Rate Analysis of Cell-Free Massive MIMO-NOMA With Three Linear Precoders

Fatemeh Rezaei, *Student Member, IEEE*, Chintha Tellambura, *Fellow, IEEE*, Ali Akbar Tadaion, *Senior Member, IEEE*, and Ali Reza Heidarpour, *Student Memeber, IEEE*

***Abstract*— Although the hybrid of cell-free (CF) massive multiple-input multiple-output (MIMO) and non-orthogonal multiple access (NOMA) promises massive spectral efficiency gains, the type of precoders employed at the access points (APs) impacts the gains. In this paper, we thus comprehensively evaluate the system performance with maximum ratio transmission (MRT), full-pilot zero-forcing (fpZF) and modified regularized ZF (mRZF) precoders. We derive their closed-form sum rate expressions by considering Rayleigh fading channels, the effects of intra-cluster pilot contamination, inter-cluster interference, and imperfect successive interference cancellation (SIC). Our results reveal that this system supports significantly more users simultaneously at the same coherence interval compared to its OMA equivalent. However, intra-cluster pilot contamination and imperfect SIC degrade the system performance when the number of users is low. Moreover, with perfect SIC, mRZF and fpZF significantly outperform MRT. Also, we show that this system with either mRZF or fpZF precoding outperforms OMA systems with MRT. The analytical findings are verified by numerical results.**

***Index Terms*— NOMA, cell-free massive MIMO, MRT, fpZF, modified RZF, achievable sum rate.**

I. Introduction

A. Background and Scope

INTERNATIONAL mobile telecommunications (IMT)-2020 standard is envisioned to support peak data rates up to 10 Gb/s for low mobility users and 1 Gb/s for high mobility users, 1 ms over-the-air latency, and a network energy efficiency improvement of 100× of 4G networks [1]. These requirements have necessitated the development of new wireless technologies that will offer higher data rates, higher energy efficiency, and ultra low latency.

One such technology is the use of distributed large antenna arrays, which can increase the ergodic sum rate [2], extend coverage [3] and improve energy efficiency [4]. Thus, distributed access points (APs) based cellular networks have recently been studied [2], [5]. Distributed APs enable the efficient utilization of spatial resources [6]. However, inter-cell interference, inherent in all cell-centric networks, becomes a major performance limiting factor [7]. To overcome this and preserve the main benefits of massive MIMO (mMIMO), cell-free (CF) mMIMO has thus been proposed [8], [9]. It consists of a large number of spatially-distributed APs to serve many single-antenna users in the same time-frequency resources [8]. Thus, each user is served by all the APs it can reach, and hence, it does not experience cell boundaries (cell-free). A central processing unit (CPU) coordinates the APs, which are connected to it through a fronthaul network. Locality of the operations of each AP is another key idea, which minimizes the fronthaul overhead. Thus, each AP performs precoding based on only the estimates of the channels from itself to all the users. These downlink channels can be estimated with the help of pilots transmitted by the users in the uplink, thereby exploiting the channel reciprocity inherent in time-division duplexing (TDD). This architecture offers increased macro-diversity and favorable propagation with negligible inter-user interference. And it outperforms the conventional cellular counterparts by exploiting the best of both collocated mMIMO and network MIMO systems [8]–[12].

Precoder design, power control, hardware impairments and other factors for CF mMIMO systems have been investigated. For instance, [8] and [9] investigate conjugate beamforming (CB) (which maximizes the signal gain at the intended user), and zero-forcing (ZF) (which nulls the inter-user interference at the expense of gain losses), and they show that significant spectral efficiency gains are achievable over conventional small-cell systems. Joint power control and load balancing using ZF and maximum ratio combining (MRC) processing is also investigated [13]. The spectral efficiency losses due to both user and AP hardware impairments have also been studied [14]; the key insight is that the negative effects of hardware impairment vanish as the number of APs increases.

Non-orthogonal multiple access (NOMA) also achieves high spectral efficiency gains [15], [16]. It however is a paradigm shift from conventional orthogonal multiple access (OMA) techniques such as time division multiple access and frequency division multiple access, where an orthogonal channel must be created for each different user. In contrast, NOMA serves multiple users simultaneously on each orthogonal channel. This can be done by exploiting channel gain disparities and then performing individual user signal detection via successive interference cancellation (SIC) [15].

The hybrid of **co-located** mMIMO and NOMA offers a great potential to support low latency massive connectivity

Manuscript received July 11, 2019; revised November 13, 2019 and January 20, 2020; accepted February 23, 2020. Date of publication March 4, 2020; date of current version June 16, 2020. The associate editor coordinating the review of this article and approving it for publication was B. Shim. *(Corresponding author: Ali Akbar Tadaion.)*

Fatemeh Rezaei is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada, and also with the Department of Electrical Engineering, Yazd University, Yazd 81746-73441, Iran.

Chintha Tellambura and Ali Reza Heidarpour are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada.

Ali Akbar Tadaion is with the Department of Electrical Engineering, Yazd University, Yazd 81746-73441, Iran.

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

Digital Object Identifier 10.1109/TCOMM.2020.2978189

requirements of the next-generation wireless networks while further improving the spectral efficiency of NOMA-based systems [17]–[19]. Thus, integrating NOMA with CF mMIMO may reap further gains and is therefore a critically important research topic.

B. Problem Statement and Contributions

Hybrid CF massive MIMO-NOMA offers tremendous potential to improve the spectral efficiency. To achieve this goal, two key factors are essential. First, precoding of superposition-coded signals to multiple user-clusters is necessary. Second, the user multiplexing in each cluster must be on the basis of their channel power gain differentials. However, so far such hybrid designs have been investigated by only few papers [20]–[22]. Thus, the main objective of this paper is to add to this literature and further investigate these systems.

In particular, for CF massive MIMO-NOMA systems, we investigate three linear precoding schemes, with the same front-hauling overhead. These are maximum ratio transmission (MRT), full-pilot zero-forcing (fpZF) and modified regularized ZF (mRZF). Of these three, while its performance in mMIMO systems has been investigated [23]–[25], mRZF has not been studied for CF massive MIMO before. The advantages of mRZF include its ability to balance the interference suppression and desired signal power and also having additional parameters to be optimized.

Normal ZF and RZF precoders require exchanging instantaneous channel state information (CSI) among the APs – the main benefit of mRZF and fpZF is the elimination of this exchange and that each AP computes its precoder with only its local CSI. This is a big advantage as mRZF and fpZF have the same front-hauling overhead as MRT [26], [27]. Unlike MRT which maximizes the signal gain at the intended cluster and ignores the inter-cluster interference, fpZF sacrifices some of the array gain to cancel the inter-cluster interference [26], [27]. On the other hand, mRZF balances the inter-cluster interference mitigation and intra-cluster power enhancement. Note that the optimal precoding is nonlinear, so these linear precoders are sub-optimal. But they offer high performance with affordable computational complexity, making them ideal for practical large MIMO systems [23].

The main contributions of this paper on CF massive MIMO-NOMA systems can be summarized as follows:

1) We derive the closed-form downlink sum rate when the APs employ MRT or fpZF precoders, considering the effects of intra-cluster pilot contamination, and inter-cluster interference. The users rely on the statistics for their effective channels for decoding, and the SIC process is imperfect.
2) We also analyze the performance when the APs employ mRZF precoding. Since the closed-form analysis of it with finite system parameters is difficult (if not impossible), we analyze the achievable rate when the number of clusters (N) and the number of antennas at each AP (L) grow infinitely large while $\frac{L}{N}$ is a finite ratio.
3) We show that NOMA allows a significant number of users to be supported simultaneously at the same coherence interval compared to its counterpart OMA. For instance, with K users in each cluster, NOMA based CF mMIMO can support K times the users that of OMA. Moreover, for a large number of users, NOMA outperforms OMA; however, for low number of users, the sum rate of NOMA is lower than that of OMA due to the effects of intra-cluster pilot contamination and imperfect SIC. It is further shown that given perfect SIC, mRZF and fpZF significantly outperform MRT. We also show that either mRZF or fpZF outperforms OMA systems with MRT.
4) Numerical results are also presented to support our findings.

C. Previous Contributions on CF Massive MIMO-NOMA

As mentioned before, the investigation of CF massive MIMO-NOMA systems has been sparse, except for [20]–[22]. In fact [20] is the first paper to study the design of such systems. It considers single-antenna APs that use conjugate beamforming precoders. It thus analyzes degradations due to estimated/imperfect CSI at APs, SIC, and statistical CSI at users. On the other hand, [21] generalizes to multiple-antenna APs and derives minimum mean square error (MMSE) uplink and downlink channel estimates, and for both these cases, the achievable rate is derived. Reference [21] also considers spatial correlation among the multiple antennas at each AP and intra-cluster pilot contamination and quantifies the adverse impact of intra-cluster pilot contamination and error propagation due to imperfect SIC at the user nodes. The only other work [22] deals with the max-min fairness based bandwidth efficiency problem. It develops an optimal algorithm and mode switching between NOMA and OMA for maximum bandwidth efficiency.

The above paragraph clarifies the differences between this paper and [20]–[22]. The focus of this paper is to comparatively evaluate the three types of practical precoders. We believe that this paper is the first to do that in the context of CF massive MIMO-NOMA.

D. General Works on MIMO NOMA

Since the NOMA literature is vast, we mention only few works here. NOMA outperforms OMA from user and system throughput perspectives [28]. In [29], multi-user power allocation, user scheduling, and error propagation in SIC for NOMA are investigated. MIMO-NOMA outperforms conventional MIMO-OMA systems [30], [31]. MIMO-NOMA outperforms MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity [30]. Furthermore, the user outage probability in a MIMO-NOMA cluster is studied in [31]. Beamforming, user clustering, and power allocation for MIMO-NOMA are further investigated in [32].

Hybrid mMIMO-NOMA, which outperforms standalone mMIMO and NOMA schemes, has been studied in [19], [33]. The outage probability with perfect user ordering and limited feedback is investigated in [34]. In [17], a user pairing and pair scheduling algorithm is proposed to enhance the spectral efficiency of mMIMO-NOMA systems. Moreover, in [18], user clustering and pilot assignment schemes for

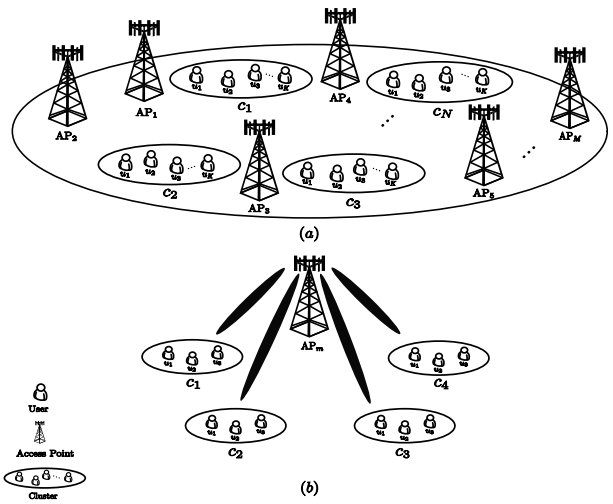

Fig. 1. (a) System model of cell-free massive MIMO-NOMA (b) An AP transmits to four clusters $\{c_1, \ldots, c_4\}$, each with three users.

multi-cell mMIMO-NOMA are investigated by employing the characteristics of correlated fading channels.

E. Structure and Notations

This paper is organized as follows. The system model is introduced in Section II. In Section III, the achievable sum rate with linear precoding techniques is derived. In Section IV, the analytical results are confirmed through simulation examples. Section V concludes the paper.

Notation: Lower-case bold and upper-case bold denote vectors and matrices, respectively. $\mathbf{I}_n$ represents the $n \times n$ identity matrix. $\mathbf{A}^{\mathrm{T}}$, $\mathbf{A}^{\mathrm{H}}$, tr[$\mathbf{A}$], and $[\mathbf{A}]_{(m,n)}$ denote transpose, Hermitian transpose, trace, and the (m,n)th element of matrix $\mathbf{A}$, respectively. $\mathbb{E}\{\cdot\}$ denotes the statistical expectation. Finally, $\mathcal{CN}(\boldsymbol{\mu}, \mathbf{R})$ is a complex Gaussian vector with mean $\boldsymbol{\mu}$ and co-variance matrix $\mathbf{R}$.

