Joint Beamwidth and Power Optimization in MmWave Hybrid Beamforming-NOMA Systems

Mojtaba Ahmadi Almasi, Lisi Jiang, Hamid Jafarkhani, and Hani Mehrpouyan

Abstract

The use of directional transmission in millimeter-Wave (mmWave) frequencies results in limited channel coherence time. In this paper, we take the limited channel coherence time into account for non-orthogonal multiple access (NOMA) in mmWave hybrid beamforming systems. Due to the limited coherence time, the beamwidth of the hybrid beamformer affects the beam-training time, which in turn directly impacts the data transmission rate. To investigate this trade-off, we utilize a combined beam-training algorithm. Then, we formulate a sum-rate expression which considers the channel coherence time and beam-training time as well as users' power and other system parameters. Further, a joint power and beamwidth optimization problem is solved by iterating between the power allocation and the beamwidth optimization. When allocating the power, we use the log-exponential reformulation and the sequential parametric convex approximation (SPCA) methods to solve the non-convex problem. Since beamwidth optimization involves too many variables, we propose an algorithm which iterates between clusters of users. Numerical results show that the optimized mmWave hybrid beamforming-NOMA system can achieve much higher sum-rates compared to NOMA with analog beamforming and traditional multiple access techniques.

Index Terms

mmWave communication, NOMA, beamwidth control, sum-rate, optimization.

This work was supported in part by the NSF Awards ECCS-1642865, ECCS-1642536, and CCF-2008786.

M. Ahmadi Almasi and H. Mehrpouyan are with the Department of Electrical and Computer Engineering, Boise State University, Boise, ID 83725, USA (e-mail: mojtabaahmadialm@u.boisestate.edu, hanimehrpouyan@boisestate.edu). L. Jiang and H. Jafarkhani are with Center for Pervasive Communications and Computing, University of California at Irvine, Irvine, CA 92697, USA (e-mail: {lisi.jiang, hamidj}@uci.edu), where M. Ahmadi Almasi was a visiting student.

I. INTRODUCTION

Current wireless communication networks operating under 6 GHz are restrained by limited spectral resources. Subsequently, it is necessary to use the millimeter-Wave (mmWave) band ranging from 30 to 300 GHz to increase the available spectrum [1]. Short wavelength and a large path loss are key characteristics of mmWave communication systems. Due to the short wavelength, a large number of antennas can be packed in a small area in mmWave devices. This feature combined with beamforming can be used to tackle severe path loss. Since a fully-digital beamforming may not be practical, various architectures have been proposed for mmWave outdoor communications, i.e., analog beamforming with multiple RF chains [2], hybrid beamforming [3]–[5], beamspace multiple-input multiple-output (MIMO) [6], and reconfigurable antenna-based MIMO [7], [8]. The fully-digital architecture needs one radio frequency (RF) chain per antenna. Hence, power consumption by large number of RF chains and hardware complexity are the main obstacles in implementing the fully-digital architecture. Although multi-user analog beamforming reduces the hardware complexity of the system and uses only the angle of arrival (AoA) information for beam alignment, it may not completely direct the total energy of a beam toward a desired receiver [2]. Accordingly, alternative methods such as beamspace MIMO [6] and reconfigurable antenna-based MIMO [7], [8] architectures reduce the number of required RF chains by dedicating one RF chain to each channel path instead of each antenna. However, these architectures are not able to change their beamwidth, which seems to be necessary and desired in the mmWave networks [9]. This is because in lens-based architectures, the lens operates like a passive phase-shifter network. Hence, it may not be possible to adjust the beamwidth. In contrast, not only does the hybrid beamforming architecture reduce the number of required RF chains, but also, thanks to the use of phase-shifters, it can adjust the transmission beamwidth. Hence, in this paper, we adopt the hybrid beamforming architecture which is a feasible solution to meet the demands in mmWave networks.

Non-orthogonal multiple access (NOMA) aims to improve the spectral efficiency and simultaneously serve more than one user at the same frequency/time/code in single-carrier and multi-carrier systems [10], [11]. Especially, NOMA transmits the users' signal at the same time slot and frequency band by using superposition coding (SC) and decodes the desired signal by exploiting successive interference cancellation (SIC) at the receiver [12]. In this paper, we leverage power-domain NOMA in which each user has a different level of power.

Recently, NOMA has been incorporated into the mmWave communication, termed mmWave-

NOMA, to enhance spectral efficiency and connectivity of the network. Here, we review the work on mmWave-NOMA networks in the downlink transmission with a single transmitter [9], [13]–[20]. In [13], a random beamforming method is studied for mmWave directional transmission. In [14], two NOMA users with different directions are assigned the same beamforming codeword using phase-shifters with finite resolution. NOMA is combined with lens-based beamspace MIMO in [15], and a power allocation algorithm is proposed. Energy-efficiency of mmWave-NOMA networks is evaluated in [16]. A joint power allocation optimization to design beamforming vectors for mmWave-NOMA networks is presented in [17]. The coverage and rate of mmWave-NOMA networks for analog beamforming in the presence of misalignment between the transmit and receive beams is analyzed in [18]. The impact of beamwidth on user pairing in mmWave-NOMA is studied in [9]. Also, [19] evaluates the effect of beam misalignment on the sum-rate performance of mmWave-NOMA networks with hybrid beamforming. Further, NOMA is utilized in lens-based mmWave reconfigurable antennas to increase the number of served users and improve the sum-rate in [20]. What is common among the above works is the assumption that there is sufficient time to train the beams.

In practical scenarios, neglecting the effect of beam-training duration may cast doubt on the performance of the mmWave-NOMA networks. Especially, since the channel coherence time in mmWave bands is limited [21], the beam-training duration should be adequately small. Thus, on one hand, a small beam-training duration results in a wide beamwidth, i.e., low beamforming gain, and noisy channel estimation. On the other hand, a long beam-training time provides robust beamforming and accurate channel estimation but imposes a delay in data transmission. This may not be desirable in delay-sensitive systems as it leaves less time for data transmission and leads to low sum-rates. There is a rich literature on fast beam-training algorithms [22]–[31]. This issue is very crucial in mmWave-NOMA networks in which more users are trained at each frequency/time resource. Beamwidth control and sum-rate trade-off in the mmWave analog beamforming-NOMA network for two users are evaluated in [32]. The impact of beam-training duration on the sum-rate of the system is determined and then an optimization problem that maximizes the sum-rate subject to the training duration and allocated power for each user is investigated. However, due to the inter-cluster interference, an extension of this architecture to the mmWave hybrid beamforming-NOMA network is quite challenging. In this paper, motivated by [32], we study the beamwidth control and sum-rate trade-off for the mmWave hybrid beamforming-NOMA network. There are two major differences between [32] and our work. First, we consider a

hybrid beamforming system which produces side lobes and as a result inter-cluster interference. Second, we do not allow receivers to have a beamwidth wider than that of their intended transmit beam. Otherwise, the receiver cannot catch the entire transmission energy. Neither the first case nor the second case is considered in [32]. The contributions of this paper are listed below:

1) We consider the well-studied mmWave hybrid beamforming combined with NOMA for limited coherence time scenarios. The system can control the beamwidth, using the phase-shifters deployed in the hybrid beamformer, and allocate power to NOMA users. To this end, a tone-based beam-training algorithm [26] compatible to our mmWave-NOMA system is utilized. The algorithm combines the exhaustive search [22] and tone-based beam-training [26] algorithms.
2) Unlike the existing multi-beam mmWave-NOMA systems, we take the channel coherence time into account. The limited coherence time leads to a trade-off between the beamwidth resolution and the data transmission rate. We also formulate a new sum-rate expression for optimization.
3) A joint power and beamwidth optimization algorithm is proposed which iterates between the power allocation and beamwidth optimization.
4) The numerical results verify the effectiveness of the joint optimization algorithm. Also, three significant results are revealed. First, at low signal-to-noise ratios (SNRs), both power allocation and beamwidth control play a major role in the sum-rate while at high SNRs, beamwidth is the only important parameter. Second, for very short channel coherence times and high SNRs the optimization is not required and predefined fixed values can be used instead. Third, a bottleneck for achieving high sum-rates is a small number of antennas, which results in a low resolution beamwidth, especially at large coherence time and low SNRs.

The rest of the paper is organized as follows. In Section II, the system model is described. Section III formulates the optimization problem. In Section IV, the allocated power and beamwidth are determined through the proposed optimization algorithm and its convergence analysis is provided. Numerical results are presented in Section V. Section VI concludes the paper.

Notations: Hereafter, $j = \sqrt{-1}$, small letters, bold letters, and bold capital letters designate scalars, vectors, and matrices, respectively. Superscripts $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^\dagger$ denote the transpose, conjugate, and transpose-conjugate operators, respectively. Further, $|\cdot|$, $\|\cdot\|$, $\|\cdot\|_2$, and $\|\cdot\|_F$ denote the absolute value, the norm-1, the norm-2, and the Frobenius norm, respectively. Also,

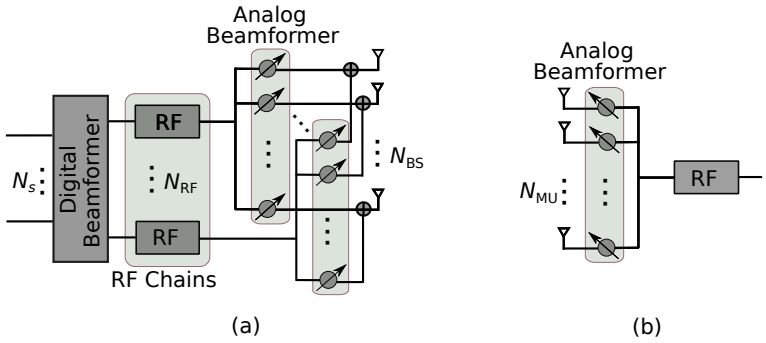

Fig. 1: Schematic of (a) the BS with hybrid beamforming structure and (b) a user equipment with analog beamforming structure.

$\mathbb{E}(\cdot)$ denotes the expectation. Finally, $\lceil \cdot \rceil$ denotes the ceiling function.

II. System Model

We assume a narrow band mmWave downlink system composed of a single cell with a base station (BS) and M_{UE} user devices. The BS is equipped with N_{RF} RF chains and N_{BS} antennas whereas each user has one RF chain and N_{UE} antennas. Each RF chain is connected to the antennas through phase-shifters. The architecture of the BS and a typical user is shown in Fig. 1. Due to the hybrid beamforming structure at the BS, the number of antennas is larger than the number of RF chains, $N_{\text{BS}} > N_{\text{RF}}$, and due to the analog beamforming at the users, we have $N_{\text{UE}} > 1$. Further, the BS transmits N_s streams simultaneously by steering N_{B} beams toward the users. To implement hybrid beamforming, the condition $N_{\text{B}} \leq N_{\text{RF}}$ should be satisfied. In this paper, however, we assume $N_{\text{B}} = N_{\text{RF}}$ to reduce the complexity and cost of the system. Indeed, if we consider sending one stream via one beam, it results in $N_s = N_{\text{RF}}$. On the other hand, to establish better connectivity by increasing the number of simultaneously served users in a dense area and further improve spectral efficiency, we use NOMA in the proposed mmWave hybrid beamforming network. Hence, each beam can serve more than one user. That is, the transmitter simultaneously sends N_{RF} streams toward $M_{\text{UE}} > N_{\text{RF}}$ users which are grouped into N_{RF} clusters, i.e., $M_{\text{UE}} = \sum_{n=1}^{N_{\text{RF}}} K_n$, where K_n denotes the number of users in the nth cluster. Note that NOMA requires the number of users in each cluster to be more than one, which should be satisfied by $K_n > 1$. Hereafter, the mth user equipment in the nth cluster is represented by $\text{UE}_{n,m}$.

