On the Application of Quasi-Degradation to MISO-NOMA Downlink

Zhiyong Chen, Zhiguo Ding, *Member, IEEE* Xuchu Dai, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—In this paper, the design of non-orthogonal multiple access (NOMA) in a multiple-input-single-output (MISO) downlink scenario is investigated. The impact of the recently developed concept, *quasi-degradation*, on NOMA downlink transmission is first studied. Then, a Hybrid NOMA (H-NOMA) precoding algorithm, based on this concept, is proposed. By exploiting the properties of H-NOMA precoding, a low-complexity sequential user pairing algorithm (SUPA) is consequently developed, to further improve the overall system performance. Both analytical and numerical results are provided to demonstrate the performance of the H-NOMA precoding through the average power consumption and outage probability, while conventional schemes, as dirty-paper coding and zero-forcing beamforming, are used as benchmarking.**

***Index Terms*—Non-orthogonal multiple access (NOMA), multiple-input-single-output (MISO), quasi-degradation, outage probability, sequential user pairing algorithm.**

I. Introduction

THE design of downlink transmission is crucial in cellular networks, where the challenge is how to support broadband downloading services, constrained by the scarce bandwidth resources. Most conventional downlink transmission schemes developed in the literature are based on the use of orthogonal multiple access (OMA), e.g., time/frequency/code division multiple access (TDMA/FDMA/CDMA) [1] and orthogonal frequency-division multiplexing access (OFDMA) [2]. By using the advantage of these orthogonal resource allocation techniques, the interference between different users can be avoided [3]. However, these schemes are far from the optimality, where a spectrum resource, allocated to a user poorly connected to the base station (BS), cannot be efficiently used. Moreover, the use of opportunistic scheduling, such as single-user (SU) selection, can ensure that the scheduled users have the largest weighted sum capacity per channel transmission block [4]. These approaches are effective to improve the system throughput, but they result in a loss of user connectivity and fairness. It is worth pointing out that several downlink precoding schemes, such as zero-forcing beamforming (ZFBF) [5], [6], can also be categorized as a special case of OMA, where spatial degrees of freedom are used for interference avoidance. For example, the use of ZFBF mitigates multi-user interference, by transmitting data in the null space of other users' channel matrices [7]. Also, it has been shown in [8] that the ZFBF strategy can achieve the optimal asymptotic sum capacity as the number of users goes to infinity. However, this approach is efficient only when the number of transmit antennas is not smaller than the number of receive antennas.

In parallel to design practical downlink transmission schemes, the information theoretic counterpart of downlink, termed as *broadcast channel*, has been investigated extensively during the last three decades. By assuming perfect channel state information at transmitter (CSIT), it has been shown that the capacity region can be achieved by using dirty-paper coding (DPC) [9]. However, DPC is difficult to be implemented in practical communication systems, due to its prohibitively high complexity. To this end, this paper is motivated by the following question: Is it possible to design a practical downlink transmission scheme, which can outperform those based on OMA and simultaneously yield a performance close to the capacity region of multiple-input single-output (MISO) broadcast channels?

A. Literature

Recently, non-orthogonal multiple access (NOMA) has been recognized as a promising multiple access technique for the 5th generation (5G) of mobile networks, due to its superior spectral efficiency compared to traditional OMA [10]–[15]. Particularly, the concept of NOMA is shown to be ideal for improving the spectral efficiency of downlink transmission. This is the reason why NOMA has been recently proposed to downlink scenarios in 3rd generation partnership project long-term evolution (3GPP-LTE) systems. Specifically, NOMA has been included in LTE Release 13, termed as multi-user superposition transmission (MUST) [16], which is a two-user downlink special case. In particular, MUST can be viewed as a hybrid multiple access scheme between OFDMA and NOMA, where NOMA is adopted by two users sharing the same sub-carrier. The implementation of NOMA is based on the combination of superposition code (SC) and successive interference cancellation (SIC), which is a method proved to be able to achieve the capacity region of degraded broadcast channels [17]. Specifically, take a two-user single-input single-output (SISO) NOMA system as an example. The BS serves the users at the same time/code/frequency channel, where

Manuscript received XXX XX, 2016. The associate editor coordinating the review of this paper and approving it for publication was XXX.

This work was supported in part by the National Natural Science Foundation of China (No. 61471334). The work of Z. Ding was supported by the UK EPSRC under grant number EP/L025272/1 and by H2020-MSCA-RISE-2015 under grant number 690750.

Z. Chen and X. Dai are with the Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences, School of Information Science and Technology, University of Science and Technology of China, No. 96 Jinzhai Road, Hefei, Anhui Province, 230026, P. R. China. Email: zhiyong@mail.ustc.edu.cn, daixc@ustc.edu.cn.

Z. Ding is with the School of Computing and Communications, Lancaster University, LA1 4YW, UK. Email: z.ding@lancaster.ac.uk.

G. K. Karagiannidis is with the Provincial Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Sichuan 611756, China, and also with the Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki GR-54124, Greece. Email: geokarag@auth.gr.

the signals are superposed with different power allocation coefficients. At the user side, the far user (i.e., the user with poor channel conditions) decodes its information by treating the other's information as noise, while the near user (i.e., the user with strong channel conditions) first decodes the information of its partner and then decodes its own information by removing partner's information from its observation. In this way, both users can have full access to all the resource blocks (RBs), moreover, the near user can decode its own information without any interference from the far user. Therefore, the overall performance is enhanced, compared to conventional OMA and linear transceiver schemes. But, note that, with regarding to the general non-degraded multiple-input multiple-output (MIMO) broadcast channels, NOMA-based schemes often yield a performance loss.

Because of its relatively low complexity compared to D-PC, recently, NOMA-based MIMO downlink transmission has attracted considerable research interest. For example, in [18], the ergodic rate maximization problem for MIMO NOMA systems with statistical CSIT was first formulated and then two algorithms to solve this problem, were proposed. In [19], the downlink sum rate maximization problem for MISO downlink with perfect CSIT was investigated. However, the ordering of users with similar distances to the BS is still unknown. In [20], users were grouped into small-size clusters, and NOMA was implemented for the users within one cluster, while MIMO detection was used to cancel inter-cluster interference. A general MIMO-NOMA framework for downlink and uplink transmission was proposed in [21] by applying the concept of signal alignment. Furthermore, in [22] and [23], the performance of MISO downlink was enhanced by applying various NOMA-based multi-user beamforming (NOMA-BF) strategies. While the aforementioned MIMO-NOMA schemes can offer efficient and practical solutions with several advantages, they yield few insights about their optimality, compared to the capacity (rate) regions of broadcast channels, particularly for MISO communications, where users have similar distances to the BS.

The concept of *quasi-degradation* was previously developed in [24], and used to characterize the gap between the optimal performance of DPC and that achieved by NOMA for the special case of two users. In this paper, we focus on the application of this concept to the general multi-user MISO downlink, by employing the idea of user pairing. We first formulate a quality-of-service (QoS) optimization problem for two users, which minimizes the total transmit power constrained by the target individual rates. Closed-form expressions for different precoding algorithms including DPC and ZFBF are obtained, which can be used to insightful performance evaluation. Furthermore, following the initial results reported in [24], a more in-depth study is first undertaken to illustrate important properties of the quasi-degradation concept. Consequently, by using these properties, a Hybrid NOMA (H-NOMA) precoding algorithm with closed-form expressions is presented, and a sequential user pairing algorithm (SUPA) is proposed, which is used in combination with H-NOMA precoding to yield a practical efficient transmission scheme for MISO downlink. Furthermore, analytical results, based on various performance metrics, such as the quasi-degradation probability, the expectation of total power consumption and the outage probability, are presented. Finally, simulations are provided to demonstrate the accuracy of analytical results and also to validate the efficiency of the proposed transmission scheme.

B. Contribution

The main contribution of this paper is listed as follows:

1) Building on the quasi-degradation concept previously proposed in [24], closed-form expressions for the optimal precoding vectors in the addressed quasi-degraded channels are obtained, whereas an iterative method was still needed in [24]. In addition, the definition for quasi-degradation is refined and generalized in order to provide both the sufficient and necessary conditions for quasi-degradation, whereas only the sufficient condition for quasi-degradation was obtained in [24]. Furthermore, a closed-form expression for the quasi-degradation probability is derived, and a novel H-NOMA precoding algorithm is developed, with the application of the quasi-degradation concept.
2) By combining H-NOMA precoding and SUPA, a practical transmission scheme is proposed. By taking advantage of the closed-form expressions of H-NOMA precoding and the efficiency of SUPA, the proposed transmission scheme can be implemented with a low complexity of $\mathcal{O}(K^2N)$, where N denotes the number of antennas at the BS and K is the number of users. Note that the complexity of the optimal NOMA with exhaustive search is $\mathcal{O}((K-1)!!N^3)$.
3) Analytical results are provided to demonstrate the performance achieved by the proposed H-NOMA precoding algorithm, compared to DPC and ZFBF, which are used as benchmarking schemes. Particularly, by comparing the proposed scheme to DPC, the optimality of the proposed downlink transmission scheme is clearly illustrated, a result that, to the best of the authors' knowledge, has not been previously presented in the literature.

C. Structure

The remainder of this paper is organized as follows. Section II briefly describes the system model and introduces some existing transmission schemes. In Section III, an in-depth study on quasi-degradation is undertaken. The proposed transmission scheme is presented in Section IV, while Section V includes analytical results. Section VI illustrates the numerical results, and finally Section VII concludes this paper.

II. Problem Description

A. System Model

Consider a downlink communication system with one BS and K (assumed to be a even number) mobile users. The BS is equipped with N ($N \geq 2$) antennas and each user is equipped with a single antenna. By adopting the idea of user pairing

[12], users m and n are assumed to be paired over the shared spectrum. The received signal at user i is given by

$$y_i = \mathbf{h}_i^H \mathbf{x} + n_i, \qquad i = m, n, \tag{1}$$

where the channel coefficient $\mathbf{h}_m$ and $\mathbf{h}_n$ are independent distributed $\mathcal{CN}(0, 2\sigma_m^2 \mathbf{I}_N)$ and $\mathcal{CN}(0, 2\sigma_n^2 \mathbf{I}_N)$[1], $n_i \sim \mathcal{CN}(0,1)$ is the additive Gaussian noise at user i, and $2\sigma_i^2$ is the variance of the channel between the BS and user i. Furthermore, $\mathbf{x}$ is the signal transmitted by the BS containing s_m and s_n, where s_i is the signal intended to user i. For example, we consider that $\mathbf{x}$ is a linear combination of the two signals, $\mathbf{x} = \mathbf{w}_m s_m + \mathbf{w}_n s_n$, i.e., superposition coding is used, and $\mathbf{w}_i$ is termed as the corresponding precoding vector. The power of s_i is assumed to be normalized, i.e., $\mathcal{E}\{s_i^2\} = 1$. By using superposition code, the design complexity at the BS is dramatically reduced in comparison to the non-linear DPC.

Next, the aim is to minimize the total transmission power of the BS in order to meet the requirement for the target rate of each user. Mathematically, it can be formulated through the following QoS optimization problem

$$\begin{aligned} \min_{\mathbf{w}_j} \quad & \sum_{j=1}^{K} P_j \\ \text{s.t.} \quad & R_j \geq R_j^{MT} \quad j = 1, 2, ..., K, \end{aligned} \tag{2}$$

where $P_j = \|\mathbf{w}_j\|^2$ is the transmission power, which must be optimized in order to convey signal s_j to user j with the rate constraint, and R_j^{MT} is the target rate of user j.

B. Existing Transmission Schemes

Several schemes for the MISO downlink have been proposed in the literature. Herein, we focus on two widely used schemes, i.e., DPC and ZFBF, which can be combined with user pairing.