II. System Model and Preliminaries

Here, we describe the system model, channel model, and transmission model in detail.

A. System and Channel Models

Consider the downlink transmission of CF mMIMO-NOMA. This system has M APs and KN single-antenna users, which are grouped into N clusters with $K(K \geq 2)$ users per cluster and NOMA is applied among the users in the same cluster. Each AP is equipped with L antennas (Fig. 1.a). The total M APs serve KN users in the same time-frequency resource block, where $KN \ll ML$. All the APs are connected to a CPU via an error-free fronthaul network to achieve coherent processing [8]. This network carries only the payload data, large-scale parameters and power control coefficients that change slowly. Each AP computes its precoder based on the estimates of the channel states between itself and the users. Importantly, it does not need the knowledge of the channel states between the users and other APs.

The downlink channel between the mth AP ($m = 1, \ldots, M$) and the kth user ($k = 1, \ldots, K$) in the nth cluster ($n = 1, \ldots, N$) is the complex Gaussian random vector

$$\mathbf{h}_{mnk} \sim \mathcal{CN}\left(\mathbf{0}, \beta_{mnk}\mathbf{I}_L\right), \tag{1}$$

where $\{\beta_{mnk}\}$ are the set of the large-scale fading coefficients. Each AP is assumed to know its own set of large-scale coefficients [8]. This assumption is justifiable because they are quasi-static and hence only need to be estimated once about every 40 coherence time intervals [35]. This model (1) also presumes that small-scale fading is Rayleigh distributed.

B. Uplink Pilot Transmission

The users transmit pilot sequences in the uplink, and the APs use them to estimate the downlink channels. In order to minimize the channel estimation overhead in NOMA, the same pilot sequence of length τ samples, is shared among the users within each cluster, but different clusters are assigned mutually orthogonal pilots[1] [20]. Then, the pilot sequence for the kth user in the nth cluster is $\sqrt{\tau}\boldsymbol{\phi}_n \in \mathcal{C}^{\tau \times 1}$ satisfying $\|\boldsymbol{\phi}_n\|^2 = 1$ where $N \leq \tau$. Accordingly, the pilot sequences allocated to the N clusters are mutually orthogonal implying that $\boldsymbol{\phi}_n^{\mathrm{H}}\boldsymbol{\phi}_j = 0$ $(n \neq j)$. The received pilot signal at the mth AP ($\mathbf{Y}_m^p \in \mathcal{C}^{L \times \tau}$) can then be expressed as

$$\mathbf{Y}_m^p = \sqrt{\tau p_p}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{h}_{mnk}\boldsymbol{\phi}_n^{\mathrm{H}} + \mathbf{N}_m, \quad \forall m \tag{2}$$

where p_p is the pilot transmit power and $\mathbf{N}_m \in \mathcal{C}^{L \times \tau}$ is a Gaussian noise matrix with i.i.d $\mathcal{CN}(0,1)$ elements. Then, the mth AP estimates $\mathbf{h}_{mnk}$ using minimum mean square error (MMSE) estimation. To do so, the received pilot signal at the mth AP (2) is projected onto $\boldsymbol{\phi}_n$ which yields

$$\tilde{\mathbf{y}}_{mn}^p = \sqrt{\tau p_p}\sum_{k=1}^{K}\mathbf{h}_{mnk} + \tilde{\mathbf{n}}_{mn}, \tag{3}$$

where $\tilde{\mathbf{n}}_{mn} = \mathbf{N}_m\boldsymbol{\phi}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$.

By using (3), the MMSE estimate of $\mathbf{h}_{mnk}$ can be expressed as $\hat{\mathbf{h}}_{mnk} = c_{mnk}\tilde{\mathbf{y}}_{mn}^p$, where c_{mnk} is given by

$$c_{mnk} = \frac{\sqrt{\tau p_p}\beta_{mnk}}{1 + \tau p_p \sum\limits_{i=1}^{K}\beta_{mni}}. \tag{4}$$

Since $\tilde{\mathbf{y}}_{mn}^p$ is Gaussian distributed, $\hat{\mathbf{h}}_{mnk}$ can be written as,

$$\hat{\mathbf{h}}_{mnk} = \sqrt{\theta_{mnk}}\nu_{mn}, \tag{5}$$

where $\nu_{mn} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$ and θ_{mnk} is equal to

$$\theta_{mnk} = \frac{1}{L}\mathbb{E}\left\{\left\|\hat{\mathbf{h}}_{mnk}\right\|^2\right\} = \frac{\tau p_p \beta_{mnk}^2}{1 + \tau p_p \sum\limits_{i=1}^{K}\beta_{mnk}}. \tag{6}$$

The channel estimation error can then be defined as $\boldsymbol{\epsilon}_{mnk} = \mathbf{h}_{mnk} - \hat{\mathbf{h}}_{mnk}$ where $\boldsymbol{\epsilon}_{mnk} \sim \mathcal{CN}(\mathbf{0}, (\beta_{mnk} - \theta_{mnk})\mathbf{I}_L)$. Since the users in the same cluster use the same pilot sequence, their

[1]It is different from OMA where all the users in all the clusters are assigned with different mutually orthogonal pilots.

channels are parallel as shown in (5). Mathematically, this can be written as

$$\hat{\mathbf{h}}_{mnk} = \frac{\beta_{mnk}}{\beta_{mni}} \hat{\mathbf{h}}_{mni}, \tag{7}$$

where β_{mnk} is already given by (1).

C. Downlink Data Transmission Model

The superposition coded data signal for the K users in the nth cluster is expressed as

$$s_n = \sum_{k=1}^{K} \sqrt{p_{nk}} s_{nk}, \quad \forall n. \tag{8}$$

In (8), s_{nk} and p_{nk} denote the data signal and the transmitted power allocated to the kth user in the nth cluster. Also, $p_{nk} = p_t \lambda_{nk}$ where p_t is the total transmit power of each AP and $\{\lambda_{nk}\}$ are the set of power coefficients that satisfies $\sum_{n=1}^{N} \sum_{k=1}^{K} \lambda_{nk} = 1$. Furthermore, the different data signals are mutually uncorrelated:

$$\mathbb{E}\{s_{nk} s_{mi}^*\} = \begin{cases} 1, & m = n \ \& \ k = i \\ 0, & \text{else}, \end{cases} \tag{9}$$

where $m, n \in \{1, 2, \ldots, N\}$ and $k, i \in \{1, 2, \ldots, K\}$. Therefore, $\mathbb{E}\{|s_n|^2\} = \sum_{k=1}^{K} p_{nk} = p_t \lambda_n$, where $\lambda_n = \sum_{k=1}^{K} \lambda_{nk}$ accounts for the power allocation coefficient for the nth cluster.

The mth AP ($m = 1, \ldots, M$) transmits the signal

$$\mathbf{x}_m = \sum_{n=1}^{N} \mathbf{w}_{mn} s_n, \tag{10}$$

where $\mathbf{w}_{mn} \in \mathcal{C}^L$ represents the spatial directivity of the signals sent to the users in the nth cluster. Note that each AP precodes the transmitted signals for all the users in the same cluster with the same beamforming vector $\mathbf{w}_{mn}$, i.e., each AP has N precoding vectors (Fig. 1.b).

Since the KN users are served simultaneously by M APs, the received signal at the kth user in the nth cluster can be expressed as

$$y_{nk} = \underbrace{\sum_{m=1}^{M} \sqrt{p_{nk}} \mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn} s_{nk}}_{\text{desired signal}} + \underbrace{\sum_{m=1}^{M} \mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn} \sum_{\substack{i=1 \\ i \neq k}}^{K} \sqrt{p_{ni}} s_{ni}}_{\text{intra-cluster interference before SIC}} + \underbrace{\sum_{m=1}^{M} \mathbf{h}_{mnk}^{\mathrm{H}} \sum_{\substack{n'=1 \\ n' \neq n}}^{N} \mathbf{w}_{mn'} s_{n'}}_{\text{inter-cluster interference}} + n_{nk}. \tag{11}$$

In (11), $n_{nk} \sim \mathcal{CN}(0, 1)$ while the first, second, and the third terms are desired signal, intra-cluster interference before SIC and inter-cluster interference, respectively.

In TDD CF massive MIMO systems, with sufficiently large number antennas at each AP (e.g., $5 - 10$ antennas per AP [36]), the instantaneous channel coefficients can be well approximated by their corresponding expected values. This phenomenon is referred to as "*channel hardening*" which substantially reduces the effective channel gain fluctuations; thereby rendering the use of downlink pilot training unnecessary [8], [36]. To apply power domain NOMA, we thus assume that the users in the nth cluster are ordered based on the mean of the effective channel gains as follows [20], [21]:

$$\Gamma_{n1} \geq \Gamma_{n2} \geq \cdots \geq \Gamma_{nK}, \tag{12}$$

where $\Gamma_{nk} = \mathbb{E}\left\{\left|\sum_{m=1}^{M} \hat{\mathbf{h}}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}\right|^2\right\}$.

The user ordering can be done by a central entity (e.g., the CPU) that collects all information of the mean of effective channel strengths and then feeds back the result of user ordering and power allocation to the APs; higher powers are allocated to the users with lower channel strength (weaker users); i.e., $p_{n1} \leq p_{n2} \leq \cdots \leq p_{nK}$ where the kth user applies SIC to decode its own signal. More precisely, the kth user decodes the signals of the users with higher powers ($\forall i > k$), and then subtracts them from the received signal y_{nk} (11). Moreover, the kth user treats the signals of the other stronger users ($\forall i < k$) as interference. Mathematically, the condition that the kth user can perform SIC and decode the data intended for the weaker users can be written as [37], [38]

$$\mathbb{E}\{\log_2(1 + \gamma_{nk}^{ni})\} \geq \mathbb{E}\{\log_2(1 + \gamma_{ni}^{ni})\}, \quad \forall i > k. \tag{13}$$

In (13), γ_{nk}^{ni} is the effective signal-to-interference-plus-noise ratio (SINR) of the ith user at the kth user in cluster n, when user k decodes the signal intended for user i.

By using this rate inequality (13), the achievable rate of the kth user in the nth cluster can then be computed as

$$\mathcal{R}_{nk} = \zeta \log_2(1 + \bar{\gamma}_{nk}^{nk}), \tag{14}$$

where $\zeta = (\tau_c - \tau)/\tau_c$ is the pre-log factor and τ_c is the coherence interval. Here, $\bar{\gamma}_{nk}^{nk}$ is defined as $\bar{\gamma}_{nk}^{nk} \triangleq \min(\gamma_{nk}^{nk}, \gamma_{ni}^{nk}), \forall i < k$ to guarantee that the ith user ($\forall i < k$) can perform SIC and decode the data of the kth user [37]. Note that, allocating higher powers to the users with lower channel strength also yields non-trivial data rate for the weak users [37].