A. Channel Model

We use the widely adopted extended Saleh-Valenzuela model as a multi-path channel (MPC) model in our mmWave hybrid beamforming-NOMA system [3], [33]. In this model, each LoS

and NLoS path is defined by a channel gain and an array steering vector at the transmitter and an array response vector at the receiver. Hence, the channel matrix between the BS and $\text{UE}_{n,m}$ in downlink is given by

$$\mathbf{H}_{n,m} = \frac{1}{\sqrt{L_{n,m}+1}} \left(\beta_{n,m,0}\mathbf{G}_{n,m,0} + \sum_{l=1}^{L_{n,m}} \beta_{n,m,l}\mathbf{G}_{n,m,l} \right), \tag{1}$$

where $\beta_{n,m,0}$ and $\beta_{n,m,l}$ denote the channel gain of LoS and NLoS channels, respectively. $\mathbf{G}_{n,m,0} \in \mathbb{C}^{N_{\text{UE}} \times N_{\text{BS}}}$ is the LoS channel matrix and $\mathbf{G}_{n,m,l}$ is the lth NLoS channel matrix. In particular, $\mathbf{G}_{n,m,l}$, $0 \leq l \leq L_{n,m}$, is given by

$$\mathbf{G}_{n,m,l} = \mathbf{a}_{\text{UE}}(\theta^{\text{az}}_{n,m,l}, \theta^{\text{el}}_{n,m,l}) \mathbf{a}^{\dagger}_{\text{BS}}(\phi^{\text{az}}_{n,m,l}, \phi^{\text{el}}_{n,m,l}), \tag{2}$$

where $\theta^{\text{az}}_{n,m,l}$ ($\theta^{\text{el}}_{n,m,l}$) and $\phi^{\text{az}}_{n,m,l}$ ($\phi^{\text{el}}_{n,m,l}$) are normalized azimuth (elevation) AoA and angle of departure (AoD), respectively. Also, $\mathbf{a}_{\text{BS}} \in \mathbb{C}^{N_{\text{BS}} \times 1}$ and $\mathbf{a}_{\text{UE}} \in \mathbb{C}^{N_{\text{UE}} \times 1}$ are the antenna array steering vector and array response vector of the BS and $\text{UE}_{n,m}$, respectively. In mmWave outdoor communications, to further reduce the interference, sectorized BSs can be employed. Mostly, each sector in the azimuth domain is much wider than that of the elevation domain [2], [4]. Reasonably, we assume that the BS separates the clusters in the azimuth domain and considers fixed elevation angles. Further, we assume that the sector-level beamwidth for the BS is defined by ω_{BS} and for each user is defined by ω_{UE}. Hence, the BS implements only azimuth beamforming and neglects elevation beamforming. In this case, the antenna configuration is a uniform linear array (ULA) and the superscript "el" is dropped. For a ULA, the steering vector is defined as

$$\mathbf{a}_{\text{BS}}(\phi_{n,m,l}) = \left[1, e^{-j\pi\phi_{n,m,l}}, \ldots, e^{-j\pi(N_{\text{BS}}-1)\phi_{n,m,l}}\right]^T, \tag{3}$$

where $\phi_{n,m,l} \in [-1, 1]$ is related to the AoD $\varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ as $\phi_{n,m,l} = \frac{2D\sin(\varphi)}{\lambda}$. Note that D denotes the antenna spacing and λ denotes the wavelength of the propagation. The antenna array response vector for $\mathbf{a}_{\text{UE}}(\theta_{n,m,l})$ can be written in a similar fashion. AoD/AoA variations over the coherence time are trivial and can be ignored [34]. Let T and T_b denote the coherence time and the time duration over which AoD/AoA remain unchanged, respectively. In [34], it is shown that the coherence time duration is far less than T_b, i.e., $T \ll T_b$, which ensures that AoD/AoA do not change over the coherence time. In this paper, the channel gain captures path loss and shadow fading effects. The assumption on AoD/AoA and the channel gain state that the channel model in (1) represents a long term channel which is widely adopted in the literature [24], [31], [34]. Ignoring AoD/AoA variations and instantaneous channel fluctuations are valid assumptions since the power allocation and the beamwidth control are done over the

coherence time. This follows from the fact that the long term channel model can be effectively used in long term resource allocation [34].

It is demonstrated that in dense urban environments, with high probability, the mmWave channels contain only one or two paths, with the dominant one that carries most of the signal energy [35]. Therefore, with a single path assumption, the MPC model described in (1) is converted to a single path channel model given by

$$\mathbf{H}_{n,m} = \beta_{n,m}\mathbf{a}_{\text{UE}}(\theta_{n,m})\mathbf{a}^{\dagger}_{\text{BS}}(\phi_{n,m}). \tag{4}$$

Hence, the BS communicates to the users through a single path channel. It is worth mentioning that the users are ordered based on their channel gain, i.e., $\beta_{n,1} \geq, \ldots, \geq \beta_{n,M}$ where $\beta_{n,m}$ is captured through channel quality indicator (CQI) [36]. Although it is assumed that the channel is single path, in some rare cases there might be more than one dominant path. To mitigate the multipath issue, rake receivers or orthogonal frequency-division multiplexing (OFDM) can be used. It should also be mentioned that due to the availability of large bandwidth, in mmWave systems, wide band transmission is preferred. For this case, the considered narrow band system should be combined with OFDM. In general, the extension of our narrow band system to the wide band is straightforward and studied in the literature. For instance, the OFDM-based NOMA has been considered in [12] and other similar work.

B. Beam-Training

Each transmission frame in mmWave directional communications depends on the channel coherence time and consists of two parts: (i) beam-training and (ii) data transmission as depicted in Fig. 2. At the first step, the channel parameters AoDs, AoAs, and effective channel are estimated by channel estimation algorithms. In this paper, we assume that the channel parameters are perfectly estimated [24], [32]. In particular, the estimation of AoDs and AoAs is performed using beam alignment algorithms and takes much more time compared to the effective channel estimation. The beam alignment algorithms should be fast, accurate, and energy-efficient. At the second step, during the remaining time, the data is transmitted. Recently, a few codebook-based beam-training algorithms have been proposed for mmWave hybrid beamforming systems [22], [23], [25], [26], [36]. Even the current fastest algorithms take a considerable portion of the coherence time that leaves a short time for data transmission and can diminish the achievable rate of a user [37]. On the other hand, a smaller beam-training duration means wider beamwidth, which supplies lower beamforming gain. Consequently, in the mmWave systems, there exists a

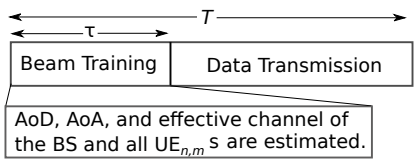

Fig. 2: Schematic of the transmission frame in the mmWave-NOMA system.

trade-off between the training duration and data transmission duration. This trade-off becomes more notable in the mmWave-NOMA networks in which more channels should be estimated. Motivated by this, finding an optimal beam-training duration and user power allocation for data transmission to increase the sum-rate of the mmWave-NOMA system will be the subject of this paper.

As mentioned before, the main part of a beam-training algorithm consists of beam alignment followed by an effective channel estimation. In general, there are two different search algorithms for beam alignment, exhaustive search [22] and hierarchical search [23]. The former algorithm examines all beam pairs in the codebook for BS and UE and determines the best pair that maximizes the beamforming gain. The training time for this algorithm is proportional to the size of the beam's search space which is given by

$$\tau = \left\lceil \frac{\omega_{\text{BS}}}{\eta} \right\rceil \left\lceil \frac{\omega_{\text{UE}}}{\mu} \right\rceil T_p, \tag{5}$$

where η and μ denote the beamwidths of the BS and the UE, respectively. Further, T_p is the time for pilot transmission. On the other hand, the hierarchical search algorithm is designed based on multi-level codebook designs and uses bisection beam search. At the first level, the algorithm chooses a wider beam with a low resolution which has a small beam search space. The algorithm refines the search iteratively using the next-level codebook within the subspace defined in the wider-level. At each level, the algorithm performs an exhaustive search to find the best pair. Compared to the exhaustive search, the hierarchical algorithm takes less training time with the same beam resolution and length of the pilot sequence at the cost of the higher probability of misalignment [36].

The exhaustive and hierarchical algorithms are designed only for single-user and multi-user scenarios. In particular, multi-user beam alignment algorithms assume that each user has a distinct AoD and might not be efficient for NOMA systems in which users are allowed to have the same AoD. Particularly, the hierarchical algorithm has a higher probability of beam misalignment at the low SNR regime [36]. This can be a major barrier in realizing the hierarchical algorithm in mmWave-NOMA networks, where the users with low SNR are paired with the users with high SNR. It seems that the exhaustive algorithm is a proper candidate for the beam alignment in the

mmWave-NOMA system since it works better at the low SNR regime [36]. In the exhaustive search algorithm, all beams are aligned with the same resolution. That is, the beamwidth of the beams at the BS is equal. In some scenarios, this may impose a limitation on designing an optimal mmWave-NOMA system. To overcome this issue, we adopt a multi-user tone-based beam-training algorithm proposed in [26]. The algorithm consists of three steps summarized as follows. At the first step, each user transmits a pilot using one omni-directional antenna with a unique frequency tone in the uplink. Given a predefined resolution $\frac{\omega_{\text{BS}}}{\eta}$ for each user, the BS searches for the best AoD that maximizes the beamforming gain. It is worth mentioning that the BS can estimate the AoDs with different predefined resolutions. At the second step, using the estimated AoDs, the BS simultaneously transmits a pilot for each user over a unique frequency tone in the downlink. Each user estimates the AoA with a predefined resolution $\frac{\omega_{\text{UE}}}{\mu}$. Finally, each user transmits an orthogonal pilot sequence to the BS, and the BS estimates the channel. We note that using a unique tone for each user requires more hardware complexity compared to the search algorithm. There are two main differences between the tone-based algorithm and the exhaustive search algorithm. First, due to using unique frequency tones, the beam alignment for each user is done independently. Hence, the BS can select different beamwidth values for different users and each user can also have a distinct beamwidth value. Second, the beam-training time is shorter than those of the exhaustive search algorithm. That is, the total training time for the tone-based algorithm is $\tau = \max\{(\lceil\frac{\omega_{\text{BS}}}{\eta_n}\rceil + \lceil\frac{\omega_{\text{UE}}}{\mu_{n,m}}\rceil)T_p\}$, where η_n and $\mu_{n,m}$ denote the beamwidth of the nth beam of the BS and the $\text{UE}_{n,m}$, respectively. It is clear when η_n and $\mu_{n,m}$ are the same as those of the exhaustive search algorithm, the training time for the tone-based algorithm is smaller than (5).

Although the algorithm in [26] is applicable to the mmWave-NOMA structure and can remarkably reduce the training time, similar to the hierarchical algorithm, it may result in a higher probability of misalignment. This is due to the use of omni-directional antennas at the first step which does not provide enough beamforming gain, especially for low-SNR users. To tackle this challenge, we modify the algorithm at the cost of sacrificing the speed of beam-training. We assume that each user steers directional beams with the predefined beamwidth μ. Then, we combine the first and second steps and perform an exhaustive beam search to find the best beam pair that achieves the highest beamforming gain. Note that the BS communicates with each user via a unique frequency tone. Further, the third step remains unchanged. Therefore, the training time becomes $\tau = \max\{\lceil\frac{\omega_{\text{BS}}}{\eta_n}\rceil\lceil\frac{\omega_{\text{UE}}}{\mu_{n,m}}\rceil T_p\}$. When beamwidth for the BS and users are the same as

those of (5), the beam-training time for the modified tone-based algorithm is similar to that of the exhaustive search algorithm. In summary, we adopt the tone-based beam alignment algorithm in [26] and instead of hierarchical search we use exhaustive search.

