Vector-Perturbation based NOMA Systems

Lin Bai, *Senior Member, IEEE,* Lina Zhu, *Student Member, IEEE,* Quan Yu, *Senior Member, IEEE,*
Jinho Choi, *Senior Member, IEEE,* and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—Non-orthogonal multiple access (NOMA) is one of the potential multiuser supporting techniques in the fifth generation (5G) commercial systems for its higher spectrum efficiency (SE) and cell-edge throughput, comparing to conventional orthogonal multiple access (OMA) techniques. Vector-perturbation (VP) is widely known as one of nonlinear precoding schemes that achieve near-capacity performance in practical wireless multi-input multi-output (MIMO) communication systems. In this paper, we propose a hybrid transmission strategy based on VP and NOMA (VP-NOMA) by designing proper beamforming matrix and power allocation strategy to minimize total transmit power for certain quality of service (QoS) requests. Rather than searching for optimal beamforming matrix with analytical expression, we propose a more intuitive suboptimal algorithm, named as iteration beamforming for VP-NOMA systems (IBVP-NOMA), to generate beamforming vectors with lower complexity and to limit the system performance degradation. Futher, different user clustering strategies are analyzed and compared to enhance the performance of VP-NOMA systems. Simulation results demonstrate that the proposed method requires lower transmit power than the NOMA system without VP.**

***Index Terms*—non-orthogonal multiple access (NOMA), vector-perturbation (VP), multiuser clustering, hybrid transmission.**

I. Introduction

A. Backgrounds

Non-orthogonal multiple access (NOMA), as one of the potential multiuser access techniques in the fifth generation (5G)commercial systems, has received consistent attentions from scholars around the world [1]-[5]. In conventional orthogonal multiple access (OMA), such as frequency division multiple access (FDMA) for the first generation systems, time division multiple access (TDMA) for the second generation, code division multiple access (CDMA) for the third generation, and orthogonal frequency-division multiple access (OFDMA) for the fourth generation, the transmission resources allocated to different users are orthogonal. In NOMA, however, multiuser signals are multiplexed by superposition coding in the power domain at a transmitter, and are decoded based on successive interference cancellation (SIC) at receivers [6]-[7]. Therefore, a higher spectrum efficiency (SE) of cellular systems can be achieved in NOMA in comparison with OMA, by exploiting the power domain. In addition, existed studies demonstrate that NOMA, with the benefits of higher spectrum efficiency and cell-edge throughput, can be effectively combined with the existing communication techniques, such as massive multiple-input multiple-output (MIMO), millimeter wave (mmWave) transmission, and many other widely studied 5G technologies.

Vector-perturbation (VP) is known as one of nonlinear precoding schemes that achieve near-capacity performance in practical wireless MIMO communication systems. This technique is motivated by the fact that, while system sum-rates in rich scattering wireless environment has been well studied for decades, few practical transmission schemes are proposed to actually realize that theoretical capacity. Although "dirty paper coding" (DPC) is a technique with the ability to cover the entire region of channel capacity, the high complexity makes it almost impossible for practical use. Therefore, VP is proposed as one of practical precoding methods which aim to minimize the transmit energy. More specifically, this scheme introduces an arbitrary vector to the original signal block. By employing nonlinear beamforming at the transmitter and proper decoding method to the received signal, the interference from an arbitrary vector is restrained and the transmit power can be efficiently reduced.

B. Related Works of NOMA and VP

Existing studies of NOMA mainly focus on two areas. One is performance analysis of various communication scenarios by applying NOMA with the purpose to obtain more accurate and practical theoretical results. For example, the application of NOMA in finite resolution analog beamforming (FRAB) is mainly studied in [8] to take full advantages of its features and achieve better performance. In [9], the outage performance of cooperative NOMA systems with full-duplex (FD) or half-duplex (HD) user relaying is investigated. The other area is the search for more efficient ways to optimize parameters in NOMA systems and enhance the system performance (including power allocation, user scheduling, and beamforming design techniques). More specifically, the concept of fair-NOMA system is introduced in [10], where NOMA always outperforms OMA by proper power allocation strategies. Various user selection and power allocation strategies for NOMA systems [11]-[15], multiuser NOMA beamforming techniques [16]-[18] are proposed to achieve better performance. In conclusion, both research areas have the same goal, i.e., accommodating higher communication demands of different systems (e.g., capacity/SE, outage probability, and fairness). There are also

This work was supported by the National Natural Science Foundation of China (NSFC, Grant Nos. 61231011, 61231013, and 61571025), and the National Key R&D Program of China, Grant No.2017YFB0503002.

L. Bai and L. Zhu are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China. L. Bai is also with the Beijing Laboratory for General Aviation Technology, Beihang University, Beijing 100191, China. L. Zhu is also with the Shenyuan Honors College of Beihang University, Beijing 100191, China (e-mail: {l.bai, zhulina}@buaa.edu.cn).

Quan Yu is with Institute of China Electronic System Engineering Corporation, Beijing 100191, China. (e-mail: quanyu@ieee.org).

J. Choi is with the School of Electronic Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Korea (e-mail: jchoi0114@gist.ac.kr).

W. Zhuang is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, N2L 3G1. (e-mail: wzhuang@uwaterloo.ca).

some studies concentrating on the comparison between NOMA and other widely studied techniques. For example, the superiority of NOMA in terms of system sum-rate is analyzed theoretically in comparison with conventional OMA [19]. A comparison between power domain NOMA (PD-NOMA) and sparse code multiple access (SCMA) in terms of performance and complexity is presented in [20].

The original proposition of VP is to reduce the transmit energy after channel inverse regularization, as given in [21]. Afterwards, it is found out that VP substantially reduces the gap to channel capacity, leading to its applications and performance analysis in various of communication systems [22]-[25], [28]-[29]. More specifically, Avner etal investigate VP precoding for the MIMO Gaussian broadcast channels (GBC) and obtain an analytical lower-bound on the achievable sum-rate [23]. Li and Masourvs propose a low-complexity two-stage VP scheme for adaptive modulation by applying constructive VP to simplify the conventional operation. In [25], the channel pre-inversion and VP methods are studied for large-scale broadcast channels, based on which the max-SINR vector perturbation (MSVP) scheme is introduced to enhance the performance in broadcast networks.

Moreover, based on the procedure of traditional VP, the transmitter is required to address a closest vector problem in an arbitrary lattice, which is widely known as an NP-hard problem. There are several schemes to tackle the problem, such as sphere decoding [26]-[27], which is of high complexity for a large number of users. To apply VP more efficiently, lattice reduction algorithms can be used in VP systems [28]. Multiuer MIMO (MU-MIMO) nonlinear precoding method (degree-2 VP) is presented in [29] to achieve low complexity with performance guaranteed.

C. Motivations and Contributions of This Work

Since NOMA is a promising technique for the next generation cellular communication and VP has the benefits of reducing transmit power and achieving near-capacity communication, it is a natural idea to apply VP on NOMA and to evaluate the performance. To the best of our knowledge, there is not similar study in this field. However, due to the features of nonlinear precoding in VP procedure, the existing NOMA beamforming and power allocation schemes cannot be directly extended to VP-NOMA systems. Meanwhile, the transmitters in VP-NOMA systems are burdened with NP-hard problems caused by VP, which results in inherently high complexity in comparison with the conventional NOMA. That motivates us to propose efficient beamforming design and power allocation strategy for VP-NOMA systems.

In this paper, we consider a multiple-input single-output (MISO) system consisting of one base station (BS) and multiple users. We propose a hybrid VP-NOMA transmission strategy by designing proper beamforming matrix and power allocation strategy to minimize total transmit power under certain quality of service (QoS) constraints. The contributions of this paper are summarized as follows:

1) A transmission framework of VP-NOMA is developed with power constraints;
2) A low-complexity greedy iteration algorithm for VP-NOMA systems (IBVP-NOMA) is proposed to generate beamforming vectors, without greatly reducing the system performance;
3) In order to further enhance the performance of VP-NOMA systems, different user clustering strategies are analyzed and compared. Using theoretical and numerical results, we demonstrate that the proposed method can achieve relatively lower transmit power than NOMA systems without VP (called NVP-NOMA systems).

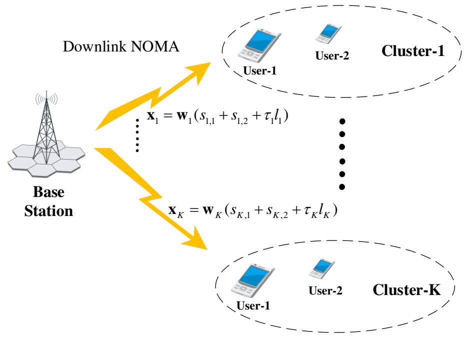
Downlink NOMA
User-1
User-2
Cluster-1
$\mathbf{x}_1 = \mathbf{w}_1(s_{1,1} + s_{1,2} + \tau_1 l_1)$
Base Station
$\mathbf{x}_K = \mathbf{w}_K(s_{K,1} + s_{K,2} + \tau_K l_K)$
User-1
User-2
Cluster-K

Fig. 1. A MU-MISO VP-NOMA downlink system.

The rest of this paper is organized as follows. Section II describes the transmission system model which incorporates both NOMA and VP methods. In Section III, an optimization problem is formulated to minimize the transmit power in VP-NOMA systems. For the sake of analysis simplification, we partially relax the constraints and transform the original problem into a series of convex sub-problems. After that, a low-complexity iteration algorithm, named as IBVP-NOMA, is proposed to solve the relaxed problems. Then, in Section IV we present user clustering strategies for VP-NOMA systems and the performance comparison between different clustering schemes. Section V provides numerical results of the VP-NOMA system performance based on the beamforming and clustering methods proposed in Section III and Section IV. This work is concluded in Section VI.

Notation: $\mathbf{A}^T$,$\mathbf{A}^H$, and $\mathbf{A}^\dagger$ represent the transpose, conjugate transpose, and the inverse of $\mathbf{A}$, respectively; diag$\{x_1, x_2, \ldots, x_n\}$ represents the diagonal matrix with the diagonal elements x_i. $\mathbb{E}\{x\}$ denotes the mean of x and $\|\cdot\|$ represents the 2-norm of a vector. Denote by $\prod_{\mathbf{A}}^{\perp}$ is the projection matrix in the orthogonal complement space of $\mathbf{A}$, and $A \setminus B$ represents exclusion of the elements of set B in set A.