For the ease of analysis, to describe DPC, a fixed encoding order (n, m) is assumed at the BS. Specifically, for a fixed pair of users, e.g., users m and n, the BS first encodes the information intended to user n, and then encodes the information intended to user m by pre-subtracting the first information. Hence, the achievable rate pair can be expressed as

$$\begin{cases} R_m = \ln(1 + \mathbf{h}_m^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_m) \\ R_n = \ln(1 + \text{SINR}_{n,n}) \end{cases}, \tag{3}$$

where

$$\text{SINR}_{n,n} = \frac{\mathbf{h}_n^H \mathbf{w}_n \mathbf{w}_n^H \mathbf{h}_n}{1 + \mathbf{h}_n^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_n}.$$

The minimum required power for reliable transmission at the target rates can be found by solving the following optimization problem

$$\begin{aligned} P_{m,n}^{\text{DPC}} = & \min_{\mathbf{w}_m, \mathbf{w}_n} \quad \|\mathbf{w}_m\|^2 + \|\mathbf{w}_n\|^2 \\ \text{s.t.} \quad & \mathbf{h}_m^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_m \geq r_m^{MT} \\ & \mathbf{h}_n^H \mathbf{w}_n \mathbf{w}_n^H \mathbf{h}_n \geq r_n^{MT} (1 + \mathbf{h}_n^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_n), \end{aligned} \tag{4}$$

where, r_i^{MT}, is the corresponding target signal-to-noise ratio (SNR) level, i.e., $\ln(1 + r_i^{MT}) = R_i^{MT}, \quad i = m, n$. The optimal solution of the problem in (4) can be derived by using Lemma 1 and Proposition 1 presented in the next section. The minimum required power for transmitting messages to the paired users using DPC can be expressed as

$$P_{m,n}^{\text{DPC}} = \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} + \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \frac{1 + r_n^{MT}}{1 + r_n^{MT} \sin^2 \theta}, \tag{5}$$

where $\theta \in [0, \pi]$ is the angle between $\mathbf{h}_m$ and $\mathbf{h}_n$, i.e.,

$$\cos^2 \theta = \frac{\mathbf{h}_n^H \mathbf{h}_m \mathbf{h}_m^H \mathbf{h}_n}{\|\mathbf{h}_m\|^2 \|\mathbf{h}_n\|^2}.$$

In ZFBF, each user's stream is coded independently and multiplied by a precoding vector for transmission. This vector should be designed to eliminate mutual interference among different streams, by taking advantage of the spatial structure between users' channel matrices. Note that ZFBF is also a kind of space-division multiple access (SDMA). Mathematically, for a fixed pair of users, the minimum required power for reliable transmission at the target rates can be formulated with the following optimization problem

$$\begin{aligned} P_{m,n}^{\text{ZF}} = & \min_{\mathbf{w}_m, \mathbf{w}_n} \quad \|\mathbf{w}_m\|^2 + \|\mathbf{w}_n\|^2 \\ \text{s.t.} \quad & (\mathbf{h}_m^H \mathbf{w}_m)^2 \geq r_m^{MT} \\ & (\mathbf{h}_n^H \mathbf{w}_n)^2 \geq r_n^{MT} \\ & (\mathbf{h}_m^H \mathbf{w}_n) = 0 \\ & (\mathbf{h}_n^H \mathbf{w}_m) = 0. \end{aligned} \tag{6}$$

The optimal solution of this problem can be trivially obtained by employing the least square property of Moore-Penrose inverse [25]. By defining the matrix $\mathbf{H} = \begin{bmatrix} \mathbf{h}_m & \mathbf{h}_n \end{bmatrix}$, the optimal solution can be evaluated as

$$\begin{aligned} \begin{bmatrix} \mathbf{w}_m & \mathbf{w}_n \end{bmatrix} &= \mathbf{H}^{H\dagger} \begin{bmatrix} \sqrt{r_m^{MT}} & 0 \\ 0 & \sqrt{r_n^{MT}} \end{bmatrix} \\ &= \frac{1}{\|\mathbf{h}_m\|^2 \|\mathbf{h}_n\|^2 \sin^2 \theta} \begin{bmatrix} \sqrt{r_m^{MT}} \mathbf{a}_m & \sqrt{r_n^{MT}} \mathbf{a}_n \end{bmatrix}, \end{aligned}$$

where $\dagger$ stands for the Moore-Penrose inverse, and

$$\begin{aligned} \mathbf{a}_m &= \|\mathbf{h}_n\|^2 \mathbf{h}_m - (\mathbf{h}_n^H \mathbf{h}_m) \mathbf{h}_n \\ \mathbf{a}_n &= -(\mathbf{h}_m^H \mathbf{h}_n) \mathbf{h}_m + \|\mathbf{h}_m\|^2 \mathbf{h}_n. \end{aligned}$$

Alternatively, the optimal precoding vectors can be expressed as

$$\begin{cases} \mathbf{w}_m = \dfrac{\sqrt{r_m^{MT}} \mathbf{a}_m}{\|\mathbf{h}_m\|^2 \|\mathbf{h}_n\|^2 \sin^2 \theta} \\ \mathbf{w}_n = \dfrac{\sqrt{r_n^{MT}} \mathbf{a}_n}{\|\mathbf{h}_m\|^2 \|\mathbf{h}_n\|^2 \sin^2 \theta} \end{cases}. \tag{7}$$

Therefore, the minimum required power for transmitting messages to these paired users using ZFBF can be calculated as

$$P_{m,n}^{\text{ZF}} = \frac{1}{\sin^2 \theta} \left(\frac{r_m^{MT}}{\|\mathbf{h}_m\|^2} + \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} \right). \tag{8}$$

[1] $\mathcal{CN}(0, 2\sigma_i^2 \mathbf{I}_N)$ stands for circularly-symmetric complex normal distributions, with mean zero and covariance matrix $2\sigma_i^2 \mathbf{I_N}$.

C. NOMA and Quasi-Degradation

For the ease of analysis, to introduce NOMA transmission scheme, a fixed decoding order (n, m) is also assumed. Specifically, for a fixed pair of users, e.g., users m and n, the BS transmits the superposition code, $\mathbf{x} = \mathbf{w}_m s_m + \mathbf{w}_n s_n$. At the user side, the SIC process is implemented. Particularly, user m first decodes s_n and subtracts this from its received signal y_m. Thus, user m can decode s_m without interference from s_n. User n does not perform SIC and simply decodes s_n by treating s_m as noise. Note that the encoding order (n, m) in DPC is consistent with the decoding order (n, m) in NOMA. Therefore, the achievable rate pair can be expressed as

$$\begin{cases} R_m = \ln(1 + \mathbf{h}_m^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_m) \\ R_n = \min\left\{\ln\left(1 + \mathrm{SINR}_{n,m}\right), \ln\left(1 + \mathrm{SINR}_{n,n}\right)\right\} \end{cases}, \tag{9}$$

where

$$\mathrm{SINR}_{n,m} = \frac{\mathbf{h}_m^H \mathbf{w}_n \mathbf{w}_n^H \mathbf{h}_m}{1 + \mathbf{h}_m^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_m}.$$

The minimum required power for reliable transmission at the target rates can be formulated through the following optimization problem

$$\begin{aligned} P_{m,n}^{\mathrm{NOMA}} = \min_{\mathbf{w}_m, \mathbf{w}_n} \quad & \|\mathbf{w}_m\|^2 + \|\mathbf{w}_n\|^2 \\ \text{s.t.} \quad & \mathbf{h}_m^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_m \geq r_m^{MT} \\ & \mathbf{h}_m^H \mathbf{w}_n \mathbf{w}_n^H \mathbf{h}_m \geq r_n^{MT}(1 + \mathbf{h}_m^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_m) \\ & \mathbf{h}_n^H \mathbf{w}_n \mathbf{w}_n^H \mathbf{h}_n \geq r_n^{MT}(1 + \mathbf{h}_n^H \mathbf{w}_m \mathbf{w}_m^H \mathbf{h}_n). \end{aligned} \tag{10}$$

In general, a closed-form solution for this problem is difficult to obtain. Fortunately, by introducing the definition of quasi-degradation in [24], it is shown that it can be achieved if the channel coefficients $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded. In this paper, this definition is refined in a more general way.

Definition 1 (Quasi-Degradation)**.** *Without loss of generality, we assume a fixed decoding order (n, m) of NOMA and a fixed encoding order (n, m) of DPC.*[2] *Given the channel coefficients $\mathbf{h}_m, \mathbf{h}_n$ and the target SNR levels r_m^{MT}, r_n^{MT}, then the broadcast channels $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded with respect to r_m^{MT} and r_n^{MT} if and only if the minimum transmission power of NOMA is equivalent to that of DPC, i.e.,*

$$P_{m,n}^{\mathrm{NOMA}} = P_{m,n}^{\mathrm{DPC}}. \tag{11}$$

Note that closed-form expressions for the precoding vectors and the explicit condition of quasi-degradation, were not provided in the original work [24].

III. Further study of Quasi-degradation

In this section, closed-form precoding vectors for quasi-degraded channels as well as an explicit sufficient and necessary condition are given in Propositions 1 and 2, respectively.

To obtain the optimal solution for quasi-degraded channels using NOMA, we first introduce the following Lemma.

Lemma 1. *If $\{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}$ is an optimal solution of* (10)*, and the broadcast channels are quasi-degraded, then, there exists an optimal solution $\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\}$ of* (4)*, such that*

$$\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\} = \{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}.$$

Proof: Denote the feasible region of the optimization problem in (4) and (10) by $\mathcal{C}_D$ and $\mathcal{C}_N$, respectively. It is clear that $\mathcal{C}_N \subseteq \mathcal{C}_D$. Therefore, the optimal solution of (10) $\{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}$ is also a feasible solution of (4). Denote the objective function value corresponding to this solution by $P_{m,n}^{\mathrm{NOMA}}$. According to the definition of quasi-degradation, $P_{m,n}^{\mathrm{NOMA}} = P_{m,n}^{\mathrm{DPC}}$, which means that $\{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}$ can achieve the optimal value of (4). In other words, $\{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}$ is also an optimal solution of (4), i.e.,

$$\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\} = \{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}.$$

and the proof is completed. □

By using Lemma 1, a closed-form optimal solution of quasi-degraded channels using NOMA is obtained in the following Proposition.

Proposition 1. *If the broadcast channels $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded with respect to r_m^{MT} and r_n^{MT}, then an optimal solution of* (10) *is*

$$\begin{cases} \mathbf{w}_m^{N^*} = \alpha_m((1 + r_n^{MT})\mathbf{e}_m - r_n^{MT}\mathbf{e}_n^H \mathbf{e}_m \mathbf{e}_n) \\ \mathbf{w}_n^{N^*} = \alpha_n \mathbf{e}_n \end{cases}, \tag{12}$$

where

$$\begin{cases} \mathbf{e}_m = \dfrac{\mathbf{h}_m}{\|\mathbf{h}_m\|}, \mathbf{e}_n = \dfrac{\mathbf{h}_n}{\|\mathbf{h}_n\|} \\ \alpha_m^2 = \dfrac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \dfrac{1}{(1 + r_n^{MT}\sin^2\theta)^2} \\ \alpha_n^2 = \dfrac{r_n^{MT}}{\|\mathbf{h}_n\|^2} + \dfrac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \dfrac{r_n^{MT}\cos^2\theta}{(1 + r_n^{MT}\sin^2\theta)^2} \end{cases}. \tag{13}$$

Proof: By employing Lemma 1, the optimal solution of (10) can be acquired by solving the optimization problem in (4). Note that this non-convex optimization problem can be transformed into a convex one, and hence strong duality holds. For more details, please see Proposition 1 in [24]. By introducing the Lagrangian multipliers λ_1^D, λ_3^D, the dual problem of (4) can be written as [26]

$$\begin{aligned} d^{D^*} = \max_{\lambda_1^D, \lambda_3^D} \quad & r_m^{MT}\lambda_1^D + r_n^{MT}\lambda_3^D \\ \text{s.t.} \quad & \mathbf{A}_1^D \succeq 0, \quad \mathbf{A}_2^D \succeq 0, \\ & \lambda_i^D \geq 0, \quad i = 1, 3, \end{aligned} \tag{14}$$

where

$$\begin{cases} \mathbf{A}_1^D = \mathbf{I}_N - \lambda_1^D \mathbf{h}_m \mathbf{h}_m^H + \lambda_3^D r_n^{MT} \mathbf{h}_n \mathbf{h}_n^H \\ \mathbf{A}_2^D = \mathbf{I}_N - \lambda_3^D \mathbf{h}_n \mathbf{h}_n^H \end{cases}. \tag{15}$$

The Karush-Kuhn-Tucker (KKT) conditions are

$$\begin{cases} \mathbf{A}_1^D \mathbf{w}_m^{D^*} = \mathbf{0}, \mathbf{A}_2^D \mathbf{w}_n^{D^*} = \mathbf{0}, & \text{(16a)} \\ \text{Primary \& dual Constraints,} & \text{(16b)} \\ \text{Complementarity Conditions.} & \text{(16c)} \end{cases}$$

[2] The same assumption is used throughout this paper, including all the Lemmas, Propositions and Theorems.

From (16a), it holds that

$$\begin{cases} \lambda_1^{D^*} = \dfrac{1}{\|\mathbf{h}_m\|^2}\dfrac{1+r_n^{MT}}{1+r_n^{MT}\sin^2\theta} \\ \lambda_3^{D^*} = \dfrac{1}{\|\mathbf{h}_n\|^2} \end{cases}.$$

Therefore, the optimal objective function value of (14) can be evaluated as

$$\begin{aligned} d^{D^*} &= r_m^{MT}\lambda_1^{D^*} + r_n^{MT}\lambda_3^{D^*} \\ &= \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} + \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2}\frac{1+r_n^{MT}}{1+r_n^{MT}\sin^2\theta}, \end{aligned} \tag{17}$$

and the KKT conditions in (16) can be derived as

$$\begin{cases} \mathbf{A}_1^D\mathbf{w}_m^{D^*} = 0 & \text{(18a)} \\ \mathbf{A}_2^D\mathbf{w}_n^{D^*} = 0 & \text{(18b)} \\ \mathbf{h}_m^H\mathbf{w}_m^{D^*}\mathbf{w}_m^{D^*H}\mathbf{h}_m = r_m^{MT} & \text{(18c)} \\ \mathbf{h}_n^H\mathbf{w}_n^{D^*}\mathbf{w}_n^{D^*H}\mathbf{h}_n = r_n^{MT}(1+\mathbf{h}_n^H\mathbf{w}_m^{D^*}\mathbf{w}_m^{D^*H}\mathbf{h}_n) & \text{(18d)} \end{cases}$$

Consequently, the solutions for (18) can be obtained after some trivial manipulations as

$$\begin{cases} \mathbf{w}_m^{D^*} = \alpha_m((1+r_n^{MT})\mathbf{e}_m - r_n^{MT}\mathbf{e}_n^H\mathbf{e}_m\mathbf{e}_n) \\ \mathbf{w}_n^{D^*} = \alpha_n\mathbf{e}_n \end{cases}, \tag{19}$$

where $\mathbf{e}_m, \mathbf{e}_n, \alpha_m, \alpha_n$ are defined in (13). Therefore, the optimal solution set of (4) $\mathcal{W}^D$ can be written as

$$\begin{aligned} \mathcal{W}^D = \Big\{ &\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\}\Big| \\ &\mathbf{w}_m^{D^*} = \alpha_m((1+r_n^{MT})\mathbf{e}_m - r_n^{MT}\mathbf{e}_n^H\mathbf{e}_m\mathbf{e}_n), \\ &\mathbf{w}_n^{D^*} = \alpha_n\mathbf{e}_n, \\ &\alpha_m = \sqrt{\alpha_m^2}e^{j\phi_m}, \alpha_n = \sqrt{\alpha_n^2}e^{j\phi_n}, \phi_m, \phi_n \in [0, 2\pi]\Big\}. \end{aligned} \tag{20}$$

Since the strong duality holds, each $\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\} \in \mathcal{W}^D$ is also global optimal. Finally, by using Lemma 1, it is concluded that there exists one optimal solution $\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\} \in \mathcal{W}^D$, such that

$$\{\mathbf{w}_m^{D^*}, \mathbf{w}_n^{D^*}\} = \{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}.$$

In other words, there must exist one optimal solution of (10) such that

$$\{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\} \in \mathcal{W}^D,$$

and the proof is completed. □

By taking advantage of the closed-form solution given in Proposition 1, an explicit sufficient and necessary condition for channels to be quasi-degraded is given in the next Proposition.

