted outage probability can

be strictly guaranteed, and the diversity gains at user 2 are the same as that in the case when user 2 is served alone. With the short term power allocation policy, user 1's outage experience is the same as that in the case when all the power is given to user 1, and the diversity gain achieved at user 2 is always one. Therefore, between the two power allocation policies, user 2 prefers the long term one since the diversity gain it can obtain is larger. However, the short term power allocation policy can ensure that user 1's QoS requirement is met instantaneously, a property important to those safety-critical and real-time applications in the IoT.

## II. SYSTEM MODEL

Consider a MIMO-NOMA downlink transmission scenario with one base station and two users, in which the base station is equipped with  $M$  antennas and each user is equipped with  $N$  antennas. The  $N \times M$  channel matrices of the two users are denoted by  $\mathbf{H}_1$  and  $\mathbf{H}_2$ , respectively. Elements of the channel matrices are independent and identically complex Gaussian distributed with zero means and unit variances. In this paper, we focus on the scenario without path loss, i.e., the two users have similar channel conditions, which is a challenging situation in which to realize the potential of NOMA. Furthermore, we assume  $M \geq N$ , a scenario in which existing MIMO-NOMA schemes, such as the ones in [13] and [14], cannot work properly.

Without loss of generality, we assume that user 1 needs to be connected quickly to transmit small packets. For example, this user can be an IoT device that needs to be served with a small predefined data rate. The base station will transmit the following vector:

$$\mathbf{x} = \mathbf{P}\mathbf{s}, \tag{1}$$

where  $\mathbf{P}$  is an  $M \times N$  precoding matrix. The information bearing vector  $\mathbf{s}$  is constructed by using the NOMA approach as follows:

$$\mathbf{s} = [\alpha_1 s_1 + \beta_1 w_1 \quad \cdots \quad \alpha_N s_N + \beta_N w_N]^T, \tag{2}$$

where  $s_i$  is the  $i$ -th stream transmitted to user 1,  $\alpha_i$  is the power allocation coefficient for  $s_i$ ,  $w_i$  and  $\beta_i$  are defined similarly, and  $\alpha_i^2 + \beta_i^2 = 1$ .

As can be seen from (1) and (2), there are two sets of parameters to be designed, the precoding matrix  $\mathbf{P}$  and the power allocation coefficients  $\alpha_i$  ( $\beta_i$ ). The aim of the proposed design is to realize two goals simultaneously. One is to meet the QoS requirement at user 1 strictly and the other is to improve user 2's experience in an opportunistic manner. Alternatively, one can view the addressed NOMA scenario as a special case of cognitive radio networks, where user 1 is a primary user whose QoS requirements need to be satisfied strictly and user 2 is served opportunistically [7].

Assume that the QR decomposition of user 2's channel matrix,  $\mathbf{H}_2$ , is given by

$$\mathbf{H}_2^H = \mathbf{Q}_2 \tilde{\mathbf{R}}_2, \tag{3}$$

where  $\mathbf{Q}_2$  is an  $M \times M$  unitary matrix, and  $\tilde{\mathbf{R}}_2$  is an  $M \times N$  matrix obtained from the QR decomposition. Define  $\mathbf{V}_2$  as an  $M \times N$  matrix collecting the  $N$  left columns of  $\mathbf{Q}_2$ , and  $\mathbf{R}_2$  is an  $N \times N$  upper submatrix of  $\tilde{\mathbf{R}}_2$ . From the QR decomposition, we know that  $\mathbf{H}_2^H = \mathbf{V}_2 \mathbf{R}_2$ . The precoding matrix  $\mathbf{P}$  is set as  $\mathbf{P} = \mathbf{V}_2$ , which is to improve the signal strength at user 2. As can be seen from the following subsection, this choice of the precoding matrix also degrades the channel conditions at user 1, which makes user 1 analogous to a cell edge user in a conventional NOMA setup.

User 2's observation can be expressed as follows:

$$\mathbf{y}_2 = \mathbf{R}_2^H \mathbf{s} + \mathbf{n}_2, \quad (4)$$

where  $\mathbf{n}_2$  is the noise vector. Since  $\mathbf{R}_2^H$  is a lower triangular matrix, successive interference cancellation (SIC) can be carried out to cancel inter-layer interference (between  $w_i$  and  $w_j$ ,  $i \neq j$ ) and intra-layer interference (between  $s_i$  and  $w_i$ ). Particularly, suppose that  $s_j$  and  $w_j$  from the previous layers,  $j < i$ , are decoded successfully, whose outage probability will be included for the calculation of the overall probability in the next section. User 2 can decode the message intended for user 1 at the  $i$ -th layer,  $s_i$ , with the following signal-to-interference-plus-noise ratio (SINR):

$$\text{SINR}_{2,i} = \frac{\alpha_i^2 [\mathbf{R}_2^H]_{i,i}^2}{\beta_i^2 [\mathbf{R}_2^H]_{i,i}^2 + \frac{1}{\rho}}, \quad (5)$$

where  $\rho$  denotes the transmit signal-to-noise ratio (SNR) and  $[\mathbf{A}]_{i,j}$  denotes the element in the  $i$ -th row and the  $j$ -th column of a matrix  $\mathbf{A}$ . Denote the targeted data rate of user  $m$  at the  $i$ -th layer by  $R_{m,i}$ . Provided that  $\log(1 + \text{SINR}_{2,i}) > R_{1,i}$ , user 2 can successfully remove user 1's message,  $s_i$ , from its  $i$ -th layer, and its own message can be decoded with the following SNR:

$$\text{SNR}_{2,i} = \rho \beta_i^2 [\mathbf{R}_2^H]_{i,i}^2. \quad (6)$$

User 1's observation is given by

$$\mathbf{y}_1 = \mathbf{H}_1 \mathbf{P} \mathbf{s} + \mathbf{n}_1, \quad (7)$$

where  $\mathbf{n}_1$  is an  $N \times 1$  noise vector. Analogously to the cell edge user in a conventional NOMA network, user 1 is not to decode  $w_i$ , which means that the use of QR based detection will result in significant performance loss, as discussed in Section III-C. Therefore, zero forcing is applied at user 1. Particularly, the system model at user 1 can be written as follows:

$$(\mathbf{H}_1 \mathbf{V}_2)^\dagger \mathbf{y}_1 = \mathbf{s} + (\mathbf{H}_1 \mathbf{V}_2)^\dagger \mathbf{n}_1, \quad (8)$$

where  $(\mathbf{H}_1 \mathbf{V}_2)^\dagger = (\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1} \mathbf{V}_2^H \mathbf{H}_1^H$ . It is worth pointing out that the dimension of  $\mathbf{V}_2$  is  $M \times N$ , and therefore  $\mathbf{H}_1 \mathbf{V}_2$  is an  $N \times N$  square matrix, which means  $(\mathbf{H}_1 \mathbf{V}_2)^\dagger = (\mathbf{H}_1 \mathbf{V}_2)^{-1}$ . It is assumed here that the channel matrices are of full rank. As a result, user 1 can decode its message at the  $i$ -th layer with the following SINR:

$$\text{SINR}_{1,i} = \frac{\alpha_i^2 z_i}{\beta_i^2 z_i + \frac{1}{\rho}}, \quad (9)$$

where  $z_i = \frac{1}{[(\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1}]_{i,i}}$ .

### A. IMPACT OF THE PROPOSED PRECODING SCHEME

The two users' experiences with the proposed precoding scheme are different. According to the previous discussions, the reception reliability at user 2 is determined by the parameter  $x_i$ , where  $x_i \triangleq [\mathbf{R}_2^H]_{i,i}^2$ . Denote an  $M \times (M - N)$  complex Gaussian matrix that is independent of  $\mathbf{H}_2$  by  $\mathbf{B}$ . The QR decomposition of  $[\mathbf{H}_2^H \ \mathbf{B}]$  is given by

$$[\mathbf{H}_2^H \ \mathbf{B}] = \mathbf{Q}_2 \tilde{\mathbf{R}}_2 = [\mathbf{V}_2 \ \bar{\mathbf{V}}_2] \underbrace{\begin{bmatrix} \mathbf{R}_2 & \mathbf{C} \\ \mathbf{0}_{(M-N) \times N} & \mathbf{D} \end{bmatrix}}_{\tilde{\mathbf{R}}_2}, \quad (10)$$

where  $\tilde{\mathbf{R}}$  is an  $M \times M$  upper triangular matrix,  $\bar{\mathbf{V}}_2$  is a submatrix of  $\mathbf{Q}_2$ ,  $\mathbf{C}$  and  $\mathbf{D}$  are the submatrices of  $\tilde{\mathbf{R}}_2$ . According to [16], the elements of  $\tilde{\mathbf{R}}$  are independent, and the square of the  $i$ -th element on its diagonal follows the chi-square distribution with  $2(M - i + 1)$  degrees of freedom, i.e.,  $f_{x_i}(x) = \frac{x^{M-i}}{(M-i)!} e^{-x}$ . Therefore, more antennas at the base station can improve the receive signal strength at user 2 which is a function of  $[\mathbf{R}_2^H]_{i,i}^2$ .

On the other hand, the reception reliability at user 1 is degraded due to the use of the precoding matrix,  $\mathbf{P}$ , as explained in the following. Note that  $\mathbf{V}_2$  is a unitary matrix obtained from the QR decomposition based on  $\mathbf{H}_2$ . Because  $\mathbf{H}_1$  and  $\mathbf{H}_2$  are independent, and also by using the fact that a unitary transformation of a Gaussian matrix does not alter its statistical properties,  $\mathbf{H}_1 \mathbf{P}$  is still an  $N \times N$  complex Gaussian matrix, which means that the use of the proposed precoding matrix *shrinks* user 1's channel matrix from an  $M \times N$  complex Gaussian matrix to another complex Gaussian matrix with smaller size. Note that the probability density function (pdf) of the effective channel gain,  $\frac{1}{[(\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1}]_{i,i}}$ ,

is given by

$$f_{\frac{1}{[(\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1}]_{i,i}}}(z) = e^{-z}, \quad (11)$$

which is no longer a function of  $M$ .

