IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 13, 2018, accepted December 19, 2018, date of publication January 3, 2019, date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2890441

Power Control for Sum Spectral Efficiency Optimization in MIMO-NOMA Systems With Linear Beamforming

SEONGGYOON PARK[1], ANH QUAN TRUONG[2], AND TIEN HOA NGUYEN[2]
[1]Radio Engineering Department, Information and Communication School, Kongju National University, Cheonan 31080, South Korea
[2]School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi 100000, Vietnam

Corresponding author: Tien Hoa Nguyen (hoa.nguyentien@hust.edu.vn)

ABSTRACT This paper considers a multiple-input multiple-output non-orthogonal multiple access (NOMA) downlink transmission system with different linear beamforming techniques, where the base station uses each cluster to serve a pair of users. In the considered NOMA cluster, we first derive the performance analysis of the system that uses a proposed user paring method, which exploits the different large-scale channel qualities of users to allocate the transmit power of the strong and weak users in each pair, to ensure that both users in each pair can contribute the best on the system performance. We further formulate a sum spectral efficiency (SE) maximization with a subject to the limited transmit power budget, which is emphasized to be non-convex. We have, then, proposed a framework that solves the above non-convex problem into two steps: lower bound this non-convex problem by a geometric program by using the arithmetic mean-geometric mean inequality and, then, employ the successive optimization approach to find the local Karush–Kuhn–Tucker point. Numerical results manifest that a NOMA-based network with zero forcing (ZF) beamforming gives the highest sum SE, while regularized ZF brings benefits to the SE of weak users.

INDEX TERMS NOMA, OMA, spectral efficiency, optimization.

I. INTRODUCTION

Multiple access (MA) schemes allow multiple users to share a channel over time, frequency or space efficiently [1]. The current access methods can broadly be divided into two categories. One of them is orthogonal frequency-division multiplexing (OFDM) that is being standardized for the most of current communication systems [2], [3]. The other one is non-orthogonal multiple access (NOMA), which is especially being oriented as the core technology of the fifth generation (5G) and possibly Next-G systems as well [4], [5]. For a system that uses the OFDM technique, one of the main drawbacks is low spectral efficiency when radio resources is still assigned to users with poor channel conditions [6]–[8]. Meanwhile, NOMA is expected to increase the system throughput, low latency, high spectral efficiency, and supply massive connectivity [9]. The great idea behind this approach comes from the fact that NOMA utilizes spectrum more efficiently than OMA by making use of the different channel conditions of each user and is capable of serving multiple users with different quality of service (QoS) requirements in the same time-, frequency-, and space-domain slot [10].

In state-of-the-art researches, the exist NOMA solutions can essentially be classified into two considerable approaches, which are power- and code-domain [11]–[13]. Developed and extended from conventional code division multiple access scheme, coding domain NOMA uses user-specific spreading sequences for sharing the entire available resource in the frequency domain as well as in the time domain [14]. Whereas, the power domain NOMA exploits channel differences between users for multiplexing through resource allocation. In this paper, we focus on the power domain NOMA. In this approach, the signals corresponding to different users are overlapped in the same frequency band. They are then decoded at the receiver through successive interference cancellation (SIC) [15], [16].

MIMO or Multiple Input Multiple Output has been investigated and the applied extensively during the last a few decades. Although the millimeter-wave (mmWave) massive MIMO has been considered as the new standard that will be applied to 5G. However, the latest standard version of the 3GPP 38 series is expected to become the stand alone version by 2020 considering only 5G systems as MIMO with

up to 16 base station (BS) antennas. However, the latest 3GPP standard of the 38 series, scheduled to become the stand alone version in 2020, only considers the 5G system as MIMO with up to 16 BS antennas [17], [18]. Therefore, MIMO will still be the core technology of 5G in the next two years. Follow the conventional approach to enhance the spectral efficiency, NOMA has also been combined with MIMO as in the studies [19]–[21] and extended to multi-cluster to archive higher network capacity and spectral efficiency [21], [22]. In such MIMO-NOMA systems, all users are often paired into clusters with two users [23] or multi-user [22], [24] in each cluster. To reduce the interference between users as well as clusters in a cell, different transmit NOMA beamforming techniques have been proposed to improve the MIMO-NOMA system performance [25], [26].

Energy efficiency is an important research topic in 5G and the future wireless networks. The optimal beamforming design for the sum spectral efficiency maximization problem with subject to the power constraints has been investigated in [28]. However, linear beamforming techniques are preferable in 5G because of their low computational complexity and because they can achieve the near optimal performance when integrating with other advanced technologies, for instance Massive MIMO. Besides, we observe increasing research interest in the performance analysis of different fading channel models [29], [55]. Even though such channels may bring some particular properties of real propagation environment, capacity analysis is challenging to obtain in closed-form. Meanwhile, Rayleigh fading channels are less complicated in term of modelling and matching well with non-line-of-sight (NLOS) conditions [45].

A. RELATED WORK

In this section, related works on NOMA beamforming and user paring techniques regarding the downlink MIMO scenarios will be discussed. In most of the current studies, a base station has been considered with multiple antennas [30] and each user is normally equipped with single antenna [31]. The authors of aforementioned studies have demonstrated that NOMA significantly improves sum spectral efficiency than conventional orthogonal multiple access (OMA) counterpart. These results continue to be asserted that this advantage still exists for a two-user [24] and multi-user on each cluster MIMO-NOMA system [22], [32]. For multiple users in a cluster, NOMA will lead to inter-clustering interference, especially very severe at the boundaries of mobile networks. Consequently, it leads to reduction in the QoS for both cluster-edge users and the user fairness [19], [20]. Regarding the sum-throughput optimization problem for intra-cluster power allocation, the authors in [33] have considered the cases with different number of users allocated in a cluster. The authors have also pointed out that two users are allocated to one cluster will achieve the highest sum-throughput performance with the users that has weak channel gains.

Beamforming techniques have been typically investigated and applied to avoid inter-user and inter-cluster interference in the downlink of NOMA-MIMO systems. Distinct approaches deploying linear beamforming to reduce interference or ensure the QoS for every users was investigated in [21], [31], and [34]–[37]. For instance, Nguyen *et al.* [37] have used zero-forcing (ZF) combining vectors to investigate the sum spectral efficiency in the joint power control and load balancing problem scenario. Choi [31] have proposed a two-stage beamforming method that uses ZF to avoid inter-cluster interference in the first stage, and then the optimal beamforming vectors were considered to minimize the total power transmission within cluster in the second stage. Wang *et al.* [20] considered a beamforming model for a MIMO-NOMA system, which allows different power allocations between clusters. This result is extended in [38] from single-cell to multi-cell, where a novel precode design has been investigated. The simulation results demonstrate that the beamforming method in MIMO-NOMA achieves higher spectral efficiency (SE) comparing to the conventional MIMO-OMA.

There have been a number of studies that conduct in the field of massive MIMO-NOMA or cognitive NOMA systems such as in [20] and [39]–[44]. However, within the framework of this paper, we discuss only the problems of user-paring and beamforming in these studies. The authors generally used the linear beamforming techniques such as ZF or maximum ratio transmission (MRT), since they have low computational complexity and even nearly optimal in Massive MIMO which is potentially used in future radio networks [45]. As reported in [46], the use of linear beamforming helps simplify the receiver. In an effort to use these linear beamforming techniques, the considered issues addressed in MIMO-NOMA systems can be decomposed into optimization problems. Nevertheless, the authors have not solved the issue of satisfying the QoS of weak users in the NOMA system.

B. MOTIVATIONS AND CONTRIBUTIONS

The works in [24] and [30]–[32] consider a fixed power allocation level for each cluster in the entire network. However, in practice, the power per cluster corresponding to the transmit power of an antenna in the MIMO-NOMA system should be considered to have a limited transmit power budget. The studies in [20] and [39]–[42] mainly compare the performance of the system using different linear beamforming methods. Nevertheless, in certain situations involving user-paring and ensuring the fairness between weak users and strong users in a pair of NOMA users, to our best knowledge, there is still no research has addressed this issue.

Motivated by the aforementioned analysis, our main contributions are summarized as follows:

- We evaluate the network spectral efficiency with different linear beamforming techniques comprising of maximum ration transmission (MRT), zero-forcing (ZF), and regularized zero-forcing (RZF). In particular, for comparison purposes, the normalized beamforming vectors are used to derive the achievable rate of each user in

the network coverage area. The array gain together with mutual interference and thermal noise is then clearly observed.

- We stress that our derived achievable rates are able to apply for any pairing methods. In general, the optimal solution to pair users is only obtained by exhaustive search as it is a combinatorial problem. Nonetheless, a heuristic pairing method which has low complexity is proposed with aiming at improving the sum spectral efficiency in the network.
- We also formulate a sum spectral efficiency maximization problem with the limited power budget constraints. The epi-graph form representation and successive optimization approach are then implemented to deal with the inherent non-convexity of this optimization problem. The computing progress can be easily performed with an arbitrary general purpose optimization toolbox to attain the KKT optimal results.
- Numerical results demonstrate many attractive observations of different beamforming techniques. Specifically, the performance of RZF and ZF are competitive and RZF can outperform the existed dominant ZF when increasing number of served users.