II. System model

A. Vector Perturbation in NOMA Downlink systems

As illustrated in Fig. 1, we consider a single-cell multiuser MISO (MU-MISO) NOMA downlink system where the BS is equipped with M transmit antennas and there are m ($m \leq M$) single-antenna users. Each user belongs to either the strong user channel set $\mathcal{K}_1$, or the weak user channel set $\mathcal{K}_2$. For cooperative transmission, K user clusters are formed based on a clustering strategy, each of which contains a strong user (called user-1) from $\mathcal{K}_1$ and a weak user (denoted by user-2) from $\mathcal{K}_2$. The paired user in the same cluster simultaneously communicate with the BS based on the traditional NOMA principle. Consider only the situation of $M = m = 2K$. The following analysis method can be extended to a scenario of $M > m$. If $M < m$, new beamforming criteria are needed to control inter-cluster interference, which remains an open topic for further research.

At the transmitter, denote by $s_{k,i}$ ($i = 1, 2$ and $1 \leq k \leq K$) the information symbols intended for users in cluster-k in one scheduling interval. With the scheduling interval index omitted, the original signals, formed by superposition coding, are given by

$$s_k = s_{k,1} + s_{k,2}, \quad k = 1, 2, \ldots, K. \tag{1}$$

By using vector perturbation, we add an arbitrary valuable to the original signal intended for each cluster, which is given by

$$u_k = s_k + \tau_k v_k, \quad k = 1, 2, \ldots, K. \tag{2}$$

In (2), $\tau_k v_k$ is the deliberately introduced interference, where τ_k is the perturbation scale interval and v_k is a complex number whose real and imaginary parts are integers. To guarantee VP feasibility, the perturbation scale interval should be chosen properly. According to [21], τ_k is usually obtained by

$$\tau_k = 2(|C|^k_{\max} + \frac{\Delta_k}{2}) \tag{3}$$

where each cluster-k, $|C|^k_{\max}$ is the absolute value of the constellation symbol with the largest magnitude, and Δ_k is the spacing between constellation points.

In order to facilitate the following analysis and design in this study, (1) and (2) are written into vector forms. Let $\mathbf{s} = [s_1, s_2, \ldots, s_K]^T = [s_{1,1} + s_{1,2}, s_{2,1} + s_{2,2}, \ldots, s_{K,1} + s_{K,2}]^T$, and $\mathbf{u} = [u_1, u_2, \ldots, u_k]^T$ denote the signal vectors before and after VP, respectively. Let $\mathbf{v} = [v_1, v_2, \ldots, v_k]^T$ be the perturbation vector. It can be derived that

$$\mathbf{u} = \mathbf{s} + \mathbf{T}\mathbf{v} \tag{4}$$

where $\mathbf{T} = \mathrm{diag}\{\tau_1, \tau_2, \ldots, \tau_K\}$, and $\mathbf{v} \in (\mathbb{Z} + j\mathbb{Z})^K$ is a K-dimensional complex vector whose real and imaginary parts are integers. If the coordinates of the data vectors $\mathbf{s}$ are N-QAM constellation points, the set

$$\left\{\mathbf{s} + \mathbf{T}\mathbf{v} \in \mathbb{C}^K | s_k \in N - \mathrm{QAM} \quad \text{and} \quad \mathbf{v} \in (\mathbb{Z} + j\mathbb{Z})^K\right\} \tag{5}$$

is a translated lattice in $\mathbb{C}^K$ [33].

Assume that the channel state information (CSI) is perfectly known at the transmitter, based on which an $M \times K$ beamforming matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K]^T$ is derived by non-linear precoding, where $\mathbf{w}_k$ is normalized beamforming vector with unit norm. The transmitted signal vector $\mathbf{x}$ after precoding is

$$\mathbf{x} = \mathbf{W}(\mathbf{s} + \mathbf{T}\mathbf{v}) \tag{6}$$

and the transmit power is given by

$$\rho_{\mathrm{VP-NOMA}} = \mathbb{E}(\|\mathbf{W}(\mathbf{s} + \mathbf{T}\mathbf{v})\|^2). \tag{7}$$

Traditionally, for a given beamforming matrix, $\mathbf{W}$, the perturbation vector $\mathbf{v}$ is designed to minimize the transmit power, i.e.,

$$\mathbf{v} = \arg\min_{\mathbf{v}'} \|\mathbf{W}(\mathbf{s} + \mathbf{T}\mathbf{v}')\|^2. \tag{8}$$

Therefore, the properties of $\mathbf{W}$ and $\mathbf{T}$ have significant effects on the total transmit power, as the perturbation vector $\mathbf{v}$ is correlated to the design of beamforming matrix. More detailed discussion is given in Section III.

The received signal for user-i in cluster k can be written as

$$\begin{aligned} y_{k,i} =& \mathbf{h}_{k,i}^H \mathbf{w}_k (s_k + \tau_k v_k) + \\ & \sum_{j \neq k} \mathbf{h}_{k,i}^H \mathbf{w}_j (s_j + s_2^{(j)} + \tau_j v_j) + n_k, i = 1, 2. \end{aligned} \tag{9}$$

where $\mathbf{h}_{k,i}$ is the channel coefficient vector between the BS and user-i in cluster-k, and n_k is i.i.d. circularly symmetric complex Gaussian additive noise with $\mathbb{E}|n_k|^2 = \sigma_n^2$.

In order to avoid inter-cluster interference, the beamforming vector $\mathbf{w}_j$ for cluster-j should satisfy

$$\mathbf{h}_{k,i}^H \mathbf{w}_j = 0, k \neq j, i = 1, 2. \tag{10}$$

Therefore, we have

$$y_{k,i} = \mathbf{h}_{k,i}^H \mathbf{w}_k (s_{k,1} + s_{k,2} + \tau_k v_k) + n_k, i = 1, 2. \tag{11}$$

Before decoding the information signals, the offset vector $\mathbf{T}\mathbf{v}$ should be eliminated based on modulo operation [21]. To simplify the analysis below, we assume that the modulo operation removes the impact of $\mathbf{T}\mathbf{v}$ perfectly. Therefore, after successful reduction modulo the various lattices, the final superposition coded information symbols can be obtained from the received signals,

$$y_{k,i} = \mathbf{h}_{k,i}^H \mathbf{w}_k (s_{k,1} + s_{k,2}) + n_k, i = 1, 2. \tag{12}$$

Clearly, the received SINR for user-i in cluster k is given by

$$\eta_{k,1} = \frac{p_{k,1} |\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}{\sigma^2} \tag{13}$$

$$\eta_{k,2} = \frac{p_{k,2} |\mathbf{h}_{k,2}^H \mathbf{w}_k|^2}{p_{k,1} |\mathbf{h}_{k,1}^H \mathbf{w}_k|^2 + \sigma^2}. \tag{14}$$

B. NOMA Beamforming Constraints

Eq (10) actually provides possible constraints of the NOMA beamforming matrix based on zero inter-cluster interference. Following an approach similar to according to those in [30],[31],[32], denote by $\overline{\mathbf{H}}_j = [\mathbf{h}_{j,1}, \mathbf{h}_{j,2}]$ the channel vectors for cluster-j. Let $\mathbf{H}_{-k} = \{\overline{\mathbf{H}}_1, \overline{\mathbf{H}}_2, \ldots, \overline{\mathbf{H}}_{k-1}, \overline{\mathbf{H}}_{k+1}, \ldots, \overline{\mathbf{H}}_K\}$ be the set of all channel vectors except $\overline{\mathbf{H}}_k$, and $\mathcal{V}_k = Span\{\mathbf{H}_{-k}\}$ and $\mathcal{V}_k^\perp$ the space generated by $\mathbf{H}_k$ and its orthogonal complement. It is trivial to observe from (10) that the beamforming vector $\mathbf{w}_k$ for cluster-k should lie in $\mathcal{V}_k^\perp$. More specifically, suppose that $\prod_{\mathbf{H}_{-k}}^\perp$ is the projection matrix in $\mathcal{V}_k^\perp$, then $\mathbf{w}_k$ is the linear combination of $\prod_{\mathbf{H}_{-k}}^\perp \mathbf{h}_{k,1}$ and $\prod_{\mathbf{H}_k}^\perp \mathbf{h}_{k,2}$, where $\prod_{\mathbf{H}_{-k}}^\perp \mathbf{h}_{k,i}$ is the projection of $\mathbf{h}_{k,i}$ in $\mathcal{V}_k^\perp$. By using Gram-Schmidt orthogonalization, we can obtain a group of standard bases in $\mathcal{V}_k^\perp$ denoted by $\widetilde{\mathbf{h}}_{k,1}$ and $\widetilde{\mathbf{h}}_{k,2}$.

Thus, the beamforming vector $\mathbf{w}_k$ can be rewritten in the following form:

$$\mathbf{w}_k = a_{k,1}\widetilde{\mathbf{h}}_{k,1} + a_{k,2}\widetilde{\mathbf{h}}_{k,2} \tag{15}$$

where $\widetilde{\mathbf{h}}_{k,1} = \frac{\prod^{\perp}_{\mathbf{H}_{-k}} \mathbf{h}_{k,1}}{\| \prod^{\perp}_{\mathbf{H}_{-k}} \mathbf{h}_{k,1} \|}$ and $\widetilde{\mathbf{h}}_{k,2} = \frac{(\mathbf{I}-\widetilde{\mathbf{h}}_{k,1}\widetilde{\mathbf{h}}_{k,1}^H)\prod^{\perp}_{\mathbf{H}_k} \mathbf{h}_{k,2}}{\|(\mathbf{I}-\widetilde{\mathbf{h}}_{k,1}\widetilde{\mathbf{h}}_{k,1}^H)\prod^{\perp}_{\mathbf{H}_{-k}} \mathbf{h}_{k,2}\|}$. In order to normalize beamforming vectors, we have $a_{k,1}^2 + a_{k,2}^2 = 1$.

C. Constellation Size Constraint

Different from the conventional vector perturbation, here we use matrix $\mathbf{T}$ instead of a single number to adjust different power allocation strategies, so that $\mathbf{s} + \mathbf{Tl}$ are uniformly distributed in $\mathbb{C}^K$. According to (3), $\mathbf{T}$ is closely related to the constellation size.

