Beam Design and User Scheduling for Nonorthogonal Multiple Access With Multiple Antennas Based on Pareto Optimality

Junyeong Seo, *Student Member, IEEE*, and Youngchul Sung, *Senior Member, IEEE*

***Abstract*—In this paper, the problem of transmit beam design and user scheduling is investigated for multiuser (MU) multiple-input single-output (MISO) nonorthogonal multiple access (NOMA) downlink. First, Pareto-optimal beam design is solved for two-user MISO broadcast channels (BCs) with successive interference cancelation (SIC), and certain properties of Pareto-optimal beam vectors are obtained. Based on this result, a beam design and user scheduling method is proposed for MU-MISO NOMA. In the proposed method, simultaneously served users are grouped into multiple clusters with two users in each cluster, and a hierarchical beam structure composed of zero forcing intercluster beamforming and two-user Pareto-optimal intracluster beamforming is applied. The proposed method has the ability to control the rates of strong and weak users and hence provides great flexibility to MU-MISO NOMA operation. In addition, the problem of Pareto-optimal beam design for general M-user MISO BCs with SIC is solved.**

***Index Terms*—Non-orthogonal multiple access (NOMA), multiuser MIMO, scheduling, Pareto-optimal design, SIC.**

I. Introduction

NOMA is a promising technology for 5G wireless networks to increase the spectral efficiency[1]. Unlike conventional orthogonal multiple access (OMA) which serves multiple users based on time, frequency and/or spatial domains, NOMA exploits the power domain that results from unequal channel conditions, under which users with strong channels are basically limited by degree-of-freedom (DoF) such as bandwidth, but users with weak channels are limited by noise [2, P. 239]. In NOMA with such channel conditions, superposition coding at the base station (BS) and SIC at certain receivers are applied to support multiple users at the same time/frequency resource block. Initially, NOMA with a single antenna in both the BS and users was studied [1], [3]–[8]. Recently, there have been efforts to extend NOMA to multiple-antenna systems. Unlike conventional multiple-antenna downlink systems which can serve as many users as the number of antennas, more users can be served in multiple-antenna NOMA systems. Although the possibility that multiple-antenna NOMA can outperform conventional multiple-antenna OMA was shown in [9], [10], many important problems need to be investigated further for multiple-antenna NOMA. In multiple-antenna NOMA, typically user grouping is done first by forming clusters as many as the number of transmit antennas, and then multiple users in each cluster share the spatial dimension and are served in the power domain [11]–[13]. Since the performance of such multiple-antenna NOMA significantly depends on the channel conditions of users across clusters and within each cluster, effective user scheduling and grouping methods should be devised in order to achieve both MU diversity and the NOMA gain from unequal channel conditions. In addition, the problem of optimal beam design and power allocation compatible to user scheduling should be solved.

A. Related Works

There have been several studies on multiple-antenna NOMA for cellular downlink especially for MU-MISO downlink, which is the main focus of this paper. Especially, much work has been conducted on the case in which two users are grouped and served by NOMA with consideration of signalling overhead and SIC error propagation. In [14], beam design was considered for two-user MISO NOMA downlink to minimize the total transmission power of BS for given target rates of two users, and the idea of quasi-degradation was developed. In [11]–[13], a hierarchical beam structure for MU-MISO NOMA was considered to provide an approach to sum rate maximization for MU-MISO NOMA by combining spatial beamforming and NOMA. In the considered hierarchical structure, simultaneously-served users are grouped into multiple clusters and two users (one strong user and one weak user) are assigned to each cluster. Then, the beam design problem for MU-MISO NOMA was simplified by designing ZF beams based on strong users' channels and allocating the same beam to the weak user as the beam of the strong user in each cluster. Then, two users in each cluster sharing the same beam are served by two-user NOMA. Under this hierarchical MU-MISO NOMA model, several user scheduling and clustering methods were proposed. In [11], highly correlated users are chosen as candidates to be clustered, and two users having the largest channel gain difference are assigned to the same cluster. In [12], strong users are selected by the semi-orthogonal user selection (SUS) algorithm [15], and then weak users are selected using a matching user selection algorithm by considering inter-cluster interference (ICI). In [13], a fairness-oriented user

Manuscript received May 19, 2017; revised November 20, 2017 and February 13, 2018; accepted March 20, 2018. Date of publication April 2, 2018; date of current version April 18, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marios Kountouris. This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2013R1A1A2A10060852) and in part by "The Cross-Ministry Giga KOREA Project" grant funded by the Korean Government (MSIT) (GK17S0400, Research and Development of Open 5G Reference Model). *(Corresponding author: Youngchul Sung.)*

The authors are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea (e-mail: jyseo@kaist.ac.kr; ysung@ee.kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2018.2821635

1053-587X © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

selection algorithms was proposed by selecting two paired users based on their NOMA data rate.

There is some work for the case beyond two users in a cluster. For example, sum-rate maximization was considered for general M-user MU-MISO BCs with SIC based on minorization-maximization in [16]. Unlike [11]–[13], this work directly considered M users without a hierarchical approach, i.e., only one cluster with M users is considered. There also exists some study on multiple-input multiple-output (MIMO) NOMA. For example, in [17] the impact of user pairing is analyzed with fixed power allocation under the assumption that ICI is removed by multiple receive antennas.

Some part of this work was presented in [18].

B. Contributions

The contributions of this paper are summarized in the below:

- First, in Section III, we solved the Pareto-optimal beam design and power allocation problem for two-user MISO BCs in which superposition coding is used at the transmitter and the interference at the strong user is eliminated by SIC while the interference at the weak user is treated as noise. The result is based on the existing work for two-user MISO interference channels [19] with fixed power allocation. The power allocation part is newly solved for BCs.
- Second, in Section IV, we derived certain properties of Pareto-optimal beams and channel conditions for Pareto-optimality for two-user MISO BCs with SIC in terms of the channel gains and the angle between channel vectors of two users. This newly obtained result provides insights into good channel conditions for two-user MISO BCs with SIC. Furthermore, the derived results lays foundation for the proposed beam design, cluster power allocation and user scheduling method for MU-MISO NOMA.
- Third, in Section V, we developed an efficient user scheduling, beam design and power allocation method for MU-MISO NOMA downlink based on the hierarchical beam structure composed of ZF inter-cluster beamforming and two-user Pareto-optimal intra-cluster beamforming. The key advantage of the proposed method compared to existing methods is that the rates of strong users and weak users can be controlled arbitrarily under intra-cluster Pareto-optimality and this feature of the proposed method provides great operational flexibility to NOMA networks.
- Finally, the problem of Pareto-optimal beam design in general M-user MISO BCs with SIC is newly solved in Section V. This problem was solved by using a new induction approach to obtain the feasible rate-tuple set for a general M-user MISO BC with SIC. This result is meaningful as an independent item and can also be used for NOMA system design.

Notations: Vectors and matrices are written in boldface with matrices in capitals. All vectors are column vectors. For a matrix $\mathbf{A}$, $\mathbf{A}^*$, $\mathbf{A}^H$ and $\mathbf{A}^T$ indicate the complex conjugate, conjugate transpose, and transpose of $\mathbf{A}$, respectively, and $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}^{\perp}(\mathbf{A})$ denotes the linear space spanned by the columns of $\mathbf{A}$ and its orthogonal complement, respectively. $\mathbf{\Pi}_{\mathbf{A}}$ and $\mathbf{\Pi}_{\mathbf{A}}^{\perp}$ are the projection matrices to $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}^{\perp}(\mathbf{A})$, respectively. $||\mathbf{x}||$ represents the 2-norm of vector $\mathbf{x}$. $\mathbf{I}$ denotes the identity matrix. $\mathbf{y} \sim \mathcal{CN}(\boldsymbol{\mu}, \mathbf{\Sigma})$ means that random vector $\mathbf{y}$ is circularly-symmetric complex Gaussian distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.

II. System Model

We consider a single-cell MU-MISO NOMA downlink system with a BS equipped with N_t transmit antennas and K single-antenna users. We assume the following:[*]

A.1 (User Partition): We assume that the K users in the cell are partitioned into two user sets: $\mathcal{K}_1$ and $\mathcal{K}_2$. Here, $\mathcal{K}_1$ consists of $K/2$ strong-channel users close to the BS and $\mathcal{K}_2$ consists of $K/2$ weak-channel users far from the BS.[†]

A.2 (User Selection and Clustering): We assume that out of the K users in the cell, $2K_c$ users are selected and simultaneously served for each scheduling interval. This simultaneous service to $2K_c$ users is done by forming K_c clusters. Each cluster is composed of two users: one from the strong user set $\mathcal{K}_1$ and the other from the weak user set $\mathcal{K}_2$. Here, $K_c\,(\leq N_t)$ is the number of clusters, and details of user selection shall be presented in Section V.

A.3 (Hierarchical Beam Design for Selected Users): In order to exploit both the spatial domain and the power domain in MU-MISO-NOMA, we consider the hierarchical beam design approach in which two users in each cluster are served in the power domain with SIC while multiple clusters are served based on the spatial domain with inter-cluster beamforming. Under the assumption, the transmit signal $\mathbf{x}$ of the BS for one scheduling interval is given by

$$\mathbf{x} = \sum_{k=1}^{K_c} \mathbf{V}^{(k)} \left(\sqrt{p_1^{(k)}}\mathbf{w}_1^{(k)} s_1^{(k)} + \sqrt{p_2^{(k)}}\mathbf{w}_2^{(k)} s_2^{(k)} \right), \quad (1)$$

where $s_i^{(k)}$ is the transmit symbol for user i in cluster k from $\mathcal{CN}(0,1)$, $\mathbf{V}^{(k)}$ is the $N_t \times N_t$ inter-cluster beamforming matrix for cluster k, $\mathbf{w}_1^{(k)}$ and $\mathbf{w}_2^{(k)}$ are respectively the $N_t \times 1$ intra-cluster beamforming vectors for users 1 and 2 in cluster k out of the feasible beamforming vector set $\mathcal{W} := \{\mathbf{w} \,|\, \|\mathbf{V}^{(k)}\mathbf{w}\|^2 \leq 1\}$, and $p_i^{(k)}$ is the power assigned to user i in cluster k such that $p_1^{(k)} + p_2^{(k)} = P_k$. Here, P_k is the transmit power allocated to cluster k, satisfying the overall transmit power constraint $\sum_{k=1}^{K_c} P_k = P_T$, where P_T is the total BS transmit power. Then, the received signals $y_1^{(k)}$ and $y_2^{(k)}$ of the two users in cluster k are given by

$$y_1^{(k)} = \tilde{\mathbf{h}}_1^{(k)H}\mathbf{V}^{(k)} \left(\sqrt{p_1^{(k)}}\mathbf{w}_1^{(k)} s_1^{(k)} + \sqrt{p_2^{(k)}}\mathbf{w}_2^{(k)} s_2^{(k)} \right) + \tilde{\mathbf{h}}_1^{(k)H} \sum_{k' \neq k}^{K_c} \mathbf{V}^{(k')} \left(\sqrt{p_1^{(k')}}\mathbf{w}_1^{(k')} s_1^{(k')} + \sqrt{p_2^{(k')}}\mathbf{w}_2^{(k')} s_2^{(k')} \right) + w_1^{(k)} \quad (2)$$

$$y_2^{(k)} = \tilde{\mathbf{h}}_2^{(k)H}\mathbf{V}^{(k)} \left(\sqrt{p_1^{(k)}}\mathbf{w}_1^{(k)} s_1^{(k)} + \sqrt{p_2^{(k)}}\mathbf{w}_2^{(k)} s_2^{(k)} \right) + \tilde{\mathbf{h}}_2^{(k)H} \sum_{k' \neq k}^{K_c} \mathbf{V}^{(k')} \left(\sqrt{p_1^{(k')}}\mathbf{w}_1^{(k')} s_1^{(k')} + \sqrt{p_2^{(k')}}\mathbf{w}_2^{(k')} s_2^{(k')} \right) + w_2^{(k)} \quad (3)$$

[*]With consideration of signalling overhead and SIC error propagation, we first consider the case of two users in each cluster in Sections II, III, IV and V, like in many other works for MU-MISO NOMA, e.g., [11]–[14]. Extension to the case of more than two users in a cluster is considered in Section VI.

[†]In this paper, we will not handle user partitioning explicitly. User partitioning is important for NOMA performance and it can be done based on the channel norm information and available communication resource blocks.

where $\tilde{\mathbf{h}}_i^{(k)}$ denotes the *actual* $N_t \times 1$ (conjugated) channel vector between the BS and user i in cluster k, and

$$w_i^{(k)} \overset{i.i.d.}{\sim} \mathcal{CN}(0, (\epsilon_i^{(k)})^2) \tag{4}$$

is the zero-mean additive white Gaussian noise (AWGN) at user i in cluster k. Note that the last two terms in each of the right-hand sides (RHSs) of (2) and (3) are ICI and AWGN. In order to control ICI, we apply spatial ZF beamforming for inter-cluster beamforming. However, due to lack of spatial dimensions, we cannot eliminate ICI completely for all users. Hence, we remove ICI for the strong users with inter-cluster beamforming to keep the strong users not interference-limited. With this approach, the inter-cluster beamforming matrix $\mathbf{V}^{(k)}$ is given by

$$\mathbf{V}^{(k)} = \mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}, \tag{5}$$

where $\tilde{\mathbf{H}}_k := [\tilde{\mathbf{h}}_1^{(1)}, \tilde{\mathbf{h}}_1^{(2)}, \ldots, \tilde{\mathbf{h}}_1^{(k-1)}, \tilde{\mathbf{h}}_1^{(k+1)}, \ldots, \tilde{\mathbf{h}}_1^{(K_c)}]$ under the assumption that user 1 is the strong user and user 2 is the weak user in each cluster. With this ZF inter-cluster beamforming, we have $\tilde{\mathbf{h}}_1^{(k)H}\mathbf{V}^{(k')} = \mathbf{0}, \forall k' \neq k$, and the ICI term in (2) for the strong user disappears.

Remark 1 (Inter-cluster ZF beamforming): The reason for ZF inter-cluster beamforming is two-fold. First, in NOMA, typically two users for pairing are selected so that the strong user has large signal-to-noise ratio (SNR) $||\tilde{\mathbf{h}}_1^{(k)}||^2/(\epsilon_1^{(k)})^2$, whereas the weak user is noise-limited. If ICI leaks into the received signal of the strong user, then the high SNR of the strong user would be degraded severely down to medium or low signal-to-interference-plus-noise ratio (SINR) and consequently a significant rate loss would occur for the strong user. Second, ZF inter-cluster beamforming makes overall beam design tractable since it does not depend on the intra-cluster beam vectors $(\mathbf{w}_1^{(k)}, \mathbf{w}_2^{(k)})$. (The use of minimum mean-square-error (MMSE) beamforming for inter-cluster beamforming is compared with the ZF inter-cluster beamforming case in Section VII. Please see Fig. 8.)

Remark 2 (Design goal and overall procedure): Once the ICI is controlled and given, the model (2)–(3) is a two-user MISO BC. Thus, our approach to overall design is to first investigate the optimal beam design and power allocation for a two-user MISO BC with SIC at the strong user's receiver for a given cluster power in Section III, to derive certain properties of optimal beams relevant to selection of two users in a cluster in Section IV, and then to develop a final algorithm for overall user selection, cluster power allocation and beam design for all clusters in Section V.