Because of the statistical CSI knowledge of the effective channels at users, i.e., $\mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}\right\}$ ($\forall m, n, k$), intra-cluster pilot contamination and channel estimation error, a perfect SIC process may not be always feasible. Hence, the received signal (11) after an imperfect SIC process can be written as follows:

$$\begin{aligned} \bar{y}_{nk} &= y_{nk} - \sum_{m=1}^{M} \sum_{i=k+1}^{K} \sqrt{p_{ni}} \mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}\right\} \hat{s}_{ni} \\ &= \underbrace{\sum_{m=1}^{M} \sqrt{p_{nk}} \mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}\right\} s_{nk}}_{T_0\text{:desired signal}} \\ &\quad + \underbrace{\sum_{m=1}^{M} \sqrt{p_{nk}} \left(\mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn} - \mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}\right\}\right) s_{nk}}_{T_1\text{:beamforming gain uncertainty}} \\ &\quad + \underbrace{\sum_{m=1}^{M} \sum_{i=1}^{k-1} \sqrt{p_{ni}} \mathbf{h}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn} s_{ni}}_{T_2\text{:intra-cluster interference after SIC}} \end{aligned}$$

$$+\underbrace{\sum_{m=1}^{M}\sum_{i=k+1}^{K}\sqrt{p_{ni}}\left(\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}s_{ni}-\mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\}\hat{s}_{ni}\right)}_{T_3:\text{residual interference due to imperfect SIC}}$$
$$+\underbrace{\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\sum_{\substack{n'=1\\n'\neq n}}^{N}\mathbf{w}_{mn'}s_{n'}}_{T_4:\text{Inter-cluster interference}}+n_{nk}, \tag{15}$$

where T_0 is the desired signal and T_1 is the beamforming gain uncertainty. Further, T_2 is intra-cluster interference caused by the signals of the users which are considered as the interference at the kth user and T_3 represents the error propagation due to the imperfect SIC. Here, $\hat{s}_{ni}$ is the detected signal (the estimate of s_{ni}) of the ith user by the kth user [39]. Without loss of generality, s_{ni} and its estimate $\hat{s}_{ni}$ can be assumed as jointly Gaussian with a certain correlation coefficient [20]. The relationship between s_{ni} and $\hat{s}_{ni}$ can then be written as

$$s_{ni}=\rho_{ni}\hat{s}_{ni}+e_{ni}, \tag{16}$$

where $\hat{s}_{ni}\sim\mathcal{CN}(0,1)$, $e_{ni}\sim\mathcal{CN}(0,\sigma_{e_{ni}}^2/[1+\sigma_{e_{ni}}^2])$ is the estimation error, statistically independent of $\hat{s}_{ni}$, and $\rho_{ni}=1/\sqrt{1+\sigma_{e_{ni}}^2}$. The correlation coefficient $0\leq\rho_{ni}\leq 1$ reflects the quality of the estimation and quantifies the severity of the SIC imperfection. The value of ρ_{ni} is determined by channel related issues (fading and shadowing) and other factors [40]; the greater its value, the greater the association between $\hat{s}_{ni}$ and s_{ni} and better the SIC performance [20]. We must note that, even with perfect estimation, e.g., $\rho_{ni}=1$ which results in $\hat{s}_{ni}=s_{ni}$, there is still intra-cluster interference, i.e., the loss due to the absence of instantaneous effective channel knowledge at the stronger users [38].

III. Downlink Achievable Rate

Here, we analyze the downlink sum rates achievable with MRT, fpZF, and mRZF. To this end, we first derive the effective SINR at each user. It then leads to the total downlink sum rate.

According to (15), since data signals intended to different users are mutually uncorrelated and the white additive noise is independent from the data symbols and the channel coefficients, it is easy to check that $T_i,\forall i$ and n_{nk} are mutually uncorrelated. Therefore, by considering the first term in (15) as the desired signal and the remaining terms as an effective noise, invoking the argument in [41], the SINR at the kth user in the nth cluster γ_{nk}^{nk} can be derived as (17), shown at the bottom of the next page, where η_{nk} is defined as $\eta_{nk}\triangleq\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}$.

Following the same principle as (17), $\gamma_{ni}^{nk},\forall i<k$ can also be calculated as (18), shown at the bottom of the next page.

The achievable rate $\mathcal{R}_{nk}$ in (14) is computed using (17) and (18) which depends on the precoding process at the APs. In order to design the precoding matrices at the APs, we first define $\bar{\mathbf{H}}_m\triangleq\mathbf{Y}_m^p\Phi$ where $\mathbf{\Phi}=[\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,\ldots,\boldsymbol{\phi}_N]\in\mathcal{C}^{\tau\times N}$ and $\mathbf{Y}_m^p\in\mathcal{C}^{L\times\tau}$ is given by (2). $\bar{\mathbf{H}}_m$ can be then written as

$$\bar{\mathbf{H}}_m=\left[\sqrt{\tau p_p}\sum_{k=1}^{K}\mathbf{h}_{m1k}+\tilde{\mathbf{n}}_m,\ldots,\sqrt{\tau p_p}\sum_{k=1}^{K}\mathbf{h}_{mNk}+\tilde{\mathbf{n}}_m\right]. \tag{19}$$

Therefore, $\bar{\mathbf{H}}_m$ has independent columns and $\bar{\mathbf{h}}_{mn}\sim\mathcal{CN}(\mathbf{0},(1+\tau p_p\sum_{k=1}^{k}\beta_{mnk})\mathbf{I}_L)$ is the nth column of $\bar{\mathbf{H}}_m$ i.e., $\bar{\mathbf{h}}_{mn}=\bar{\mathbf{H}}_m\mathbf{e}_n$ where $\mathbf{e}_n$ denotes the nth column of $\mathbf{I}_N$. Then, the channel estimate can be written as

$$\hat{\mathbf{h}}_{mnk}=c_{mnk}\bar{\mathbf{h}}_{mn}. \tag{20}$$

We next assume that, the APs use either MRT, fpZF or mRZF with an average normalization $\{\mathbb{E}[||\mathbf{w}_{mn}||^2]\}=1,\forall m,n$ [42], to precode data signals. It is more analytically tractable to have an average normalization over many coherence blocks rather than the normalization in every coherence block; i,e., $||\mathbf{w}_{mn}||^2=1$. Note that, the difference between these two normalizations is small when there is substantial channel hardening [42]. For each precoder, we then derive the closed-form achievable sum rate.

A. MRT Beamforming

With MRT, the mth AP computes the following precoding vector for the kth user in the nth cluster:

$$\mathbf{w}_{mn}=\frac{\bar{\mathbf{h}}_{mn}}{\sqrt{\mathbb{E}\left\{\|\,\bar{\mathbf{h}}_{mn}\,\|^2\right\}}}. \tag{21}$$

This precoder (21) is used for all the users in the nth cluster.

Theorem 1: In the CF massive MIMO-NOMA system with MRT precoding, γ_{nk}^{nk} in (17)*, for finite values of M,L,N and K, is given by* (22)*, as shown at the bottom of the next page.*

Proof: See Appendix A for derivations. ∎

Since (22) connects several factors together, it can be used to provide several remarks and guidelines relevant for practical NOMA-based CF mMIMO. We briefly describe them next.

Remark 1: With MRT precoding, the signal power increases as the number of antennas at each AP, L, increases, thanks to the array gain. On the other hand, the interference due to the pilot contamination and imperfect SIC (the first term in the denominator) proportionally increases as L increases.

Remark 2: Beamforming gain uncertainty and inter-cluster interference (the second term in the denominator), are not affected by L.

Remark 3: As opposed to the number of clusters N, which is restricted by the maximum orthogonal pilot sequence length ($N\leq\tau$), the number of users K within each cluster can be increased greatly. However, the downside is the increase of pilot contamination and imperfect SIC.

B. Full-Pilot ZF Beamforming

In this case, the precoding vector at the mth AP for the users in the nth cluster can be formulated as

$$\mathbf{w}_{mn}=\frac{\bar{\mathbf{H}}_m\left(\bar{\mathbf{H}}_m^{\mathrm{H}}\bar{\mathbf{H}}_m\right)^{-1}\mathbf{e}_n}{\sqrt{\mathbb{E}\left\{\left\|\bar{\mathbf{H}}_m(\bar{\mathbf{H}}_m^{\mathrm{H}}\bar{\mathbf{H}}_m)^{-1}\mathbf{e}_n\right\|^2\right\}}}, \tag{23}$$

where each AP has N precoding vectors, one per cluster (pilot).

In order to find the effective SINR γ_{nk}^{nk} given in (17), we first need to derive the value of $\eta_{nk}, \forall n$. To obtain $\eta_{nk}, \forall n$, we need to compute $\hat{\mathbf{h}}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn'}$. To this end, we first obtain the normalization term in (23). By employing Lemma 2.10 of [43] and using the properties of $N \times N$ central complex Wishart matrix with L ($L \geq N+1$) degrees of freedom, we obtain

$$\begin{aligned} &\mathbb{E}\left\{\| \bar{\mathbf{H}}_m (\bar{\mathbf{H}}_m^{\mathrm{H}} \bar{\mathbf{H}}_m)^{-1} \mathbf{e}_n \|^2\right\} \\ &\quad = \mathbb{E}\left\{(\bar{\mathbf{H}}_m^{\mathrm{H}} \bar{\mathbf{H}}_m)_{n,n}^{-1}\right\} \\ &\quad = \frac{1}{(L-N)(1+\tau p_p \sum_{k=1}^{K} \beta_{mnk})}. \end{aligned} \tag{24}$$

Finally, by employing (20), (23) and (24), we have

$$\begin{aligned} \hat{\mathbf{h}}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn'} &= c_{mnk} \mathbf{e}_n^{\mathrm{H}} \mathbf{e}_{n'} \sqrt{(L-N)\left(1+\tau p_p \sum_{k=1}^{K} \beta_{mnk}\right)} \\ &= \begin{cases} \sqrt{\theta_{mnk}(L-N)}, & n = n' \\ 0, & n \neq n' \end{cases} \end{aligned} \tag{25}$$

From (25), we find that fpZF precoder suppresses the inter-cluster interference by only utilizing local CSI rather than CSI shared among the APs, a big advantage. Moreover, fpZF has the same front-hauling overhead as MRT.

Theorem 2: In the CF massive MIMO-NOMA system with fpZF precoder, γ_{nk}^{nk} in (17)*, for any finite M, L, N and K, is given by* (26)*, as shown at the bottom of the next page.*

Proof: See Appendix B for the derivations. ∎

Remark 4: Similar to MRT, by employing fpZF precoding, the signal power increases as the number of antennas at each AP L increases, thanks to the array gain (which is $L-N$). On the other hand, the interference due to the pilot contamination and imperfect SIC (the first term in the denominator) proportionally increases as L increases.

Remark 5: Unlike MRT which only aims to maximize the signal-to-noise ratio (SNR) and ignores the inter-cluster interference, fpZF aims to suppress the inter-cluster interference by sacrificing array gain.

Remark 6: For a fixed number of clusters N, if the number of antennas at each AP tends to infinity ($L \to \infty$), the SINR for both the MRT and fpZF precoding (22) *and* (26) *can be approximated as*

$$\begin{aligned} &\lim_{L\to\infty} \gamma_{nk}^{nk} \\ &= \frac{p_{nk}\left(\sum_{m=1}^{M} \sqrt{\theta_{mnk}}\right)^2}{\left(\sum_{m=1}^{M} \sqrt{\theta_{mnk}}\right)^2 \left(\sum_{i=1}^{k-1} p_{ni} + \sum_{i=k+1}^{K} p_{ni}(2-2\rho_{ni})\right)}. \end{aligned} \tag{27}$$

which shows that, the gain of adding more antennas at each AP disappears. Besides, for fixed values of L and N, by increasing the number of users at each cluster, pilot contamination and SIC become the dominant interferences. Therefore, the performance of the CF massive MIMO-NOMA system is limited by pilot contamination and imperfect SIC.