The rest of this paper is organized as the following: The considered system model is first depicted in Section II. The achievable rate for each user and a heuristic pairing method are further investigated. Section III analyzes the downlink achievable rate by using linear beamforming techniques. Meanwhile, the sum SE optimization problem is represented in Section IV, following by a solution to obtain its KKT local optimum by utilizing epi-graph form representation and successive optimization approach. Section V provides extensive numerical results to approve our theoretical analysis in the previous sections. Finally, we give some main conclusions in Section VI.

Notations: Vectors (matrices) are denoted by bold face small (big) letters. The superscripts T and H stand for the transpose and conjugate transpose. $\mathbf{I}_K$ is the $K \times K$ identity matrix. $\mathbb{E}\{\cdot\}$ is the expectation operator. The notation $\|.\|$ is used for the Euclidean norm. $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C})$ means $\mathbf{n}$ is a circularly symmetric complex Gaussian random vector with covariance matrix $\mathbf{C}$.

II. SYSTEM MODEL

In this paper, we consider a downlink multiuser MIMO single cell using NOMA. The system contains a BS equipped with N antennas for the purpose of being able to serve more than $2N$ single antenna users. The BS uses beamforming to reduce inter-cluster interference, and each vector beamforming can support a group of two users. For the sake of simplicity, we consider $2N$ single antenna users in the NOMA cell. [1]

[1]For conveniences in notation and take merits in constructing a mathematical framework, we only consider $2N$ single-antenna users which are divided into N groups, each comprising of two users. However, the extension to an arbitrary number of users with more than two users in one group is straightforward by using the mathematical framework constructed in this paper thanks to the flexibility of NOMA technology.

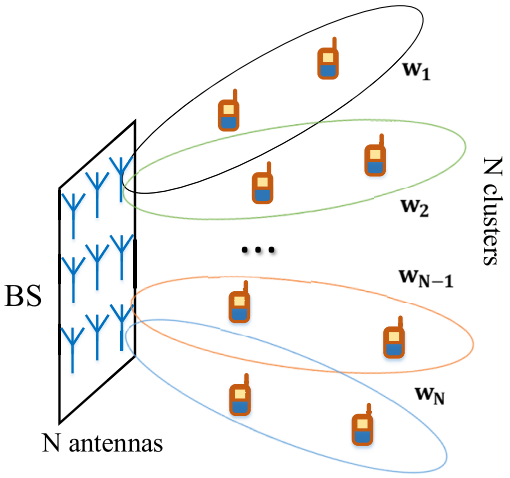


FIGURE 1. The considered NOMA system with $2N$ users divided in N clusters.

We further assume that these users are divided in N clusters as shown in Fig. 1. Even though channel varies in both time and frequency, we assume that the resource is spitted in coherence intervals where the channels are flat. Every cluster $n, n = 1, \ldots, N$ has two users and the channel of user k ,$k \in \{1, 2\}$, denotes as a column vector $\mathbf{h}_{n,k} \in \mathbb{C}^N$ comprising of path loss and shadow fading. We also assume that all users share the same time and frequency resource.

A. DOWNLINK TRANSMISSION

In the downlink data transmission phase, BS transmits a couple of data symbols includes $s_{n,1}, s_{n,2} \in \mathbb{C}$ to the two users in cluster n with $\mathbb{E}\{|s_{n,1}|^2\} = 1$ and $\mathbb{E}\{|s_{n,2}|^2\} = 1$. The transmitted signal for cluster n is then formulated as

$$\mathbf{x}_n = \mathbf{w}_n\sqrt{p_{n,1}}s_{n,1} + \mathbf{w}_n\sqrt{p_{n,2}}s_{n,2}, \tag{1}$$

where $\mathbf{w}_n \in \mathbb{C}^N$ denotes the beamforming vectors, which BS assigns to the two users in cluster n, respectively. $p_{n,1}$ and $p_{n,2}$ are the transmit power assigned to these users satisfying

$$0 \le p_{n,1} + p_{n,2} \le P_{\max,n}. \tag{2}$$

In the above equations, $P_{\max,n}$ is the maximum power levels, which BS can allocate to the users in cluster n. Therefore, the transmitted signal from BS to all users in the coverage area is

$$\mathbf{x} = \sum_{n'=1}^{N} \mathbf{x}_{n'} = \sum_{n'=1}^{N} \mathbf{w}_{n'}\left(\sqrt{p_{n',1}}s_{n',1} + \sqrt{p_{n',2}}s_{n',2}\right). \tag{3}$$

Without the loss of generality, in every cluster, user 1 is always assumed to be the weaker user and user 2 is the stronger one. In cluster n, $n = 1, \ldots, N$, the received signal at user 1 is expressed as

$$\begin{aligned} y_{n,1} &= \mathbf{h}_{n,1}^H\mathbf{x} + n_{n,1} \\ &\overset{(a)}{=} \mathbf{h}_{n,1}^H \sum_{n=1}^{N} \mathbf{w}_n\left(\sqrt{p_{n',1}}s_{n',1} + \sqrt{p_{n',2}}s_{n',2}\right) + n_{n,1} \end{aligned}$$

$$\overset{(b)}{=} \mathbf{h}_{n,1}^H \mathbf{w}_n \sqrt{p_{n,1}} s_{n,1} + \mathbf{h}_{n,1}^H \mathbf{w}_n \sqrt{p_{n,2}} s_{n,2} + \mathbf{h}_{n,1}^H \left(\sum_{n'=1, n' \neq n}^{N} \mathbf{w}_{n'} \sqrt{p_{n',1}} s_{n',1} + \mathbf{w}_{n'} \sqrt{p_{n',2}} s_{n',2} \right) + n_{n,1}. \quad (4)$$

In (4), $n_{n,1}$ is complex Gaussian noise and distributed as $n_{n,1} \sim \mathcal{CN}(0, \sigma_{\mathrm{DL}}^2)$, where σ_{DL}^2 is the noise variance. We obtain (a) by plugging (3) into the first equation of (4). The first term of (b) contains the desired signal, while the second term is intra-cluster interference. The remaining is inter-cluster interference and additive noise. Similarly, the received signal at user 2 in cluster n is given as

$$y_{n,2} = \mathbf{h}_{n,2}^H \mathbf{w}_n \sqrt{p_{n,2}} s_{n,2} + \mathbf{h}_{n,2}^H \mathbf{w}_n \sqrt{p_{n,1}} s_{n,1} + \mathbf{h}_{n,2}^H \left(\sum_{n'=1, n' \neq n}^{N} \mathbf{w}_{n'} \sqrt{p_{n',1}} s_{n',1} + \mathbf{w}_{n'} \sqrt{p_{n',2}} s_{n',2} \right) + n_{n,2}. \quad (5)$$

The first term of (5) denotes the desired signal, while the second term is intra-cluster interference from user 1. The last terms are inter-cluster interference and noise. We stress that both (4) and (5) can use an arbitrary beamforming technique. In the following subsection, we apply linear beamforming techniques, which has low computational complexity.

B. PERFORMANCE ANALYSIS USING NOMA

In cluster n, the weaker user, in other words, the user 1 is assumed to ignore successive interference cancellation, thus the achievable rate of this user is formulated as

$$R_{n,1} = \log_2 \left(1 + \mathrm{SINR}_{n,1}\right) \text{ [b/s/Hz]}, \quad (6)$$

where the signal-to-interference-and-noise ratio (SINR) value is

$$\mathrm{SINR}_{n,1} = \frac{\left|\mathbf{h}_{n,1}^H \mathbf{w}_n\right|^2 p_{n,1}}{\left|\mathbf{h}_{n,1}^H \mathbf{w}_n\right|^2 p_{n,2} + \sum\limits_{\substack{n'=1, \\ n' \neq n}}^{N} \left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 p_{n'} + \sigma_{\mathrm{DL}}^2}, \quad (7)$$

where $p_{n'}$ is denoted as sum power in cluster n'.

The first term in the denominator of (7) contains the mutual interference from user 1 inside cluster n as a consequence of using conventional decoding. In addition, the second part contains mutual interference from the other clusters, which is caused by ineffectiveness of a beamforming technique. The last term is inherent thermal noise.

For user 2, before decoding the desired signal, the successive interference cancellation is first deployed to remove intra-cluster interference from user 1.[2] Consequently, its achievable rate is computed as

$$R_{n,2} = \log_2 \left(1 + \mathrm{SINR}_{n,2}\right) \text{ [b/s/Hz]}, \quad (8)$$

where the SINR value is

$$\mathrm{SINR}_{n,2} = \frac{\left|\mathbf{h}_{n,2}^H \mathbf{w}_n\right|^2 p_{n,2}}{\sum\limits_{\substack{n'=1, \\ n' \neq n}}^{N} \left|\mathbf{h}_{n,2}^H \mathbf{w}_{n'}\right|^2 p_{n'} + \sigma_{\mathrm{DL}}^2}. \quad (9)$$

Unlikely to user 1, the intra-cluster interference is completely removed thanks to the perfect successive interference cancellation. Meanwhile, both (7) and (9) involve mutual interference users in the other clusters, and therefore an effective linear beamforming technique among all others should be testified as demonstrated in the next section.