Let $\mathbb{E}(|s_{k,i}|^2) = p_{k,i}$ denote the average per-symbol energy of an N-QAM constellation symbol for user-i in the k-th cluster. Let $p_k = p_{k,1} + p_{k,2}$ denote the total transmit power allocated to cluster-k, in order to form a translated lattice in $\mathbb{C}^K$, $p_{k,i}$ $(i = 1, 2)$ should satisfy

$$Np_{k,1} = p_{k,2} \tag{16}$$

and $p_k = p_{k,1} + p_{k,2} = (N + 1)p_{k,1}$.

III. Beamforming Strategy to Minimize Transmit Power

A. Problem Formulation

As discussed, perturbation vector $\mathbf{v}$ is chosen based on (8) to minimize the transmit power. Actually, this is a typical problem of finding the closest lattice point in an unbounded lattice, which can be easily solved by sphere decoding (here, it is referred to as sphere encoding). Therefore, the average transmit power determined by (7) is given by

$$\varepsilon_{\text{VP-NOMA}|\mathbf{W}} = \mathbb{E}(\min_{\mathbf{v}} \|\mathbf{W}(\mathbf{s} + \mathbf{Tv})\|^2). \tag{17}$$

To satisfy the service quality for all strong and weak user groups, the received SINRs $\eta_{k,1}$ and $\eta_{k,2}$ should be larger than threshold SINRs with values G_1 and G_2. The requirement ensures that the users can decode their own signals correctly. Therefore, the constraints of $\eta_{k,1}$ and $\eta_{k,2}$ should be $\eta_{k,1} \geq G_1$ and $\eta_{k,2} \geq G_2$. In terms of minimizing the transmit power, the optimization problem is formulated as

Problem 1:

$$\varepsilon_{\text{VP-NOMA}*} := \min_{\substack{(p_{k,1}, p_{k,2}, \mathbf{W}) \\ \mathbf{W} \in \mathbb{C}^{M \times K} \\ k=1,2,\ldots,K}} \varepsilon_{\text{VP-NOMA}|\mathbf{W}}$$

$$s.t. \quad |\mathbf{h}_{k,1}^H \mathbf{w}_k|^2 \geq |\mathbf{h}_{k,2}^H \mathbf{w}_k|^2 \tag{1a}$$

$$\frac{p_{k,1}|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}{\sigma_n^2} \geq G_1 \tag{1b}$$

$$\frac{p_{k,2}|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2}{p_{k,1}|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2 + \sigma_n^2} \geq G_2 \tag{1c}$$

$$\mathbf{h}_{k,i}^H \mathbf{w}_j = 0, k \neq j, i = 1, 2 \tag{1d}$$

$$k = 1, 2, \ldots, K.$$

B. Performance Bound Analysis of VP-NOMA

Before solving the optimization problem given by **Problem 1**, firstly we analyze the potential performance of VP-NOMA, i.e., the upper-bound of transmit power in VP-NOMA systems, to evaluate its practical significance for further study. Since precise solution is not necessary when considering the upper-bound of transmit power in VP-NOMA systems, the original problem is relaxed to facilitate analysis.

To begin with, we replace the object function in **Problem 1** with the lower-bound of (17) given by [33]

$$\varepsilon'_{\text{VP-NOMA}} = \frac{K\Gamma(K+1)^{\frac{1}{K}}}{\pi(K+1)} \left[(\prod_{k=1}^{K} \tau_k^2) \det(\mathbf{W}^H \mathbf{W}) \right]^{\frac{1}{K}} \tag{18}$$

where $\Gamma(\cdot)$ is the gamma function. Here, τ_k can be obtained by (3), which has constant value determined only by p_k and the constellation number N,

$$\tau_k = \sqrt{6p_k \frac{N^2}{N^2 - 1}}. \tag{19}$$

Eq (18) is referred to be the approximation (actually the lower-bound) of the actual transmit power in VP-NOMA systems. In the following, we basically concentrate on minimizing this approximation expression given by (18) rather than the accurate one in (17), as it remains a challenging (if not impossible) to find a closed-form of $\min \|\mathbf{W}(\mathbf{s} + \mathbf{Tv})\|^2$. For notation simplicity, (17) is referred to as the transmit power in VP-NOMA systems in the following analysis. As a comparison, the transmit power computed by (17), whose perturbation vector $\mathbf{v}$ is obtained by sphere encoding, is illustrated in Section V.

Let $F_L = \frac{\Gamma(\frac{L}{2}+1)^{\frac{2}{L}}}{(L+2)\pi}$ according to [33], it has been known that $\lim_{L\to\infty} F(L) = \frac{1}{2\pi e}$. Therefore, we can derive that

$$\begin{aligned} \varepsilon'^{\infty}_{\text{VP-NOMA}} &= \lim_{K\to\infty} 2KF_{2K} \left[(\prod_{k=1}^{K} \tau_k^2) \det(\mathbf{W}^H \mathbf{W}) \right]^{\frac{1}{K}} \\ &= \frac{K}{\pi e} \left[(\prod_{k=1}^{K} \tau_k^2) \det(\mathbf{W}^H \mathbf{W}) \right]^{\frac{1}{K}} \\ &\leq \frac{K}{\pi e} (\prod_{k=1}^{K} \tau_k^2)^{\frac{1}{K}}. \end{aligned} \tag{20}$$

The component $(\prod_{k=1}^{K} \tau_k^2)^{\frac{1}{K}}$ does not increase unrestrictedly, which indicates that the average transmit power per antenna is almost fixed when K is large enough. Eq (20) actually provides a potential upper-bound of (18). However, it can be seen that this upper-bound is a loose one, and has limited contribution to further analysis.

To obtain a tighter performance bound of VP-NOMA, one natural idea is to find a potential solution in the feasible region determined by constraints $(1a)$-$(1d)$. Letting $\tilde{\mathbf{w}}_k = \sqrt{p_k}\mathbf{w}_k$, we rewrite the constraints in **Problem 1** into matrix form as

$$\widetilde{\mathbf{W}}^H\mathbf{H} = \mathbf{G} \tag{21}$$

where $\mathbf{G}$ is the threshold matrix given by

$$\mathbf{G} = \left(\begin{array}{ccccccc} g_{1,1} & g_{1,2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & g_{2,1} & g_{2,2} & \cdots & 0 & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & \ldots & 0 & 0 & g_{K,1} & g_{K,2} \end{array}\right). \tag{22}$$

Clearly, based on constraints $(1a)$ - $(1d)$, $\mathbf{G}$ has following properties

$$\begin{array}{rcl} |\tilde{\mathbf{w}}_k^H\mathbf{h}_{k,1}|^2 & = & |g_{k,1}|^2 \geq G_1\sigma^2(N+1) \\ |\tilde{\mathbf{w}}_k^H\mathbf{h}_{k,2}|^2 & = & |g_{k,2}|^2 \geq \frac{G_2\sigma^2(N+1)}{N-G_2} \\ |g_{k,1}|^2 & \geq & |g_{k,2}|^2 \\ & & \qquad\qquad k = 1,2,\ldots,K \end{array} \tag{23}$$

and

$$\tilde{\mathbf{w}}_j^H\mathbf{h}_{k,i} = 0, \quad k \neq j, i = 1,2. \tag{24}$$

According to (23)-(24), we can intuitively derive a feasible solution to **Problem 1** by assuming that $g_{k,1} = \sqrt{G_1\sigma^2(N+1)}$ and $g_{k,2} = \sqrt{\frac{G_2\sigma^2(N+1)}{N-G_2}}$. Thus, one possible design of the beamforming matrix is given by

$$\widetilde{\mathbf{W}}^* = (\mathbf{G}\mathbf{H}^\dagger)^H \tag{25}$$

where $\mathbf{H} = \{\overline{\mathbf{H}}_1, \overline{\mathbf{H}}_2, \ldots, \overline{\mathbf{H}}_K\}$ is the channel matrix between the BS and all users. The power allocated to cluster-k is $p_k = \|\tilde{\mathbf{w}}_k{}^*\|^2$. Therefore, the transmit power is calculated by

$$\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}} = \frac{6N^2}{N^2-1}\frac{K\Gamma(K+1)^{\frac{1}{K}}}{\pi(K+1)}\det(\widetilde{\mathbf{W}}^{*H}\widetilde{\mathbf{W}}^*)^{\frac{1}{K}}. \tag{26}$$

In (26), $\widetilde{\mathbf{W}}^*$ is not optimal.The optimal solution to the original problem corresponds to a lower transmit power, while (26) determines a tighter upper-bound of the transmit power in the VP-NOMA system.

To obtain the expectation of $\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}}$ in (26), we consider a Rayleigh fading environment, i.e., $\mathbf{h}_{k,1} \sim \mathcal{CN}(0, \sigma_1^2\mathbf{I})$ and $\mathbf{h}_{k,2} \sim \mathcal{CN}(0, \sigma_2^2\mathbf{I})$. Denote by $\mathbf{H}_{\mathrm{r}} = \mathbf{HD}$ the normalized channel matrix whose elements are standard complex Gaussian random variables. Here, $\mathbf{D} = \mathrm{diag}\{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_K\}$ is a $2K \times 2K$ diagonal matrix and $\mathbf{D}_i = \mathrm{diag}\{\frac{1}{\sqrt{\sigma_1^2}}, \frac{1}{\sqrt{\sigma_2^2}}\}$. Then, (26) can be rewritten as

$$\begin{aligned} &\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}} \\ &= \frac{6N^2}{N^2-1}\frac{K\Gamma(K+1)^{\frac{1}{K}}}{\pi(K+1)}\det(\mathbf{GD}(\mathbf{H}_{\mathrm{r}}^H\mathbf{H}_{\mathrm{r}})^\dagger(\mathbf{GD})^H)^{\frac{1}{K}}. \end{aligned} \tag{27}$$

Based on QR decomposition, assuming that $(\mathbf{GD})^H = \mathbf{QR}$, and $\widetilde{\mathbf{H}}_{\mathrm{r}} = \mathbf{H}_{\mathrm{r}}\mathbf{Q}$, (27) can be derived that

$$\begin{aligned} &\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}} \\ &= \frac{6N^2}{N^2-1}\frac{K\Gamma(K+1)^{\frac{1}{K}}}{\pi(K+1)}(\det(\mathbf{R}_1^H\mathbf{R}_1)\det(\mathbf{A}_{\mathrm{r}11}))^{\frac{1}{K}} \end{aligned} \tag{28}$$

where $\mathbf{R}_1$ and $\mathbf{A}_{\mathrm{r}11}$ are $K \times K$ sub-matrixes of $\mathbf{R}$ (=$[\mathbf{R}_1 \quad \mathbf{0}]^T$) and $\mathbf{A}_{\mathrm{r}}$ (=$(\widetilde{\mathbf{H}}_{\mathrm{r}}^H\widetilde{\mathbf{H}}_{\mathrm{r}})^\dagger$), with

$$\mathbf{A}_{\mathrm{r}} = \left(\begin{array}{cc} \mathbf{A}_{\mathrm{r}11} & \mathbf{A}_{\mathrm{r}12} \\ \mathbf{A}_{\mathrm{r}21} & \mathbf{A}_{\mathrm{r}22} \end{array}\right). \tag{29}$$