In contrast, in OMA, where each user is assigned with an orthogonal pilot, SINR γ_{nk}^{nk} for MRT and fpZF is given as

$$\gamma_{nk,\mathrm{OMA}}^{nk,\mathrm{MRT}} = \frac{L p_{nk}\left(\sum_{m=1}^{M} \sqrt{\theta_{mnk}}\right)^2}{\left(\sum_{n'=1}^{N}\sum_{k'=1}^{K} p_{n'k'}\right) \sum_{m=1}^{M} \beta_{mnk} + 1} \tag{28}$$

and

$$\gamma_{nk,\mathrm{OMA}}^{nk,\mathrm{fpZF}} = \frac{(L-KN)\, p_{nk}\left(\sum_{m=1}^{M} \sqrt{\theta_{mnk}}\right)^2}{\left(\sum_{n'=1}^{N}\sum_{k'=1}^{K} p_{n'k'}\right) \sum_{m=1}^{M} (\beta_{mnk} - \theta_{mnk}) + 1}. \tag{29}$$

Therefore, deploying more antennas at each AP is always beneficial for OMA.

$$\gamma_{nk}^{nk} = \frac{p_{nk}\left|\mathbb{E}\left\{\eta_{nk}\right\}\right|^2}{p_{nk}\mathbb{E}\left\{\left|(\eta_{nk} - \mathbb{E}\left\{\eta_{nk}\right\})\right|^2\right\} + \sum_{i=1}^{k-1} p_{ni}\mathbb{E}\left\{|\eta_{nk}|^2\right\} + \sum_{i=k+1}^{K} p_{ni}\mathbb{E}\left\{\left|\eta_{nk} s_{ni} - \mathbb{E}\left\{\eta_{nk}\right\}\hat{s}_{ni}\right|^2\right\} + \sum_{\substack{n'=1 \\ n'\neq n}}^{N} p_{n'}\mathbb{E}\left\{\left|\eta_{n'k}\right|^2\right\} + 1}. \tag{17}$$

$$\gamma_{ni}^{nk} = \frac{p_{nk}\left|\mathbb{E}\left\{\eta_{ni}\right\}\right|^2}{p_{nk}\mathbb{E}\left\{\left|(\eta_{ni} - \mathbb{E}\left\{\eta_{ni}\right\})\right|^2\right\} + \sum_{j=1}^{k-1} p_{nj}\mathbb{E}\left\{|\eta_{ni}|^2\right\} + \sum_{j=k+1}^{K} p_{nj}\mathbb{E}\left\{\left|\eta_{ni} s_{nj} - \mathbb{E}\left\{\eta_{ni}\right\}\hat{s}_{nj}\right|^2\right\} + \sum_{\substack{n'=1 \\ n'\neq n}}^{N} p_{n'}\mathbb{E}\left\{\left|\eta_{n'i}\right|^2\right\} + 1}. \tag{18}$$

$$\gamma_{nk}^{nk,\mathrm{MRT}} = \frac{L p_{nk}\left(\sum_{m=1}^{M} \sqrt{\theta_{mnk}}\right)^2}{L\left(\sum_{m=1}^{M} \sqrt{\theta_{mnk}}\right)^2 \left(\sum_{i=1}^{k-1} p_{ni} + \sum_{i=k+1}^{K} p_{ni}(2-2\rho_{ni})\right) + \left(\sum_{n'=1}^{N} p_{n'}\right) \sum_{m=1}^{M} \beta_{mnk} + 1}. \tag{22}$$

C. Modified RZF Beamforming

The mRZF precoder tries to balance inter-cluster interference mitigation and intra-cluster power enhancement. Since closed-form analysis of it is difficult (if not impossible), we analyze its achievable rate in the asymptotic regime, where the number of clusters N and the number of antennas at each AP L grows infinitely large while keeping a finite ratio; i.e., $1 \leq \lim_{L,N\to\infty} \frac{L}{N} \leq \infty$.[2] This asymptotic expression can nevertheless be used with finite values of L and N [23], [24].

The precoding vector of mRZF at the mth AP for the users in the nth cluster can be expressed as

$$\mathbf{w}_{mn} = \frac{\left(\bar{\mathbf{H}}_m \bar{\mathbf{H}}_m^{\mathrm{H}} + L\alpha \mathbf{I}_L\right)^{-1} \bar{\mathbf{h}}_{mn}}{\sqrt{\mathbb{E}\left\{\left\|(\bar{\mathbf{H}}_m \bar{\mathbf{H}}_m^{\mathrm{H}} + L\alpha \mathbf{I}_L)^{-1} \bar{\mathbf{h}}_{mn}\right\|^2\right\}}}, \tag{30}$$

where $\bar{\mathbf{H}}_m$ is given in (19) and $\alpha > 0$ is the regularization parameter that can be optimized [44]. Finding optimal value for α, however, is outside the scope of this paper and is left for future work. Here, we assume that α is scaled with L to ensure that it converges to a constant value as L and N tend to infinity.

It can be seen from (30), in our design the precoder of each AP only utilizes local CSI rather than global CSI knowledge shared between the APs. Hence, mRZF has the same front-hauling overhead as that of MRT and fpZF.

Theorem 3: In the CF massive MIMO-NOMA system with mRZF precoder, γ_{nk}^{nk} in (17)*, when L and N grow large such that $1 \leq \lim_{L,N\to\infty} \frac{L}{N} \leq \infty$, is given by* (31)*, as shown at the bottom of the next page, where $\mathbf{\Theta}_{mn} = (1 + \tau p_p \sum_{k=1}^{k} \beta_{mnk})\mathbf{I}_L$, $a_{mnk} = \sqrt{\frac{\beta_{mnk} - \theta_{mnk}}{1+\tau p_p \sum_{k=1}^{K} \beta_{mnk}}}$ and $e_{mn}^o = e_{mn}$ in which*

$$e_{mn} = \frac{1}{L}\mathrm{tr}\left[\mathbf{\Theta}_{mn}\mathbf{T}_m\right] \tag{32}$$

$$\mathbf{T}_m = \left(\frac{1}{L}\sum_{j=1}^{N} \frac{\mathbf{\Theta}_{mj}}{1+e_{mj}} + \alpha \mathbf{I}_L\right)^{-1} \tag{33}$$

$$\psi_{mn}^o = \frac{1}{L}\frac{e'_{mn}}{(1+e_{mn})^2} \tag{34}$$

$$\Upsilon_{mn} = \frac{1}{L}\sum_{\substack{n'=1 \\ n'\neq n}}^{N} \frac{p_{n'} e'_{n',mn}}{\psi_{mn'}^o (1+e_{mn'})^2} \tag{35}$$

[2]This assumption implies that the coherence time of the channel τ_c scales linearly with N. However, as it will be shown in Section IV, our analysis provides a tight approximation for the achievable rate even for small values of N or equivalently τ_c.

in which $\mathbf{e}'_m = [e'_{m1}, \ldots, e'_{mN}]^{\mathrm{T}}$ and $\mathbf{e}'_{mn} = [e'_{1,mn}, \ldots, e'_{N,mn}]^{\mathrm{T}}$ are given by

$$\mathbf{e}'_m = (\mathbf{I}_N - \mathbf{J}_m)^{-1}\mathbf{v}_m, \tag{36}$$

$$\mathbf{e}'_{mn} = (\mathbf{I}_N - \mathbf{J}_m)^{-1}\mathbf{v}_{mn}, \tag{37}$$

and $\mathbf{J}_m$, $\mathbf{v}_m$ and $\mathbf{v}_{mn}$ are derived as follows,

$$[\mathbf{J}_m]_{ij} = \frac{\frac{1}{L}\mathrm{tr}\left[\mathbf{\Theta}_{mi}\mathbf{T}_m\mathbf{\Theta}_{mj}\mathbf{T}_m\right]}{L(1+e_{mj})^2} \tag{38}$$

$$\mathbf{v}_m = \left[\frac{1}{L}\mathrm{tr}\left[\mathbf{\Theta}_{m1}\mathbf{T}_m^2\right], \ldots, \frac{1}{L}\mathrm{tr}\left[\mathbf{\Theta}_{mN}\mathbf{T}_m^2\right]\right]^{\mathrm{T}}. \tag{39}$$

$$\mathbf{v}_{mn} = \left[\frac{1}{L}\mathrm{tr}\left[\mathbf{\Theta}_{m1}\mathbf{T}_m\mathbf{\Theta}_{mn}\mathbf{T}_m\right], \ldots, \frac{1}{L}\mathrm{tr}\left[\mathbf{\Theta}_{mN}\mathbf{T}_m\mathbf{\Theta}_{mn}\mathbf{T}_m\right]\right]^{\mathrm{T}}. \tag{40}$$

Besides, the initial values of $e_{mn}; \forall n$ to calculate (32) *and* (33) *are set to $e_{mn}^0 = \frac{1}{\alpha}, n = 1, \ldots, N$ [23].*

Proof: See Appendix C for derivations. ∎

Remark 7: In order to reduce the amount of overhead exchanged over the fronthaul network, channel estimation is performed locally at each AP and MRT, fpZF and mRZF precoders are designed based on the local CSI at each AP (19)*. In this case, pilot signals are not shared over the fronthaul link. Indeed, only channel statistics (large scale parameters which changes slowly [8]) are sent to the CPU for user ordering and power allocation process. Thus, assuming a given system parameters M, L, N, and K, all three precoders have the same front-hauling overhead. In particular, for each realization of the user locations, the number of MNK statistical parameters need to be exchanged between the CPU and the APs.*

IV. Simulation Results

Herein, we provide simulation results to evaluate the performance of the CF mMIMO with NOMA or OMA. In NOMA, the same pilot sequence is shared among users within each cluster, but different clusters are assigned mutually orthogonal pilots. In OMA, all the users are assigned with different mutually orthogonal pilots. Therefore, the minimum pilot sequence lengths for NOMA and OMA are $\tau_{\mathrm{NOMA}} = N$ and $\tau_{\mathrm{OMA}} = KN$, respectively. The pre-log factor for NOMA and OMA cases are also defined as $\zeta_{\mathrm{NOMA}} = (\tau_c - N)/\tau_c$ and $\zeta_{\mathrm{OMA}} = (\tau_c - KN)/\tau_c$, respectively.

In our simulations, the APs are uniformly distributed within an area of size $D \times D$ m^2. Furthermore, users are clustered based on their spatial locations and all the clusters are uniformly distributed at random in the given area (the users in the same cluster are also distributed uniformly at random around the center point of the cluster).

$$\gamma_{nk}^{nk,\mathrm{fpZF}} = \frac{(L-N)\, p_{nk}\left(\sum_{m=1}^{M}\sqrt{\theta_{mnk}}\right)^2}{(L-N)\left(\sum_{m=1}^{M}\sqrt{\theta_{mnk}}\right)^2\left(\sum_{i=1}^{k-1} p_{ni} + \sum_{i=k+1}^{K} p_{ni}(2-2\rho_{ni})\right) + \left(\sum_{n'=1}^{N} p_{n'}\right)\sum_{m=1}^{M}(\beta_{mnk} - \theta_{mnk}) + 1}. \tag{26}$$

A. Large-Scale Fading Model

Here, we assume that the large-scale fading coefficient β_{mnk} in (1) includes both the path-loss effect and shadowing. Thus, β_{mnk} can be wrtten as [8]

$$\beta_{mnk} = \mathrm{PL}_{mnk} + z_{mnk} \quad \mathrm{dB}, \tag{41}$$

where PL_{mnk} is the path loss, z_{mnk} represents the shadow fading generated from a log-normal distribution with standard deviation σ_{sh}. A three-slope model is considered for the path loss [45], the path loss exponent equals (i) 3.5 if the distance between the mth AP and the kth user in the nth cluster (d_{mnk}) is greater than d_1, (ii) equals 2 if $d_0 < d_{mnk} \leq d_1$, and (iii) equals 0 if $d_{mnk} \leq d_0$ for some d_0 and d_1. When $d_{mnk} > d_1$, the Hata-COSTA231 propagation model is employed [8]. Therefore, the path loss can be written as

$$\mathrm{PL}_{mnk} = \begin{cases} -\mathcal{L} - 35\log_{10}(d_{mnk}), & d_{mnk} > d_1 \\ -\mathcal{L} - 15\log_{10}(d_1) & \\ \quad -20\log_{10}(d_{mnk}), & d_0 < d_{mnk} \leq d_1 \\ -\mathcal{L} - 15\log_{10}(d_1) & \\ \quad -20\log_{10}(d_0), & d_{mnk} \leq d_0 \end{cases} \tag{42}$$

where

$$\mathcal{L} \triangleq 46.3 + 33.9\log_{10}(f) - 13.82\log_{10}(h_{AP}) - (1.1\log_{10}(f) - 0.7)h_u + (1.56\log_{10}(f) - 0.8). \tag{43}$$

In (43), f is the carrier frequency (in MHz). Also, h_{AP} and h_u are the AP antenna and user antenna heights (in m), respectively. Note that when $d_{mnk} \leq d_1$, there is no shadowing.

B. Parameters and Setup

The simulation parameters are reported in Table I. The noise variance is given by $\sigma_w^2 = 290 \times k_b \times$ bandwidth $\times$ noise figure, where k_b is the Boltzmann constant. To each cluster, the mth AP allocates power $\frac{p_t}{N}$. We further assume that $K = 2$; i.e., each cluster has two users, and the total power allocated for the nth cluster p_n is divided between the two users within a cluster based on a $3:7$ ratio ($\forall n$) implying that $\lambda_{n1} = 0.3\lambda_n, \lambda_{n2} = 0.7\lambda_n$. However, these power coefficients are not necessarily optimal and can be further optimized.