C. USER CLUSTERING

In this subsection, we investigate a method to divide $2N$ users to N clusters. In detail, we calculate channel gains of each user which is defined by

$$g_k = \|\mathbf{h}_k\|_2, \quad \forall, k = 1, \ldots, 2N, \quad (10)$$

and arrange these gains in an increasing order as

$$g_1 \leq g_2 \leq \cdots \leq g_{2N-1} \leq g_{2N}. \quad (11)$$

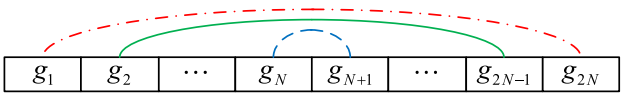


FIGURE 2. The clustering method used to divided $2N$ users in N clusters.

The clustering method is described in Fig. 2. This method pairs users with the criteria of maximum differences in channel gains. The user with the weakest channel gain and that one with the strongest channel gain (i.e., g_1 and g_{2N}) will form a cluster. Continually, the user with the second weakest channel gain pairs with the penultimate one and so on. Lastly, user N and user $(N+1)$ will be grouped in a pair. The main idea behind this clustering method is that in each cluster, the wide gain gap results in the less contamination from the weaker user to the stronger user. Consequently, it is expected that the total sum spectral efficiency in the entire network will be improved.

[2]In this paper, the framework is based on the perfect channel state information (CSI) which is aligned with many the previous works in the literature. This assumption is optimistic for NOMA to mitigate mutual interference efficiently and is reasonable if base station is equipped with the small number of antennas. We stress that the imperfect CSI is an important consideration in case of the massive number of base station antennas. This is because not only the channel estimation is a difficult task but also the system capacity may reduce rapidly when the orthogonal pilot signals are less than the number of users in the coverage area. NOMA systems under the imperfect CSI assumption are beyond the scope of this paper and we leave them for the future work.

III. DOWNLINK PERFORMANCE ANALYSIS WITH DIFFERENT LINEAR BEAMFORMING TECHNIQUES

In the scope of this paper, we assume that perfect channel state information is achieved and shared among BS and $2N$ users. For further analysis and comparison, we now evaluate the performance of every user in the network by utilizing the normalized versions of linear beamforing techniques [47]. This is the fundamental difference from the previous works on NOMA-based systems which focus on only one linear beamforming technique since the normalization allows us to observe and compare the performance among them. It also provides benefits to deal with the power allocation problems.

A. SPECTRAL EFFICIENCY WITH MAXIMUM RATIO TRANSMISSION

The first and simplest technique is MRT, which is defined for two users in cluster n as

$$\mathbf{w}_n^{\mathrm{MRT}} = \frac{\mathbf{h}_{n,2}}{\|\mathbf{h}_{n,2}\|_2} \in \mathbb{C}^N. \tag{12}$$

The main purpose of this beamforming technique is to amplify the designed signal from a particular user. However, the major drawback of this approach is that it does not support removing the mutual interference cancellation. By using MRT beamforming in (12), the achievable rate of user 1 in cluster n is now expressed as

$$R_{n,1}^{\mathrm{MRT}} = \log_2\left(1 + \mathrm{SINR}_{n,1}^{\mathrm{MRT}}\right) \text{ [b/s/Hz]}, \tag{13}$$

where the $\mathrm{SINR}_{n,1}^{\mathrm{MRT}}$ is expressed as

$$\mathrm{SINR}_{n,1}^{\mathrm{MRT}} = \frac{\frac{\left|\mathbf{h}_{n,1}^H \mathbf{h}_{n,2}\right|^2}{\|\mathbf{h}_{n,2}\|_2^2} p_{n,1}}{\frac{\left|\mathbf{h}_{n,1}^H \mathbf{h}_{n,2}\right|^2}{\|\mathbf{h}_{n,2}\|_2^2} p_{n,2} + \sum\limits_{\substack{n'=1,\\ n'\neq n}}^{N} \frac{\left|\mathbf{h}_{n,1}^H \mathbf{h}_{n',2}\right|^2}{\|\mathbf{h}_{n',2}\|_2^2} p_{n'} + \sigma_{\mathrm{DL}}^2}, \tag{14}$$

Meanwhile, the achievable rate of user 2 in cluster n is given as

$$R_{n,2}^{\mathrm{MRT}} = \log_2\left(1 + \mathrm{SINR}_{n,2}^{\mathrm{MRT}}\right) \text{ [b/s/Hz]}, \tag{15}$$

where the $\mathrm{SINR}_{n,2}^{\mathrm{MRT}}$ is

$$\mathrm{SINR}_{n,2}^{\mathrm{MRT}} = \frac{\frac{\left|\mathbf{h}_{n,2}^H \mathbf{h}_{n,2}\right|^2}{\|\mathbf{h}_{n,2}\|_2^2} p_{n,2}}{\sum\limits_{\substack{n'=1\\ n'\neq n}}^{M} \frac{\left|\mathbf{h}_{n,2}^H \mathbf{h}_{n',2}\right|^2}{\|\mathbf{h}_{n',2}\|_2^2} p_{n'} + \sigma_{\mathrm{DL}}^2}. \tag{16}$$

B. SPECTRAL EFFICIENCY WITH ZERO FORCING BEAMFORMING

The second linear beamforming technique considered here is ZF, which is a more effective solution to cancel interference than MRT. For the purpose of maximizing the total sum spectral efficiency under NOMA, the ZF beamforming vector defined for cluster n is

$$\mathbf{w}_n^{\mathrm{ZF}} = \frac{\mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}}{\left\|\mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\right\|_2} \in \mathbb{C}^N, \tag{17}$$

where $\mathbf{H} \in \mathbb{C}^{N\times N}$ is $\mathbf{H} = \left[\mathbf{h}_{1,2}, \ldots, \mathbf{h}_{N,2}\right]$ and $\mathbf{r}_n^{\mathrm{ZF}} \in \mathbb{C}^N$ is the n-th column of matrix $\left(\mathbf{H}^H\mathbf{H}\right)^{-1}$. We stress that the above definition of ZF beamforming focuses on the stronger user in each cluster which contributes significantly to the sum spectral efficiency. Furthermore, this beamforming technique produces the following property

$$\mathbf{h}_{n,2}^H \mathbf{w}_{n'}^{\mathrm{ZF}} = \begin{cases} 0, & \text{if } n' \neq n, \\ \dfrac{1}{\left\|\mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\right\|_2}, & \text{if } n' = n, \end{cases} \tag{18}$$

which is able to cancel severe mutual interference from the other clusters.

Utilizing ZF beamforming defined in (17), the achievable rate of user 1 in cluster n is

$$R_{n,1}^{\mathrm{ZF}} = \log_2\left(1 + \mathrm{SINR}_{n,1}^{\mathrm{ZF}}\right) \text{ [b/s/Hz]}, \tag{19}$$

where the $\mathrm{SINR}_{n,1}^{\mathrm{ZF}}$ is expressed as

$$\frac{\frac{\left|\mathbf{h}_{n,1}^H \mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\|_2^2} p_{n,1}}{\frac{\left|\mathbf{h}_{n,1}^H \mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\|_2^2} p_{n,2} + \sum\limits_{\substack{n'=1,\\ n'\neq n}}^{N} \frac{\left|\mathbf{h}_{n,1}^H \mathbf{H}\mathbf{r}_{n'}^{\mathrm{ZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_{n'}^{\mathrm{ZF}}\|_2^2} p_{n'} + \sigma_{\mathrm{DL}}^2}. \tag{20}$$

Meanwhile, the achievable rate of user 2 in cluster n is given as

$$R_{n,2}^{\mathrm{ZF}} = \log_2\left(1 + \mathrm{SINR}_{n,2}^{\mathrm{ZF}}\right) \text{ [b/s/Hz]}, \tag{21}$$

where the $\mathrm{SINR}_{n,2}^{\mathrm{ZF}}$ is

$$\mathrm{SINR}_{n,2}^{\mathrm{ZF}} = \frac{p_{n,2}}{\|\mathbf{H}\mathbf{r}_n^{\mathrm{ZF}}\|_2^2 \sigma_{\mathrm{DL}}^2}. \tag{22}$$

In comparison to (16), the SINR expression for the second user in (22) is much simpler because ZF eliminates mutual interference from other users effectively. Therefore, ZF is a beamforming technique which is expected to yield high spectral efficiency for strong users.

C. SPECTRAL EFFICIENCY WITH REGULARIZED ZERO FORCING

The last effective beamforming technique considered in this paper is RZF, which is defined as

$$\mathbf{w}_n^{\mathrm{RZF}} = \frac{\mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}}{\left\|\mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\right\|_2} \in \mathbb{C}^N. \tag{23}$$

where $\mathbf{r}_n^{\mathrm{RZF}}$ are the n-th column of the matrix $\left(\mathbf{H}^H\mathbf{H} + \alpha\mathbf{I}_N\right)^{-1}$ in which α is an non-negative constant that ensures the existence of inverse matrix. Notice that RZF beamforing vector

is more flexible than ZF counterpart thanks to controlling the value of α. In the case $\alpha = 0$, then RZF becomes ZF.

