6-1-2019

Impact of Beam Misalignment on Hybrid Beamforming NOMA for mmWave Communications

Mojtaba Ahmadi Almasi
Boise State University

Mojtaba Vaezi
Villanova University

Hani Mehrpouyan
Boise State University

 doi: 10.1109/TCOMM.2019.2904507

Impact of Beam Misalignment on Hybrid Beamforming NOMA for mmWave Communications

Mojtaba Ahmadi Almasi, Mojtaba Vaezi, Hani Mehrpouyan

Abstract

This paper analyzes the effect of beam misalignment on rate performance in downlink of hybrid beamforming-based non-orthogonal multiple access (HB-NOMA) systems. First an HB-NOMA framework is designed in multiuser millimeter wave (mmWave) communications. A sum-rate maximization problem is formulated for HB-NOMA, and an algorithm is introduced to design digital and analog precoders and efficient power allocation. Then, regarding perfectly aligned line-of-sight (LoS) channels, a lower bound for the achievable rate is derived. Next, when the users experience misaligned LoS or non-LoS (NLoS) channels, the impact of beam misalignment is evaluated. To this end, a misalignment factor is modeled and each misaligned effective channel is described in terms of the perfectly aligned effective channel parameters and the misalignment factor. Further, a lower bound for the achievable rate is extracted. We then derive an upper bound for the rate gap expression between the aligned and misaligned HB-NOMA systems. The analyses reveal that a large misalignment can remarkably degrade the rate. Extensive numerical simulations are conducted to verify the findings.

Index Terms

Millimeter wave, hybrid beamforming, NOMA, beam misalignment, achievable rate.

This project was supported in part by the NSF ERAS under Grant 1642865. This work was presented in part at VTC Fall 2018 [1].

M. A. 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).

M. Vaezi is with the Department of Electrical and Computer Engineering, Villanova University, Villanova, PA 19085, USA. He is also a Visiting Research Collaborator at Princeton University (e-mail: mvaezi@villanova.edu).

I. Introduction

Millimeter wave (mmWave) communications has emerged as one of the key solutions for the fifth-generation (5G) wireless networks. The existence of large unused spectrum at mmWave band (30-300 GHz) offers the potential for significant throughput gains. Shorter wavelengths of the mmWave band, on the other hand, allow for the deployment of large numbers of antenna elements at both the base station (BS) and mobile users, which, in turn, enables mmWave systems to support higher degrees of *multiplexing* gain in the multiple-input multiple-output (MIMO) and *multiuser MIMO* systems [2]–[5]. To this end, the BS needs to apply some form of *beamforming*. This beamforming can be done in the baseband, radio frequency (RF), or a combination of the two. While *baseband beamforming* (fully-digital) offers a better control over the entries of the precoding matrix, it is unlikely with current semiconductor technologies due to high hardware cost and power consumption. *Analog beamforming* is an alternative to the baseband beamforming which controls the phase of the signal transmitted at each antenna using analog phase-shifters implemented in the RF domain. Fully-analog beamforming which uses one *RF chain*, see, e.g. [6], can, however, support only one data stream.

In order to transmit multiple streams and keep the hardware complexity and energy consumption low, by exploiting several RF chains, *hybrid analog/digital beamforming* mmWave systems are designed [7], [8]. In [9] and [10], the concept of beamspace MIMO is introduced where several RF chains are connected to a lens antenna array via switches. Recently, multi-beam lens-based reconfigurable antenna MIMO systems have been proposed to overcome severe path loss and shadowing in mmWave frequencies [11], [12]. In the aforementioned systems, each beam is considered to serve only one user. The works in [13] and [14] show that exploiting hybrid beamforming in multiuser systems achieves a higher spectral efficiency. Also, [15] enhances the spectral efficiency by supporting several users through multi-beam reconfigurable antenna. Nevertheless, the number of served users are far less than the number of users envisioned for 5G networks.

Non-orthogonal multiple access (NOMA) is another enabling technique for 5G networks that augments the number of users and spectral efficiency in multiuser scenarios [16]–[23]. Unlike orthogonal multiple access (OMA) techniques, such as time division multiple access (TDMA), frequency division multiple access (FDMA), and code division multiple access (CDMA) which can support only one user per time, frequency, or code, respectively, NOMA can support multiple

users in the same time/frequency/code/beam. NOMA can be realized in the code, power, or other domains [24]. In the power domain, NOMA employs *superposition coding* at the transmitter. This technique exploits the channel gain difference between users to multiplex their signal. Subsequently, *successive interference cancellation* (SIC) is applied at the receiver such that the user with better channel first decodes the signal of the user with worse channel and then subtracts it from the received signal to decode its own signal [16]–[25]. Beside superiority of NOMA over OMA techniques in terms of number of supported users and spectral efficiency, OMA techniques may not be a practical option for mmWave communications [26]. As an example, TDMA, which serves users through orthogonal time slots but the same spectrum, requires precise and fast timing synchronization. This is because symbol rate in 5G network is far higher than the current networks. Therefore, employing TDMA in mmWave 5G network might be challenging. Exploiting FDMA in mmWave 5G network also can bring about implementation issues. In FDMA, the existing large frequency band is divided into several orthogonal frequency bands. It is expected that FDMA to serve all users via the orthogonal bands at the same time slot. However, due to highly directional beams, the current mmWave systems are not able to cover all users' locations and only a few users will be supported. Further, frequency band division causes the allocated bandwidth for each user in a dense mmWave network to become small. So, mmWave networks may not have enough bandwidth to support the users with the required high data-rate. The obstacles related to using CDMA in mmWave frequencies have been explained in [26]. The propagation characteristics of mmWave frequencies are another reason to incorporate the hybrid beamforming systems and NOMA. Transmission in mmWave band suffers from high path loss and thus users in different locations may experience very different channel gains. This implies that mmWave band better suits power domain NOMA which offers a larger spectral efficiency when the channel gain difference between the users is high. Severe shadowing and blockage are other factors that make mmWave links vulnerable to outage [2], [3], [5]. Although the large unused spectrum in mmWave bands is envisioned a promising solution for high data-rate transmission in 5G networks, high path loss and outage due to shadowing and blockage make mmWave links prone to temporary shutdowns. Hence, when the link exists, increasing the spectral efficiency will lead to higher data-rate. This would meet the required unprecedented throughput of $1000\times$ current networks in 5G networks.

Integration of NOMA into mmWave systems, which allows multiple users to share the same beam or the same RF chain, has been received considerable research interests [26]–[33]. In [27],

a random beamforming technique is designed for mmWave NOMA systems where the BS randomly radiates a directional beam toward paired users. In [28], it is shown that mismatch between the users' channel vector and finite resolution analog beamforming[1] simplifies utilizing NOMA in MIMO mmWave systems. In [29], a combination of beamspace MIMO and NOMA is proposed to ensure that the number of served users is not limited to the number of RF chains. In [30], NOMA is studied for hybrid mmWave MIMO systems, where a power allocation algorithm has been provided in order to maximize energy efficiency. In all aforementioned works, NOMA is combined with mmWave systems assuming only baseband precoders/combiners. The works in [26], [31]–[33] have recently studied NOMA in hybrid beamforming systems. Ref. [26] proposes a beam splitting NOMA scheme for hybrid beamforming mmWave systems. In order to increase the spectral efficiency, some users are served with a common RF chain but the grated beams. This technique is only proper when the angle of the directional beams serving the users is large enough. Also, beam grating divides the power of a strong mmWave beam. Hence, far users cannot capture the required power. In [31], designing beamforming vectors and allocating power for just two users have been studied. In [32], it is demonstrated that due to the utilization of HB, the digital precoder of the BS is not perfectly aligned with the user's effective channel. Then, a power allocation algorithm that maximizes the sum-rate has been proposed. Only two users in each beam is considered; moreover, the work fails to study the effect of analog beamforming on the rate performance. Newly, Zhou $et\ al.$ have proposed an angle-based user pairing strategy [33]. The strategy repeatedly switches between NOMA and OMA techniques. Such that, when beamwidth of mainlobe of BS is not smaller than the angle difference between two users, they are considered as NOMA users. Otherwise, they are treated as OMA users. Then, the coverage probability and the sum-rate are evaluated. Regularly switching between NOMA and OMA will add more hardware complexity to the system. Also, as it is mentioned, OMA techniques may not be a practical choice for mmWave systems. *In mmWave systems, due to the directional nature of beams in mmWave systems, beam misalignment between the BS and users is inevitable [34]*. Most of the reviewed works consider neither the effect of phase-shifters employed in the analog beamformer of a HB system nor the effect of beam misalignment.

In this paper, we investigate the impact of exploiting NOMA in multiuser HB systems termed HB-NOMA. At the outset, it is supposed that HB-NOMA users are paired with respect to their

[1]Finite resolution analog beamforming is due to the use of a finite number of phase-shifters in the analog beamformer.

locations and effective channels which is widely adopted by recent research works [26]–[33]. The achievable rate is evaluated when the BS and users' beam are aligned and misaligned. Essentially, the perfect beam alignment is attributed to the existence of LoS channel aligned in the same direction between the BS and users which allows the users to steer their beam directly toward the BS. The imperfect beam alignment (misalignment) occurs due to practical phenomena such as misaligned LoS channels and NLoS channels which are caused by shadowing and blockage. To the best of authors' knowledge, this paper is the first research work that studies the effect of integration of hybrid beamforming and NOMA on the achievable rate in the presence of beam alignment and misalignment. The contribution of this paper is summarized as follows.

1) We incorporate the 5G enabling technology NOMA and a multiuser HB system studied in [14]. Since we aim to evaluate the impact of beam misalignment on the downlink of HB-NOMA systems, a sum-rate expression is formulated. Specifically, we revise the sum-rate expression in [14] with regard to the NOMA technique. Then, an algorithm is introduced to maximize the system sum-rate subject to a total power constraint, in three steps. To get the first and second steps, we design the analog and digital precoders only regarding LoS channels using the well-known strong effective channel-based effective channel precoder. The third step is a location-based static power allocation.
2) As the maximized sum-rate directly depends on the effective channels of users, we first study the rate for perfect beam alignment where all users exploit LoS channels. A lower bound is derived for the achievable rate of an HB-NOMA user. The bound reveals that the interference is just due to using NOMA in which SC technique at transmitter and SIC at the receiver are exploited. That is to say, the interference on a user is caused by NOMA users located inside the same cluster called intra-cluster interference. Indeed, HB slightly amplifies the noise term which is led by analog devices used in the beamformer. The analysis shows that for the perfect alignment, the HB-NOMA users can achieve a rate which is close to that of NOMA with the fully-digital beamforming systems.
3) We study the achievable rate of the maximized sum-rate for misaligned beams between the BS and users in the presence of misaligned LoS and NLoS channels. Toward this goal, the beam misalignment problem is modeled by a beam misalignment factor. Considering the derived factor, the effective channel of the users with misaligned LoS or NLoS channel is described in terms of the aligned effective channel parameter and the misalignment factor.

4) We extract a lower bound for the achievable rate using the effective channel model. Three terms, i.e., intra-cluster interference, inter-cluster interference, and noise, constrain the achievable rate. Unfortunately, these terms are directly or indirectly associated with misalignment factors. It is concluded that in HB-NOMA with the precoder based on the strongest effective channel the achievable rate of a user depends on both the effective channel gain and beam alignment issue. This is opposite to the fully-digital NOMA systems in which only the effective channel gain affects the rate. Then, an upper bound for rate gap between the aligned and misaligned HB-NOMA user is found.

To confirm the analyses and the derived expressions, numerical simulations are done. Different HB-NOMA system parameters are evaluated. The simulations indicate that the HB-NOMA outperforms OMA.

The paper is organized as follows: Section II presents the system model of HB-NOMA and formulates a sum-rate expression. In Section III, we maximize the sum-rate for perfect beam alignment then analyze the rate performance. Section IV studies the rate performance for beam misaligned HB-NOMA. In Section V, we present simulation results investigating the rate performance of HB-NOMA. Section VI concludes the paper.