C. Performance Evaluation

Here, we first compare the performance of the CF massive MIMO-NOMA and OMA systems for MRT, fpZF and mRZF precoding in terms of sum rate given by [37]

$$\mathcal{R} = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathcal{R}_{nk}. \tag{44}$$

TABLE I
SIMULATION SETTINGS

Parameter	Value	Parameter	Value
Carrier frequency	1.9 GHz	τ_c	56
Bandwidth	20 MHz	p_p	20 dBm
Noise figure	9 dB	p_t	23 dBm
D, d_1, d_0	$1000, 50, 10$ m	σ_{sh}	8 dB
$h_{\mathrm{AP}}, h_{\mathrm{u}}$	$65, 15$ m	ρ_{ni}	0.1

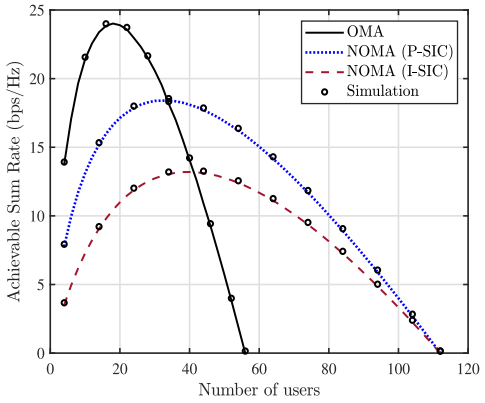

Fig. 2. The achievable sum rate versus the number of users for 25 APs ($M = 25$) and 8 antennas per AP ($L = 8$) with MRT precoding. Legends P-SIC and I-SIC stand for perfect and imperfect SIC operations. The curves are generated using the derived SINR expressions. Several simulation points are superimposed on each curve.

where $\mathcal{R}_{nk}$ is given in (14) in which $\bar{\gamma}_{nk}^{nk}$ is the closed-form effective SINR for MRT, fpZF and mRZF calculated using (22), (26) and (31), respectively. Several Monte-Carlo simulated sum rates are also superimposed on each curve in order to validate the derived analytical expressions.

Fig. 2 demonstrates the achievable sum rate (44) of NOMA and OMA systems with MRT precoding as a function of the number of users. We set $\tau_c = 56$. We observe that the maximum number of users simultaneously served in OMA in the same resource block is $K_{\max}^{\mathrm{OMA}} = 56$. In contrast, NOMA doubles this. This is due to the fact that in OMA each user is assigned with an orthogonal pilot while in NOMA same pilot sequence is shared among users within each cluster. For a large number of users, NOMA outperforms OMA; however, with fewer users, NOMA achieves a lower sum rate than OMA due to the effects of intra-cluster pilot contamination and imperfect SIC. We also observe that, SIC imperfection considerably degrades the performance of NOMA; for instance, compared to the perfect SIC, the residual interference caused by imperfect SIC degrades the achievable sum rate by 4.9 bps/Hz for 40 simultaneously served users.

$$\gamma_{nk}^{nk,\mathrm{mRZF}} = \frac{p_{nk}\left(\sum_{m=1}^{M} \frac{1}{\sqrt{\psi_{mn}^o}} \frac{c_{mnk} e_{mn}^o}{1+e_{mn}^o}\right)^2}{\left(\sum_{m=1}^{M} \frac{1}{\sqrt{\psi_{mn}^o}} \frac{c_{mnk} e_{mn}^o}{1+e_{mn}^o}\right)^2 \left(\sum_{i=1}^{k-1} p_{ni} + \sum_{i=k+1}^{K} p_{ni}(2-2\rho_{ni})\right) + \sum_{m=1}^{M} \Upsilon_{mn}\left(\frac{c_{mnk}^2}{(1+e_{mn}^o)^2} + a_{mnk}^2\right) + 1}. \tag{31}$$

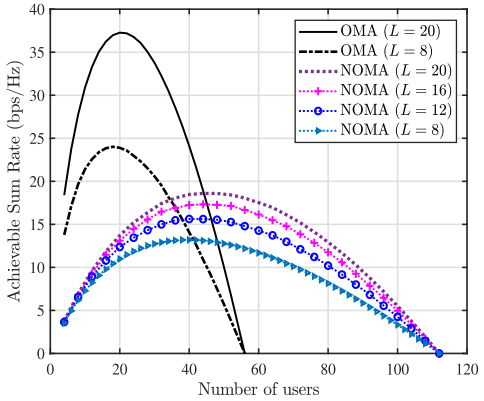

Fig. 3. The achievable sum rate versus the number of users for 25 APs, with MRT precoding (Imperfect SIC). The curves are generated using analysis.

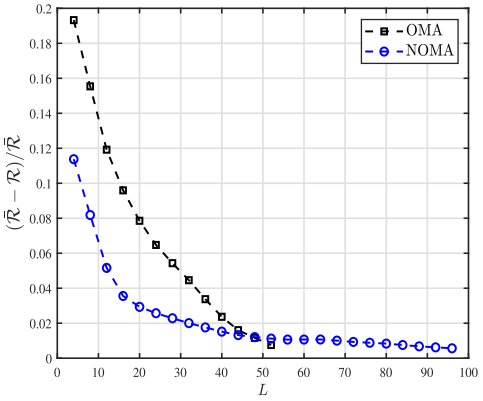

Fig. 5. The mRZF precoder; the relative error of the sum rate $\mathcal{R}$ via (31) compared to the ergodic sum rate; $(\bar{\mathcal{R}}-\mathcal{R})/\bar{\mathcal{R}}$ versus the number of antennas per AP, $L = KN$, for 25 APs and $\alpha = 0.8$.

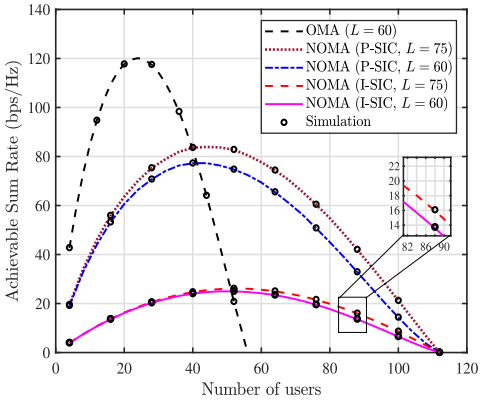

Fig. 4. The achievable sum rate as a function of the number of users for 25 APs with fpZF precoding. The curves are generated using analysis and Monte Carlo simulation.

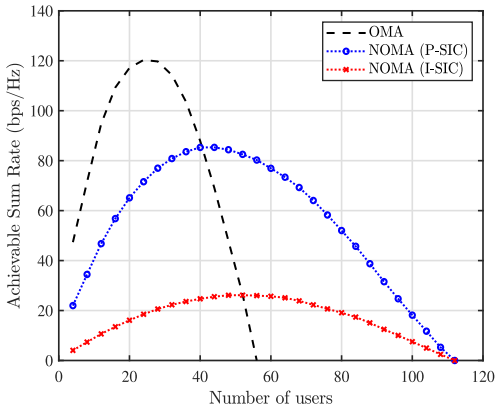

Fig. 6. The achievable sum rate versus the number of users for 25 APs and 60 antennas per AP and mRZF precoding with $\alpha = 0.8$. The curves are generated using the derived results.

In Fig. 3, the impact of deploying more antennas at the APs is investigated. As expected, adding more antennas at the APs results in better sum rate, thanks to the array gain. However, despite OMA, the gain of adding more antennas diminishes for NOMA system since the interference due to the pilot contamination and imperfect SIC proportionally increases as L increases. This observation confirms Remarks 1 and 6.

Fig. 4 shows the achievable sum rate of NOMA and OMA systems as a function of the number of users. Here, fpZF precoding and at least $N + 1 \leq L$ antennas (26) per AP are considered. We see that the residual interference caused by imperfect SIC significantly degrades the achievable sum rate. More specifically, for 40 users and 60 antennas per AP, the gap between I-SIC and P-SIC curves is about 53 bps/Hz. For imperfect SIC, the gain due to antennas at APs is almost negligible. This is because, for a low number of users (clusters), the SINR converges to (27), where imperfect SIC is the dominant term. Furthermore, as the number of users increases, both the desired power and interference power due to imperfect SIC and pilot contamination increase proportionally to $L - N$.

For the mRZF precoder, the SINR asymptotic $\gamma_{nk}^{nk,\text{mRZF}}$ in (31) must be validated. Thus, Fig. 5 shows the error of the sum rate $\mathcal{R}$ computed based on (31) compared to the ergodic sum rate via simulations of the SINR. Clearly, the SINR asymptotic holds even for small values of L and becomes more accurate as L increases.

Fig. 6 shows the achievable sum rate of NOMA and OMA systems for different number of users with mRZF precoding and finite system dimensions. In this scheme, we simply consider $\alpha = 0.8$; however, the optimal value of α that accounts for the CSI imperfection could be further investigated [23], [46]. As expected, with imperfect SIC, the resulting error propagation and intra-cluster pilot contamination limit the system performance. Just as in the case of MRT and fpZF, for a large number of users, NOMA outperforms OMA; however, with a few users, the converse is true.

In Fig. 7 and Fig. 8, we compare the performance of the three precoders for perfect and imperfect SIC, respectively. We observe that, for perfect SIC, mRZF always achieves higher rates than fpZF and MRT. The reason is that mRZF tries to balance the inter-cluster interference mitigation and intra-cluster power enhancement. As well, fpZF and mRZF outperform MRT as they are able to cancel the inter-cluster interference. Therefore, although fpZF and mRZF have the same front-hauling overhead as MRT, they can achieve higher rates. However, for a very large number of users, MRT

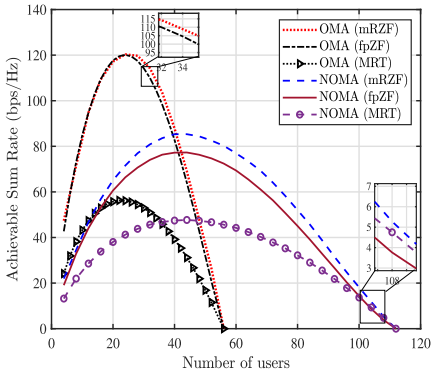

Fig. 7. The achievable sum rate versus the number of users for 25 APs and 60 antennas per AP (Perfect SIC). The curves are generated using the derived results only.

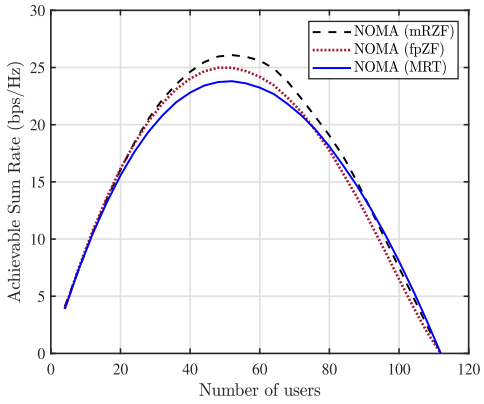

Fig. 8. The achievable sum rate versus the number of users for 25 APs and 60 antennas per AP (Imperfect SIC). The curves are generated using analysis.

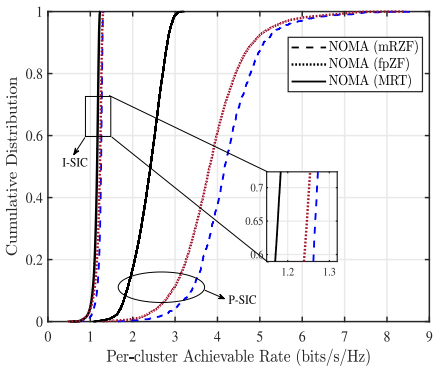

Fig. 9. Cumulative distribution of per-cluster achievable rate for 25 APs and 60 antennas per AP and 20 clusters (NOMA). The curves are generated using analysis.

outperforms fpZF. Furthermore, CF hybrid mMIMO-NOMA with either mRZF or fpZF outperforms OMA systems with MRT as twice number of users could be served at the price of a negligible performance loss for a lightly loaded system. We also observe from Fig. 8, with imperfect SIC, MRT and mRZF perform roughly the same for a large number of users. This is because both fpZF and mRZF sacrifice much of the array gain to suppress the inter-cluster interference; however mRZF outperforms fpZF since it also considers the desired power. We also note that α may be optimized to enhance the performance of mRZF precoding.