III. Two-User MISO Broadcast Channel With SIC: Pareto-Optimal Design

In this section, we focus on optimal beam vector design and power allocation for a two-user MISO BC with SIC at the strong user's receiver from the perspective of *Pareto-optimality*. Here, we assume that the cluster power P_k is given with the problem of cluster power allocation postponed to Section V-B. With the cluster index (k) omitted (the index will be used if necessary), the two-user model (2)–(3) for cluster k is given by

$$y_i = \mathbf{h}_i^H(\sqrt{p_i}\mathbf{w}_i s_i + \sqrt{p_j}\mathbf{w}_j s_j) + n_i \quad i,j \in \{1,2\}, \quad j \neq i, \tag{6}$$

where $p_1 + p_2 \leq P_k$ with the total power P_k allocated to the cluster;

$$n_i \sim \mathcal{CN}(0, \sigma_i^2) \tag{7}$$

is the sum of ICI and AWGN for user i; and $\mathbf{h}_i$ is the *projected effective* channel of user i (in cluster k) given by

$$\mathbf{h}_i = \mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\tilde{\mathbf{h}}_i^{(k)}, \ i = 1,2 \tag{8}$$

from (2), (3) and (5) since $\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k} = [\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}]^H$ for orthogonal projection. The feasible set for $\mathbf{w}_i$ is given by $\mathcal{W} = \{\mathbf{w} \,|\, \|\mathbf{w}\|^2 \leq 1\}$ since the Pareto-optimal beam $\mathbf{w}_i$ under the model (6) lies in the linear space spanned by $\mathbf{h}_1 = \mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\tilde{\mathbf{h}}_1^{(k)}$ and $\mathbf{h}_2 = \mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\tilde{\mathbf{h}}_2^{(k)}$ [20], and hence $\|\mathbf{w}_i^{(k)}\| = \|\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\mathbf{w}_i^{(k)}\| = \|\mathbf{V}^{(k)}\mathbf{w}_i^{(k)}\|$ for Pareto-optimal beams.

Under the NOMA framework, we assume that user 1 is the strong user and user 2 is the weak user, i.e., $\|\mathbf{h}_1\|^2/\sigma_1^2 > \|\mathbf{h}_2\|^2/\sigma_2^2$, and that user 1 decodes the interference from user 2 and subtracts it before decoding its own data while user 2 treats the interference as noise. With this assumption, the rates of the two users are given by

$$\begin{aligned} R_1(\mathbf{w}_1, p_1) &= \log_2\left(1 + \frac{s_1(\mathbf{w}_1, p_1)}{\sigma_1^2}\right) \\ R_2(\mathbf{w}_1, \mathbf{w}_2, p_1, p_2) &\\ = \log_2\left(1 + \min\left\{\frac{r_1(\mathbf{w}_2, p_2)}{s_1(\mathbf{w}_1, p_1) + \sigma_1^2}, \frac{s_2(\mathbf{w}_2, p_2)}{r_2(\mathbf{w}_1, p_1) + \sigma_2^2}\right\}\right), & \end{aligned} \tag{9}$$

where the signal power and the interference power are respectively given by

$$s_i(\mathbf{w}_i, p_i) := p_i|\mathbf{h}_i^H\mathbf{w}_i|^2 \text{ and } r_i(\mathbf{w}_j, p_j) := p_j|\mathbf{h}_i^H\mathbf{w}_j|^2. \tag{10}$$

Note in (9) that for the rate of user 1, the interference from user 2 is not incorporated due to SIC and the rate of user 2 is bounded by not only the SINR of user 2 but also the required 'SINR' for user 1 to decode the message of user 2 for SIC before decoding its own data. Then, for the given (effective) channel vectors $(\mathbf{h}_1, \mathbf{h}_2)$ and the cluster power P_k, the achievable rate region $\mathcal{R}$ of the two-user MISO-NOMA BC is defined as the union of the rate-tuples that can be achieved by all feasible beam vectors and power allocation:

$$\mathcal{R} := \bigcup_{\substack{(\mathbf{w}_1,\mathbf{w}_2)\in\mathcal{W}^2 \\ p_1,p_2:\, p_1,p_2\geq 0, \\ p_1+p_2=P_k}} (R_1(\mathbf{w}_1, p_1), R_2(\mathbf{w}_1, \mathbf{w}_2, p_1, p_2)). \tag{11}$$

The *Pareto boundary* of the rate region $\mathcal{R}$ is the outer boundary of $\mathcal{R}$ for which the rate of any one user cannot be increased without decreasing the rate of the other user and Pareto-optimality has been used widely as a general optimal beam design criterion for MU-MISO networks with linear precoding [21]. A pair of beam vectors $(\mathbf{w}_1, \mathbf{w}_2)$ not achieving a Pareto-boundary point is not optimal since both users' rates can be increased by a better designed beam pair. Note that the sum-rate optimal point is one of the points on the Pareto-boundary where the Pareto-boundary and the minus 45°-degree line touch in the (R_1, R_2) plane, and the Pareto-optimality provides a general optimality criterion beyond the sum-rate optimality because we can change the rate operating point arbitrarily and optimally. It

is known that the Pareto-boundary can be found by maximizing R_2 for each given feasible R_1^* [21], i.e.,

$$\max_{\substack{(\mathbf{w}_1,\mathbf{w}_2)\in\mathcal{W}^2 \\ p_1,p_2:\, p_1,p_2\geq 0,\, p_1+p_2=P_k}} R_2(\mathbf{w}_1,\mathbf{w}_2,p_1,p_2) \quad \text{subject to} \quad R_1(\mathbf{w}_1,p_i)=R_1^*. \tag{12}$$

By exploiting the relationship between the rates and the SINRs in (9), the problem (12) can be rewritten in terms of SINR as

$$\max_{\substack{(\mathbf{w}_1,\mathbf{w}_2)\in\mathcal{W}^2 \\ p_1,p_2:\, p_1,p_2\geq 0, \\ p_1+p_2=P_k}} \gamma_2 := \min\left\{\frac{r_1(\mathbf{w}_2,p_2)}{s_1(\mathbf{w}_1,p_1)+\sigma_1^2}, \frac{s_2(\mathbf{w}_2,p_2)}{r_2(\mathbf{w}_1,p_1)+\sigma_2^2}\right\} \quad \text{subject to} \quad \frac{s_1(\mathbf{w}_1,p_1)}{\sigma_1^2}=\gamma_1^*, \tag{13}$$

where γ_1^* is a given feasible target SINR for user 1. An efficient solution to the problem (13) exploits an efficient parameterization of the beam vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. Note that the number of design variables in $\mathbf{w}_1$ and $\mathbf{w}_2$ is $2N_t$ complex numbers. However, one can realize that it is sufficient that both beam vectors are linear combinations of $\mathbf{h}_1$ and $\mathbf{h}_2$ (equivalently, $\Pi_{\mathbf{h}_2}\mathbf{h}_1$ and $\Pi_{\mathbf{h}_2}^{\perp}\mathbf{h}_1$). A component in the beam vector not in the span of $\mathbf{h}_1$ and $\mathbf{h}_2$ does not affect either the signal power or the interference power and hence does not affect either R_1 or R_2 [22]. Thus, it is known from [20] that the Pareto-optimal beam vectors for the problem (13) can be parameterized as [19], [20]

$$\mathbf{w}_1(\alpha_1,\beta_1)=\alpha_1\frac{\Pi_{\mathbf{h}_2}\mathbf{h}_1}{\|\Pi_{\mathbf{h}_2}\mathbf{h}_1\|}+\beta_1\frac{\Pi_{\mathbf{h}_2}^{\perp}\mathbf{h}_1}{\|\Pi_{\mathbf{h}_2}^{\perp}\mathbf{h}_1\|}, \tag{14}$$

$$\mathbf{w}_2(\alpha_2)=\alpha_2\frac{\Pi_{\mathbf{h}_1}\mathbf{h}_2}{\|\Pi_{\mathbf{h}_1}\mathbf{h}_2\|}+\sqrt{1-\alpha_2^2}\frac{\Pi_{\mathbf{h}_1}^{\perp}\mathbf{h}_2}{\|\Pi_{\mathbf{h}_1}^{\perp}\mathbf{h}_2\|}, \tag{15}$$

where

$$(\alpha_1,\beta_1)\in\mathcal{F}:=\{(\alpha,\beta)|\alpha,\beta\geq 0,\alpha^2+\beta^2\leq 1\},\ \alpha_2\in[0,1].$$

Unlike the conventional parametrization without SIC in which both users use full power [21], [23], in the parameterization (14)–(15) user 1 may not use full power whereas user 2 uses full power. This is because full power use of user 2 helps both SIC at user 1 and its own SINR at user 2, but full power use of user 1 is beneficial for its own rate but detrimental to user 2's rate since user 2 treats interference as noise. Substituting (14)–(15) into (10), we have

$$s_1(\mathbf{w}_1)=p_1\left(\alpha_1\|\Pi_{\mathbf{h}_2}\mathbf{h}_1\|+\beta_1\|\Pi_{\mathbf{h}_2}^{\perp}\mathbf{h}_1\|\right)^2 = p_1\|\mathbf{h}_1\|^2(\sqrt{\theta}\alpha_1+\sqrt{1-\theta}\beta_1)^2 \tag{16}$$

$$r_2(\mathbf{w}_1)=p_1\alpha_1^2\frac{|\mathbf{h}_2^H\mathbf{h}_1|^2}{\|\Pi_{\mathbf{h}_2}\mathbf{h}_1\|^2}=p_1\|\mathbf{h}_2\|^2\alpha_1^2 \tag{17}$$

$$s_2(\mathbf{w}_2)=p_2\|\mathbf{h}_2\|^2(\sqrt{\theta}\alpha_2+\sqrt{1-\theta}\sqrt{1-\alpha_2^2})^2 \tag{18}$$

$$r_1(\mathbf{w}_2)=p_2\alpha_2^2\frac{|\mathbf{h}_2^H\mathbf{h}_1|^2}{\|\Pi_{\mathbf{h}_1}\mathbf{h}_2\|^2}=p_2\|\mathbf{h}_1\|^2\alpha_2^2, \tag{19}$$

where the angle parameter θ between two effective channel vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ is defined as

$$\theta:=\frac{|\mathbf{h}_1^H\mathbf{h}_2|^2}{\|\mathbf{h}_1\|^2\|\mathbf{h}_2\|^2}\ \in[0,\,1]. \tag{20}$$

Substituting (16)–(19) into the problem (13) and taking square-root operation yield the equivalent problem of (21)–(22), shown at the bottom of the this page. For later use, we define the following channel quality factor λ_i and the normalized target SINR value for user 1, Γ:

$$\lambda_i:=\frac{||\mathbf{h}_i||^2}{\sigma_i^2},\ i=1,2, \qquad \Gamma:=\gamma_1^*/\lambda_1\in[0,P_k]. \tag{23}$$

Here, λ_i indicates the SNR quality of user i's channel, whereas θ in (20) is a measure of the angle between the two users' channels. Note that the actual target SINR for user 1 γ_1^* is given by $\gamma_1^*=\Gamma\lambda_1$ and the feasible range for Γ is $\Gamma\in[0,P_k]$, where the maximum P_k occurs when $\mathbf{w}_1=\mathbf{h}_1/||\mathbf{h}_1||$ and $p_1=P_k$ since $\gamma_1=s_1(\mathbf{w}_1,p_1)/\sigma_1^2=p_1|\mathbf{h}_1^H\mathbf{w}_1|^2/\sigma_1^2$.

The optimization problem (21)–(22) with fixed power allocation $p_1=p_2=1$ was solved in [19] under the framework of a two-user MISO interference channel. In the MISO interference channel case, two transmitters in the network neither cooperate nor share transmit power and hence the two transmit power values p_1 and p_2 are fixed. On the other hand, in the MISO BC case the two power values p_1 and p_2 can be designed at the BS under the constraints $p_1,p_2\geq 0$ and $p_1+p_2=P_k$ to maximize the performance.

A. Pareto-Optimal Beam Design in Two-User MISO BC With SIC With Power Allocation

In this section, we consider the problem (21)–(22) of Pareto-optimal beam design and power allocation for the two-user MISO-NOMA BC. We obtain the solution to this problem by exploiting the result in [19]. Note that once the power allocation values p_1 and p_2 are fixed, the corresponding optimal solution can be obtained from [19]. Therefore, we represent the optimal solution as a function of power allocation and then optimize the power allocation. For given $p_1\in[0,P_k]$, first, note that α_1 appears only in the constraint (22) and the denominator in the second term in the RHS of (21). Thus, optimal α_1 can be found

$$\max_{\substack{(\alpha_1,\beta_1)\in\mathcal{F} \\ \alpha_2\in[0,1] \\ 0\leq p_1\leq P_k}} \gamma_2=\min\left\{\frac{\sqrt{P_k-p_1}\|\mathbf{h}_1\|\alpha_2}{\sqrt{\sigma_1^2(1+\gamma_1^*)}}, \frac{\sqrt{P_k-p_1}\|\mathbf{h}_2\|(\sqrt{\theta}\alpha_2+\sqrt{1-\theta}\sqrt{1-\alpha_2^2})}{\sqrt{p_1\|\mathbf{h}_2\|^2\alpha_1^2+\sigma_2^2}}\right\} \tag{21}$$

$$\text{subject to} \quad \sqrt{p_1}\|\mathbf{h}_1\|(\sqrt{\theta}\alpha_1+\sqrt{1-\theta}\beta_1)=\sqrt{\gamma_1^*\sigma_1^2}. \tag{22}$$

by solving the following problem:

$$\min_{(\alpha_1,\beta_1)\in\mathcal{F}} \quad \alpha_1$$

$$\text{subject to} \quad \sqrt{p_1}\|\mathbf{h}_1\|(\sqrt{\theta}\alpha_1+\sqrt{1-\theta}\beta_1)=\sqrt{\gamma_1^*\sigma_1^2}, \quad (24)$$

and the corresponding solution is given by $\alpha_1^*(p_1)=$

$$\begin{cases}0 & \text{if} \quad \Gamma\le p_1(1-\theta),\\ \sqrt{\theta\Gamma/p_1}-\sqrt{(1-\theta)(1-\Gamma/p_1)} & \text{if} \quad \Gamma>p_1(1-\theta).\end{cases} \quad (25)$$

Then, optimal $\beta_1^*(p_1)$ can be obtained from (22) by using $\alpha_1^*(p_1)$. Furthermore, optimal solution α_2^* to (21)–(22) can be obtained as a function of p_1 (detailed derivation can be founded in [24] based on the result in [19]):

$$\alpha_2^*(p_1)=\begin{cases}1 & \text{if } p_1\in\mathcal{P}_1,\\ \frac{c(p_1)}{\sqrt{c^2(p_1)+[a(p_1)-b(p_1)]^2}} & \text{if } p_1\in\mathcal{P}_2,\\ \frac{b(p_1)}{\sqrt{b^2(p_1)+c^2(p_1)}}=\sqrt{\theta} & \text{if } p_1\in\mathcal{P}_3,\end{cases} \quad (26)$$

where

$$a(p_1):=\sqrt{P_k-p_1}\,\frac{\|\mathbf{h}_1\|}{\sqrt{\sigma_1^2(1+\gamma_1^*)}}, \quad (27)$$

$$b(p_1):=\sqrt{P_k-p_1}\,\frac{\|\mathbf{h}_2\|\sqrt{\theta}}{\sqrt{p_1\|\mathbf{h}_2\|^2(\alpha_1^*(p_1))^2+\sigma_2^2}}, \quad (28)$$

$$c(p_1):=\sqrt{P_k-p_1}\,\frac{\|\mathbf{h}_2\|\sqrt{1-\theta}}{\sqrt{p_1\|\mathbf{h}_2\|^2(\alpha_1^*(p_1))^2+\sigma_2^2}}. \quad (29)$$

and $\mathcal{P}_1:=\{p_1|a(p_1)\le b(p_1)\}$, $\mathcal{P}_2:=\{p_1|b(p_1)<a(p_1)\le b(p_1)+c^2(p_1)/b(p_1)\}$, and $\mathcal{P}_3:=\{p_1|a(p_1)>b(p_1)+c^2(p_1)/b(p_1)\}$, and the corresponding SINR for user 2 $\gamma_2^*(p_1)$ is given by $\gamma_2^*(p_1)=$

$$\begin{cases}\gamma_2^{*(1)}(p_1)=(P_k-p_1)\frac{\|\mathbf{h}_1\|^2}{\sigma_1^2(1+\gamma_1^*)} & \text{if } p_1\in\mathcal{P}_1,\\ \gamma_2^{*(2)}(p_1)=(P_k-p_1)\frac{\|\mathbf{h}_1\|^2}{\sigma_1^2(1+\gamma_1^*)}[\alpha_2^*(p_1)]^2 & \text{if } p_1\in\mathcal{P}_2,\\ \gamma_2^{*(3)}(p_1)=(P_k-p_1)\frac{\|\mathbf{h}_2\|^2}{\|\mathbf{h}_2\|^2p_1[\alpha_1^*(p_1)]^2+\sigma_2^2} & \text{if } p_1\in\mathcal{P}_3.\end{cases} \quad (30)$$

With the optimal design variables α_1,β_1 and α_2 expressed as functions of p_1, the original problem (21)–(22) finally reduces to

$$\max_{0\le p_1\le P_k}\ \gamma_2^*(p_1), \quad (31)$$

where $\gamma_2^*(p_1)$ is given by (30). Note that if we knew which $\gamma_2^{*(i)}$ in (30) to use for optimization by directly computing $a(p_1)$, $b(p_1)$, $c(p_1)$ and their relationship, it would be easy to solve the problem (31). However, the parameters $a(p_1)$, $b(p_1)$ and $c(p_1)$ to determine $\gamma_2^{*(i)}$ to use are functions of the design variable p_1. Nevertheless, this is possible and the following proposition is useful in solving (31).