Notations: Hereafter, $j = \sqrt{-1}$, small letters, bold letters and bold capital letters will designate scalars, vectors, and matrices, respectively. Superscripts $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^\dagger$ denote the transpose, conjugate and transpose-conjugate operators, respectively. Further, $|\cdot|$, $\|\cdot\|$, and $\|\cdot\|_2$ denote the absolute value, norm-1 of $(\cdot)$, and norm-2 of vector $(\cdot)$, respectively. Indeed, $\|\cdot\|_F$ denotes the Frobenius norm of matrix $(\cdot)$. Finally, $\mathbb{E}[\cdot]$ denotes the expected value of $(\cdot)$.

II. System Model and Rate Formulation

A. System Model for HB-NOMA

We assume a narrow band mmWave downlink system composed of a BS and multiple users as shown in Fig. 1. The BS is equipped with N_{RF} chains and N_{BS} antennas whereas each user has one RF chain and N_{U} antennas. Each RF chain is connected to the antennas through phase-shifters. We also assume that the BS communicates with each user via only one stream. This will be justified later in the present section. In traditional multiuser systems based on the hybrid beamforming the maximum number of users that can be simultaneously served by the BS equals the number of BS RF chains [14].

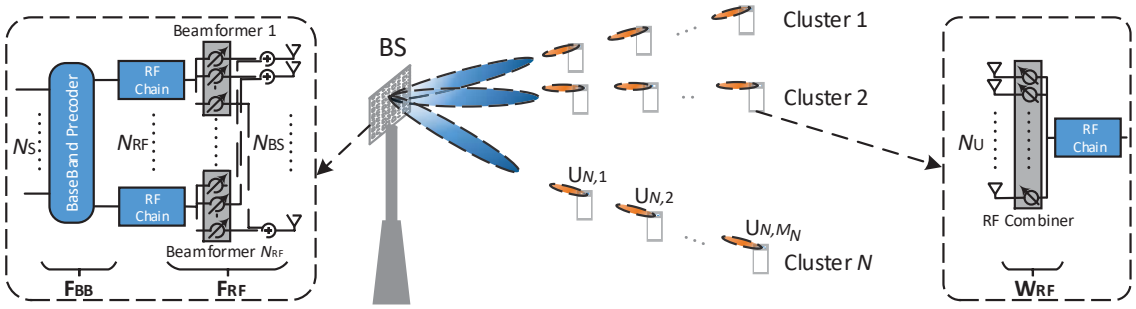


Fig. 1: HB-NOMA with one BS and huge number of users grouped into N clusters each with M_n NOMA users. $N_{\mathrm{S}}, N_{\mathrm{RF}}, N_{\mathrm{BS}}$, and N_{U} are the numbers of multiplexed streams, RF chains, BS antennas, and user antennas, respectively.

In order to establish a better connectivity in dense areas and further improve the sum-rate, this paper develops HB-NOMA system. The system is practical and takes the parameters of the promising hybrid beamforming into account. To achieve this, we utilize NOMA in hybrid beamforming multiuser systems where each beam is allowed to serve more than one user. The transmitter simultaneously sends N_{S} streams toward $\sum_{n=1}^{N} M_n$ users which are grouped into $N \leq N_{\mathrm{RF}}$ clusters. M_n denotes the number of users in the nth cluster. The users in each cluster can be scheduled by using the efficient approaches presented in [35], [36]. Without loss of generality, we assume $N_{\mathrm{S}} = N$. Hence, $\sum_{n=1}^{N} M_n \gg N_{\mathrm{RF}}$; i.e., an HB-NOMA system can simultaneously serve $\sum_{n=1}^{N} M_n$ users which is much larger than the number of RF chains. In the following we formulate the transmit and received signals for the HB-NOMA system.

1) Superposition coding: On the downlink of the HB-NOMA system, first, the transmit symbols are superposition coded at the BS. Let $\mathbf{s} = [s_1, s_2, \ldots, s_N]^T$ denote the information signal vector such that $\mathbb{E}[s_n s_n^*] = \frac{1}{N}$. Each $s_n = \sum_{m=1}^{M_n} \sqrt{P_{n,m}} s_{n,m}$ is the superposition coded signal performed by NOMA with $P_{n,m}$ and $s_{n,m}$ being transmit power and transmit information signal for the mth user in the nth cluster. Then, the hybrid beamforming is done in two stages. In the first stage, the transmitter applies an $N \times N$ baseband precoder $\mathbf{F}_{\mathrm{BB}}$ using its N_{RF} RF chains. This stage then is followed by an $N_{\mathrm{BS}} \times N$ RF precoder $\mathbf{F}_{\mathrm{RF}}$ using analog phase-shifters. Thus, the transmit signal vector after superposition coding is given by

$$[x_1, x_2, \ldots, x_N]^T = \mathbf{F}_{\mathrm{RF}} \mathbf{F}_{\mathrm{BB}} [s_1, s_2, \ldots, s_N]^T, \tag{1}$$

where x_n denotes the transmit signal toward the nth cluster. Hereafter, $\mathrm{U}_{n,m}$ denotes the mth user in the nth cluster. Since $\mathbf{F}_{\mathrm{RF}}$ is implemented by using analog phase-shifters it is assumed that all elements of $\mathbf{F}_{\mathrm{RF}}$ have an equal norm, i.e., $|\left(\mathbf{F}_{\mathrm{RF}}\right)_{n,m}|^2 = N_{\mathrm{BS}}^{-1}$. Also, the total power of the hybrid transmitter is limited to $\left\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\right\|_F^2 = N$ [8], [14].

2) Successive interference cancellation: The received signal at $\mathrm{U}_{n,m}$ is given by

$$\mathbf{r}_{n,m} = \mathbf{H}_{n,m}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{n}_{n,m}, \tag{2}$$

where $\mathbf{H}_{n,m}$ of size $N_{\mathrm{U}} \times N_{\mathrm{BS}}$ denotes the mmWave channel between the BS and $\mathrm{U}_{n,m}$ such that $\mathbb{E}\left[\left\|\mathbf{H}_{n,m}\right\|_F^2\right] = N_{\mathrm{BS}}N_{\mathrm{U}}$. $\mathbf{n}_{n,m} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is the additive white Gaussian noise vector of size $N_{\mathrm{U}} \times 1$. Each component of $\mathbf{n}_{n,m}$ has zero-mean and σ^2 variance. $\mathbf{I}$ denotes the identity matrix of size $N_{\mathrm{U}} \times N_{\mathrm{U}}$. At $\mathrm{U}_{n,m}$, the RF combiner is used to process the received vector as

$$\begin{aligned} y_{n,m} = {} & \underbrace{\mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}^{n}\sqrt{P_{n,m}}s_{n,m}}_{\text{desired signal}} + \underbrace{\mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}^{n}\sum_{k=1,k\neq m}^{M}\sqrt{P_{n,k}}s_{n,k}}_{\text{intra-cluster interference}} \\ & + \underbrace{\mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\sum_{\ell=1,\ell\neq n}^{N}\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB}}^{\ell}\sum_{q=1}^{M}\sqrt{P_{\ell,q}}s_{\ell,q}}_{\text{inter-cluster interference}} + \underbrace{\mathbf{w}_{n,m}^{\dagger}\mathbf{n}_{n,m}}_{\text{noise}}, \end{aligned} \tag{3}$$

where $\mathbf{w}_{n,m} \in \mathbb{C}^{N_{\mathrm{U}}\times 1}$ denotes the combiner at $\mathrm{U}_{n,m}$. After combining, each user decodes the intended signal by using SIC as follows. The first user of each cluster, which has the highest channel gain, is allocated the lowest power and the M_nth user, which has the lowest channel gain, is allocated the highest power. At the receiver side, $\mathrm{U}_{n,m}$ decodes the intended signal of $\mathrm{U}_{n,k'}$, i.e., $s_{n,k'}$, for $k' = m+1, m+2, \ldots, M_n$ and subtracts it from the received signal $y_{n,m}$. However, NOMA treats the intended signal of $\mathrm{U}_{n,k}$ for $k = 1, 2, \ldots, m-1$ as intra-cluster interference. In this paper, SIC process is assumed to be ideal. When SIC is non-ideal, the user cannot completely remove the signals of some of $\mathrm{U}_{n,k'}$ for $k' = m_1, m+2, \ldots, M_n$ which degrades the performance of the system [37]. The effect of non-ideal SIC on NOMA has recently been studied in [38]. The effect of non-ideal SIC on HB-NOMA will be evaluated in the authors' future work. To this end, the BS should send the order of superposition coding to all users in the cluster. Usually NOMA users are selected to have very different channel gains, specially in mmWave frequencies in which path loss is higher that sub-6 GHz frequencies. So, the order of decoding can be estimated from the user's distance to the BS or its channel gain, correspondingly. We note that the order of encoding is related to the channel gain as indicated in Section III-A.

3) Channel model: In mmWave communications, the extended Saleh-Valenzuela model as a multi-path channel (MPC) model has been widely adopted for hybrid beamforming systems [8], [13], [14]. In this model, each LoS and NLoS path is described by a channel gain and array steering/response vector at the transmitter/receiver. Here, the number of paths between the BS and $\mathrm{U}_{n,m}$ is defined by $A_{n,m}$. The channel matrix is given by

$$\mathbf{H}_{n,m} = \sqrt{\frac{N_{\mathrm{BS}} N_{\mathrm{U}}}{A_{n,m}}} \sum_{\alpha=1}^{A_{n,m}} \beta_{n,m,\alpha} \mathbf{a}_{\mathrm{U}}(\vartheta^{\mathrm{Az}}_{n,m,\alpha}, \vartheta^{\mathrm{El}}_{n,m,\alpha}) \mathbf{a}^{\dagger}_{\mathrm{BS}}(\varphi^{\mathrm{Az}}_{n,m,\alpha}, \varphi^{\mathrm{El}}_{n,m,\alpha}), \tag{4}$$

where $\beta_{n,m,\alpha} = g_{n,m,\alpha} d_{n,m,\alpha}^{\frac{-\nu}{2}}$ with $g_{n,m,\alpha}$ is the complex gain with zero-mean and unit-variance for the αth MPC, $d_{n,m,\alpha}$ is the distance between the BS and $\mathrm{U}_{n,m,\alpha}$, and ν is the path loss factor. $\vartheta^{\mathrm{Az}}_{n,m,\alpha}$ ($\vartheta^{\mathrm{El}}_{n,m,\alpha}$) and $\varphi^{\mathrm{Az}}_{n,m,\alpha}$ ($\varphi^{\mathrm{El}}_{n,m,\alpha}$) are normalized azimuth (elevation) angle of arrival (AoA) and angle of departure (AoD), respectively. Also, $\mathbf{a}_{\mathrm{BS}}$ and $\mathbf{a}_{\mathrm{U}}$ are the antenna array steering/response vector of the BS/$\mathrm{U}_{n,m}$. In mmWave outdoor communications, to further reduce the interference, sectorized BSs are likely employed [39]. Mostly, each sector in azimuth domain is much wider than elevation domain [39]. Reasonably, we assume that the BS separates the clusters in azimuth domain and considers fixed elevation angles. 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

$$\mathbf{a}_{\mathrm{BS}}(\varphi_{n,m,\alpha}) = \frac{1}{\sqrt{N_{\mathrm{BS}}}} \left[1, e^{-j\pi\varphi_{n,m,\alpha}}, \ldots, e^{-j\pi(N_{\mathrm{BS}}-1)\varphi_{n,m,\alpha}}\right]^{T}. \tag{5}$$

where $\varphi_{n,m,\alpha} \in [-1, 1]$ is related to the AoD $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ as $\varphi_{n,m,\alpha} = \frac{2D\sin(\phi)}{\lambda}$ [8], [14]. Note that D denotes the antenna spacing and λ denotes the wavelength of the propagation. The antenna array response vector for $\mathbf{a}_{\mathrm{U}}(\vartheta_{n,m,\alpha})$ can be written in a similar fashion.