Proposition 2. *The broadcast channels $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded with respect to r_m^{MT} and r_n^{MT}, if and only if*

$$Q(u) \le \frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2}, \tag{21}$$

where

$$Q(u) = \frac{1+r_m^{MT}}{u} - \frac{r_m^{MT}u}{(1+r_n^{MT}(1-u))^2}, u = \cos^2\theta.$$

Proof: First, we prove the sufficiency. If $\mathbf{h}_m$ and $\mathbf{h}_n$ with respect to r_m^{MT} and r_n^{MT} are quasi-degraded, by adopting Proposition 1, the optimal solution of the optimization problem in (10) is $\{\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}\}$. Hence, it must satisfy the constraints in (10). Specifically, we focus on the second constraint in (10)

$$\mathbf{h}_m^H\mathbf{w}_n^{N^*}\mathbf{w}_n^{N^*H}\mathbf{h}_m \ge r_n^{MT}(1+\mathbf{h}_m^H\mathbf{w}_m^{N^*}\mathbf{w}_m^{N^*H}\mathbf{h}_m). \tag{22}$$

By substituting the expressions of $\mathbf{w}_m^{N^*}, \mathbf{w}_n^{N^*}$ given in (12) to the inequality in (22), one can obtain (21).

Then, we prove the necessity in two steps. Specifically, if $\mathbf{h}_m$ and $\mathbf{h}_n$ satisfy $Q(u) \le \frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2}$, then we prove that the optimal value of (4) is equivalent to that of (10), i.e., $P_{m,n}^{\text{NOMA}} = P_{m,n}^{\text{DPC}}$. By the definition of quasi-degradation in Definition 1, we can conclude that $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded. In the following we focus on proving $P_{m,n}^{\text{NOMA}} = P_{m,n}^{\text{DPC}}$ by using contradictive method.

Given a pair of specific precoding vectors $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ as

$$\begin{cases} \mathbf{w}_m^c = \alpha_m((1+r_n^{MT})\mathbf{e}_m - r_n^{MT}\mathbf{e}_n^H\mathbf{e}_m\mathbf{e}_n) \\ \mathbf{w}_n^c = \alpha_n\mathbf{e}_n \end{cases}, \tag{23}$$

where $\mathbf{e}_m, \mathbf{e}_n, \alpha_m, \alpha_n$ are defined in Proposition 1. Surprisingly, in what follows, we show that $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is indeed the optimal solution of both the optimization problem in (4) and (10). Firstly, we show that $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is feasible for both (4) and (10). It is observed that $\{\mathbf{w}_m^c, \mathbf{w}_n^c\} \in \mathcal{C}_D$, i.e., $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is a feasible solution of (4). Since $Q(u) \le \frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2}$, we also have $\{\mathbf{w}_m^c, \mathbf{w}_n^c\} \in \mathcal{C}_N$, i.e., $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is also a feasible solution of (10). Secondly, we show that $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is optimal for both (4) and (10). Suppose that there exists a pair of precoding vectors $\{\mathbf{w}_m^x, \mathbf{w}_n^x\} \in \mathcal{C}_N$ which satisfies

$$\|\mathbf{w}_m^x\|^2 + \|\mathbf{w}_n^x\|^2 < \|\mathbf{w}_m^c\|^2 + \|\mathbf{w}_n^c\|^2. \tag{24}$$

Since $\{\mathbf{w}_m^x, \mathbf{w}_n^x\} \in \mathcal{C}_N$ and $\mathcal{C}_N \subseteq \mathcal{C}_D$, then $\{\mathbf{w}_m^x, \mathbf{w}_n^x\} \in \mathcal{C}_D$. By evaluating the right side of (24), we obtain

$$\|\mathbf{w}_m^x\|^2 + \|\mathbf{w}_n^x\|^2 < \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} + \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2}\frac{1+r_n^{MT}}{1+r_n^{MT}\sin^2\theta}. \tag{25}$$

Note that the right hand of (25) is the optimal value of Lagrangian dual optimization of DPC optimization problem in (4). Hence, the following holds

$$\|\mathbf{w}_m^x\|^2 + \|\mathbf{w}_n^x\|^2 < d^{D^*}. \tag{26}$$

This contradicts with the weak duality of DPC optimization problem in (4). Hence, we conclude that $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is the optimal solution of (4), and $P_{m,n}^{\text{DPC}} = d^{D^*}$.

On the other hand, since $\mathcal{C}_N \subseteq \mathcal{C}_D$, it holds that

$$P_{m,n}^{\text{NOMA}} \ge P_{m,n}^{\text{DPC}} = d^{D^*}. \tag{27}$$

By combining (26) and (27),

$$\|\mathbf{w}_m^x\|^2 + \|\mathbf{w}_n^x\|^2 < P_{m,n}^{\text{NOMA}}. \tag{28}$$

This contradicts with the definition of the NOMA optimization problem in (10). Hence, we conclude that $\{\mathbf{w}_m^c, \mathbf{w}_n^c\}$ is also the optimal solution of (10).

Based on these evidence, we confirm that $P_{m,n}^{\text{NOMA}} = P_{m,n}^{\text{DPC}}$. By using Definition 1, it is concluded that $\mathbf{h}_m, \mathbf{h}_n$ are quasi-degraded and the proof is completed. □

Remark 1: Note that $Q(u) \in [1, \infty]$ is monotonically decreasing with respect to $u \in [0, 1]$. Hence, by assuming the decoding order in the last section, $\|\mathbf{h}_m\|^2 \geq \|\mathbf{h}_n\|^2$ becomes a necessary condition that the broadcast channels $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded.

Remark 2: Given the ratio $\frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2}$, the broadcast channels becomes quasi-degraded by simply choosing the angle that satisfies the equality $u \geq u_T$[3], where u_T is the threshold value which ensures $Q(u_T) = \frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2}$. The value of u_T can be obtained efficiently by Newton's iterative method, or alternatively directly calculated by applying the root formula of the cubic equation.

To better illustrate the concept of quasi-degraded channels, we provide the following example.

Example 1: Assume that the channel realizations are given as

$$\mathbf{h}_m = [2,1]^T, \mathbf{h}_n = [1,0]^T, r_m^{MT} = 1, r_n^{MT} = 1.$$

We assume the decoding order (n, m), i.e., user m needs to decode the message intended for user n before decoding its own.

Since $u = \cos^2\theta = \frac{\mathbf{h}_n^H \mathbf{h}_m \mathbf{h}_m^H \mathbf{h}_n}{\|\mathbf{h}_m\|^2 \|\mathbf{h}_n\|^2} = 0.8$, and $Q(u) = 1.833 \leq \frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2} = 5$, based on Proposition 2, we can conclude that these broadcast channels $\mathbf{h}_m$ and $\mathbf{h}_n$ with respect to r_m^{MT} and r_n^{MT} are quasi-degraded. According to Proposition 1,

$$\mathbf{w}_m = [0.333, 0.333]^T, \mathbf{w}_n = [1.054, 0]^T.$$

It is easy to check that the constraints in (10) are all satisfied. Consequently, the power consumption is, $P^{\text{NOMA}} = \|\mathbf{w}_m\|^2 + \|\mathbf{w}_n\|^2 = 1.333$. On the other hand, the optimal power consumption using DPC is

$$P^{\text{DPC}} = \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} + \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \frac{1 + r_n^{MT}}{1 + r_n^{MT} \sin^2\theta} = 1.333.$$

It is observed that $P^{\text{NOMA}} = P^{\text{DPC}}$, which means NOMA can achieve the optimal performance provided that the channels are quasi-degraded.

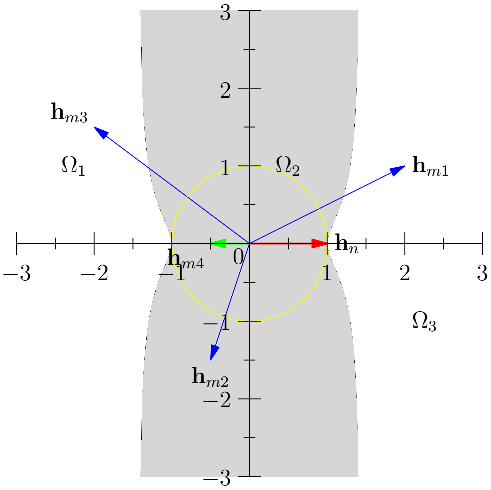


Fig. 1: Quasi-degraded region of $\mathbf{h}_m$ of 2-dimensional quasi-degraded channels, for fixed $\mathbf{h}_n$.

In Fig. 1, the two-dimensional quasi-degraded region of $\mathbf{h}_m$ with a *fixed* $\mathbf{h}_n = [1,0]^T$ is illustrated. First, since the decoding order is (n, m), $\|\mathbf{h}_m\|^2 \geq 1$ becomes a necessary condition of quasi-degradation in this example, according to Remark 1. Therefore, $\mathbf{h}_{m4} = [-0.5, 0]^T$ and $\mathbf{h}_n$ are not quasi-degraded[4]. Second, it is easy to check that $(\mathbf{h}_{m1}, \mathbf{h}_n)$ and $(\mathbf{h}_{m3}, \mathbf{h}_n)$ are both quasi-degraded. Finally, it is also checked that $(\mathbf{h}_{m2}, \mathbf{h}_n)$ is not quasi-degraded. In general, the curve $Q(u) - \frac{\|\mathbf{h}_m\|^2}{\|\mathbf{h}_n\|^2} = 0$ divides $\mathbb{R}^2$ into three subsets, Ω_1, Ω_2, and Ω_3. It is shown that $(\mathbf{h}_m, \mathbf{h}_n)$ is quasi-degraded, if and only if $\mathbf{h}_m \in \Omega_1 \cup \Omega_3$, for the considered $\mathbf{h}_n$, $\mathbf{h}_n = [1,0]^T$.

IV. Hybrid NOMA precoding with user pairing

In this section, we first propose a hybrid NOMA precoding algorithm, by focusing on the case with two users. Based on the closed-form expressions of H-NOMA precoding, a user pairing algorithm is then proposed.

A. Hybrid NOMA (H-NOMA) Precoding Algorithm

The motivations behind proposing the hybrid NOMA precoding scheme are as follows:

1) By using Definition 1 and Proposition 1, NOMA is optimal and a closed-form solution of its precoding exists, when the channels are quasi-degraded. When this is not true, by using NOMA, performance loss is inevitable and closed-form solutions cannot be obtained. Hence, the advantage of the closed-form solution in Proposition 1 cannot be exploited, and the precoding vectors can only be obtained by solving the optimization problem via iterative algorithms [24], but with a high computational complexity.

2) From Proposition 2, one can conclude that the channels are quasi-degraded with a much lower probability, if the channel vectors are orthogonal or close to orthogonal.

3) Compared to NOMA, ZFBF is optimal (or near optimal) when the channels are orthogonal (or quasi orthogonal). It is also noted that the closed-form precoding vectors of ZFBF are given in (7), thus can be efficiently calculated .