According to the properties of inverse Wishart distribution, if $\mathbf{A}_{\mathrm{r}11} = \mathbf{X}^\dagger$, then $\mathbf{X} \sim \mathcal{W}(\mathbf{I}, K)$ is a complex Wishart distributed matrix. We finally obtain

$$\begin{aligned} &\mathbb{E}(\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}}) \\ &= \frac{6N^2}{N^2-1}\det(\mathbf{R}_1^H\mathbf{R}_1)^{\frac{1}{K}}\frac{K\Gamma(K+1)^{\frac{1}{K}}}{\pi(K+1)}\prod_{l=0}^{K-1}\frac{\Gamma(K-\frac{1}{K}-l)}{\Gamma(K-l)}. \end{aligned} \tag{30}$$

When the number of user clusters, K, tends to infinity, the upper-bound of total transmit power in VP-NOMA systems satisfies $\lim_{K\to\infty}\mathbb{E}(\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}}) = \iota\frac{\alpha}{\pi}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$, where $\iota = \lim_{K\to\infty}\frac{6N^2}{N^2-1}\det(\mathbf{R}_1^H\mathbf{R}_1)^{\frac{1}{K}}$ and $\alpha = K/M$. For $K = M$, $\lim_{K\to\infty}\mathbb{E}(\varepsilon_{\mathrm{VP-NOMA}}^{'\mathrm{up}}) = \frac{\iota}{\pi}$ [33]. This result confirms that the average transmit power per antenna in VP-based systems decreases linearly with the a increase of user number.

C. Problem Reduction

As analyzed in the preceding section, the beamforming matrix given by $\widetilde{\mathbf{W}}^*$ is only one of the feasible solutions to **Problem 1**. In this subsection, we try to obtain better beamforming matrix and power allocation strategy to enhance the system performance.

It is inherently difficult to derive the optimal solution to **Problem 1**, as the object function defined by (17) is an NP-hard problem. However, it is possible to transform the original problem into a more succinct form and propose a practical approach, named as iteration beamforming for VP-NOMA (IBVP-NOMA), to search for suboptimal solutions.

1) Object Function Approximation: For the sake of analyzing convenience, we transform the expression of the transmit power given by (18) into a more intuitive form. Let $\beta_k = \mathbf{w}_k - \sum_{i=1}^{k-1}\frac{\beta_i^H\mathbf{w}_k}{\|\beta_i\|^2}\beta_i, k = 1,2,\ldots,K$ denote orthogonal bases of $\mathbf{W}$ derived from Gram-Schmidt orthogonalization, and $\beta_1 = \mathbf{w}_1$. We have

$$\det(\mathbf{W}^H\mathbf{W}) = \prod_{k=1}^{K}\|\beta_k\|^2. \tag{31}$$

Thus, (18) can be transformed to

$$\varepsilon_{\mathrm{LB-NOMA}} = \frac{K\Gamma(K+1)^{\frac{1}{K}}}{\pi(K+1)}\left[\prod_{k=1}^{K}(\tau_k^2\|\beta_k\|^2)\right]^{\frac{1}{K}}. \tag{32}$$

The object function given by (32) is much more succinct than the original one. Instead of **Problem 1**, we now mainly consider a simplified problem given as follow:

Problem 2:

$$\min_{\substack{(p_{k,1},p_{k,2},\beta_k) \\ k=1,2,\ldots,K}}\left[\prod_{k=1}^{K}(\tau_k^2\|\beta_k\|^2)\right]^{\frac{1}{K}}$$

$$s.t. \quad \beta_k = \mathbf{w}_k - \sum_{i=1}^{k-1} \frac{\beta_i^H \mathbf{w}_k}{\|\beta_i\|^2} \beta_i \qquad (2a)$$

$$\beta_1 = \mathbf{w}_1 \qquad (2b)$$

$$(1b) - (1d) \qquad (2c)$$

$$k = 1, 2, \ldots, K.$$

It is not straightforward to deal with **Problem 2** directly. However, observing that β_k is a function of $\{\beta_1, \beta_2, \ldots, \beta_{k-1}\}$ and $\mathbf{w}_k$, once the suboptimal solutions $\{\beta_1^*, \beta_2^*, \ldots, \beta_{k-1}^*\}$ are in hand, the suboptimal solution β_k^* can be obtained by solving $\min \tau_k^2 \|\beta_k\|^2$ (equivalent to $\min \prod_{i=1}^{k} (\tau_i^2 \|\beta_i\|^2)$ as $\{\beta_1^*, \beta_2^*, \ldots, \beta_{k-1}^*\}$ are already given). That is, by iteration, we can find the suboptimal beamforming vectors $\{\beta_1^*, \beta_2^*, \ldots, \beta_K^*\}$ (or $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K\}$) to **Problem 2** when $k = K$.

Furthermore, noticing that

$$\|\beta_k\|^2 = 1 - \sum_{i=1}^{k-1} \|\beta_i^H \mathbf{w}_k\|^2 = 1 - \|\mathbf{B_k}^H \mathbf{w}_k\|^2 \qquad (33)$$

where $\mathbf{B}_k = [\beta_1, \beta_2, \ldots, \beta_{k-1}]$, we substitute (15) into (33). **Problem 2** is divided into K sub-problems:

Problem 3:

$$\min_{(p_{k,1}, p_{k,2}, a_{k,1}, a_{k,2})} \tau_k^2 (1 - \|\widetilde{\mathbf{B}}_k \mathbf{a}_k\|^2)$$

$$s.t. \quad a_{k,1}^2 + a_{k,2}^2 = 1 \qquad (3a)$$

$$(2a) - (2c) \qquad (3b)$$

$$k = 2, 3, \ldots, K$$

where $\widetilde{\mathbf{B}}_k = \mathbf{B}_k^H [\widetilde{\mathbf{h}}_{k,1}, \widetilde{\mathbf{h}}_{k,2}] = \mathbf{B}_k^H \widetilde{\mathbf{H}}_k$ and $\mathbf{a}_k = [a_{k,1}, a_{k,2}]^T$. Note that, when $k = 1$, the object function of **Problem 3** is given by $\min\limits_{p_{1,1}, p_{1,2}, a_{1,1}, a_{1,2}} \tau_1^2$.

Although the problem approximation is not optimal, though, the proposed method provides an efficient way to deal with original VP-NOMA beamforming problem. The simulation results in Section V show that the solution obtained by problem approximation is close enough to optimal one.

2) Constraint of p_k: In the following, we aims to simplify the constraints given by $(1a) - (1c)$ and obtain the constraint of p_k.

Firstly, denote by $\mathbf{h}'_{k,1} = \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,1}$ and $\mathbf{h}'_{k,2} = \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,2}$ the projection of $\mathbf{h}_{k,1}$ and $\mathbf{h}_{k,2}$ in $\mathcal{V}_k^{\perp}$, where $\prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,i}$ and $\mathcal{V}_k^{\perp}$ are defined in Section II. According to (15), we have following results:

$$|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2 = |\mathbf{h}'^H_{k,1} \mathbf{w}_k|^2 = a_{k,1}^2 \|\mathbf{h}'_{k,1}\|^2 = \lambda_{k,1} a_{k,1}^2 \qquad (34)$$

$$\begin{aligned} |\mathbf{h}_{k,2}^H \mathbf{w}_k|^2 &= |\mathbf{h}'^H_{k,2} \mathbf{w}_k|^2 \\ &= \|\mathbf{h}'_{k,2}\|^2 (a_{k,1}\sqrt{\phi_k} + a_{k,2}\sqrt{1-\phi_k})^2 \qquad (35) \\ &= \lambda_{k,2} (a_{k,1}\sqrt{\phi_k} + a_{k,2}\sqrt{1-\phi_k})^2 \end{aligned}$$

where $\phi_k = \frac{|\mathbf{h}'^H_{k,2} \mathbf{h}'_{k,1}|^2}{\|\mathbf{h}'_{k,2}\|^2 \|\mathbf{h}'_{k,1}\|^2}$ is the angle parameter between $\mathbf{h}'_{k,1}$ and $\mathbf{h}'_{k,2}$.

Secondly, noticing that $Np_{k,1} = p_{k,2}$ and $p_k = p_{k,1} + p_{k,2}$, according to $(1b)$ and $(1c)$, p_k should satisfies

$$\frac{p_k}{N+1} \geq \frac{G_1 \sigma_n^2}{|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}; \qquad (36)$$

$$\frac{Np_k}{N+1} \geq G_2 \left(\frac{\sigma_n^2}{|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2} + p_{k,1} \right). \qquad (37)$$

Therefore, the constraint of p_k is given by

$$p_k \geq \max \left\{ \frac{G_1 \sigma_n^2 (N+1)}{|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}, \frac{G_2 \sigma_n^2 (N+1)}{(N - G_2)|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2} \right\}. \qquad (38)$$

Obviously, in terms of transmit power minimization, we can choose

$$p_k^* = \max \left\{ \frac{G_1 \sigma_n^2 (N+1)}{|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}, \frac{G_2 \sigma_n^2 (N+1)}{(N - G_2)|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2} \right\}. \qquad (39)$$

More specifically, let $\xi_k = \frac{|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}{|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2}$ and $\mu' = \frac{G_1(N - G_2)}{G_2}$, based on constraints $(1a) - (1c)$, it can be easily derived that

$$p_k^* = \begin{cases} \frac{G_2 \sigma^2 (N+1)}{(N - G_2)|\mathbf{h}_{k,2}^H \mathbf{w}_k|^2}, & \text{if} \quad \xi_k \geq \max\{1, \mu'\} \\ \frac{G_1 \sigma^2 (N+1)}{|\mathbf{h}_{k,1}^H \mathbf{w}_k|^2}, & \text{if} \quad 1 \leq \xi_k \leq \mu' \quad \text{and} \quad \mu' \geq 1 \end{cases}. \qquad (40)$$