The impact of the proposed precoding scheme can be clearly illustrated by using the following extreme example. Consider a special case with  $N = 1$ , where the channel matrices become  $1 \times M$  vectors, denoted by  $\mathbf{h}_1$  and  $\mathbf{h}_2$ , respectively. After applying  $\mathbf{P}$ , the effective channel gain at user 2 is  $|\mathbf{h}_2|^2$  which becomes stronger by increasing  $M$ . On the other hand, the effective channel gain at user 1 is always exponentially distributed, and the use of more antennas at the base station does not improve the transmission reliability at user 1.

### B. POWER ALLOCATION POLICIES

Because the precoding matrix degrades user 1's channel conditions, the power allocation coefficients  $\alpha_i$  ( $\beta_i$ ) need to be carefully designed to ensure that user 1's QoS requirements are met, which motivates the following two power allocation policies.

1) POWER ALLOCATION POLICY I

This approach is to meet user 1’s QoS requirements in the long term. Recall that the targeted data rate for user 1 to decode its message at the  $i$ -th layer ( $s_i$ ) is denoted by  $R_{1,i}$ . As a result, the outage probability for user 1 to decode  $s_i$  is given by

$$P_{1,i}^o \triangleq P(\log(1 + \text{SINR}_{1,i}) < R_{1,i}). \quad (12)$$

The power allocation coefficients,  $\alpha_i$  and  $\beta_i$ , are designed to satisfy the following constraint:

$$P_{1,i}^o \leq P_{1,i,\text{target}}, \quad (13)$$

where  $P_{1,i,\text{target}}$  denotes the targeted outage probability. A closed-form expression for the power allocation coefficients is given in Section III-A. The advantage of this type of power allocation is that there is no need to update power allocation coefficients frequently, but it cannot satisfy user 1’s QoS requirements instantaneously.

2) POWER ALLOCATION POLICY II

This approach is to meet user 1’s QoS requirements instantaneously. This type of power allocation is quite similar to the cognitive radio inspired power allocation policy proposed in [7]. Particularly, the power allocation coefficients are defined to ensure that the targeted data rate of user 1 is met instantaneously, i.e.,

$$\log(1 + \text{SINR}_{1,i}) \geq R_{1,i}. \quad (14)$$

By defining  $\epsilon_{k,i} = 2^{R_{k,i}} - 1$ , the above constraint yields the following power allocation policy:

$$\beta_i^2 = \max \left\{ 0, \frac{z_i - \frac{\epsilon_{1,i}}{\rho}}{z_i(1 + \epsilon_{1,i})} \right\}. \quad (15)$$

The above policy is sufficient for those scenarios addressed in [7] and [13], where users are *ordered* according to their channel conditions. For the scenario addressed in this paper,  $z_i < x_i$  does not always hold, and it is possible that the effective channel gain of user 1 is stronger. If  $z_i > x_i$ , the value of  $\beta_i$  in (15) is a very poor choice for user 2, as it is guaranteed that SIC at user 2 will fail. To ensure that SIC at user 2 can be carried out successfully, we revise the power allocation strategy as follows:

$$\beta_i^2 = \min \left\{ \max \left\{ 0, \frac{z_i - \frac{\epsilon_{1,i}}{\rho}}{z_i(1 + \epsilon_{1,i})} \right\}, \max \left\{ 0, \frac{x_i - \frac{\epsilon_{1,i}}{\rho}}{x_i(1 + \epsilon_{1,i})} \right\} \right\}. \quad (16)$$

This is to ensure that user 1’s message can be decoded by both users with the best effort. In an extreme case with  $x_i \rightarrow 0$  and a fixed  $z_i$ ,  $\beta_i = 0$ , i.e., the cognitive radio user, user 2, will not be admitted.

Note that it is also possible to reverse the decoding order when  $z_i > x_i$ . In this case, we need to ensure that the

following two conditions are satisfied. One is to ensure that user 1 can decode the message intended for user 2,  $w_i$ ,

$$\log \left( 1 + \frac{z_i \beta_i^2}{z_i \alpha_i + \frac{1}{\rho}} \right) \geq R_{2,i}, \quad (17)$$

and the other is to ensure user 1 can decode its own message,

$$\log \left( 1 + \rho z_i \alpha_i^2 \right) \geq R_{1,i}. \quad (18)$$

As a result, user 1’s outage experience becomes a function of  $R_{2,i}$  which is user 2’s targeted data rate, since user 1 needs to decode user 2’s message first. When user 2’s targeted data rate is varying, there is uncertainty as to whether user 1’s QoS requirements can be met strictly. Therefore, this type of reverse NOMA decoding is not considered in this paper.

III. OUTAGE PERFORMANCE AT USER 1

User 1’s outage performance will be studied in the following two subsections with the two different power allocation policies.

A. POWER ALLOCATION POLICY I

Recall that the outage probability for user 1 to detect  $s_i$  can be expressed as follows:

$$P_{1,i}^o = P \left( x_i < \frac{\frac{\epsilon_{1,i}}{\rho}}{\alpha_i^2 - \beta_i^2 \epsilon_{1,i}} \right), \quad (19)$$

When power allocation policy I is adopted, the power allocation coefficients are set to meet user 1’s QoS requirements in the long term, which means that neither  $\alpha_i$  or  $\beta_i$  is a function of instantaneous channel gains. Therefore, the outage probability in this case is the cumulative distribution function (CDF) of  $\frac{1}{[(\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1}]_{i,i}}$ .

By using the pdf in (11), the outage probability can be expressed as follows:

$$P_{1,i}^o = 1 - e^{-\frac{\frac{\epsilon_{1,i}}{\rho}}{\alpha_i^2 - \beta_i^2 \epsilon_{1,i}}}. \quad (20)$$

In order to ensure  $P_{1,i}^o \leq P_{1,i,\text{target}}$ , the power allocation coefficients need to be set as follows:

$$\beta_i^2 = \frac{1 + \frac{\epsilon_{1,i}}{\rho \ln(1 - P_{1,i,\text{target}})}}{1 + \epsilon_{1,i}}. \quad (21)$$

Note that  $1 - P_{1,i,\text{target}} \leq 1$ , which means  $\ln(1 - P_{1,i,\text{target}}) \leq 0$ . Therefore, the choice of  $\beta_i$  in (21) is always smaller than or equal to one, i.e.,  $\beta_i \leq 1$ . In order to ensure the choice of  $\beta_i$  in (21) positive or equivalently  $\frac{1 + \frac{\epsilon_{1,i}}{\rho \ln(1 - P_{1,i,\text{target}})}}{1 + \epsilon_{1,i}} > 0$ , the following constraint is imposed on the targeted outage probability:

$$1 > P_{1,i,\text{target}} > 1 - e^{-\frac{\epsilon_{1,i}}{\rho}}. \quad (22)$$

We ignore the choice of  $P_{1,i,\text{target}} = 1$ , since this choice does not consider user 1’s QoS requirements. The righthand side of the above equation is a lower bound on the targeted



outage probability which is achieved by giving all the power to user 1. Or in other words, if the targeted outage probability is smaller than or equal to  $\left(1 - e^{-\frac{\epsilon_{1,i}}{\rho}}\right)$ , we will have  $\beta_i = 0$  and the addressed NOMA scenario is degraded to the case in which only user 1 is served. Therefore, in the remainder of this paper, it is assumed that the targeted outage probability is chosen to be larger than  $\left(1 - e^{-\frac{\epsilon_{1,i}}{\rho}}\right)$  when power allocation policy I is used, in order to avoid the trivial case of  $\beta_i = 0$ .

### B. POWER ALLOCATION POLICY II

While the power allocation coefficients are set to ensure  $\log(1 + \text{SINR}_{1,i}) \geq R_{1,i}$ , outage can still occur at user 1 since this ideal choice of power allocation might not be feasible due to deep fading, i.e., a situation with very small channel gains can result in  $\beta_i = 0$ . Rewrite the expression of  $\beta_i$  in (16) as follows:

$$\beta_i^2 = \min \left\{ \beta_{i,z}^2, \beta_{i,x}^2 \right\}, \quad (23)$$

where  $\beta_{i,z}^2 = \max \left\{ 0, \frac{z_i - \frac{\epsilon_{1,i}}{\rho}}{z_i(1 + \epsilon_{1,i})} \right\}$ ,  $\beta_{i,x}^2 = \max \left\{ 0, \frac{x_i - \frac{\epsilon_{1,i}}{\rho}}{x_i(1 + \epsilon_{1,i})} \right\}$ ,  $\alpha_{i,x}$  and  $\alpha_{i,z}$  are defined similarly. Note that when  $x_i > z_i$ ,  $\beta_{i,x} \geq \beta_{i,z}$ , otherwise  $\beta_{i,x} \leq \beta_{i,z}$ . With these definitions, the outage probability for user 1 to detect  $s_i$  can be expressed as follows:

$$\begin{aligned} P_{1,i}^0 &= \text{P} \left( \log(1 + \text{SINR}_{1,i}) < R_{1,i} \right) \\ &= \text{P} \left( \underbrace{x_i > z_i, \log \left( 1 + \frac{z_i \alpha_{i,z}^2}{z_i \beta_{i,z}^2 + \frac{1}{\rho}} \right)}_{T_1} < R_{1,i} \right) \\ &\quad + \text{P} \left( \underbrace{z_i > x_i, \log \left( 1 + \frac{z_i \alpha_{i,x}^2}{z_i \beta_{i,x}^2 + \frac{1}{\rho}} \right)}_{T_2} < R_{1,i} \right). \end{aligned} \quad (24)$$

First consider the case of  $x_i > z_i$ . If  $\beta_{i,z} \neq 0$ , we have  $\log \left( 1 + \frac{z_i \alpha_{i,z}^2}{z_i \beta_{i,z}^2 + \frac{1}{\rho}} \right) = R_{1,i}$ , which means that no outage occurs. Therefore, the outage event when  $x_i > z_i$  is due to  $\beta_i = 0$ , and therefore,  $T_1$  can be simplified as follows:

$$T_1 = \text{P} \left( x_i > z_i, z_i < \frac{\epsilon_{1,1}}{\rho} \right). \quad (25)$$