Proposition 1: For given $\mathbf{h}_1$, $\mathbf{h}_2$, σ_1^2, σ_2^2, P_k and γ_1^*, we have the following regarding the optimal solution p_1^{opt} to the problem (31). If $\theta\Gamma<\tau$ or if $\theta\Gamma\ge\tau\ge0$ and $P_k\ge\Gamma+\frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}})$, then $p_1^{opt}\in\mathcal{P}_2$. Otherwise, $p_1^{opt}\in\mathcal{P}_3$. Here, $\tau:=\theta^{-1}\left(\lambda_1^{-1}+\Gamma\right)-\lambda_2^{-1}$, and Γ and λ_i are defined in (23).

Proof: See Appendix. ■

Due to Proposition 1 we know which of the three cases in (30) is applicable to the given combination of $\mathbf{h}_1$, $\mathbf{h}_2$, σ_1^2, σ_2^2, P_k, and γ_1^*. Once the set $\mathcal{P}_i$ to which p_1^{opt} belongs is determined, optimal p_1^{opt} can be found by maximizing the corresponding $\gamma_2^{*(i)}(p_1)$ in (30) with respect to p_1. A closed-form solution from solving $\frac{d\gamma_2^{*(i)}(p_1)}{dp_1}=0$ seems complicated but the solution can easily be found by a numerical method. The proposed algorithm to design Pareto-optimal beam vectors and power allocation is summarized in Algorithm 1. The Pareto-boundary of a two-user MISO-NOMA BC can be computed by sweeping $R_1^*=\log(1+\gamma_1^*)$ and computing the corresponding maximum $R_2^*=\log(1+\gamma_2^*)$. An example is shown in Fig. 1. It is seen that power allocation significantly enlarges the achievable rate region over the fixed-power beam-only design and optimal power allocation is crucial for MISO-NOMA BC.

Algorithm 1: Pareto-optimal design for 2-user MISO-BC with SIC.

function $[\sqrt{p_1}\mathbf{w}_1,\sqrt{p_2}\mathbf{w}_2]=\mathcal{D}(\mathbf{h}_1,\mathbf{h}_2,\sigma_1^2,\sigma_2^2,\gamma_1^*,P_k)$
  $\lambda_1=\|\mathbf{h}_1\|^2/\sigma_1^2$, $\lambda_2=\|\mathbf{h}_2\|^2/\sigma_2^2$,
  $\theta=\frac{|\mathbf{h}_1^H\mathbf{h}_2|^2}{\|\mathbf{h}_1\|^2\|\mathbf{h}_2\|^2}$, $\Gamma=\gamma_1^*/\lambda_1$, and
  $\tau=\theta^{-1}\left(\lambda_1^{-1}+\Gamma\right)-\lambda_2^{-1}$
  if $\theta\Gamma<\tau$ **then**
    obtain p_1^{opt} maximizing $\gamma_2^{*(2)}$
  else if $\tau\ge0$ and
    $P_k\ge\Gamma+\frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})\left(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}}\right)$ **then**
    obtain p_1^{opt} maximizing $\gamma_2^{*(2)}$
  else
    obtain p_1^{opt} maximizing $\gamma_2^{*(3)}$
  end if
  Obtain α_1^*,β_1^* and α_2^* using (25), (22) and (26) with $p_1=p_1^{opt}$, and obtain $\mathbf{w}_1$ and $\mathbf{w}_2$ from (14) and (15) with α_1^*,β_1^* and α_2^*.
end function

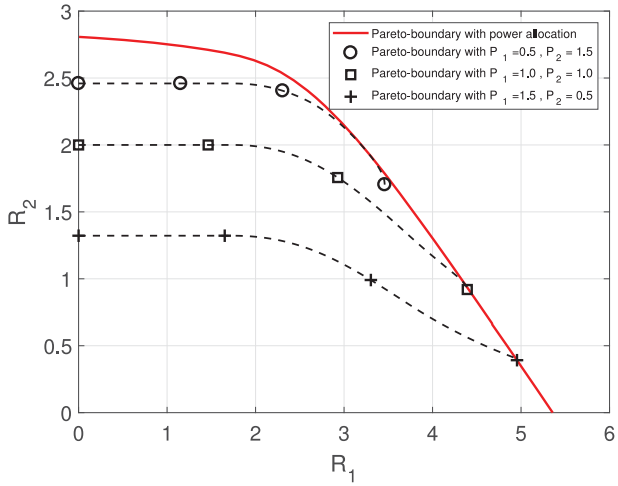

Fig. 1. Pareto-boundary: fixed power versus power allocation ($\|\mathbf{h}_1\|^2/\sigma_1^2=20$, $\|\mathbf{h}_2\|^2/\sigma_2^2=3$, $\theta=0.5$ and $P_k=2$).

IV. Two-User MISO Broadcast Channels with SIC: Properties of Pareto-Optimal Beams

In the previous section, we developed a Pareto-optimal beam design and power allocation algorithm for given channel vectors

$\mathbf{h}_1$ and $\mathbf{h}_2$. The performance of the Pareto-optimal design is a function of the two (effective) channel vectors $\mathbf{h}_1$ and $\mathbf{h}_2$. In the conventional ZF downlink beamforming with no SIC at the receivers, two users with orthogonal channel vectors are preferred since non-orthogonality between the two channel vectors reduces the effective SINR of ZF beamforming [15]. However, in the considered MISO-NOMA framework in which user 1 intends to decode the interference from user 2 and subtracts it before decoding its own data whereas user 2 treats the interference from user 1 as noise, orthogonality between $\mathbf{h}_1$ and $\mathbf{h}_2$, i.e., $\theta \approx 0$, and corresponding orthogonal beam vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ (see (14) and (15) with $\mathbf{h}_1 \perp \mathbf{h}_2$) do not necessarily imply high performance. Intuitively, if θ is small, user 2 receives less interference from user 1, but user 1 has difficulty in decoding the message of user 2 for SIC under the NOMA framework. In this section, we investigate the properties of Pareto-optimal beams in the two-user MISO-NOMA BC before proceeding to overall user scheduling in the next section.

To gain some insight into good channel conditions for two-user MISO-NOMA BCs, let us first consider the fixed power allocation case with $p_1 = p_2 = 1$ and investigate the impact of the angle θ between the two channel vectors when the magnitudes are given. For given $||\mathbf{h}_1||$, $||\mathbf{h}_2||$ and γ_1^*, the SINR of user 2 γ_2^* in (30) can be rewritten as a function of θ as

$$\gamma_2^*(\theta) = \begin{cases} \gamma_2^{*(1)} \ (= \frac{\lambda_1}{1+\Gamma\lambda_1}) & \text{for case 1,} \\ \gamma_2^{*(2)}(\theta) \ (= \frac{\lambda_1}{1+\Gamma\lambda_1}[\alpha_2^*(\theta)]^2) & \text{for case 2,} \\ \gamma_2^{*(3)}(\theta) \ (= \frac{\lambda_2}{\lambda_2[\alpha_1^*(\theta)]^2+1}) & \text{for case 3,} \end{cases} \tag{32}$$

where

$$\alpha_1^*(\theta) = \begin{cases} 0 & \text{if} \quad \Gamma \le 1-\theta \\ \sqrt{\theta\Gamma} - \sqrt{(1-\theta)(1-\Gamma)} & \text{if} \quad \Gamma > 1-\theta \end{cases} \tag{33}$$

$$\alpha_2^*(\theta) = \begin{cases} 1 & \text{for case 1} \\ \frac{c(\theta)}{\sqrt{c^2(\theta)+(a-b(\theta))^2}} & \text{for case 2} \\ \sqrt{\theta} & \text{for case 3,} \end{cases} \tag{34}$$

and

$$a := \frac{\|\mathbf{h}_1\|}{\sqrt{\sigma_1^2(1+\gamma_1^*)}}, \tag{35}$$

$$b(\theta) := \frac{\|\mathbf{h}_2\|\sqrt{\theta}}{\sqrt{\|\mathbf{h}_2\|^2(\alpha_1^*(\theta))^2+\sigma_2^2}}, \tag{36}$$

$$c(\theta) := \frac{\|\mathbf{h}_2\|\sqrt{1-\theta}}{\sqrt{\|\mathbf{h}_2\|^2(\alpha_1^*(\theta))^2+\sigma_2^2}}. \tag{37}$$

Here, case 1, case 2 and case 3 mean the cases of $p_1 \in \mathcal{P}_1$, $p_1 \in \mathcal{P}_2$, $p_1 \in \mathcal{P}_3$, respectively, with $p_1 = p_2 = 1$. Regarding optimal θ that maximizes $\gamma_2^*(\theta)$ in (32), we have the following proposition:

Proposition 2: Let λ_1, λ_2 and Γ be given. If $\Gamma \in [\Gamma_1, \Gamma_2]$, where

$$\begin{aligned} \Gamma_1 &= \frac{1}{2}\left(1+\lambda_2^{-1}-\lambda_1^{-1}\right) \\ &\quad - \frac{1}{2}\sqrt{(1+\lambda_1^{-1}+\lambda_2^{-1})^2 - 4\lambda_2^{-1}(1+\lambda_2^{-1})} \\ \Gamma_2 &= \frac{1}{2}\left(1+\lambda_2^{-1}-\lambda_1^{-1}\right) \\ &\quad + \frac{1}{2}\sqrt{(1+\lambda_1^{-1}+\lambda_2^{-1})^2 - 4\lambda_2^{-1}(1+\lambda_2^{-1})}, \end{aligned} \tag{38}$$

then optimal θ that maximizes $\gamma_2^*(\theta)$ in (32) is given by the region (39) shown at the bottom of this page, where $z_1 = \lambda_1^{-1} + 1 - \Gamma$, $z_2 = \lambda_2^{-1} + 1 - \Gamma$, and

$$\theta_0 = \begin{cases} \frac{\lambda_1}{\lambda_2}\frac{1}{1+\Gamma\lambda_1}, & \text{if} \ \frac{\lambda_1}{\lambda_2}\frac{1}{1+\Gamma\lambda_1} \le 1-\Gamma, \\ \frac{z_1 z_2 + 2\Gamma(1-\Gamma) - \sqrt{4\Gamma(1-\Gamma)[\Gamma(1-\Gamma)+z_1 z_2 - z_2^2]}}{z_1^2+4\Gamma(1-\Gamma)}, & \text{o.w.} \end{cases} \tag{40}$$

If $\Gamma \notin [\Gamma_1, \Gamma_2]$, optimal θ is given by the region

$$\begin{aligned} &\{\theta \,|\, \tfrac{\lambda_2}{\lambda_1}(1+\Gamma\lambda_1) \le \theta \le 1-\Gamma\} \\ &\qquad \text{if } \Gamma \le \tfrac{\lambda_2^{-1}-\lambda_1^{-1}}{1+\lambda_2^{-1}} \text{ and } \Gamma \notin [\Gamma_1, \Gamma_2], \\ &\text{or } \{\theta \,|\, \tfrac{\partial \gamma_2^{(2)}(\theta)}{\partial \theta} = 0\} \\ &\qquad \text{if } \Gamma > \tfrac{\lambda_2^{-1}-\lambda_1^{-1}}{1+\lambda_2^{-1}} \text{ and } \Gamma \notin [\Gamma_1, \Gamma_2]. \end{aligned} \tag{41}$$

Proof. See Appendix. ■

From Proposition 2 we obtain a more insightful corollary as follows:

Corollary 1: Let λ_1, λ_2 and Γ be given. When $\lambda_1 = \lambda_2$, optimal θ for the 2-user MISO BC with SIC is given by the set $\{\theta \,|\, \theta_0 \le \theta \le 1\}$ with θ_0 reduced to

$$\theta_0 = \begin{cases} \frac{1}{1+\Gamma\lambda_1} & \text{if} \ \frac{1}{1+\Gamma\lambda_1} \le 1-\Gamma \\ \frac{z_1^2}{z_1^2+4\Gamma(1-\Gamma)} & \text{if} \ \frac{1}{1+\Gamma\lambda_1} > 1-\Gamma \end{cases}. \tag{42}$$

Proof: See Appendix. ■

Corollary 1 states that in two-user MISO BCs with SIC with the same channel magnitudes $\lambda_1 = \lambda_2$ and the same power $p_1 = p_2 = 1$, two aligned channel vectors are preferred to two orthogonal channels and channel alignment beyond a certain angle is all optimal.