Eq (40) provides the constraints of $a_{k,1}$ and $a_{k,2}$, and the optimal selection of p_k when substituting (34) and (35) into (40). For notation convenience, we define two sets $\mathcal{R}_{k,1}$ and $\mathcal{R}_{k,2}$ to represent the constraints in (40), which is written as follows:

$$\mathcal{R}_{k,1} = \{a_{k,1} | \xi_k \geq \max\{1, \mu'\}\} \qquad (41)$$

$$\mathcal{R}_{k,2} = \{a_{k,1} | 1 \leq \xi_k \leq \mu' \quad \text{and} \quad \mu' \geq 1\} \qquad (42)$$

and $a_{k,2} = \sqrt{1 - a_{k,1}^2}$. Observing that τ_k^2 is directly proportional to p_k, the form of **Problem 3** is reduced as follow:

Problem 4:

$$\min_{(a_{k,1}, a_{k,2})} p_k^* (1 - \|\widetilde{\mathbf{B}}_k^H \mathbf{a}_k\|^2)$$

$$s.t. \quad a_{k,1}^2 + a_{k,2}^2 = 1 \quad and \quad (2a) \qquad (4a)$$

$$k = 2, 3, \ldots, K.$$

The rest of constraints in $(3b)$ is contained in p_k^*, which is given by (40). Also, when $k = 1$, the objective function of **Problem 4** becomes $\min\limits_{a_{1,1}, a_{1,2}, p_1} p_1^*$. **Problem 4** is a convex optimization problem and is relatively easy to solve. The detailed deduction of $\mathcal{R}_{k,1}$ and $\mathcal{R}_{k,2}$ is shown in Appendix A.

D. Algorithm of IBVP-NOMA

As analyzed, the original optimization problem can be divided into K sub-problems given by **Problem 3**. We now propose an iteration algorithm to solve **Problem 4** based on the analysis in the previous subsection.

Step 1: Randomly initialize the strong and weak user channel sets $\mathcal{K}_1 = \{\mathbf{h}_{1,1}, \mathbf{h}_{2,1}, \dots, \mathbf{h}_{K,1}\}$, $\mathcal{K}_2 = \{\mathbf{h}_{1,2}, \mathbf{h}_{2,2}, \dots, \mathbf{h}_{K,2}\}$. Note that this algorithm does not include user clustering, i.e., we assume that $\mathbf{h}_{k,1}$ and $\mathbf{h}_{k,2}$ are the selected channels of strong and weak user pair in cluster-k. This assumption does not hold in Section IV.

To begin with, the optimal beamforming vector $\mathbf{w}_1^*$ with β_1^* is obtained by solving **Problem 4** when $k = 1$. Note that in this case the object function should be replaced by $\min\limits_{a_{1,1}, a_{1,2}, p_1} p_1^*$.

Step 2: For the k-th pair of users ($\mathbf{h}_{k,1} \in \mathcal{K}_1, \mathbf{h}_{k,2} \in \mathcal{K}_2$), assume that we have already computed the optimal beamforming vectors $\mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_{k-1}^*$ and their orthogonalization vectors $\beta_1^*, \beta_2^*, \dots, \beta_{k-1}^*$. The parameters $a_{k,1}^*$, $a_{k,2}^*$ and p_k^* for cluster-k can be determined by solving **Problem 4**, based on which the optimal $\mathbf{w}_k^*$ and β_k^* are obtained according to (15).

Step 3: Check whether $k = K$, if not, return to step 2; Otherwise, the algorithm is completed.

The detailed algorithm is illustrated in Algorithm 1.

It can be seen that the complexity of Algorithm 1 is $O(K)$. Thus, we can obtain the suboptimal beamforming matrix $\mathbf{W}^*$ as well as power allocation strategy $\{p_1^*, p_2^*, \dots, p_k^*\}$ by applying IBVP-NOMA proposed for VP-NOMA systems.

IV. User clustering Strategies Based on VP-NOMA

So far, the VP-NOMA beamforming strategy for certain user clusters has been analyzed. To further enhance the performance of VP-NOMA systems, we focus on user clustering, i.e., for a given strong user with channel $\mathbf{h}_{k,1}$ from set $\mathcal{K}_1$, how to select the paired weak user channel $\mathbf{h}_{\pi(k),2}$ from set $\mathcal{K}_2$. Here $\pi(k)$ represents the index of weak user channel selected for cluster-k.

In this section, we study different user clustering strategies in VP-NOMA systems based on IBVP-NOMA, for certain strong and weak user groups $\mathcal{K}_1$ and $\mathcal{K}_2$.

Obviously, the optimal clustering strategy can be obtained by exhaustive search, i.e., by estimating the performance of all possible combinations of strong and weak user channel pairs. The system chooses the optimal user that minimizes the transmit power. When IBVP-NOMA is applied in user clustering, the complexity of exhaustive searching is $K(K-1)^2$ (for each combination of strong and weak user pair, the number of iteration is $K-1$). Therefore, with the an increasing of K, the complexity of exhaustive searching increases dramatically, which makes the algorithm less practical.

A more efficient clustering strategy can be derived based on the IBVP-NOMA for certain user clusters, referred to as greedy IBVP-NOMA clustering (GIBC). Rather than attempting every combinations of strong and weak user pairs, the strategy only searches for the optimal paired weak user in the candidate weak user set. The detailed steps of GIBC is given as follows:

Algorithm 1 Algorithm of IBVP-NOMA

```
Input: channel vectors of all user clusters $\mathbf{H} = \{\overline{\mathbf{H}}_1, \overline{\mathbf{H}}_2, \dots, \overline{\mathbf{H}}_K\}$, the target receiving SINRs $G_1$ and $G_2$ for strong users and weak users, the noise power $\sigma^2$, the constellation number $N$
Initialization: $\mathbf{h}'_{k,1} = \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,1}$, $\mathbf{h}'_{k,2} = \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,2}$;
        $\lambda_{k,1} = \|\mathbf{h}'_{1,k}\|^2$, $\lambda_{k,2} = \|\mathbf{h}'_{2,k}\|^2$;
        $\widetilde{\mathbf{h}}_{k,1} = \frac{\prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,1}}{\| \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,1}\|}$, $\widetilde{\mathbf{h}}_{k,2} = \frac{(I - \widetilde{\mathbf{h}}_{k,1}\widetilde{\mathbf{h}}_{k,1}^H) \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,2}}{\|(I - \widetilde{\mathbf{h}}_{k,1}\widetilde{\mathbf{h}}_{k,1}^H) \prod_{\mathbf{H}_{-k}}^{\perp} \mathbf{h}_{k,2}\|}$;
        $\phi = \frac{|\mathbf{h}'^H_{k,2}\mathbf{h}'_{k,1}|^2}{\|\mathbf{h}'_{k,2}\|^2 \|\mathbf{h}'_{k,1}\|^2}$ and $k = 1$.
  if $k \leq K$ then
      $p_k \leftarrow \frac{G_2\sigma^2(N+1)}{(N-G_2)\lambda_{k,2}(a_{k,1}\sqrt{\phi_k} + a_{k,2}\sqrt{1-\phi_k})^2}$,
      if $k = 1$ then
          obtain $a_{1,1}^{(1)}$, $a_{1,2}^{(1)}$, and $p_1^{(1)}$ by $\gamma_1^* = \min\limits_{a_{1,1}, a_{1,2}, p_1} p_1$, where $a_{k,1} \in \mathcal{R}_{k,1}$
      else
          obtain $a_{k,1}^{(1)}$, $a_{k,2}^{(1)}$, and $p_k^{(1)}$ by $\gamma_1^* = \min_{(a_{k,1}, a_{k,2})} p_k(1 - \|\widetilde{\mathbf{B}}_k^H \mathbf{a}_k\|^2)$, where $a_{k,1} \in \mathcal{R}_{k,1}$
      end if
      $p_k \leftarrow \frac{G_1\sigma^2(N+1)}{\lambda_{k,1}a_{k,1}^2}$,
      if $k = 1$ then
          obtain $a_{1,1}^{(2)}$, $a_{1,2}^{(2)}$, and $p_1^{(2)}$ by $\gamma_2^* = \min\limits_{a_{1,1}, a_{1,2}, p_1} p_1$, where $a_{k,1} \in \mathcal{R}_{k,2}$
      else
          obtain $a_{k,1}^{(2)}$, $a_{k,2}^{(2)}$, and $p_k^{(2)}$ by $\gamma_2^* = \min_{(a_{k,1}, a_{k,2})} p_k(1 - \|\widetilde{\mathbf{B}}_k^H \mathbf{a}_k\|^2)$, where $a_{k,1} \in \mathcal{R}_{k,2}$
      end if
      $i \leftarrow \arg\min\{\gamma_1^*, \gamma_2^*\}$,
      $\{a_{k,1}^*, a_{k,2}^*, p_k^*\} \leftarrow \{a_{k,1}^{(i)}, a_{k,2}^{(i)}, p_k^{(i)}\}$
      obtain $\mathbf{w}_k^*$ and $\beta_k^*$ according to (15) and $(2a)$ or $(2b)$,
      $\mathbf{B}_k \leftarrow [\beta_1^*, \beta_2^*, \dots, \beta_{k-1}^*]$,
      $\widetilde{\mathbf{B}}_k \leftarrow \mathbf{B}_k^H[\widetilde{\mathbf{h}}_{k,1}, \widetilde{\mathbf{h}}_{k,2}]$
      $k \leftarrow k + 1$
  end if
  Output: beamforming vectors $\{\mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_k^*\}$,
      the average per-symbol energy for all user clusters $\{p_1^*, p_2^*, \dots, p_k^*\}$
```

Step 1: Randomly initialize the channel vectors for strong and weak user groups $\mathcal{H}_\mathrm{s}^0$ and $\mathcal{H}_\mathrm{w}^0$. Denote by $\mathcal{H}_\mathrm{w}$ the channels of weak users that have already been clustered, and $\mathcal{H}_\mathrm{w}^-$ is the set of candidate weak user channels. Therefore, $\mathcal{S} = \{\mathcal{H}_\mathrm{w}, \mathcal{H}_\mathrm{w}^-\}$ includes the whole weak user channels. For the sake of analysis simplicity, assume that the users are always paired based on the orders in set $\mathcal{H}_\mathrm{s}^0$ and $\mathcal{S}$. Initially, $\mathcal{S} = \mathcal{H}_\mathrm{w}^- = \mathcal{H}_\mathrm{w}^0$ and $\mathcal{H}_\mathrm{w} = \emptyset$.