C. Data Transmission

In mmWave-NOMA systems, during the data transmission, the transmit symbols are superposition coded at the BS. Then, at the user side, unintended symbols are removed via SIC. More details on these two processes are provided as follows. Let $\mathbf{s} \in \mathbb{C}^{N_{\text{RF}} \times 1}$ denote the information signal vector such that its nth element s_n satisfies $\mathbb{E}\left[s_n s_n^*\right] = \frac{1}{N_{\text{RF}}}$ for $n = 1, 2, \ldots, N_{\text{RF}}$. At the baseband of the BS, the superposition coded signal of the nth stream is given by $s_n = \sum_{m=1}^{K_n} \sqrt{P_{n,m}} z_{n,m}$ where $P_{n,m}$ and $z_{n,m}$ are the allocated power and transmit symbol for the mth user in the nth cluster, respectively. Then, the hybrid beamforming is done in digital and analog precoding stages. The BS applies the digital precoder $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{N_{\text{RF}} \times N_{\text{RF}}}$ using RF chains, and then applies the analog precoder $\mathbf{F}_{\text{RF}} \in \mathbb{C}^{N_{\text{BS}} \times N_{\text{RF}}}$ using phase-shifters. Thus, the transmit signal vector after superposition coding and beamforming, $\mathbf{x} \in \mathbb{C}^{N_{\text{BS}} \times 1}$, is expressed as

$$\mathbf{x}^T = \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}} \mathbf{s}^T. \tag{6}$$

Each element of all beamforming vectors has a constant magnitude of $\frac{1}{\sqrt{N_{\text{BS}}}}$. Further, the total power of the hybrid beamforming is constrained to $\left\|\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}\right\|_F^2 = N_{\text{RF}}$. On the other hand, the received signal by $\text{UE}_{n,m}$, $\mathbf{r}_{n,m} \in \mathbb{C}^{N_{\text{UE}} \times 1}$, is given by

$$\mathbf{r}_{n,m} = \mathbf{H}_{n,m} \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}} \mathbf{s} + \mathbf{n}_{n,m}, \tag{7}$$

where $\mathbf{n}_{n,m} \in \mathbb{C}^{N_{\text{UE}} \times 1}$ is the additive white Gaussian noise vector with zero-mean and σ^2 variance for each element, i.e., $\mathcal{CN}(0, \sigma^2)$. Then, the received vector at $\text{UE}_{n,m}$ followed by the analog combiner $\mathbf{w}_{n,m} \in \mathbb{C}^{N_{\text{UE}} \times 1}$ is obtained as

$$\begin{aligned} y_{n,m} = & \underbrace{\sqrt{P_{n,m}} \mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB}}^{n} z_{n,m}}_{\text{desired signal}} + \underbrace{\sum_{k \neq m}^{K_n} \sqrt{P_{n,k}} \mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB}}^{n} z_{n,k}}_{\text{intra-cluster interference}} \\ & + \underbrace{\sum_{q \neq n}^{N_{\text{RF}}} \sum_{\ell=1}^{K_q} \sqrt{P_{q,\ell}} \mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB}}^{\ell} z_{q,\ell}}_{\text{inter-cluster interference}} + \underbrace{\mathbf{w}_{n,m}^{\dagger} \mathbf{n}_{n,m}}_{\text{noise}}. \end{aligned} \tag{8}$$

Each user decodes the intended signal by using SIC. As such, after applying SIC, the received signal at $\text{UE}_{n,1}$ is given by

$$y_{n,1} = \sqrt{P_{n,1}}\mathbf{w}_{n,1}^{\dagger}\mathbf{H}_{n,1}\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB}}^{n}z_{n,1} + \sum_{q\neq n}^{N_{\text{RF}}}\sum_{\ell=1}^{K_q}\sqrt{P_{q,\ell}}\mathbf{w}_{n,1}^{\dagger}\mathbf{H}_{n,1}\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB}}^{\ell}z_{q,\ell} + \mathbf{w}_{n,1}^{\dagger}\mathbf{n}_{n,1}, \tag{9}$$

and the received signal at $\text{UE}_{n,m}$, for $m > 1$, is given by

$$\begin{aligned} y_{n,m} =& \sqrt{P_{n,m}}\mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB}}^{n}z_{n,m} + \underbrace{\sum_{k=1}^{m-1}\sqrt{P_{n,k}}\mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB}}^{n}z_{n,k}}_{\text{residual intra-cluster interference}} \\ &+ \sum_{q\neq n}^{N_{\text{RF}}}\sum_{\ell=1}^{K_q}\sqrt{P_{q,\ell}}\mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB}}^{\ell}z_{q,\ell} + \mathbf{w}_{n,m}^{\dagger}\mathbf{n}_{n,m}. \end{aligned} \tag{10}$$

One can observe that the desired signal of the first user in (9) is contaminated by the inter-cluster interference and noise, whereas the desired signal of the other users represented by (10) is contaminated by the residual intra-cluster and inter-cluster interference and noise.

D. Clustering

In this section, we describe a simple, yet effective clustering method for two NOMA users per cluster which is the case in our system model. The reason for choosing two users per cluster will be explained in the next sections. Before proceeding, we define cluster-head and far UE terms. In a cluster, we call the closer UE to the BS the cluster-head and the other UE the far UE. A clustering algorithm for two NOMA users per cluster, mainly designed based on the following two key points, has been proposed in [38, Algorithm 1]: (i) A key point to maximize the sum-rate in NOMA is to ensure that the high channel gain users are selected as the cluster-heads. (ii) The channel gain difference between the cluster-head and the far UE should be sufficiently high.

Before applying the clustering method in [38], we select $2N_{\text{RF}}$ users and divide them into two groups. The first group consists of the N_{RF} users with the highest channel gains denoted by $\text{UE}_{n,1}$ for $n = 1, 2, \cdots, N_{\text{RF}}$. The second group includes the remaining N_{RF} users denoted by $\text{UE}_{n,2}$ for $n = 1, 2, \cdots, N_{\text{RF}}$. Further, the users of the first group are called the cluster-heads and the users of the second group are called the far users. The following conditions result in good performance.

Condition 1: The cluster-heads are located in distinctive directions.

Condition 2: The far users $\mathrm{UE}_{n,2}$ have the lowest channel gains among all the users and are paired with $\mathrm{UE}_{1,1}$, $\mathrm{UE}_{2,1}$, $\cdots$, $\mathrm{UE}_{N_{\mathrm{RF}},1}$, respectively.

Then, we use the clustering algorithm proposed in [38] in our mmWave-NOMA network. To make sure that Conditions 1 and 2 hold, we replace the users that violate them. Since the probability of existing high channel gain users in mmWave cells is almost one, new cluster-heads that will not violate Condition 1 are always available. Since the sum-rate is mainly determined by the channel gain of the cluster-heads, replacing the users that violate Condition 2 will not affect the sum-rate dramatically. Therefore, to ease the calculations, for the rest of the paper, it is assumed that $\mathrm{UE}_{n,1}$ and $\mathrm{UE}_{n,2}$ are clustered together.

E. Hybrid Beamforming Gain and SINR

After the clustering is performed, an efficient beamforming is used to reduce/eliminate the inter-cluster interference. We use the zero-forcing beamforming (ZFBF) method which is widely adopted in the literature [15], [19], [38]–[40]. This method is low-complex and highly efficient. In fact, it is shown that when the channels of the users inside a cluster are highly correlated, ZFBF can significantly suppress the inter-cluster interference. In ideal cases, i.e., a perfect correlation, ZFBF is able to completely eliminate the inter-cluster interference. First, we describe an ideal beamforming gain which is the same as that of an ideal ZFBF. Then, to take the practical issues into account, we describe a non-ideal beamforming gain which reflects the impact of the imperfect channel correlation in ZFBF. We note that when the channels between the users are not highly correlated, the singular value decomposition (SVD) method is used to design the beamforming matrix [15].

Let us define $\mathbf{f}_n = \mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}^n$ as the hybrid beamforming vector of the nth beam at the BS. An ideal hybrid beamformer leads to $|\mathbf{a}_{\mathrm{BS}}^{\dagger}(\phi_{n,m})\mathbf{f}_n| = \sqrt{G_{\mathrm{BS}}^{\mathrm{id}}(\phi_{n,m},\eta_n)}$ in which $G_{\mathrm{BS}}^{\mathrm{id}}$ is the beamforming gain of the ideal beamformer at the BS and η_n denotes the beamwidth of the nth beam. It is worth mentioning that in this paper the parameter $G_{\mathrm{BS}}^{\mathrm{id}}$ is irrespective of how the hybrid beamforming is designed. Essentially, the value of $G_{\mathrm{BS}}^{\mathrm{id}}$ depends on the beamformer $\mathbf{f}_n$, where $||\mathbf{f}_n|| = 1$, and the size of the transmit antenna array. Also, note that an ideal beamforming vector is obtained when there is no channel estimation error and perfect beam alignment is done while considering an infinite resolution for the phase-shifters. Further, the beamwidth depends on the design of the analog beamformer and the digital beamformer. In particular, the beamforming

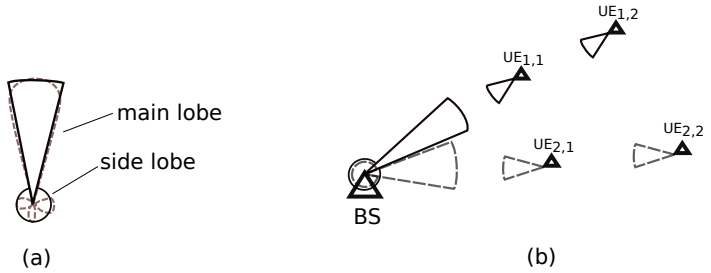

Fig. 3: (a) A non-ideal beam is modeled with a constant main lobe gain and side lobe gain, (b) The impact of the side lobe gain of each beam on the UEs located in the other cluster.

gain is defined as

$$G_{\text{BS}}^{\text{id}}(\phi_{n,m},\eta_n) = \begin{cases} \frac{2\pi}{\eta_n}, & \text{if } |\phi_{n,m}| \leq \frac{\eta_n}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Further, the beamforming gain of the ideal analog beamformer at $\text{UE}_{n,m}$ is assumed to be $|\mathbf{w}_{n,m}^{\dagger}\mathbf{a}_{\text{UE}}(\theta_{n,m})| = \sqrt{G_{\text{UE}}^{\text{id}}(\theta_{n,m},\mu_{n,m})}$ in which $G_{\text{UE}}^{\text{id}}$ is the gain of the ideal analog beamformer and $\mu_{n,m}$ denotes the beamwidth of $\text{UE}_{n,m}$. Similar to $G_{\text{BS}}^{\text{id}}$, the ideal beamforming gain is defined as

$$G_{\text{UE}}^{\text{id}}(\theta_{n,m},\mu_{n,m}) = \begin{cases} \frac{2\pi}{\mu_{n,m}}, & \text{if } |\theta_{n,m}| \leq \frac{\mu_{n,m}}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Note that ideal beamforming at the BS and users results in the complete cancellation of the inter-cluster interference represented in (8).

In practice, achieving the ideal beamforming gain may not be possible because of the application of suboptimal solutions for the analog beamformer, finite resolution of the phase-shifters, channel estimation error, and beam misalignment. These problems reduce the gain in the main lobe and introduce a side lobe for each beam. Hence, the beamforming model should take these issues into account [18], [23]. A more practical model for the beamforming gain of the nth beam is given by

$$G_{\text{BS}}(\phi_{n,m},\eta_n) = \begin{cases} \frac{2\pi-(2\pi-\eta_n)\xi}{\eta_n}, & \text{if } |\phi_{n,m}| \leq \frac{\eta_n}{2}, \\ \xi, & \text{otherwise,} \end{cases} \tag{13}$$

where $0 \leq \xi < 1$ with $\xi \ll 1$ for narrow beams, which is widely adopted in the literature [25], [29], [41]. Note that there is another common model for the beamforming gain with side lobe level varying with the beamwidth [18]. In this paper, to make the analysis tractable, we use the model described above that satisfies the total power of 1, i.e., $\int_0^{2\pi} G_{\text{BS}}(\phi_{n,m},\eta_n)\text{d}\phi_{n,m} =$

$\frac{\eta_n}{2\pi}\frac{2\pi-(2\pi-\eta_n)\xi}{\eta_n}+\frac{2\pi-\eta_n}{2\pi}\xi=1$. Similarly, the model for the beamforming gain of $\text{UE}_{n,m}$ is given by

$$G_{\text{UE}}(\theta_{n,m},\mu_{n,m})=\begin{cases}\frac{2\pi-(2\pi-\mu_{n,m})\xi}{\mu_{n,m}}, & \text{if } |\theta_{n,m}|\leq\frac{\mu_{n,m}}{2},\\ \xi, & \text{otherwise.}\end{cases} \tag{14}$$

In the above equations, the main lobe's gain is distributed uniformly in the entire beamwidth and the side lobe's gain is assumed to be constant [26], [27] as demonstrated in Fig. 3.(a). However, in reality, the main lobe's gain changes over the beamwidth and the side lobe's gain depends on the size of the beamwidth. For example, for a narrower beam, the side lobe's gain is higher [42]. Further, the side lobe results in interference that impacts the UEs located in other clusters as shown in Fig. 3.(b). In our formulation, this interference is modeled by the inter-cluster interference term in (8).