Although Proposition 2 and Corollary 1 provide some insight into good channel conditions in two-user MISO BCs with SIC, the assumptions for Proposition 2 and Corollary 1 are not valid in the actual NOMA situation in which power allocation is applied. Unfortunately, the optimal power p_1^{opt} was not obtained in closed form in the previous section and this puts difficulty on analysis of the impact of channel angle on the performance. Hence, in the actual case, to enable analysis we derive the SINR γ_2^* for user 2 as a function of θ by assuming the simple power allocation method that assigns minimum power $p_{1,\min}(= \gamma_1^*/\lambda_1 = \Gamma)$ to achieve the target SINR γ_1^* for user 1 and assigns the rest of power to user 2. The simple power allocation strategy is based

$$\left\{ \theta_0 \le \theta \le \frac{z_1 z_2 + 2\Gamma(1-\Gamma) + \sqrt{4\Gamma(1-\Gamma)[\Gamma(1-\Gamma)+z_1 z_2 - z_2^2]}}{z_1^2+4\Gamma(1-\Gamma)} \right\} \tag{39}$$

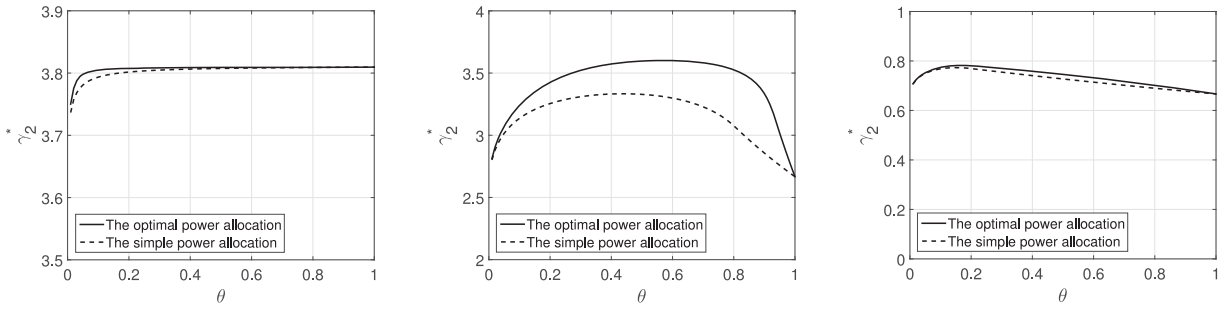

Fig. 2. γ_2^* in (43) as a function of θ ($\lambda_1 = \|\mathbf{h}_1\|^2/\sigma_1^2 = 10$, $\gamma_1^* = 20$, $P_k = 10$): (a) $\lambda_2 = \|\mathbf{h}_2\|^2/\sigma_2^2 = 10$, (b) $\lambda_2 = 1$, and (c) $\lambda_2 = 0.1$.

on the assumption that user 1 has a strong channel and is limited by channel's DoF such as bandwidth, whereas user 2 with a weak channel is limited by noise and needs to receive more power. For this simple power allocation method, substituting $p_1 = p_{1,\min} = \Gamma$ into (30) and applying some manipulation, we obtain the optimal γ_2^* in closed form, given by

$$\gamma_2^* = \begin{cases} \frac{P_k-\Gamma}{\Gamma}\frac{1}{\lambda_1^{-1}\Gamma^{-1}+1}\left[1+\frac{\theta}{1-\theta}\left(\sqrt{\frac{1+\lambda_2^{-1}\Gamma^{-1}\theta^{-1}}{1+\lambda_1^{-1}\Gamma^{-1}}}-1\right)^2\right]^{-1}, & \text{if } \theta \le \theta_1 \\ \frac{P_k-\Gamma}{\Gamma}\frac{1}{\theta+\lambda_2^{-1}\Gamma^{-1}}, & \text{if } \theta > \theta_1, \end{cases} \tag{43}$$

where $\theta_1 := \frac{1}{2}[-\lambda_2^{-1}\Gamma^{-1} + \sqrt{\lambda_2^{-2}\Gamma^{-2} + 4(\lambda_1^{-1}\Gamma^{-1}+1)}]$.

Examples of γ_2^* in (43) as a function of θ are shown in Fig. 2 together with the optimal γ_2^* obtained by running Algorithm 1. It is seen that the behavior of γ_2^* depends on the relative magnitude of λ_1 and λ_2 through the two performance limiting factors: the SIC processing at user 1 and the SINR of user 2, as seen in (13). In the case of $\lambda_1 = \lambda_2 \gg 0$, the performance of user 2 is not limited by noise at user 2 but is limited by signal-to-interference ratio (SIR). Thus, in this case it is preferred that channel vectors are aligned and more power is allocated to user 2 under the constraint that the required SINR for user 1 is satisfied. By doing so, SIC at user 1 is easy and SIR at user 2 is high. Therefore, it is seen in Fig. 2(a) that the performance increases as $\theta \uparrow 1$. However, in the medium asymmetric case of $\lambda_1 > \lambda_2$ like in Fig. 2(b), there is a trade-off between the two performance limiting factors. When two channels are orthogonal, SIC at user 1 is difficult. On the other hand, when two channels are aligned, interference from user 1 to user 2 at user 2 is high. Hence, the performance is good when the two user channels are neither too orthogonal nor too aligned. This behavior is clearly seen in Fig. 2(b). Note that in this case, there is a noticeable performance gap between the optimal power allocation and the simple power allocation, as seen in Fig. 2(b).

In addition to the above simple power allocation result, we have another result exploiting the fact $\lambda_1 \gg \lambda_2$ in actual NOMA, given by the following proposition:

Proposition 3: For given λ_1 and γ_1^*, if $\theta \neq 0$, then as $\lambda_2 \to 0$, the optimal power value $p_1^{opt} \to p_{1,\min} = \Gamma$; the corresponding γ_2^* converges to $\gamma_2^* = \frac{P_k-\Gamma}{\Gamma}\frac{1}{\theta+\lambda_2^{-1}\Gamma^{-1}}$; and the beam vectors converge to $\sqrt{p_1}\mathbf{w}_1 = \sqrt{\Gamma}\frac{\mathbf{h}_1}{\|\mathbf{h}_1\|}$ and $\sqrt{p_2}\mathbf{w}_2 = \sqrt{P_k-\Gamma}\frac{\mathbf{h}_2}{\|\mathbf{h}_2\|}$. That is, both users use matched-filtering beams.

Proof: See Appendix. ■

Note that γ_2^* in Proposition 3 coincides with the second formula in (43). This is because θ_1 in (43) converges to 0^+ as $\lambda_2 \to 0$ for given λ_1 and Γ, and the second formula in (43) is valid in this case. By Proposition 3, if λ_2 is sufficiently small compared to λ_1, matched filtering beams for both users with minimum power to user 1 satisfying the target SINR are optimal regardless of the angle between the two channel vectors. This is because if $\lambda_1 \gg \lambda_2$, the limitation for γ_2^* results from the SINR of user 2 at user 2. Hence, maximum power should be delivered to user 2 with matched filtering beam $\mathbf{w}_2$ by assigning minimum power to user 1 with matched filtering beam $\mathbf{w}_1$. Therefore, it is seen in Fig. 2(c) that the simple power allocation method almost achieves the optimal performance for all θ.

V. Overall User Scheduling, Beam Design and Power Allocation for MU-MISO-NOMA Downlink

In this section, we present our user scheduling/pairing, beam design and cluster power allocation method for MU-MISO-NOMA downlink under the considered hierarchical beam design strategy in which ZF is applied for inter-cluster beamforming and the two-user Pareto-optimal beam design in Sections III and IV is applied for intra-cluster beamforming. The joint design of user scheduling, transmission beams and cluster power allocation is difficult. To circumvent this difficulty, we approach the problem in two steps. In the first step, we perform user selection under the assumption of equal cluster power allocation. In the second step, with the selected users from the first step, we perform power allocation and beam design. Although the proposed two-step approach is suboptimal, it provides an effective solution to the complicated joint problem of user scheduling, beam design and cluster power allocation.

A. User Selection

For MU-MISO-NOMA downlink scheduling with the considered hierarchical beam design strategy, two major aspects should be taken into account simultaneously to guarantee good system performance: One is controlling ICI to reduce interference from other clusters and the other is pairing and beam design for the paired users in each cluster. Recall that the gain of NOMA lies in the case that strong users are in the DoF(such as bandwidth)-limited regime and weak users are limited by noise [2, P. 239]. In MU-MISO NOMA, the beneficial situation for NOMA should be maintained, i.e., the high channel quality for strong users should be maintained by proper interference control, and low SINRs of weak users should be leveraged by

assigning more power to weak users. Considering these factors, we propose a two-step user scheduling-and-pairing method under the assumption that all channel vector information is available at the BS and the thermal noise variance is known. In the first step, $K_c\,(\leq N_t)$ strong users are selected from the strong user set $\mathcal{K}_1$. This is done by running the semi-orthogonal user selection (SUS) algorithm [15] targeting selection of N_t users from the strong user set $\mathcal{K}_1$. Then, the SUS algorithm returns $K_c\,(\leq N_t)$ users with roughly orthogonal channel vectors out of the $K/2$ users in $\mathcal{K}_1$. Depending on the size $|\mathcal{K}_1|$ and the semi-orthogonality parameter, the SUS algorithm may return users less than N_t although we target selecting N_t users [15]. We set these K_c users returned by the SUS algorithm as the K_c strong users in K_c clusters (one user for each cluster). In the second step, weak users are selected from the weak user set $\mathcal{K}_2$ so that the selected weak users are well matched to the already selected strong users for good performance based on the results in Section III and IV. The detail of the proposed method is described in Algorithm 2.

Remark 3 (Rate control between strong and weak users): Note that rate control between the strong and weak users for each cluster can be done by changing the target SINR for the strong user. For this we defined the parameter Γ as the ratio of the strong user's target SINR $\gamma_1^{*(k)}$ to the strong user's channel SNR $\lambda_1^{(k)}$ in (23), i.e., for cluster k

$$\Gamma_k := \frac{\gamma_1^{*(k)}}{\lambda_1^{(k)}} = \frac{\gamma_1^{*(k)}}{||\mathbf{h}_1^{(k)}||^2/(\sigma_1^{(k)})^2} \tag{44}$$

(The cluster index is now added.) The feasible range of Γ_k is given by $[0, P_k]$, where the value zero corresponds to the case of no power assigned to the strong user and the maximum P_k occurs when $\mathbf{w}_1^{(k)} = \mathbf{h}_1^{(k)}/||\mathbf{h}_1^{(k)}||$ and $p_1^{(k)} = P_k$. Thus, the feasible range of Γ_k varies with the cluster index k when the cluster power values P_k's are different from each other. By normalizing with P_k, we define the rate control parameter η as

$$\eta := \Gamma_k/P_k, \quad 0 \leq \eta \leq 1. \tag{45}$$

Now, controlling the single parameter η common to all clusters, we can control the rate balance between the strong and weak users in all clusters. $\eta = 0$ corresponds to all power to the weak users in all clusters and $\eta = 1$ corresponds to all power to the strong users in all clusters. The parameter η appears in the 3rd line of Step 2 of Algorithm 2.

Remark 4 (Use of effective channels): The computation of $\gamma_2^*(u)$ and the Pareto-optimal beam design for two users in each cluster in steps S.2-1 and S.2-2 in Algorithm 2 are based on the *projected effective channels*. Note that the projected effective channels $\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)}$ and $\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_2^{(k)}$ lie in $\mathcal{L}^{\perp}(\tilde{\mathbf{H}}_k)$. Thus, the corresponding Pareto-optimal beams $\mathbf{w}_1^{(k)}$ and $\mathbf{w}_2^{(k)}$ lie in $\mathcal{L}^{\perp}(\tilde{\mathbf{H}}_k)$ by the property of Pareto-optimal beams (see (14) and (15))[22], and hence we have $\mathbf{w}_i^{(k)} = \mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\mathbf{w}_i^{(k)}$, $i = 1, 2$.

Remark 5 (Dimension of the ZF space): If $K_c = N_t$, then $\tilde{\mathbf{H}}_k$ has nullity of one, and $\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)}$ and $\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_2^{(k)}$ are aligned, as shown in Fig. 3 (a). In this case, only Pareto-optimal power control between the two users in each cluster is applied by the proposed design method. (Note that Algorithm 1 is still applicable in case of two aligned input channel vectors.) On the other hand, if the number K_c of the returned users by the SUS algorithm in Step 1 is less than N_t (which is often true for large N_t with small K[15]), then $\tilde{\mathbf{H}}_k$ has nullity larger than or equal to two for all k. In this case, the projected effective channels $\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)}$ and $\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_2^{(k)}$ span a 2-dimensional (2-D) space and the full 2-D Pareto-optimal beam design is applicable for each cluster, as shown in Fig. 3(b). This is another advantage of the proposed method over the previous methods [11]–[13] based on simple spatial ZF beam design ignoring the case of $K_c < N_t$. In fact, we can impose the constraint $K_c \leq N_t - 1$ intentionally

Algorithm 2: User Selection.

Step 1: Run the SUS algorithm [15] targeting selection of N_t users from the strong user set $\mathcal{K}_1$. Then, K_c is determined.

Step 2: Weak user selection

Initialization:
Actual channel vectors of the users in $\mathcal{K}_2$: $\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_{|\mathcal{K}_2|}$.
Total power P_T and rate-control parameter η are given.
Set $P_k = P_T/K_c, \forall k$ and set $\Gamma_k = \eta P_k$.
Actual channel vectors of the strong users from Step 1: $\tilde{\mathbf{h}}_1^{(1)}, \ldots, \tilde{\mathbf{h}}_1^{(K_c)}$.
$\mathcal{K}_2 \leftarrow \{1, \ldots, K/2\}$ ▷the original weak user set
$\mathcal{S}_1$ ▷the set of selected strong users from step 1
$\mathcal{S}_2 \leftarrow \phi$ ▷the set of selected weak users
$\widehat{\mathbf{W}} = \left[\frac{\mathbf{\Pi}_{\tilde{\mathbf{H}}_1}^{\perp}\tilde{\mathbf{h}}_1^{(1)}}{||\mathbf{\Pi}_{\tilde{\mathbf{H}}_1}^{\perp}\tilde{\mathbf{h}}_1^{(1)}||}, \ldots, \frac{\mathbf{\Pi}_{\tilde{\mathbf{H}}_{K_c}}^{\perp}\tilde{\mathbf{h}}_1^{(K_c)}}{||\mathbf{\Pi}_{\tilde{\mathbf{H}}_{K_c}}^{\perp}\tilde{\mathbf{h}}_1^{(K_c)}||}\right]$,
$\widetilde{\mathbf{W}}_1 = [\,]$ and $\widetilde{\mathbf{W}}_2 = [\,]$,
Execution loop:
for $k = 1$ to K_c **do**
(S.2-1) Compute γ_2^* for all candidates for the weak user for cluster k.
for $u = 1$ to $|\mathcal{K}_2|$ **do**

$$\hat{\sigma}_u^2 = \sum_{l<k}(|\tilde{\mathbf{g}}_u^H\widetilde{\mathbf{W}}_1(l)|^2 + |\tilde{\mathbf{g}}_u^H\widetilde{\mathbf{W}}_2(l)|^2) + \sum_{l>k}P_k|\tilde{\mathbf{g}}_u^H\widehat{\mathbf{W}}(l)|^2 + \epsilon_u^2, \tag{46}$$

where $\widehat{\mathbf{W}}(l), \widetilde{\mathbf{W}}_1(l)$ and $\widetilde{\mathbf{W}}_2(l)$ are the l-th columns of $\widehat{\mathbf{W}}, \widetilde{\mathbf{W}}_1$ and $\widetilde{\mathbf{W}}_2$ respectively.
Compute $\gamma_2^*(u)$ as described in Section III-A or Section IV with
$\lambda_1 = \frac{\|\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)}\|^2}{\sigma_1^2}$, $\lambda_2 = \frac{\|\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{g}}_u\|^2}{\hat{\sigma}_u^2}$, and
$\theta = \frac{\left|(\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)})^H(\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{g}}_u)\right|^2}{\|\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)}\|^2\|\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{g}}_u\|^2}$.
end for
(S.2-2) Select the weak user for cluster k.
$u^* = \arg\max_{u\in\mathcal{K}_2}\gamma_2^*(u)$
$\mathcal{S}_2 \leftarrow \mathcal{S}_2 \cup \{u^*\}$ and $\tilde{\mathbf{h}}_2^{(k)} = \tilde{\mathbf{g}}_{u^*}$
Run Algorithm 1: $[\sqrt{p_1}\mathbf{w}_1^{(k)}, \sqrt{p_2}\mathbf{w}_2^{(k)}]$
$=\mathcal{D}(\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_1^{(k)}, \mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\tilde{\mathbf{h}}_2^{(k)}, \sigma_1^2, \hat{\sigma}_{u^*}^2, \lambda_1\Gamma_k, P_k)$
(S.2-3) Update:
$\widetilde{\mathbf{W}}_1 \leftarrow [\widetilde{\mathbf{W}}_1, \sqrt{p_1}\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\mathbf{w}_1^{(k)}]$
$\widetilde{\mathbf{W}}_2 \leftarrow [\widetilde{\mathbf{W}}_2, \sqrt{p_2}\mathbf{\Pi}_{\tilde{\mathbf{H}}_k}^{\perp}\mathbf{w}_2^{(k)}]$
$\mathcal{K}_2 \leftarrow \mathcal{K}_2 \setminus \{u^*\}$
end for

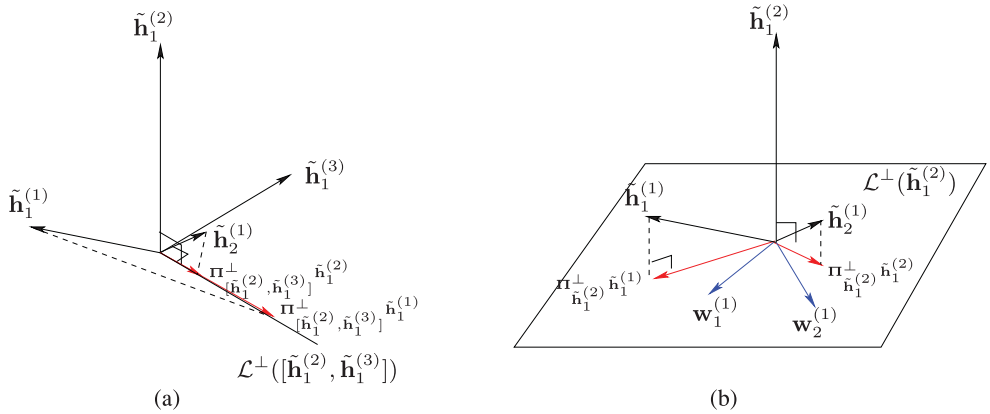

Fig. 3. Beam design in the example of $N_t = 3$: (a) $K_c = N_t$ and (b) $K_c = N_t - 1$.

in order to exploit the freedom of full 2-D beam design for the two users in each cluster.