To further investigate these three precoders, we also plot the cumulative distributions of their rates per-cluster in Fig. 9. It shows that the per-cluster rate degrades due to imperfect SIC. In particular, in 90%-likely performance, all the three precoders behave almost the same and achieve low per-cluster rate of ~ 1 bps/Hz. In the case of perfect SIC, however, mRZF and fpZF significantly outperform MRT. Thus, the rate loss due to the imperfect SIC is more considerable for fpZF and mRZF compared to that of MRT. More precisely, the 90%-likely per-cluster rate loss is, respectively, about 1.8 bps/Hz and 2.2 bps/Hz for fpZF and mRZF, which are more than twice than that of MRT (0.8 bps/Hz).

V. Conclusion

In this paper, we analyzed the achievable rate of three standard precoders for NOMA-aided CF mMIMO. We thus derived closed-form sum rates for MRT and fpZF by considering joint effects of intra-cluster pilot contamination, inter-cluster interference and imperfect SIC. However, since a closed-form rate analysis is intractable for finite mRZF precoder, we analyzed the asymptotic regime, where the number of clusters and the number of antennas at each AP grow extremely large while keeping a finite ratio.

We showed that NOMA based CF mMIMO can support significantly more users compared to the OMA-hybrid at the same time-frequency resources. For a large number of users, the NOMA-hybrid outperforms the OMA-hybrid; however, for a few users, the former achieves lower sum rate than the latter due to the effects of intra-cluster pilot contamination and imperfect SIC. It was further shown that, with perfect SIC, mRZF and fpZF significantly outperforms MRT despite of having the same front-hauling overhead. Finally, CF hybrid mMIMO-NOMA with either fpZF or mRZF outperforms the OMA systems with MRT.

Future research includes the optimization of α of mRZF precoder, optimal user clustering and power allocations.

Appendix A
Derivation of $\gamma_{nk}^{nk,\mathrm{MRT}}$ (22)

By invoking (20) in (21), the MRT precoding vector can be written as

$$\mathbf{w}_{mn} = \frac{\hat{\mathbf{h}}_{mnk}}{\sqrt{L\theta_{mnk}}}. \tag{45}$$

We first proceed to compute $\mathbb{E}\{\eta_{nk}\}$ as follows:

$$\begin{aligned}\mathbb{E}\{\eta_{nk}\} &= \mathbb{E}\left\{\sum_{m=1}^{M} \mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\} \\ &= \mathbb{E}\left\{\sum_{m=1}^{M} (\hat{\mathbf{h}}_{mnk}^{\mathrm{H}} + \boldsymbol{\epsilon}_{mnk}^{\mathrm{H}})\mathbf{w}_{mn}\right\} \\ &= \sum_{m=1}^{M} \sqrt{L\theta_{mnk}}. \end{aligned} \tag{46}$$

Since $\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}$ are independent for different values of m and the variance of a sum of independent random variables is

equal to the sum of the variances, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left|(\eta_{nk}-\mathbb{E}\{\eta_{nk}\})\right|^2\right\}\\
&=\mathbb{E}\left\{\left|\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}-\mathbb{E}\left\{\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\}\right|^2\right\}\\
&=\sum_{m=1}^{M}\mathbb{E}\left\{\left|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right|^2\right\}-\sum_{m=1}^{M}\left|\mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\}\right|^2\\
&=\sum_{m=1}^{M}\mathbb{E}\left\{\left|(\hat{\mathbf{h}}_{mnk}^{\mathrm{H}}+\boldsymbol{\epsilon}_{mnk}^{\mathrm{H}})\frac{\hat{\mathbf{h}}_{mnk}}{\sqrt{L\theta_{mnk}}}\right|^2\right\}-\sum_{m=1}^{M}L\theta_{mnk}\\
&=\sum_{m=1}^{M}\mathbb{E}\left\{\left|\hat{\mathbf{h}}_{mnk}^{\mathrm{H}}\frac{\hat{\mathbf{h}}_{mnk}}{\sqrt{L\theta_{mnk}}}\right|^2\right\}+\sum_{m=1}^{M}\mathbb{E}\left\{\left|\boldsymbol{\epsilon}_{mnk}^{\mathrm{H}}\frac{\hat{\mathbf{h}}_{mnk}}{\sqrt{L\theta_{mnk}}}\right|^2\right\}\\
&\quad-\sum_{m=1}^{M}L\theta_{mnk}\\
&\overset{(a)}{=}(L+1)\sum_{m=1}^{M}\theta_{mnk}+\sum_{m=1}^{M}(\beta_{mnk}-\theta_{mnk})-L\sum_{m=1}^{M}\theta_{mnk}\\
&=\sum_{m=1}^{M}\beta_{mnk},
\end{aligned}\tag{47}$$

where (a) comes from the fact that, for any vector $\mathbf{z}\sim\mathcal{CN}(\mathbf{0},\theta\mathbf{I}_L)$, we have [43]

$$\mathbb{E}\left\{|\mathbf{z}^{\mathrm{H}}\mathbf{z}|^2\right\}=(L^2+L)\theta^2.\tag{48}$$

To find $\mathbb{E}\left\{|\eta_{nk}|^2\right\}$, we use $\mathrm{Var}\{X\}=\mathbb{E}\{X^2\}-(\mathbb{E}\{X\})^2$. Thus, we have

$$\begin{aligned}
\mathbb{E}\left\{|\eta_{nk}|^2\right\}&=\mathbb{E}\left\{\left|(\eta_{nk}-\mathbb{E}\left\{\eta_{nk}\right\})\right|^2\right\}+(\mathbb{E}\{\eta_{nk}\})^2\\
&=\sum_{m=1}^{M}\beta_{mnk}+L\left(\sum_{m=1}^{M}\sqrt{\theta_{mnk}}\right)^2.
\end{aligned}\tag{49}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left\{|\eta_{n'k}|^2\right\}&=\mathbb{E}\left\{\left|\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}\right|^2\right\}\\
&=\sum_{m=1}^{M}\mathbb{E}\left\{\left|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}\right|^2\right\}\\
&\quad+\sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\mathbb{E}\left\{\left|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}\left(\mathbf{h}_{m'nk}^{\mathrm{H}}\mathbf{w}_{m'n'}\right)\right|\right\}\\
&=\sum_{m=1}^{M}\beta_{mnk},
\end{aligned}\tag{50}$$

where the second term in the second equality is discarded as $\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}(\forall m\in 1,\ldots,M)$ are independent of each other.

Moreover, since $\mathbb{E}\{s_{ni}^{*}\hat{s}_{ni}\}=\mathbb{E}\{s_{ni}\hat{s}_{ni}^{*}\}=\rho_{ni}$, we have

$$\begin{aligned}
\mathbb{E}\left\{\left|\eta_{nk}s_{ni}-\mathbb{E}\left\{\eta_{nk}\right\}\hat{s}_{ni}\right|^2\right\}&=\mathbb{E}\left\{|\eta_{nk}|^2\right\}\\
&\quad+(1-2\rho_{ni})\mathbb{E}^2\{\eta_{nk}\}.
\end{aligned}\tag{51}$$

Finally, by substituting the above formulations into (17), $\gamma_{nk}^{nk,\mathrm{MRT}}$ can be obtained as (22).

Appendix B
Derivation of $\gamma_{nk}^{nk,\mathrm{ZF}}$ (26)

For fpZF precoding, we first compute $\mathbb{E}\{\eta_{nk}\}$ as follows:

$$\begin{aligned}
\mathbb{E}\{\eta_{nk}\}&=\mathbb{E}\left\{\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\}\\
&=\mathbb{E}\left\{\sum_{m=1}^{M}\left(\hat{\mathbf{h}}_{mnk}^{\mathrm{H}}+\boldsymbol{\epsilon}_{mnk}^{\mathrm{H}}\right)\mathbf{w}_{mn}\right\}\\
&=\sum_{m=1}^{M}\sqrt{(L-N)\theta_{mnk}}.
\end{aligned}\tag{52}$$

In the denominator, we have

$$\mathbb{E}\left\{\left|(\eta_{nk}-\mathbb{E}\{\eta_{nk}\})\right|^2\right\}=\mathbb{E}\left\{|\eta_{nk}|^2\right\}-\mathbb{E}^2\left\{\eta_{nk}\right\}.\tag{53}$$

Referring to (51) and (53), we just need to calculate $\mathbb{E}\left\{|\eta_{n'k}|^2\right\},\forall n'$. Employing the result in (25), we have

$$\begin{aligned}
\mathbb{E}\left\{|\eta_{n'k}|^2\right\}&=\mathbb{E}\left\{\left|\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}\right|^2\right\}\\
&=\mathbb{E}\left\{\left|\sum_{m=1}^{M}\left(\hat{\mathbf{h}}_{mnk}^{\mathrm{H}}+\boldsymbol{\epsilon}_{mnk}^{\mathrm{H}}\right)\mathbf{w}_{mn'}\right|^2\right\}\\
&=\begin{cases}(L-N)\left(\sum\limits_{m=1}^{M}\sqrt{\theta_{mnk}}\right)^2\\\quad+\sum\limits_{m=1}^{M}(\beta_{mnk}-\theta_{mnk}), & n=n'\\ \sum\limits_{m=1}^{M}(\beta_{mnk}-\theta_{mnk}), & n\neq n'\end{cases}
\end{aligned}\tag{54}$$

By substituting (52) and (54) into (17), $\gamma_{nk}^{nk,\mathrm{fpZF}}$ is in (26).

Appendix C
Derivation of $\gamma_{nk}^{nk,\mathrm{mRZF}}$ (31)

To derive SINR $\gamma_{nk}^{nk,\mathrm{mRZF}}$, we need the following terms:

- Desired signal power

$$p_d=p_{nk}\left(\sum_{m=1}^{M}\mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\}\right)^2.\tag{55}$$

- Variance of the beamforming gain uncertainty

$$p_{I1}=p_{nk}\left(\sum_{m=1}^{M}\mathbb{E}\left\{|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}|^2\right\}-\sum_{m=1}^{M}\left|\mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right\}\right|^2\right).\tag{56}$$

Inter-cluster

- interference

$$p_{I2}=\sum_{\substack{n'=1\\n'\neq n}}^{N}p_{n'}\mathbb{E}\left\{\left|\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}\right|^2\right\}.\tag{57}$$

In order to calculate the intra-cluster interference due to the pilot contamination and imperfect SIC, we need

$\mathbb{E}\{|\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}|^2\}$, which can be derived by exploiting (56) and the definition of the variance.

We now proceed to calculate (55), (56) and (57) when $L, N \to \infty$ while keeping a finite ratio.

Let

$$\bar{\mathbf{h}}_{mn} = \sqrt{L}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}, \tag{58}$$

where $\bar{\mathbf{g}}_{mn}$ has i.i.d complex entries with zero mean and variance of $\frac{1}{L}$ ($\bar{\mathbf{g}}_{mn} \sim \mathcal{CN}(\mathbf{0}, \frac{1}{L}\mathbf{I}_L)$) and $\boldsymbol{\Theta}_{mn} = (1+\tau p_p \sum_{k=1}^{k}\beta_{mnk})\mathbf{I}_L$. Also, assume that $\boldsymbol{\Theta}_{mn};\ \forall m,n$ and $\frac{1}{L}\bar{\mathbf{H}}_m\bar{\mathbf{H}}_m^{\mathrm{H}} = \sum_{n=1}^{N}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2};\ \forall m$ have uniformly bounded spectral norms [23]. Moreover, we define

$$\boldsymbol{\Sigma}_m \triangleq \left(\bar{\mathbf{H}}_m\bar{\mathbf{H}}_m^{\mathrm{H}} + L\alpha\mathbf{I}_L\right), \tag{59}$$

$$\boldsymbol{\Sigma}_{mn} \triangleq \left(\bar{\mathbf{H}}_{mn}\bar{\mathbf{H}}_{mn}^{\mathrm{H}} + L\alpha\mathbf{I}_L\right), \tag{60}$$

$$\begin{aligned}\psi_{mn} &\triangleq \mathbb{E}\left\{\left\|(\bar{\mathbf{H}}_m\bar{\mathbf{H}}_m^{\mathrm{H}} + L\alpha\mathbf{I}_L)^{-1}\bar{\mathbf{h}}_{mn}\right\|^2\right\}\\ &= \mathbb{E}\left\{\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-2}\bar{\mathbf{h}}_{mn}\right\}.\end{aligned} \tag{61}$$

where $\bar{\mathbf{H}}_{mn}$ is equal to $\bar{\mathbf{H}}_m$ with the nth column removed.