By combining the advantage of NOMA and ZFBF, the H-NOMA precoding algorithm is proposed. The key idea is that BS uses the NOMA approach if the channels are quasi-degraded, otherwise ZFBF is used. The H-NOMA precoding algorithm at the BS is described in Algorithm 1, where S is the control bit for switching between two modes. Therefore, user m can choose its decoder properly according to S. Consequently, the required power for transmitting messages

Algorithm 1 Hybrid NOMA precoding algorithm (H-NOMA)

```
INPUT: $\mathbf{h}_m, \mathbf{h}_n, r_m^{MT}, r_n^{MT}$, $\|\mathbf{h}_m\| \geq \|\mathbf{h}_n\|$
OUTPUT: $\mathbf{w}_m, \mathbf{w}_n, S$
if $\mathbf{h}_m$ and $\mathbf{h}_n$ w.r.t. $(r_m^{MT}, r_n^{MT})$ are quasi-degraded then
    Calculate $\mathbf{w}_m, \mathbf{w}_n$ by (12), $S = 1$
else
    Calculate $\mathbf{w}_m, \mathbf{w}_n$ by (7), $S = 0$
end if
```

[3] Mathematically, $\theta \leq \arccos\sqrt{u_T}$ or $\theta \geq \pi - \arccos\sqrt{u_T}$

[4] In fact, they are quasi-degraded by assuming a reverse decoding order

to a paired users m and n using the proposed H-NOMA precoding algorithm can be expressed as

$$P_{m,n}^{\mathrm{H-NOMA}} = \begin{cases} P_1, & \text{if} \quad u \geq u_T, \\ P_2, & \qquad\quad \text{else}, \end{cases} \tag{29}$$

where

$$\begin{cases} P_1 = \dfrac{r_n^{MT}}{\|\mathbf{h}_n\|^2} + \dfrac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \dfrac{1 + r_n^{MT}}{1 + r_n^{MT} \sin^2 \theta} \\ P_2 = \dfrac{1}{\sin^2 \theta} (\dfrac{r_m^{MT}}{\|\mathbf{h}_m\|^2} + \dfrac{r_n^{MT}}{\|\mathbf{h}_n\|^2}) \end{cases}. \tag{30}$$

By using (30), we can have the following properties.

1) Note that

$$\frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} \leq \frac{1}{\sin^2 \theta} \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2}, \tag{31}$$

and

$$\begin{aligned} \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \frac{1 + r_n^{MT}}{1 + r_n^{MT} \sin^2 \theta} &\leq \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \frac{1 + r_n^{MT}}{\sin^2 \theta + r_n^{MT} \sin^2 \theta} \\ &\leq \frac{r_m^{MT}}{\|\mathbf{h}_m\|^2} \frac{1}{\sin^2 \theta}. \end{aligned} \tag{32}$$

By substituting (31) and (32) into (30), we conclude that

$$P_1 \leq P_2,$$

and the equality holds when $\theta = \frac{\pi}{2}$.

2) If the weak user is paired with another user to have quasi-degraded channels, the equivalent power consumption of this user is $\frac{r_n^{MT}}{\|\mathbf{h}_n\|^2}$. If the weak user is not paired to form quasi-degraded or orthogonal channels, the equivalent power consumption of this user is $\frac{1}{\sin^2 \theta} \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2}$. Consequently, the power consumption difference between these two modes can be written as

$$\frac{1}{\sin^2 \theta} \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} - \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2} = \frac{r_n^{MT}}{\|\mathbf{h}_n\|^2 \tan^2 \theta}.$$

Therefore, we conclude that a weak user with small $\|\mathbf{h}_n\|$ or small θ can lead to a significant performance loss if it is not paired forming quasi-degraded channels or orthogonal channels.

These properties can be utilized for developing an efficient user pairing algorithm, which will be discussed in details in the next subsection.

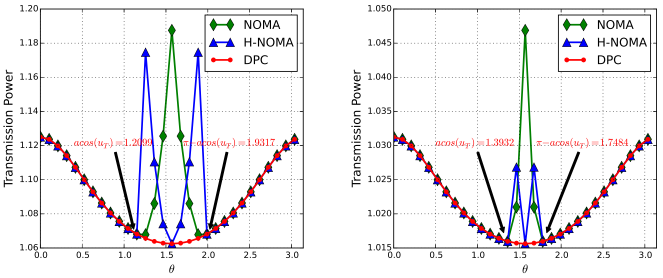
To further understand the performance of the proposed H-NOMA precoding, a two-dimensional example is illustrated.

Example 2: Assume that we have the following channel realization

$$\begin{aligned} &\mathbf{h}_m = \sigma_m [\cos \theta, \sin \theta]^T, \mathbf{h}_n = \sigma_n [1, 0]^T, \\ &r_m^{MT} = 1, r_n^{MT} = 1, \sigma_n = 1. \end{aligned}$$

Again, we assume the decoding order (n, m).

For this example, it is easy to verify that θ is the angle between $\mathbf{h}_m$ and $\mathbf{h}_n$. The total required power of this example versus θ is illustrated in Fig. 2. It is evident that DPC is optimal for all θ, and H-NOMA precoding outperforms the optimal NOMA introduced in [24], when θ is close to $\pi/2$, while there exists a performance loss at both side peak. It is also observed that both H-NOMA and the optimal NOMA [24] are optimal when $\mathbf{h}_m$ and $\mathbf{h}_n$ are quasi-degraded. Note that the performance loss compared to DPC becomes small as σ_m becomes larger (see Fig. 2a and Fig. 2b).



(a) $\sigma_m = 4$

(b) $\sigma_m = 8$

Fig. 2: Total power consumption versus the angle between $\mathbf{h}_m$ and $\mathbf{h}_n$.

The advantages of H-NOMA precoding can be summarised as follows:

1) The computation complexity is low. Since the closed-form expressions for the solutions can be derived, the complexity can be reduced to $\mathcal{O}(N)$, while the complexity of optimal NOMA is $\mathcal{O}(N^3)$.
2) H-NOMA precoding suffers almost no performance degradation compared to the optimal NOMA scheme, as illustrated in Fig.2.

B. Sequential User Pairing Algorithm (SUPA)

In this subsection, we develop a user pairing algorithm for H-NOMA precoding, named as Sequential User Pairing Algorithm (SUPA), to further reduce the total power consumption. In order to obtain a minimum total power consumption at the BS, some useful properties of H-NOMA precoding algorithm are firstly exploited, followed by the basic principles which an efficient pairing algorithm should abide. By focusing on the power consumption expression for H-NOMA precoding in (29), it is worthwhile noticing the following properties:

1) When the channels are orthogonal or quasi-degraded, the total power consumption is minimized and there is no performance loss compared to DPC.
2) When the channels are not orthogonal or quasi-degraded, performance loss appears.

It is worth pointing out that users can have strong correlated channels, regardless whether the channels are quasi-degraded or not. However, the impact of the channel correlation is different for the two cases, as illustrated in the following:

1) When the channels are quasi-degraded, the channel correlation does not affect the system performance.
2) When the channels are not quasi-degraded, the performance loss compared to DPC becomes significant, if the norms of the channels are small or the correlation coefficient u is large.

It is important to note that it is less likely to have a large correlation coefficient u when the channels are not orthogonal or quasi-degraded. Specifically, as revealed in Proposition 2 , when u is larger, $Q(u)$ becomes smaller, and the channels are more likely to be quasi-degraded. In this case, performance loss can be avoided. Then, by applying these properties, the users can be paired by following the basic principles given below:

a) The users having orthogonal channels should be paired in a group and be served by using ZFBF, since this strategy yields the optimal solution for orthogonal channels.
b) One user should be paired with another user to form quasi-degraded channels, if possible, since $P_1 \leq P_2$ always holds given a fixed θ, according to (30).
c) The weak users, i.e. those having small $\|\mathbf{h}\|^2$ and often being far away from the BS, should be paired first, since these users can lead to a significant performance loss when they cannot be paired forming either orthogonal or quasi-degradation channels, according to (30). Simulation results provided in the next section also confirm that there is a significant performance loss, when the remaining users have weak channel gains.
d) If one user cannot find a partner to form orthogonal or quasi-degraded channels, then it should be paired with the user with the minimum channel correlation and be served with ZFBF. The reason is that there is a positive correlation between the performance loss caused by ZFBF and the channel correlation, by observing the equation in (8).

Hence, a heuristic user pairing algorithm named as Sequential User Pairing Algorithm (SUPA) is proposed in the following. Algorithm 2 provides a detailed description of the proposed SUPA. The number of users K is assumed to be even for ease of illustration. When the number of users is odd, the last remained user can be trivially served by OMA. Specifically, if the channel vector and the corresponding target SNR of the last remained user are $\mathbf{h}_l$ and r_l^{MT}, respectively, then the precoding vector is

Algorithm 2 Sequential User Pairing Algorithm (SUPA)

```
INPUT: $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K$ and $r_1^{MT}, r_2^{MT}, ..., r_K^{MT}$
OUTPUT: $\Pi$
Sort $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K$, such that $\|\mathbf{h}_{i_1}\| \leq ... \leq \|\mathbf{h}_{i_K}\|$
for $t = 1 : K-1$ do
    for $v = t+1 : K$ do
        if users $i_t$ and $i_v$ have orthogonal channels, and user $i_v$ has not been paired then
            $\Pi = \Pi \cup \{(i_v, i_t, 0)\}$, break;
        else if users $i_t$ and $i_v$ have quasi-degraded channels, and user $i_v$ has not been paired then
            $\Pi = \Pi \cup \{(i_v, i_t, 1)\}$, break;
        end if
    end for
    find $v = \mathrm{argmin}_v \ \frac{\mathbf{h}_{i_v}^H \mathbf{h}_{i_t} \mathbf{h}_{i_t}^H \mathbf{h}_{i_v}}{\|\mathbf{h}_{i_v}\|^2 \|\mathbf{h}_{i_v}\|^2}$
    $\Pi = \Pi \cup \{(i_v, i_t, 0)\}$
end for
```

$$\mathbf{w}_l = \frac{\sqrt{r_l^{MT}}}{\|\mathbf{h}_l\|^2}\mathbf{h}_l,$$

and the minimum power consumption of user l can be written as

$$P_l = \|\mathbf{w}_l\|^2 = \frac{r_l^{MT}}{\|\mathbf{h}_l\|^2}.$$

Note that the computational complexity of determining channels to be orthogonal or quasi-degraded is only $\mathcal{O}(N)$, and the number of "for" loops is $\frac{1}{2}K(K-1)$. Therefore, the overall computational complexity of SUPA is $\mathcal{O}(K^2N)$. The output of Algorithm 2, Π, is termed the pairing configuration in this paper, which can be mapped to a permutation κ of K elements. Mathematically, Π can be defined as

$$\Pi = \{(k, \kappa(k), S) \mid \|\mathbf{h}_k\| \geq \|\mathbf{h}_{\kappa(k)}\|, S = 0, 1\}.$$

The number of all the possible Π is

$$(K-1)!! = (K-1)(K-3)...1,$$

since it is equivalent to the number of ways to select $K/2$ disjoint pairs from K items. For example, with $K = 4$ users, after ignoring the third item S in Π, we have the following 3 pairing configurations

$$\Big\{\{(1,2),(3,4)\},\{(1,3),(2,4)\},\{(1,4),(2,3)\}\Big\}.$$

and the corresponding permutations are

$$\Big\{\{2,1,4,3\},\{3,4,1,2\},\{4,3,2,1\}\Big\}.$$

Algorithm 3 H-NOMA/SUPA

```
INPUT: $(\mathbf{h}_1, ..., \mathbf{h}_K)$, $(s_1 ..., s_K)$ and $(r_1^{MT}, ..., r_K^{MT})$
OUTPUT: $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{K/2}$
 1: Execute SUPA to obtain pairing configuration $\Pi$
 2: for $i = 1 : K/2$ do
 3:     $(m, n, S) \leftarrow (\Pi(i)_1, \Pi(i)_2, \Pi(i)_3)$
 4:     if $S = 1$ then
 5:         Calculate $\mathbf{w}_m, \mathbf{w}_n$ by (12)
 6:     else if $S = 0$ then
 7:         Calculate $\mathbf{w}_m, \mathbf{w}_n$ by (7)
 8:     end if
 9:     $\mathbf{x}_i \leftarrow \mathbf{w}_m s_m + \mathbf{w}_n s_n$
10:     Transmit $\mathbf{x}_i$ in $i$-th time/frequency slot
11: end for
```

C. H-NOMA/SUPA: A Practical Transmission Scheme

A practical transmission scheme can be proposed by combining the H-NOMA precoding algorithm with SUPA, which is described in detail in Algorithm 3. The notation $\Pi(i)_j$ stands for the j-th item in the i-th element in the set Π. Note that TDMA/FDMA is considered in Algorithm 3 (Line No. 10), it is efficient when N is small. When considering the scenario with sufficient number of antennas at the BS, e.g., $N \geq K$, the overall performance can be further boosted by combining SDMA with the proposed H-NOMA/SUPA scheme, which is beyond the scope of this paper. By exploiting the advantage of the closed-form expressions of the precoding vectors and the effective implementation of SUPA, the computational complexity of the proposed H-NOMA/SUPA scheme is only $\mathcal{O}(K^2N)$, whereas the complexity of the optimal NOMA, in combination with exhaustive search, is $\mathcal{O}((K-1)!!N^3)$.

V. Performance Analysis

In this section, we first characterize the probability for Rayleigh channels to be quasi-degraded. Then, the performance of H-NOMA precoding is investigated, by using the average power consumption and the outage probability. Finally, the performance of some existed precoding algorithms, including ZFBF and DPC, is also investigated and compared with the proposed H-NOMA precoding.