The second factor  $T_2$  can be expressed as follows:

$$\begin{aligned} T_2 &= \text{P} \left( x_i < z_i < \frac{\frac{\epsilon_{1,i}}{\rho}}{\alpha_{i,x}^2 - \beta_{i,x}^2 \epsilon_{1,i}} \right) \\ &= \text{P} \left( x_i < z_i < \frac{\frac{\epsilon_{1,i}}{\rho}}{1 - \max \left\{ 0, \frac{x_i - \frac{\epsilon_{1,i}}{\rho}}{x_i(1 + \epsilon_{1,i})} \right\}} (1 + \epsilon_{1,i}) \right). \end{aligned} \quad (26)$$

In order to explicitly show the outage events, the factor  $T_2$  can be written as follows:

$$\begin{aligned} T_2 &= \text{P} \left( x_i < z_i < \frac{\frac{\epsilon_{1,i}}{\rho}}{\min \left\{ 1, \frac{\frac{\epsilon_{1,i}}{\rho}}{x_i} \right\}} \right) \\ &= \text{P} \left( x_i < z_i < \max \left\{ \frac{\epsilon_{1,i}}{\rho}, x_i \right\} \right) \\ &= \text{P} \left( x_i < z_i < \frac{\epsilon_{1,i}}{\rho} \right) + \text{P} \left( x_i < z_i < x_i, \frac{\epsilon_{1,i}}{\rho} < x_i \right). \end{aligned} \quad (27)$$

Since the second probability in the above equation is zero, the overall outage probability can be calculated as follows:

$$\begin{aligned} P_{1,i}^0 &= \text{P} \left( x_i > z_i, z_i < \frac{\epsilon_{1,1}}{\rho} \right) + \text{P} \left( x_i < z_i < \frac{\epsilon_{1,i}}{\rho} \right) \\ &= \text{P} \left( z_i < \frac{\epsilon_{1,1}}{\rho} \right) = 1 - e^{-\frac{\epsilon_{1,1}}{\rho}}, \end{aligned} \quad (28)$$

which means that the diversity gain for user 1 to decode  $s_i$  is one.

*Remark 1:* Consider a benchmark scheme with  $\beta_i = 0$ , i.e., user 2 is not served at all. By using zero forcing detection at user 1's receiver, it is straightforward to show that the outage probability achieved by this benchmark scheme is exactly the same as the one in (28). The reason for this phenomenon is that the purpose of power allocation policy II is to ensure that the QoS requirements of user 1 are met instantaneously, while user 2 is served under the condition that the outage probability at user 1 is not degraded compared to the case with user 1 served alone.

*Remark 2:* Another interesting benchmark scheme is to consider that the precoding matrix is designed for user 1 by using the zero forcing approach, i.e.,  $\mathbf{P} = (\mathbf{H}_1^H \mathbf{H}_1)^{-1} \mathbf{H}_1^H$ . It is straightforward to show that this benchmark scheme achieves a diversity gain of  $(M - N + 1)$  for each stream, by following steps similar to those in [13]. This diversity gain loss is because the precoding matrix  $\mathbf{P}$  proposed in this paper shrinks user 1's channel matrix from an  $N \times M$  complex Gaussian matrix to an  $N \times N$  complex Gaussian matrix. This degradation is caused on purpose in order to ensure that the two users' channel conditions become very different.

### C. WHEN USER 1 ADOPTS THE QR BASED APPROACH

Instead of zero forcing, user 1 can also use the QR based approach for detection. In the following, we will show that the performance of the QR based approach is worse than that of the zero forcing one introduced in the previous section.

Suppose that the effective channel matrix at user 1 has the QR decomposition  $\mathbf{H}_1 \mathbf{V}_2 = \mathbf{Q}_1 \mathbf{R}_1$ , and therefore, the observation at user 1 can be expressed as follows:

$$\mathbf{Q}_1^H \mathbf{y}_1 = \mathbf{R}_1 \mathbf{s} + \mathbf{Q}_1^H \mathbf{n}_1. \quad (29)$$

Recall that  $\mathbf{H}_1 \mathbf{V}_2$  is an  $N \times N$  complex Gaussian matrix. Therefore,  $[\mathbf{R}_1]_{i,i}^2$  is chi-squared distributed with  $2(N - i + 1)$  degrees of freedom. Unlike user 2, user 1 needs to decode

the  $i$ -th stream first before decoding the  $j$ -th stream,  $N \geq i > j \geq 1$ , since  $\mathbf{R}_1$  is an upper triangular matrix and  $\mathbf{R}_2^H$  is a lower triangular matrix. Since user 1 does not need to decode the messages intended for user 2, the system model for the  $i$ -th stream at user 1 can be rewritten as follows:

$$\tilde{y}_{1,i} = [\mathbf{R}_1]_{i,i}\alpha_i s_i + [\mathbf{R}_1]_{i,i}\beta_i w_i + \sum_{j=i+1}^N ([\mathbf{R}_1]_{i,j}\alpha_j s_j + [\mathbf{R}_1]_{i,j}\beta_j w_j) + n_{1,i}, \quad (30)$$

where  $\tilde{y}_{1,i}$  is the  $i$ -th element of  $\mathbf{Q}_1^H \mathbf{y}_1$  and  $n_{1,i}$  is defined similarly. Consider an ideal case in which  $s_j$  has been decoded correctly. By using this assumption, the outage probability at user 1 can be lower bounded as follows:

$$\begin{aligned} P_{1,i}^o &\geq \mathbb{P} \left( \frac{[\mathbf{R}_1]_{i,i}^2 \alpha_i^2}{[\mathbf{R}_1]_{i,i}^2 \beta_i^2 + \sum_{j=i+1}^N [\mathbf{R}_1]_{i,j}^2 \beta_j^2 + \frac{1}{\rho}} < \epsilon_{1,i} \right) \\ &\geq \mathbb{P} \left( \frac{[\mathbf{R}_1]_{i,i}^2 \alpha_i^2}{\sum_{j=i+1}^N [\mathbf{R}_1]_{i,j}^2 \beta_j^2 + \frac{1}{\rho}} < \epsilon_{1,i} \right), \end{aligned} \quad (31)$$

where  $1 \leq i < N$ . In order to gain some insight, we focus on the case with power allocation policy I, and assume  $\beta_j = \beta_i$ , for  $i \neq j$ . Define  $u_i = \sum_{j=i+1}^N [\mathbf{R}_1]_{i,j}^2$ . According to [16], the entries of  $\mathbf{R}_1$  are independent, and  $[\mathbf{R}_1]_{i,j}^2$ , with  $i < j$  is exponentially distributed, which means  $u_i$  is chi-square distributed with  $2(N - i)$  degrees of freedom, i.e.,  $f_{u_i}(u) = \frac{u^{N-i-1}}{(N-i-1)!} e^{-u}$ . It is straightforward to verify that user 1's outage probability becomes a non-zero constant, regardless of how large the SNR is. Since  $P_{1,i}^o$  is lower bounded by a non-zero constant, this means that, when the QR based approach is used, the outage probability at user 1 never goes to zero, even if the transmission power becomes infinite. Recall that the use of zero forcing can effectively cancel the inter-layer interference at user 1. For this reason, only zero forcing detection is considered at user 1 in this paper.

#### IV. OUTAGE PERFORMANCE AT USER 2

Since user 2 experiences differently with different power allocation policies, two subsections are provided in the following to study the two scenarios.

##### A. POWER ALLOCATION POLICY I

Recall that SIC is carried out at user 2 to remove both intra-layer and inter-layer interference. The outage event for user 2 to decode its own message at the  $i$ -th layer can be expressed as follows:

$$\mathcal{O}_{2,i} \triangleq \bigcup_{m \in \{1, \dots, i\}} \tilde{\mathcal{O}}_{2,m},$$

where  $\tilde{\mathcal{O}}_{2,m}$  denotes the event that user 2 cannot successfully decode the messages at the  $m$ -th layer,  $s_m$  and  $w_m$ , while all the messages in the previous layers,  $s_n$  and  $w_n$ , for  $1 \leq n < m$ , can be decoded correctly. Note that  $\tilde{\mathcal{O}}_{2,m} \cap \tilde{\mathcal{O}}_{2,n} = \emptyset$ , for  $m \neq n$ .

Since there are two messages at each layer, the outage event  $\tilde{\mathcal{O}}_{2,m}$  can be further expressed as follows:

$$\tilde{\mathcal{O}}_{2,m} = \bar{E}_{m,1} \bigcup \bar{E}_{m,2},$$

where the two events are defined as follows:

- $\bar{E}_{m,1}$ : the event that user 2 cannot decode  $s_m$ , but can decode all the messages from the previous layers,  $s_n$  and  $w_n$ , for  $1 \leq n \leq (m - 1)$ ;
- $\bar{E}_{m,2}$ : the event that user 2 cannot decode  $w_m$ , but can decode  $s_m$ , as well as  $s_n$  and  $w_n$ , for  $1 \leq n \leq (m - 1)$ .

Note that  $\bar{E}_{m,1} \cap \bar{E}_{m,2} = \emptyset$ .

By using the above definitions, the outage probability for user 2 to decode its own message at the  $i$ -th layer can be expressed as follows:

$$P_{2,i}^o = \sum_{m=1}^i (\mathbb{P}(\bar{E}_{m,1}) + \mathbb{P}(\bar{E}_{m,2})).$$

The first type of outage probability  $\mathbb{P}(\bar{E}_{m,1})$  can be expressed as follows:

$$\begin{aligned} \mathbb{P}(\bar{E}_{m,1}) &= \mathbb{P}(\log(1 + \text{SINR}_{2,m'}) < R_{1,m}, \\ &\quad \log(1 + \text{SINR}_{2,n'}) > R_{1,n}, \\ &\quad \log(1 + \text{SNR}_{2,n}) > R_{2,n}, \\ &\quad \forall n \in \{1, \dots, m - 1\}). \end{aligned} \quad (32)$$

Similarly the outage probability  $\mathbb{P}(\bar{E}_{m,2})$  can be expressed as follows:

$$\begin{aligned} \mathbb{P}(\bar{E}_{m,2}) &= \mathbb{P}(\log(1 + \text{SNR}_{2,m}) < R_{2,m}, \\ &\quad \log(1 + \text{SINR}_{2,m'}) > R_{1,m}, \\ &\quad \log(1 + \text{SINR}_{2,n'}) > R_{1,n}, \\ &\quad \log(1 + \text{SNR}_{2,n}) > R_{2,n}, \\ &\quad \forall n \in \{1, \dots, m - 1\}). \end{aligned} \quad (33)$$

Note that for the case of  $m = 1$ , the above outage probabilities can be simplified as  $\mathbb{P}(\bar{E}_{1,1}) = \mathbb{P}(\log(1 + \text{SINR}_{2,1'}) < R_{1,1})$  and  $\mathbb{P}(\bar{E}_{1,2}) = \mathbb{P}(\log(1 + \text{SNR}_{2,1}) < R_{2,1}, \log(1 + \text{SINR}_{2,1'}) > R_{1,1})$ .