It is mentioned that transmission at mmWave systems is done through directional beams. Since the BS is equipped with HB system, the beamforming can be conducted as follows. When both LoS and NLoS components are available, because LoS component is stronger than NLoS it is reasonable to steer the beam toward LoS component. When only NLoS channels are available, the beam would be steered toward the strongest NLoS component. Thus, only one stream is sent for each cluster. This will also lead to low hardware cost and power consumption due to using one RF chain per stream. Therefore, with a single path component, i.e., $A_{n,m} = 1$, the MPC model described in (4) is converted to a single path channel given by

$$\mathbf{H}_{n,m} = \sqrt{N_{\mathrm{BS}} N_{\mathrm{U}}} \beta_{n,m} \mathbf{a}_{\mathrm{U}}(\vartheta_{n,m}) \mathbf{a}^{\dagger}_{\mathrm{BS}}(\varphi_{n,m}). \tag{6}$$

B. Rate Formulation

In (3), after applying superposition coding at the transmitter, each user experiences two types of interference. Intra-cluster interference which is due to other users within the cluster and inter-cluster interference which is due to users within other clusters. Suppressing the intra-cluster interference directly depends on efficient power allocation and deploying SIC which is discussed in the previous section. To mitigate the inter-cluster interference, the transmitter needs to design a proper beamforming matrix which will be discussed in Section III-A. Hence, the rate for $\mathrm{U}_{n,m}$ is expressed as

$$R_{n,m} = \log_2 \left(1 + \frac{P_{n,m} \left| \mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{F}_{\mathrm{RF}} \mathbf{f}_{\mathrm{BB}}^{n} \right|^2}{I_{\mathrm{intra}}^{n,m} + I_{\mathrm{inter}}^{n,m} + \sigma^2} \right), \tag{7}$$

where $I_{\mathrm{intra}}^{n,m}$ is given by

$$I_{\mathrm{intra}}^{n,m} = \sum_{k=1}^{m-1} P_{n,k} \left| \mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{F}_{\mathrm{RF}} \mathbf{f}_{\mathrm{BB}}^{n} \right|^2, \tag{8}$$

denotes the intra-cluster. Also, $I_{\mathrm{inter}}^{n,m}$ is defined as

$$I_{\mathrm{inter}}^{n,m} = \sum_{\ell=1, \ell \neq n}^{N} \sum_{q=1}^{M_n} P_{\ell,q} \left| \mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{F}_{\mathrm{RF}} \mathbf{f}_{\mathrm{BB}}^{\ell} \right|^2, \tag{9}$$

denotes the inter-cluster interference.

III. Perfect Beam Alignment: Rate Maximization and Analysis

A. The Maximization Algorithm

To optimize the sum-rate performance, hybrid precoder $\mathbf{F}_{\mathrm{RF}}$, and $\mathbf{F}_{\mathrm{BB}}$, combiner $\mathbf{w}_{n,m}$ and transmit power $P_{n,m}$ for $m = 1, 2, \ldots, M_n$ and $n = 1, 2, \ldots, N$ should be found from

$$\underset{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}, \mathbf{w}_{n,m}, P_{n,m}}{\text{maximize}} \quad \sum_{n=1}^{N} \sum_{m=1}^{M_n} R_{n,m} \tag{10a}$$

$$\text{subject to} \quad \left| \left(\mathbf{F}_{\mathrm{RF}} \right)_{n,m} \right|^2 = N_{\mathrm{BS}}^{-1}, \tag{10b}$$

$$\left\| \mathbf{F}_{\mathrm{RF}} \mathbf{F}_{\mathrm{BB}} \right\|_F^2 = N, \tag{10c}$$

$$\left| \mathbf{w}_{n,m} \right|^2 = N_{\mathrm{U}}^{-1}, \tag{10d}$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M_n} P_{n,m} \leq P, \tag{10e}$$

$$P_{n,m} > 0, \tag{10f}$$

where P equals to the total transmit power. In the above optimization problem, the constraints (10b) and (10d) ensure that all elements of $\mathbf{F}_{\text{RF}}$ and $\mathbf{w}_n$ have an equal norm. Further, the constraint (10c) ensures that the total power of the hybrid transmitter is limited to N. The constraint (10e) guarantees that the total transmit power is limited to P. Finally, (10f) ensures that the allocated power to $\text{U}_{n,m}$ is greater than zero. One would add fairness constrain to the maximization problem. Ref. [38] discusses a viable solution in this case. In particular, a weighted sum-rate which considers a special priority for each user is utilized. Also, to ensure that all the users achieve a predefined minimum rate R_{min}, another constrain can be included in the problem (10) such that $R_{n,m} \geq R_{\text{min}}$. In this case, an iterative algorithm that properly allocates the power is required [40]. Without loss of generality, here, we assume that all the users satisfy $R_{n,m} \geq R_{\text{min}}$.

It is mentioned that transmission in mmWave bands happens through LoS and NLoS channels. In particular, the users which are located far from the BS will mostly be supported via NLoS channels [5]. Let first focus on only LoS channels. We assume that all channels are LoS and the effective channels are perfectly aligned as shown in Fig. 1. By perfect alignment we mean that $\mathbf{a}_{\text{BS}}(\varphi_{n,m})$ is identical for all users in the nth cluster, i.e., $\mathbf{a}_{\text{BS}}(\varphi_{n,1}) = \mathbf{a}_{\text{BS}}(\varphi_{n,2}) = \cdots = \mathbf{a}_{\text{BS}}(\varphi_{n,M_n})$ for $n = 1, 2, \ldots, N$.

In general, there are two extreme cases to design baseband precoder for mmWave-NOMA systems, strong effective channel-based and singular value decomposition (SVD)-based precoder methods [29]. The strong effective channel-based is designed for only LoS channels and the SVD-based precoder is designed for only NLoS channels. Further, to the best of authors' knowledge, it is not shown how to design the SVD-based RF precoder for hybrid beamforming system. Here, in order to understand the behavior of beam misalignment in HB-NOMA systems we choose the strong effective channel-based precoder which is widely used in the literature [29], [30], [32].

The maximization problem in (10) is non-convex and finding the optimal solution is not trivial. To ease, we present an efficient and simple algorithm in three steps as described below.

In the first step, the BS and $\text{U}_{n,m}$ solve the following problem

$$\underset{\mathbf{w}_{n,m}, \mathbf{f}_{\text{RF}}^{n,m}}{\text{maximize}} \quad \left|\mathbf{w}_{n,m}^{\dagger} \mathbf{H}_{n,m} \mathbf{f}_{\text{RF}}^{n,m}\right| \qquad \text{subject to (10b) and (10d).} \tag{11}$$

Since the channel $\mathbf{H}_{n,m}$ has only one path, and given the continuous beamsteering capability assumption, in view of (4), $\mathbf{w}_{n,m} = \mathbf{a}_{\text{U}}(\vartheta_{n,m})$ and $\mathbf{f}_{\text{RF}}^{n,m} = \mathbf{a}_{\text{BS}}(\varphi_{n,m})$, are the optimal solutions [14]. We design the RF (analog) and baseband (digital) precoders using the adopted strong effective channel-based method. Hence, in order to design the RF precoder, the BS selects the first user

of each cluster. The RF precoder of the first user of the nth cluster makes the nth column of the RF precoding matrix, i.e., $\mathbf{f}_{\text{RF}}^{n,1}$, gives the RF precoding matrix as

$$\mathbf{F}_{\text{RF}} = \left[\mathbf{f}_{\text{RF}}^{1,1}, \mathbf{f}_{\text{RF}}^{2,1}, \ldots, \mathbf{f}_{\text{RF}}^{N,1}\right]. \tag{12}$$

The first user is determined based on the locations of the user as follows:

$$|\beta_{n,1}| \geq |\beta_{n,2}| \geq \cdots \geq |\beta_{n,M_n}|, \quad \text{for} \quad n = 1, 2, \ldots, N, \tag{13}$$

where $\beta_{n,m}$ is the gain factor defined in (4). To determine the first user, the BS does not need to know the channel gain of the users. Recall that the channel gain $\beta_{n,m}$, defined in (4), mainly depends on distance between the BS and $\text{U}_{n,m}$ (d) and path loss factor (ν). Since the path loss factor is identical for all users, the first user of each cluster can be determined as the closest user to the BS such that its channel gain has the highest amplitude among the users in the same cluster. While the purpose of ordering in (13) is to define the first user, to realize NOMA, another ordering method based on the effective channel gain is presented in the third step. It should be stressed that the main reason to design the digital precoder with respect to the strongest channel is that the strongest user must decode the other users' signal before its signal. So, the power of this user's signal is not affected by other clusters' signal. More details will be provided in Section IV.

In the second step, the effective channel for $\text{U}_{n,m}$ is expressed as

$$\bar{\mathbf{h}}_{n,m}^{\dagger} = \mathbf{w}_{n,m}^{\dagger}\mathbf{H}_{n,m}\mathbf{F}_{\text{RF}} = \sqrt{N_{\text{BS}}N_{\text{U}}}\beta_{n,m}\mathbf{a}_{\text{BS}}^{\dagger}(\varphi_{n,m})\mathbf{F}_{\text{RF}}. \tag{14}$$

Regarding the strongest channel-based method, we write the effective channel matrix as

$$\overline{\mathbf{H}} = \left[\bar{\mathbf{h}}_{1,1}, \bar{\mathbf{h}}_{2,1}, \ldots, \bar{\mathbf{h}}_{N,1}\right]^{\dagger}, \tag{15}$$

where $\bar{\mathbf{h}}_{n,1}$ denotes the effective channel vector of $\text{U}_{n,1}$.

Designing a proper digital precoder $\mathbf{F}_{\text{BB}}$ can reduce the inter-cluster interference. In brief, designing the baseband precoder becomes equivalent to solving

$$\underset{\{\mathbf{f}_{\text{BB}}^{\ell}\}_{\ell \neq n}}{\text{minimize}} \; I_{\text{inter}}^{n,m} \qquad \text{subject to (10c)}. \tag{16}$$

where $I_{\text{inter}}^{n,m}$ is defined in (8). We notice that so far we have designed the analog beamformer and combiner. The only unknown parameter is the digital beamformer. In this paper, we adopt zero-forcing beamforming (ZFBF) which makes a balance between implementation complexity and performance [41], [42]. Based on ZFBF, the solution for (16) is obtained as [14]

$$\mathbf{F}_{\text{BB}} = \overline{\mathbf{H}}^{\dagger}\left(\overline{\mathbf{H}}\overline{\mathbf{H}}^{\dagger}\right)^{-1}\mathbf{\Gamma}, \tag{17}$$

where the diagonal elements of $\mathbf{\Gamma}$ are given by [14]

$$\mathbf{\Gamma}_{n,n} = \sqrt{\frac{N_{\text{BS}} N_{\text{U}}}{\left(\mathbf{F}^{-1}\right)_{n,n}}} \left|\beta_{n,1}\right|, \quad \text{for} \quad n = 1, 2, \ldots, N. \tag{18}$$

where $\mathbf{F} = \mathbf{F}_{\text{RF}}^{\dagger}\mathbf{F}_{\text{RF}}$. The determined precoder in (17) indicates that inter-cluster interference on first users is zero, i.e., $\bar{\mathbf{h}}_{n,1}^{\dagger}\mathbf{f}_{\text{BB}}^{\ell} = 0$ for $n = 1, 2, \ldots, N$ and $\ell \neq n$. That is, inter-cluster interference is perfectly eliminated on the first users. This completes our justification about the orienting the beams toward the first users and choosing their effective channel vector in designing $\mathbf{F}_{\text{BB}}$.

In the third step, the BS first reorders the users then allocates the power. The reordering process is done based on the effective channel vectors as

$$\left\|\bar{\mathbf{h}}_{n,1}\right\| \geq \left\|\bar{\mathbf{h}}_{n,2}\right\| \geq \cdots \geq \left\|\bar{\mathbf{h}}_{n,M_n}\right\|, \quad \text{for} \quad n = 1, 2, \ldots, N. \tag{19}$$

Notice that in (13) we aimed to find the first users based on the large-scale gain. However, in HB-NOMA the power allocation is conducted based on order of the effective channel gains. It is not irrational to assume that the BS knows the effective channels. This can be done through the channel quality indicator (CQI) messages [43]. Each user feeds the effective channel back to the BS then it sorts the users.