A. Quasi-degradation Probability

Lemma 2. *The probability density function (pdf) of $\xi = \frac{\|\mathbf{g}_1\|^2}{\|\mathbf{g}_2\|^2}$, where $\mathbf{g}_1 \sim \mathcal{CN}(0, 2\mathbf{I}_N)$ and $\mathbf{g}_2 \sim \mathcal{CN}(0, 2\mathbf{I}_N)$ are independent ($N \geq 2$), is given by*

$$f_\xi(x) = \frac{(2N-1)!}{((N-1)!)^2}\frac{x^{N-1}}{(x+1)^{2N}}. \tag{33}$$

Proof: See Appendix A □

Lemma 3. *The pdf of $u = \cos^2\theta = \frac{(\mathbf{g}_1^H\mathbf{g}_2\mathbf{g}_2^H\mathbf{g}_1)}{\|\mathbf{g}_1\|^2\|\mathbf{g}_2\|^2}$, where $\mathbf{g}_1 \sim \mathcal{CN}(0, 2\mathbf{I}_N)$ and $\mathbf{g}_2 \sim \mathcal{CN}(0, 2\mathbf{I}_N)$ are independent ($N \geq 2$), is given by*

$$f_u(x) = (N-1)(1-x)^{N-2}, \tag{34}$$

which is a Beta distribution with parameter $(1, N-1)$ [27].

Proof: See Appendix B □

Based on Lemma 2 and Lemma 3, we can provide the following theorem for calculating the quasi-degradation probability.

Theorem 1. *For independent Rayleigh distributed broadcast channels, i.e., $\mathbf{h}_1 \sim \mathcal{CN}(0, 2\sigma_1^2\mathbf{I}_N), \mathbf{h}_2 \sim \mathcal{CN}(0, 2\sigma_2^2\mathbf{I}_N)$, and with a pair of fixed target SNR, (r_1, r_2), the probability of $\mathbf{h}_1, \mathbf{h}_2$ to be quasi-degraded is*

$$P_{QD} = \frac{(2N-1)!}{((N-1)!)^2}\sum_{k=0}^{N-1}\tbinom{N-1}{k}\frac{(-1)^k}{N+K}G(k, r_1, r_2, q, N), \tag{35}$$

where the channel quotient q is defined as $q = \frac{\sigma_1^2}{\sigma_2^2}$ and

$$\begin{aligned} G(k, r_1, r_2, q, N) &= \int_0^1 \frac{1}{(1+q^{-1}Q(u))^{N+k}}f_u(u)du, \\ Q(u) &= \frac{1+r_1}{u} - \frac{r_1u}{(1+r_2-r_2u)^2}. \end{aligned} \tag{36}$$

Proof: See Appendix C. □

B. Average Power Consumption

Theorem 2. *For independent Rayleigh distributed broadcast channels $\mathbf{h}_1 \sim \mathcal{CN}(0, 2\sigma_1^2\mathbf{I}_N), \mathbf{h}_2 \sim \mathcal{CN}(0, 2\sigma_2^2\mathbf{I}_N)$ and with a pair of fixed target SNR, (r_1, r_2), the expectation of the required power using DPC is given by*

$$\mathcal{E}\{P_{1,2}^{\text{DPC}}\} = \frac{r_2}{2\sigma_2^2(N-1)} + \frac{r_1}{2\sigma_1^2(N-1)}A, \tag{37}$$

where

$$\begin{aligned} A &= (-1)^{N-2}(1+r_2)(N-1)r_2^{1-N}B, \\ B &= \begin{cases} \ln(1+r_2) + \sum\limits_{k=1}^{N-2}\frac{(-1)^k}{k}r_2^k, & N \geq 3 \\ \ln(1+r_2) & N = 2 \end{cases}. \end{aligned} \tag{38}$$

Proof: See Appendix D. □

Theorem 3. *For independent Rayleigh distributed broadcast channels $\mathbf{h}_1 \sim \mathcal{CN}(0, 2\sigma_1^2\mathbf{I}_N), \mathbf{h}_2 \sim \mathcal{CN}(0, 2\sigma_2^2\mathbf{I}_N)$ and with a pair of fixed target SNR, (r_1, r_2), the expectation of the required power using ZFBF is given by*

$$\mathcal{E}\{P_{1,2}^{\text{ZF}}\} = \frac{r_2}{2\sigma_2^2(N-2)} + \frac{r_1}{2\sigma_1^2(N-2)} \tag{39}$$

Proof: See Appendix E. □

Theorem 4. *For independent Rayleigh distributed broadcast channels $\mathbf{h}_1 \sim \mathcal{CN}(0, 2\sigma_1^2\mathbf{I}_N), \mathbf{h}_2 \sim \mathcal{CN}(0, 2\sigma_2^2\mathbf{I}_N)$ and with a pair of fixed target SNR, (r_1, r_2), the expectation of the required power using H-NOMA precoding algorithm is given by*

$$\begin{aligned} \mathcal{E}\{P_{1,2}^{\text{H-NOMA}}\} &= \frac{r_2}{2\sigma_2^2(N-1)}\int_0^\infty F_2(x)f_\xi(x)dx \\ &+ \frac{r_1}{2\sigma_1^2(N-1)}\int_0^\infty F_1(x)f_\xi(x)dx, \end{aligned} \tag{40}$$

where

$$\begin{aligned} F_2(x) &= (1-Q^{-1}(qx))^{N-1} + \frac{N-1}{N-2}(1-(1-Q^{-1}(qx))^{N-2}), \\ F_1(x) &= (-1)^{N-2}(1+r_2)(N-1)r_2^{1-N}C(x)+ \\ &\quad \frac{N-1}{N-2}(1-(1-Q^{-1}(qx))^{N-2}), \\ C(x) &= \begin{cases} \ln(1+r_2(1-Q^{-1}(qx))) \\ +\sum\limits_{k=1}^{N-2}\frac{(-1)^k}{k}r_2^k(1-Q^{-1}(qx))^k, & N \geq 3 \\ \ln(1+r_2(1-Q^{-1}(qx))), & N = 2 \end{cases}. \end{aligned}$$

Note that here, $Q^{-1}(x) = 1$ if $x < 1$.

Proof: See Appendix F. □

A closed-form solution to the integral in (40) is difficult to obtain. However, a closed-form approximation for this integral can be derived by applying the following

$$\int_0^\infty (1-Q^{-1}(qx))^k f_\xi(x)dx \approx [1-Q^{-1}(q)]^k, \tag{41}$$

where $k = 0, 1, ..., N-1$. Note that $Q^{-1}(qx) \approx (1+r_1)/(qx)$ holds when q is large, by applying Taylor's expansion, one can trivially check that the approximation is accurate for large q and large N. Simulations show that this approximation works quite well when $q \geq 2$.

To measure the gap between H-NOMA precoding and DPC , we introduce the following theorem.

Theorem 5. *The average required power gap between the proposed H-NOMA precoding and DPC vanishes as $q^{-1} \to 0$, i.e.,*

$$\lim_{q^{-1}\to 0}\mathcal{E}\{P_{1,2}^{\text{H-NOMA}}\} - \mathcal{E}\{P_{1,2}^{\text{DPC}}\} = 0. \tag{42}$$

Moreover, the gap is a second order infinitesimal of q^{-1}, i.e.,

$$\lim_{u_T\to 0}\frac{\mathcal{E}\{P_{1,2}^{\text{H-NOMA}}\}-\mathcal{E}\{P_{1,2}^{\text{DPC}}\}}{q^{-2}}=\delta, \tag{43}$$

where

$$\delta=(\frac{r_2}{4\sigma_2^2}+\frac{r_1}{4\sigma_1^2(1+r_2)})(1+r_1)^2\frac{(N+1)N}{(N-1)(N-2)}.$$

Proof: See Appendix G. □

Theorem 5 reveals that the performance loss of H-NOMA precoding vanishes quickly as q becomes large, compared to the optimal DPC.

Regarding the gap between H-NOMA precoding and ZFBF, by using Theorems 3 and 4, we conclude that: a) the gap becomes larger as q increases. b) the gap does not vanish even when $q\to 1$. By combining Theorems 3, 4 and 5, it is also worthwhile to note that all gaps (including those between H-NOMA precoding and ZFBF, H-NOMA precoding and DPC, and between ZFBF and DPC) vanish when $N\to\infty$.

C. Outage Probability

To further analyse the asymptotic performance of different algorithms for the QoS optimization problem, we first define the outage probability with respect to a maximal transmit power as follows:

Definition 2 (Outage Probability)**.** *The outage probability with respect to a maximal transmit power P is defined as*

$$Pr_{out}(P)=Pr\{P_{1,2}\geq P\},$$

where $P_{1,2}$ denotes the minimum required power to support the predetermined QoS requirement for a paired users with channels $(\mathbf{h}_1,\mathbf{h}_2)$.

Consequently, the power diversity can be defined as follows:

Definition 3 (Power Diversity)**.** *The power diversity d is defined as*

$$d=-\lim_{P\to\infty}\frac{\ln Pr_{out}(P)}{\ln P}.$$

There is a strong connection between the two concepts (Definition 2 and Definition 3) and the conventional concepts about the outage probability and diversity [28]. For example, the outage probability defined in this paper indicates that there is not sufficient power at the BS to build desirable precoding in order to satisfy the predefined rates. In other words, when the outage event in Definition 2 happens, then, the targeted data rates cannot be supported, which is the same with the conventional definition of outage probability. On the other side, the power diversity defined in this paper indicates how fast the outage probability drops with respect to the power, which is similar to the conventional definition of diversity.

The following theorem provides the asymptotic behaviour of the outage probability achieved by H-NOMA precoding.

Theorem 6. *The proposed H-NOMA precoding algorithm can achieve the optimal power diversity of N, i.e.,*

$$-\lim_{P\to\infty}\frac{\ln Pr_{out}^{\text{H-NOMA}}(P)}{\ln P}=-\lim_{P\to\infty}\frac{\ln Pr_{out}^{\text{DPC}}(P)}{\ln P}=N, \tag{44}$$

while ZFBF precoding algorithm can only achieve a power diversity of $N-1$, i.e.,

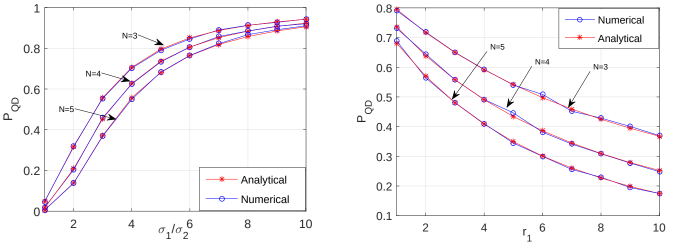
$$-\lim_{P\to\infty}\frac{\ln Pr_{out}^{\text{ZF}}(P)}{\ln P}=N-1, \tag{45}$$

Proof: See Appendix H. □

Theorem 6 reveals that the proposed H-NOMA precoding is actually asymptotic optimal, while ZFBF is only asymptotic suboptimal.

VI. Simulation Results and Discussion

In this section, the performance of the proposed H-NOMA precoding algorithm is investigated and validated through computer simulations by using criteria, as the quasi-degradation probability, the average power consumption as well as the outage probability. The performance of the proposed practical transmission scheme (H-NOMA/SUPA) is also simulated, in comparison with traditional transmission schemes.



(a) $r_1=r_2=1$, and $N=3,4,5$ (b) $r_2=1$, $\frac{\sigma_1}{\sigma_2}=5$ and $N=3,4,5$.

Fig. 3: Quasi-degradation probability.

In Figs. 3a and 3b, the quasi-degradation probability achieved by the proposed H-NOMA precoding is plotted as a function of $\frac{\sigma_1}{\sigma_2}$ and r_1, respectively, for several values of N. The analytical results are based on Theorem 1. As it can be observed, the probability for channels to be quasi-degraded is monotonically increasing with $\frac{\sigma_1}{\sigma_2}$. Specifically, when $\frac{\sigma_1}{\sigma_2}=10$, it is already around 90%, which means that users' channel vectors become quasi-degraded very frequently. Note that $\frac{\sigma_1}{\sigma_2}=10$ means that the ratio of two variances of channel coefficients is 100, and the ratio of the users' distances is only around 3.2 for the case with the path loss component of 4. It can be also observed that the analytical results fit perfectly with the numerical ones, which validates the accuracy of Theorem 1.

In Fig. 4, four different precoding schemes, including DPC, ZFBC, the NOMA scheme proposed in [24] and H-NOMA precoding, are compared by using the average power consumption as the criterion, where the results for several values of N are also shown in the figure. The analytical results are based on Theorem 2, 3 and 4. The curves for H-NOMA precoding are plotted by using the approximation in (41). The optimal NOMA is obtained via iterative algorithm introduced in [24]. It is evident from this figure that the proposed H-NOMA precoding algorithm results in a slight performance loss, compared to the optimal DPC scheme. In comparison with ZFBF, the use of H-NOMA precoding yields a significant performance gain, particularly when N is small. The performance of ZFBF is

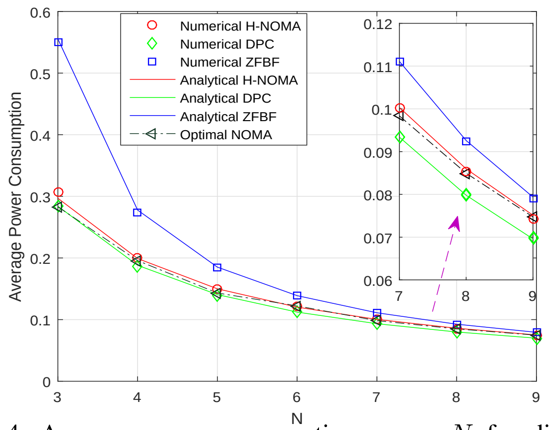
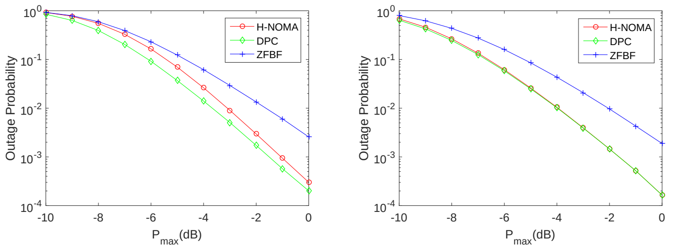


Fig. 4: Average power consumption versus N for different precoding algorithms, with $r_1 = r_2 = 1$, and $\frac{\sigma_1}{\sigma_2} = 3$.

significantly improved by increasing N, and achieves a similar performance to that of H-NOMA precoding and DPC, which is consistent to the conclusion made in [8]. The analytical results and the numerical results fit perfectly for ZFBF and DPC, and the approximation for H-NOMA precoding is also very tight, even for small values of N. It can also be observed that the optimal NOMA in [24] achieves similar performance with that of low-complexity H-NOMA precoding.