Hence, using (9), (13), and (14), the signal-to-interference-plus-noise ratio (SINR) of $\text{UE}_{n,1}$ in the nth beam is expressed as

$$\gamma_{n,1}=\frac{P_{n,1}\beta_{n,1}^2 G_{\text{BS}}(\phi_{n,1},\eta_n)G_{\text{UE}}(\theta_{n,1},\mu_{n,1})}{\sum\limits_{q\neq n}^{N_{\text{RF}}}\sum\limits_{\ell=1}^{K_q}P_{q,\ell}\beta_{n,1}^2 G_{\text{UE}}(\theta_{n,1},\mu_{n,1})\xi+\sigma^2}, \tag{15}$$

and, using (10), (13), and (14), the SINR of $\text{UE}_{n,m}$, $m>1$, is given by

$$\gamma_{n,m}=\frac{P_{n,m}\beta_{n,m}^2 G_{\text{BS}}(\phi_{n,m},\eta_n)G_{\text{UE}}(\theta_{n,m},\mu_{n,m})}{\sum\limits_{k=1}^{m-1}P_{n,k}\beta_{n,m}^2 G_{\text{BS}}(\phi_{n,m},\eta_n)G_{\text{UE}}(\theta_{n,m},\mu_{n,m})+\sum\limits_{q\neq n}^{N_{\text{RF}}}\sum\limits_{\ell=1}^{K_q}P_{q,\ell}\beta_{n,m}^2 G_{\text{UE}}(\phi_{n,m},\mu_{n,m})\xi+\sigma^2}. \tag{16}$$

Due to the single cell assumption, we can conclude that the users do not receive any interference from the side lobe and only receive signal from the main lobe. Further, it is assumed that codebooks for a specific level (beam resolution) are designed efficiently such that the steered beams by the BS do not overlap [22]. Hence, each user receives the desired signal and intra-cluster interference sent through the main lobe of the desired beam and the inter-cluster interference sent through the side lobe of the other beams as visualized in Fig. 3.(b). Further, in the modified beam-training algorithm described in Section II-B and adopted in this section, the beams directed by the BS can have different beamwidth values, i.e., the beams have different resolutions. In this case, during the data transmission, beam overlap may occur, which can impose severe inter-cluster interference. To avoid this, we assume that there is a proper angle gap between the two neighboring beams. In mmWave hybrid beamforming, a limited number

of RF chains is used, i.e., the number of beams is limited [19]. Therefore, the direction of each beam is selected to satisfy the required angle gap between the beams. Further, each cluster's users are served via a common beam directed by the BS. Therefore, the training time for $\text{UE}_{n,m}$ is given by $\tau_{n,m} = \lceil \frac{\omega_{\text{BS}}}{\eta_n} \rceil \lceil \frac{\omega_{\text{UE}}}{\mu_{n,m}} \rceil T_p$ as explained in Section II-B. Accordingly, the achievable rate for $\text{UE}_{n,m}$ can be calculated as

$$R_{n,m} = \left(1 - \frac{\tau}{T}\right) \log_2(1 + \gamma_{n,m}), \tag{17}$$

where T denotes the channel coherence time as indicated in Fig. 2. It is worth mentioning that the chosen frame duration is smaller than the channel coherence time.

III. Problem Formulation

Here, NOMA is performed for two UEs per cluster which is compatible with the multi-user superposition transmission schemes recently adopted by 3GPP [43], [44]. Further, the BS is assumed to generate only two beams. Extension to more than two clusters will be addressed in future work. To optimize the sum-rate performance, η_n, $\mu_{n,m}$, and $P_{n,m}$ should be optimized according to

$$\underset{\boldsymbol{\eta}, \boldsymbol{\mu}, \mathbf{P}}{\text{maximize}} \quad \sum_{n=1}^{2} \sum_{m=1}^{2} R_{n,m} \tag{18a}$$

$$\text{subject to} \quad \eta_{\text{min}} \leq \eta_n \leq \omega_{\text{BS}}, \tag{18b}$$

$$\mu_{\text{min}} \leq \mu_{n,m} \leq \min\{\omega_{\text{UE}}, \eta_n\}, \tag{18c}$$

$$\tau \leq T, \tag{18d}$$

$$\tau \geq \frac{\omega_{\text{BS}}}{\eta_n} \frac{\omega_{\text{UE}}}{\mu_{n,m}} T_p, \tag{18e}$$

$$R_{n,m} \geq R_{\text{min}}, \tag{18f}$$

$$\sum_{n=1}^{2} \sum_{m=1}^{2} P_{n,m} \leq P_{\text{tot}}, \tag{18g}$$

$$P_{n,m} > 0, \tag{18h}$$

where $\boldsymbol{\eta} = [\eta_1, \eta_2]$, $\boldsymbol{\mu} = [\mu_{1,1}, \mu_{1,2}, \mu_{2,1}, \mu_{2,2}]$, $\mathbf{P} = [P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2}]$, and P_{tot} denotes the total power of the BS. The smallest beamwidth resolutions for the BS and UE are denoted by η_{min} and μ_{min}, respectively. Here, we assume $\eta_{\text{min}} = \mu_{\text{min}}$. The beamwidth resolution relates to the number of antennas. Usually, the number of antennas at a BS is larger than those of UEs. Thus, the BS can generate narrower beams. However, we assume that the minimum beamwidths of the BS

and UEs are identical. For the sake of simplicity, we relax $\tau \geq \lceil \frac{\omega_{\text{BS}}}{\eta_n} \rceil \lceil \frac{\omega_{\text{UE}}}{\mu_{n,m}} \rceil T_p$ to $\tau \geq \frac{\omega_{\text{BS}}}{\eta_n} \frac{\omega_{\text{UE}}}{\mu_{n,m}} T_p$ in (18e). After we obtain the optimal $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, we can recalculate $\tau = \max\{\lceil \frac{\omega_{\text{BS}}}{\eta_n} \rceil \lceil \frac{\omega_{\text{UE}}}{\mu_{n,m}} \rceil T_p\}$.

IV. Joint beamwidth control and power allocation

Problem (18) is an intractable non-convex optimization problem and needs to be decomposed. We propose an algorithm which iterates between the power allocation and the beamwidth optimization. When allocating the power, we fix the beamwidths and when optimizing the beamwidths, we keep the powers fixed. We assume that the BS and the users are aligned after the training process, which means $|\phi_{n,m}| \leq \frac{\eta_n}{2}$ and $|\theta_{n,m}| \leq \frac{\mu_n}{2}$. We also assume the users within the same cluster have the same beamwidth, i.e., $\mu_{n,1} = \mu_{n,2} = \mu_n$. However, the users in different clusters do not necessarily have the same beamwidth.

A. Power allocation

When the beamwidth and the training time are fixed, the beamforming gains are also fixed. Then, problem (18) is simplified to:

$$\underset{\mathbf{P}}{\text{maximize}} \qquad \sum_{n=1}^{2} \sum_{m=1}^{2} R_{n,m} \tag{19a}$$

$$\text{subject to} \quad (18f)\text{–}(18h). \tag{19b}$$

Although Problem (19) has been greatly simplified compared to Problem (18), its objective is still complicated and non-convex. To transform Problem (19) into a tractable form, we use the log-exponential reformation idea in [45]. Introducing slack variables $\boldsymbol{S} = [x_{n,m}, d_{n,m}], n = 1,2, m = 1,2$, we can transform the objective in Problem (19) into a linear form by $\sum_{n=1}^{2} \sum_{m=1}^{2} \log_2 \frac{2^{x_{n,m}}}{2^{d_{n,m}}}$. For the sake of brevity, we denote $G_{\text{BS}}(\eta_n) G_{\text{UE}}(\mu_n)$ by G_n and $G_{\text{UE}}(\mu_n)$ by G_{UE}^n and make the following definitions:

$$\text{SI}_{n,1} \triangleq P_{n,1} \beta_{n,1}^2 G_n + \sum_{\ell=1, q \neq n}^{2} P_{q,\ell} \beta_{n,1}^2 G_{\text{UE}}^n \xi + \sigma^2, \tag{20}$$

$$\text{SI}_{n,2} \triangleq P_{n,2} \beta_{n,2}^2 G_n + P_{n,1} \beta_{n,2}^2 G_n + \sum_{\ell=1, q \neq n}^{2} P_{q,\ell} \beta_{n,2}^2 G_{\text{UE}}^n \xi + \sigma^2, \tag{21}$$

$$\text{I}_{n,1} \triangleq \sum_{\ell=1, q \neq n}^{2} P_{q,\ell} \beta_{n,1}^2 G_{\text{UE}}^n \xi + \sigma^2, \tag{22}$$

$$\mathrm{I}_{n,2} \triangleq P_{n,1}\beta_{n,2}^2 G_n + \sum_{\ell=1,q\neq n}^{2} P_{q,\ell}\beta_{n,2}^2 G_{\mathrm{UE}}^n \xi + \sigma^2. \tag{23}$$

Then, Problem (19) can be rewritten as

$$\begin{aligned}
\underset{\boldsymbol{S},\mathbf{P}}{\text{maximize}} \quad & (1-\frac{\tau}{T})\sum_{n=1}^{2}\sum_{m=1}^{2}(x_{n,m}-d_{n,m}) && \text{(24a)}\\
\text{subject to} \quad & 2^{x_{n,1}} \leq \mathrm{SI}_{n,1}, && \text{(24b)}\\
& 2^{x_{n,2}} \leq \mathrm{SI}_{n,2}, && \text{(24c)}\\
& 2^{d_{n,1}} \geq \mathrm{I}_{n,1}, && \text{(24d)}\\
& 2^{d_{n,2}} \geq \mathrm{I}_{n,2}, && \text{(24e)}\\
& (1-\frac{\tau}{T})(x_{n,m}-d_{n,m}) \geq R_{\min}, && \text{(24f)}\\
& (18g)-(18h). && \text{(24g)}
\end{aligned}$$

In Problem (24), the optimum is achieved when the constraints (24b)-(24e) satisfy with equality. Let us use (24b) as an example to show that the equality should be satisfied at the optimum. Assuming the opposite, we can increase $x_{n,1}$ while keeping other variables fixed. This results in increasing the cost function and contradicts the optimality assumption. Since constraints (24b)-(24e) achieve equality at the optimum, the non-convex objective of Problem (19) is equivalently decomposed into (24a) and constraints (24b)-(24e).