By using the SINR expression in (5) and the above definitions,  $\mathbb{P}(\bar{E}_{m,1})$  can be expressed as follows:

$$\begin{aligned} \mathbb{P}(\bar{E}_{m,1}) &= \mathbb{P} \left( \log \left( 1 + \frac{\alpha_m^2 x_m}{\beta_m^2 x_m + \frac{1}{\rho}} \right) < R_{1,m}, \right. \\ &\quad \log \left( 1 + \frac{\alpha_n^2 x_n}{\beta_n^2 x_n + \frac{1}{\rho}} \right) > R_{1,n}, \\ &\quad \left. \log(1 + \beta_n^2 x_n) > R_{2,n}, \right. \\ &\quad \left. \forall n \in \{1, \dots, m - 1\} \right). \end{aligned} \quad (34)$$

Provided that power allocation policy I is used, the power coefficients are not functions of instantaneous channel gains, which yields the following:

$$\begin{aligned} \mathbb{P}(\bar{E}_{m,1}) &= \mathbb{P} \left( x_m < \frac{\frac{\epsilon_{1,m}}{\rho}}{\alpha_m^2 - \beta_m^2 \epsilon_{1,m}} \right) \\ &\quad \times \prod_{n=1}^{m-1} \mathbb{P} \left( x_n > \frac{\frac{\epsilon_{1,n}}{\rho}}{\alpha_n^2 - \beta_n^2 \epsilon_{1,n}}, x_n > \frac{\epsilon_{2,n}}{\rho \beta_n^2} \right), \end{aligned} \quad (35)$$

for  $\alpha_i^2 > \beta_i^2 \epsilon_{1,i}, \forall i \in \{1, \dots, m\}$ ; otherwise the probability is one. Note that (35) follows from the fact that the elements on the diagonal of  $\mathbf{R}_2, x_m$ , are independent. It can be verified that the choice of  $\beta_i$  in (21) can always ensure  $\alpha_i^2 > \beta_i^2 \epsilon_{1,i}$  since

$$\alpha_i^2 - \beta_i^2 \epsilon_{1,i} = 1 - \beta_i^2(1 + \epsilon_{1,i}) \geq -\frac{\epsilon_{1,i}}{\rho \ln(1 - P_{1,i,target})} > 0, \quad (36)$$

where  $P_{1,i,target} < 1$  as defined in (22).

By applying the pdf of  $x_m$ , the above probability can be obtained as follows:

$$P(\bar{E}_{m,1}) = \frac{\gamma(M - m + 1, \xi_m)}{(M - m)!} \times \prod_{n=1}^{m-1} \left[ 1 - \frac{\gamma(M - n + 1, \max\{\xi_n, \frac{\epsilon_{2,n}}{\rho\beta_n^2}\})}{(M - n)!} \right], \quad (37)$$

where  $\xi_m = \frac{\epsilon_{1,m}}{\alpha_m^2 - \beta_m^2 \epsilon_{1,m}}$  and  $\gamma(\cdot)$  denotes the incomplete gamma function [17].

Similarly the probability of  $P(\bar{E}_{m,2})$  can be calculated as follows:

$$P(\bar{E}_{m,2}) = \prod_{n=1}^{m-1} \left[ 1 - \frac{\gamma(M - n + 1, \max\{\xi_n, \frac{\epsilon_{2,n}}{\rho\beta_n^2}\})}{(M - n)!} \right] \times \frac{\left[ \gamma(M - m + 1, \frac{\epsilon_{2,m}}{\rho\beta_m^2}) - \gamma(M - m + 1, \xi_m) \right]}{(M - m)!}, \quad (38)$$

if  $\frac{\epsilon_{2,m}}{\rho\beta_m^2} \geq \xi_m$ ; otherwise  $P(\bar{E}_{m,2}) = 0$ .

Hence, the outage probability for user 2 to decode its own message at the  $i$ -th layer can be expressed as follows:

$$P_{2,i}^o = \sum_{m=1}^i \frac{\gamma(M - m + 1, \max\{\xi_m, \frac{\epsilon_{2,m}}{\rho\beta_m^2}\})}{(M - m)!} \times \prod_{n=1}^{m-1} \left[ 1 - \frac{\gamma(M - n + 1, \max\{\xi_n, \frac{\epsilon_{2,n}}{\rho\beta_n^2}\})}{(M - n)!} \right]. \quad (39)$$

At high SNR, i.e.,  $\rho$  approaches infinity, for a fixed  $P_{1,i,target}$  which is constrained as in (22) and not a function of  $\rho$ , it is straightforward to show that both  $\xi_m$  and  $\frac{\epsilon_{2,m}}{\rho\beta_m^2}$  approach zero. Therefore, the outage probability can be approximated as follows:

$$P_{2,i}^o = \sum_{m=1}^i \left[ 1 - e^{-\gamma_m} \left( \sum_{j=0}^{M-m} \frac{\gamma_m^j}{j!} \right) \right] \times \prod_{n=1}^{m-1} \left[ e^{-\gamma_n} \left( \sum_{j=0}^{M-n} \frac{\gamma_n^j}{j!} \right) \right] \approx \sum_{m=1}^i \frac{\gamma_m^{M-m+1}}{(M - m + 1)!} \approx \frac{\gamma_m^{M-i+1}}{(M - i + 1)!}, \quad (40)$$

where  $\gamma_m = \max\{\xi_m, \frac{\epsilon_{2,m}}{\rho\beta_m^2}\}$ . By using this high SNR approximation, one can readily find that the diversity gain for user 2 to decode  $w_i$  is  $(M - i + 1)$ .

*Remark 3:* In Section V, we will also use another choice of the targeted outage probability, i.e.,  $P_{1,i,target} = 1 - e^{-\frac{x\epsilon_{1,i}}{\rho}}$ , where  $x$  is not a function of  $\rho$  and  $x > 1$ . This targeted outage probability becomes a function of  $\rho$ . First note that this choice of  $P_{1,i,target}$  still fits the range defined in (22). Although this choice of  $P_{1,i,target}$  is a function of  $\rho$ , the approximation developed in (40) is still applicable, as explained in the following. With  $P_{1,i,target} = 1 - e^{-\frac{x\epsilon_{1,i}}{\rho}}$ , the power allocation coefficient  $\beta_i$  becomes  $\beta_i^2 = \frac{1 - \frac{1}{x}}{1 + \epsilon_{1,i}}$ . When  $\rho$  approaches infinity,  $\xi_m = \frac{x\epsilon_{1,i}}{\rho}$  approaches zero, and the same conclusion can be made for  $\frac{\epsilon_{2,m}}{\rho\beta_m^2}$ . As a result, the diversity order shown in (40) is also applicable to the case with  $P_{1,i,target} = 1 - e^{-\frac{x\epsilon_{1,i}}{\rho}}$ .

### B. POWER ALLOCATION POLICY II

With this type of power allocation, the power allocation coefficients become functions of the instantaneous channel gains, and this fact makes the evaluation of the outage probability very challenging, as explained in the following. First define  $y_{ii} = \left[ (\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1} \right]_{i,i}$ . As a result, the power allocation coefficient for user 2 can be expressed as follows:

$$\beta_i^2 = \max \left\{ 0, \min \left\{ \frac{y_{ii} \left( \frac{1}{y_{ii}} - \frac{\epsilon_{1,i}}{\rho} \right)}{(1 + \epsilon_{1,i})}, \frac{x_i - \frac{\epsilon_{1,i}}{\rho}}{x_i(1 + \epsilon_{1,i})} \right\} \right\}. \quad (41)$$

Even if we can reduce the expression of  $\beta$  to  $\beta_i^2 = \max \left\{ 0, \frac{y_{ii} \left( \frac{1}{y_{ii}} - \frac{\epsilon_{1,i}}{\rho} \right)}{(1 + \epsilon_{1,i})} \right\}$ , a policy conventionally used in [7], the power allocation coefficient  $\beta_i$  is still a function of  $y_{ii}$ , which means that the outage probability for user 2 to detect  $w_i$  can be written as follows:

$$P_{2,i}^o = \int \dots \int_{y_{11}, \dots, y_{ii}} \sum_{m=1}^i \left[ 1 - e^{-\gamma_m} \left( \sum_{j=0}^{M-m} \frac{\gamma_m^j}{j!} \right) \right] \times \prod_{n=1}^{m-1} \left[ e^{-\gamma_n} \left( \sum_{j=0}^{M-n} \frac{\gamma_n^j}{j!} \right) \right] \times f_{y_{11}, \dots, y_{ii}}(y_{11}, \dots, y_{ii}) dy_{11} \dots dy_{ii}, \quad (42)$$

where the outage probability expression in (40) is used, and  $f_{y_{11}, \dots, y_{ii}}(y_{11}, \dots, y_{ii})$  is the joint pdf of  $(y_{11}, \dots, y_{ii})$ . Recall that  $y_{ii}$  is the  $i$ -th element on the diagonal of the inverse Wishart matrix  $\mathbf{W}^{-1} \triangleq (\mathbf{V}_2^H \mathbf{H}_1^H \mathbf{H}_1 \mathbf{V}_2)^{-1}$ . Note that the joint pdf can be obtained by calculating the marginal pdf of  $\mathbf{W}^{-1}$  as follows [18]:

$$f_{y_{11}, \dots, y_{ii}}(y_1, \dots, y_i) = \int \dots \int_{y_j, \forall i \neq j} \frac{(\det \mathbf{W}^{-1})^{2N}}{\Gamma} e^{-\text{tr}(\mathbf{W}^{-1})} dy_{12} \dots dy_{N(N-1)} \quad (43)$$

where  $\Gamma = \pi^{\frac{N(N-1)}{2}} \prod_{j=1}^N \Gamma(N - j + 1)$ .