(a) $r_1 = r_2 = 1$, $N = 5$, and $\frac{\sigma_1}{\sigma_2} = 2$. (b) $r_1 = r_2 = 1$, $N = 5$, and $\frac{\sigma_1}{\sigma_2} = 4$.

Fig. 5: Outage probability versus P_{max} for different precoding algorithms.

In Figs. 5a and 5b, the outage performance achieved by different precoding algorithms is depicted as a function of the transmission power. Since the Gaussian noise n_i in this paper is assumed normalized, P_{max} is actually the signal-to-noise ratio, hence, is represented in dB. By increasing the transmission power, the outage performance of all schemes is improved, since the BS acquires more power to build desirable precoding vectors and satisfy the rate constraints. It is observed that H-NOMA precoding and DPC can achieve better outage performance than ZFBF, which can be explained as follows. As deduced by Theorem 6, H-NOMA precoding and DPC can achieve the maximal power diversity of N while ZFBF can only achieve a suboptimal power diversity of $N - 1$. It is also worthwhile noting that the gap between H-NOMA precoding and DPC vanishes as $\frac{\sigma_1}{\sigma_2}$ becomes large.

In order to illustrate the performance of the proposed SUPA, a downlink communication system is assumed. The BS is located at the center of a disk with radius R, and K mobile users are randomly deployed within the disk with the uniform distribution. In this case, the variance of the channel between the BS and the i-th user is modelled as [21], [29]

$$2\sigma_i^2 = \min(d_i^{-\alpha}, d_0^{-\alpha}),$$

where d_i is the distance between the BS and the i-th user, α is the path loss exponent, and the parameter d_0 is introduced to avoid the singularity for path loss when the d_i is small.

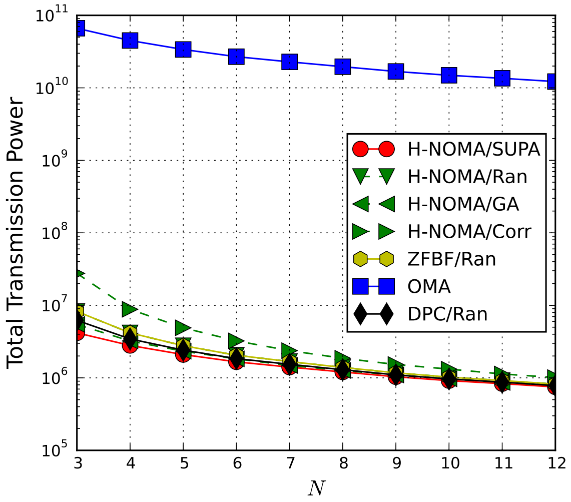


Fig. 6: Total power consumption versus N for several transmission schemes, with $K = 30$, $R = 10, d_0 = 1, \alpha = 3$, and $r_i^{MT} = 1, i = 1, 2, ..., K$.

Fig. 6 illustrates the total power consumption of the BS versus N, to serve $K = 30$ users uniformly deployed in a disk using different transmission schemes. Figs. 7, 8, 9 and 10 show the total power consumption versus K with different choices of the target SINR constraint, as illustrated in the following.

1) In Fig. 7, all users have the same SINR constraint, i.e.,

$$r_i^{MT} = 1, \quad i = 1, 2, ..., K.$$

2) In Fig. 8, the target SINR constraint is set as

$$r_i^{MT} = \begin{cases} 1, & \text{if} \quad d_i < \dfrac{R}{2} \\ 0.5, & \text{if} \quad d_i \geq \dfrac{R}{2} \end{cases}.$$

3) In Fig. 9, the target SINR constraint is set as

$$r_i^{MT} = \begin{cases} 1, & \text{if} \quad d_i < \dfrac{R}{2} \\ 0.01, & \text{if} \quad d_i \geq \dfrac{R}{2} \end{cases}.$$

4) In Fig. 10, the target SINR constraint is set as

$$r_i^{MT} = \frac{1}{1 + \sqrt{d_i}}.$$

Fig. 11 plots the sum rate versus total transmission power. The individual rate is optimized according to the max-min problem. Throughout these simulations, i.e., Figs. 6, 7, 8, 9, 10 and 11, the total bandwidth is assumed to be unity. To validate the effectiveness of the proposed SUPA, several other user pairing algorithms are also simulated for comparing, including RANdom Pairing algorithm (Ran), Greedy Algorithm (GA), and CORRelation-based pairing algorithm (CORR) [22], [23].

Specifically, the key idea of GA is to pair users with the minimum power consumption. Mathematically, for a fixed user i, we paired it with user j, if

$$j = \mathrm{argmin}_{j \neq i} \quad P_{i,j}^{\mathrm{H-NOMA}},$$

where $P_{i,j}^{\mathrm{H-NOMA}}$ is defined in (29). Similarly, the key idea of CORR algorithm is to pair users with the maximum channel correlation. Mathematically, for a fixed user i, we pair it with user j, if

$$j = \mathrm{argmax}_{j \neq i} \frac{\mathbf{h}_i^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{h}_i}{\|\mathbf{h}_i\|^2 \|\mathbf{h}_j\|^2}.$$

By observing there figures, we have the following remarks.

1) The proposed practical transmission scheme H-NOMA/SUPA is numerically robust and realizes significant throughput improvement and power reduction, compared with conventional OMA.
2) SUPA outperforms the greedy algorithm, since that GA does not consider the performance loss caused by the weak users.
3) SUPA outperforms all the other pairing algorithms, since they do not take the advantage of quasi-degradation. Note that the proposed SUPA scheme can even outperforms DPC with random user pairing.
4) The correlation-based pairing algorithm performs worse than the random pairing algorithm, since that it breaks the channels orthogonality, i.e., user with orthogonal channels are never paired. Therefore, it can be concluded that the CORR algorithm is not suitable to the proposed precoding scheme, H-NOMA.
5) By using a random user pairing algorithm, the H-NOMA precoding scheme presents nearly the same performance as that of ZFBF. The reason is that the users paired into one group may not have quasi-degraded channels, since the users are paired randomly and the property of quasi-degraded channels is not utilized.

VII. Conclusion

In this paper, we have considered the design of low-complexity transmission schemes, based on NOMA in MISO downlink scenarios. We first have undertaken an in-depth study on the properties of quasi-degraded channels, where closed-form precoding vectors as well as explicit sufficient and necessary condition for channels to be quasi-degraded, were developed. Then, based on these in-depth studies, the H-NOMA precoding algorithm was presented, and then combined with the proposed user pairing algorithm, to yield a practical transmission scheme. Analytical results about the quasi-degradation probability, the average power consumption and the outage probability, have been developed for better evaluating the proposed precoding schemes. Finally, we have validated the analytical results and the efficiency of the proposed low-complexity transmission scheme, by using computer simulations. In this paper, TDMA/FDMA was considered as inter-group transmission scheme, which is efficient for the scenario with $N < K$. When considering the case with sufficient antennas at the BS, i.e., $N \geq K$, an important future direction is to study the system design of combining the proposed H-NOMA/SUPA scheme with advanced inter-group transmission schemes, such as SDMA, including ZFBF, relaxed ZFBF, leakage-based BF and SINR-based BF.

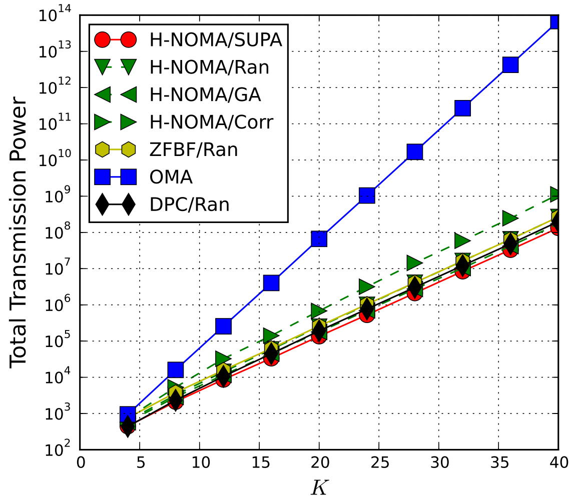


Fig. 7: Total power consumption versus K for several transmission schemes, with $N = 3$, $R = 10, d_0 = 1, \alpha = 3$, and $r_i^{MT} = 1, i = 1, 2, ..., K$.

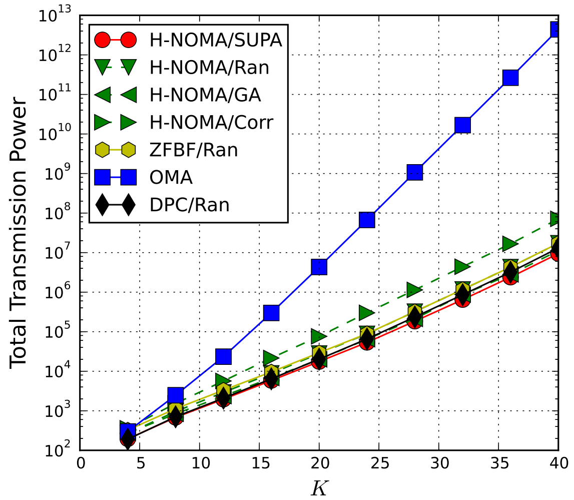


Fig. 8: Total power consumption versus K for several transmission schemes, with $N = 3$, $R = 10, d_0 = 1, \alpha = 3$.

Appendix A
Proof of Lemma 2

Since $\mathbf{g}_i \sim \mathcal{CN}(0, 2\mathbf{I}_N)$, $\|\mathbf{g}_i\|^2$ is a Chi-Square random variable with $2N$ degrees of freedom, i.e.,

$$f_{\|\mathbf{g}_i\|^2}(x) = \frac{1}{2^N (N-1)!} e^{-1/2x} x^{N-1}, \quad i = 1, 2.$$

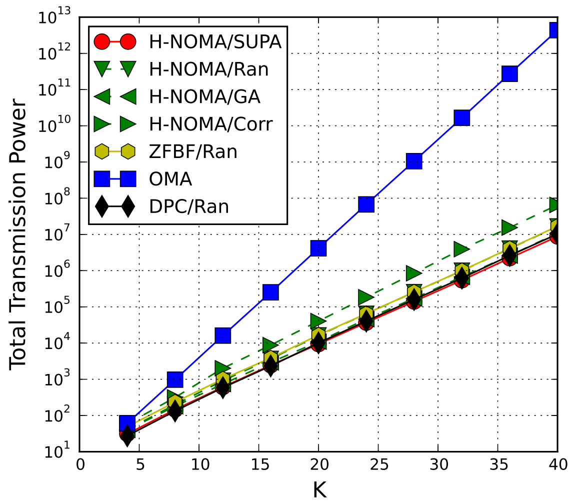


Fig. 9: Total power consumption versus K for several transmission schemes, with $N = 3,\ R = 10, d_0 = 1, \alpha = 3$.

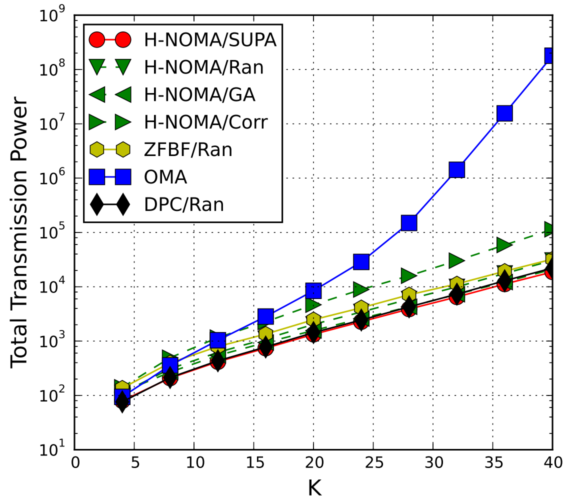


Fig. 10: Total power consumption versus K for several transmission schemes, with $N = 3,\ R = 10, d_0 = 1, \alpha = 3$.

The distribution of the quotient can be evaluated as

$$\begin{aligned} f_\xi(x) &= \int_0^\infty f_{\|\mathbf{g}_i\|^2}(xu) f_{\|\mathbf{g}_i\|^2}(u) u du \\ &= \frac{x^{N-1}}{(2^N(N-1)!)^2} \int_0^\infty e^{-1/2(x+1)u} u^{2N-1} du \\ &= \frac{x^{N-1}}{(2^N(N-1)!)^2} \frac{2^{2N}(2N-1)!}{(x+1)^{2N}} \\ &= \frac{(2N-1)!}{((N-1)!)^2} \frac{x^{N-1}}{(x+1)^{2N}}, \end{aligned}$$

and the proof is completed.