By using the beamforming vector defined in (23), the achievable rate of user 1 in cluster n is formulated as

$$R_{n,1}^{\mathrm{RZF}} = \log_2\left(1 + \mathrm{SINR}_{n,1}^{\mathrm{RZF}}\right) \text{ [b/s/Hz]}, \tag{24}$$

where $\mathrm{SINR}_{n,1}^{\mathrm{RZF}}$ is given as

$$\frac{\frac{\left|\mathbf{h}_{n,1}^H \mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\|_2^2} p_{n,1}}{\frac{\left|\mathbf{h}_{n,1}^H \mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\|_2^2} p_{n,2} + \sum\limits_{\substack{n'=1,\\ n'\neq n}}^{N} \frac{\left|\mathbf{h}_{n,1}^H \mathbf{H}\mathbf{r}_{n'}^{\mathrm{RZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_{n'}^{\mathrm{RZF}}\|_2^2} p_{n'} + \sigma_{\mathrm{DL}}^2}. \tag{25}$$

The achievable rate of user 2 in cluster n is given as

$$R_{n,2}^{\mathrm{RZF}} = \log_2\left(1 + \mathrm{SINR}_{n,2}^{\mathrm{ZF}}\right) \text{ [b/s/Hz]}, \tag{26}$$

where the signal-to-interference-and-noise ratio (SINR) value is

$$\mathrm{SINR}_{n,2}^{\mathrm{RZF}} = \frac{\frac{\left|\mathbf{h}_{n,2}^H \mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_n^{\mathrm{RZF}}\|_2^2} p_{n,2}}{\sum\limits_{\substack{n'=1\\ n'\neq n}}^{N} \frac{\left|\mathbf{h}_{n,2}^H \mathbf{H}\mathbf{r}_{n'}^{\mathrm{RZF}}\right|^2}{\|\mathbf{H}\mathbf{r}_{n'}^{\mathrm{RZF}}\|_2^2} p_{n'} + \sigma_{\mathrm{DL}}^2}. \tag{27}$$

IV. SUM SPECTRAL EFFICIENCY MAXIMIZATION

In this section, we maximize the total sum rate of this MIMO-NOMA system subject to the power budget constraints. The optimization problem is defined as

$$\begin{aligned} &\underset{\{p_{n,1}, p_{n,2} \geq 0\}}{\text{maximize}} \sum_{n=1}^{N} \left(R_{n,1} + R_{n,2}\right) \\ &\text{subject to } p_{n,1} + p_{n,2} \leq P_{\max,n} \quad \forall n. \end{aligned} \tag{28}$$

We emphasize that (28) is non-covex problem which the proof can be obtained by applying the main steps in [48]. Therefore, the optimal solution to (28) is difficult to obtain with any nontrivial setups. In order to apply the general purpose optimization toolboxes such as CVX [49], we first plug (6) and (8) into (28) and then obtain the following sum spectral efficiency problem.

$$\begin{aligned} &\underset{\{p_{n,1}, p_{n,2} \geq 0\}}{\text{maximize}} \sum_{n=1}^{N} \log_2\Big((1 + \mathrm{SINR}_{n,1})(1 + \mathrm{SINR}_{n,2})\Big) \\ &\text{subject to } p_{n,1} + p_{n,2} \leq P_{\max,n} \quad \forall n. \end{aligned} \tag{29}$$

By converting from the quality of service constraint to SINR constraints and using the epigraph form representation [50], the optimization problem (29) is equivalent to

$$\begin{aligned} &\underset{\substack{\{p_{n,1}, p_{n,2} \geq 0\}\\ \{\lambda_{n,1}, \lambda_{n,2} \geq 0\}}}{\text{maximize}} \prod_{n=1}^{N} \lambda_{n,1}\lambda_{n,2} \\ &\text{subject to } 1 + \mathrm{SINR}_{n,1} \geq \lambda_{n,1} \quad \forall n, \\ &\qquad\qquad 1 + \mathrm{SINR}_{n,2} \geq \lambda_{n,2} \quad \forall n, \\ &\qquad\qquad p_{n,1} + p_{n,2} \leq P_{\max,n} \quad \forall n. \end{aligned} \tag{30}$$

For generality purpose, cluster n acquires a general beamforming vector $\mathbf{w}_n$ which may be selected from the set $\{\mathbf{w}_n^{\mathrm{MRT}}, \mathbf{w}_n^{\mathrm{ZF}}, \mathbf{w}_n^{\mathrm{RZF}}\}$. From the SINR expressions, an observation of the optimization problem (30) is made in Lemma 1.

Lemma 1: The optimization problem (30) *is a signomial program.*

Proof: The main proof is to show that the SINR constraints are signomial (please see Appendix for the definition of a signomial function). For cluster n, the SINR constraint of user 1 can be reformulated as

$$\begin{aligned} &\lambda_{n,1}\left|\mathbf{h}_{n,1}^H \mathbf{w}_n\right|^2 p_{n,2} + \lambda_{n,1} \sum_{\substack{n'=1,\\ n'\neq n}}^{N} \left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 p_{n'} + \lambda_{n,1}\sigma_{\mathrm{DL}}^2 \\ &\qquad - \sum_{n'=1}^{N} \left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 p_{n'} - \sigma_{\mathrm{DL}}^2 \leq 0, \end{aligned} \tag{31}$$

whose the left-hand side is signomial due to the non-negative factors. Similarly, the SINR contraint of user 2 in cluster n is reformulated as

$$\begin{aligned} &\lambda_{n,2} \sum_{\substack{n'=1\\ n'\neq n}}^{N} \left|\mathbf{h}_{n,2}^H \mathbf{w}_{n'}\right|^2 p_{n'} + \lambda_{n,2}\sigma_{\mathrm{DL}}^2 - \left|\mathbf{h}_{n,2}^H \mathbf{w}_n\right|^2 p_{n,2} \\ &\qquad - \sum_{\substack{n'=1\\ n'\neq n}}^{N} \left|\mathbf{h}_{n,2}^H \mathbf{w}_{n'}\right|^2 p_{n'} - \sigma_{\mathrm{DL}}^2 \leq 0, \end{aligned} \tag{32}$$

which also formulates a signomial constraint. Therefore, we complete the proof. ■

Lemma 1 indicates that the optimization problem (30) still preserves the non-convexity feature. To tackle this issue, we will apply the successive optimization approach to find a local solution to (30) [45], [51]. We first introduce the arithmetic mean-geometric mean inequality as in problem (28). To this end, the signomial SINR constraints of each cluster must be converted to monomial ones by using the weighted arithmetic mean-geometric mean inequality [51] as shown in Lemma 2.

Lemma 2 [51, Lemma 1]: Assume that $h(x)$ is a posynomial function which is defined from N monomials $\{z_1(x), \ldots, z_N(x)\}$ as

$$h(x) = \sum_{n=1}^{N} z_n(x), \tag{33}$$

then the function $h(x)$ is lower bounded by a monomial function $\tilde{h}(x)$ as

$$h(x) \geq \tilde{h}(x) = \prod_{n=1}^{N} \left(z_n(x)/\alpha_n\right)^{\alpha_n}, \tag{34}$$

where α_n is a non-negative weight regarding to $z_n(x)$. The best approximation to $h(x_0)$ near the point x_0 in the sense of the first order Taylor expansion, say $\tilde{h}(x_0)$, if the weight α_n is selected as

$$\alpha_n = \frac{z_n(x_0)}{\sum_{n'=1}^{N} z_n'(x_0)}. \tag{35}$$

We now reformulate the SINR constraint of user 1 in cluster n as

$$\lambda_{n,1}\left|\mathbf{h}_{n,1}^H\mathbf{w}_n\right|^2 p_{n,2}+\lambda_{n,1}\sum_{\substack{n'=1,\\ n'\neq n}}^{N}\left|\mathbf{h}_{n,1}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\lambda_{n,1}\sigma_{\mathrm{DL}}^2 \leq \sum_{n'=1}^{N}\left|\mathbf{h}_{n,1}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\sigma_{\mathrm{DL}}^2. \quad (36)$$

We denote the right-hand side of (36) as

$$h_n(p_{n',1},p_{n',2})=\sum_{n'=1}^{N}\left|\mathbf{h}_{n,1}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\sigma_{\mathrm{DL}}^2, \quad (37)$$

then applying the arithmetic mean-geometric mean in Lemma 2, we lower bound $h_n(p_{n',1},p_{n',2})$ as

$$h_n(p_{n',1},p_{n',2}) \geq \tilde{h}_n(p_{n',1},p_{n',2}) = \left(\frac{\sigma_{\mathrm{DL}}^2}{\alpha_0^n}\right)^{\alpha_0^n}\prod_{n'=1}^{N}\left(\frac{\left|\mathbf{h}_{n,1}^H\mathbf{w}_{n'}\right|^2 p_{n',1}}{\alpha_{n',1}^n}\right)^{\alpha_{n',1}^n} \times\left(\frac{\left|\mathbf{h}_{n,1}^H\mathbf{w}_{n'}\right|^2 p_{n',1}}{\alpha_{n',2}^n}\right)^{\alpha_{n',2}^n}, \quad (38)$$

where the non-negative weights $\alpha_0^n,\alpha_{n',1}^n,\alpha_{n',2}^n$ associate with user 1 in cluster n and satisfy

$$\alpha_0^n+\sum_{n'=1}^{N}\alpha_{n',1}^n+\sum_{n'=1}^{N}\alpha_{n',2}^n=1. \quad (39)$$