Because of the correlation among  $y_{ii}$  shown in (43) and the complicated form of the power allocation coefficients in (41), a closed-form expression for the outage probability for user 2 to decode  $w_i$  cannot be found. In the following, we will focus on the development of upper and lower bounds on the outage probability, which will be used for the analysis of the diversity gain achieved by the proposed MIMO-NOMA scheme at user 2.

1) UPPER AND LOWER BOUNDS ON THE OUTAGE PROBABILITY

By using the definitions provided in Section IV-A, the outage probability for user 2 to decode its own message at the  $i$ -th layer can be expressed as follows:

$$P_{2,i}^o = \sum_{m=1}^i (P(\bar{E}_{m,1}) + P(\bar{E}_{m,2})). \quad (44)$$

In the following, we first focus on the development of an upper bound on the outage probability. The probability,  $P(\bar{E}_{m,1})$ , can be upper bounded as follows:

$$P(\bar{E}_{m,1}) \leq P(\log(1 + \text{SINR}_{2,m'}) < R_{1,m}). \quad (45)$$

Similarly we can upper bound  $P(\bar{E}_{m,2})$  as follows:

$$P(\bar{E}_{m,2}) \leq P(\log(1 + \text{SNR}_{2,m}) < R_{2,m}, \log(1 + \text{SINR}_{2,m'}) > R_{1,m}). \quad (46)$$

By using the SINR expression in (5), the upper bound on  $P(\bar{E}_{m,1})$  can be expressed as follows:

$$P(\bar{E}_{m,1}) \leq \underbrace{P\left(z_m < x_m < \frac{\epsilon_{1,m}}{\alpha_m^2 - \beta_m^2 \epsilon_{1,m}}\right)}_{Q_1} + \underbrace{P\left(z_m > x_m, x_m < \frac{\epsilon_{1,m}}{\alpha_m^2 - \beta_m^2 \epsilon_{1,m}}\right)}_{Q_2}. \quad (47)$$

The reason to have the two probabilities,  $Q_1$  and  $Q_2$ , is that the power allocation coefficient  $\beta_i$  has different forms depending on the relationship between  $x_m$  and  $z_m$ .

By substituting the expression for  $\beta_i$  when  $z_m < x_m$ , the factor  $Q_1$  can be expressed as follows:

$$Q_1 = P\left(z_m < x_m < \frac{\epsilon_{1,m}}{1 - \beta_m^2(1 + \epsilon_{1,m})}\right) = P\left(z_m < x_m < \frac{\epsilon_{1,m}}{1 - \max\left\{0, \frac{(z_m - \epsilon_{1,m})}{z_m(1 + \epsilon_{1,m})}\right\}(1 + \epsilon_{1,m})}\right). \quad (48)$$

To simplify the outage probability, the max function needs to be removed, and we have the following:

$$Q_1 = P\left(z_m < x_m < \frac{\epsilon_{1,m}}{\min\left\{1, \frac{\epsilon_{1,m}}{z_m}\right\}}\right) = P\left(z_m < x_m < \max\left\{\frac{\epsilon_{1,m}}{\rho}, z_m\right\}\right) = P\left(z_m < x_m < \frac{\epsilon_{1,m}}{\rho}\right) + P\left(\frac{\epsilon_{1,m}}{\rho} < z_m < x_m < z_m\right). \quad (49)$$

Note that the second probability in the above equation is zero. Since an upper bound is of interest, we have

$$Q_1 \leq P\left(z_m < \frac{\epsilon_{1,m}}{\rho}\right) \sim \frac{1}{\rho}. \quad (50)$$

The factor  $Q_2$  can be calculated as follows:

$$Q_2 = P\left(z_m > x_m, x_m < \frac{\epsilon_{1,m}}{1 - \max\left\{0, \frac{(x_m - \frac{\epsilon_{1,m}}{\rho})}{x_m(1 + \epsilon_{1,m})}\right\}(1 + \epsilon_{1,m})}\right) = P\left(z_m > x_m, x_m < \max\left\{\frac{\epsilon_{1,m}}{\rho}, x_m\right\}\right). \quad (51)$$

By using the two possible choices of  $x_m$ , the above probability can be further upper bounded as follows:

$$Q_2 = P\left(z_m > x_m, x_m < \frac{\epsilon_{1,m}}{\rho}, \frac{\epsilon_{1,m}}{\rho} > x_m\right) \leq P\left(x_m < \frac{\epsilon_{1,m}}{\rho}\right) \sim \frac{1}{\rho^{M-m+1}}, \quad (52)$$

where the first equation follows from the fact that  $\max\left\{\frac{\epsilon_{1,m}}{\rho}, x_m\right\} = x_m$  for  $\frac{\epsilon_{1,m}}{\rho} < x_m$ , a situation in which user 2 can decode  $s_m$  for sure, i.e.,

$$P\left(z_m > x_m, x_m < \max\left\{\frac{\epsilon_{1,m}}{\rho}, x_m\right\}\right) = 0,$$

for  $\frac{\epsilon_{1,m}}{\rho} < x_m$ .

On the other hand,  $P(\bar{E}_{m,2})$  can be calculated as follows:

$$P(\bar{E}_{m,2}) \leq \underbrace{P\left(z_m < x_m, \frac{\epsilon_{1,m}}{\alpha_m^2 - \beta_m^2 \epsilon_{1,m}} < x_m < \frac{\epsilon_{2,m}}{\beta_m^2 \rho}\right)}_{Q_3} + \underbrace{P\left(z_m > x_m, \frac{\epsilon_{1,m}}{\alpha_m^2 - \beta_m^2 \epsilon_{1,m}} < x_m < \frac{\epsilon_{2,m}}{\beta_m^2 \rho}\right)}_{Q_4}. \quad (53)$$



The factor  $Q_3$  can be written as follows:

$$\begin{aligned}
Q_3 &= \mathbb{P} \left( \frac{\frac{\epsilon_{1,m}}{\rho}}{1 - \max \left\{ 0, \frac{\left( z_m - \frac{\epsilon_{1,m}}{\rho} \right)}{z_m(1+\epsilon_{1,m})} \right\}} < x_m \right. \\
&\quad \left. < \frac{\epsilon_{2,m}}{\max \left\{ 0, \frac{\left( z_m - \frac{\epsilon_{1,m}}{\rho} \right)}{z_m(1+\epsilon_{1,m})} \right\} \rho}, z_m < x_m \right) \\
&= \mathbb{P} \left( z_m < \frac{\epsilon_{1,m}}{\rho}, \frac{\epsilon_{1,m}}{\rho} < x_m, z_m < x_m \right) \\
&\quad + \mathbb{P} \left( z_m > \frac{\epsilon_{1,m}}{\rho}, z_m < x_m < \frac{\epsilon_{2,m}}{\frac{\left( z_m - \frac{\epsilon_{1,m}}{\rho} \right)}{z_m(1+\epsilon_{1,m})} \rho}, z_m < x_m \right) \\
&\leq \mathbb{P} \left( z_m < \frac{\epsilon_{1,m}}{\rho} \right) + \mathbb{P} \left( z_m < \frac{\epsilon_{2,m}}{\frac{\left( z_m - \frac{\epsilon_{1,m}}{\rho} \right)}{z_m(1+\epsilon_{1,m})} \rho} \right). \quad (54)
\end{aligned}$$

It is interesting to observe that the second probability in the above equation can be rewritten as follows:

$$\mathbb{P} \left( z_m < \frac{\epsilon_{2,m}}{\frac{\left( z_m - \frac{\epsilon_{1,m}}{\rho} \right)}{z_m(1+\epsilon_{1,m})} \rho} \right) = \mathbb{P} \left( z_m < \frac{\epsilon_{1,m} + \epsilon_{2,m} + \epsilon_{1,m}\epsilon_{2,m}}{\rho} \right). \quad (55)$$

Therefore, the factor  $Q_3$  can be upper bounded as follows:

$$\begin{aligned}
Q_3 &\leq \mathbb{P} \left( z_m < \frac{\epsilon_{1,m}}{\rho} \right) + \mathbb{P} \left( z_m < \frac{\epsilon_{1,m} + \epsilon_{2,m} + \epsilon_{1,m}\epsilon_{2,m}}{\rho} \right) \\
&\sim \frac{1}{\rho}. \quad (56)
\end{aligned}$$

Furthermore, the factor  $Q_4$  can be calculated as follows:

$$\begin{aligned}
Q_4 &= \mathbb{P} \left( \frac{\frac{\epsilon_{1,m}}{\rho}}{1 - \max \left\{ 0, \frac{\left( x_m - \frac{\epsilon_{1,m}}{\rho} \right)}{x_m(1+\epsilon_{1,m})} \right\}} \right. \\
&\quad \left. < x_m < \frac{\epsilon_{2,m}}{\max \left\{ 0, \frac{\left( x_m - \frac{\epsilon_{1,m}}{\rho} \right)}{x_m(1+\epsilon_{1,m})} \right\} \rho}, z_m > x_m \right) \\
&\stackrel{(a)}{=} \mathbb{P} \left( x_m < \frac{\epsilon_{1,m}}{\rho}, \frac{\epsilon_{1,m}}{\rho} < x_m, z_m > x_m \right) \\
&\quad + \mathbb{P} \left( x_m > \frac{\epsilon_{1,m}}{\rho}, x_m < \frac{\epsilon_{2,m}}{\frac{\left( x_m - \frac{\epsilon_{1,m}}{\rho} \right)}{x_m(1+\epsilon_{1,m})} \rho}, z_m > x_m \right) \\
&\leq \mathbb{P} \left( x_m < \frac{\epsilon_{1,m}}{\rho} \right) + \mathbb{P} \left( x_m < \frac{\epsilon_{2,m}}{\frac{\left( x_m - \frac{\epsilon_{1,m}}{\rho} \right)}{x_m(1+\epsilon_{1,m})} \rho} \right), \quad (57)
\end{aligned}$$

where the step (a) is due to the fact that, when  $x_m > \frac{\epsilon_{1,m}}{\rho}$  and  $z_m > x_m$ , user 2 can always decode  $s_m$ , since  $\log(1 + \text{SINR}_{2,m'}) = R_{1,m}$ .