- Desired signal power
  Regarding (55), we derive $\mathbb{E}\{\mathbf{h}_{mnk}\mathbf{w}_{mn}\}$ as follows:

$$\begin{aligned}&\mathbb{E}\left\{\mathbf{h}_{mnk}\mathbf{w}_{mn}\right\}\\ &= \frac{1}{\sqrt{\psi_{mn}}}\mathbb{E}\left\{\mathbf{h}_{mnk}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-1}\bar{\mathbf{h}}_{mn}\right\}\\ &\overset{(a)}{=} \frac{1}{\sqrt{\psi_{mn}}}\mathbb{E}\left\{\frac{\mathbf{h}_{mnk}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}{1+\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}\right\}\\ &= \frac{1}{\sqrt{\psi_{mn}}}\mathbb{E}\left\{\frac{\left(c_{mnk}\bar{\mathbf{h}}_{mn}+\boldsymbol{\epsilon}_{mnk}\right)^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}{1+\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\Sigma_{mn}^{-1}\bar{\mathbf{h}}_{mn}}\right\}\\ &= \frac{1}{\sqrt{\psi_{mn}}}\mathbb{E}\\ &\quad\left\{\frac{c_{mnk}\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}{1+\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}+\frac{\boldsymbol{\epsilon}_{mnk}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}{1+\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}\right\},\end{aligned} \tag{62}$$

  where (a) comes from the matrix inversion Lemma [23], [47]. In order to solve (62), we use the following Lemma.
  Lemma 1: Let $\mathbf{A} \in \mathcal{C}^{L\times L}$ *and* $\mathbf{x},\mathbf{y} \sim \mathcal{CN}(\mathbf{0}, \frac{1}{L}\mathbf{I}_L)$. *Assume that* $\mathbf{A}$ *has uniformly bounded spectral norm (with respect to* L*) and that* $\mathbf{x}$ *and* $\mathbf{y}$ *are mutually independent and independent of* $\mathbf{A}$. *Now, as* $L \longrightarrow \infty$,

$$a.\ \mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x} \overset{a.s.}{\to} \frac{1}{L}\mathrm{tr}\left[\mathbf{A}\right],\ b.\ \mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{y} \overset{a.s.}{\to} 0, \tag{63}$$

  where $\overset{a.s.}{\to}$ denotes almost sure convergence as $L \longrightarrow \infty$. By substituting $\bar{\mathbf{h}}_{mn}$ with (58) and applying Lemma 1.a and Theorem 2 of [23] to the first term of (62), we have

$$\begin{aligned}\frac{c_{mnk}\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}}{1+\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn}} &= \frac{c_{mnk}L\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2}\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}}{1+L\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2}\boldsymbol{\Sigma}_{mn}^{-1}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}}\\ &= \frac{c_{mnk}\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2}\mathbf{C}_{mn}^{-1}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}}{1+\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2}\mathbf{C}_{mn}^{-1}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}}\\ &\overset{a.s.}{\to} \frac{c_{mnk}e_{mn}^{o}}{1+e_{mn}^{o}},\end{aligned} \tag{64}$$

  where $\mathbf{C}_{mn} = \boldsymbol{\Gamma}_{mn} + \alpha\mathbf{I}_L$ with $\boldsymbol{\Gamma}_{mn} = \frac{1}{L}\bar{\mathbf{H}}_{mn}\bar{\mathbf{H}}_{mn}^{\mathrm{H}}$. Besides, e_{mn}^{o} is given in (32). Similarly, Since, $\hat{\mathbf{h}}_{mnk} = c_{mnk}\bar{\mathbf{h}}_{mn}$ is independent of $\boldsymbol{\epsilon}_{mnk}$, by applying Lemma 1.b to the second term of (62), we have

$$\boldsymbol{\epsilon}_{mnk}^{\mathrm{H}}\boldsymbol{\Sigma}_{mn}^{-1}\bar{\mathbf{h}}_{mn} \overset{a.s.}{\to} 0. \tag{65}$$

  Regarding (61), to find the value of ψ_{mn}, we need to calculate $\mathbb{E}\left\{\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-2}\bar{\mathbf{h}}_{mn}\right\}$. By employing matrix inversion Lemma, Theorems 1 and 2 of [23], we obtain

$$\begin{aligned}\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-2}\bar{\mathbf{h}}_{mn} &\overset{(a)}{=} \frac{1}{L}\frac{\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2}\mathbf{C}_{mn}^{-2}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn}}{(1+\bar{\mathbf{g}}_{mn}^{\mathrm{H}}\boldsymbol{\Theta}_{mn}^{1/2}\mathbf{C}_{mn}^{-1}\boldsymbol{\Theta}_{mn}^{1/2}\bar{\mathbf{g}}_{mn})^2}\\ &\overset{a.s.}{\to} \frac{1}{L}\frac{\frac{1}{L}\mathrm{tr}\left[\boldsymbol{\Theta}_{mn}\mathbf{C}_{mn}^{-2}\right]}{(1+\frac{1}{L}\mathrm{tr}\left[\boldsymbol{\Theta}_{mn}\mathbf{C}_{mn}^{-1}\right])^2}\\ &\overset{a.s.}{\to} \frac{1}{L}\frac{\frac{1}{L}\mathrm{tr}\left[\boldsymbol{\Theta}_{mn}\mathbf{T}'_m\right]}{(1+\frac{1}{L}\mathrm{tr}\left[\boldsymbol{\Theta}_{mn}\mathbf{T}_m\right])^2},\end{aligned} \tag{66}$$

  where (a) comes from the matrix inversion Lemma which is applied twice, $\mathbf{T}_{mn}$ is defined in (33) and

$$\mathbf{T}'_m = \mathbf{T}_m\left[\frac{1}{L}\sum_{j=1}^{N}\frac{\boldsymbol{\Theta}_{mj}e'_{mj}}{(1+e_{mj})^2}+\mathbf{I}_L\right]\mathbf{T}_m, \tag{67}$$

  where

$$e'_{mj} = \frac{1}{L}\mathrm{tr}\left[\boldsymbol{\Theta}_{mj}\mathbf{T}'_m\right]. \tag{68}$$

  By employing Theorem 2 of [23], we finally have

$$\bar{\mathbf{h}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-2}\bar{\mathbf{h}}_{mn} \overset{a.s.}{\to} \psi_{mn}^{o}, \tag{69}$$

  where ψ_m^o is given in (34). Therefore, by invoking (64) and (69) in (55), desired signal power can be written as

$$p_d \overset{a.s.}{\to} p_{n,k}\left(\sum_{m=1}^{M}\frac{1}{\sqrt{\psi_{mn}^{o}}}\frac{c_{mnk}e_{mn}^{o}}{1+e_{mn}^{o}}\right)^2. \tag{70}$$

- Variance of the beamforming gain uncertainty
  Regarding (56), We need to calculate the term $(\mathbb{E}\left\{|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}|^2\right\})$. By employing (62) and (64), we have

$$|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}|^2 \overset{a.s.}{\to} \frac{1}{\psi_{mn}^{o}}\frac{c_{mnk}^2(e_{mn}^{o})^2}{(1+e_{mn}^{o})^2}. \tag{71}$$

  Substituting (64), (69) and (71) in (56), we obtain

$$\begin{aligned}p_{I_1} &\overset{a.s.}{\to} p_{nk}\left(\sum_{m=1}^{M}\frac{1}{\psi_{mn}^{o}}\frac{c_{mnk}^2(e_{mn}^{o})^2}{(1+e_{mn}^{o})^2}\right.\\ &\quad\left.-\sum_{m=1}^{M}\frac{1}{\psi_{mn}^{o}}\frac{c_{mnk}^2(e_{mn}^{o})^2}{(1+e_{mn}^{o})^2}\right)\\ &\overset{a.s.}{\to} 0.\end{aligned} \tag{72}$$

  We also find

$$\mathbb{E}\left\{\left|\sum_{m=1}^{M}\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn}\right|^2\right\} \overset{a.s.}{\to} \left(\sum_{m=1}^{M}\frac{1}{\sqrt{\psi_{mn}^{o}}}\frac{c_{mnk}e_{mn}^{o}}{1+e_{mn}^{o}}\right)^2. \tag{73}$$

- Inter-cluster interference

In order to derive (57), we define

$$\mathbf{P}_{mn} \triangleq \mathrm{Diag}\left(\frac{p_1}{\psi^o_{m1}},\ldots,\frac{p_{n-1}}{\psi^o_{m(n-1)}},\frac{p_{n+1}}{\psi^o_{m(n+1)}},\ldots,\frac{p_N}{\psi^o_{mN}}\right). \tag{74}$$

By applying the matrix inversion Lemma, Lemma 1 and (74), (57) can be written as

$$\begin{aligned} p_{I2} &= \sum_{\substack{n'=1\\ n'\neq n}}^{N} p_{n'} \sum_{m=1}^{M} \mathbb{E}\left\{\left|\mathbf{h}^{\mathrm{H}}_{mnk}\mathbf{w}_{mn'}\right|^2\right\} \\ &= \sum_{m=1}^{M} \mathbb{E}\left\{\mathbf{h}^{\mathrm{H}}_{mnk}\boldsymbol{\Sigma}^{-1}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\boldsymbol{\Sigma}^{-1}_m\mathbf{h}_{mnk}\right\}. \end{aligned} \tag{75}$$

Since $\mathbf{h}_{mnk} = \hat{\mathbf{h}}_{mnk} + \boldsymbol{\epsilon}_{mnk}$ and $\hat{\mathbf{h}}_{mnk} = c_{mnk}\bar{\mathbf{h}}_{mn}$, we find that the estimated channels of the users in the same cluster are parallel. Accordingly, $\mathbf{h}_{mnk}$ can be written as

$$\mathbf{h}_{mnk} = \sqrt{L}\boldsymbol{\Theta}^{1/2}_{mn}[c_{mnk}\bar{\mathbf{g}}_{mn} + a_{mnk}\hat{\mathbf{g}}_{mnk}], \tag{76}$$

where $a_{mnk} = \sqrt{\frac{\beta_{mnk}-\theta_{mnk}}{1+\tau p_p\sum_{k=1}^{K}\beta_{mnk}}}$ and $\hat{\mathbf{g}}_{mnk} \sim \mathcal{CN}(\mathbf{0}, \frac{1}{L}\mathbf{I}_L)$ is independent of $\bar{\mathbf{g}}_{mn}$. By substituting (76) in (75), we find

$$\begin{aligned} &\mathbf{h}^{\mathrm{H}}_{mnk}\boldsymbol{\Sigma}^{-1}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\boldsymbol{\Sigma}^{-1}_m\mathbf{h}_{mnk} \\ &\quad= \frac{1}{L}c^2_{mnk}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn} \\ &\quad+\frac{1}{L}a^2_{mnk}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\hat{\mathbf{g}}_{mnk} \\ &\quad+\frac{1}{L}c_{mnk}a_{mnk}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\hat{\mathbf{g}}_{mnk} \\ &\quad+\frac{1}{L}c_{mnk}a_{mnk}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn}. \end{aligned} \tag{77}$$