Appendix B
Proof of Lemma 3

We can rewrite u as

$$u = \frac{\mathbf{g}_1^H \mathbf{g}_2 \mathbf{g}_2^H \mathbf{g}_1}{\|\mathbf{g}_1\|^2 \|\mathbf{g}_2\|^2} = \frac{\mathbf{g}_1^H \mathbf{\Pi}_{\mathbf{g}_2} \mathbf{g}_1}{\mathbf{g}_1^H \mathbf{g}_1}, \tag{46}$$

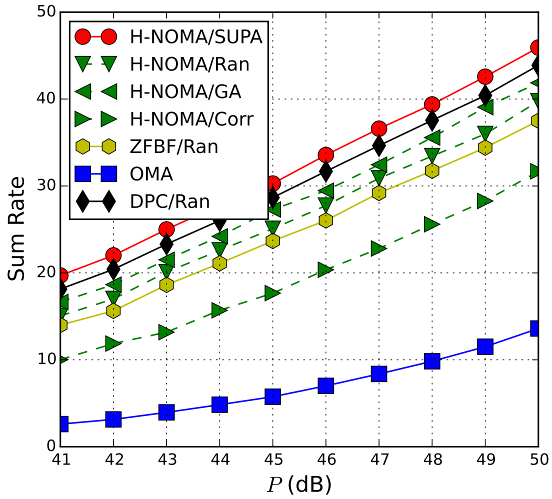


Fig. 11: Sum rate versus total transmission power P for several transmission schemes, with $N = 3, K = 10,\ R = 10, d_0 = 1, \alpha = 3$.

where $\mathbf{\Pi}_{\mathbf{g}_2} = \mathbf{g}_2(\mathbf{g}_2^H\mathbf{g}_2)^{-1}\mathbf{g}_2^H$ is the projection matrix of $\mathbf{g}_2$. Then,

$$\begin{aligned} &Pr\{u \le X\} = Pr\{\frac{\mathbf{g}_1^H \mathbf{\Pi}_{\mathbf{g}_2} \mathbf{g}_1}{\mathbf{g}_1^H \mathbf{g}_1} \le X\} \\ &= Pr\{\frac{\mathbf{g}_1^H \mathbf{\Pi}_{\mathbf{g}_2} \mathbf{g}_1}{\mathbf{g}_1^H (\mathbf{I} - \mathbf{\Pi}_{\mathbf{g}_2}) \mathbf{g}_1} \le \frac{X}{1-X}\} \\ &= \int_0^\infty \int_{\frac{1-X}{X}y}^\infty \frac{t^{N-2} e^{-1/2t}}{2^{N-1}(N-2)!} \frac{e^{-1/2y}}{2} dt dy \\ &= \int_0^\infty \int_0^X \frac{((1/u-1)y)^{N-2}}{2^N(N-2)!} e^{-1/2(1/u)y} y u^{-2} du dy \\ &= \int_0^X (N-1)(1-u)^{N-2} du. \end{aligned} \tag{47}$$

The third equality holds since $\mathbf{g}_1^H \mathbf{\Pi}_{\mathbf{g}_2} \mathbf{g}_1$ and $\mathbf{g}_1^H(\mathbf{I} - \mathbf{\Pi}_{\mathbf{g}_2})\mathbf{g}_1$ are independent Chi-square random variables with degrees of freedom 2 and $2(N-1)$, respectively. The fourth equality is obtained by using $t = y(1/u - 1)$. Therefore, the pdf of u can be written as

$$f_u(x) = (N-1)(1-x)^{N-2},$$

and Lemma 3 is proved.

Appendix C
Proof of Theorem 1

According to Proposition 2, $\mathbf{h}_1, \mathbf{h}_2$ are quasi-degraded if and only if $Q(u) \le \frac{\|\mathbf{h}_1\|^2}{\|\mathbf{h}_2\|^2}$. Denote $\mathbf{g}_1 = \mathbf{h}_1/\sigma_1$ and $\mathbf{g}_2 = \mathbf{h}_2/\sigma_2$. Hence, the quasi-degradation probability can be calculated as

$$\begin{aligned} Pr_{QD} &= Pr\{Q(u) \le \frac{\|\mathbf{h}_1\|^2}{\|\mathbf{h}_2\|^2}\} \\ &= Pr\{\frac{\|\mathbf{g}_1\|^2}{\|\mathbf{g}_2\|^2} \ge q^{-1}Q(u)\} \\ &= \int_0^1 \int_{q^{-1}Q(u)}^\infty \frac{(2N-1)!}{((N-1)!)^2} \frac{x^{N-1}}{(x+1)^{2N}} f_u(u) dx du \\ &= \frac{(2N-1)!}{((N-1)!)^2} \sum_{k=0}^{N-1} \binom{N-1}{k} \frac{(-1)^k}{N+K} G(k, r_1, r_2, q, N), \end{aligned} \tag{48}$$

where the third equality follows by Lemma 2, and the proof is completed.

Appendix D
Proof of Theorem 2

We first calculate the following expectations $\mathcal{E}\{\frac{r_2}{\|\mathbf{h}_2\|^2}\}$, $\mathcal{E}\{\frac{r_1}{\|\mathbf{h}_1\|^2}\}$, and $\mathcal{E}\{\frac{1+r_2}{1+r_2(1-u)}\}$.

$$\begin{aligned}\mathcal{E}\{\frac{r_2}{\|\mathbf{h}_2\|^2}\} &= \frac{r_2}{\sigma_2^2}\int_0^\infty \frac{1}{x}\frac{e^{-1/2x}x^{N-1}}{2^N(N-1)!}dx \\ &= \frac{r_2}{2\sigma_2^2(N-1)}.\end{aligned} \tag{49}$$

Similarly, we have

$$\mathcal{E}\{\frac{r_1}{\|\mathbf{h}_1\|^2}\} = \frac{r_1}{2\sigma_1^2(N-1)}. \tag{50}$$

According to Lemma 3, the pdf of u is a Beta function, we have

$$\begin{aligned}&\mathcal{E}\{\frac{1+r_2}{1+r_2(1-u)}\} \\ &= (1+r_2)\int_0^1 \frac{(N-1)(1-u)^{N-2}}{1+r_2(1-u)}du \\ &= (1+r_2)(N-1)\int_0^1 \frac{x^{N-2}}{1+r_2x}dx = A.\end{aligned} \tag{51}$$

Note that the random variables $\|\mathbf{h}_1\|^2$, $\|\mathbf{h}_2\|^2$, and u are independent. Hence, we have

$$\mathcal{E}\{P_{1,2}^{\mathrm{DPC}}\} = \frac{r_2}{2\sigma_2^2(N-1)} + \frac{r_1}{2\sigma_1^2(N-1)}A, \tag{52}$$

and the proof is completed.

Appendix E
Proof of Theorem 3

We first show that

$$\begin{aligned}E\{\frac{1}{\sin^2\theta}\} &= \int_0^1 \frac{1}{1-u}(N-1)(1-u)^{N-2}du \\ &= (N-1)\int_0^1 x^{N-3}dx = \frac{N-1}{N-2}.\end{aligned} \tag{53}$$

Hence, by combining it with (49) and (50), we obtain

$$\begin{aligned}\mathcal{E}\{P_{1,2}^{\mathrm{ZF}}\} &= \frac{N-1}{N-2}(\frac{r_2}{2\sigma_2^2(N-1)} + \frac{r_1}{2\sigma_1^2(N-1)}) \\ &= \frac{r_2}{2\sigma_2^2(N-2)} + \frac{r_1}{2\sigma_1^2(N-2)},\end{aligned} \tag{54}$$

and Theorem 3 is proved.

Appendix F
Proof of Theorem 4

According to (29), we have

$$\begin{aligned}&\mathcal{E}\{P_{1,2}^{\mathrm{H-NOMA}}\} = \int_0^\infty f_\xi(x) \\ &(\int_{Q^{-1}(qx)}^1 (\frac{r_2}{2\sigma_2^2(N-1)} + \frac{r_1}{2\sigma_1^2(N-1)}\frac{1+r_2}{1+r_2(1-u)})f_u(u)du \\ &+ \int_0^{Q^{-1}(qx)} (\frac{r_2}{2\sigma_2^2(N-1)} + \frac{r_1}{2\sigma_1^2(N-1)})\frac{1}{1-u}f_u(u)du)dx \\ &= \frac{r_2}{2\sigma_2^2(N-1)}\int_0^\infty F_2(x)f_\xi(x)dx \\ &+ \frac{r_1}{2\sigma_1^2(N-1)}\int_0^\infty F_1(x)f_\xi(x)dx,\end{aligned} \tag{55}$$

where

$$\begin{aligned}F_2(x) &= \int_{Q^{-1}(qx)}^1 f_U(u)du + \int_0^{Q^{-1}(qx)} \frac{1}{1-u}f_u(u)du, \\ F_1(x) &= \int_{Q^{-1}(qx)}^1 \frac{1+r_2}{1+r_2(1-u)}f_u(u)du \\ &+ \int_0^{Q^{-1}(qx)} \frac{1}{1-u}f_u(u)du.\end{aligned} \tag{56}$$

By substituting $f_u(u)$ by (34) to (56), we obtain the results in (4) and the proof is completed.

Appendix G
Proof of Theorem 5

Denote $z = Q^{-1}(qx)$, since $q^{-1} \to 0$, we have $z = (1 + r_1)/(qx) \to 0$. As $z \to 0$, we have the following facts. First,

$$\begin{aligned}F_2(x) &= (1-z)^{N-1} + \frac{N-1}{N-2}(1-(1-z)^{N-2}) \\ &= 1 + \frac{N-1}{2}z^2 + o(z^2).\end{aligned} \tag{57}$$

Second, by employing Taylor expansion, we obtain

$$\begin{aligned}B &= 1 + \sum_{k=N-1}^\infty (-1)^{k+1}\frac{r_2^k}{k}, \\ C(x) &= 1 + \sum_{k=N-1}^\infty (-1)^{k+1}\frac{(r_2(1-z))^k}{k}.\end{aligned} \tag{58}$$

Hence,

$$\begin{aligned}C(x) - B &= \sum_{k=N-1}^\infty (-r_2)^k(z - \frac{k-1}{2}z^2 + o(z^2)) \\ &= z\sum_{k=N-1}^\infty (-r_2)^k - z^2\sum_{k=N-1}^\infty (-r_2)^k\frac{k-1}{2} + o(z^2) \\ &= z(\frac{(-r_2)^{N-1}}{1+r_2}) \\ &- z^2\frac{(-r_2)^{N-1}}{1+r_2}(\frac{N}{2} - 1 - \frac{r_2}{2+2r_2}) + o(z^2).\end{aligned} \tag{59}$$

Finally, by using the equation in (59), we have

$$\begin{aligned}F_1(x) - A &= (-1)^{N-2}(1+r_2)(N-1)r_2^{1-N}(C(x)-B) \\ &+ \frac{N-1}{N-2}((N-2)z - \binom{N-2}{2}z^2) + o(z^2) \\ &= \frac{N-1}{2(1+r_2)}z^2 + o(z^2).\end{aligned} \tag{60}$$

Therefore, by applying Theorem 2 and 4, we have

$$\begin{aligned}&\mathcal{E}\{P_{1,2}^{\mathrm{H-NOMA}}\} - \mathcal{E}\{P_{1,2}^{\mathrm{DPC}}\} \\ &= \frac{r_2}{2\sigma_2^2(N-1)}\int_0^\infty (F_2(x)-1)f(x)dx \\ &+ \frac{r_1}{2\sigma_1^2(N-1)}\int_0^\infty (F_1(x)-A)f(x)dx \\ &= (\frac{r_2}{4\sigma_2^2} + \frac{r_1}{4\sigma_1^2(1+r_2)})(1+r_1)^2q^{-2}\int_0^\infty x^{-2}f(x)dx + o(q^{-2}) \\ &= (\frac{r_2}{4\sigma_2^2} + \frac{r_1}{4\sigma_1^2(1+r_2)})(1+r_1)^2\frac{(N+1)N}{(N-1)(N-2)}q^{-2} + o(q^{-2}) \\ &= \delta q^{-2} + o(q^{-2}),\end{aligned} \tag{61}$$

where $f(x) = f_\xi(x)$ is the pdf of $\frac{\|\mathbf{g}_1\|^2}{\|\mathbf{g}_2\|^2}$. The theorem is proved.