According to (55), the factor  $Q_4$  can be upper bounded as follows:

$$\begin{aligned}
Q_4 &\leq \mathbb{P} \left( x_m < \frac{\epsilon_{1,m}}{\rho} \right) + \mathbb{P} \left( x_m < \frac{\epsilon_{1,m} + \epsilon_{2,m} + \epsilon_{1,m}\epsilon_{2,m}}{\rho} \right) \\
&\sim \frac{1}{\rho^{M-m+1}}. \quad (58)
\end{aligned}$$

By combining (44), (50), (52), (56) and (58), we can conclude that a lower bound on the diversity gain at user 2, obtained from the upper bound on the outage probability, is 1.

A lower bound on the outage probability can be obtained as follows:

$$\mathbb{P}_{2,i}^o \geq \mathbb{P}(\bar{E}_{1,2}) \geq Q_3, \quad (59)$$

by focusing the case of  $m = 1$ . Following (54), the factor  $Q_3$  with  $m = 1$  can be calculated as follows:

$$\begin{aligned}
Q_3 &\geq \mathbb{P} \left( z_1 < \frac{\epsilon_{1,1}}{\rho}, \frac{\epsilon_{1,1}}{\rho} < x_1 \right) \\
&= \mathbb{P} \left( z_1 < \frac{\epsilon_{1,1}}{\rho} \right) - \mathbb{P} \left( z_1 < \frac{\epsilon_{1,1}}{\rho}, x_1 < \frac{\epsilon_{1,1}}{\rho} \right) \\
&= \mathbb{P} \left( z_1 < \frac{\epsilon_{1,1}}{\rho} \right) \left( 1 - \mathbb{P} \left( x_1 < \frac{\epsilon_{1,1}}{\rho} \right) \right) \sim \frac{1}{\rho}, \quad (60)
\end{aligned}$$

since  $z_1$  is independent of  $x_1$ ,  $\mathbb{P} \left( z_1 < \frac{\epsilon_{1,1}}{\rho} \right) \sim \frac{1}{\rho}$  and  $\mathbb{P} \left( x_1 < \frac{\epsilon_{1,1}}{\rho} \right) \sim \frac{1}{\rho^M}$ .

Since both upper and lower bounds converge, we can conclude that the diversity gain for user 2 to decode  $w_i$  is one, when power allocation policy II is used.

*Remark 4:* Note that the diversity gain obtained above is the same as that at user 1, when power allocation policy II is used. This is consistent with the conclusion made in [7], where the diversity gain at the user with stronger channel conditions is determined by the channel conditions of its partner, when the cognitive radio inspired power allocation policy is applied.

## V. NUMERICAL STUDIES

In this section, the performance of the proposed MIMO-NOMA scheme is evaluated by using simulation results. We will first compare the proposed scheme with some existing MIMO-NOMA and MIMO-OMA schemes. Then, additional simulation results are provided to demonstrate the impact of different choices of the system parameters, where analytical results developed in the paper will also be verified.

### A. COMPARISON TO BENCHMARK SCHEMES

To simplify the simulation comparison, power allocation policy I is used in this subsection. The targeted data rates for two users are set as  $R_{1,i} = 1$  bit per channel user (BPCU) and  $R_{2,i} = 4$  BPCU, for all  $1 \leq i \leq N$ , respectively. The targeted outage probabilities for the two users

are set as  $P_{1,i,target} = 1 - e^{-\frac{2\epsilon_{1,i}}{\rho}}$ , and  $P_{2,i,target} = 1 - e^{-\frac{2\epsilon_{2,i}}{\rho}}$ , respectively. Using these targeted outage probabilities and the step in (21), the power allocation coefficients can be obtained. Note that these chosen  $P_{k,i,target}$  values are still within the range defined in (22). Since the use of power allocation policy I guarantees the QoS requirements at user 1, we will focus on the outage performance at user 2 in this subsection.

We first compare the proposed scheme to those MIMO-NOMA schemes developed in [13] and [14], which are termed ZF-NOMA and SA-NOMA, respectively. Since both schemes were proposed for scenarios with different system parameters, they need be tailored to the scenario addressed in this paper as explained in the following. Recall that ZF-NOMA proposed in [13] requires  $N \geq M$ , and SA-NOMA proposed in [14] requires  $N > \frac{M}{2}$ . In order to ensure that both schemes are applicable, we focus on a scenario with  $N = M$ , but it is important to point out that the scheme proposed in this paper is applicable to a scenario with a small  $N$ .

For ZF-NOMA, the precoding matrix is set as an identity matrix, i.e.,  $\mathbf{P} = \mathbf{I}_M$ , and both users use zero forcing for detection. The SINR for user 2 to decode the message intended to user 1 at the  $i$ -th layer can be written as  $SINR_{ZF,i} =$

$$\frac{\alpha_i^2}{\frac{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{i,i}}{\beta_i^2} + 1}.$$

If user 2 can decode its partner's message successfully, it can decode its own with the following SNR:

$$SNR_{ZF,i} = \frac{\rho \beta_i^2}{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{i,i}}.$$

In order to have a fair comparison,

for ZF-NOMA, the effective channel gains are ordered, i.e.,

$$\frac{1}{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{1,1}} \geq \dots \geq \frac{1}{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{M,M}}.$$

It is interesting to point out that for the case of  $M = N$ , SA-NOMA achieves the same performance as ZF-NOMA, as shown in the following. For SA-NOMA, signal alignment is used, where the two users' detection matrices,  $\mathbf{U}_1$  and  $\mathbf{U}_2$ , are obtained from the equation  $[\mathbf{H}_1^H \ -\mathbf{H}_2^H] [\mathbf{U}_1 \ \mathbf{U}_2]^H = \mathbf{0}_{M \times M}$ , where both detection matrices are  $N \times N$ . As a result, the users' effective channel matrices become the same, i.e.,  $\mathbf{U}_1 \mathbf{H}_1 = \mathbf{U}_2 \mathbf{H}_2$ . Therefore, the SINR for user 2 to decode the message intended for user 1 at the  $i$ -th layer can be written as  $SINR_{SA,i} = \frac{\alpha_i^2}{\beta_i^2 + \frac{1}{\rho} [(\mathbf{U}_2 \mathbf{H}_2)^{-1} \mathbf{U}_2 \mathbf{U}_2^H (\mathbf{U}_2 \mathbf{H}_2)^{-H}]_{i,i}} =$

$$SINR_{ZF,i},$$

where both  $\mathbf{H}_i$  and  $\mathbf{U}_i$  are assumed to be invertible. Similarly, provided that user 2 can decode its partner's message, it can decode its own with the following SNR:

$$SNR_{SA,i} = \frac{\rho \beta_i^2}{[(\mathbf{U}_2 \mathbf{H}_2)^{-1} \mathbf{U}_2 \mathbf{U}_2^H (\mathbf{U}_2 \mathbf{H}_2)^{-H}]_{i,i}} = SNR_{ZF,i}.$$

Therefore,

the two schemes achieve the same outage performance in the addressed scenario with  $M = N$ .

In Fig. 1, the performance of these MIMO-NOMA schemes is shown as functions of the transmit SNR.

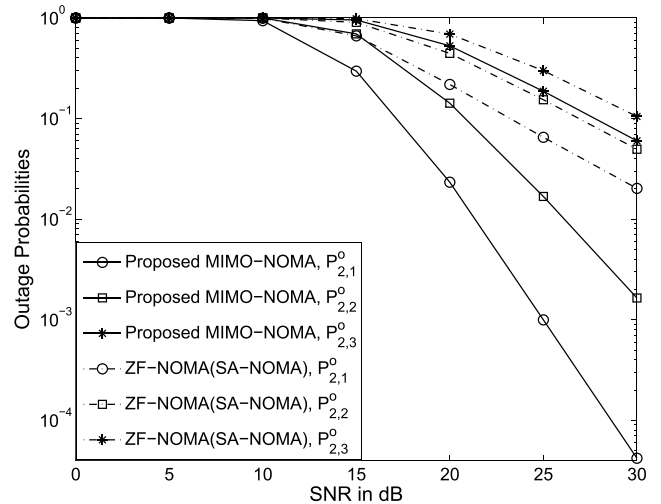


FIGURE 1. Comparison to the existing MIMO-NOMA schemes.  $M = N = 3$ .

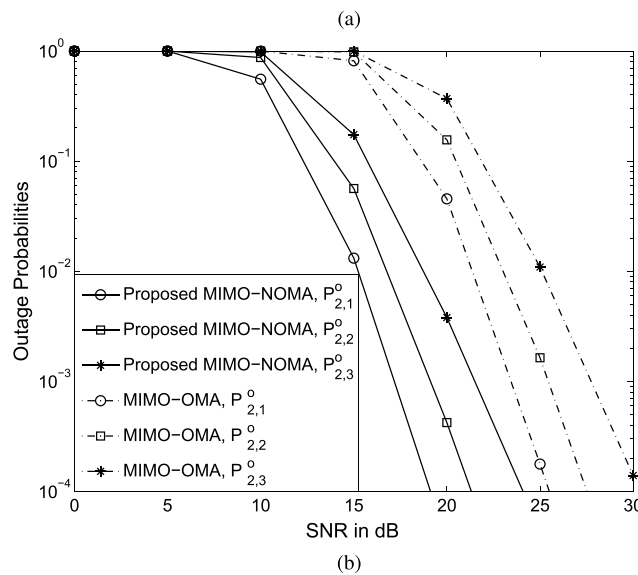
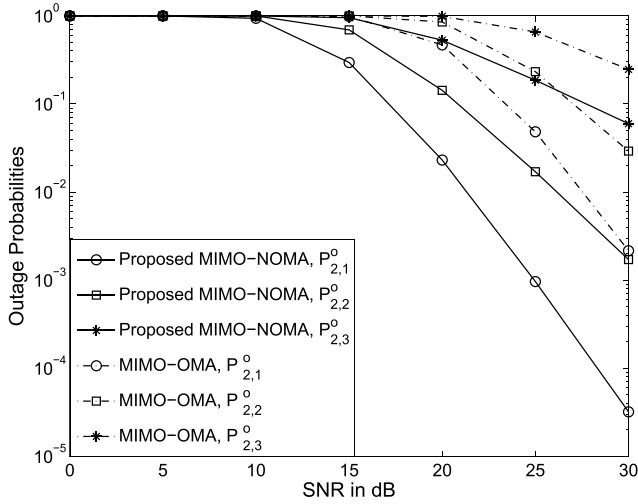
$P_{1,i,target} = 1 - e^{-\frac{2\epsilon_{1,i}}{\rho}}$ , and  $P_{2,i,target} = 1 - e^{-\frac{2\epsilon_{2,i}}{\rho}}$ .  $R_{1,i} = 1$  BPCU and  $R_{2,i} = 4$  BPCU,  $\forall 1 \leq i \leq N$ .