Unfortunately, constraints (24d) and (24e) are still non-convex. To relax the non-convex constraint to convex constraints, we use a sequential parametric convex approximation method (SPCA) [46]. In this method, the non-convex feasible set is sequentially approximated by an inner convex approximation. Using (24d) as an example, at Iteration k, since function $2^{d_{n,1}}$ is a convex function, i.e., $2^y - 2^x \geq 2^x \log 2(y-x)$, we have a lower bound of $2^{d_{n,1}}$ as:

$$2^{d_{n,1}^*[k-1]} \log 2(d_{n,1} - d_{n,1}^*[k-1]) + 2^{d_{n,1}^*[k-1]} \leq 2^{d_{n,1}}, \tag{25}$$

where $d_{n,1}^*[k-1]$ is the optimal solution at Iteration $k-1$. Based on (25), we can relax (24d) into a convex constraint as

$$2^{d_{n,1}^*[k-1]} \log 2(d_{n,1} - d_{n,1}^*[k-1]) + 2^{d_{n,1}^*[k-1]} \geq \sum_{\ell=1,q\neq n}^{2} P_{q,\ell}\beta_{n,1}^2 G_{\mathrm{UE}}^n \xi + \sigma^2. \tag{26}$$

Using the same method for (24e), we have

$$2^{d_{n,2}^*[k-1]} \log 2(d_{n,2} - d_{n,2}^*[k-1]) + 2^{d_{n,2}^*[k-1]} \geq P_{n,1}\beta_{n,2}^2 G_n + \sum_{\ell=1,q\neq n}^{2} P_{q,\ell}\beta_{n,2}^2 G_{\mathrm{UE}}^n \xi + \sigma^2. \tag{27}$$

At each iteration, we can relax Problem (24) into the following convex problem:

$$\underset{\boldsymbol{S},\mathbf{P}}{\text{maximize}} \quad (1-\frac{\tau}{T})\sum_{n=1}^{2}\sum_{m=1}^{2}(x_{n,m}-d_{n,m}) \tag{28a}$$

$$\text{subject to} \quad (24b), \tag{28b}$$

$$(24c), \tag{28c}$$

$$(26)-(27), \tag{28d}$$

$$(18g)-(18h). \tag{28e}$$

This is a convex problem, which can be efficiently solved by off-the-shelf solutions, such as CVX [47].

By relaxing Problem (24) to Problem (28) in each iteration, we can propose an iterative algorithm to provide an approximation solution for Problem (24). Detailed steps are presented in Alg. 1. According to [46], Alg. 1 converges.

Algorithm 1 Power allocation

Set the sum-rate $R_{\text{sum}}[-1] \leftarrow 0$, the maximal iteration number $k_{max} \leftarrow 1000$ and the convergence threshold $\epsilon \leftarrow 10^{-3}$;
repeat
  Choose a feasible start point $\mathbf{P}^*[0]$;
  $x^*_{n,m}[0] \leftarrow \log_2(P^*_{n,m}[0]\beta^2_{n,m}G_n)$,
  $d^*_{n,1}[0] \leftarrow \log_2(\sum_{\ell=1,q\neq n}^{2} P^*_{q,\ell}[0]\beta^2_{n,1}G^n_{\text{UE}}\xi+\sigma^2)$;
  $d^*_{n,2}[0] \leftarrow \log_2(P^*_{n,1}[0]\beta^2_{n,2}G_n+\sum_{\ell=1,q\neq n}^{2} P^*_{q,\ell}[0]\beta^2_{n,2}G^n_{\text{UE}}\xi+\sigma^2)$;
until (24f) is satisfied
$k \leftarrow 0$;
while $R_{\text{sum}}[k]-R_{\text{sum}}[k-1] \geq \epsilon R_{\text{sum}}[k-1]$ and $k \leq k_{max}$ **do**
  $k \leftarrow k+1$;
  Solve (28) to obtain $\boldsymbol{S}^*[k]$ and $\mathbf{P}^*[k]$;
end while
Return $\mathbf{P}^*[k]$.

B. Beamwidth optimization

When the powers are fixed, the problem to optimize the beamwidth can be rewritten as

$$\underset{\boldsymbol{\eta},\,\boldsymbol{\mu}}{\text{maximize}} \quad \sum_{n=1}^{2}\sum_{m=1}^{2}R_{n,m} \tag{29a}$$

$$\text{subject to} \quad (18b)-(18f). \tag{29b}$$

Problem (29) is a very complicated problem with non-convex objective and constraints. To simplify the problem, we first perform the following variable substitutions:

$$\mu_n = \frac{A}{f_n}, \tag{30}$$

$$\eta_n = \frac{A}{h_n}, \tag{31}$$

where $f_n \triangleq G_{\text{UE}}(\mu_n) - \xi$, $h_n \triangleq G_{\text{BS}}(\eta_n) - \xi$ and $A = 2\pi - 2\pi\xi$ is a constant. By assuming $G_{\text{BS}}(\eta_n)G_{\text{UE}}(\mu_n) \approx \frac{(2\pi-2\pi\xi)^2}{\eta_n\mu_n} + \xi^2$, we have $\eta_n\mu_n = \frac{A^2}{h_n f_n}$ and $G_{\text{BS}}(\eta_n)G_{\text{UE}}(\mu_n) = h_n f_n + \xi^2$. Then, the SINR for $\text{UE}_{n,1}$ and $\text{UE}_{n,2}$ can be rewritten as

$$\gamma_{n,1} = \frac{P_{n,1}\beta_{n,1}^2(h_n f_n + \xi^2)}{\sum_{\ell=1,q\neq n}^{2} P_{q,\ell}\beta_{n,1}^2(f_n + \xi)\xi + \sigma^2}, \tag{32}$$

$$\gamma_{n,2} = \frac{P_{n,2}\beta_{n,2}^2(h_n f_n + \xi^2)}{P_{n,1}\beta_{n,2}^2(h_n f_n + \xi^2) + \sum_{\ell=1,q\neq n}^{2} P_{q,\ell}\beta_{n,2}^2(f_n + \xi)\xi + \sigma^2}. \tag{33}$$

Instead of finding the optimal beamwidth, we find the optimal f_n and h_n. We can rewrite Problem (29) as

$$\underset{\boldsymbol{h},\,\boldsymbol{f}}{\text{maximize}} \quad \sum_{n=1}^{2}\sum_{m=1}^{2} R_{n,m} \tag{34a}$$

$$\text{subject to} \quad \frac{A}{\omega_{\text{BS}}} \leq h_n \leq \frac{A}{\eta_{\text{min}}}, \tag{34b}$$

$$\frac{A}{\omega_{\text{UE}}} \leq f_n \leq \frac{A}{\mu_{\text{min}}}, \tag{34c}$$

$$h_n \leq f_n, \tag{34d}$$

$$R_{n,m} \geq R_{\text{min}}, \tag{34e}$$

$$\tau \leq T, \tag{34f}$$

$$\tau = \max\{\frac{\omega_{\text{BS}}\omega_{\text{UE}} h_n f_n}{A^2} T_p\}, \tag{34g}$$

where $\boldsymbol{h} = [h_1, h_2]$ and $\boldsymbol{f} = [f_1, f_2]$. Problem (34) is still intractable with a non-convex objective. To further decompose the problem, we will iterate between the two clusters, i.e., we first fix Cluster 2 to optimize the beamwidths in Cluster 1 and then fix Cluster 1 to optimize the beamwidths in Cluster 2.

C. Optimal beamwidth search for each cluster

Let us assume the beamwidths of Cluster 2 are fixed and optimize the beamwidths in Cluster 1 as an example. The optimization for Cluster 2 is similar. In this case, the SINRs of $\text{UE}_{2,1}$ and

$\text{UE}_{2,2}$ are fixed. We denote them by $\gamma_{2,1}^{\text{fix}}$ and $\gamma_{2,2}^{\text{fix}}$. We also denote the corresponding variables f_2 and h_2 as f_2^{fix} and h_2^{fix}. Then, the beamwidth optimization problem for Cluster 1 is as follows:

$$\underset{h_1,\, f_1}{\text{maximize}} \quad \sum_{m=1}^{2} R_{1,m} + (1-\frac{\tau}{T}) \sum_{m=1}^{2} \log_2(1+\gamma_{2,m}^{\text{fix}}) \tag{35a}$$

$$\text{subject to} \quad \frac{A}{\omega_{\text{BS}}} \leq h_1 \leq \frac{A}{\eta_{\text{min}}}, \tag{35b}$$

$$\frac{A}{\omega_{\text{UE}}} \leq f_1 \leq \frac{A}{\mu_{\text{min}}}, \tag{35c}$$

$$h_1 \leq f_1, \tag{35d}$$

$$R_{1,m} \geq R_{\text{min}}, \tag{35e}$$

$$(1-\frac{\tau}{T}) \log_2(1+\gamma_{2,m}^{\text{fix}}) \geq R_{\text{min}}, \tag{35f}$$

$$\tau \leq T, \tag{35g}$$

$$\tau = \max\{\frac{\omega_{\text{BS}}\omega_{\text{UE}} h_1 f_1}{A^2} T_p, \frac{\omega_{\text{BS}}\omega_{\text{UE}} h_2^{\text{fix}} f_2^{\text{fix}}}{A^2} T_p\}. \tag{35h}$$

To simplify Problem (35), we discuss how to pick the optimal value for τ and remove it from the objective function. There are two cases for the optimal τ:

- Case 1: $h_1 f_1 < h_2^{\text{fix}} f_2^{\text{fix}}$. In this case, the τ should be set to $\tau^* = \frac{\omega_{\text{BS}}\omega_{\text{UE}} h_2^{\text{fix}} f_2^{\text{fix}}}{A^2} T_p$. Then, the objective function should be $(1-\frac{\tau^*}{T}) \sum_{n=1}^{2} \sum_{m=1}^{2} \log_2(1+\gamma_{n,m})$.
- Case 2: $h_1 f_1 \geq h_2^{\text{fix}} f_2^{\text{fix}}$. In this case, the τ should be set according to the value of $h_1 f_1$, which is $\tau = \frac{\omega_{\text{BS}}\omega_{\text{UE}} h_1 f_1}{A^2}$. Then, the objective function should be $(1-\frac{\omega_{\text{BS}}\omega_{\text{UE}} h_1 f_1 T_p}{A^2 T}) \sum_{n=1}^{2} \sum_{m=1}^{2} \log_2(1+\gamma_{n,m})$.

Since the solution for the two cases are different and the objective function may change, the search for the optimal beamwidths is complicated and needs to be simplified. To simplify, first, we introduce a variable $g_n = h_n f_n$. Then, the SINR $\text{UE}_{n,1}$ and $\text{UE}_{n,2}$ can be rewritten as

$$\gamma_{n,1} = \frac{P_{n,1}\beta_{n,1}^2 (g_n + \xi^2)}{\sum_{\ell=1, q\neq n}^{2} P_{q,\ell}\beta_{n,1}^2 (f_n + \xi)\xi + \sigma^2}, \tag{36}$$

$$\gamma_{n,2} = \frac{P_{n,2}\beta_{n,2}^2 (g_n + \xi^2)}{P_{n,1}\beta_{n,2}^2 (g_n + \xi^2) + \sum_{\ell=1, q\neq n}^{2} P_{q,\ell}\beta_{n,2}^2 (f_n + \xi)\xi + \sigma^2}. \tag{37}$$

Since we fix the parameters for Cluster 2, $\log_2(1+\gamma_{2,1}) + \log_2(1+\gamma_{2,2})$ is a constant, which we denote by C. Then, we define a function $F(g_1, f_1) \triangleq \log_2(1+\gamma_{1,1}) + \log_2(1+\gamma_{1,2}) + C$ which has the following property:

Proposition 1. *For $F(g_1, f_1)$ with its domain defined by $(f_1,g_1) \in [lb_f,ub_f] \times [lb_g, \min\{f_1^2,ub_g\}]$, $0 < lb_g \leq lb_f^2$ and $0 < ub_g \leq ub_f^2$, the maximum point lies on the boundary $g_1 = f_1^2$, $f_1 \in [lb_f,\sqrt{ub_g}]$.*

Proof. See Appendix A. □

Proposition 1 implies that if we want to find the maximum point of $F(g_1,f_1)$, we only need to search on the boundary $g_1 = f_1^2,f_1 \in [lb_f,ub_f]$. This simplifies $F(g_1,f_1)$ to $F(f_1^2,f_1)$. We further define $F_b(f_1) \triangleq F(f_1^2,f_1)$. Then, to find the maximum point of $F(g_1,f_1)$, we can perform a line search for $F_b(f_1)$ on $f_1 \in [lb_f,ub_f]$.