The SINR constraint of user 1 in the optimization problem (30) is now approximated to

$$\frac{\lambda_{n,1}}{\tilde{h}_n(p_{n',1},p_{n',2})}\left|\mathbf{h}_{n,1}^H\mathbf{w}_n\right|^2 p_{n,2}+\frac{\lambda_{n,1}}{\tilde{h}_n(p_{n',1},p_{n',2})}\times\sum_{\substack{n'=1,\\ n'\neq n}}^{N}\left|\mathbf{h}_{n,1}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\frac{\lambda_{n,1}\sigma_{\mathrm{DL}}^2}{\tilde{h}_n(p_{n',1},p_{n',2})}\leq 1, \quad (40)$$

which is posynomial. Similarly, the SINR constraint of user 2 in cluster n is reformulated as

$$\lambda_{n,2}\sum_{n'=1,n'\neq n}^{M}\left|\mathbf{h}_{n,2}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\lambda_{n,2}\sigma_{\mathrm{DL}}^2 \leq \left|\mathbf{h}_{n,2}^H\mathbf{w}_n\right|^2 p_{n,2}+\sum_{\substack{n'=1,\\ n'\neq n}}^{N}\left|\mathbf{h}_{n,2}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\sigma_{\mathrm{DL}}^2. \quad (41)$$

For the SINR constraint of user 2 in cluster n, let us denote

$$g_n(p_{n',1},p_{n',2})=\left|\mathbf{h}_{n,2}^H\mathbf{w}_n\right|^2 p_{n,2}+\sum_{\substack{n'=1,\\ n'\neq n}}^{N}\left|\mathbf{h}_{n,2}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\sigma_{\mathrm{DL}}^2, \quad (42)$$

and applying the arithmetic mean-geometric mean inequality, a lower bound of $g_n(p_{n',1},p_{n',2})$ is

$$g_n(p_{n',1},p_{n',2}) \geq \tilde{g}_n(p_{n',1},p_{n',2}) = \left(\frac{\sigma_{\mathrm{DL}}^2}{\tilde{\alpha}_{0,0}^n}\right)^{\tilde{\alpha}_{0,0}^n}\left(\frac{\left|\mathbf{h}_{n,2}^H\mathbf{w}_n\right|^2 p_{n,2}}{\tilde{\alpha}_{0,1}^n}\right)^{\tilde{\alpha}_{0,1}^n} \times\prod_{n'=1,n'\neq n}^{N}\left(\frac{\left|\mathbf{h}_{n,2}^H\mathbf{w}_{n'}\right|^2 p_{n',1}}{\tilde{\alpha}_{n',1}^n}\right)^{\tilde{\alpha}_{n',1}^n} \times\prod_{n'=1,n'\neq n}^{N}\left(\frac{\left|\mathbf{h}_{n,2}^H\mathbf{w}_{n'}\right|^2 p_{n',1}}{\tilde{\alpha}_{n',2}^n}\right)^{\tilde{\alpha}_{n',2}^n}, \quad (43)$$

where the non-negative weights $\tilde{\alpha}_{0,0}^n,\tilde{\alpha}_{0,1}^n,\tilde{\alpha}_{n',1}^n,\tilde{\alpha}_{n',2}^n,\forall n'$, associate with user 2 in cluster n and satisfy

$$\tilde{\alpha}_{0,0}^n+\tilde{\alpha}_{0,1}^n+\sum_{n'=1}^{N}\tilde{\alpha}_{n,1}^n+\sum_{n'=1}^{N}\tilde{\alpha}_{n',2}^n=1. \quad (44)$$

To the end, the SINR constraint of user 2 in cluster n is now approximated as

$$\frac{\lambda_{n,2}}{\tilde{g}_n(p_{n',1},p_{n',2})}\sum_{\substack{n'=1,\\ n'\neq n}}^{N}\left|\mathbf{h}_{n,2}^H\mathbf{w}_{n'}\right|^2 p_{n'}+\frac{\lambda_{n,2}\sigma_{\mathrm{DL}}^2}{\tilde{g}_n(p_{n',1},p_{n',2})}\leq 1. \quad (45)$$

The optimal solution to (30) is bounded from below by the solution of the following problem

$$\begin{aligned}\underset{\substack{\{p_{n,1},p_{n,2}\geq 0\}\\ \{\lambda_{n,1},\lambda_{n,2}\geq 0\}}}{\text{maximize}}\ &\prod_{n=1}^{N}\lambda_{n,1}\lambda_{n,2}\\ \text{subject to}\ &\text{Constraint (40)}\quad \forall n,\\ &\text{Constraint (45)}\quad \forall n,\\ &p_{n,1}+p_{n,2}\leq P_{\max,n}\quad \forall n.\end{aligned} \quad (46)$$

We stress that the global solution to (46) is able to obtain in limited time due to its convexity as shown in Theorem 1.

Theorem 1: The optimization problem (46) *is a geometric program, so its global optimal solution is obtained with polynomial complexity.*

Proof: The proof is straightforward since the objective function of (46) is a monomial function, while the constraints are posynomial. Therefore, the optimization problem (46) follows the standard form of a geometric program as demonstrated in Definition 2 [50]. ■

We now apply the successive optimization approach to find a local solution to (30) in an iterative manner. Specifically, from an initial set of the data power $\{p_{n,1}^{(0)},p_{n,2}^{(0)}\}$ in the feasible set, at the i-th iteration, the weight values for user 1 in cluster n

is computed

$$\alpha_0^{n,(i)} = \frac{\sigma_{\mathrm{DL}}^2}{\sum_{n'=1}^{N} \left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 \left(p_{n',1}^{(i-1)} + p_{n',2}^{(i-1)}\right) + \sigma_{\mathrm{DL}}^2}, \tag{47}$$

$$\alpha_{n',1}^{n,(i)} = \frac{\left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 p_{n',1}^{(i-1)}}{\sum_{n'=1}^{N} \left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 \left(p_{n',1}^{(i-1)} + p_{n',2}^{(i-1)}\right) + \sigma_{\mathrm{DL}}^2}, \quad \forall n', \tag{48}$$

$$\alpha_{n',2}^{n,(i)} = \frac{\left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 p_{n',2}^{(i-1)}}{\sum_{n'=1}^{N} \left|\mathbf{h}_{n,1}^H \mathbf{w}_{n'}\right|^2 \left(p_{n',1}^{(i-1)} + p_{n',2}^{(i-1)}\right) + \sigma_{\mathrm{DL}}^2}, \quad \forall n'. \tag{49}$$

Meanwhile, the weight values for user 2 in cluster n is computed as

$$\tilde{\alpha}_{0,0}^{n,(i)} = \frac{\sigma_{\mathrm{DL}}^2}{\gamma^{n,(i-1)}}, \tag{50}$$

$$\tilde{\alpha}_{0,1}^{n,(i)} = \frac{\left|\mathbf{h}_{n,2}^H \mathbf{w}_n\right|^2 p_{n,2}^{(i-1)}}{\gamma^{n,(i-1)}}, \tag{51}$$

$$\tilde{\alpha}_{n',1}^{n,(i)} = \frac{\left|\mathbf{h}_{n,2}^H \mathbf{w}_{n'}\right|^2 p_{n',1}^{(i-1)}}{\gamma^{n,(i-1)}}, \quad \forall n', \tag{52}$$

$$\tilde{\alpha}_{n',2}^{n,(i)} = \frac{\left|\mathbf{h}_{n,2}^H \mathbf{w}_{n'}\right|^2 p_{n',2}^{(i-1)}}{\gamma^{n,(i-1)}}, \quad \forall n', \tag{53}$$

where the value $\gamma^{n,(i-1)}$ is computed from the previous iteration as

$$\gamma^{n,(i-1)} = \left|\mathbf{h}_{n,2}^H \mathbf{w}_n\right|^2 p_{n,2}^{(i-1)} + \sum_{\substack{n'=1,\\ n' \neq n}}^{N} \left|\mathbf{h}_{n,2}^H \mathbf{w}_{n'}\right|^2 p_{n'}^{(i-1)} + \sigma_{\mathrm{DL}}^2. \tag{54}$$

After that, we solve the geometric program (46) to attain the optimal solutions $\{p_{n,1}^{(i)}, p_{n,2}^{(i)}\}, \forall n$. At the end of each iteration, the weight values for the next iteration are computed from $\{p_{n,1}^{(i)}, p_{n,2}^{(i)}\}$ by utilizing (47)-(49) and (50)-(53). This iterative process will be terminated, for example, the variation between two consecutive iterations is small. In a nutshell, the proposed power control to maximize the sum spectral efficiency is summarized in Algorithm 1. Moreover, we manifest the convergence of this algorithm as shown in Theorem (2).