$$f_t(t)=\int_0^t f_{\frac{1}{\|\mathbf{g}_2\|^2}}(t_2y)f_{\frac{1}{\|\mathbf{g}_1\|^2}}(t_1(t-y))dy=\frac{1}{(2^N(N-1)!)^2}\int_0^t e^{-\frac{1}{2yt_2}}(t_2y)^{-N-1}e^{-\frac{1}{2t_1(t-y)}}(t_1(t-y))^{-N-1}dy. \tag{62}$$

$$\begin{aligned}\lim_{\epsilon\to0}Pr\{t>\frac{1}{\epsilon}\}&=\lim_{\epsilon\to0}\int_{\frac{1}{\epsilon}}^{\infty}f_t(t)dt=\lim_{\epsilon\to0}\frac{1}{(2^N(N-1)!)^2}\int_{\frac{1}{\epsilon}}^{\infty}\int_0^t e^{-\frac{1}{2t_2y}}(t_2y)^{-N-1}e^{-\frac{1}{2t_1(t-y)}}(t_1(t-y))^{-N-1}dydt\\&=\lim_{\epsilon\to0}\frac{1}{(2^N(N-1)!)^2}\{\int_{\frac{1}{\epsilon}}^{\infty}e^{-\frac{1}{2t_2y}}(t_2y)^{-N-1}(t_1)^{-1}\int_0^{\infty}e^{-\frac{1}{2}u}u^{N-1}dudy+\int_0^{\frac{1}{\epsilon}}e^{-\frac{1}{2t_2y}}(t_2y)^{-N-1}(t_1)^{-1}\int_0^{t_1^{-1}\epsilon}e^{-\frac{1}{2}u}u^{N-1}dudy\}\\&=\lim_{\epsilon\to0}\frac{1}{2^N(N-1)!}(t_1)^{-1}\{\int_{\frac{1}{\epsilon}}^{\infty}e^{-\frac{1}{2t_2y}}(t_2y)^{-N-1}dy+\int_0^{\frac{1}{\epsilon}}e^{-\frac{1}{2t_2y}}(t_2y)^{-N-1}(1-e^{-\epsilon/(2t_1)}\sum_{r=0}^{N-1}\frac{(\epsilon/(2t_1))^r}{r!})dy\}\\&=\lim_{\epsilon\to0}t_1^{-1}t_2^{-1}(1-e^{-\epsilon/(2t_1)}e^{-\epsilon/(2t_2)}\sum_{r=0}^{N-1}\frac{(\epsilon/(2t_2))^r}{r!}\sum_{r=0}^{N-1}\frac{(\epsilon/(2t_1))^r}{r!})=(\frac{1}{N!}t_1^{-1}t_2^{-1}((2t_1)^{-N}+(2t_2)^{-N}))\epsilon^N+o(\epsilon^N)=\beta\epsilon^N+o(\epsilon^N).\end{aligned} \tag{63}$$

Appendix H
Proof of Theorem 6

Recalling the power consumption for H-NOMA in (29), we have

$$P_{1,2}^{\mathrm{H-NOMA}}\geq\frac{r_2}{\|\mathbf{h}_2\|^2}+\frac{r_1}{\|\mathbf{h}_1\|^2},$$
$$P_{1,2}^{\mathrm{H-NOMA}}\leq\left\{\begin{array}{ll}\frac{r_2}{\|\mathbf{h}_2\|^2}+\frac{r_1(1+r_2)}{\|\mathbf{h}_1\|^2}, & u>u_T\\ (\frac{r_2}{\|\mathbf{h}_2\|^2}+\frac{r_1}{\|\mathbf{h}_1\|^2})\frac{1}{1-u}, & u\leq u_T\end{array}\right.. \tag{64}$$

Define $\delta_2=\frac{r_2}{1-u_T}$ and $\delta_1=\max(r_1(1+r_2),\frac{r_1}{1-u_T})$, we can rewrite the upper bound as

$$P_{1,2}^{\mathrm{H-NOMA}}\leq\frac{\delta_2}{\|\mathbf{h}_2\|^2}+\frac{\delta_1}{\|\mathbf{h}_1\|^2}. \tag{65}$$

Similarly, by recalling the power consumption for DPC in (5), we have

$$\frac{r_2}{\|\mathbf{h}_2\|^2}+\frac{r_1}{\|\mathbf{h}_1\|^2}\leq P_{1,2}^{\mathrm{DPC}}\leq\frac{r_2}{\|\mathbf{h}_2\|^2}+\frac{r_1(1+r_2)}{\|\mathbf{h}_1\|^2}. \tag{66}$$

Denote $\mathbf{g}_i=\mathbf{h}_i/\sigma_i, i=1,2$, we can bound $P_{1,2}^{\mathrm{H-NOMA}}$ and $P_{1,2}^{\mathrm{DPC}}$ as follows:

$$\begin{aligned}\frac{r_2\sigma_2^2}{\|\mathbf{g}_2\|^2}+\frac{r_1\sigma_1^2}{\|\mathbf{g}_1\|^2}\leq P_{1,2}^{\mathrm{DPC}}\leq\frac{r_2\sigma_2^2}{\|\mathbf{g}_2\|^2}+\frac{r_1(1+r_2)\sigma_1^2}{\|\mathbf{g}_1\|^2}.\\ \frac{r_2\sigma_2^2}{\|\mathbf{g}_2\|^2}+\frac{r_1\sigma_1^2}{\|\mathbf{g}_1\|^2}\leq P_{1,2}^{\mathrm{H-NOMA}}\leq\frac{\delta_2\sigma_2^2}{\|\mathbf{g}_2\|^2}+\frac{\delta_1\sigma_1^2}{\|\mathbf{g}_1\|^2}.\end{aligned} \tag{67}$$

Let t_1,t_2 be two positive constant, and denote $t=\frac{1}{t_2\|\mathbf{g}_2\|^2}+\frac{1}{t_1\|\mathbf{g}_1\|^2}$. We first show that

$$\lim_{\epsilon\to0}Pr\{t>\frac{1}{\epsilon}\}=\beta\epsilon^N+o(\epsilon^N). \tag{68}$$

Since $\|\mathbf{g}_1\|^2,\|\mathbf{g}_2\|^2$ are two independent Chi-square variables with degrees of freedom $2N$, the pdf of $\frac{1}{\|\mathbf{g}_i\|^2}, i=1,2$, can be expressed as

$$f_{\frac{1}{\|\mathbf{g}_i\|^2}}(y)=\frac{1}{2^N(N-1)!}e^{-\frac{1}{2y}}y^{-N-1}. \tag{69}$$

Therefore, the pdf of t is given in (62).

Therefore, we can calculate the probability as written in (63). Combining (67) and (68), we obtain

$$-\lim_{P\to\infty}\frac{\ln Pr_{out}^{\mathrm{H-NOMA}}(P)}{\ln P}=-\lim_{P\to\infty}\frac{\ln Pr_{out}^{\mathrm{DPC}}(P)}{\ln P}=N. \tag{70}$$

With regard to ZFBF precoding, we have

$$\begin{aligned}P_{1,2}^{\mathrm{ZF}}&=(\frac{r_2}{\|\mathbf{h}_2\|^2}+\frac{r_1}{\|\mathbf{h}_1\|^2})\frac{1}{1-u}\\&=\frac{r_2\sigma_2^2}{\mathbf{g}_2^H(\mathbf{I}-\mathbf{\Pi}_{\mathbf{g}_1})\mathbf{g}_2}+\frac{r_1\sigma_1^2}{\mathbf{g}_1^H(\mathbf{I}-\mathbf{\Pi}_{\mathbf{g}_2})\mathbf{g}_1}=\frac{r_2}{\|\mathbf{k}_2\|^2}+\frac{r_1}{\|\mathbf{k}_1\|^2},\end{aligned} \tag{71}$$

where $\mathbf{k}_2=(\mathbf{I}-\mathbf{\Pi}_{\mathbf{g}_1})\mathbf{g}_2$ and $\mathbf{k}_1=(\mathbf{I}-\mathbf{\Pi}_{\mathbf{g}_2})\mathbf{g}_1$. It is clear that $\|\mathbf{k}_1\|^2$ and $\|\mathbf{k}_2\|^2$ are two independent Chi-square random variables with degrees of freedom $2(N-1)$. According to (68), we have

$$\lim_{\epsilon\to0}Pr\{P_{1,2}^{\mathrm{ZF}}>\frac{1}{\epsilon}\}=\beta\epsilon^{N-1}+o(\epsilon^{N-1}). \tag{72}$$

Consequently, we obtain

$$-\lim_{P\to\infty}\frac{\ln Pr_{out}^{\mathrm{ZF}}(P)}{\ln P}=N-1, \tag{73}$$

and the proof is completed.

References

[1] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. New York: Cambridge university press, 2005.

[2] R. V. Nee and R. Prasad, *OFDM for wireless multimedia communications*. MA: Artech House, Inc., 2000.

[3] Z. Shen, J. G. Andrews, and B. L. Evans, "Optimal power allocation in multiuser OFDM systems," in *Proc. IEEE GLOBECOM*, 2003, pp. 337–341.

[4] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC*, 1995, pp. 331–335.

[5] X. Wang and X.-D. Zhang, "Linear transmission for rate optimization in MIMO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 3247–3257, 2010.

[6] Y. Wu, M. Wang, C. Xiao, Z. Ding, and X. Gao, "Linear precoding for MIMO broadcast channels with finite-alphabet constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2906–2920, 2012.

[7] Q. H. Spencer, M. Haardt *et al.*, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, 2004.

[8] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J.Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, 2006.

[9] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[10] 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 Vehicular Technology Conference (VTC Spring)*, Jun. 2013.

[11] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, “5G network capacity: key elements and technologies,” *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–78, 2014.

[12] Z. Ding, P. Fan, and H. V. Poor, “Impact of user pairing on 5G non-orthogonal multiple access,” *IEEE Trans. Veh. Technol.*, to be published.

[13] 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, 2014.

[14] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, H. V. Poor *et al.*, “Application of non-orthogonal multiple access in LTE and 5G networks,” 2015, arXiv preprint arXiv:1511.08610.

[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, September 2015.

[16] 3rd Generation Partnership Project (3GPP), *Study on downlink multiuser superposition transmission for LTE*, Mar. 2015.

[17] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: John Wiley & Sons, 2012.

[18] 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.

[19] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, “A minorization-maximization method for optimizing sum rate in non-orthogonal multiple access systems,” *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan 2016.

[20] 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.

[21] Z. Ding, R. Schober, and H. V. Poor, “A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, June 2016.

[22] B. Kim, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, “Non-orthogonal multiple access in a downlink multiuser beamforming system,” in *Proc. IEEE MILCOM 2013*, 2013, pp. 1278–1283.

[23] 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, 2015.

[24] Z. Chen, Z. Ding, P. Xu, and X. Dai, “Optimal precoding for a QoS optimization problem in 2-user MISO-NOMA downlink,” *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, June 2016.

[25] R. A. Horn and C. R. Johnson, *Matrix analysis*. New York: Cambridge university press, 2012.

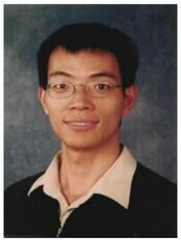
[26] S. Boyd and L. Vandenberghe, *Convex optimization*. New York: Cambridge university press, 2004.

[27] T. Anderson, *Multivariate statistical analysis*. New York: John Wiley & Sons, 1958.

[28] L. Zheng and D. N. Tse, “Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels,” *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, 2003.

[29] B. Sklar, “Rayleigh fading channels in mobile digital communication systems. I. characterization,” *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 136–146, Sep. 1997.

Zhiyong Chen received his B.Eng. degree in Electronic and Information Engineering from University of Science and Technology of China in 2012. He is now pursuing the Ph.D. degree in Communication and Information Systems at School of Information Science and Technology, University of Science and Technology of China. His research interests include wireless communication, signal processing, multiple-input multiple-output (MIMO) systems, non-orthogonal multiple access (NOMA) systems and 5G communication systems.



Zhiguo Ding (S’03-M’05) 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 Queens University Belfast, Imperial College, and Newcastle University. Since 2014, he has been with Lancaster University as a Chair Professor. His research interests are 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He received the best paper award in the IET Communication 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. 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 Wireless Communications and Mobile Computing Journal.

Xuchu Dai received the B.Eng. degree in electrical engineering from Airforce Engineering University, Xian, China, in 1984, and the M.Eng. and Ph.D. degrees in communication and information systems from the University of Science and Technology of China, Hefei, China, in 1991 and 1998, respectively. He was with the Hong Kong University of Science and Technology as a Post-Doctoral Researcher from 2000 to 2002. He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His current research interests include wireless communication systems, blind adaptive signal processing, and signal detection.

George K. Karagiannidis (M’96-S’03-F’14) was born in Pithagorion, Samos Island, Greece. He received the University Diploma (5 years) and PhD degree, both in electrical and computer engineering from the University of Patras, in 1987 and 1999, respectively. From 2000 to 2004, he was a Senior Researcher at the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In June 2004, he joined the faculty of Aristotle University of Thessaloniki, Greece where he is currently Professor in the Electrical Computer Engineering Dept. and Director of Digital Telecommunications Systems and Networks Laboratory. He is also Honorary Professor at South West Jiaotong University, Chengdu, China. His research interests are in the broad area of Digital Communications Systems with emphasis on Wireless Communications, Optical Wireless Communications, Wireless Power Transfer and Applications, Molecular Communications, Communications and Robotics and Wireless Security. He is the author or co-author of more than 400 technical papers published in scientific journals and presented at international conferences. He is also author of the Greek edition of a book on Telecommunications Systems and co-author of the book Advanced Optical Wireless Communications Systems, Cambridge Publications, 2012. Dr. Karagiannidis has been involved as General Chair, Technical Program Chair and member of Technical Program Committees in several IEEE and non-IEEE conferences. In the past he was Editor in IEEE Transactions on Communications, Senior Editor of IEEE Communications Letters, Editor of the EURASIP Journal of Wireless Communications Networks and several times Guest Editor in IEEE Selected Areas in Communications. From 2012 to 2015 he was the Editor-in Chief of IEEE Communications Letters. Dr. Karagiannidis has been selected as a 2015 Thomson Reuters Highly Cited Researcher and he Listed in Thomson Reuters 2015 Worlds Most Influential Scientific Minds.