As can be seen from the figure, for all MIMO-NOMA schemes considered, the outage performance at the  $i$ -th layer is better than that at the  $j$ -th layer, for  $i < j$ , which can be explained as follows. For the proposed scheme, the effective channel gain at the  $i$ -th layer,  $[\mathbf{R}_2^H]_{i,i}^2$ , is statistically stronger than that at the  $j$ -th layer, since  $[\mathbf{R}_2^H]_{i,i}^2$  is chi-square distributed with  $2(M - i + 1)$  degrees of freedom. For the two existing MIMO-NOMA schemes, we have ordered the effective channel gains as  $\frac{1}{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{1,1}} \geq \dots \geq \frac{1}{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{M,M}}$ .

Furthermore, it is important to observe that at all layers, the proposed scheme outperforms the existing MIMO-NOMA schemes. Particularly, the figure demonstrates that for the proposed scheme, the slope of the outage probability curves is changing, which means change of the diversity gains. On the other hand, all the outage probability curves for the existing schemes have the same slope, which is mainly due to the correlation among the effective channel gains,  $\frac{1}{[\mathbf{H}_2^H \mathbf{H}_2]^{-1}]_{i,i}}$ .

OMA is another important benchmark scheme. Recall that in this paper, user 1 is viewed as a primary user in a conventional cognitive radio network. If OMA is used, the bandwidth resource allocated to user 1 cannot be reused. The use of NOMA means that user 2, which can be viewed as a secondary user, is admitted into the bandwidth occupied by user 1. Because the proposed power allocation policies can ensure that user 1's QoS requirements are met, whatever can be transmitted to user 2, such as  $R_{2,i}(1 - P_{2,i}^0)$ , will be a net performance gain over OMA. Or in other words, the benefit of the proposed MIMO-NOMA scheme over OMA is clear if we ask user 2, a user with stronger channel conditions, to be admitted into the bandwidth allocated to user 1.

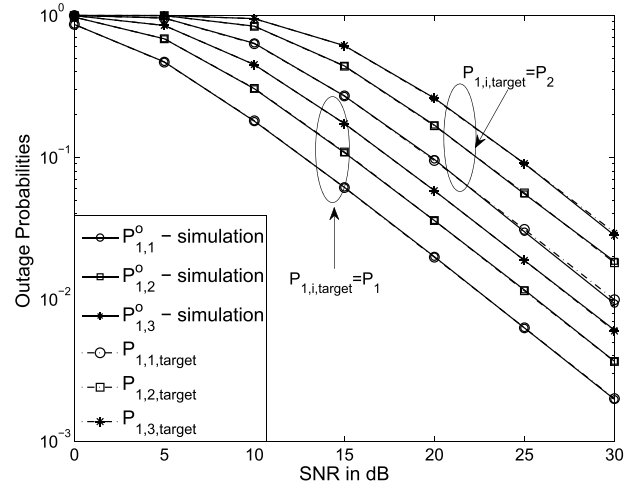
In the following, we consider a comparison that is more difficult for NOMA. Particularly, assume that there are two time



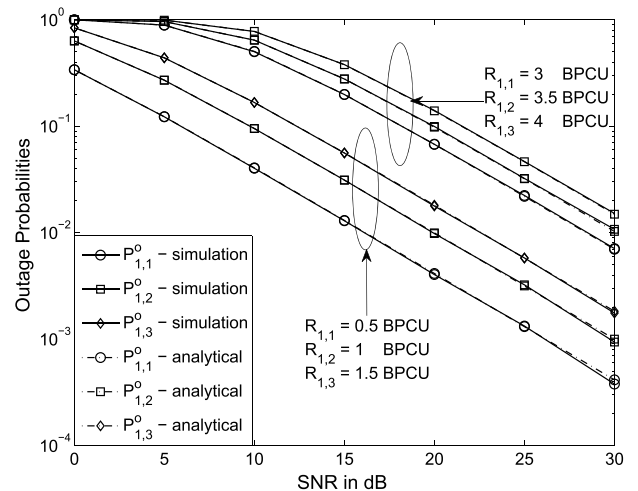
**FIGURE 2.** Comparison to MIMO-OMA.  $P_{1,i,target} = 1 - e^{-\frac{2\epsilon_{1,i}}{\rho}}$ , and  $P_{2,i,target} = 1 - e^{-\frac{2\epsilon_{2,i}}{\rho}}$ .  $R_{1,i} = 1$  BPCU and  $R_{2,i} = 4$  BPCU,  $\forall 1 \leq i \leq N$ . (a)  $M = N = 3$ . (b)  $M = 6$  and  $N = 3$ .

slots (or frequency-channels/spreading-codes). For OMA, time slot  $i$  is allocated to user  $i$ . For NOMA, the two users are served at the same time. Comparing NOMA to this type of OMA is challenging, since user 1, a user with weaker channel conditions, is admitted into the time slot allocated to user 2 and user 2 cannot enjoy the interference free communications experienced in the OMA case. As a result, the performance gain of NOMA over OMA becomes less obvious.

In Fig. 2, the performance of the proposed MIMO-NOMA scheme is compared to that of the MIMO-OMA scheme described above. Precoding for the considered MIMO-OMA scheme is designed by using the same QR approach as discussed in Section II. Particularly, during the second time slot, user 2 is served and the precoding matrix is designed according to the QR decomposition of  $\mathbf{H}_2$ , which means that the data rate for user 2 to decode its message at the  $i$ -th layer is  $\frac{1}{2} \log(1 + \rho[\mathbf{R}_2^H]_{i,i}^2)$ , where the factor  $\frac{1}{2}$  is due to the



**FIGURE 3.** Outage performance at user 1 with power allocation policy I.  $M = N = 3$ .  $P_1 = 1 - e^{-\frac{2\epsilon_{1,i}}{\rho}}$ , and  $P_2 = 1 - e^{-\frac{10\epsilon_{2,i}}{\rho}}$ .  $R_{1,1} = 1$  BPCU,  $R_{1,2} = 1.5$  BPCU, and  $R_{1,3} = 2$  BPCU.

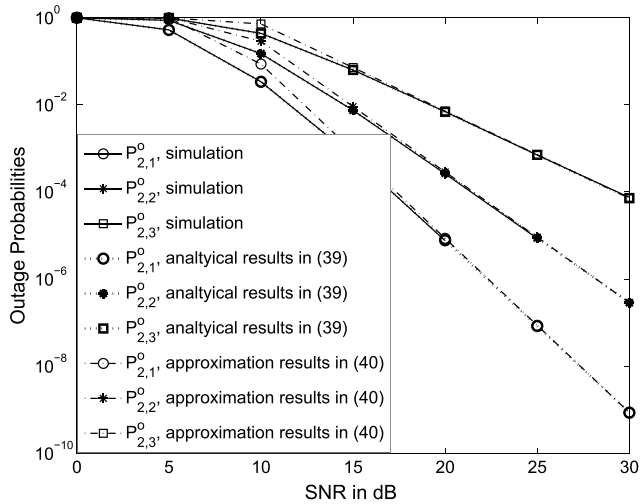


**FIGURE 4.** Outage performance at user 1 with power allocation policy II.  $M = N = 3$ . The analytical results are based on (28).

use of OMA. As can be seen from the two sub-figures in Fig. 2, the proposed MIMO-NOMA scheme can achieve better outage performance compared to MIMO-OMA, and the performance gap between the two schemes can be increased by introducing more antennas at the base station, i.e., increasing  $M$ . An interesting observation is that the slope of the outage probability curves for MIMO-NOMA is the same as that for MIMO-OMA, which means that the diversity gains achieved by the two schemes are the same. This phenomenon is expected since for both schemes, the outage probabilities are determined by the same effective channel gain,  $[\mathbf{R}_2^H]_{i,i}^2$ .

### B. IMPACT OF DIFFERENT SYSTEM PARAMETERS ON USERS' OUTAGE PERFORMANCE

In Figs. 3 and 4, user 1's outage performance achieved by the proposed MIMO-NOMA scheme is shown with different choices of the targeted data rates. Particularly, when power allocation policy I is used, Fig. 3 shows that the outage probability curves achieved by the proposed MIMO-NOMA

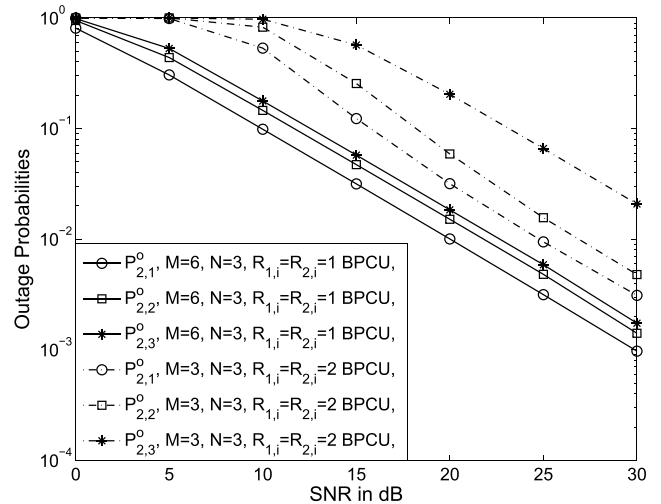


**FIGURE 5. Outage performance at user 2 with power allocation policy I.**  $M = N = 3, P_1 = 1 - e^{-\frac{2\epsilon_{1,i}}{\rho}}, P_2 = 1 - e^{-\frac{10\epsilon_{2,i}}{\rho}}, R_{1,i} = 1$  BPCU, and  $R_{2,i} = 2$  BPCU. The analytical results are based on (39) and (40).