Theorem 2: The solution of Algorithm 1 converges to a fixed point which is the Karush-Kuhn-Tucker (KKT) point of the optimization problem (28).

Proof: The proof basically follows the main steps as shown in [45] and [52]. The guidance is as follows. By utilizing the arithmetic mean-geometric mean inequality to lower bound the SINR constraints as in (40) and (45), we guarantee that the global optimal solution to (46) is also a feasible point to the original problem (28). The successive optimization approach with solving a geometric program in each iteration produces a non-decreasing objective function to (46). Meanwhile, the limited power budget constraints ensure a convex feasible set, so Algorithm 1 must converge to a limited point. By doing a matching process for the KKT conditions of the two problems (28) and (46) at the limited point, we conclude that it is the KKT point of (28). ■

Algorithm 1 Successive Approximation Algorithm for (30)

Input: Set $i = 1$; Select the maximum power for each cluster $P_{\max,n}, \forall n$; Select the initial values of powers $p_{n,1}^{(0)}, p_{n,2}^{(0)}$ for $\forall n$; Compute the weight values for all users corresponding to $\{p_{n,1}^{(0)}, p_{n,2}^{(0)}\}$ by applying (47)-(49) and (50)-(53).

1. *Iteration i*: Solve the geometric program (46) to obtain $\{p_{n,1}^{(i)}, p_{n,2}^{(i)}\}$. Then update the weight values by applying (47)-(49) and (50)-(53)
2. If Stopping criterion satisfied → Stop. Otherwise, go to Step 3.
3. Set $\lambda_{n,1}^{\mathrm{opt}} = \lambda_{n,1}^{(i)}, \lambda_{n,2}^{\mathrm{opt}} = \lambda_{n,2}^{(i)}, \forall n$, and $p_{n,1}^{\mathrm{opt}} = p_{n,1}^{(i)}, p_{n,2}^{\mathrm{opt}} = p_{n,2}^{(i)} \forall n$; Set $i = i + 1$, back to Step 1.

Output: The solutions $\lambda_{n,1}^{\mathrm{opt}}, \lambda_{n,2}^{\mathrm{opt}}$ and $p_{n,1}^{\mathrm{opt}}, p_{n,2}^{\mathrm{opt}}, \forall n$.

TABLE 1. Simulated network parameters.

Propagation min radius	3.5 [m]
Propagation max radius	250 [m]
Beamforming technology	MRT, RZF, ZF
α weight in RZF beamforming	σ_{DL}
Operating frequency	5.9 [GHz]
Maximum transmit power of device and cellular user	26 [dBm]
Noise power	-96 [dBm]

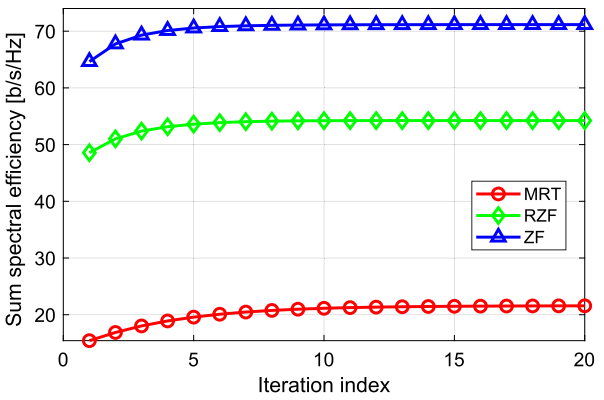


FIGURE 3. Convergence of Algorithm 1 versus different linear beamforming techniques. The network serves 40 users.

V. SIMULATION RESULTS

In this section, we execute extensively numerical simulations in order to testify and compare the performance among all considered beam-forming techniques in the previous sections. The system parameters are given in Table 1. The content includes comparing how fast the algorithm converges among the three beam-forming techniques which are shown in Fig. 3.

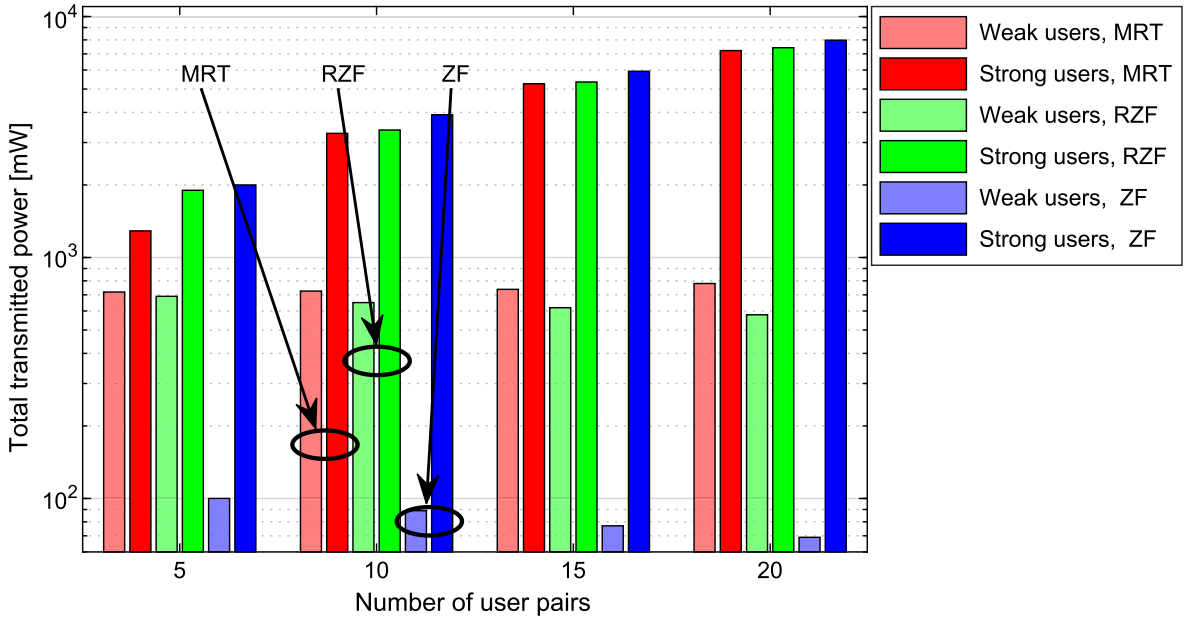


FIGURE 4. Total transmitted power of weak and strong users versus the number of pairs in the network.

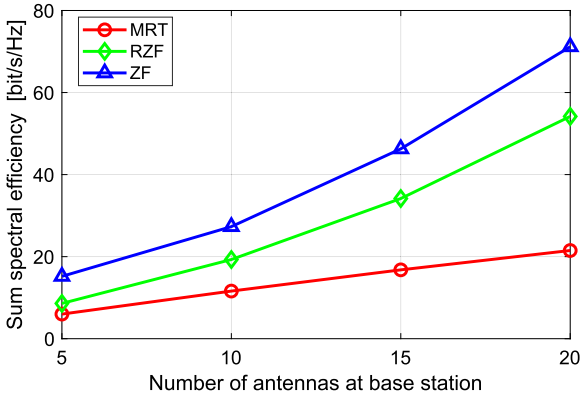


FIGURE 5. Sum spectral efficiency of strong users for different linear beamforming techniques.

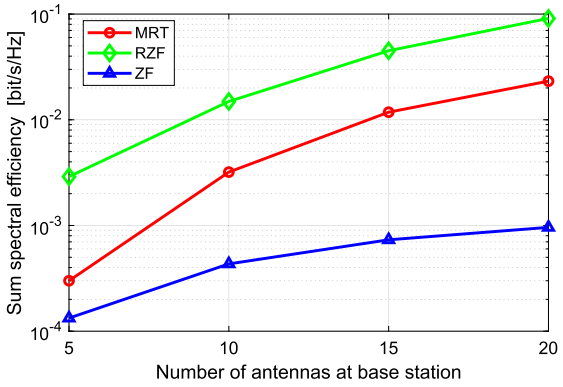


FIGURE 6. Sum spectral efficiency of weak users for different linear beamforming techniques.

Fig. 4 illustrates the difference of transmitted power that the base station allocates to all users. The sum spectral efficiency of strong and weak users is given in Fig. 5 and Fig. 6, respectively. The remaining figures demonstrate per user spectral efficiency in the network.

1) CHANNEL MODEL

In the scope of this study, the propagation channels from the base station to all users follow uncorrelated Rayleigh distribution which are formulated as:

$$\mathbf{h}_{n,n'} \sim \mathcal{CN}\left(\mathbf{0}, \beta_{n,n'}\mathbf{I}_N\right), \quad \forall n = 1, \ldots N, n' = 1, 2, \tag{55}$$

where $\beta_{n,n'}$ represent large-scale fading coefficients modelling the path loss and shadow fading. This is defined as

$$\beta_{n,n'} = \xi_{n,n'} 10^{\frac{\sigma z_{n,n'}}{10}}, \quad \forall n, n', \tag{56}$$

where σ is the standard deviation of shadow fading, which is set to 7 dB in this paper. $z_{n,n'}$ is distributed as $\mathcal{CN}(0, 1)$. For the path loss coefficients $\xi_{n,n'}$ we use the model which is given by

$$\xi_{n,n'} = -148.1 - 37.6 \log_{10} d_{n,n'}, \tag{57}$$

where $\xi_{n,n'}$ is in dB and $d_{n,n'}$ (measured in km) represents the distance between the base station and users in cluter n'.