Remark 6 (ICI estimation for weak users): In the step (46) of computing the power of ICI plus AWGN for each candidate weak user for cluster k, already designed beam vectors are used up to cluster $k-1$ and the beam estimates $\sqrt{P_k}\frac{\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_{k+1}}\tilde{\mathbf{h}}_1^{(k+1)}}{\|\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_{k+1}}\tilde{\mathbf{h}}_1^{(k+1)}\|},\ldots,\sqrt{P_k}\frac{\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_{K_c}}\tilde{\mathbf{h}}_1^{(K_c)}}{\|\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_{K_c}}\tilde{\mathbf{h}}_1^{(K_c)}\|}$ are used for undesigned clusters $k+1,\ldots,K_c$.

Remark 7 (Effective channel gain loss for strong users and selection of weak users): Note that there is length reduction from the actual channel $\tilde{\mathbf{h}}_1^{(k)}$ to the effective channel $\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\tilde{\mathbf{h}}_1^{(k)}$ of each strong user. This reduction is the typical effective gain loss associated ZF, but the loss is not significant because the strong channels $\tilde{\mathbf{h}}_1^{(1)},\ldots,\tilde{\mathbf{h}}_1^{(K_c)}$ are semi-orthogonal by the SUS algorithm[15]. Only the weak users experience ICI whereas ICI to the strong users is completely removed in the proposed method to be consistent with the NOMA design principle. However, the weak users are selected by considering all the factors, i.e., the ICI, projection onto $\mathcal{L}^{\perp}(\tilde{\mathbf{H}}_k)$ and the friendliness with the strong users to yield good performance.

B. Cluster Power Allocation and Beam Redesign

In the previous section, we considered user selection and beam design under the assumption of equal cluster power allocation, as shown in the fourth line of Step 2 in Algorithm 2. Now, let us consider performance improvement by cluster power allocation and resulting beam redesign for the given set of selected strong and weak users from Section V-A. Since the rate balance between the strong and weak users is determined by fixing the parameter η in (45) common to all clusters, one reasonable approach to cluster power allocation is to maximize the cluster sum rate for the given η that is used in the user selection step. To tackle this problem, we exploit our result Proposition 3 with the assumption of sufficiently unequal channel gains for strong and weak users. By Proposition 3, the rates of user 1 (the strong user) and user 2 (the weak user) in cluster k can be expressed as

$$R_1^{(k)} = \log_2\left(1+\lambda_1^{(k)}\Gamma_k\right)$$
$$R_2^{(k)} = \log_2\left(1+\frac{P_k-\Gamma_k}{\Gamma_k}\frac{1}{\theta^{(k)}+(\lambda_2^{(k)})^{-1}\Gamma_k^{-1}}\right) \quad (47)$$

where P_k is the total power assigned to cluster k, Γ_k is the normalized target SINR for user 1 in cluster k shown in (44), $\lambda_1^{(k)} = \|\mathbf{h}_1^{(k)}\|/(\sigma_1^{(k)})^2$, $\lambda_2^{(k)} = \|\mathbf{h}_2^{(k)}\|/(\sigma_2^{(k)})^2$, and $\theta^{(k)} = \frac{|\mathbf{h}_1^{(k)H}\mathbf{h}_2^{(k)}|^2}{\|\mathbf{h}_1^{(k)}\|^2\|\mathbf{h}_2^{(k)}\|^2}$. Here, $\mathbf{h}_1^{(k)}(=\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\tilde{\mathbf{h}}_1^{(k)})$ and $\mathbf{h}_2^{(k)}(=\mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\tilde{\mathbf{h}}_2^{(k)})$ are the projected effective channel vectors of users 1 and 2 in cluster k, respectively. Since ICI is cancelled by SIC at user 1, the variance $(\sigma_1^{(k)})^2$ of ICI plus AWGN of user 1 is the same as the AWGN variance $(\epsilon_1^{(k)})^2$ (see (4) and (7)). However, the variance $(\sigma_2^{(k)})^2$ of ICI plus AWGN of user 2 is given from (3) by

$$(\sigma_2^{(k)})^2 = (\epsilon_2^{(k)})^2 + \sum_{k'\neq k}\left(p_1^{(k')}|\tilde{\mathbf{h}}_2^{(k)H}\tilde{\mathbf{w}}_1^{(k')}|^2 + p_2^{(k')}|\tilde{\mathbf{h}}_2^{(k)H}\tilde{\mathbf{w}}_2^{(k')}|^2|\right), \quad (48)$$

where $\tilde{\mathbf{w}}_1^{(k)} := \mathbf{\Pi}^{\perp}_{\tilde{\mathbf{H}}_k}\mathbf{w}_1^{(k)}$. Furthermore, Proposition 3 states that the Pareto-optimal beam vectors in each cluster can be approximated by

$$\sqrt{p_1^{(k)}}\tilde{\mathbf{w}}_1^{(k)} = \sqrt{\Gamma_k}\frac{\mathbf{h}_1^{(k)}}{\|\mathbf{h}_1^{(k)}\|}, \sqrt{p_2^{(k)}}\tilde{\mathbf{w}}_2^{(k)} = \sqrt{P_k-\Gamma_k}\frac{\mathbf{h}_2^{(k)}}{\|\mathbf{h}_2^{(k)}\|}. \quad (49)$$

Based on (45), (48), and (49), $(\sigma_2^{(k)})^2$ can be rewritten as

$$(\sigma_2^{(k)})^2 = (\epsilon_2^{(k)})^2 + \|\mathbf{h}_2^{(k)}\|^2\sum_{k'\neq k}P_{k'}\nu_{k,k'} \quad (50)$$

where $\nu_{k,k'} := (\eta\frac{|\tilde{\mathbf{h}}_2^{(k)H}\mathbf{h}_1^{(k')}|^2}{\|\mathbf{h}_2^{(k)}\|^2\|\mathbf{h}_1^{(k')}\|^2} + (1-\eta)\frac{|\tilde{\mathbf{h}}_2^{(k)H}\mathbf{h}_2^{(k')}|^2}{\|\mathbf{h}_2^{(k)}\|^2\|\mathbf{h}_2^{(k')}\|^2})$. Then, the rates in (47) are expressed in terms of P_k's as

$$R_1^{(k)} = \log_2\left(1+\bar{\lambda}_1^{(k)}\eta P_k\right)$$
$$R_2^{(k)} = \log_2\left(1+\frac{(1-\eta)\bar{\lambda}_2^{(k)}P_k}{1+\bar{\lambda}_2^{(k)}[\sum_{k'\neq k}P_{k'}\nu_{k,k'}+\theta^{(k)}\eta P_k]}\right) \quad (51)$$

where $\bar{\lambda}_i^{(k)} := \|\mathbf{h}_i^{(k)}\|^2/(\epsilon_i^{(k)})^2$ for $i=1,2$. Thus, the cluster sum rate maximization for given η is formulated as the problem (52)–(53), shown at the bottom of the next page. The objective function in (52) can be expressed as the difference of two

functions, $f(\mathbf{P}) - g(\mathbf{P})$, where

$$f(\mathbf{P}) = \sum_{k=1}^{K_c} \left[\log_2(1 + \bar{\lambda}_1^{(k)} \eta P_k) + \log_2 \left(1 + \bar{\lambda}_2^{(k)} \left[(1 - \eta + \eta\theta^{(k)}) P_k + \sum_{k' \neq k} P_{k'} \nu_{k,k'} \right] \right) \right],$$

$$g(\mathbf{P}) = \sum_{k=1}^{K_c} \log_2 \left(1 + \bar{\lambda}_2^{(k)} \left[\sum_{k' \neq k} P_{k'} \nu_{k,k'} + \theta^{(k)} \eta P_k \right] \right),$$

and $\mathbf{P} := \{P_1, \ldots, P_{K_c}\}$. Therefore, the problem (52)–(53) is equivalent to

$$\min -[f(\mathbf{P}) - g(\mathbf{P})], \ \mathbf{P} \in \left\{ \mathbf{P} \,|\, 0 \leq P_i, \sum_1^{K_c} P_k \leq P_T \right\}. \quad (54)$$

Note that $-\log_2(\cdot)$ is a convex function and the inner terms in the logarithms in $f(\mathbf{P})$ and $g(\mathbf{P})$ are linear functions of $\mathbf{P}$. From the fact that the sum of convex functions is convex, $-f(\mathbf{P})$ and $-g(\mathbf{P})$ are convex functions of $\mathbf{P}$. Hence, the problem is a difference of convex functions programming (DC programming) [25], and this problem can easily be solved by the prismatic branch and bound (PBnB) algorithm [26].

Once the new cluster power allocation for all clusters is obtained from (54), we can redesign the intra-cluster beam vectors $\{(\mathbf{w}_1^{(k)}, \mathbf{w}_2^{(k)}), k = 1, \ldots, K_c\}$ for the already selected users from Section V-A by using Algorithm 1.

VI. Pareto-Optimal Beam Design Beyond the Two-User Case

Up to now, we have considered grouping two users in each cluster. In this section, we consider a general MU-MISO BC with SIC in which more than two users are grouped in a cluster and devise a method for designing Pareto-optimal beams in this general case.

In an M-user MISO BC with beamforming and SIC, the system model is given by

$$y_j = \mathbf{h}_j^H \left(\sum_{i=1}^{M} \sqrt{p_i} \mathbf{w}_i s_i \right) + n_j = \mathbf{h}_j^H \left(\sum_{i=1}^{M} \bar{\mathbf{w}}_i s_i \right) + n_j, \quad j \in \{1, 2, \ldots, M\}, \quad (55)$$

where $\bar{\mathbf{w}}_i := \sqrt{p_i}\mathbf{w}_i$. (The notations are similarly defined like in Section III.) With the assumption that the users are ordered according to their channel SNR as $\|\mathbf{h}_1\|^2/\sigma_1^2 \geq \|\mathbf{h}_2\|^2/\sigma_2^2 \geq \ldots \geq \|\mathbf{h}_M\|^2/\sigma_M^2$, SIC at the receivers is applied such that user j decodes the interference from users $M, M-1, \ldots, j+1$ and cancels it. Then, the rates of the users are given by

$$R_1(\bar{\mathbf{w}}_1) = \log_2 \left(1 + \frac{|\mathbf{h}_1^H \bar{\mathbf{w}}_1|^2}{\sigma_1^2} \right) \quad (56)$$

$$R_i(\bar{\mathbf{w}}_1 \ldots, \bar{\mathbf{w}}_i) = \log_2 \left(1 + \min \left\{ \mathrm{SINR}_1^i, \ldots, \mathrm{SINR}_i^i \right\} \right) \quad i = 2, \ldots, M,$$

where SINR_j^i is the required SINR for the signal of user i such that user j can decode the message of user i, given by

$$\mathrm{SINR}_j^i = \frac{|\mathbf{h}_j^H \bar{\mathbf{w}}_i|}{\sum_{i'=1}^{i-1} |\mathbf{h}_j^H \bar{\mathbf{w}}_{i'}|^2 + \sigma_j^2}. \quad (57)$$

Here, it is assumed that user j decodes and cancels the messages in the order of $s_M, s_{M-1}, \ldots, s_i$. Hence, for SINR_j^i the interference from the messages of users $M, M-1, \ldots, i+1$ are cancelled but the interference from the messages of users $1, 2, \ldots, i-1$ remains. Similarly to the two-user case, the Pareto-optimal beam design problem can be cast as

$$\begin{aligned} \max_{(\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_M)} \quad & R_M(\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_M) \\ \text{subject to} \quad & R_1(\bar{\mathbf{w}}_1) = R_1^* \\ & R_i(\bar{\mathbf{w}}_1, \ldots \bar{\mathbf{w}}_i) = R_i^*, \quad i = 2, \ldots, M-1 \\ & \sum_{i=1}^{M} \|\bar{\mathbf{w}}_i\|^2 \leq P, \end{aligned} \quad (58)$$

where $(R_1^*, \ldots, R_{M-1}^*)$ is a rate-tuple of given target rates of users $1, \ldots, M-1$ out of the feasible target rate-tuple set $\mathcal{Q}_{M-1}$.

In case of a general M-user MISO BC channel, Pareto-optimal parameterization can be extended from the result in [27]. However, in the general M-user MISO BC with multi-user SIC, an efficient parameterization of Pareto-optimal beams is not straightforward. Hence, we tackle the optimization problem (58) directly with a convex programming approach. If the feasible target rate-tuple set $\mathcal{Q}_{M-1}$ is known, then (58) can be solved by using the reformulation technique in [16] and convex-concave procedure (CCP) [28], and the Pareto-boundary can be computed. However, it is not straightforward to know $\mathcal{Q}_{M-1}$ since the rates depend on the beam vectors and SINR in a complicated manner. In order to obtain the Pareto-boundary of the M-user MISO BC with SIC, we introduce an induction approach to compute the feasible target rate-tuple set $\mathcal{Q}_{M-1}$. That is, when $M = 2$, (i.e., the two-user case), $\mathcal{Q}_1$ is given by $\{R_1^* \,|\, 0 \leq R_1^* \leq R_1^{*,\max}\}$, where $R_1^{*,\max} = \log_2(1 + \frac{\|\mathbf{h}_1\|^2}{\sigma_1^2} P)$, which corresponds to the case that the total power is allocated only to user 1. In the case of $M = 3$, $\mathcal{Q}_2$ is obtained when the total power P is allocated only to users 1 and 2. In this

$$\max_{\{P_1, \ldots, P_{K_c}\}} \sum_{k=1}^{K_c} \left(\log_2 \left(1 + \bar{\lambda}_1^{(k)} \eta P_k \right) + \log_2 \left(1 + \frac{(1-\eta)\bar{\lambda}_2^{(k)} P_k}{1 + \bar{\lambda}_2^{(k)} \sum_{k' \neq k} P_{k'} \nu_{k,k'} + \theta^{(k)} \bar{\lambda}_2^{(k)} \eta P_k} \right) \right) \quad (52)$$

$$\text{subject to} \sum_{k=1}^{K_c} P_k \leq P_T. \quad (53)$$

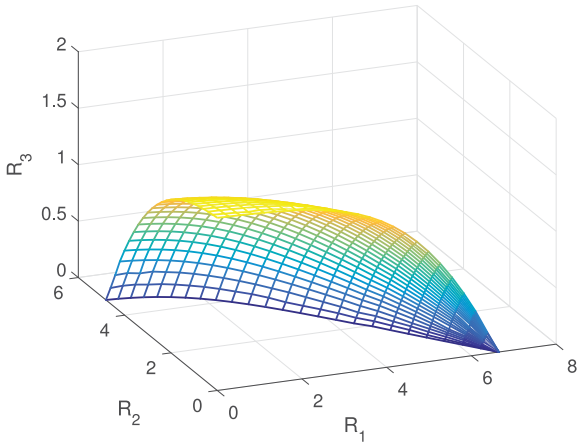

Fig. 4. Pareto-boundary of a 3-user MISO BC with SIC: $N_t = 4$, $\lambda_1 = 3.1$, $\lambda_2 = 0.85$, $\lambda_3 = 0.06$, $10\log_{10} P = 15$.