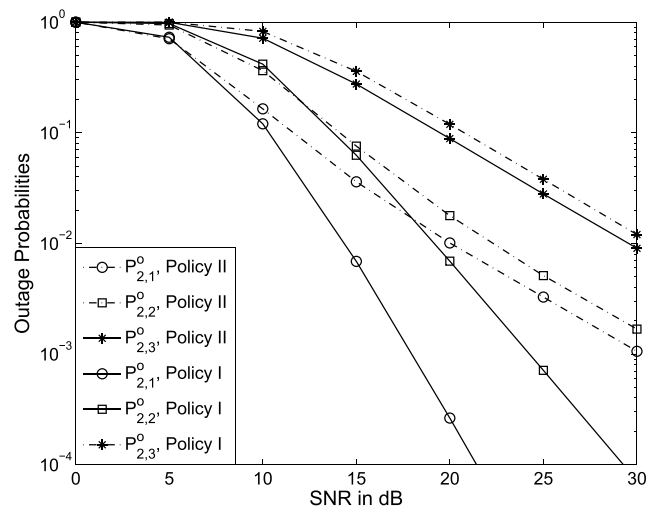
transmission scheme match perfectly with those for the targeted outage probability, which demonstrates that the proposed transmission scheme can strictly guarantee the QoS requirements at user 1 in the long term. When power allocation policy II is used, Fig. 4 demonstrates that the simulation results match perfectly with the analytical results developed in (28). Therefore, the use of power allocation policy II guarantees that the outage probability at user 1 is  $(1 - e^{-\frac{\epsilon_{1,i}}{\rho}})$ , which is equivalent to the outage performance for the case in which all the power is allocated to user 1.

In Fig. 5, the outage probabilities at user 2 are shown as functions of the transmit SNR, when power allocation policy I is used. As can be observed from the figure, the curves for the simulation results match perfectly with the ones for the analytical result developed in (39), which demonstrates the accuracy of this exact expression for the outage probability. The curves for the approximation results developed in (40) match the simulation curves only at high SNR, which is due to the fact that this approximation is obtained with the high SNR assumption. Another important observation from this figure is that the slope of the outage probability curve for  $P_{2,i}^o$  is larger than that of  $P_{2,j}^o$ , for  $i < j$ . This is because a larger diversity gain can be obtained at the  $i$ -th layer, compared to that at the  $j$ -th layer, as discussed in Remark 3 in Section IV-A.

Fig. 6 demonstrates the outage performance at user 2 when power allocation policy II is used. As shown in the figure, the outage performance of user 2 can be improved by increasing the number of antennas at the base station or decreasing the targeted data rates. An important observation from this figure is that at high SNR, the slope of all the curves is the same, which means that the same diversity gain is achieved at all layers, regardless of the choice of the number of antennas at the base station. This confirms the analytical results developed in Section IV-B1, in which it is shown that the diversity



**FIGURE 6. Outage performance at user 2 with power allocation policy II.**  $M = N = 3$ .



**FIGURE 7. Comparison between the two power allocation policies.**  $M = N = 3, R_{1,i} = 1$  BPCU,  $R_{2,i} = 2$  BPCU,  $P_1 = 1 - e^{-\frac{2\epsilon_{1,i}}{\rho}}$  and  $P_2 = 1 - e^{-\frac{10\epsilon_{2,i}}{\rho}}$ .

gain is one for all layers. In Fig. 7, the outage performance experienced by user 2 with different power allocation policies is compared. Particularly, this figure shows that the use of power allocation policy I is preferable for user 2, since better outage performance can be achieved. However, it is important to point out that the use of power allocation policy II can meet the QoS requirements of user 1 instantaneously as shown in (14), whereas power allocation policy I can only meet the long term QoS requirements.

## VI. CONCLUSIONS

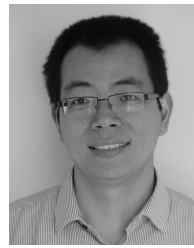
In this paper, we have considered a MIMO-NOMA downlink transmission scenario, in which a new precoding and power allocation strategy has been proposed to ensure that the potential of NOMA can be realized even if the participating users' channel conditions are similar. Particularly, the precoding matrix has been designed to degrade user 1's effective channel gains while improving the signal strength at user 2.



Two types of power allocation policies have been developed to meet user 1's QoS requirements in a long and short term, respectively. Analytical and numerical results have also been provided to demonstrate the advantages and disadvantages of the two power allocation policies. The outage performance of user 2 has been analyzed by using some bounding techniques, whereas an important future direction is to find a closed-form expression for this outage probability by applying the order statistics of the diagonal elements of an inverse Wishart matrix. It should be pointed out that the proposed scheme can be extended to general scenarios with more than two users in a straightforward manner, e.g., following the power allocation policies proposed in this paper, we can design the precoding matrix according to one user's channel condition and adjust the power allocation coefficients of the other users to meet their dynamic QoS requirements.

## REFERENCES

- [1] *NGMN 5G White Paper*, NGMN Alliance, Feb. 2015.
- [2] "5G radio access: Requirements, concepts and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G White paper, Jul. 2014.
- [3] 3GPP, "Study on downlink multiuser superposition transmission for LTE," 3rd Generat. Partnership Project, Tech. Rep., Mar. 2015.
- [4] 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.*, Dresden, Germany, Jun. 2013, pp. 1–5.
- [5] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
- [6] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.
- [7] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmission," *IEEE Trans. Veh. Technol.*, to appear in 2016.
- [8] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, to appear in 2016.
- [9] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.
- [10] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.
- [11] Q. Sun, S. Han, C. L. I., and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Aug. 2015.
- [12] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, to appear in 2016.
- [13] 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.
- [14] Z. Ding, R. Schober, and H. V. Poor, "On the design of MIMO-NOMA downlink and uplink transmission," in *Proc. IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, Jun. 2016, to appear.
- [15] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I., and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.
- [16] A. M. Tulino and S. Verdú, *Foundations and Trends in Communications and Information Theory: Random Matrix Theory and Wireless Communications*. Hanover, MA, USA: Now Publishers, 2004.
- [17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York, NY, USA: Academic, 2000.
- [18] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse Wishart distributed matrices," *IEE Proc.-Radar, Sonar Navigat.*, vol. 147, no. 4, pp. 162–168, Aug. 2000.



**ZHIGUO DING** (S'03–M'05–SM'15) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications, in 2000, and the Ph.D. degree in electrical engineering from Imperial College London, in 2005. From 2005 to 2014, he was with Queen's University Belfast, Imperial College, and Newcastle University. Since 2014, he has been with Lancaster University as a Chair Professor. From 2012 to 2016, he was also with Princeton

University as an Academic Visitor.

His research interests are 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He serves as an Editor of the *IEEE TRANSACTIONS ON COMMUNICATIONS*, the *IEEE TRANSACTIONS ON VEHICULAR NETWORKS*, the *IEEE WIRELESS COMMUNICATION LETTERS*, the *IEEE COMMUNICATION LETTERS*, and the *Journal of Wireless Communications and Mobile Computing*. He received the best paper award in the IET Communications Conference on Wireless, Mobile and Computing in 2009, the IEEE Communication Letter Exemplary Reviewer in 2012, and the EU Marie Curie Fellowship from 2012 to 2014.



**LINGLONG DAI** (M'11–SM'14) received the B.S. degree from Zhejiang University, in 2003, the M.S. (Hons.) degree from the China Academy of Telecommunications Technology, in 2006, and the Ph.D. (Hons.) degree from Tsinghua University, Beijing, China, in 2011. From 2011 to 2013, he was a Post-Doctoral Fellow with the Department of Electronic Engineering, Tsinghua University, where he has been an Assistant Professor since 2013. His research interests are in wireless communications,

with a focus on multicarrier techniques, multi-antenna techniques, and multiuser techniques. He has published over 40 IEEE journal papers on the IEEE JSAC (8 papers), the IEEE WCOM, COMMAG, TSP, TWC, TCOM, TVT, and CL, and over 30 IEEE conference papers in IEEE ICC and GLOBECOM. He has received the Outstanding Ph.D. Graduate of Tsinghua University Award in 2011, the Excellent Doctoral Dissertation of Beijing Award in 2012, the IEEE ICC Best Paper Award in 2013, the National Excellent Doctoral Dissertation Nomination Award in 2013, the IEEE ICC Best Paper Award in 2014, the URSI Young Scientist Award in 2014, the IEEE Transactions on Broadcasting Best Paper Award in 2015, and the IEEE RADIO Young Scientist Award in 2015. He serves as a Co-Chair of the IEEE Special Interest Group on Signal Processing Techniques in 5G Communication Systems.



**H. VINCENT POOR** (S'72–M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University, in 1977. From 1977 to 1990, he was a faculty member with the University of Illinois at Urbana–Champaign. Since 1990, he has been a faculty member with Princeton University, where he is the Michael Henry Strater University Professor and the Dean of the School of Engineering and Applied Science. He has also held visiting appointments at several universities, including

most recently with Stanford University and Imperial College. His research interests are in the area of wireless networks and related fields. Among his publications in these areas is the recent book *Mechanisms and Games for Dynamic Spectrum Allocation* (Cambridge University Press, 2014).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and a Foreign Member of the Royal Society. He is also a Fellow of the American Academy of Arts and Sciences, the National Academy of Inventors, and other national and international academies. He received the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2014 URSI Booker Gold Medal, the 2015 EURASIP Athanasios Papoulis Award, the 2016 John Fritz Medal, and honorary doctorates from Aalborg University, Aalto University, HKUST, and the University of Edinburgh.

• • •