2) PERFORMANCE EVALUATION

Fig. 3 demonstrates the convergence property of our proposed solution to the sum spectral efficiency optimization as shown in Theorem 2. For all considered linear beamforming techniques, Algorithm 1 converges very fast which requires less than 20 iterations to reach the KKT point. Thanks to carefully allocating the transmit power to each user, MRT can significantly improve 40% the sum spectral efficiency better than the initial point in the feasible set. This superior gain comes from the fact that MRT does not suppress mutual interference well, so the system performs badly without power control. In the case of using a more advanced beamforming technique as RZF and ZF, the improvement is about 12% and 10%, respectively. At the KKT point,

it observes the better performance of RZF than MRT with the gap of 32.7 b/s/Hz. Among the three linear precoding techniques, ZF is the optimal selection to maximize the sum capacity of a NOMA-based network.

Fig. 4 describes the total transmitted power of a network by using different linear beamforming techniques. If the BS is equipped with 5 antennas that is able to serve 10 users, the transmitted power of weak users using either RZF or MRT is nearly 7× higher than the consumption of ZF. When the number of user pairs increases from 5 up to 20, the gap between ZF and the others grows significantly. In particular, utilizing ZF only needs the total transmitted power for the weak users around 69.12 mW. However it consumes 587.5 mW when RZF is deployed. This observation is of paramount importance for a system with many user terminals.

Figs. 5 and 6 represent the sum spectral efficiency of weak and strong users versus the number of pairs, respectively. As visualization in Fig. 5, while considering the sum spectral efficiency, ZF provides the highest total rate of about 71.16 bit/s/Hz in order to serve 20 pairs of users. The amount of capacity is 54.32 bit/s/Hz in case of using RZF. Meanwhile, MRT, which is the worst choice, can only achieve 21.57 bit/s/Hz. However, the weaker group suffers from the mechanism of assigning beamforming vector to each pair in the NOMA-based network, which leads to a significant lower capacity than what is provided for strong users. The total spectral efficiency of all stronger users using ZF is approximately 93632 times higher than that of the weak ones. Although RZF is not the best choice for the net spectral efficiency, it enhances the performance of weak users. As we can see in this figure, there is a trade off in providing throughput to either strong or weak users. RZF in fact does not fully support strong users, but it raises the performance of weak priority group. In particular, the network utilizing RZF can provide 9.11×10^{-02} bit/s/Hz for all weak users. This does much better than that one with ZF since it is only able to provide 9.63×10^{-04} b/s/Hz in total.

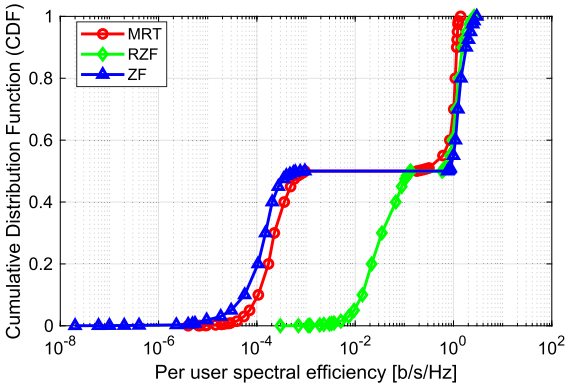


FIGURE 7. Per user spectral efficiency [b/s/Hz] for different linear beamforming techniques (both strong and weak users).

Fig. 7 illustrates the cumulative distribution function (CDF) for the per user spectral efficiency in a simulated network serving 40 users, which is explicitly divided into two

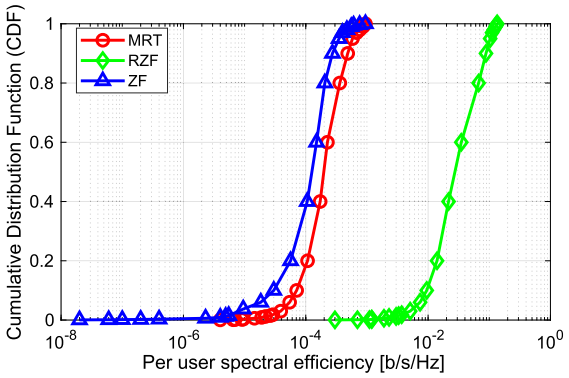


FIGURE 8. Per user spectral efficiency [b/s/Hz] for different linear beamforming techniques (weak users only).

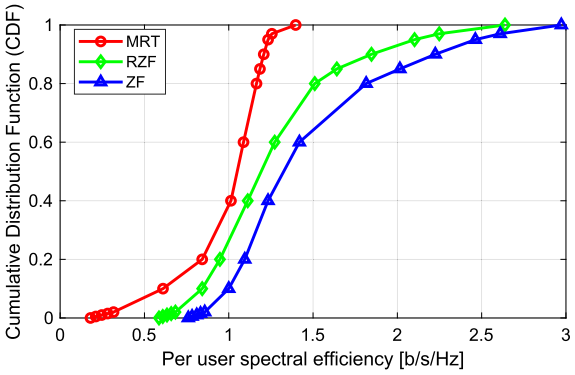


FIGURE 9. Per user spectral efficiency [b/s/Hz] for different linear beamforming techniques (strong users only).

distinct subsets: strong and weak users. Interestingly, ZF gets the lowest spectral efficiency at 95%-likely. MRT outperforms ZF about 2.4×, while the improvement from RZF is significantly up to and 317× compared with the baseline. This comes from the fact that ZF forces many users to be served with almost zeros rate such that the total sum spectral efficiency is maximal. For clearer observations, from Fig. 7 we provide the separate CDF of every weak users as in Fig. 9 and of every strong users as in Fig. 8. At the median point, the network can provide 1.05 b/s/Hz, 1.18 b/s/Hz, and 1.32 b/s/Hz for every strong users by deploying MRT, RZF, and ZF, respectively. By slightly deducting the capacity of every strong users, RZF yields the superiority in spectral efficiency for weak users among all considered linear beamforming techniques which is indicated in Fig. 8.

For better observations on the contributions of our proposed power allocation method, we plot the sum spectral efficiency with different power allocation methods as shown in Fig. 10. The first benchmark is uniform power control where all users transmit full power as proposed in [53]. The second benchmark allows each user to peak up a random power in the feasible set, which was used in [54]. Thanks to effectively mitigating mutual interference by NOMA, the two benchmarks in the previous works perform well when there are a few antennas equipped at BS. For the network where BS has 5 antennas, the loss by allocating uniformly power is

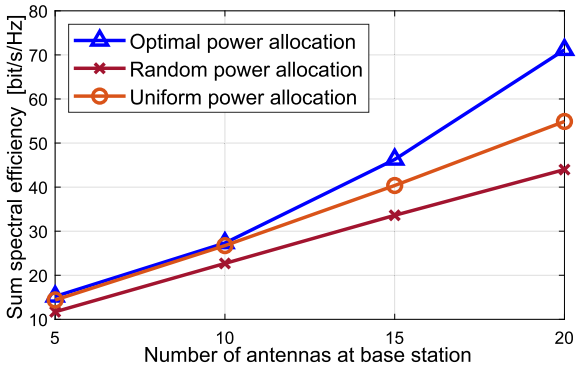


FIGURE 10. Sum spectral efficiency versus different power allocation benchmarks. The network uses ZF beamforming.

only 5.33%, while it is 22.96% if random power allocation is selected. Therefore, the random power allocation only serves as the baseline for comparison. The competitive performance among methods is because the use of only NOMA is sufficient when there are not many users in the coverage area. However, the superiority of our proposed control solution can be observed if increasing the number of base station antennas to tolerate more users. In particular, the gain from the optimal power allocation is up to 29.62% and 62.10% compared with uniform and random power allocation, respectively.

VI. CONCLUSIONS

In this paper, we have investigated the effectiveness of linear beamforming techniques for MIMO-NOMA downlink transmission systems with the target of maximizing the sum spectral efficiency. The performance analysis was rigorously made for a single-cell network utilizing the proposed user paring method based on the prior information of different large-scale channel qualities. We have formulated a non-convex sum spectral efficiency maximization taking care of the limited transmit power budget. An efficient frame work to find the local KKT solution in polynomial time was proposed. Numerical results manifest that ZF is the optimal selection for the total spectral efficiency of strong users and the entire network as well. In spite of losing the performance on strong users, RZF is the best solution to guarantee QoS for the weak users. In many situations, RZF outperforms ZF to achieve better per user spectral efficiency. Among the three considered beamforming techniques, MRT yields the lowest performance, but this is a good candidate for the large-scale networks due to its simplicity and scalability.

APPENDIX

The appendix provides the two useful definitions which are widely used in this paper [45], [50]. In more detail, Definition 1 gives the concept of signomial, polynomial, and monomial functions, while the stand form of a geometric program is stated in Definition 2.