Next, we define the function $G(g_1,f_1) \triangleq (1 - \frac{\omega_{\text{BS}}\omega_{\text{UE}}g_1}{A^2}T_p)F(g_1,f_1)$. Function $G(g_1,f_1)$ has the following property:

Proposition 2. *For $G(g_1, f_1)$ with its domain defined by $(f_1,g_1) \in [\sqrt{lb_g},ub_f]\times[lb_g, \min\{f_1^2,ub_g\}]$, $ub_f,lb_g,ub_g > 0$, $ub_g < ub_f^2$, the maximum point lies on the boundary $g_1 = f_1^2,f_1 \in [\sqrt{lb_g},\sqrt{ub_g}]$.*

Proof. See Appendix B. □

Proposition 2 implies that if we want to find the maximum point of $G(g_1,f_1)$, we only need to search on the boundary $g_1 = f_1^2,f_1 \in [lb_f,ub_f]$. This simplifies $G(g_1,f_1)$ into $G(f_1^2,f_1)$. We define $G_b(f_1) \triangleq G(f_1^2,f_1)$. Then, to find the maximum point of $G(g_1,f_1)$, we can perform the line search for $G_b(f_1)$ on $f_1 \in [lb_f,ub_f]$.

To find the maximum point for Problem (35), we plot its feasible region with boundaries colored in green and blue in Figs. 4 and 5. According to Propositions 1 and 2, the maximum point lies on the blue boundary and we only need to search on the blue boundary. However, the objective function varies along the blue boundary. To conduct an effective search, we need to divide the blue boundary into two different subsets. Moreover, different initial conditions lead to different division strategies. There are two cases:

- Case 1: $g_1^{(0)} < g_2^{\text{fix}}$, where $g_1^{(0)}$ is the initial point. In this case, along the blue boundary, when we increase g_1 from $g_1^{(0)}$ to g_2^{fix}, the objective function is $(1 - \frac{\tau^*}{T})F_b(f_1)$. If we continue to increase g_1, the objective function changes to $G_b(f_1)$. Then, the blue boundary is divided as shown in Fig. 4. On Subset a, we perform a line search over $(1 - \frac{\tau^*}{T})F_b(f_1)$ to find a maximum point $(f_1^{(F)},(f_1^{(F)})^2)$. On Subset b, we perform a line search over $G_b(f_1)$ to find the maximum point $(f_1^{(G)},(f_1^{(G)})^2)$. Then, we compare the values of $(1 - \frac{\tau^*}{T})F_b(f_1^{(F)})$ and $G_b(f_1^{(G)})$, to pick the larger one as the optimal solution (f_1^*,g_1^*).

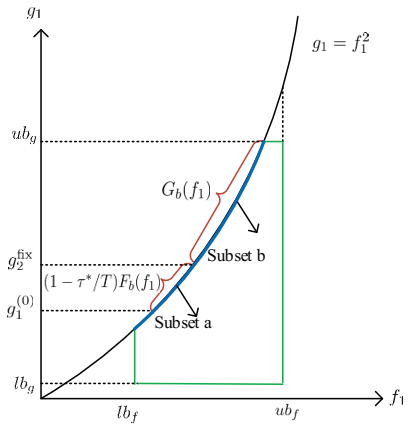

Fig. 4: The search region division for Case 1.

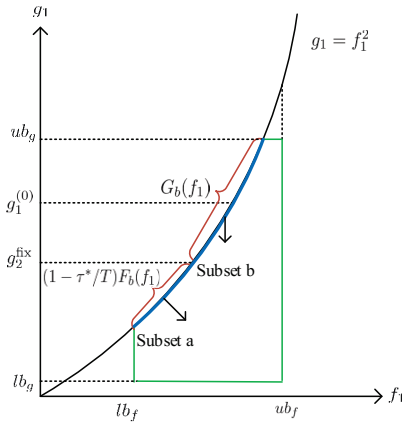

Fig. 5: The search region division for Case 2.

- Case 2: $g_1^{(0)} \geq g_2^{\text{fix}}$, where $g_1^{(0)}$ is the initial point. In this case, when we decrease g_1 to be less than g_2^{fix}, the objective function changes from $G_b(f_1)$ to $(1-\frac{\tau^*}{T})F_b(f_1)$. Then, the blue boundary is divided as shown in Fig. 5. On Subset a, we perform a line search over $(1-\frac{\tau^*}{T})F_b(f_1)$ to find a maximum point $(f_1^{(F)},(f_1^{(F)})^2)$. On Subset b, we perform a line search over $G_b(f_1)$ to find the maximum point $(f_1^{(G)},(f_1^{(G)})^2)$. Then, we compare the values of $(1-\frac{\tau^*}{T})F_b(f_1^{(F)})$ and $G_b(f_1^{(G)})$, to pick the larger one as the optimal solution (f_1^*,g_1^*).

While performing the line search, we also need to consider the minimum rate constraint. The details of line search for $(1-\frac{\tau^*}{T})F_b(f_1)$ and $G_b(f_1)$ are described in Alg. 2 and Alg. 3, respectively.

Algorithm 2 Line search over $(1-\frac{\tau^*}{T})F_b(f_1)$

1: Input τ^*, the search interval $[f_{\min},f_{\max}]$ and the step size Δ;
2: Initialize $R_{\max} \leftarrow 0$ and $f_1^{(F)} \leftarrow f_{\min}$;
3: **for** $f_1 = f_{\min} : \Delta : f_{\max}$ **do**
4: **if** $(1-\frac{\tau^*}{T})F_b(f_1) > R_{\max}$ and $(1-\frac{\tau^*}{T})\log_2(1+\gamma_{1,1}) \geq R_{\min}$ and $(1-\frac{\tau^*}{T})\log_2(1+\gamma_{1,2}) \geq R_{\min}$ **then**
5: $R_{\max} \leftarrow (1-\frac{\tau^*}{T})F_b(f_1)$;
6: $f_1^{(F)} \leftarrow f_1$;
7: **end if**
8: **end for**
9: Return $R_{\max}$, $f_1^{(F)}$.

D. Joint beamwidth optimization

For a fixed power allocation, our final beamwidth search algorithm, iterating between the beamwidth search for the two clusters, is presented in Alg. 4.

Algorithm 3 Line search over $G_b(f_1)$

Input the search interval $[f_{\min},f_{\max}]$ and the step size Δ;
Initialize $R_{\max} \leftarrow 0$ and $f_1^{(G)} \leftarrow f_{\min}$;
for $f_1 = f_{\min} : \Delta : f_{\max}$ **do**
    $\tau \leftarrow \frac{\omega_{\text{BS}}\omega_{\text{UE}} f_1^2}{A^2} T_p$
    if $G_b(f_1) > R_{\max}$ and $(1-\frac{\tau}{T})\log_2(1+\gamma_{1,m}) \geq R_{\min}$ and $(1-\frac{\tau}{T})\log_2(1+\gamma_{2,m}^{\text{fix}}) \geq R_{\min}$ **then**
        $R_{\max} \leftarrow G_b(f_1)$;
        $f_1^{(G)} \leftarrow f_1$;
    end if
end for
Return $R_{\max}$, $f_1^{(G)}$.

Algorithm 4 Joint Beamwidth Optimization

Set the sum-rate $R_{sum}[-1] \leftarrow 0$, the maximal iteration number $k_{max} \leftarrow 1000$ and the convergence threshold $\epsilon \leftarrow 10^{-3}$;
Input: $R_{sum}[0]$, μ_2,η_2,$\mu_1^{(0)}$,$\eta_1(0)$, ω_{BS}, ω_{UE}, $\mu_{\min}$ and $\eta_{\min}$;
Calculate $f_1^{(0)}$,$g_1^{(0)}$,$f_2^{(0)}$ and $g_2^{(0)}$ according to (30), (31) and $g_n = f_n h_n$;
Calculate lb_f, ub_f, lb_g and ub_g based on ω_{BS}, ω_{UE}, $\mu_{\min}$ and $\eta_{\min}$;
$k \leftarrow 0$;
while $R_{sum}[k] - R_{sum}[k-1] \geq \epsilon R_{sum}[k-1]$ and $k \leq k_{max}$ **do**
    $k \leftarrow k+1$;
    if $g_1^{(k-1)} < g_2^{(k-1)}$ **then**
        $\tau^* \leftarrow \frac{\omega_{\text{BS}}\omega_{\text{UE}} g_2^{(k-1)} T_p}{A^2}$;
        Do line search for $(1-\frac{\tau^*}{T})F_b(f_1)$ on interval $f_1 \in [\sqrt{g_1^{(k-1)}},\sqrt{g_2^{(k-1)}}]$ using Alg. 2 to get the maximum point $(f_1^{(F)},(f_1^{(F)})^2)$;
        Do line search for $G_b(f_1)$ on interval $f_1 \in [\sqrt{g_2^{(k-1)}},\sqrt{ub_g}]$ using Alg. 3 to get the maximum point $(f_1^{(G)},(f_1^{(G)})^2)$;
        if $F_b(f_1^{(F)}) == 0$ and $G_b(f_1^{(G)}) == 0$ **then**
            $(f_1^{(k)},g_1^{(k)}) \leftarrow (f_1^{(k-1)},g_1^{(k-1)})$;
        end if
        Compare the value of $(1-\frac{\tau^*}{T})F_b(f_1^{(F)})$ and $G_b(f_1^{(G)})$, and pick the larger one as the optimal solution $(f_1^{(k)},g_1^{(k)})$;
    else if $g_1^{(k-1)} \geq g_2^{(k-1)}$ **then**
        Do line search for $G_b(f_1)$ on interval $f_1 \in [\sqrt{g_2^{(k-1)}},\sqrt{ub_g}]$ using Alg. 3 to get the maximum point $(f_1^{(G)},(f_1^{(G)})^2)$;
        $\tau^* \leftarrow \frac{\omega_{\text{BS}}\omega_{\text{UE}} g_2^{(k-1)} T_p}{A^2}$;

Do line search for $(1-\frac{\tau^*}{T})F_b(f_1)$ on interval $f_1 \in [lb_f, \sqrt{g_2^{(k-1)}}]$ using Alg. 2 to get the maximum point $(f_1^{(F)}, (f_1^{(F)})^2)$;
if $F_b(f_1^{(F)}) == 0$ and $G_b(f_1^{(G)}) == 0$ **then**
$(f_1^{(k)}, g_1^{(k)}) \leftarrow (f_1^{(k-1)}, g_1^{(k-1)})$;
end if
Compare the value of $(1-\frac{\tau^*}{T})F_b(f_1^{(F)})$ and $G_b(f_1^{(G)})$, and pick the larger one as the optimal solution $(f_1^{(k)}, g_1^{(k)})$;
end if
if $g_2^{(k-1)} < g_1^{(k)}$ **then**
$\tau^* \leftarrow \frac{\omega_{\text{BS}}\omega_{\text{UE}} g_1^{(k)} T_p}{A^2}$;
Do line search for $(1-\frac{\tau^*}{T})F_b(f_2)$ on interval $f_2 \in [\sqrt{g_2^{(k-1)}}, \sqrt{g_1^{(k)}}]$ using Alg. 2 to get the maximum point $(f_2^{(F)}, (f_2^{(F)})^2)$;
Do line search for $G_b(f_2)$ on interval $f_2 \in [\sqrt{g_1^{(k)}}, \sqrt{ub_g}]$ using Alg. 3 to get the maximum point $(f_2^{(G)}, (f_2^{(G)})^2)$;
if $F_b(f_2^{(F)}) == 0$ and $G_b(f_2^{(G)}) == 0$ **then**
$(f_2^{(k)}, g_2^{(k)}) \leftarrow (f_2^{(k-1)}, g_2^{(k-1)})$;
end if
Compare the value of $(1-\frac{\tau^*}{T})F_b(f_2^{(F)})$ and $G_b(f_2^{(G)})$, and pick the larger one as the optimal solution $(f_2^{(k)}, g_2^{(k)})$;
else if $g_2^{(k-1)} \geq g_1^{(k)}$ **then**
Do line search for $G_b(f_2)$ on interval $f_2 \in [\sqrt{g_1^{(k)}}, \sqrt{ub_g}]$ using Alg. 3 to get the maximum point $(f_2^{(G)}, (f_2^{(G)})^2)$;
$\tau^* \leftarrow \frac{\omega_{\text{BS}}\omega_{\text{UE}} g_1^{(k)} T_p}{A^2}$;
Do line search for $(1-\frac{\tau^*}{T})F_b(f_2)$ on interval $f_1 \in [lb_f, \sqrt{g_1^{(k)}}]$ using Alg. 2 to get the maximum point $(f_2^{(F)}, (f_2^{(F)})^2)$;
if $F_b(f_2^{(F)}) == 0$ and $G_b(f_2^{(G)}) == 0$ **then**
$(f_2^{(k)}, g_2^{(k)}) \leftarrow (f_2^{(k-1)}, g_2^{(k-1)})$;
end if
Compare the value of $(1-\frac{\tau^*}{T})F_b(f_2^{(F)})$ and $G_b(f_2^{(G)})$, and pick the larger one as the optimal solution $(f_2^{(k)}, g_2^{(k)})$;
end if
Calculate $R_{sum}[k]$;
end while
Calculate μ_n and η_n based on f_n, h_n, where $h_n = g_n/f_n$;
Return μ_n^* and η_n^*.