Step 2: Let $\mathcal{H}_\mathrm{w} = \{\mathbf{h}_{\pi(1),2}, \mathbf{h}_{\pi(2),2}, \dots, \mathbf{h}_{\pi(k-1),2}\}$ be the already paired weak user set, where $\mathbf{h}_{\pi(i),2}$

is paired with the strong user channel $\mathbf{h}_{i,1}$ ($1 \leq i \leq k-1$). In order to select the corresponding weak user for the k-th strong user with channel vector $\mathbf{h}_{k,1}$, we need to evaluate the transmit power when the candidate weak user with $\mathbf{h}_{j,2}$ ($\in \mathcal{H}_{\mathrm{w}}^{-}$) is selected. Assume that the beamforming vectors $\mathbf{w}_{\pi(1)}, \mathbf{w}_{\pi(2)}, \ldots, \mathbf{w}_{\pi(k-1)}$ and power $p_1, p_2, \ldots, p_{k-1}$ allocated to the past $k-1$ clusters are already obtained by applying the proposed IBVP-NOMA algorithm. For each candidate weak user channel $\mathbf{h}_{j,2}$, we exclude $\mathbf{h}_{j,2}$ from $\mathcal{H}_{\mathrm{w}}^{-}$ and obtain $\mathcal{H}'_{\mathrm{w}} = \mathcal{H}_{\mathrm{w}}^{-} \setminus \mathbf{h}_{j,2}$. Then an ordered weak user set $\mathcal{S}$ corresponding to $\mathbf{h}_{j,2}$ is given by $\mathcal{S} = \{\mathcal{H}_{\mathrm{w}}, \mathbf{h}_{j,2}, \mathcal{H}'_{\mathrm{w}}\}$, where the l-th channel element in $\mathcal{S}$ is paired with $\mathbf{h}_{l,1}$ in the strong user set.

For given $\mathcal{H}_{\mathrm{s}}^{0}$ and $\mathcal{S}$, IBVP-NOMA is applied with the input strong and weak user channel matrixes generated by $\mathcal{H}_{\mathrm{s}}^{0}$ and $\mathcal{S}$, respectively. With Algorithm 1, we can obtain the transmit power ε_j and the corresponding beamforming vector $\mathbf{w}_j$ for each of the candidate weak user channels.

Step 3: Based on all values of evaluated transmit power ε_j, simply select the weak user channel that minimizes the total transmit power, i.e.,

$$\mathbf{h}_{\pi(k),2} = \arg \min_{\mathbf{h}_{j,2} \in \mathcal{H}_{\mathrm{w}}^{-}} \{\varepsilon_j\}. \tag{43}$$

The suboptimal beamforming vector $\mathbf{w}_{\pi(k)}$ can be easily derived according to the selected $\mathbf{h}_{\pi(k),2}$. Let $\mathcal{H}_{\mathrm{w}}^{-} := \mathcal{H}_{\mathrm{w}}^{-} \backslash \mathbf{h}_{\pi(k),2}$, $\mathcal{H}_{\mathrm{w}} := \mathcal{H}_{\mathrm{w}} \cup \mathbf{h}_{\pi(k),2}$, $k := k+1$. If $\mathcal{H}_{\mathrm{w}}^{-} \neq \emptyset$, return to step 2; Otherwise, the algorithm is completed.

Note that, for each strong user channel vector $\mathbf{h}_{k,1}$, $K-k$ rounds of IBVP-NOMA are required to traverse the whole candidate weak user group. Further, with $\mathbf{w}_{\pi(1)}, \mathbf{w}_{\pi(2)}, \ldots, \mathbf{w}_{\pi(k-1)}$ in hand, the algorithm of IBVP-NOMA needs extra $K-k$ rounds of iteration to obtain the transmit power ε_j for a candidate $\mathbf{h}_{j,2}$. Clearly, the total iteration rounds of GIBC is $\frac{K(K^2-1)}{3}$, which is slightly smaller than that of exhaustive searching. More specifically, GIBC creates a candidate clustering matrix during every repetition of Step 2, which causes extra iteration complexity. Actually, the procedure can be simplified to reduce complexity by not applying IBVP-NOMA to the whole channel matrixes, leading to simplified greedy IBVP-NOMA clustering (S-GIBC). The details are given in Algorithm 2

Clearly, for each strong user channel vector $\mathbf{h}_{k,1}$, S-GIBC still requires $K-k$ rounds of IBVP-NOMA, but no extra iteration is needed to determine the paired weak user $\mathbf{h}_{\pi(k),2}$ and the corresponding beamforming vector $\mathbf{w}_{\pi(k)}$. Therefore, the number of iteration rounds in S-GIBC is $\frac{K(K-1)}{2}$. This simplified method achieves lower clustering complexity. However, in S-GIBC, IBVP-NOMA is performed only on a subset of user group, leading to performance degradation in comparison with GIBC.

Further, we consider another widely used clustering strategy which is based on user channel correlation. It is universally acknowledged that NOMA has more superiority in high user channel correlations. Thus, we take this alternative clustering method for comparison to evaluate the performance of S-GIBC. By modifying the approach proposed in [34], when considering user channel correlation, the system selects the weak channel that has the maximum channel correlation with each strong user channel, given by:

Algorithm 2 S-GIBC algorithm

```
Input: channel vectors of strong and weak users $\mathcal{H}_{\mathrm{s}}^{0} = \{\mathbf{h}_{1,1}, \mathbf{h}_{2,1}, \ldots, \mathbf{h}_{K,1}\}$, $\mathcal{H}_{\mathrm{w}}^{0} = \{\mathbf{h}_{1,2}, \mathbf{h}_{2,2}, \ldots, \mathbf{h}_{K,2}\}$, the target receiving SINRs $G_1$ and $G_2$ for strong users and weak users, the noise power $\sigma_1^2$ and $\sigma_2^2$, the constellation number $N$
Initialization: $\mathcal{H}_{\mathrm{w}}^{-} = \mathcal{H}_{\mathrm{w}}^{0}$, $\mathcal{H}_{\mathrm{w}} = \emptyset$, $k = 1$
  if $k \leq K$ then
      while $\mathcal{H}_{\mathrm{w}}^{-} \neq \emptyset$ do
          Select $\mathbf{h}_{i,2} \in \mathcal{H}_{\mathrm{w}}^{-}$.
          $\mathbf{H}_k \leftarrow \{\mathcal{H}_{\mathrm{s}}^{0}, \mathcal{H}_{\mathrm{w}}^{0}\} \setminus \{\mathbf{h}_{i,2}, \mathbf{h}_{k,1}\}$, $\widetilde{\mathbf{B}}_k \leftarrow \mathbf{B}_k^H[\widetilde{\mathbf{h}}_{k,1}, \widetilde{\mathbf{h}}_{i,2}]$,
          calculate $\lambda_{k,1}$, $\lambda_{i,2}$ and $\phi_i$ according to Algorithm 1.
          $p_i \leftarrow \frac{G_2\sigma^2(N+1)}{(N-G_2)\lambda_{i,2}(a_{k,1}\sqrt{\phi_k}+a_{i,2}\sqrt{1-\phi_i})^2}$,
          if $k = 1$ then
              obtain $a_{1,1}^{(1)}$, $a_{i,2}^{(1)}$, and $p_i^{(1)}$ by $\gamma_1^* = \min_{a_{1,1}, a_{i,2}, p_i} p_i$, where $a_{k,1} \in \mathcal{R}_{k,1}$
          else
              obtain $a_{k,1}^{(1)}$, $a_{i,2}^{(1)}$, and $p_i^{(1)}$ by $\gamma_1^* = \min_{(a_{k,1}, a_{i,2})} p_k(1 - \|\widetilde{\mathbf{B}}_k^H \mathbf{a}_k\|^2)$, where $a_{k,1} \in \mathcal{R}_{k,1}$
          end if
          $p_i \leftarrow \frac{G_1\sigma^2(N+1)}{\lambda_{k,1}a_{k,1}^2}$,
          if $k = 1$ then
              obtain $a_{1,1}^{(2)}$, $a_{i,2}^{(2)}$, and $p_i^{(2)}$ by $\gamma_2^* = \min_{a_{1,1}, a_{i,2}, p_i} p_i$, where $a_{k,1} \in \mathcal{R}_{k,2}$
          else
              obtain $a_{k,1}^{(2)}$, $a_{i,2}^{(2)}$, and $p_i^{(2)}$ by $\gamma_2^* = \min_{(a_{k,1}, a_{i,2})} p_i(1 - \|\widetilde{\mathbf{B}}_k^H \mathbf{a}_k\|^2)$, where $a_{k,1} \in \mathcal{R}_{k,2}$
          end if
          $i \leftarrow \arg\min\{\gamma_1^*, \gamma_2^*\}$, and $\varepsilon_i = \min\{\gamma_1^*, \gamma_2^*\}$, $\{a_{k,1}^*, a_{j,2}^*, p_k^*\} \leftarrow \{a_{k,1}^{(i)}, a_{k,2}^{(i)}, p_k^{(i)}\}$
          obtain $\mathbf{w}_i$ and $\beta_i$ according to (15) and $(2a)$ or $(2b)$.
      end while
      $\mathbf{h}_{\pi(k),2} = \arg\min_{\mathbf{h}_{i,2} \in \mathcal{H}_{\mathrm{w}}^{-}} \{\varepsilon_i\}$, obtain the corresponding $\mathbf{w}_{\pi(k)}$ and $\beta_{\pi(k)}$.
      $\mathbf{B}_{k+1} \leftarrow [\beta_{\pi(1)}, \beta_{\pi(2)}, \ldots, \beta_{\pi(k)}]$,
      $\mathcal{H}_{\mathrm{w}}^{-} := \mathcal{H}_{\mathrm{w}}^{-} \setminus \mathbf{h}_{\pi(k),2}$, $\mathcal{H}_{\mathrm{w}} := \mathcal{H}_{\mathrm{w}} \cup \mathbf{h}_{\pi(k),2}$, $k \leftarrow k+1$
  end if
  Output: Weak user clustering order $\{\mathbf{h}_{\pi(1),2}, \mathbf{h}_{\pi(2),2}, \ldots, \mathbf{h}_{\pi(K),2}\}$, beamforming vectors $\{\mathbf{w}_{\pi(1)}, \mathbf{w}_{\pi(2)}, \ldots, \mathbf{w}_{\pi(K)}\}$.
```

$$\mathbf{h}_{\pi(k),2} = \arg \max_{\mathbf{h}_{i,2} \in \mathcal{H}_{\mathrm{weak}}^{-}} \phi_i \tag{44}$$

where $\phi_i = \frac{|\mathbf{h}_{i,2}^H \mathbf{h}'_{k,1}|^2}{\|\mathbf{h}'_{i,2}\|^2 \|\mathbf{h}'_{k,1}\|^2}$. We can see the number of iteration rounds is $\frac{K(K-1)}{2}$, the same as that in S-GIBC. As the processing of maximum-channel-correlation clustering (MCCC) during each iteration is simpler, it is necessary to evaluate the performance of the two methods and take a trade-off between them if possible.