The optimal power allocation in (10) can be done by solving the following problem.

$$\underset{P_{n,m}}{\text{maximize}} \sum_{n=1}^{N}\sum_{m=1}^{M_n} R_{n,m} \qquad \text{subject to (10e) and (10f).} \tag{20}$$

To solve the problem, we propose a two-stage solution. First the BS divides the power between the clusters considering their users' channel gain as follows.

$$P_n = \frac{\sum_{m=1}^{M_n}\left\|\bar{\mathbf{h}}_{n,m}\right\|^2}{\sum_{n=1}^{N}\sum_{m=1}^{M_n}\left\|\bar{\mathbf{h}}_{n,m}\right\|^2} P, \quad \text{for} \quad n = 1, 2, \ldots, N. \tag{21}$$

Then a fixed power allocation is utilized for the users in each cluster respecting the constraint $\sum_{m=1}^{M_n} P_{n,m} = P_n$. To determine $P_{n,m}$, one solution is to allocate a certain amount of power for each $\text{U}_{n,m}$ except the first one that only satisfies $R_{n,m} = R_{\text{min}}$, then the remaining is assigned to $\text{U}_{n,1}$. This power allocation process is in consist with the concept of NOMA in which, to achieve higher sum-rate, the stronger user should receive more power [16]–[19]. On the other hand, recall that mmWave channels are vulnerable to blockage and shadowing. Especially, for

the weak users which are located far from the BS, this issue becomes worse. So, the weak users may not be able to achieve the required minimum rate. Another solution is to give priority to the fairness issue. To this, we need to allocate less power to the strong users and more power to the weak users. It turns out, fairness works against achieving maximum rate. Thus, our solution to achieve maximum rate and compensate for the mmWave propagation issues is to assign the same amount of power for all the users, i.e.,

$$P_{n,1} = P_{n,2} = \cdots = P_{n,M_n}. \tag{22}$$

B. The Achievable Rate Analysis

In this section, the achievable rate of $\mathrm{U}_{n,m}$ is evaluated with respect to the designed parameters. We derive a lower bound which characterizes insightful results on the achievable rate of HB-NOMA.

Theorem 1. With perfect beam alignment, a lower bound on the achievable rate of $\mathrm{U}_{n,m}$ is given by

$$\overline{R}_{n,m} \geq \log_2 \left(1 + \frac{P_{n,m} N_{\mathrm{BS}} N_{\mathrm{U}} \left|\beta_{n,m}\right|^2}{\sum_{k=1}^{m-1} P_{n,k} N_{\mathrm{BS}} N_{\mathrm{U}} \left|\beta_{n,m}\right|^2 + \sigma^2 \kappa_{\min}^{-1}(\mathbf{F})} \right), \tag{23}$$

$\kappa_{\min}(\mathbf{F})$ denotes the minimum eigenvalue of $\mathbf{F}$.

Proof. Please see Appendix A. □

Remark 1. Theorem 1 indicates that when the alignment between the users in each cluster is perfect, still two terms degrade the sum-rate performance of every HB-NOMA user. The first term $\sum_{k=1}^{m-1} P_{n,k} N_{\mathrm{BS}} N_{\mathrm{U}} \left|\beta_{n,m}\right|^2$ is due to using NOMA scheme which leads to inevitable intra-cluster interference. The second term $\kappa_{\min}^{-1}(\mathbf{F})$ is due to realizing the beamforming with digital and analog components, i.e., hybrid beamforming instead of fully-digital components. It is worth mentioning that in the fully-digital beamforming the first term exists but the second term is always one. Therefore, even under perfect beam alignment assumption the hybrid beamforming intrinsically imposes small loss on the achievable rate.

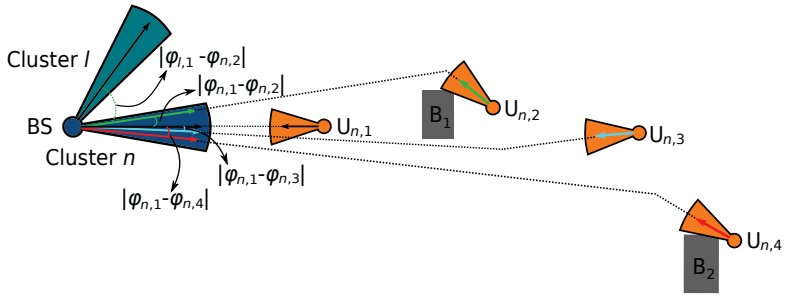


Fig. 2: Beam misalignment in mmWave communications due to the NLoS channels. The NLoS channels are caused by blockages B1 and B2.

IV. Beam Misalignment: Modeling, Rate Analysis, and Rate Gap

In the previous section we designed the precoders when only LoS channels exist and the users are perfectly aligned. The precoders are found based on the strongest effective channel. Perfect alignment is an ideal assumption. In fact, AoDs/AoAs are random variable and with almost surely the probability of occurring different AoDs/AoAs even in LoS channels is one which leads to $\mathbf{a}_{\text{BS}}(\varphi_{n,1}) \neq \mathbf{a}_{\text{BS}}(\varphi_{n,2}) \neq \cdots \neq \mathbf{a}_{\text{BS}}(\varphi_{n,M_n})$ for $n = 1, 2, \ldots, N$. On the other hand, recall that in mmWave frequencies, due to shadowing and blockage, NLoS channels are inevitable [5]. These channels force the users to indirectly steer their beam toward the BS as illustrated by Fig. 2. So, the misalignment between the effective channel of the first user and the users with misaligned LoS and NLoS channel in each cluster causes the digital baseband precoder cannot eliminate the inter-cluster interference. As a result, the achievable rate is degraded. In this section, first the misalignment is modeled. Second, using the derived model, a lower bound is found for the rate. Finally, an upper bound is extracted for the rate gap between the perfect alignment and misalignment.

Remark 2. While our findings in this section are general and hold for misaligned LoS and NLoS channels, we only concentrate on NLoS channels. Thus, by LoS channel we mean a perfectly aligned channel. Also, it is assumed that all users expect the first one in all clusters have NLoS channels. In order to distinguish effective channel of the users with aligned LoS channels from NLoS channels, hereafter, we denote $\bar{\mathbf{h}}_{n,m}$ as effective channel of the user with perfect beam alignment and $\tilde{\mathbf{h}}_{n,m}$ as effective channel of the user with imperfect beam alignment. Also, $\overline{R}_{n,m}$ and $\tilde{R}_{n,m}$ denote the rate of $\text{U}_{n,m}$ with LoS and NLoS channel, respectively.

A. Beam Misalignment Modeling

In what follows, we study the impact of imperfect beam alignment on the rate. Before that, we calculate the norm of the effective channel defined in (14). Defining

$$\left|\mathbf{a}_{\text{BS}}^{\dagger}(\varphi_{n,m})\mathbf{a}_{\text{BS}}(\varphi_{\ell,1})\right|^2 = K_{N_{\text{BS}}}(\varphi_{\ell,1}-\varphi_{n,m}), \tag{24}$$

where $K_{N_{\text{BS}}}$ is Fejér kernel of order N_{BS} [44], we get

$$\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2 = N_{\text{BS}}N_{\text{U}}\left|\beta_{n,m}\right|^2\sum_{\ell=1}^{N}K_{N_{\text{BS}}}\left(\varphi_{\ell,1}-\varphi_{n,m}\right). \tag{25}$$

Now, we model the correlation between the effective channels for $\text{U}_{n,m}$ and $\text{U}_{n,1}$ and between $\text{U}_{n,m}$ and $\text{U}_{\ell,1}$ with $\ell \neq n$ by defining them as intra-cluster misalignment factor and inter-cluster misalignment factor, respectively. Notice that we consider the worst scenario. That is, $\text{U}_{n,m}$ for $m = 2, 3, \ldots, M_n$ receives the signal through NLoS channel, while only $\text{U}_{n,1}$ for $n = 1, 2, \ldots, N$ receives through LoS channel. Assuming LoS channel for the first users is reasonable, since in mmWave communications the users close to the BS experience LoS channels with high probability [5].

Lemma 1. The misalignment effective channel of $\text{U}_{n,m}$ and $\text{U}_{n,1}$ can be modeled as

$$\hat{\tilde{\mathbf{h}}}_{n,m} = \rho_{n,m}\hat{\tilde{\mathbf{h}}}_{n,1} + \sqrt{1-\rho_{n,m}^2}\hat{\mathbf{g}}_{\text{BS}}^{-n}, \tag{26}$$

where $\hat{\tilde{\mathbf{h}}}_{n,m}$ denotes the normalized imperfect effective channel, $\rho_{n,m}$ denotes the misalignment factor obtained as

$$\rho_{n,m} = \frac{\sum_{i=1}^{N}\kappa_i(\mathbf{F})\left|\mathbf{a}_{\text{BS}}^{\dagger}(\varphi_{n,m})\mathbf{v}_1^i\mathbf{v}_1^{i\dagger}\mathbf{a}_{\text{BS}}(\varphi_{n,1})\right|}{\sqrt{\sum_{\ell=1}^{N}K_{N_{\text{BS}}}\left(\varphi_{\ell,1}-\varphi_{n,m}\right)}\sqrt{\sum_{\ell=1}^{N}K_{N_{\text{BS}}}\left(\varphi_{\ell,1}-\varphi_{n,1}\right)}}, \tag{27}$$

where $\kappa_i(\mathbf{F})$ is the ith eigenvalue of $\mathbf{F}$. $\hat{\mathbf{g}}_{\text{BS}}^{-n}$ is a normalized vector located in the subspace generated by linear combination of $\mathbf{a}_{\text{BS}}(\varphi_{\ell,1})$ for $\ell \neq n$, such that $\hat{\mathbf{g}}_{\text{BS}}^{-n} = \frac{\mathbf{g}_{\text{BS}}^{-n}}{\left\|\mathbf{g}_{\text{BS}}^{-n}\right\|}$, where $\mathbf{g}_{\text{BS}}^{-n} = \sqrt{N_{\text{BS}}N_{\text{U}}}\mathbf{F}_{\text{RF}}^{\dagger}\sum_{\ell=1,\ell\neq n}^{N}\beta_{\ell,1}\mathbf{a}_{\text{BS}}(\varphi_{\ell,1})$.

Proof. Please see Appendix B. □

B. Rate Analysis

Now we are ready to find a lower bound for the achievable rate of $\mathrm{U}_{n,m}$.

Theorem 2. With imperfect beam alignment, a lower bound on the achievable rate of $\mathrm{U}_{n,m}$, is given by

$$\tilde{R}_{n,m} \geq \log_2\left(1+\frac{P_{n,m}\rho_{n,m}^2 N_{\mathrm{BS}} N_{\mathrm{U}}\left|\beta_{n,m}\right|^2}{\zeta_{\mathrm{intra}}^{n,m}+\zeta_{\mathrm{inter}}^{n,m}+\zeta_{\mathrm{noise}}^{n,m}}\right), \tag{28}$$

where $\zeta_{\mathrm{intra}}^{n,m}=\sum_{k=1}^{m-1} P_{n,k}\rho_{n,m}^2 N_{\mathrm{BS}} N_{\mathrm{U}}\left|\beta_{n,m}\right|^2$ and $\zeta_{\mathrm{inter}}^{n,m}=\left(1-\rho_{n,m}^2\right) N_{\mathrm{BS}} N_{\mathrm{U}}\left|\beta_{n,m}\right|^2 \kappa_{\max}\left(\mathbf{S}\right)\kappa_{\min}^{-1}(\mathbf{F})\times K_{N_{\mathrm{BS}},1}$ in which $\kappa_{\max}\left(\mathbf{S}\right)$ is the maximum eigenvalue of $\mathbf{S}=\mathbf{F}_{\mathrm{BB}}^{-n,W}\mathbf{F}_{\mathrm{BB}}^{-n,W\dagger}$, $\mathbf{F}_{\mathrm{BB}}^{-n,W}$ denotes the wieghted $\mathbf{F}_{\mathrm{BB}}$ after eliminating the nth column where the columns are scaled by P_ℓ $\forall \ell \neq n$. Also, for some m we define

$$K_{N_{\mathrm{BS}},m}=\sum_{\ell=1}^{N} K_{N_{\mathrm{BS}}}\left(\varphi_{\ell,1}-\varphi_{n,m}\right), \tag{29}$$

where $K_{N_{\mathrm{BS}}}\left(\varphi_{\ell,1}-\varphi_{n,m}\right)$ denotes the Fejér kernel in (24). Finally, $\zeta_{\mathrm{noise}}^{n,m}$ is expressed as $\zeta_{\mathrm{noise}}^{n,m}=\sigma^2\kappa_{\min}^{-1}(\mathbf{F})K_{N_{\mathrm{BS}},1}K_{N_{\mathrm{BS}},m}^{-1}$, where $K_{N_{\mathrm{BS}},m}$ is defined in (29).