E. The joint algorithm

Based on Alg. 1 and Alg. 4, we can propose a joint optimization algorithm, which iterates between the power allocation and the beamwidth optimization. The details of the algorithm are described in Alg. 5.

Algorithm 5 Joint optimization

1: Set the sum-rate $R_{sum}[-1] \leftarrow 0$, the maximal iteration number $n_{max} \leftarrow 1000$ and the convergence threshold $\epsilon \leftarrow 10^{-3}$;
2: Choose a feasible start point $\mathbf{P}^*[0]$, $\boldsymbol{\mu}^*[0]$, $\boldsymbol{\eta}^*[0]$ and $\boldsymbol{\phi}^*[0]$;
3: $n \leftarrow 0$;
4: **while** $R_{sum}[n] - R_{sum}[n-1] \geq \epsilon R_{sum}[n-1]$ and $n \leq n_{max}$ **do**
5: $\quad n \leftarrow n + 1$;
6: $\quad$ Search the optimal beamwidth using Alg. 4 with $\mathbf{P}^*[n]$ to obtain $\boldsymbol{\mu}^*[n]$, $\boldsymbol{\eta}^*[n]$ and $\boldsymbol{\phi}^*[n]$;
7: $\quad$ Solve Problem (19) using Alg. 1 with $\boldsymbol{\mu}^*[n-1]$, $\boldsymbol{\eta}^*[n-1]$, $\boldsymbol{\phi}^*[n-1]$ to obtain $\mathbf{P}^*[n]$;
8: **end while**
9: Return $\mathbf{P}^*[n]$, $\boldsymbol{\mu}^*[n]$, $\boldsymbol{\eta}^*[n]$.

F. Convergence and complexity analysis

To prove the convergence of Alg. 5, we first need to prove the convergence of Alg. 1 and Alg. 4. The convergence of Alg. 1 has been proved in [46]. In Alg. 4, to maximize the sum-rate, we optimize the beamwidth for one cluster while keeping the other cluster fixed. Such a step cannot decrease the sum-rate and generates a non-decreasing sequence of sum-rate values. Therefore, the convergence of Alg. 4 is guaranteed because the algorithm generates a sequence of non-decreasing sum-rates with an upper bound (the maximum sum-rate).

In Alg. 5, when allocating the power, we increase the sum-rate while keeping the beams in the feasible region. When optimizing the beamwidth, we search the feasible region for the beams to find the maximum sum-rate while guaranteeing the minimum rate constraint and keeping the powers in the feasible region. By doing so, we generate a monotonically increasing sequence with an upper bound (the maximum sum-rate), which proves the convergence.

Here, we provide the complexity analysis of the proposed algorithm. In our algorithm, we iteratively optimize the power allocation and beamwidth. In the power allocation algorithm, we use SPCA to gradually convexify the original non-convex problem. In each iteration, the complexity mainly lies in solving Problem (28). We use an off-the-shelf solution, i.e., CVX to solve Problem (28), which uses the interior-point method. The computational complexity

of CVX is $\mathcal{O}((3M_{\text{UE}})^{3.5})$, where M_{UE} is the total number of users and $3M_{\text{UE}}$ is the number of variables in Problem (28). In the beamwidth optimization, we iteratively optimize the beamwidth for each sector. In each iteration, the main complexity lies in the line-search algorithm, with the complexity $\mathcal{O}(\frac{ub_f - lb_f}{\Delta})$, ub_f and lb_f are the upper bound and lower bound of variable f in Problem (35), respectively, and Δ is the stepsize of the line-search algorithm.

V. Simulation Results

In this section, we present the simulation results of the joint power and beamwidth optimization algorithm. Four UEs are considered which are divided in two clusters each with two UEs. It is assumed that the UEs inside each cluster have different distances from the BS. Four multiple access techniques are investigated. The first technique is OMA in which UEs are served in different time slots. The second technique is a combination of OMA and NOMA called NOMA-OMA. In NOMA-OMA, UEs that belong to the same cluster are supported by a fixed-power NOMA and each cluster is supported by OMA at each time slot. The third technique is Fixed-NOMA in which all UEs are served by a fixed-power NOMA at one time slot. Finally, the fourth technique is the jointly optimized power and beamwidth NOMA system presented in Section IV, called Optimized-NOMA. For all techniques, first, the beams are trained and then the data transmission is done.

To evaluate the performance of the Optimized-NOMA, the parameters are set as follows. The minimum rate for all UEs is assumed to be $R_{\text{min}} = 0.1$ bits/s/Hz. Further, for the Fixed-NOMA, we allocate $\frac{1}{5}$ of the total power to the cluster-head and $\frac{4}{5}$ of the total power to the far UE as done in [48]. Also, the power is equally divided between the two clusters. The SNR used in the simulations indicates the transmit SNR, i.e., SNR=$\frac{P_{\text{tot}}}{\sigma^2}$, $\sigma^2 = 1$. In the first cluster, the channel gains of the near and far UEs from the BS are $\beta_{1,1}^2 = -17$dB and $\beta_{1,2}^2 = -26.5$dB. In the second cluster, the channel gains of the near and far UEs are $\beta_{2,1}^2 = -19$dB and $\beta_{2,2}^2 = -25$dB. Also, the side lobe level is constant and is given as $\xi = 0.1$. For the Optimized-NOMA, we use $\omega_{\text{BS}} = \omega_{\text{UE}} = 120^\circ$ and $\eta_{\text{min}} = \mu_{\text{min}} = 5^\circ$ unless it is mentioned otherwise. Further, the convergence threshold is set to $\epsilon = 10^{-3}$.

Fig. 6 demonstrates the performance of the sum-rate versus SNR. It is assumed that $T = 5 \times 10^3 T_p$ which indicates a large channel coherence time and $\eta = \mu_1 = \mu_2 = 10^\circ$. For all SNRs, by increasing SNR the sum-rate increases. The Optimized-NOMA achieves the highest sum-rate. Especially, at low SNRs, the performance gap is larger. For instance, at SNR $= 0$dB the gap between the Optimized-NOMA and Fixed-NOMA is more than 5 bits/s/Hz which reveals

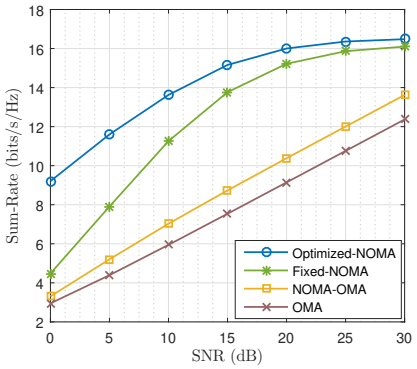

Fig. 6: Performance of the sum-rate versus SNR for a large channel coherence time, i.e., $T = 5 \times 10^3 T_p$.

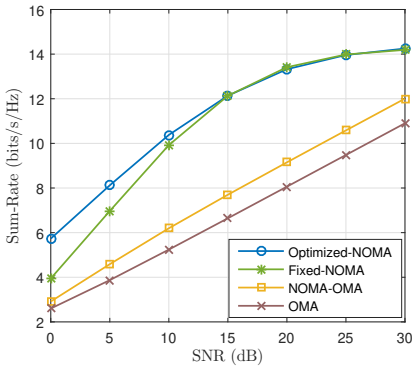

Fig. 7: Performance of the sum-rate versus SNR for a short channel coherence time $T = 1 \times 10^3 T_p$.

our joint optimization algorithm designs the powers and beamwidths very efficiently. As SNR increases, the gap decreases which is due to the fact that in the Fixed-NOMA the BS and UEs steer strong beams even if the powers and beamwidths are not optimized. The Fixed-NOMA technique preforms better than the NOMA-OMA and OMA techniques. The reason is that the Fixed-NOMA serves all the users at the same time and takes the advantages of the spectrum sharing among UEs.

In Fig. 7, we repeat the same simulation as in Fig. 6 for a relatively short channel coherence time, i.e., $T = 1 \times 10^3 T_p$. Similarly, by increasing SNR, the sum-rate increases for all the techniques. However, compared to Fig. 6, at low SNRs, the rate gap between the Optimized-NOMA and Fixed-NOMA is small. Moreover, at high SNR regions, these two techniques achieve identical sum-rates. This is because when the channel coherence time is short, the optimization algorithm does not allocate a large portion of T to the beam-training, e.g., τ is small. Thus, the optimized beamwidths are not narrow enough to provide higher gain. Also, at high SNRs, the optimized powers have trivial effects on the sum-rate compared to the predefined fixed values which is an interesting observation. This observation indicates that for a short channel coherence time like $T = 1 \times 10^3 T_p$ and high SNR, the optimization is not required and fixed-NOMA can be used instead. For a smaller coherence time, the optimized-NOMA shows better performance only at low SNRs. Nevertheless, severe path loss and shadowing in mmWave bands makes the low SNR regime very crucial. Especially, NOMA is supposed to consider near and far users, where the far users likely receive signal through NLoS low SNR channels [35]. We emphasize that at high SNRs, by increasing the coherence time, the rate gap between the optimized-NOMA and the fixed-NOMA becomes larger (See Fig. 6).

Fig. 8 shows the sum-rate performance versus the normalized channel coherence time, i.e.,

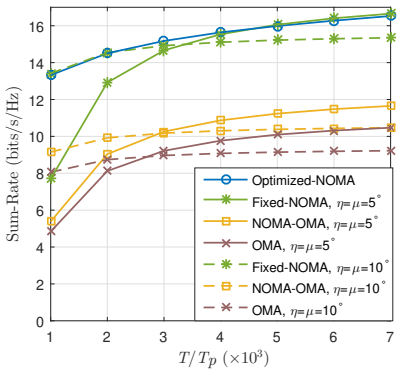

Fig. 8: Performance of the sum-rate versus T/T_p for SNR=20dB.

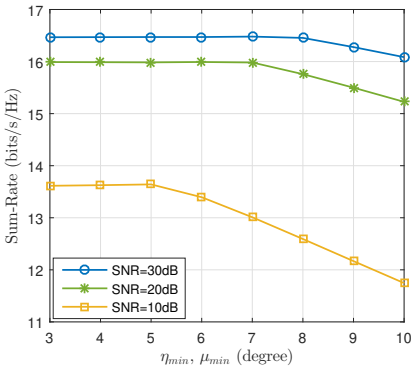

Fig. 9: Performance of the sum-rate of the Optimized-NOMA versus minimum BS and UE beamwidth (η_{min} and μ_{min}) for $T = 5 \times 10^3 T_p$ and various SNRs.