Definition 1: A multivariate function $h(x_1, \ldots, x_{M_1}) = \sum_{n=1}^{M_2} c_n \prod_{m=1}^{M_1} x_m^{b_{n,m}}$ *defined in* $\mathbb{R}_+^{M_1}$ *is signomial with* M_2 *terms* ($M_2 \geq 2$) *if the exponents* $b_{n,m}$ *are real numbers and the coefficients* c_n *are real, but at least one is negative. If all* $c_n, \forall n$, *are positive, the function* $h(x_1, \ldots, x_{N_1})$ *is polynomial. In the case* $h(x_1, \ldots, x_{M_1}) = c_n \prod_{m=1}^{M_1} x_m^{b_{n,m}}$, *then it is a monomial function if* $c_n > 0$.

Definition 2: A geometric program has the following standard form

$$\begin{aligned} &\underset{\mathbf{x}\in\mathcal{X}}{\text{maximize}}\, f_0(\mathbf{x}) \\ &\text{subject to}\, f_n(\mathbf{x}) \leq 1 \quad \forall n = 1, \ldots N, \\ &\qquad\qquad h_m(\mathbf{x}) = 1 \quad \forall m = 1, \ldots, M, \end{aligned} \tag{58}$$

where $f_0(\mathbf{x})$ *can be either monomial or polynomial. The functions* $f_n(\mathbf{x}), n = 1, \ldots, N$ *are polynomial and* $h_m(\mathbf{x}), m = 1, \ldots, M$ *are monomial. The feasible set* $\mathcal{X}$ *is convex.*

REFERENCES

[1] K. Fazel and S. Kaiser, *MMulti-Carrier and Spread Spectrum Systems: From OFDM and MC-CDMA to LTE and WiMAX*. Hoboken, NJ, USA: Wiley, 2008.

[2] G. L. Stüber, J. R. Barry, S. W. McLaughlin, Y. Li, M. A. Ingram, and T. G. Pratt, "Broadband MIMO-OFDM wireless communications," *Proc. IEEE*, vol. 92, no. 2, pp. 271–294, Feb. 2004.

[3] R. Van Nee, V. Jones, G. Awater, A. Van Zelst, J. Gardner, and G. Steele, "The 802.11n MIMO-OFDM standard for wireless LAN and beyond," *Wireless Pers. Commun.*, vol. 37, nos. 3–4, pp. 445–453, 2006.

[4] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[5] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *Proc. IEEE CSCN*, Oct. 2015, pp. 18–23.

[6] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32–38, Apr. 2018.

[7] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[8] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, "A tutorial on nonorthogonal multiple access for 5G and beyond," *Wireless Commun. Mobile Comput.*, vol. 2018, Feb. 2018, Art. no. 9713450.

[9] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

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

[11] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[12] Y. Cai, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, "Modulation and multiple access for 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 629–646, 1st Quart., 2018.

[13] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, Feb. 2017.

[14] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[15] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[16] F. Zhou, Y. Wu, Y.-C. Liang, Z. Li, Y. Wang, and K.-K. Wong, "State of the art, taxonomy, and open issues on cognitive radio networks with NOMA," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 100–108, Apr. 2018.

[17] J. Jeon, "NR wide bandwidth operations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 42–46, Mar. 2018.

[18] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

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

[20] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[21] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, Mar. 2017.

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

[23] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[24] 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.

[25] M. B. Shahab, M. F. Kader, and S. Y. Shin, "A virtual user pairing scheme to optimally utilize the spectrum of unpaired users in non-orthogonal multiple access," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1766–1770, Dec. 2016.

[26] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Robust beamforming techniques for non-orthogonal multiple access systems with bounded channel uncertainties," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2033–2036, Sep. 2017.

[27] Z. Zhu, Z. Chu, N. Wang, S. Huang, Z. Wang, and I. Lee, "Beamforming and power splitting designs for AN-aided secure multi-user MIMO SWIPT systems," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2861–2874, Dec. 2017.

[28] Q. Li, Q. Zhang, and J. Qin, "Secure relay beamforming for SWIPT in amplify-and-forward two-way relay networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9006–9019, Nov. 2016.

[29] L. Yang, M. O. Hasna, and I. S. Ansari, "Physical layer security for TAS/MRC systems with and without co-channel interference over $\nu-\mu$ fading channels," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12421–12426, Dec. 2018.

[30] T. Yoon, T. H. Nguyen, X. T. Nguyen, D. Yoo, B. Jang, and V. D. Nguyen, "Resource allocation for NOMA-based D2D systems coexisting with cellular networks," *IEEE Access*, vol. 6, pp. 66293–66304, 2018.

[31] J. Choi, "Minimum power multicast beamforming with superposition coding for multi-resolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[32] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, 2018.

[33] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

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

[35] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, "Low complexity beamforming and user selection schemes for 5G MIMO-NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2708–2722, Dec. 2017.

[36] Z. Chen, Z. Ding, and X. Dai, "Beamforming for combating inter-cluster and intra-cluster interference in hybrid NOMA systems," *IEEE Access*, vol. 4, pp. 4452–4463, Aug. 2016.

[37] T. H. Nguyen, T. K. Nguyen, H. D. Han, and V. D. Nguyen, "Optimal power control and load balancing for uplink cell-free multi-user massive MIMO," *IEEE Access*, vol. 6, pp. 14462–14473, 2018.

[38] V. D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O.-S. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2681–2695, Dec. 2017.

[39] Y. Li and G. A. A. Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, Dec. 2018.

[40] X. Chen, F. Gong, G. Li, H. Zhang, and P. Song, "User pairing and pair scheduling in massive MIMO-NOMA systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 788–791, Apr. 2018.

[41] J. Ma, C. Liang, C. Xu, and L. Ping, "On orthogonal and superimposed pilot schemes in massive MIMO-NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2696–2707, Dec. 2017.

[42] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.

[43] N. T. Hoa, N. T. Hieu, N. Van Duc, G. Gelle, and H. Choo, "Second order suboptimal power allocation for OFDM-based cognitive radio systems," in *Proc. 7th Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2013, p. 50.

[44] T. Nguyen, H. Nguyen, V. Nguyen, G. Gelle, and H. Choo, "Second order suboptimal power allocation for MIMO-OFDM based cognitive radio systems," *KSII Trans. Internet Inf. Syst.*, vol. 8, pp. 2647–2662, Aug. 2014.

[45] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint pilot design and uplink power allocation in multi-cell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 2000–2015, Mar. 2018.

[46] Y. Jiang, M. K. Varanasi, and J. Li, "Performance analysis of ZF and MMSE equalizers for MIMO systems: An in-depth study of the high SNR regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2008–2026, Apr. 2011.

[47] T. Van Chien and E. Björnson, *Massive MIMO Communications*. Springer, 2017, pp. 77–116.

[48] V. S. Annapureddy and V. V. Veeravalli, "Sum capacity of MIMO interference channels in the low interference regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 5, pp. 2565–2581, May 2011.

[49] M. Grant and S. Boyd. (Dec. 2017). *CVX: MATLAB Software for Disciplined Convex Programming, Academic Users*. [Online]. Available: http://cvxr.com/cvx

[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[51] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, Jul. 2007.

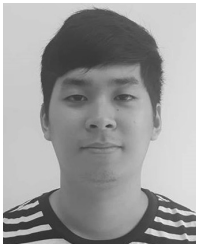
[52] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 26, pp. 681–683, Jul. 1978.

[53] T. Van Chien, C. Mollén, and E. Björnson, "Large-scale-fading decoding in cellular massive MIMO systems with spatially correlated channels," *IEEE Trans. Commun.*, to be published, doi: 10.1109/TCOMM.2018.2889090.

[54] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," in *Proc. Int. Zurich Seminar Inf. Commun. (IZS8)*, Zürich, Switzerland, Feb. 2018, pp. 46–47.

[55] T. Van Chien, E. Björnson, and E. G. Larsson, "Multi-cell massive MIMO performance with double scattering channels," in *Proc. IEEE Int. Workshop Comput.-Aided Model. Anal. Design Commun. Links Netw. (CAMAD)*, Toronto, ON, Canada, Oct. 2016, pp. 231–236.

SEONGGYOON PARK received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1985, 1987, and 1994, respectively. He is currently a Professor with the Radio Engineering Department, Information and Communication School, Kongju National University, Cheonan, South Korea. His research interests include 5G and wireless communications technology, radio compatibility and interference analysis, and public safety communications.



ANH QUAN TRUONG is currently pursuing the degree in electronics and telecommunications with the Hanoi University of Science and Technology, Vietnam. He has been a member of the Signal Processing Laboratory, for one year. His research interests include convex optimization work in multiple-input multiple-output 4G and future 5G, machine learning, and application of computer vision.

TIEN HOA NGUYEN received the Dip.Ing. degree in electronics and communication engineering from Hanover University and the Ph.D. degree from the Department Wireless Communication Technique, Hanoi University of Science and Technology, in 2016. He was involved in image processing with the R&D Department and in the development of SDR-based drivers with Bosch. He had also three-year experiment with R&D Team, MIMOon, to develop the embedded signal processing and radio modules for 4G and 5G mobile networks. He is currently a Lecturer with the Hanoi University of Science and Technology. His research interests include resource allocation in cognitive radios, massive multiple-input multiple-output, and vehicular communication systems.