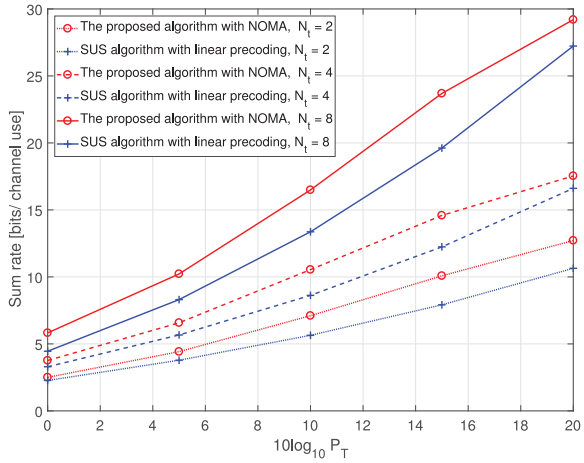

Fig. 5. Total sum rate ($K = 50$).

case, $\mathcal{Q}_2$ is given by $\{(R_1^*, R_2^*) \,|\, 0 \le R_1^* \le R_1^{*,\max}, 0 \le R_2^* \le R_2^{*,\max}(R_1^*)\}$, where $R_2^{*,\max}(R_1^*)$ for given R_1^* can be obtained by $\mathcal{D}(\mathbf{h}_1, \mathbf{h}_2, \sigma_1^2, \sigma_2^2, 2^{R_1^*} - 1, P)$ in Algorithm 1. Then, $\mathcal{Q}_3$ is computed by solving (58) to obtain maximum R_3 for every $(R_1^*, R_2^*) \in \mathcal{Q}_2$ with the total power allocated only to users 1, 2, and 3. In this way of induction, we can obtain the feasible target rate-tuple set $\mathcal{Q}_{M-1}$ for arbitrary M. Once $\mathcal{Q}_{M-1}$ is determined, we can obtain the Pareto-boundary-achieving beam vectors $\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_M$ by solving (58) for a given target rate-tuple in $\mathcal{Q}_{M-1}$. As an illustration, Fig. 4 shows the Pareto-boundary of a 3-user MISO BC with SIC.

Discussion: In the case that we just construct one cluster with M users for MU-MISO NOMA, we can simply apply the proposed M-user Pareto-optimal beam design method. Now, consider the multi-cluster hierarchical beam design strategy (using certain inter-cluster beamforming and Pareto-optimal intra-cluster beamforming) for MU-MISO NOMA with more than two users grouped in each cluster. Although we know how to design Pareto-optimal intra-cluster beam vectors, it is not clear how to design inter-cluster beamforming across multiple clusters in this case. We can still try ZF inter-cluster beamforming eliminating ICI for some strong users in each cluster. For example, suppose that $N_t = 8$ and three users are grouped in each cluster. If we cancel ICI only for the strongest channel user in each cluster by ZF inter-cluster beamforming, maximum 8 clusters can be formulated and hence we can support 24 users. On the other hand, if we cancel ICI for the strongest and second strongest channel users in each cluster, only 4 clusters can be formulated and hence we can support 12 users in this case. Although more users can be supported in the first case, the performance of the second strongest user can be degraded severely from ICI in the first case. Hence, there is a trade-off between inter-cluster ICI cancellation and the number of supportable users for ZF inter-cluster beamforming. So, in the case of more than two users in each cluster, it is more reasonable to handle ICI not only by transmit beamforming but also by receive beamforming by using MIMO. We leave the multi-cluster hierarchical beam design with general cluster size based on MU-MIMO and associated user selection as a future work.

VII. Numerical Results

In this section, we provide some numerical results to evaluate the performance of the proposed scheduling, beam design and power allocation method described in Section V for MU-MISO-NOMA downlink. The basic simulation setting for the numerical results in this section is as follows. K NOMA users in a cell was divided into two subsets $\mathcal{K}_1$ and $\mathcal{K}_2$ and the AWGN variance was one for all users, i.e., $(\epsilon_i^{(k)})^2 = 1, \forall i, k$. Each element of each channel vector in the strong user set $\mathcal{K}_1$ with $K/2$ users was randomly and independently generated from $\mathcal{CN}(0, \sigma_{h,1}^2)$ with $\sigma_{h,1}^2 = 1$ and each element of each channel vector in the weak user set $\mathcal{K}_2$ with $K/2$ users was randomly and independently generated from $\mathcal{CN}(0, \sigma_{h,2}^2)$ with $\sigma_{h,2}^2 = 0.01$. (Hence, we have $\lambda_1/\lambda_2 = 100 = 20$ dB.)[†]

First, we evaluated the gain of the proposed method over conventional SUS-based MU-MISO scheduling [15] and the result is shown in Fig. 5. The simulation setup for Fig. 5 is as follows. For the conventional SUS-based MU-MISO scheduling we considered two scheduling intervals. At the first interval, user scheduling out of $\mathcal{K}_1$ was performed by running the SUS algorithm for MU-MISO with N_t transmit antennas and the scheduled users were served by ZF downlink beamforming[15]. At the second interval, user scheduling out of $\mathcal{K}_2$ was performed by running the SUS algorithm for MU-MISO with N_t transmit antennas and the scheduled users were served by ZF downlink beamforming. The average sum rates for $\mathcal{K}_1$ and $\mathcal{K}_2$ were obtained by averaging the rates of 1000 independent channel realizations. For the overall sum rate of the conventional SUS method, the two rates of $\mathcal{K}_1$ and $\mathcal{K}_2$ were averaged. For the proposed NOMA method, the same 1000 channel realizations used for the conventional method were used but the proposed NOMA scheduling was performed over the overall user set $\mathcal{K}_1 \cup \mathcal{K}_2$ in a single scheduling interval. For the SUS algorithm applied separately to $\mathcal{K}_1$ and $\mathcal{K}_2$ and Step 1 of Algorithm 2, the semi-orthogonality parameter δ (defined in [15]) should be chosen [15] and we used optimal δ for each $K/2$ provided from Fig. 2 of [15]. In addition, for the proposed method, we need to set the rate controlling parameter η defined in (45) to balance the rates from the two groups $\mathcal{K}_1$ and $\mathcal{K}_2$. Here, η was chosen properly for each P_T. It is seen in Fig. 5 that the

[†]Such a model is approximately valid in the following scenario. We group users near the BS as the strong user group and users near the cell edge as the weak user group and support these two-group users using NOMA in one resource block, whereas users in the middle of the cell are supported as a separate group by using a separate resource block.

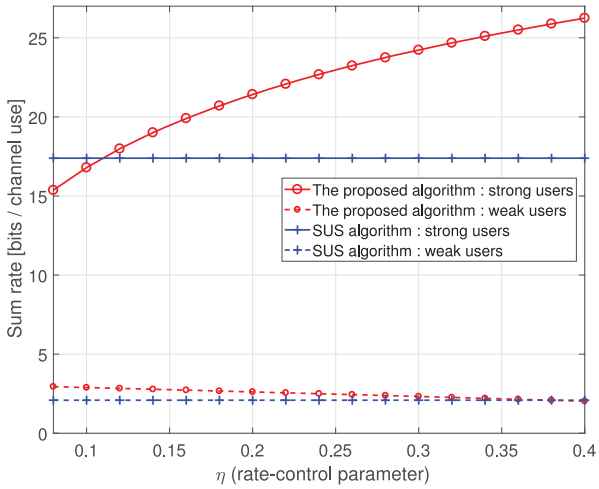

Fig. 6. Sum rate versus η: $N_t = 8, K = 50$ and $10\log_{10} P_T = 15$.

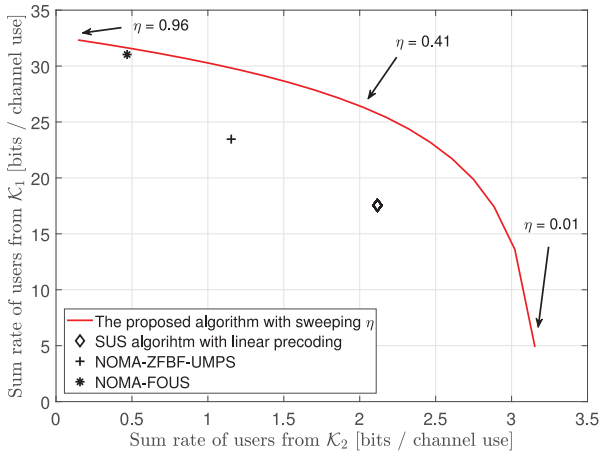

Fig. 7. 2D plot of strong user rate and weak user rate: $N_t = 8$, $K = 50$ and $10\log_{10} P_T = 15$.

proposed MU-MISO NOMA method outperforms the conventional MU-MISO downlink based on the SUS user scheduling. (The corresponding separate sum rate curves with respect to the total transmit power for $\mathcal{K}_1$ and $\mathcal{K}_2$ are available in [24].)

The key feature of our Pareto-optimality-based intra-cluster beam design is that we have control over the rate operating point. Hence, we investigated the rate balancing property between the two groups $\mathcal{K}_1$ and $\mathcal{K}_2$ by controlling the rate control parameter $\eta \in [0, 1]$ defined in (45) (larger η means larger rates for strong users). Fig. 6 shows the rates of strong and weak users versus η. The simulation parameters are the same as those for Fig. 5. For reference, the rates of $\mathcal{K}_1$ and $\mathcal{K}_2$ separately obtained by the conventional MU-MISO SUS algorithm are shown. It is seen that by abandoning the improvement for weak users but maintaining the weak-user performance at the level of the conventional SUS method, significant rate gain can be attained for strong users.

Next, we compared the proposed algorithm with existing algorithms proposed for MU-MISO NOMA downlink. We considered the NOMA-FOUS algorithm in [13] and the NOMA-ZFBF-UMPS algorithm in [12]. Since the simulation setting is different from those in [13] and [12], we slightly modified the two existing algorithms so that the strong user is selected from $\mathcal{K}_1$ and the weak user is selected from $\mathcal{K}_2$, although the original two algorithms consider only one set of users. Fig. 7 shows the resulting 2D plot of strong and weak user rates with varying the rate control parameter η. As shown in the figure, the proposed method based on Pareto-optimal intra-cluster beam design can achieve any combination of the strong and weak user rates along the red solid line. However, each of the existing methods can yield only one point in the 2D rate plane and cannot control the rate balance between strong and weak users. Note that the rate-tuple points of the existing methods are strictly within the red solid line achieved by the proposed method. Hence, there exists net rate gain by the proposed method over the existing methods in addition to its rate controllability.

Finally, we investigated an MMSE inter-cluster beam design approach. For the MMSE approach, we used the SUS algorithm to first select the strong users from the strong user set $\mathcal{K}_1$. (The same set of selected strong users is used for the proposed method.) Based on the selected strong users' channel vectors, we designed MMSE beams for the strong users. Then, the strong user's beam is shared by the weak user in the same

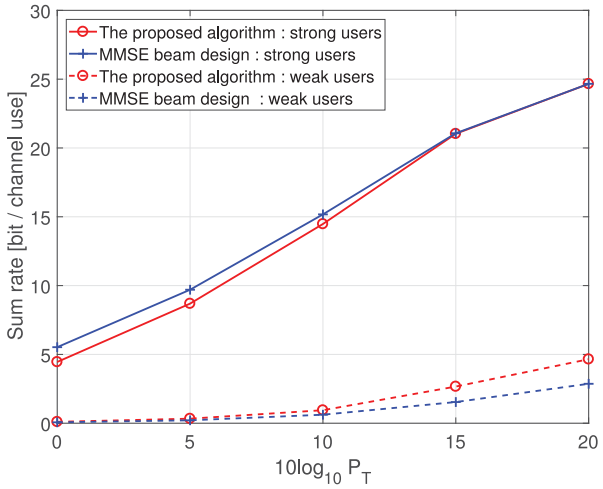

Fig. 8. Separate sum rate from $\mathcal{K}_1$ and $\mathcal{K}_2$ ($N_t = 8$, $K = 50$).

cluster and the power distribution within each cluster is obtained to be Pareto-optimal for the same given η as that used for the ZF inter-cluster beamforming case. Under this beam design strategy, we selected the weak user for each cluster for maximum performance under the considered setup. Fig. 8 shows the result. It is seen that indeed there is some gain by the MMSE approach for the strong users at low SNR. However, as SNR increases, the strong user performance converges to that of the ZF inter-cluster beamforming case, as expected. It is interesting to note that the proposed method combining ZF inter-cluster beamforming and two-user Pareto-optimal intra-cluster beamforming outperforms the considered MMSE approach in terms of the weak user rate at high SNR. This gain results from the fact that for $N_t = 8$ the SUS algorithm generates clusters less than N_t (for the reason of this, please see Appendix C of [29]), situation like Fig. 3(b) occurs, and thus the proposed method combining ZF inter-cluster beamforming and two-user Pareto-optimal intra-cluster beamforming exploits the 2D design freedom to construct the intra-cluster beams.

VIII. Conclusion

In this paper, we have considered the problem of transmit beam design and user scheduling for MU-MISO NOMA

downlink and proposed an efficient beam design and user scheduling method based on a hierarchical beam structure exploiting both the spatial and power domains available in MU-MISO NOMA downlink. The proposed method has the ability of rate control between strong and weak users and hence provides great flexibility to NOMA network operation.