Proof. Please see Appendix C. □

Remark 3. Since for $\mathrm{U}_{n,1}$ the factor $\rho_{n,1}$ is one, we have $\bar{\mathbf{h}}_{n,1}=\tilde{\mathbf{h}}_{n,1}$. Thus, Theorem 1 is still valid for these users.

Remark 4. Theorem 2 states that the achievable rate of each user depends on the intra-cluster and inter-cluster misalignment factors, and a weak alignment reduces the power of the effective channel of that user. Intra-cluster and inter-cluster power allocation are other parameters that affect the achievable rate as seen in (28). Further, the bound shows that the maximum eigenvalue of the baseband precoder is important in maximizing the achievable rate. That is to say, the effective channel matrix should be designed in a way that the eigenvalues of the baseband precoder are as close as possible to each other. This is because if eigenvalues are far from each other, the maximum eigenvalue will be large. This increases the value of $\zeta_{\mathrm{inter}}^{n,m}$ which causes less achievable rate.

To gain some insight into the effect of beam misalignment, we extract a lower bound for the rate gap when $\mathrm{U}_{n,m}$ receives the signal via LoS and NLoS channel.

Theorem 3. The rate gap between the perfect aligned and misaligned $\mathrm{U}_{n,m}$ is given by

$$\begin{aligned}\Delta R_{n,m} &\triangleq \overline{R}_{n,m} - \tilde{R}_{n,m} \\ &\leq \log_2\left(1 + \frac{\left(1-\rho_{n,m}^2\right)\kappa_{\max}\left(\mathbf{S}\right) + \sigma^2 K_{N_{\mathrm{BS}},m}^{-1} N_{\mathrm{BS}}^{-1} N_{\mathrm{U}}^{-1} \left|\beta_{n,m}\right|^{-2}}{\rho_{n,m}^2 K_{N_{\mathrm{BS}},1}^{-1} \kappa_{\min}(\mathbf{F}) \sum_{k=1}^{m-1} P_{n,k}}\right). \end{aligned} \tag{30}$$

Proof. Please see Appendix D. □

The upper bound in Theorem 3 explicitly shows the effect of the parameters of HB-NOMA system on the rate performance. A low misalignment factor can substantially increase the rate gap.

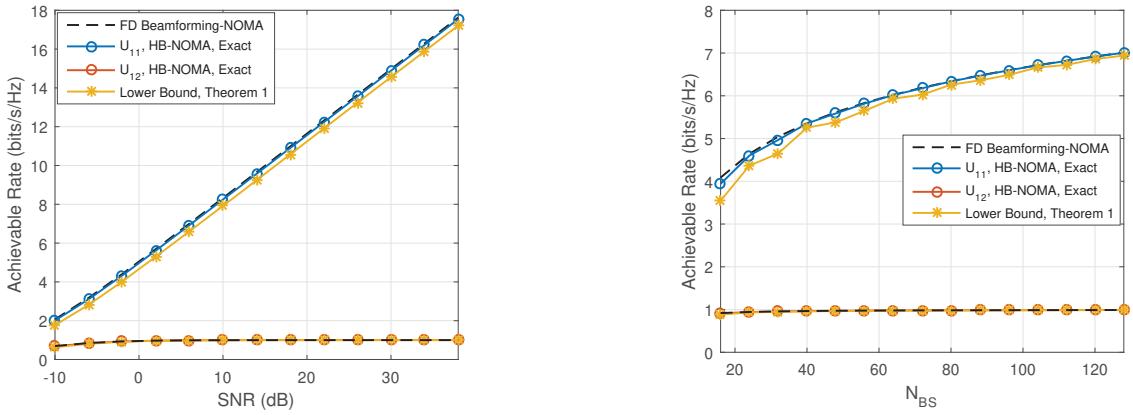
Remark 5. In Section III-A the users are assumed to have LoS channels and to be perfectly aligned in a same direction. Particularly, Eq. (19) orders the users with respect to the their effective channel. Actually, these effective channels are the strongest path between the BS and users. However, when the users are not aligned in the same direction, the effective channels are not necessarily the strongest. This is because the users have to orient their antenna array response vector toward the beam direction of the first user rather than the best direction. Hence, to properly perform SIC, we revise the ordering considering the misalignment effective channel, i.e.,

$$\left\|\tilde{\mathbf{h}}_{n,1}\right\| \geq \left\|\tilde{\mathbf{h}}_{n,2}\right\| \geq \cdots \geq \left\|\tilde{\mathbf{h}}_{n,M_n}\right\|, \quad \text{for} \quad n = 1, 2, \ldots, N. \tag{31}$$

Further, in (21) the aligned effective channel should be replaced by the misaligned effective channel.

V. Numerical Results

In this section we simulate the HB-NOMA system regarding the various design parameters to confirm the analytical derivations in Theorems 1-3. For simulations, since large scaling fading and path loss put more restriction on mmWave systems, the small scale fading is negligible. The defualt number of antennas N_{BS} N_{MU} for the BS and all users is assumed 32 and 8, respectively, unless it is mentioned. The misalignment is described as a random variable uniformly distributed by parameter b, i.e., $\varphi_{n,1} - \varphi_{n,m} \in [-b, b]$. We first present the results of the HB-NOMA with perfect alignment. Then, the effect of misalignment on the rate performance is shown. Finally, the sum-rate of HB-NOMA with OMA is illustrated.



(a) (b)

Fig. 3: Evaluation of rate performance of the strong channel-based precoder in HB-NOMA with perfect alignment (LoS channels) in terms of (a) SNR and (b) N_{BS}.

A. Perfect Beam Alignment

Figure 3 studies the performance of the derived bound in Theorem 1 for aligned users. The users are not affected by the inter-cluster interference from other clusters. It is supposed that the number of users is two and channel gain of the strong and weak user is 0 and -2 dB, respectively. Fig. 3(a) reveals that the HB-NOMA approximately achieves the rate the same as that of fully-digital beamforming (FD beamforming) for a wide range of SNR. In particular, a small gap between the exact value of HB-NOMA and the lower bound is observed for the strong user ($\mathrm{U}_{1,1}$). This is because the complicated expression of the noise term in (23) is replaced by a simple but greater term. For the weak user ($\mathrm{U}_{1,2}$) the bound is very tight due to two reasons. First, in the SINR of the weak user, the noise term is dominated by the interference term. Therefore, the effect of noise term is neglected. Second, the interference term is modeled very accurately. Fig. 3(b) studies the achievable rate for various N_{BS}. For small N_{BS}s, the fully-digital outperforms the HB-NOMA. When N_{BS} is samll, the RF precoder is not able to steer a highly direct beam toward the users. By increasing N_{BS}, the beam becomes narrow and the users capture much more power. Again, for the weak user, the lower bound is accurate at all N_{BS} regions. For the strong user, the bound does not approach to the exact value but, for $N_{\mathrm{BS}} > 60$, the bound is approximately the same as to the exact HB-NOMA.

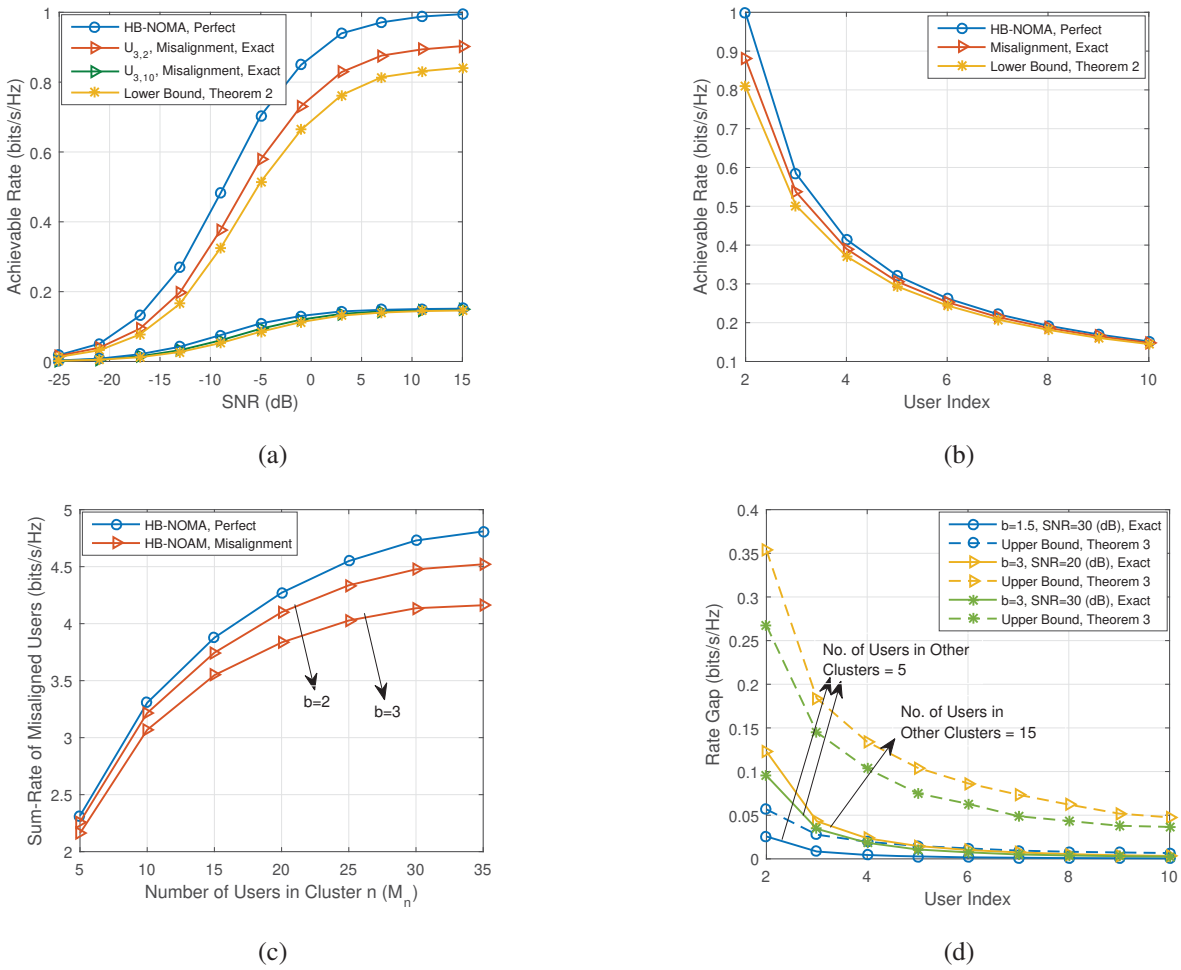


Fig. 4: Evaluation of the misalignment on the rate performance of HB-NOMA versus (a) SNR, (b) user index, and (c) number of users per cluster (M_n). Also, (d) demonstrates the rate gap among the different misaligned users.

B. Beam Misalignment

The beam misalignment effect is depicted by Fig. 4. We consider five clusters in which $\varphi_{1,1} = 10°$, $\varphi_{2,1} = 30°$, $\varphi_{3,1} = 50°$, $\varphi_{4,1} = 65°$, and $\varphi_{5,1} = 80°$. All simulations have been done for the middle cluster (third cluster) which is likely imposed the same interference from all the other clusters. Also, the channel gain of the strongest user is 0 dB and the next user's gain drops 1 dB. For instance, the channel gain of $\mathrm{U}_{n,m}$ is $-(m-1)$ dB. Fig. 4(a), (b), and (d) the number of users in the third cluster is 10.

In Fig. 4(a) the achievable rate of two misaligned users $\mathrm{U}_{3,2}$ (the strong user) and $\mathrm{U}_{3,10}$ (the weak user) versus SNR is shown where the channel gains are -1 and -9 dB, respectively. The

misalignment parameter is assumed $b = 3$. The number of users in all the other clusters is equal to five. Two different observations are obtained. Increasing the SNR leads to a larger rate gap between perfectly aligned and the misaligned HB-NOMA for the strong user, whereas for the weak users both HB-NOMAs achieve almost the same rate for all SNRs. This demonstrates that the effect of misalignment on the strong users is greater than the weak users. In other words, the weak users should deal with the intra-cluster interference while the strong users should deal with the inter-cluster interference. The other observation is that the lower bound is loose for the strong users but tight for the weak user. The observation indicates that our derived normalized effective channel model in Lemma 1 is precise for those users which are intra-cluster interference limited. That is, our finding is able to exactly model the intra-cluster interference. However, the loose lower bound for the strong user indicates that the inter-cluster interference is a little inaccurate which is due to approximating an $N-1$ dimensional subspace with one dimensional space provided in Appendix B.