V. Simulation Results

In this section, we evaluate the performance of VP-NOMA and NVP-NOMA systems in terms of minimizing the transmit power while satisfying certain QoS of users. Meanwhile, the efficiency of different user clustering strategies are simulated and compared when applied to VP-NOMA systems as well as to NVP-NOMA systems. During the simulation, the VP-NOMA system is assumed to be under independent Rayleigh fading environment, i.e., the strong user channel $\mathbf{h}_{k,1} \sim \mathcal{CN}(0, \sigma_1^2 \mathbf{I})$ with $\sigma_1^2 = 1$, and the weak user channel $\mathbf{h}_{k,2} \sim \mathcal{CN}(0, \sigma_2^2 \mathbf{I})$ with $\sigma_2^2 = 0.1$. The target SINRs of strong and weak users are supposed to be 9dB and 3dB, respectively.

A. The performance of VP-NOMA

Fig. 2 illustrates the transmit power of VP-NOMA and NVP-NOMA systems. The solid line shows the performance of NVP-NOMA systems, whose beamforming technique is presented in [32]. The solid-dashed and dashed lines show the accurate and the approximation of average transmit power in VP-NOMA systems, respectively. Here, the accurate expression is determined by (17), where the perturbation vector $\mathbf{v}$ is obtained by sphere encoding. The approximation of transmit power is defined by (18), with the beamforming matrix obtained by IBVP-NOMA in Subsection III-D. Clearly, (18) tells the lower-bound of the accurate transmit power in VP-NOMA systems. Besides, the results indicate that the system applying VP-NOMA greatly outperforms the one without VP, and the performance gap continues to grow with the increasing of antenna scales.

In Fig. 3, the upper-bound and lower-bound of transmit power in VP-NOMA systems are estimated respectively. The upper-bound, given by (26), is derived according to the non-optimal beamforming design proposed in Subsection III-B. The lower-bound of transmit power in VP-NOMA systems is actually the optimal solution of **Problem 2**, which is obtained by numerical method. It is observed that the gap between the transmit powers based on IBVP-NOMA and (26) becomes considerably small when K is large enough. Note that the beamforming design in Subsection III-B is much more simpler, it can be a great substitution of IBVP-NOMA.

B. Different User Clustering Strategies in VP-NOMA Systems

According to Section IV, there are generally four user clustering strategies that can be applied to VP-NOMA systems: exhaustive searching, GIBC based on minimum transmit power, S-GIBC, and MCCC. We compare the efficiency of different user clustering strategies in terms of average transmit power per antenna, and discuss their strengthes and weaknesses.

Fig. 4 shows the average transmission power of VP-NOMA and NVP-NOMA systems by applying different user clustering schemes. The performance of NVP-NOMA systems is estimated based on the clustering method proposed in [32]). Tt is

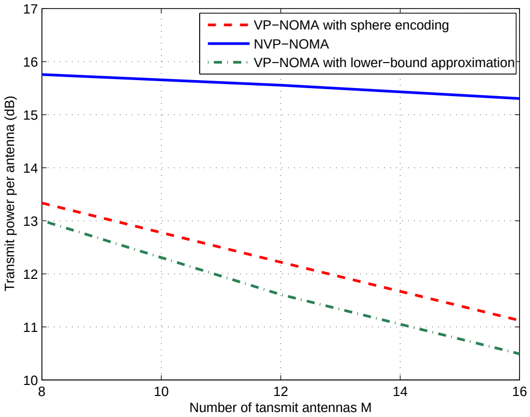

Fig. 2. The average transmit power per antenna of VP-NOMA and NVP-NOMA systems, $G_1 = 9dB$, $G_2 = 3dB$, $M = K$.

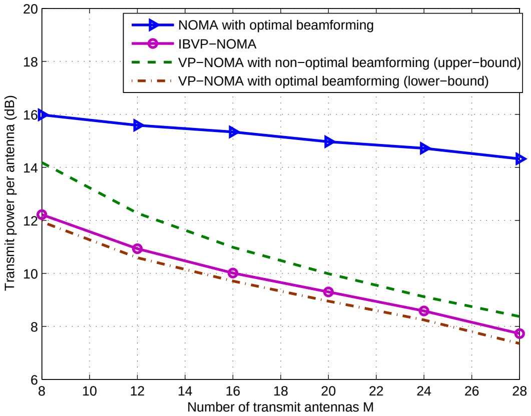

Fig. 3. The average transmit power per antenna of VP-NOMA, $G_1 = 9dB$, $G_2 = 3dB$, $M = K$.

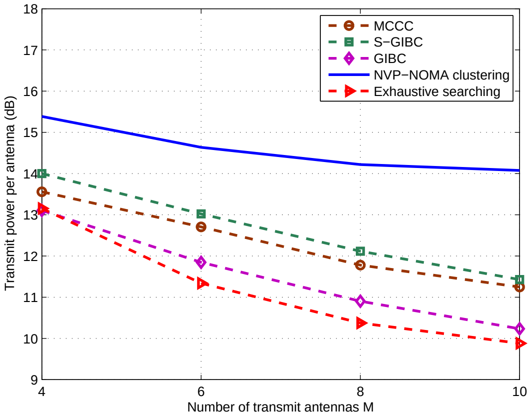

Fig. 4. The average transmit power per antenna of VP-NOMA in different user clustering methods, $G_1 = 8$, $G_2 = 2$, $M = K$.

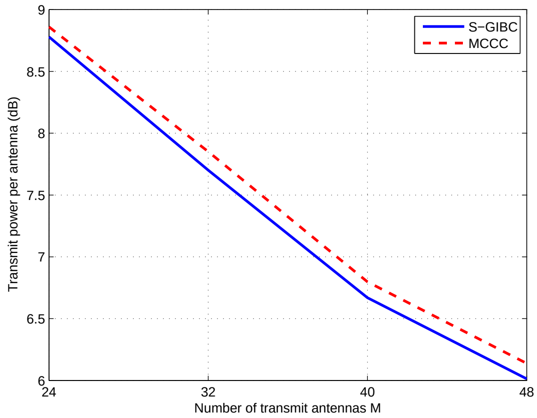

Fig. 5. The average transmit power per antenna of VP-NOMA in simplified greedy VP-NOMA clustering and maximum-channel-correlation clustering methods, $G_1 = 8$, $G_2 = 2$, $M = K$.

observed that VP-NOMA with all of the clustering methods mentioned in Section IV (exhaustive searching, GIBC, S-GIBC, and MCCC) greatly outperforms NVP-NOMA, which demonstrates the priority of VP-NOMA in comparison with NVP-NOMA. Further, the results indicate that GIBC and S-GIBC enhances the performance of VP-NOMA systems at relatively lower complexity comparing to exhaustive searching. However, S-GIBC cannot achieve optimal system performance as the exhaustive searching does, which is the cost of lower processing complexity.

In Fig. 5, we compare MCCC and S-GIBC in terms of transmit power per antenna. The results indicate that the system average transmit power per antenna with MCCC is slightly lower than that with S-GIBC for relatively small K, while the situation converges with the increasing of user number. Therefore, MCCC takes a prior place in small-scale multiuser transmission scenarios. When K becomes large, a trade-off between these two methods need to be considered to balance the performance and efficiency. For example, if the hardware capabilities is within toleration, S-GIBC is a better choice.

VI. Conclusion

In this paper, we propose a hybrid VP-NOMA transmission strategy is proposed by properly design beamforming matrix and power allocation to minimize total transmit power under QoS constraints. Firstly, we model a general MU-MISO VP-NOMA system and exploit the performance bound in comparison with NVP-NOMA systems. In terms of transmit power minimization, a low-complexity greedy iteration algorithm for VP-NOMA systems, named as IBVP-NOMA, is proposed to generate beamforming vectors, without significant system performance degradation. We also propose several user clustering strategies based on IBVP-NOMA to further enhance the system performance. The system numerical results under hybrid VP-NOMA beamforming method with different user clustering schemes, such as GIBC, S-GIBC, and MCCC are presented and compared in this paper, which demonstrate that VP-NOMA systems can achieve relatively lower transmit power than NVP-NOMA systems. Further, a trade-off between S-GIBC and MCCC need to be considered to balance the system performance and efficiency.

Appendix A
The deduction of $\mathcal{R}_{k,1}$ and $\mathcal{R}_{k,2}$

Firstly, we derive region $\mathcal{R}_{k,1}$ defined in (41), and cluster index k is omitted in the following analysis for notation simplicity. Substituting (34) and (35) into (41), ξ can be rewritten by

$$\xi = \frac{|\mathbf{h}_1^H \mathbf{w}|^2}{|\mathbf{h}_2^H \mathbf{w}|^2} = \frac{\lambda_1 a_1^2}{\lambda_2 (a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)})^2}. \tag{45}$$

Letting $\mu = \max\{1, \frac{G_1(N-G_2)}{G_2}\}$, and $\zeta = \frac{\mu\lambda_2}{\lambda_1}$, the constraint in Eq (41) can be written as

$$|a_1| \geq A_l \tag{46}$$

where $A_l = \sqrt{\zeta}|a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)}|$. Region $\mathcal{R}_1$ defined by (46) is derived according to the following two cases.