APPENDIX

Proof of Proposition 1: The set $\mathcal{P}_i$ to which p_1^{opt} belongs is dependent on the relationship among $a(p_1), b(p_1)$ and $d(p_1) := b(p_1) + c^2(p_1)/b(p_1)$, given in terms of Γ, θ and λ_i by

$$a(p_1) = \sqrt{P-p_1}\sqrt{\frac{\|\mathbf{h}_1\|^2}{\sigma_1^2(1+\gamma_1^*)}} = \sqrt{(P-p_1)\frac{\lambda_1}{1+\Gamma\lambda_1}} \quad (59)$$

$$b(p_1) = \sqrt{P-p_1}\sqrt{\frac{\|\mathbf{h}_2\|^2\theta}{\|\mathbf{h}_2\|^2 p_1[\alpha_1^*(p_1)]^2+\sigma_2^2}} = \sqrt{(P-p_1)\frac{\lambda_2\theta}{\lambda_2 p_1[\alpha_1^*(p_1)]^2+1}} \quad (60)$$

$$d(p_1) = \sqrt{P-p_1}\sqrt{\frac{\|\mathbf{h}_2\|^2/\theta}{\|\mathbf{h}_2\|^2 p_1[\alpha_1^*(p_1)]^2+\sigma_2^2}} = \sqrt{(P-p_1)\frac{\lambda_2/\theta}{\lambda_2 p_1[\alpha_1^*(p_1)]^2+1}}. \quad (61)$$

The three sets $\mathcal{P}_1, \mathcal{P}_2$ and $\mathcal{P}_3$ can be rewritten by squaring $a(p_1)$, $b(p_1)$ and $d(p_1)$ and dropping the common factor $(P-p_1)$ as $\mathcal{P}_1 = \{p_1|\bar{a} \le \bar{b}(p_1)\}$, $\mathcal{P}_2 = \{p_1|\bar{b}(p_1) < \bar{a} \le \bar{d}(p_1)\}$, and $\mathcal{P}_3 = \{p_1|\bar{a} > \bar{d}(p_1)\}$, where

$$\bar{a} = \frac{\lambda_1}{1+\Gamma\lambda_1},\ \bar{b}(p_1) = \frac{\lambda_2\theta}{\lambda_2 p_1[\alpha_1^*(p_1)]^2+1}, \text{ and} \quad (62)$$

$$\bar{d}(p_1) = \frac{\lambda_2/\theta}{\lambda_2 p_1[\alpha_1^*(p_1)]^2+1}. \quad (63)$$

First, we show $p_1^{opt} \notin \mathcal{P}_1$. Let $p_{1,\min}$ denote the minimum p_1 to achieve γ_1^* with $\mathbf{w}_1 = \mathbf{h}_1/\|\mathbf{h}_1\|$. Then, $p_{1,\min} = \gamma_1^*/\lambda_1 = \Gamma$. Hence, the second condition in (25), i.e., $\Gamma > p_{1,\min}(1-\theta)$ or $\Gamma = p_{1,\min}(1-\theta)$ is satisfied since $0 \le \theta \le 1$, and $\alpha_1^*(p_{1,\min}) = \sqrt{\theta}$ from (25). Hence, we have

$$\bar{a} = \frac{\lambda_1}{1+\Gamma\lambda_1} = \frac{1}{\frac{1}{\lambda_1}+\Gamma} \quad (64)$$

$$\bar{b}(p_{1,\min}) = \frac{\lambda_2\theta}{\lambda_2\Gamma\theta+1} = \frac{1}{\frac{1}{\theta\lambda_2}+\Gamma}. \quad (65)$$

By the NOMA condition $\lambda_1 > \lambda_2$, we have $\frac{1}{\lambda_1} < \frac{1}{\theta\lambda_2}$ since $0 \le \theta \le 1$ and thus $\bar{a} > \bar{b}(p_{1,\min})$. In case of $\Gamma = p_{1,\min}(1-\theta)$, we have $\theta = 0$ and thus $\bar{b}(p_{1,\min}) = 0$ and $\bar{a} > \bar{b}(p_{1,\min})$. Hence, $p_{1,\min} \notin \mathcal{P}_1$. Note that $\bar{a}$ is constant over p_1. It can be shown from (25) that the term $p_1[\alpha_1^*(p_1)]^2$ in the denominator of $\bar{b}(p_1)$ in (62) is monotone decreasing with respect to p_1, and hence $\bar{b}(p_1)$ is monotone increasing with respect to p_1. If $\bar{a} > \bar{b}(p_1)$ for all p_1, $\mathcal{P}_1$ is empty. Otherwise, there exists p_1, denoted as $p_{1,a}$, such that $\bar{a} = \bar{b}(p_1)$, as p_1 increases, given by

$$p_{1,a} = \{p_1|\bar{a} = \bar{b}(p_1)\} = \left\{p_1|\frac{\lambda_1}{1+\Gamma\lambda_1} = \frac{\lambda_2\theta}{\lambda_2 p_1[\alpha_1^*(p_1)]^2+1}\right\} = \Gamma + \frac{1}{1-\theta}\left(\sqrt{\theta\Gamma} - \sqrt{\theta\Gamma + \lambda_1^{-1}\theta - \lambda_2^{-1}}\right)^2. \quad (66)$$

At $p_1 = p_{1,a}$, we have $\gamma_2^{*(2)} = \gamma_2^{*(1)}$ from (30) since $\alpha_2^*(p_1) = 1$ at the boundary of $\mathcal{P}_1$ ($p_1 \ge p_{1,a}$ side) and $\mathcal{P}_2$ ($p_1 < p_{1,a}$ side), i.e., $\bar{a} = \bar{b}(p_1)$. Furthermore, it can be shown that $\frac{\partial\gamma_2^{*(2)}(p_1)}{\partial p_1}|_{p_1 \to p_{1,a}^-} = \frac{-1}{c_1}$, where c_1 is a non-negative constant with respect to p_1. Hence, there exists $p_1 \in \mathcal{P}_2$ such that $\gamma_2^{*(2)}(p_1) > \gamma_2^{*(2)}(p_{1,a}) = \gamma_2^{*(1)}(p_{1,a})$. Since $\gamma_2^{*(1)}$ is a monotone decreasing function of p_1 as seen in (30), optimal γ_2^* does not occur in $\mathcal{P}_1$, i.e., $p_1^{opt} \notin \mathcal{P}_1$.

Next, we check the condition that $\mathcal{P}_3$ is empty. Since the term $p_1[\alpha_1^*(p_1)]^2$ in the denominator of $\bar{d}(p_1)$ in (63) is monotone decreasing with respect to p_1, and thus $\bar{d}(p_1)$ is monotone increasing with respect to p_1. Therefore, if $\bar{a} < \bar{d}(p_{1,\min})$, then $\mathcal{P}_3$ is empty. Since $\alpha_1^*(p_{1,\min}) = \sqrt{\theta}$ from (25) and $p_{1,\min} = \Gamma$, the condition is rewritten from (62) and (63) as

$$\bar{a} < \bar{d}(p_{1,\min}) \Leftrightarrow \frac{\lambda_1}{1+\Gamma\lambda_1} < \frac{\lambda_2/\theta}{\lambda_2\Gamma\theta+1} \quad (67)$$

$$\Leftrightarrow \theta\Gamma < \frac{1}{\theta}\left(\frac{1}{\lambda_1}+\Gamma\right) - \frac{1}{\lambda_2} =: \tau. \quad (68)$$

In this case, $\mathcal{P}_3 = \emptyset$ and $p_1^{opt} \in \mathcal{P}_2$ since $p_1^{opt} \notin \mathcal{P}_1$.

Now assume $\theta\Gamma \ge \tau$. Then, $\mathcal{P}_3$ is not empty. Furthermore, we have a sufficient condition for $\forall\, p_1 \in \mathcal{P}_3$ as follows:

$$\bar{d}(p_1) < \bar{a}, \quad \forall p_1 \Leftarrow \lambda_2/\theta < \frac{\lambda_1}{1+\Gamma\lambda_1} \quad (69)$$

$$\Leftrightarrow \frac{1}{\theta}\left(\frac{1}{\alpha_1}+\Gamma\right) - \frac{1}{\lambda_2} < 0 \quad (70)$$

$$\Leftrightarrow \tau < 0, \quad (71)$$

because λ_2/θ is an upper bound of $\bar{d}(p_1)$ (see (63)). In this case, $p_1^{opt} \in \mathcal{P}_3$.

Finally, if $\theta\Gamma \ge \tau$ and $\tau \ge 0$, compute p_1, denoted by $p_{1,b}$, such that $\bar{a} = \bar{d}(p_1)$, given by

$$p_{1,b} = \{p_1|\bar{a} = \bar{d}(p_1)\} \quad (72)$$

$$= \left\{p_1|\frac{\lambda_1}{1+\Gamma\lambda_1} = \frac{\lambda_2/\theta}{\lambda_2 p_1[\alpha_1^*(p_1)]^2+1}\right\} \quad (73)$$

$$= \Gamma + \frac{1}{1-\theta}\left(\sqrt{\theta\Gamma} - \sqrt{\tau}\right)^2. \quad (74)$$

If

$$P < p_{1,b}, \quad (75)$$

then $\bar{a} > \bar{d}(p_1)$ for $p_1 \le P$ since $\bar{d}(p_1)$ is a monotone increasing function of p_1. Hence, in this case, $\forall p_1 \in \mathcal{P}_3$ and $p_1^{opt} \in \mathcal{P}_3$. On the other hand, if $p_{1,b} \le P$, we have both nonempty $\mathcal{P}_2 = \{p_1 \ge p_{1,b}\}$ and $\mathcal{P}_3 = \{p_1 < p_{1,b}\}$. In this case, we compute

the derivatives of $\gamma_2^{*(2)}$ and $\gamma_2^{*(3)}$ at point $p_{1,b}$, which are given by

$$\left.\frac{\partial \gamma_2^{*(2)}}{\partial p_1}\right|_{p_1=p_{1,b}^+} = c_2\left[\left(\lambda_2\sqrt{\tau}\frac{1-\theta}{\sqrt{\theta\Gamma}-\sqrt{\tau}}\right)P - \left(\lambda_2\sqrt{\tau}\frac{1-\theta}{\sqrt{\theta\Gamma}-\sqrt{\tau}}\cdot\Gamma + \lambda_2\sqrt{\tau}\cdot\sqrt{\theta\Gamma}+1\right)\right]$$

$$\left.\frac{\partial \gamma_2^{*(3)}}{\partial p_1}\right|_{p_1=p_{1,b}^-} = c_3\left[\left(\lambda_2\sqrt{\tau}\frac{1-\theta}{\sqrt{\theta\Gamma}-\sqrt{\tau}}\right)P - \left(\lambda_2\sqrt{\tau}\frac{1-\theta}{\sqrt{\theta\Gamma}-\sqrt{\tau}}\cdot\Gamma + \lambda_2\sqrt{\tau}\cdot\sqrt{\theta\Gamma}+1\right)\right],$$

where c_2 and c_3 are non-negative constants. The two derivatives have the same sign. If the two derivatives are positive, then γ_2^* increases as p_1 crosses $p_{1,b}$ from the left to the right and hence $p_1^{opt} \in \mathcal{P}_2$. Otherwise, γ_2^* increases as p_1 crosses $p_{1,b}$ from the right to the left and hence $p_1^{opt} \in \mathcal{P}_3$. Equivalently, we have

$$p_1^{opt} \in \mathcal{P}_2 \text{ if } P \geq \Gamma + \frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})\left(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}}\right)$$
$$p_1^{opt} \in \mathcal{P}_3 \text{ if } P < \Gamma + \frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})\left(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}}\right). \tag{76}$$

Since $p_{1,b} = \Gamma + \frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})^2 < \Gamma + \frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}})$, the set $\{P < p_{1,b}\}$ mentioned in (75) is a subset of the set $\{P < \Gamma + \frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}})\}$. Thus, the case of $P < p_{1,b}$ is covered by (76). The only two cases for $p_1^{opt} \in \mathcal{P}_2$ are $[\theta\Gamma < \tau]$ or $[\theta\Gamma \geq \tau \geq 0$ and $P \geq \Gamma + \frac{1}{1-\theta}(\sqrt{\theta\Gamma}-\sqrt{\tau})(\sqrt{\theta\Gamma}+\frac{1}{\lambda_2\sqrt{\tau}})]$. Hence, the claim follows. ■

To prove Proposition 2, we introduce the following lemma.

Lemma 1: Define

$$\mathcal{N}_1 := \{\theta \mid a \leq b(\theta)\} \tag{77}$$
$$\mathcal{N}_2 := \{\theta \mid b(\theta) < a \leq b(\theta) + c^2(\theta)/b(\theta)\} \tag{78}$$
$$\mathcal{N}_3 := \{\theta \mid a > b(\theta) + c^2(\theta)/b(\theta)\}, \tag{79}$$

where a, $b(\theta)$ and $c(\theta)$ are defined in (35)–(37) Then, for given λ_1, λ_2 and Γ, every θ in $\mathcal{N}_1$ achieves the same optimal $\gamma_2^*(\theta)$ in (32), if $\mathcal{N}_1$ is not empty.

Proof: Let us assume that $\mathcal{N}_1$ is not empty. For every $\theta \in \mathcal{N}_1$, $\gamma_2^*(\theta) = \gamma_2^{*(1)} == \frac{\lambda_1}{1+\lambda_1\Gamma}$. From the fact that $\gamma_2^{*(1)} = \frac{\lambda_1}{1+\lambda_1\Gamma}$ and $\gamma_2^{*(2)}(\theta) = \frac{\lambda_1}{1+\lambda_1\Gamma}[\alpha_2^*(\theta)]^2$, and $0 \leq \alpha_2^*(\theta) < 1$ for $\theta \in \mathcal{N}_2$, it is obvious that $\gamma_2^{*(2)}(\theta) < \gamma_2^{*(1)}$ for all $\theta \in \mathcal{N}_2$. Hence, $\gamma_2^*(\theta)$ for $\theta \in \mathcal{N}_2$ is less than $\gamma_2^*(\theta)$ for $\theta \in \mathcal{N}_1$. Furthermore, for any $\theta \in \mathcal{N}_3$, we have $\gamma_2^{*(3)}(\theta) \overset{(a)}{=} \frac{\lambda_2}{\lambda_2[\alpha_1^*(\theta)]^2+1} \overset{(b)}{\leq} \frac{\lambda_2/\theta}{\lambda_2[\alpha_1^*(\theta)]^2+1} \overset{(c)}{=} [b(\theta)+c^2(\theta)/b(\theta)]^2 \overset{(d)}{<} a^2 \overset{(e)}{=} \gamma_2^{*(1)}$. Here, step (a) is by (32), step (b) holds because $\theta \in [0,1]$, step (c) is by direction computation based on $b(\theta)$ and $c(\theta)$, step (d) holds because $\theta \in \mathcal{N}_3$, and step (e) is by (32). Consequently, we have the claim. ■

Proof of Proposition 2: The necessary and sufficient condition for $\mathcal{N}_1$ defined in (77) being non-empty is given by

$$a^2 \leq \max_{0\leq\theta\leq 1} b^2(\theta), \tag{80}$$

where

$$a^2 = \frac{\lambda_1}{1+\Gamma\lambda_1},\ b^2(\theta) = \frac{\lambda_2\theta}{\lambda_2[\alpha_1^*(\theta)]^2+1},\ c^2(\theta) = \frac{\lambda_2(1-\theta)}{\lambda_2[\alpha_1^*(\theta)]^2+1}. \tag{81}$$

Since $b(\theta)$ is maximized at $\theta = \frac{[\lambda_2(1-\Gamma)+1]^2}{\lambda_2^2\Gamma(1-\Gamma)+[\lambda_2(1-\Gamma)+1]^2}$ and the corresponding maximum value is $\max_{0\leq\theta\leq1} b^2(\theta) = \lambda_2(1+\frac{\lambda_2\Gamma}{\lambda_2+1})$, the condition (80) becomes

$$\frac{\lambda_1}{1+\lambda_1\Gamma} \leq \lambda_2\left(1+\frac{\lambda_2\Gamma}{\lambda_2+1}\right) \Longleftrightarrow \Gamma_1 \leq \Gamma \leq \Gamma_2,$$

where

$$\Gamma_1 := \frac{1}{2}\left(1+\lambda_2^{-1}-\lambda_1^{-1}\right) - \frac{1}{2}\sqrt{(1+\lambda_1^{-1}+\lambda_2^{-1})^2 - 4\lambda_2^{-1}(1+\lambda_2^{-1})}$$

and

$$\Gamma_2 := \frac{1}{2}\left(1+\lambda_2^{-1}-\lambda_1^{-1}\right) + \frac{1}{2}\sqrt{(1+\lambda_1^{-1}+\lambda_2^{-1})^2 - 4\lambda_2^{-1}(1+\lambda_2^{-1})}.$$

In the case of non-empty $\mathcal{N}_1$, by substituting $\alpha_1^*(\theta)$ in (33) or (85) into $b^2(\theta)$, $b^2(\theta)$ is given by

$$b^2(\theta) = \begin{cases} \lambda_2\theta & \text{if } \theta \leq 1-\Gamma \\ \frac{\lambda_2\theta}{\lambda_2[\sqrt{\theta\Gamma}-\sqrt{(1-\theta)(1-\Gamma)}]^2+1} & \text{if } \theta > 1-\Gamma \end{cases} \tag{82}$$

When $\theta \leq 1-\Gamma$, $b^2(\theta)$ is linear and it can be shown that $b^2(\theta)$ is quasi-concave function when $\theta > 1-\Gamma$[§]. If $a^2 > \lambda_2(1-\Gamma)$, there doesn't exist θ satisfying $a^2 \leq \lambda_2\theta$ (for $\theta \leq 1-\Gamma$) and

[§]When $\theta > 1-\Gamma$, $b^2(\theta)$ can be written as $f^2(\theta)/g(\theta)$, where $f(\theta) = \sqrt{\lambda_2\theta}$ and $g(\theta) = \lambda_2[\sqrt{\theta\Gamma} - \sqrt{(1-\theta)(1-\Gamma)}]^2 + 1$. Since $f(\theta)$ is the concave function and $g(\theta)$ is the convex function (it can be proved easily by taking secondary derivative), we can conclude that $f^2(\theta)/g(\theta)$ is quasi concave [30].

$$\left\{\theta \mid \frac{z_1z_2+2\Gamma(1-\Gamma)-\sqrt{4\Gamma(1-\Gamma)[\Gamma(1-\Gamma)+z_1z_2-z_2^2]}}{z_1^2+4\Gamma(1-\Gamma)} \leq \theta \leq \frac{z_1z_2+2\Gamma(1-\Gamma)+\sqrt{4\Gamma(1-\Gamma)[\Gamma(1-\Gamma)+z_1z_2-z_2^2]}}{z_1^2+4\Gamma(1-\Gamma)}\right\}. \tag{83}$$

hence the set $\mathcal{N}_1 = \{\theta \mid a^2 \le b^2(\theta)\}$ is given by (83), shown at the bottom of the previous page.