To gain more details, we have simulated the achievable rate of all the misaligned users for SNR=15 dB in Fig. 4(b). Also, the number of users in the other clusters is set to 15. The mentioned two observations can be seen from this figure, too. However, for strong user, the rate gap between the perfect HB-NOMA and misaligned HB-NOMA is smaller than that of Fig. 4(a). Another important observation gained form Fig. 4(b) is the impact of the power allocation among the clusters. Based on the proposed power allocation scheme in (21), to achieve higher rate, more power is assigned to the other clusters than the third cluster which causes $\mathrm{U}_{3,2}$ to achieve the rate 0.91 bits/s/Hz. Whereas, for the previous scenario more power is allocated to the third cluster which has more users. Therefore, the rate of $\mathrm{U}_{3,2}$ is 0.88 bits/s/Hz. This shows that due to the misalignment the strong clusters leads to higher inter-cluster interference.

Fig. 4(c) compares the sum-rate performance of all the misaligned users with the perfectly aligned HB-NOMA users. Likewise Fig. 4(b), we set SNR=15 dB and 15 users for all the clusters except the third. The number of users in the third cluster varies from 5 to 35. Notice that the sum-rate is shown only for the misaligned users, e.g., rate of the first user is neglected. By increasing the number of users, the allocated power to the cluster increases. In consequence, the total rate increases. However, the difference between the aligned and misaligned HB-NOMA becomes worse. Although more users in a cluster means more power is allocated to, the number of users which have inter-cluster interference limited increases as well. As a result, it brings about higher rate lost. Indeed, by making the misalignment parameter worse (b=6), the rate lost

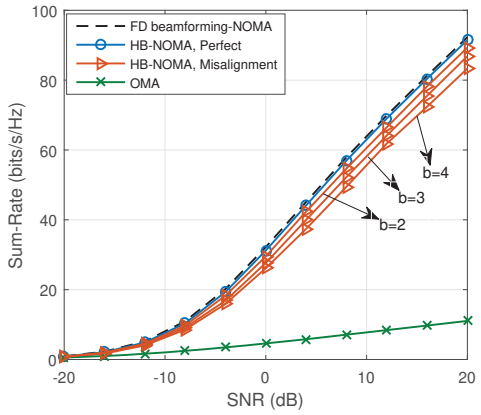


Fig. 5: Sum-rate comparison of the three different systems. The fully-digital and hybrid beamforming systems serve the users using NOMA. The analog system supports the users by exploiting OMA.

becomes bigger. It can be concluded that to avoid higher rate lost, HB-NOMA needs to schedule equal number of users per cluster to serve.

The upper bound evaluation for gap rate between the perfect alignment and misalignment is demonstrated by Fig. 4(d). The number of users in other clusters is 5 or 15. For SNR=30 dB and b=3, the gap is not substantial and the bound is close to the actual value. When b becomes larger, the gap between the stronger users is bigger than the weaker users. When number of the users of the other cluster increases and simultaneously SNR is reduced, only the stronger users' gap increases. To clarify, for $\mathrm{U}_{3,2}$ to $\mathrm{U}_{3,5}$, the gap becomes larger, while for the remaining users it is unchanged. The bounds for b=6 are not very close to the exact rate gap curves. The main reason is that in the deriving process of the bound in the second line of (46) in Appendix D, the effect of the inter-cluster interference term is skipped. However, for high misalignment values the interference is considerable. This causes the extracted bound to be less accurate for higher misalignment.

Our HB-NOMA is compared with the traditional OMA technique in Fig. 5. We choose TDMA for OMA. To gain some insights, three different mmWave systems is evaluated. These systems are fully-digital beamforming, hybrid beamforming and analog beamforming. For fully-digital we assume $N_{\mathrm{BS}} = N_{\mathrm{RF}}$=32 which serve 8 clusters. Likewise, for hybrid beamforming we have N_{BS}=32 but N_{RF}=8. Both fully-digital and hybrid systems support 8 clusters of users. The first cluster has AoD of 10° and AoD of the next clusters increases by 10°. Further, the users inside of each clusters are distributed in a way that the maximum channel gain difference between the

strongest and weakest user is 18 dB. Indeed, the channel gain of the strongest user is 0 dB. The first cluster contains 4 users and each next cluster serves two users more than the previous cluster. Totally, thanks to NOMA technique, both systems support 88 users in each time slot. For OMA, we assume the analog beamforming system equipped with only one RF chain is able to serve one user per time slot. For $\mathrm{U}_{n,m}$, the achievable rate of OMA is $\log_2(1+P|\mathbf{w}_{n,m}\mathbf{H}_{n,m}\mathbf{f}_{\mathrm{RF}}|^2/\sigma^2)$. As expected fully-digital NOMA system achieves the highest sum-rate performance. The HB-NOMA with perfect alignment achieves approximately the same rate as the full-digital. For b=2, the misaligned HB-NOMA shows a very close performance to the perfect HB-NOMA. By increasing b, the performance slightly decreases. There is a huge rate difference between HB-NOMA and OMA. We conclude that, even in the presence of misalignment, HB-NOMA outperforms OMA.

VI. Conclusion

A hybrid beamforming-based NOMA has been designed for the downlink of a single-cell mmWave communication system. To study the achievable rate of an HB-NOMA user, we first formulated an optimization problem for the sum-rate of all users in the cell and then proposed an algorithm to solve it in three steps based on the strongest user precoder design. In order to evaluate the sum-rate, we found a lower bound for the achievable rate of each user under perfect and imperfect beam alignment between the effective channel of the users in each cluster. The lower bound analysis demonstrates that perfect HB-NOMA achieves a sum-rate close to that with fully-digital precoder. For the imperfect correlation, the relationship between the effective channels of the first user and other users inside a cluster was modeled. The bound for the misalignment shows that it is highly function of the mislaigned angle. Such that, a large misalignment angle can cause a significant reduction in the achievable rate. Further, for each user, the rate gap between the perfect and imperfect alignment is bounded. The simulation results confirmed our findings.

Appendix A
Proof of Theorem 1

Proof. Given the perfect alignment assumption and (14), the effective channel vector for $\mathrm{U}_{n,m}$ becomes

$$\bar{\mathbf{h}}_{n,m}^{\dagger} = \sqrt{N_{\mathrm{BS}}N_{\mathrm{U}}}\beta_{n,m}\mathbf{a}_{\mathrm{BS}}^{\dagger}(\varphi_{n,m})\mathbf{F}_{\mathrm{RF}} = \beta_{n,m}\beta_{n,1}^{-1}\bar{\mathbf{h}}_{n,1}^{\dagger}. \tag{32}$$

On the other hand, we have

$$\bar{\mathbf{h}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}=\begin{cases}\mathbf{\Gamma}_{n,n}, & \text{for } n,\ell=1,2,\ldots,N,\\ 0, & \text{for } \ell\neq n.\end{cases} \tag{33}$$

Therefore, using (32) and (33) the numerator in (7) becomes

$$P_{n,m}\left|\beta_{n,m}\right|^{2}\left|\beta_{n,1}\right|^{-2}\mathbf{\Gamma}_{n,n}^{2}. \tag{34}$$

Also, the intra-cluster interference in (8) becomes $I_{\text{intra}}^{n,m}=\sum_{k=1}^{m-1}P_{n,k}\left|\beta_{n,m}\right|^{2}\left|\beta_{n,1}\right|^{-2}\mathbf{\Gamma}_{n,n}^{2}$, and the inter-cluster interference term becomes zero, i.e., $I_{\text{inter}}^{n,m}=0$.

Now, substituting (34), and the determined $I_{\text{intra}}^{n,m}$ and $I_{\text{inter}}^{n,m}$ in (7) gives

$$\begin{aligned}\overline{R}_{n,m}&=\log_2\left(1+\frac{P_{n,m}\left|\beta_{n,m}\right|^{2}\left|\beta_{n,1}\right|^{-2}\mathbf{\Gamma}_{n,n}^{2}}{\sum_{k=1}^{m-1}P_{n,k}\left|\beta_{n,m}\right|^{2}\left|\beta_{n,1}\right|^{-2}\mathbf{\Gamma}_{n,n}^{2}+\sigma^{2}}\right)\\&\overset{(a)}{=}\log_2\left(1+\frac{P_{n,m}N_{\mathrm{BS}}N_{\mathrm{U}}\left|\beta_{n,m}\right|^{2}}{\sum_{k=1}^{m-1}P_{n,k}N_{\mathrm{BS}}N_{\mathrm{U}}\left|\beta_{n,m}\right|^{2}+\sigma^{2}\left(\mathbf{F}^{-1}\right)_{n,n}}\right)\\&\overset{(b)}{\geq}\log_2\left(1+\frac{P_{n,m}N_{\mathrm{BS}}N_{\mathrm{U}}\left|\beta_{n,m}\right|^{2}}{\sum_{k=1}^{m-1}P_{n,k}N_{\mathrm{BS}}N_{\mathrm{U}}\left|\beta_{n,m}\right|^{2}+\sigma^{2}\kappa_{\min}^{-1}(\mathbf{F})}\right),\end{aligned} \tag{35}$$

(a) follows by plugging (18) into the expression in the first line of (35) and using simple manipulations. To get (b), we note that $\mathbf{F}_{\mathrm{RF}}$ is full-rank matrix which means $\mathbf{F}=\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{RF}}^{\dagger}$ is positive definite. Then, we have $\left(\mathbf{F}^{-1}\right)_{n,n}\leq\kappa_{\max}\left(\mathbf{F}^{-1}\right)=\kappa_{\min}^{-1}\left(\mathbf{F}\right)$ in which $\kappa_{\max}(\cdot)$ and $\kappa_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of $(\cdot)$. □

APPENDIX B
PROOF OF LEMMA 1

Proof. Suppose that the effective channel vectors are fed back by using infinite-resolution codebooks. Also, let $\hat{\mathbf{h}}_{n,m}$ denote the normalized effective channel vector for $\mathrm{U}_{n,m}$, i.e.,

$$\hat{\tilde{\mathbf{h}}}_{n,m}=\frac{\tilde{\mathbf{h}}_{n,m}}{\|\tilde{\mathbf{h}}_{n,m}\|}. \tag{36}$$

The angle between two complex-valued vectors $\tilde{\mathbf{h}}_{n,m}$ and $\tilde{\mathbf{h}}_{n,1} \in V_{\mathbb{C}}$, denoted by Φ_{C}, is obtained as $\cos\Phi_{\text{C}} \triangleq \rho_{n,m} e^{j\omega_{n,m}} = \hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger} \hat{\tilde{\mathbf{h}}}_{n,m}$, where $(\rho_{n,m} \leq 1)$ is equal to $\rho_{n,m} = \cos\Phi_{\text{H}}(\hat{\tilde{\mathbf{h}}}_{n,1}, \hat{\tilde{\mathbf{h}}}_{n,m}) = \left| \hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger} \hat{\tilde{\mathbf{h}}}_{n,m} \right|$, in which $\Phi_{\text{H}}(\hat{\tilde{\mathbf{h}}}_{n,1}, \hat{\tilde{\mathbf{h}}}_{n,m})$, $0 \leq \Phi_{\text{H}} \leq \frac{\pi}{2}$, is the Hermitian angle between two complex-valued vectors $\tilde{\mathbf{h}}_{n,1}$ and $\tilde{\mathbf{h}}_{n,m}$ and $\omega_{n,m}$, $-\pi \leq \omega_{n,m} \leq \pi$, is called their pseudo-angle [45]. The factor $\rho_{n,m}$ describes the angle between the two lines in the complex-valued vector space $V_{\mathbb{C}}$ [45].