1) $0 \leq a_1 \leq 1$:
In this case, a_1 is subject to $a_1 \geq \sqrt{\zeta}(a_1\sqrt{\theta} + \sqrt{(1-a_1^2)(1-\phi)})$. We have following results

$$\begin{cases} a_1 \geq \sqrt{\frac{c_1}{c_1+1}}, & \text{if} \quad \phi < \frac{1}{\zeta} \\ a_1 = 1, & \text{if} \quad \phi = \frac{1}{\zeta} \\ a_1 \in \emptyset, & \text{if} \quad \phi > \frac{1}{\zeta} \end{cases} \tag{47}$$

where $c_1 = \frac{\zeta(1-\phi)}{(1-\sqrt{\zeta\phi})^2}$;

2) $-1 \leq a_1 < 0$:
In this case, we have $-a_1 \geq \sqrt{\zeta}|a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)}|$. The region of a_1 is given by

$$\begin{cases} a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1'}{c_1'+1}}\}, \text{if} \quad \phi \leq \frac{1}{\zeta} \\ \min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1}{c_1+1}}\} \leq a_1 \\ \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1'}{c_1'+1}}\}, \text{if} \quad \phi > \frac{1}{\zeta} \end{cases} \tag{48}$$

where $c_1' = \frac{\zeta(1-\phi)}{(1+\sqrt{\zeta\phi})^2}$. Note that for $c_1 \geq c_1'$, the region given in (48) is not empty.

Region $\mathcal{R}_1$ is obtained by combining the two cases together:

- $\sqrt{\frac{c_1}{c_1+1}} \leq a_1 \leq 1$ or
  $-1 \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1'}{c_1'+1}}\}$
  if $\phi < \frac{1}{\zeta}$;
- $a_1 = 1$ or
  $-1 \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1'}{c_1'+1}}\}$
  if $\phi = \frac{1}{\zeta}$;
- $\min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1}{c_1+1}}\} \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_1'}{c_1'+1}}\}$
  if $\phi > \frac{1}{\zeta}$.

Similarly, letting $\mu' = \frac{G_1(N-G_2)}{G_2}$, and $\zeta' = \frac{\mu'\lambda_2}{\lambda_1}$, the constraint in Eq (42) is given by:

$$B_l \leq |a_1| \leq B_r \tag{49}$$

where $B_l = |a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)}|$, and $B_r = \sqrt{\zeta'}|a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)}|$. For the right side of (46), the region of a_1 can be derived by discussing the following two cases.

1) $0 \leq a_1 \leq 1$:
Here, a_1 is subject to $a_1 \leq \sqrt{\zeta'}(a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)})$. We have following results

$$\begin{cases} 0 \leq a_1 \leq \sqrt{\frac{c_2}{c_2+1}}, & \text{if} \quad \phi < \frac{1}{\zeta'} \\ a_1 \geq 0, & \text{if} \quad \phi \geq \frac{1}{\zeta'} \end{cases} \tag{50}$$

where $c_2 = \frac{\zeta'(1-\phi)}{(1-\sqrt{\zeta'\phi})^2}$;

2) $-1 \leq a_1 < 0$:
As $-a_1 \leq \sqrt{\zeta}|a_1\sqrt{\phi} + \sqrt{(1-a_1^2)(1-\phi)}|$, the region of a_1 can be obtained by

$$\begin{cases} a_1 \geq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\}, & \text{if} \quad \phi < \frac{1}{\zeta} \\ a_1 \geq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\} & \text{or} \quad a_1 = -1, \\ & \text{if} \quad \phi = \frac{1}{\zeta} \\ a_1 \leq \min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2}{c_2+1}}\} & \text{or} \\ a_1 \geq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\}, & \text{if} \quad \phi > \frac{1}{\zeta} \end{cases} \tag{51}$$

where $c_2' = \frac{\zeta'(1-\phi)}{(1+\sqrt{\zeta'\phi})^2}$.

The analysis for the left constraint in (46) is the same as that for $\mathcal{R}_1$, which is omitted here. Thus, we derive region $\mathcal{R}_2$ as follow:

- $\sqrt{\frac{c_0}{c_0+1}} \leq a_1 \leq \sqrt{\frac{c_2}{c_2+1}}$ or $\max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\} \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_0'}{c_0'+1}}\}$, if $\phi < \frac{1}{\zeta'}$;
- $\sqrt{\frac{c_0}{c_0+1}} \leq a_1 \leq 1$ or $a_1 = -1$ or $\max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\} \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_0'}{c_0'+1}}\}$, if $\phi = \frac{1}{\zeta'}$;
- $\sqrt{\frac{c_0}{c_0+1}} \leq a_1 \leq 1$ or $-1 \leq a_1 \leq \min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2}{c_2+1}}\}$ or $\max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\} \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_0'}{c_0'+1}}\}$, if $\frac{1}{\zeta'} < \phi < \frac{1}{\zeta_0}$;
- $0 \leq a_1 \leq 1$ or $a_1 \leq \min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2}{c_2+1}}\}$ or $\max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\} \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_0'}{c_0'+1}}\}$, if $\phi = \frac{1}{\zeta_0}$;
- $\max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2'}{c_2'+1}}\} \leq a_1 \leq \max\{-\sqrt{1-\phi}, -\sqrt{\frac{c_0'}{c_0'+1}}\}$ or $\min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_0}{c_0+1}}\} \leq a_1 \leq \min\{-\sqrt{1-\phi}, -\sqrt{\frac{c_2}{c_2+1}}\}$ if $\phi > \frac{1}{\zeta_0}$.

Here, $\zeta_0 = \frac{\lambda_2}{\lambda_1}$, $c_0 = \frac{\zeta_0(1-\phi)}{(1-\sqrt{\zeta_0\phi})^2}$, and $c_0' = \frac{\zeta_0(1-\phi)}{(1+\sqrt{\zeta_0\phi})^2}$.

References

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

[2] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313-316, Feb. 2014.

[3] Z.Ding, Y.Liu, J.Choi, M.Elkashlan, C.-L.I,and H.V.Poor, "Application of non-orthogonal multiple access in LTE and 5G Networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185-191, Feb. 2017.

[4] 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, Sept. 2015.

[5] A. Benjebbour, K. Saito, Anxin Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials," in *Proc. IEEE WINCOM*, pp. 1-6, Oct. 2015.

[6] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE VTC Fall*, pp. 1-5, Sept. 2013.

[7] L. Bai and J. Choi, *Low Complexity MIMO Detection*. Springer, 2012.

[8] Z. Ding, L. Dai, R. Schober, and H. V. Poor, "NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1879-1882, Aug. 2017.

[9] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, "Outage performance of full/half-duplex user relaying in NOMA systems," in *Proc. IEEE ICC*, pp. 1-6, May 2017.

[10] J. A. Oviedo, and H. R. Sadjadpour, "A fair power allocation approach to NOMA in multiuser SISO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7974-7985, Sept. 2017.

[11] T. Seyama, T. Dateki, and H. Seki, "Efficient selection of user sets for downlink non-orthogonal multiple access," in *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, pp. 476C480, Aug 2012.

[12] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686-7698, Nov. 2016.

[13] S. N. Datta and S. Kalyanasundaram, "Optimal power allocation and user selection in non-orthogonal multiple access systems," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, April 2016.

[14] L. Bai, L. Zhu, J. Choi, and W. Zhuang, "An efficient hybrid transmission method: using nonorthogonal multiple access and multiuser diversity," *IEEE Trans. Veh. Technol.*, DOI: 10.1109/TVT.2017.2767573.

[15] G. Nain, S. S. Das and A. Chatterjee, "Low complexity user selection with optimal power allocation in downlink NOMA," *IEEE Commun. Lett.*, DOI: 10.1109/LWC.2017.2762303.

[16] G. Nain, S. S. Das and A. Chatterjee, "On generalized downlink beamforming with NOMA," *IEEE Journal of Communications and Networks*, vol. 19, no. 4, pp. 319-328, Aug. 2017.

[17] J. Choi, "Multiuser precoding with limited cooperation for large-scale MIMO multicell downlink," *IEEE Trans. Wireless Commun.*, vol.14, No.3, pp.1295-1308, Mar. 2015.

[18] 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.

[19] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191-5202, Oct. 2017.

[20] M. Moltafet, N. Mokari, M. R. Javan, and P. Azmi. "Comparison study between PD-NOMA and SCMA," *IEEE Trans. Veh. Technol.*, DOI: 10.1109/TVT.2017.2759910

[21] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537-544, Mar. 2005.

[22] R. R. Mller, D. Guo, and A. L. Moustakas, "Vector precoding for wireless MIMO systems and its replica analysis," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 530-540, Apr. 2008.

[23] Y. Avner, B. M. Zaidel, and S. Shamai (Shitz), "On Vector Perturbation Precoding for the MIMO Gaussian Broadcast Channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5999-6027, Mar. 2015.

[24] A. Li, and C. Masouros, "A two-stage vector perturbation scheme for adaptive modulation in downlink MU-MIMO," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7785-7791, Sept. 2015.

[25] D. A. Karpuk, and P. Moss, "Channel pre-inversion and max-SINR vector perturbation for large-scale broadcast channels," *IEEE Trans. Broadcasting*, vol. 63, no. 3, pp. 494-506, Sept. 2017.

[26] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1639-1642, Jul. 1999.

[27] M. O. Damen, H. E. Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2389-2402, Oct. 2003.

[28] C. Windpassinger, R. F. H. Fischer, and J. B. Huber, "Lattice-reductionaided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2057-2060, Dec. 2004.

[29] Y. Ma, A. Yamani, N. Yi, and R. Tafazolli, "Low-complexity MU-MIMO nonlinear precoding using degree-2 sparse vector perturbation," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 497-509, Mar. 2016.

[30] K. M. Ho, D. Gesbert, E. Jorswieck, and R. Mochaourab, "Beamforming on the MISO interference channel with multi-user decoding capability," in *Proc. Signals, Systems and Computers (ASILOMAR)*, pp. 1196 - 1201, Nov. 2010.

[31] J. Lindblom, E. Karipidis, and E. G. Larsson, "Efficient computation of pareto optimal beamforming vectors for the MISO interference channel with successive interference cancellation," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4782-4795, Oct. 2013.

[32] Jinho Choi, "A suboptimal approach for minimum transmit power NOMA beamforming," in *Proc. IEEE VTC Fall*, pp. 1-5, Sept. 2017.

[33] D. J. Ryan, I. B. Collings, I. V. L. Clarkson, and R. W. Heath Jr., "performance of vector perturbation multiuser MIMO systems with limited feedback," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2633-2644, Sept. 2009.

[34] 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 Military Communications Conference (MILCOM 2013)*, pp. 1278-1283, Nov 2013.