Otherwise, the minimum of θ satisfying $a^2 \le b^2(\theta)$ is the point such that $a^2 = \lambda_2\theta$ and $\mathcal{N}_1$ becomes

$$\left\{\theta \,|\, \frac{\lambda_1}{\lambda_2}\frac{1}{1+\lambda_1\Gamma} \le \theta \le \frac{z_1 z_2 + 2\Gamma(1-\Gamma) + \sqrt{4\Gamma(1-\Gamma)[\Gamma(1-\Gamma)+z_1 z_2 - z_2^2]}}{z_1^2 + 4\Gamma(1-\Gamma)}\right\},$$

where $z_1 = \lambda_1^{-1} + 1 - \Gamma$ and $z_2 = \lambda_2^{-1} + 1 - \Gamma$. ((83) is obtained by solving $\frac{\lambda_1}{1+\Gamma\lambda_1} \le \frac{\lambda_2\theta}{\lambda_2[\sqrt{\theta\Gamma} - \sqrt{(1-\theta)(1-\Gamma)}]^2+1}$ reducing to a quadratic inequality). In the case of non-empty $\mathcal{N}_1$, by Lemma 1, $\mathcal{N}_1$ is optimal and we obtain (39).

Next, consider the case that $\mathcal{N}_1$ is empty. At $\theta = 1$, we have $a(1) = \sqrt{\frac{1}{\Gamma+\frac{1}{\lambda_1}}} > \sqrt{\frac{1}{\Gamma+\frac{1}{\lambda_2}}} = b(1) + c^2(1)/b(1)$ by the NOMA assumption $\lambda_1 > \lambda_2$ and hence $\theta = 1 \in \mathcal{N}_3$ by the definition of $\mathcal{N}_3$ in (79). We also have

$$\lim_{\theta\to 0} b(\theta) + c^2(\theta)/b(\theta) = \lim_{\theta\to 0}\sqrt{\frac{\lambda_2\theta^{-1}}{\lambda_2[\alpha_1^*(\theta)]^2+1}} = \infty, \quad (84)$$

which can easily be seen from $\alpha_1^*(0) = \sqrt{1-\Gamma}$. Thus, $\theta = 0 \in \mathcal{N}_2$. Furthermore, $b(\theta) + c^2(\theta)/b(\theta) = \sqrt{\frac{\lambda_2\theta^{-1}}{\lambda_2[\alpha_1^*(\theta)]^2+1}}$ is a monotone decreasing function of θ since

$$\alpha_1^*(\theta) = \begin{cases} 0 & \text{if} \quad \theta \le \theta_I \triangleq 1-\Gamma \\ \sqrt{\theta\Gamma} - \sqrt{(1-\theta)(1-\Gamma)} & \text{if} \quad \theta > \theta_I \end{cases} \quad (85)$$

is a monotone increasing function of θ. Hence, there exists θ_a such that $a(\theta_a) = b(\theta_a) + c^2(\theta_a)/b(\theta_a)$ to yield $\mathcal{N}_2 = \{\theta|\theta \le \theta_a\}$ and $\mathcal{N}_3 = \{\theta|\theta > \theta_a\}$.

Now recall that

$$\gamma_2^{*(1)} = \frac{\lambda_1}{1+\lambda_1\Gamma},\ \gamma_2^{*(2)}(\theta) = \frac{\lambda_1}{1+\lambda_1\Gamma}[\alpha_2^*(\theta)]^2,$$
$$\text{and}\ \gamma_2^{*(3)}(\theta) = \frac{\lambda_2}{\lambda_2[\alpha_1^*(\theta)]^2+1}. \quad (86)$$

If $\theta_a \le \theta_I$, then the optimal θ set for maximizing γ_2^* is given by $\{\theta|\theta_a \le \theta \le \theta_I\}$. This is because $\gamma_2^{*(2)}(\theta_a) = \gamma_2^{*(3)}(\theta_a)$, because $\gamma_2^{*(3)}(\theta)$ is monotone decreasing with respect to θ as seen in (86) since $\alpha_1^*(\theta)$ is a monotone increasing function of θ, and because $\gamma_2^{*(2)}(\theta)$ is monotone increasing with respect to θ for $\theta \le \theta_a$ since $\alpha_2^*(\theta)$ is a monotone increasing function of θ for $\theta \le \min\{\theta_a, \theta_I\} = \theta_a$ (this can be shown by substituting $\alpha_1^* = 0$ for $\theta \le \theta_I$ into $\alpha_2^*(\theta)$ and taking derivative of $\alpha_2^*(\theta)$ with respect to θ and showing the derivative is positive for $\theta \le \theta_a$). Hence, in this case the optimal γ_2^* occurs at θ_a but for all θ in $\{\theta|\theta_a \le \theta \le \theta_I\}$, $\alpha_1^*(\theta) = 0$ and the corresponding optimal $\gamma_2^* = \gamma_2^{*(3)} = \lambda_2$ from (86). In this case, from the assumption $\theta_a \le \theta_I$, θ_a is computed based on (81) with $\alpha_1^*(\theta) = 0$ as

$$\theta_a = \theta \text{ s.t. } a = b(\theta) + c^2(\theta)/b(\theta) \quad (87)$$
$$= \theta \text{ s.t. } \frac{\lambda_1}{1+\lambda_1\Gamma} = \frac{\lambda_2}{\theta} \quad (88)$$
$$= \frac{\lambda_2}{\lambda_1}(1+\lambda_1\Gamma) \quad (89)$$

and the condition $\theta_a \le \theta_I$ reduces to

$$\frac{\lambda_2}{\lambda_1}(1+\lambda_1\Gamma) \le 1-\Gamma \iff \Gamma \le \frac{{\lambda_2}^{-1} - {\lambda_1}^{-1}}{1+{\lambda_2}^{-1}}. \quad (90)$$

On the other hand, if $\theta_a > \theta_I$, i.e., $\Gamma > \frac{{\lambda_2}^{-1}-{\lambda_1}^{-1}}{1+{\lambda_2}^{-1}}$, then optimal θ exists between θ_I and θ_a because $\gamma_2^{*(2)}$ is an increasing function for $\theta < \min\{\theta_a, \theta_I\} = \theta_I$ and $\gamma_3^{*(3)}$ is a decreasing function for $\theta > \theta_a$. Since optimal θ lies in $\mathcal{N}_2$ in this case, it is obtained by solving

$$\frac{\partial\gamma_2^{*(2)}(\theta)}{\partial\theta} = 0. \quad (91)$$

Therefore, the claim follows. ■

Proof of Corollary 1: With the assumption of $\lambda_1 = \lambda_2$, we have $\Gamma_1 = 0$ and $\Gamma_2 = 1$ from (38) and hence the condition $\Gamma \in [\Gamma_1, \Gamma_2]$ reduces to $\Gamma \in [0,1]$ which is always valid for $p_1 = p_2 = 1$ (see the definition of Γ in (23)). Furthermore, with the assumption, we have $z_1 = z_2$ and $\sqrt{4\Gamma(1-\Gamma)[\Gamma(1-\Gamma)+z_1z_2-z_2^2]}$ in (39) is given by $2\Gamma(1-\Gamma)$. From (39), the optimal θ set is given by $\{\theta \,|\, \theta_0 \le \theta \le 1\}$, where

$$\theta_0 = \begin{cases} \frac{1}{1+\Gamma\lambda_1} & \text{if} \ \ \frac{1}{1+\Gamma\lambda_1} \le 1-\Gamma \\ \frac{z_1^2}{z_1^2+4\Gamma(1-\Gamma)} & \text{if} \ \ \frac{1}{1+\Gamma\lambda_1} > 1-\Gamma \end{cases} \quad (92)$$

■

Proof of Proposition 3: As $\lambda_2 \to 0$. the threshold τ in Proposition 1 converges to $-\infty$ and thus neither of the two conditions for $p_1^{opt} \in \mathcal{P}_2$ in Proposition 1 is satisfied. Hence, by Proposition 1, $p_1^{opt} \in \mathcal{P}_3$. Since $p_1^{opt} \in \mathcal{P}_3$, p_1^{opt} can be obtained in closed form by maximizing $\gamma_2^{*(3)}(p_1)$ in (30) and is given by

$$\bar{p}_1^{opt} = -P + 2\Gamma + \frac{\psi_1^2 - \psi_1\sqrt{\psi_2^2 + 2\lambda_2^{-1}\psi_1 + \lambda_2^{-2}}}{2\theta(1-\theta)\Gamma} \quad (93)$$

where $\psi_1 := \theta\Gamma + (1-\theta)(P-\Gamma) + \lambda_2^{-1}$ and $\psi_2 := \theta\Gamma - (1-\theta)(P-\Gamma)$. Using L'Hospital's rule, we can show that $\lim_{\lambda_2\to 0}\bar{p}_1^{opt} = \Gamma(= p_{1,\min})$. With $p_1 = \Gamma$, we have $\alpha_1^* = \sqrt{\theta}$ from (25) and consequently $\beta_1^* = \sqrt{1-\theta}$ from the constraint eq. in (24), and $\alpha_2^* = \sqrt{\theta}$ from (26), and by substituting these values into $\gamma_2^{*(3)}(p_1)$ in (30) and (14) and (15), we have $\gamma_2^* = \frac{P-\Gamma}{\Gamma}\frac{1}{\theta+\lambda_2^{-1}\Gamma^{-1}}$, $\sqrt{p_1}\mathbf{w}_1 = \sqrt{\Gamma}\frac{\mathbf{h}_1}{\|\mathbf{h}_1\|}$ and $\sqrt{p_2}\mathbf{w}_2 = \sqrt{P-\Gamma}\frac{\mathbf{h}_2}{\|\mathbf{h}_2\|}$. ■

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 Veh. Technol. Conf.*, Jun. 2013, pp. 1–5.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[3] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[4] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[5] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Aug. 2015, pp. 1127–1131.

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

[7] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. IEEE Int. Symp. Wireless Commun. Syst.*, 2014, pp. 781–785.

[8] J. So and Y. Sung, "Improving non-orthogonal multiple access by forming relaying broadcast channels," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1816–1819, Sep. 2016.

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

[10] X. Chen, A. Benjebbour, Y. Lan, A. Li, and H. Jiang, "Impact of rank optimization on downlink non-orthogonal multiple access (NOMA) with SU-MIMO," in *Proc. IEEE Int. Conf. Commun. Syst.*, 2014, pp. 233–237.

[11] B. Kim *et al.*"Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. Mil. Commun. Conf.*, 2013, pp. 1278–1283.

[12] S. Liu, C. Zhang, and G. Lyu, "User selection and power schedule for downlink non-orthogonal multiple access (NOMA) system," in *Proc. IEEE Int. Conf. Commun. Workshop*, 2015, pp. 2561–2565.

[13] A. Sayed-Ahmed and M. Elsabrouty, "User selection and power allocation for guaranteed SIC detection in downlink beamforming non-orthogonal multiple access," in *Proc. IEEE Wireless Days*, 2017, pp. 188–193.

[14] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, Dec. 2016.

[15] 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, Mar. 2006.

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

[17] 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.

[18] J. Seo and Y. Sung, "A new transceiver architecture for multi-user MIMO communication based on mixture of linear and non-linear reception," in *Proc. Int. Workshop Signal Process. Adv. Wireless Commun.*, Sapporo, Japan, Jun. 2017, pp. 679–683.

[19] 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.

[20] K. Ho, D. Gesbert, E. Jorswieck, and R. Mochaourab, "Beamforming on the MISO interference channel with multi-user decoding capability," arXiv:1107.0416, 2011.

[21] E. A. Jorswieck and E. G. Larsson, "Linear precoding in multiple antenna broadcast channels: Efficient computation of the achievable rate region," in *Proc. Int. ITG Workshop Smart Antennas*, 2008, pp. 21–28.

[22] E. A. Jorswieck, E. G. Larsson, and D. Danev, "Complete characterization of the Pareto boundary for the MISO interference channel," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5292–5296, Oct. 2008.

[23] J. Lindblom, E. Karipidis, and E. G. Larsson, "Closed-form parameterization of the Pareto boundary for the two-user MISO interference channel," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 3372–3375.

[24] J. Seo and Y. Sung, "Beam design and user scheduling for non-orthogonal multiple access with multiple antennas based on Pareto-optimality," arXiv:1705.06959, 2017.

[25] R. Horst and H. Tuy, *Global Optimization: Deterministic Approaches*. New York, NY, USA: Springer, 2013.

[26] Y. Xu, T. Le-Ngoc, and S. Panigrahi, "Global concave minimization for optimal spectrum balancing in multi-user DSL networks," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2875–2885, Jul. 2008.

[27] R. Mochaourab and E. A. Jorswieck, "Optimal beamforming in interference networks with perfect local channel information," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1128–1141, Mar. 2011.

[28] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.

[29] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in mm-wave multi-user MISO downlink," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, Feb. 2016.

[30] C. Bector, "Programming problems with convex fractional functions," *Oper. Res.*, vol. 16, pp. 383–391, Apr. 1968.

Junyeong Seo (S'11) received the B.S. and M.S. degrees in electrical engineering, in 2011 and 2013, respectively, from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, where he is currently working toward the Ph.D. degree at the Wireless Information Systems Research Group. His research interests include design and analysis of large-scale MIMO systems, new multiple access methods, and signal processing for next wireless communication.

Youngchul Sung (S'92–M'93–SM'09) received B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1993 and 1995, respectively, and the Ph.D. degree in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2005. From 1995 to 2000, he was with the LG Electronics Ltd., Seoul, South Korea. From 2005 to 2007, he was a Senior Engineer with the Corporate R&D Center of Qualcomm, Inc., San Diego, CA, USA, and participated in design of Qualcomm's 3GPP R6 WCDMA basestation modem. Since 2007, he has been on the faculty with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. His research interests include signal processing for communications, statistical signal processing and inference, and asymptotic statistics with applications to wireless communications and related areas.

Dr. Sung is currently a Chair of the IEEE Communications Society Asia-Pacific Board and Chapters Coordination Committee. He is in the Signal Processing and Communications Electronics Technical Committee of IEEE Communications Society, and the Signal and Information Processing Theory and Methods Technical Committee of Asia-Pacific Signal and Information Processing Association. He is currently an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2012 to 2014 and a Guest Editor for the 2012 IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS special issue "Theories and Methods for Advanced Wireless Relays."