T/T_p for the moderate SNR$=$ 20dB. In this simulation, two sets of beamwidths are considered for the first three techniques: (i) $\eta = \mu = 5^\circ$ (narrow beamwidth) and (ii) $\eta = \mu = 10^\circ$ (relatively wide beamwidth). The Optimized-NOMA outperforms the other techniques for both sets of the beamwidths. For the Fixed-NOMA with $\eta = \mu = 10^\circ$ and short normalized channel coherence times, the performance is very close to that of the Optimized-NOMA. This is expected as we explained before. However, for $\eta = \mu = 5^\circ$ and short normalized channel coherence times, the sum-rate of the Fixed-NOMA is much smaller than that of the Optimized-NOMA. This is because at small T/T_p, the Fixed-NOMA assigns more time to the beam-training and leaves less time for the data transmission. As the normalized time goes up, more time is available for data transmission, and the narrow beam provides a higher sum-rate. This statement also is supported by Fig. 8 where at large channel coherence times, the Optimized-NOMA selects the minimum beamwidth. Hence, the Optimized-NOMA and the Fixed-NOMA with $\eta = \mu = 5^\circ$ achieve identical sum-rates at large normalized channel coherence times. Using Optimized-NOMA, a wide beamwidth is preferred for short T/T_p while a narrow beamwidth is preferred for large T/T_p. In Fig. 9, we simulate the performance of the sum-rate versus the minimum beamwidth of BS and UEs. The simulation is done for SNRs 10dB, 20dB, and 30dB and $T = 5 \times 10^3 T_p$. In practice, the number of antennas at the BS and UEs is limited and even for large T/T_p, a narrow beamwidth may not be generated. At high SNRs, e.g. 30dB, increasing the minimum beamwidth does not affect the sum-rate severely. As such, compared to $\eta_{\text{min}} = \mu_{\text{min}} = 3^\circ$, at 10°, the sum-rate is reduced only by 0.5 bits/s/Hz. At SNR$=$ 20dB, the sum-rate drops about 0.8 bits/s/Hz which is larger than the drop at SNR$=$ 30dB. The minimum beamwidth has a major effect at SNR$=$ 10dB. When SNR is low, narrow beams can still provide high gains

to compensate for the low SNR. As the minimum beamwidth increases and the SNR is low, the optimization algorithm cannot select narrow beams. As a result, the sum-rate dramatically decreases. In this case, simulation results indicate that decreasing the beamwidth from 3° to 10° decreases the sum-rate by about 2 bits/s/Hz.

VI. Conclusions

In this paper, NOMA is incorporated into mmWave hybrid beamforming systems. We also consider the beam-training time because of the limited channel coherence time in mmWave directional communications. By combining the exhaustive search and tone-based beam-training algorithms, a new beam-training algorithm is employed. The formulated sum-rate expression consists of the channel coherence time and beam-training time. To maximize the sum-rate, a joint power allocation and beamwidth control optimization problem is solved by an algorithm which iterates between the power allocation and the beamwidth optimization. The non-convex power allocation is solved by the log-reformulation and SPCA. The beamwidth optimization is solved by iterating between the two clusters. A boundary-search algorithm is proposed to reduce the search complexity for the beamwidth in each cluster. The numerical results demonstrate that an efficient power allocation and beam-training time can lead to higher sum-rates compared to the conventional mmWave-NOMA without optimized parameters, NOMA-OMA, and OMA. The only exception is that for a short channel coherence time and high SNR, the optimized-NOMA and the fixed-NOMA have identical sum-rate performance. Also, at low SNRs, the size of the antenna array is a major obstacle in achieving higher sum-rates.

Appendix A
Proof of Proposition 1

Since $F(g_1,f_1)$ is a continuous function defined on a bounded closed set, it has a maximum point according to the extreme value theorem. Also, according to the critical point theorem, the maximum point should either be a stationary point or a boundary point. It is easy to observe that $F(g_1,f_1)$ is a monotonic increasing function of g_1 and a monotonic decreasing function of f_1. This means $\frac{\partial F}{\partial g_1} > 0$ and $\frac{\partial F}{\partial f_1} < 0$, i.e., there is no stationary point for $F(g_1,f_1)$ on the defined domain. Then, the maximum point should lie on the five boundaries: (i) $f_1 = lb_f,\ g_1 \in [lb_g,lb_f^2]$, (ii) $g_1 = lb_g,\ f_1 \in [lb_f,ub_f]$, (iii) $f_1 = ub_f,\ g_1 \in [lb_g,ub_g]$, (iv) $g_1 = ub_g,\ f_1 \in [\sqrt{ub_g},ub_f]$, and (v) $g_1 = f_1^2,\ f_1 \in [lb_f,\sqrt{ub_g}]$.

For the boundary (i), since $F(g_1,f_1)$ is a monotonic increasing function of g_1, the maximum point can only lie on the point (lb_f,lb_f^2) which belongs to the boundary (v) as well. For the

boundary (ii), since $F(g_1,f_1)$ is a monotonic increasing function of g_1, we can pick $g_1 > lb_g$ to increase the value of $F(g_1,f_1)$. This implies that the maximum point cannot lie on the boundary (ii). Similarly, the maximum point cannot lie on the boundary (iii) either. For the boundary (iv), since $F(g_1,f_1)$ is a monotonic decreasing function of f_1, the maximum point can only lie on the point $(\sqrt{ub_g},ub_g)$ which belongs to boundary (v) as well. Note that the possible maximum points on the boundaries (i) and (iv) also belong to the boundary (v). Therefore, the maximum point must lie on the boundary (v) and the proof is complete.

Appendix B
Proof of Proposition 2

Since $G(g_1,f_1)$ is a continuous function defined on a bounded closed set, it has a maximum point according to the extreme value theorem. Also, according to the critical point theorem, the maximum point should either be a stationary point or a boundary point. It is easy to observe that $G(g_1,f_1)$ is a monotonic decreasing function of f_1. $\frac{\partial G}{\partial f_1} < 0$, i.e., there is no stationary point for $G(g_1,f_1)$ on the defined domain. Then, the maximum point should lie on the four boundaries: (i) $g_1 = lb_g,\ f_1 \in [\sqrt{lb_g},ub_f]$, (ii) $f_1 = ub_f,\ g_1 \in [lb_g,ub_g]$, (iii) $g_1 = ub_g,\ f_1 \in [\sqrt{ub_g},ub_f]$, and (iv) $g_1 = lb_f^2,\ f_1 \in [\sqrt{lb_g},\sqrt{ub_g}]$.

For the boundary (i), since $G(g_1,f_1)$ is a monotonic decreasing function of f_1, the maximum point can only lie on the point $(\sqrt{lb_g},lb_g)$ which belongs to the boundary (iv) as well. For the boundary (ii), since $G(g_1,f_1)$ is a monotonic decreasing function of f_1, we can pick $f_1 < ub_f$ to increase the value of $G(g_1,f_1)$, which implies that the maximum point cannot lie on this boundary. For the boundary (iii), since $G(g_1,f_1)$ is a monotonic decreasing function of f_1, the maximum point must lie on the point $(\sqrt{ub_g},ub_g)$, which also belongs to the boundary (iv). Since the possible maximum points on the boundaries (i) and (iii) also belong to the boundary (iv), the maximum point must lie on the boundary (iv) and the proof is complete.

REFERENCES

[1] T. S. Rappaport *et al.*, *Millimeter wave wireless communications*. Pearson Education, 2014.

[2] L. Jiang and H. Jafarkhani, "Multi-user analog beamforming in millimeter wave MIMO systems based on path angle information," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 608–619, Jan. 2019.

[3] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[4] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.

[5] L. Jiang and H. Jafarkhani, "Mmwave amplify-and-forward MIMO relay networks with hybrid precoding/combining design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1333–1346, Feb. 2020.

[6] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, July 2013.

[7] M. A. Almasi, H. Mehrpouyan, V. Vakilian, N. Behdad, and H. Jafarkhani, "A new reconfigurable antenna MIMO architecture for mmwave communication," *in Proc. IEEE Int. Conf. Commun.*, pp. 1–7, May 2018.

[8] B. He and H. Jafarkhani, "Low-complexity reconfigurable MIMO for millimeter wave communications," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5278–5291, Nov. 2018.

[9] Z. Wei, D. W. K. Ng, and J. Yuan, "NOMA for hybrid mmwave communication systems with beamwidth control," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 567–583, June 2019.

[10] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond.* Springer, 2019.

[11] M. Ganji and H. Jafarkhani, "Improving NOMA multi-carrier systems with intentional frequency offsets," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1060–1063, Aug. 2019.

[12] 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 VTC Spring*, pp. 1–5, June 2013.

[13] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Feb. 2017.

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

[15] B. Wang *et al.*, "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.

[16] 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.

[17] Z. Xiao *et al.*, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May 2018.

[18] Y. Zhou, V. W. S. Wong, and R. Schober, "Coverage and rate analysis of millimeter wave NOMA networks with beam misalignment," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8211–8227, Dec. 2018.

[19] M. A. Almasi, M. Vaezi, and H. Mehrpouyan, "Impact of beam misalignment on hybrid beamforming NOMA for mmwave communications," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4505–4518, June 2019.

[20] M. A. Almasi, R. Amiri, M. Vaezi, and H. Mehrpouyan, "Lens-based millimeter wave reconfigurable antenna NOMA," *in Proc. IEEE Int. Conf. Commun. Workshops*, pp. 1–5, May 2019.

[21] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[22] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[23] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[24] H. Shokri-Ghadikolaei, L. Gkatzikis, and C. Fischione, "Beam-searching and transmission scheduling in millimeter wave communications," *in Proc. IEEE Int. Conf. Commun.*, pp. 1292–1297, June 2015.

[25] J. Palacios, D. De Donno, D. Giustiniano, and J. Widmer, "Speeding up mmwave beam training through low-complexity hybrid transceivers," *in Proc. IEEE 27th PIMRC*, pp. 1–7, Sept. 2016.

[26] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1576–1590, July 2017.

[27] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for

millimeter wave beam alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5689–5701, Sept. 2017.

[28] N. Eshraghi, V. Shah-Mansouri, and B. Maham, "Fair beamwidth selection and resource allocation for indoor millimeter-wave networks," *in Proc. IEEE Int. Conf. Commun.*, pp. 1–6, May 2017.

[29] C. Pradhan, H. Chen, Y. Li, and B. Vucetic, "Joint beamwidth and energy optimization for multi-user millimeter wave communications," *in Proc. IEEE Int. Conf. Commun. Workshops*, pp. 1–6, May 2018.

[30] R. A. Hassan and N. Michelusi, "Multi-user beam-alignment for millimeter-wave networks," *in Proc. Inf. Theory and Appl. Workshop*, pp. 1–7, Feb. 2018.

[31] M. Hussain and N. Michelusi, "Energy-efficient interactive beam alignment for millimeter-wave networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 838–851, Feb. 2019.

[32] W. Hao, F. Zhou, Z. Chu, P. Xiao, R. Tafazolli, and N. Al-Dhahir, "Beam alignment for MIMO-NOMA millimeter wave communication systems," *in Proc. IEEE Int. Conf. Commun.*, pp. 1–6, May 2019.

[33] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[34] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[35] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[36] C. Liu, M. Li, S. V. Hanly, I. B. Collings, and P. Whiting, "Millimeter wave beam alignment: Large deviations analysis and design insights," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1619–1631, July 2017.

[37] D. De Donno, J. Palacios, and J. Widmer, "Millimeter-wave beam training acceleration through low-complexity hybrid transceivers," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3646–3660, June 2017.

[38] 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, Dec. 2017.

[39] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-Wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, Nov. 2018.

[40] J. Seo, Y. Sung, and H. Jafarkhani, "A high-diversity transceiver design for MISO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2591–2606, May 2019.

[41] J. Wildman *et al.*, "On the joint impact of beamwidth and orientation error on throughput in directional wireless poisson networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 7072–7085, Dec. 2014.

[42] C. A. Balanis, *Antenna theory: analysis and design*. John wiley & sons, 2016.

[43] L. Dai, B. Wang, Y. Yuan, S. Han, C. 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, Sept. 2015.

[44] 3GPP R1-154999, "TP for classification of MUST schemes," *TSG-RAN WG1 #82, Beijing, China*, Aug. 24-28, 2015.

[45] W.-C. Li, T.-H. Chang, C. Lin, and C.-Y. Chi, "Coordinated beamforming for multiuser MISO interference channel under rate outage constraints," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1087–1103, Mar. 2012.

[46] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journal of Global Optimization*, vol. 47, no. 1, pp. 29–51, May 2010.

[47] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.

[48] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.