To ease the analysis, the angle $\omega_{n,m}$ is neglected [45]. Hence, we find the angle between two lines which are defined by the two vectors $\hat{\tilde{\mathbf{h}}}_{n,1}$ and $\hat{\tilde{\mathbf{h}}}_{n,m}$. Considering these two vectors as two lines in the space $V_{\mathbb{C}}$ would be optimistic. However, the simulation results reveal that the derived misalignment model is still effective. Such that, the extracted lower bound for the sum-rate using the misalignment model is close to the exact value of the sum-rate.

For $\ell = n$, the misalignment factor $\rho_{n,m}$ can be calculated as

$$\begin{aligned}
\rho_{n,m} \triangleq \left| \hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger} \hat{\tilde{\mathbf{h}}}_{n,m} \right| &\overset{(a)}{=} \frac{N_{\text{BS}} N_{\text{U}} \left| \beta_{n,m} \beta_{n,1} \mathbf{a}_{\text{BS}}^{\dagger}(\varphi_{n,m}) \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{RF}}^{\dagger} \mathbf{a}_{\text{BS}}(\varphi_{n,1}) \right|}{\|\tilde{\mathbf{h}}_{n,m}\| \|\tilde{\mathbf{h}}_{n,1}\|} \\
&\overset{(b)}{=} \frac{N_{\text{BS}} N_{\text{U}} \left| \beta_{n,m} \beta_{n,1} \mathbf{a}_{\text{BS}}^{\dagger}(\varphi_{n,m}) \mathbf{V}_1 \mathbf{\Lambda}_1 \mathbf{V}_1^{\dagger} \mathbf{a}_{\text{BS}}(\varphi_{n,1}) \right|}{\|\tilde{\mathbf{h}}_{n,m}\| \|\tilde{\mathbf{h}}_{n,1}\|} \\
&\overset{(c)}{=} \frac{\sum_{i=1}^{N} \kappa_i \left| \mathbf{a}_{\text{BS}}^{\dagger}(\varphi_{n,m}) \mathbf{v}_1^i \mathbf{v}_1^{i\dagger} \mathbf{a}_{\text{BS}}(\varphi_{n,1}) \right|}{\sqrt{\sum_{\ell=1}^{N} K_{N_{\text{BS}}}(\varphi_{\ell,1} - \varphi_{n,m})} \sqrt{\sum_{\ell=1}^{N} K_{N_{\text{BS}}}(\varphi_{\ell,1} - \varphi_{n,1})}}.
\end{aligned} \tag{37}$$

To get (a), the expression in (14) is used. To get (b), we apply SVD to the Hermitian matrix $\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{RF}}^{\dagger}$ which gives $\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{RF}}^{\dagger} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^{\dagger}$ where $\mathbf{V}$ of size $N_{\text{BS}} \times N_{\text{BS}}$ is a unitary matrix and $\mathbf{\Lambda}$ of size $N_{\text{BS}} \times N_{\text{BS}}$ is a diagonal matrix of singular values ordered in decreasing order. We then partition two matrices $\mathbf{V}$ and $\mathbf{\Lambda}$ as

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix}, \quad \mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{38}$$

where $\mathbf{V}_1$ is of size $N_{\text{BS}} \times N$ and $\mathbf{\Lambda}_1$ and is of size $N \times N$. We note that rank($\mathbf{F}_{\text{RF}}$)$= N$. Term (c) follows from the fact that $\mathbf{\Lambda}_1$ is a diagonal matrix with elements κ_i for $i = 1, 2, \ldots, N$. Notice that $\mathbf{v}_1^i$ represents the ith column.

For $\ell \neq n$, it is reasonable to assume that $\sqrt{1 - \rho_{n,m}^2}$ percentage of the amplitude of $\tilde{\mathbf{h}}_{n,m}$ leakages into the subspace generated by the other first users. To determine the subspace, we start

with considering the impact of the misalignment imposed by the other first users on $\mathrm{U}_{n,m}$, i.e, $\sum_{\ell=1,\ell\neq n}^{N}\left|\tilde{\mathbf{h}}_{\ell,1}^{\dagger}\tilde{\mathbf{h}}_{n,m}\right|^2$. Using the definition of vector norm, we rewrite this expression as following:

$$\begin{aligned}\sum_{\ell=1,\ell\neq n}^{N}\left|\tilde{\mathbf{h}}_{\ell,1}^{\dagger}\tilde{\mathbf{h}}_{n,m}\right|^2 &= \left\|\tilde{\mathbf{h}}_{n,m}^{\dagger}\left[\tilde{\mathbf{h}}_{1,1}\ \cdots\ \tilde{\mathbf{h}}_{n-1,1}\ \ \tilde{\mathbf{h}}_{n+1,1}\ \cdots\ \tilde{\mathbf{h}}_{N,1}\right]\right\|^2\\ &\overset{(a)}{=} N_{\mathrm{BS}}N_{\mathrm{U}}\Big\|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{F}_{\mathrm{RF}}^{\dagger}\big[\beta_{1,1}\mathbf{a}_{\mathrm{BS}}\left(\varphi_{1,1}\right)\ \cdots\ \beta_{n-1,1}\mathbf{a}_{\mathrm{BS}}\left(\varphi_{n-1,1}\right)\\ &\qquad\qquad \beta_{n+1,1}\mathbf{a}_{\mathrm{BS}}\left(\varphi_{n+1,1}\right)\ \cdots\ \beta_{N,1}\mathbf{a}_{\mathrm{BS}}\left(\varphi_{N,1}\right)\big]\Big\|^2\\ &\overset{(b)}{=} N_{\mathrm{BS}}N_{\mathrm{U}}\left\|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{F}_{\mathrm{RF}}^{\dagger}\mathbf{A}_{\mathrm{BS}}^{-n}\right\|^2. \end{aligned}\tag{39}$$

To get (a), we replace $\tilde{\mathbf{h}}_{\ell,1}$ by (14). Since $\mathbf{a}_{\mathrm{BS}}\left(\varphi_{n,1}\right)$s are independent vectors, $\mathbf{G}_{\mathrm{BS}}^{-n}=\sqrt{N_{\mathrm{BS}}N_{\mathrm{U}}}\mathbf{F}_{\mathrm{RF}}^{\dagger}\mathbf{A}_{\mathrm{BS}}^{-n}$ determines an $N-1$ dimensional subspace. We represent the weighted linear combination of $\hat{\tilde{\mathbf{h}}}_{\ell,1}^{\dagger}$ by a new vector $\mathbf{g}_{\mathrm{BS}}^{-n}$ which is located in the subspace $\mathbf{G}_{\mathrm{BS}}^{-n}$. So, we get $\mathbf{g}_{\mathrm{BS}}^{-n}=\sqrt{N_{\mathrm{BS}}N_{\mathrm{U}}}\mathbf{F}_{\mathrm{RF}}\times\sum_{\ell=1,\ell\neq n}^{N}\sqrt{P_\ell}\beta_{\ell,1}\mathbf{a}_{\mathrm{BS}}(\varphi_{\ell,1})$. To get (26), we only need to normalize $\mathbf{g}_{\mathrm{BS}}^{-n}$. □

APPENDIX C
PROOF OF THEOREM 2

Proof. Using (26), we obtain the following expressions. First,

$$\begin{aligned}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2 &= \rho_{n,m}^2\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\left(1-\rho_{n,m}^2\right)\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left|\mathbf{g}_{\mathrm{BS}}^{-n\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2\\ &\overset{(a)}{=}\rho_{n,m}^2\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2\overset{(b)}{=}\rho_{n,m}^2\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left\|\tilde{\mathbf{h}}_{n,1}\right\|^{-2}\mathbf{\Gamma}_{n,n}^2,\end{aligned}\tag{40}$$

in which (a) follows since $\mathbf{g}_{\mathrm{BS}}^{-n\dagger}\mathbf{f}_{\mathrm{BB}}^{n}=0$ and (b) follows from (33). Second,

$$\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2=\left(1-\rho_{n,m}^2\right)\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\mathbf{g}}_{\mathrm{BS}}^{-n\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2,\quad\text{for}\quad\ell\neq n.\tag{41}$$

Next, Using (18), (33), (25), (36), and (40), (8) becomes

$$I_{\text{intra}}^{n,m}=\sum_{k=1}^{m-1}P_{n,k}\rho_{n,m}^2N_{\mathrm{BS}}N_{\mathrm{U}}\left|\beta_{n,m}\right|^2\left(\mathbf{F}^{-1}\right)_{n,n}^{-1}K_{N_{\mathrm{BS}},m}K_{N_{\mathrm{BS}},1}^{-1},\tag{42}$$

where $K_{N_{\mathrm{BS}},1}$ and $K_{N_{\mathrm{BS}},m}$ are defined in (29). Likewise,using (33), (25), (36), and (41), (9) becomes

$$I_{\text{inter}}^{n,m}=\left(1-\rho_{n,m}^2\right)N_{\mathrm{BS}}N_{\mathrm{U}}\left|\beta_{n,m}\right|^2\sum_{\ell\neq n}^{N}P_\ell\left|\hat{\mathbf{g}}_{\mathrm{BS}}^{-n\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2K_{N_{\mathrm{BS}},m}.\tag{43}$$

Further, after substituting (40), (42) and (43) into (7), we get

$$\tilde{R}_{n,m} = \log_2\left(1 + \frac{\Psi}{I_{\text{intra}}^{n,m} + I_{\text{inter}}^{n,m} + \sigma^2}\right) \overset{(a)}{\geq} \log_2\left(1 + \frac{\Psi}{I_{\text{intra}}^{n,m} + \varsigma_{\text{inter}}^{n,m} + \sigma^2}\right), \tag{44}$$

where $\Psi = P_{n,m}\rho_{n,m}^2 N_{\text{BS}} N_{\text{U}} \left|\beta_{n,m}\right|^2 \left(\mathbf{F}^{-1}\right)_{n,n}^{-1} K_{N_{\text{BS}},m} K_{N_{\text{BS}},1}^{-1}$, and $\varsigma_{\text{inter}}^{n,m} = \left(1-\rho_{n,m}^2\right) N_{\text{BS}} N_{\text{U}} \left|\beta_{n,m}\right|^2 \times \kappa_{\max}(\mathbf{S}) K_{N_{\text{BS}},m}$. To get (a), we have the following lemma.

Lemma 2. An upper bound of $\sum_{\ell=1,\ell\neq n}^{N} P_\ell \left|\hat{\mathbf{g}}_{\text{BS}}^{-n\dagger} \mathbf{f}_{\text{BB}}^{\ell}\right|^2$ is the maximum eigenvalue of $\mathbf{S}$, i.e., $\kappa_{\max}(\mathbf{S})$.

Proof. We rewrite $\sum_{\ell=1,\ell\neq n}^{N} P_\ell \left|\hat{\mathbf{g}}_{\text{BS}}^{-n\dagger} \mathbf{f}_{\text{BB}}^{\ell}\right|^2 = \left\|\mathbf{g}_{\text{BS}}^{-n\dagger} \mathbf{F}_{\text{BB}}^{-n,W}\right\|_2^2$. Maximizing $\left\|\hat{\mathbf{g}}_{\text{BS}}^{-n\dagger} \mathbf{F}_{\text{BB}}^{-n,W}\right\|_2^2$ given $\left\|\hat{\mathbf{g}}_{\text{BS}}^{-n}\right\| = 1$ is similar to maximizing a beamforming vector for maximum ratio transmission systems [46], [47]. Hence, the maximum value of $\hat{\mathbf{g}}_{\text{BS}}^{-n}$ is the dominant right singular vector of $\mathbf{F}_{\text{BB}}^{-n,W}$ [46], [47]. Thus, the maximum of $\left\|\hat{\mathbf{g}}_{\text{BS}}^{-n\dagger} \mathbf{F}_{\text{BB}}^{-n,W}\right\|_2^2$ is equal to the maximum eigenvalue of $\mathbf{S}$. □

Lemma 2 indicates that $I_{\text{inter}}^{n,m} \leq \varsigma_{\text{inter}}^{n,m}$. After some manipulations

$$\begin{aligned}\tilde{R}_{n,m} &\geq \log_2\left(1 + \frac{P_{n,m}\rho_{n,m}^2 N_{\text{BS}} N_{\text{U}} \left|\beta_{n,m}\right|^2}{\zeta_{\text{intra}}^{n,m} + \left(\varsigma_{\text{inter}}^{n,m} + \sigma^2\right)\left(\mathbf{F}^{-1}\right)_{n,n} K_{N_{\text{BS}},m}^{-1} K_{N_{\text{BS}},1}}\right) \\ &\overset{(a)}{\geq} \log_2\left(1 + \frac{P_{n,m}\rho_{n,m}^2 N_{\text{BS}} N_{\text{U}} \left|\beta_{n,m}\right|^2}{\zeta_{\text{intra}}^{n,m} + \zeta_{\text{inter}}^{n,m} + \sigma^2 \kappa_{\min}^{-1}(\mathbf{F}) K_{N_{\text{BS}},m}^{-1} K_{N_{\text{BS}},1}}\right),\end{aligned} \tag{45}$$

where in the first line, $\zeta_{\text{intra}}^{n,m} = \sum_{k=1}^{m-1} P_{n,k}\rho_{n,m}^2 N_{\text{BS}} N_{\text{U}} \left|\beta_{n,m}\right|^2$ and in the second line, $\zeta_{\text{inter}}^{n,m} = \left(1-\rho_{n,m}^2\right) \times N_{\text{BS}} N_{\text{U}} \left|\beta_{n,m}\right|^2 \kappa_{\max}(\mathbf{S}) \kappa_{\min}^{-1}(\mathbf{F}) K_{N_{\text{BS}},1}$. To get (a), we note that $\left(\mathbf{F}^{-1}\right)_{n,n} \leq \kappa_{\min}^{-1}(\mathbf{F})$. □

APPENDIX D
PROOF OF THEOREM 3

Proof. We start with (7) to define the achievable rate of $\text{U}_{n,m}$ for the perfect correlation and the imperfect correlation, i.e., $\overline{R}_{n,m}$ and $\tilde{R}_{n,m}$, respectively. This gives

$$\begin{aligned}\Delta R_{n,m} &\triangleq \overline{R}_{n,m} - \tilde{R}_{n,m} \\ &= \log_2\left(1 + \frac{P_{n,m}\left|\bar{\mathbf{h}}_{n,m}^{\dagger} \mathbf{f}_{\text{BB}}^{n}\right|^2}{\sum_{k=1}^{m-1} P_{n,k}\left|\bar{\mathbf{h}}_{n,m}^{\dagger} \mathbf{f}_{\text{BB}}^{n}\right|^2 + \sigma^2}\right) -\end{aligned}$$

$$\log_2\left(1+\frac{P_{n,m}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2}{\sum_{k=1}^{m-1}P_{n,k}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\sigma^2}\right)$$

$$=\log_2\left(\frac{\sum_{k=1}^{m}P_{n,k}\left|\bar{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sigma^2}{\sum_{k=1}^{m-1}P_{n,k}\left|\bar{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sigma^2}\right)-\log_2\left(\frac{\sum_{k=1}^{m}P_{n,k}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\sigma^2}{\sum_{k=1}^{m-1}P_{n,k}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\sigma^2}\right)$$

$$\overset{(a)}{\leq}\log_2\left(\frac{\sum_{k=1}^{m}P_{n,k}\left|\bar{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sigma^2}{\sum_{k=1}^{m}P_{n,k}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sigma^2}\right)-\log_2\left(\frac{\sum_{k=1}^{m-1}P_{n,k}\left|\bar{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sigma^2}{\sum_{k=1}^{m-1}P_{n,k}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\sigma^2}\right)$$

$$\overset{(b)}{\leq}\log_2\left(\frac{\left\|\bar{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\bar{\mathbf{h}}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2}{\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\tilde{\mathbf{h}}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2}\right)-\log_2\left(\frac{\left\|\bar{\mathbf{h}}_{n,m}\right\|^2\sum_{k=1}^{m-1}P_{n,k}\left|\hat{\bar{\mathbf{h}}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+1}{\Upsilon}\right), \tag{46}$$

where $\Upsilon=\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\sum_{k=1}^{m-1}P_{n,k}\left|\hat{\tilde{\mathbf{h}}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\hat{\tilde{\mathbf{h}}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\sigma^2$. To get (a) we remove positive quantity $\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\tilde{\mathbf{h}}_{n,m}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2$ from the second term. Then, we exchange the denominator of the first term with the numerator of the second one. (b) follows from the fact that for $u>v$, it gives $\log\left(\frac{u}{v}\right)>\log\left(\frac{u+c}{v+c}\right)$ $(c>0)$, and applying the normalized vector $\tilde{\mathbf{h}}_{n,m}$ defined in (36) for both perfect and imperfect effective channel vectors.

Noting that $\hat{\bar{\mathbf{h}}}_{n,1}=\hat{\bar{\mathbf{h}}}_{n,m}$ and using (40) it yields

$$\Delta R\leq\log_2\left(\frac{\left\|\bar{\mathbf{h}}_{n,m}\right\|^2}{\rho_{n,m}^2\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2}\right)-\log_2\left(\sum_{k=1}^{m-1}P_{n,k}\left\|\bar{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\bar{\mathbf{h}}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+\sigma^2\right)$$
$$+\log_2\left(\sum_{k=1}^{m-1}P_{n,k}\rho_{n,m}^2\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\left|\hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+(1-\rho_{n,m}^2)\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\hat{\mathbf{g}}_{\mathrm{BS}}^{-n\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\sigma^2\right)$$
$$\overset{(a)}{=}-\log_2\left(\sum_{k=1}^{m-1}P_{n,k}\rho_{n,m}^2\left|\hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2\right)$$
$$+\log_2\left(\sum_{k=1}^{m-1}P_{n,k}\rho_{n,m}^2\left|\hat{\tilde{\mathbf{h}}}_{n,1}^{\dagger}\mathbf{f}_{\mathrm{BB}}^{n}\right|^2+(1-\rho_{n,m}^2)\sum_{\ell=1,\ell\neq n}^{N}P_{\ell}\left|\hat{\mathbf{g}}_{\mathrm{BS}}^{-n\dagger}\mathbf{f}_{\mathrm{BB}}^{\ell}\right|^2+\frac{\sigma^2}{\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2}\right)$$

$$\overset{(b)}{\leq} \log_2\left(1+\frac{\left(1-\rho_{n,m}^2\right)\kappa_{\max}\left(\mathbf{S}\right)+\sigma^2\left\|\tilde{\mathbf{h}}_{n,m}\right\|^{-2}}{\rho_{n,m}^2 K_{N_{\mathrm{BS}},1}^{-1}\left(\mathbf{F}^{-1}\right)_{n,n}^{-1}\sum_{k=1}^{m-1}P_{n,k}}\right)$$

$$\overset{(c)}{\leq} \log_2\left(1+\frac{\left(1-\rho_{n,m}^2\right)\kappa_{\max}\left(\mathbf{S}\right)+\sigma^2 K_{N_{\mathrm{BS}},m}^{-1}N_{\mathrm{BS}}^{-1}N_{\mathrm{U}}^{-1}\left|\beta_{n,m}\right|^{-2}}{\rho_{n,m}^2 K_{N_{\mathrm{BS}},1}^{-1}\kappa_{\min}(\mathbf{F})\sum_{k=1}^{m-1}P_{n,k}}\right), \tag{47}$$

in which (a) follows by rewriting the first term as $\log_2\left(\rho_{n,m}^{-2}\left\|\bar{\mathbf{h}}_{n,m}\right\|^2\right)-\log_2\left(\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\right)$. Then, we sum up the expression $\log_2\left(\rho_{n,m}^{-2}\left\|\bar{\mathbf{h}}_{n,m}\right\|^2\right)$ with the second term and the expression $-\log_2\left(\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2\right)$ with the third term. To get (b), we again sum up the first term with the second term. We then use Lemma 2 to get $\kappa_{\max}(\mathbf{S})$ and (33) and (18) to get $K_{N_{\mathrm{BS}},1}^{-1}\left(\mathbf{F}^{-1}\right)_{n,n}^{-1}$. To obtain (c), first we use $\left\|\tilde{\mathbf{h}}_{n,m}\right\|^2=K_{N_{\mathrm{BS}},m}N_{\mathrm{BS}}N_{\mathrm{U}}|\beta_{n,m}|^2$. Next we use the inequality $\left(\mathbf{F}^{-1}\right)_{n,n}\leq\kappa_{\min}^{-1}(\mathbf{F})$. □

REFERENCES

[1] M. A. Almasi and H. Mehrpouyan, "Non-orthogonal multiple access based on hybrid beamforming for mmwave systems," *in Proc. IEEE Veh. Technol. Conf., Fall*, Aug. 2018.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[3] T. S. Rappaport, R. W. Heath Jr, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Pearson Education, 2014.

[4] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka *et al.*, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[5] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[6] J. Kim and I. Lee, "802.11 WLAN: history and new enabling MIMO techniques for next generation standards," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 134–140, Mar. 2015.

[7] O. El Ayach, R. W. Heath, S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems," *in Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 3724–3729, Jun. 2012.

[8] O. El 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.

[9] 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, Jul. 2013.

[10] A. M. Sayeed and N. Behdad, "Continuous aperture phased MIMO: A new architecture for optimum line-of-sight links," *in Proc. IEEE Int. Symp. Antennas Propagation*, pp. 293–296, Jul. 2011.

[11] M. A. Almasi, H. Mehrpouyan, V. Vakilian, N. Behdad, and H. Jafarkhani, "Reconfigurable antennas in mmWave MIMO systems," *arXiv preprint arXiv:1710.05111*, 2017.

[12] ——, "A new reconfigurable antenna MIMO architecture for mmWave communication," *in Proc. IEEE Int. Conf. on Commun. (ICC)*, May 2018.

[13] A. Alkhateeb, R. W. Heath, and G. Leus, "Achievable rates of multi-user millimeter wave systems with hybrid precoding," *in Proc. IEEE Int. Conf. on Commun. Workshop (ICC Workshops)*, pp. 1232–1237, Jun. 2015.

[14] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[15] M. A. Almasi, H. Mehrpouyan, D. Matolak, C. Pan, and M. Elkashlan, "Reconfigurable antenna multiple access for 5G mmwave systems," *in Proc. IEEE Int. Conf. on Commun. Workshops (ICC Workshops)*, pp. 1–6, May 2018.

[16] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," *in Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, pp. 611–615, Sep. 2013.

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

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

[19] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, Jun. 2015.

[20] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, 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.

[21] 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.

[22] 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.

[23] ——, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84–87, Jan. 2017.

[24] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple Access Techniques for 5G Wireless Networks and Beyond*. Springer, 2018.

[25] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[26] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmWave systems," *arXiv preprint arXiv:1806.04919*, 2018.

[27] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, 2017.

[28] 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.*, Aug. 2017.

[29] 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.

[30] 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.

[31] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X.-G. Xia, "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.

[32] W. Wu and D. Liu, "Non-orthogonal multiple access based hybrid beamforming in 5G mmWave systems," *in Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, pp. 1–7, Oct. 2017.

[33] 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.*, pp. 1–1, Oct. 2018.

[34] J. Wildman, P. H. J. Nardelli, M. Latva-aho, and S. Weber, "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.

[35] M. Min, D. Kim, H. Kim, and G. Im, "Opportunistic two-stage feedback and scheduling for MIMO downlink systems," *IEEE Trans. Commun.*, vol. 61, no. 1, pp. 312–324, Jan. 2013.

[36] G. Lee and Y. Sung, "A new approach to user scheduling in massive multi-user MIMO broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1481–1495, Apr. 2018.

[37] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *arXiv preprint arXiv:1809.07224*, 2018.

[38] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1717–1731, Apr. 2018.

[39] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[40] Q. Zhang, Q. Li, and J. Qin, "Robust beamforming for nonorthogonal multiple-access systems in MISO channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 231–10 236, Dec. 2016.

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

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

[43] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, "Exploiting multiple-antenna techniques for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2207–2220, Oct. 2017.

[44] R. Strichartz, *The Way of Analysis*, ser. Jones and Bartlett books in mathematics. Jones and Bartlett Publishers, 2000.

[45] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Mathematica*, vol. 69, no. 1, pp. 95–103, 2001.

[46] D. J. Love, R. W. Heath, and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[47] P. A. Dighe, R. K. Mallik, and S. S. Jamuar, "Analysis of transmit-receive diversity in Rayleigh fading," *IEEE Trans. Commun.*, vol. 51, no. 4, pp. 694–703, Apr. 2003.