where $\mathbf{C}_m = \boldsymbol{\Gamma}_m + \alpha\mathbf{I}_L$ with $\boldsymbol{\Gamma}_m = \frac{1}{L}\bar{\mathbf{H}}_m\bar{\mathbf{H}}^{\mathrm{H}}_m$. The equation in (77) can be further written as (78), shown at the bottom of this page,. Employing Lemma 2 of [23], $\mathbf{C}^{-1}_m - \mathbf{C}^{-1}_{mn} = -\mathbf{C}^{-1}_m(\mathbf{C}_m - \mathbf{C}_{mn})\mathbf{C}^{-1}_{mn}$ with $\mathbf{C}_m - \mathbf{C}_{mn} = \boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}$. Then, (78) can be written as (79), shown at the bottom of this page, where $\mathbf{B}_{mn} = \boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_{mn}\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}$ and $\mathbf{A}_{mn} = \boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}$. Then, by applying Lemma 7 of [23] to each quadratic form in (79), we obtain

$$\begin{aligned} a.\ \bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{A}_{mn}\bar{\mathbf{g}}_{mn} &\xrightarrow{a.s.} \frac{u_{mn}}{1+u_{mn}}, \\ b.\ \hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{A}_{mn}\bar{\mathbf{g}}_{mn} &\xrightarrow{a.s.} 0, \\ c.\ \bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\bar{\mathbf{g}}_{mn} &\xrightarrow{a.s.} \frac{u'_{mn}}{1+u_{mn}}, \\ d.\ \hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{B}_{mn}\hat{\mathbf{g}}_{mnk} &\xrightarrow{a.s.} u'_{mn}, \\ e.\ \hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{B}_{mn}\bar{\mathbf{g}}_{mn} &\xrightarrow{a.s.} 0, \end{aligned} \tag{80}$$

where $u_{mn} = \frac{1}{L}\mathrm{tr}[\boldsymbol{\Theta}_{mn}\mathbf{C}^{-1}_{mn}]$ and $u'_{mn} = \frac{1}{L}\mathrm{tr}[\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_{mn}\boldsymbol{\Theta}_{mn}\mathbf{C}^{-1}_{mn}\bar{\mathbf{H}}_{mn}]$. By substituting (80) in (79),

$$\begin{aligned} \mathbf{h}^{\mathrm{H}}_{mnk}\boldsymbol{\Sigma}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\boldsymbol{\Sigma}_m\mathbf{h}_{mnk} &= \frac{1}{L}c^2_{mnk}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_{mn}\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn} \\ &+\frac{1}{L}c^2_{mnk}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}(\mathbf{C}^{-1}_m - \mathbf{C}^{-1}_{mn})\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn} \\ &+\frac{1}{L}a^2_{mnk}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_{mn}\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\hat{\mathbf{g}}_{mnk} \\ &+\frac{1}{L}a^2_{mnk}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\boldsymbol{\Theta}^{1/2}_{mn}(\mathbf{C}^{-1}_m - \mathbf{C}^{-1}_{mn})\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\hat{\mathbf{g}}_{mnk} \\ &+\frac{1}{L}c_{mnk}a_{mnk}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_{mn}\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\hat{\mathbf{g}}_{mnk} \\ &+\frac{1}{L}c_{mnk}a_{mnk}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\boldsymbol{\Theta}^{1/2}_{mn}(\mathbf{C}^{-1}_m - \mathbf{C}^{-1}_{mn})\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\hat{\mathbf{g}}_{mnk} \\ &+\frac{1}{L}c_{mnk}a_{mnk}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\boldsymbol{\Theta}^{1/2}_{mn}\mathbf{C}^{-1}_{mn}\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn} \\ &+\frac{1}{L}c_{mnk}a_{mnk}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\boldsymbol{\Theta}^{1/2}_{mn}(\mathbf{C}^{-1}_m - \mathbf{C}^{-1}_{mn})\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\mathbf{C}^{-1}_m\boldsymbol{\Theta}^{1/2}_{mn}\bar{\mathbf{g}}_{mn}. \end{aligned} \tag{78}$$

$$\begin{aligned} \mathbf{h}^{\mathrm{H}}_{mnk}\boldsymbol{\Sigma}_m\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}^{\mathrm{H}}_{mn}\boldsymbol{\Sigma}_m\mathbf{h}_{mnk} &= c^2_{mnk}\left(\frac{1}{L}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\bar{\mathbf{g}}_{mn} - \frac{1}{L}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{A}_{mn}\bar{\mathbf{g}}_{mn}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\bar{\mathbf{g}}_{mn}\right) \\ &+a^2_{mnk}\left(\frac{1}{L}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{B}_{mn}\hat{\mathbf{g}}_{mnk} - \frac{1}{L}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{A}_{mn}\bar{\mathbf{g}}_{mn}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\hat{\mathbf{g}}_{mnk}\right) \\ &+a_{mnk}c_{mnk}\left(\frac{1}{L}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\hat{\mathbf{g}}_{mnk} - \frac{1}{L}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{A}_{mn}\bar{\mathbf{g}}_{mn}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\hat{\mathbf{g}}_{mnk}\right) \\ &+a_{mnk}c_{mnk}\left(\frac{1}{L}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{B}_{mn}\bar{\mathbf{g}}_{mn} - \frac{1}{L}\hat{\mathbf{g}}^{\mathrm{H}}_{mnk}\mathbf{A}_{mn}\bar{\mathbf{g}}_{mn}\bar{\mathbf{g}}^{\mathrm{H}}_{mn}\mathbf{B}_{mn}\bar{\mathbf{g}}_{mn}\right). \end{aligned} \tag{79}$$

we obtain

$$\begin{aligned}
&\mathbf{h}_{mnk}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-1}\bar{\mathbf{H}}_{mn}\mathbf{P}_{mn}\bar{\mathbf{H}}_{mn}^{\mathrm{H}}\boldsymbol{\Sigma}_m^{-1}\mathbf{h}_{mnk}\\
&\overset{a.s.}{\to} c_{mnk}^2\left(\frac{1}{L}\frac{u'_{mn}}{1+u_{mn}}-\frac{1}{L}\frac{u_{mn}u'_{mn}}{(1+u_{mn})^2}\right)\\
&\quad +a_{mnk}^2(\frac{1}{L}u'_{mn})\\
&\overset{a.s.}{\to} c_{mnk}^2\frac{1}{L}\frac{u'_{mn}}{(1+u_{mn})^2}+a_{mnk}^2\frac{1}{L}u'_{mn}, \qquad (81)
\end{aligned}$$

By employing Lemma 6 of [23], we have

$$\begin{aligned}
u_{mn} &\overset{a.s.}{\to} \frac{1}{L}\mathrm{tr}[\boldsymbol{\Theta}_{mn}\mathbf{C}_m^{-1}]\\
&\overset{a.s.}{\to} e_{mn}^o\\
\frac{1}{L}u'_{mn} &\overset{a.s.}{\to} \bar{\Upsilon}_{mn}, \qquad (82)
\end{aligned}$$

in which e_{mn}^o is given in (32) and $\bar{\Upsilon}_{mn} = \frac{1}{L^2}\mathrm{tr}[\mathbf{P}_{mn}\bar{\mathbf{H}}_{mn}^{\mathrm{H}}\mathbf{C}_m^{-1}\boldsymbol{\Theta}_{mn}\mathbf{C}_m^{-1}\bar{\mathbf{H}}_{mn}]$, which can be written as

$$\begin{aligned}
\bar{\Upsilon}_{mn} = \frac{1}{L}\sum_{\substack{n'=1\\n'\neq n}}^{N} p_{mn'}\bar{\mathbf{g}}_{mn'}^{\mathrm{H}}\boldsymbol{\Theta}_{mn'}^{1/2}\\
\mathbf{C}_m^{-1}\boldsymbol{\Theta}_{mn}\mathbf{C}_m^{-1}\boldsymbol{\Theta}_{mn'}^{1/2}\bar{\mathbf{g}}_{mn'}. \qquad (83)
\end{aligned}$$

Based on Theorem 2 of [23], we have

$$\bar{\Upsilon}_{mn} \overset{a.s.}{\to} \Upsilon_{mn}, \qquad (84)$$

where Υ_{mn} is given in (35).

Inserting (81) and (82) in (57), we obtain

$$\begin{aligned}
&\sum_{\substack{n'=1\\n'\neq n}}^{N} p_{n'}\sum_{m=1}^{M}\mathbb{E}\left\{\left|\mathbf{h}_{mnk}^{\mathrm{H}}\mathbf{w}_{mn'}\right|^2\right\}\\
&\qquad\overset{a.s.}{\to}\sum_{m=1}^{M}\Upsilon_{mn}\left(\frac{c_{mnk}^2}{(1+e_{mn}^o)^2}+a_{mnk}^2\right). \qquad (85)
\end{aligned}$$

Finally, by substituting (70), (72), (73) and (85) in (17), $\gamma_{nk}^{nk,\mathrm{mRZF}}$ is obtained as (31).

REFERENCES

[1] *IMT Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document Rec. ITU 2083-0, 2015.

[2] M. Matthaiou, C. Zhong, M. R. McKay, and T. Ratnarajah, "Sum rate analysis of ZF receivers in distributed MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 180–191, Feb. 2013.

[3] D. Wang, J. Wang, X. You, Y. Wang, M. Chen, and X. Hou, "Spectral efficiency of distributed MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2112–2127, Oct. 2013.

[4] C. He, J. Yin, Y. He, M. Huang, and B. Zhao, "Energy efficiency of distributed massive MIMO systems," *J. Commun. Netw.*, vol. 18, pp. 649–657, Aug. 2016.

[5] Y. Huang, G. Zheng, M. Bengtsson, K.-K. Wong, L. Yang, and B. Ottersten, "Distributed multicell beamforming with limited intercell coordination," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 728–738, Feb. 2011.

[6] J. Wang and L. Dai, "Asymptotic rate analysis of downlink multi-user systems with co-located and distributed antennas," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3046–3058, Jun. 2015.

[7] G. Interdonato, E. Björnson, H. Quoc Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," 2018, *arXiv:1804.03421*. [Online]. Available: http://arxiv.org/abs/1804.03421

[8] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[9] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[10] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive MIMO with zero-forcing precoding design," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1871–1874, Aug. 2017.

[11] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[12] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 616–619, Apr. 2019.

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

[14] J. Zhang, Y. Wei, E. Bjornson, Y. Han, and S. Jin, "Performance analysis and power control of cell-free massive MIMO systems with hardware impairments," *IEEE Access*, vol. 6, pp. 55302–55314, 2018.

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

[16] M. 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, 2017.

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

[18] D. Kudathanthirige and G. A. A. Baduge, "NOMA-aided multicell downlink massive MIMO," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 612–627, Jun. 2019.

[19] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, Jun. 2019.

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

[21] Y. Li and G. Amarasuriya, "NOMA-aided massive MIMO downlink with distributed antenna arrays," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[22] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "NOMA/OMA mode selection-based cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[23] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.

[24] H. Huh, A. M. Tulino, and G. Caire, "Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2911–2934, May 2012.

[25] F. Rezaei and A. Tadaion, "Multi-layer beamforming in uplink/downlink massive MIMO systems with multi-antenna users," *Signal Process.*, vol. 164, pp. 58–66, Nov. 2019.

[26] T. Van Chien, E. Bjornson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6384–6399, Sep. 2016.

[27] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Downlink spectral efficiency of cell-free massive MIMO with full-pilot zero-forcing," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 1003–1007.

[28] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, "System-level performance of downlink NOMA combined with SU-MIMO for future LTE enhancements," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 706–710.

[29] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Nov. 2013, pp. 770–774.

[30] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.
[31] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.
[32] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32–38, Apr. 2018.
[33] M. Vaezi, G. Amarasuriya, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," 2019, *arXiv:1903.10489*. [Online]. Available: http://arxiv.org/abs/1903.10489
[34] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.
[35] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[36] Z. Chen and E. Bjornson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.
[37] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.
[38] H. V. Cheng, E. Bjornson, and E. G. Larsson, "Performance analysis of NOMA in training-based multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 372–385, Jan. 2018.
[39] L. Bariah, S. Muhaidat, and A. Al-Dweik, "Error probability analysis of non-orthogonal multiple access over nakagami-m fading channels," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1586–1599, Feb. 2019.
[40] P. Li, R. C. de Lamare, and R. Fa, "Multiple feedback successive interference cancellation detection for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2434–2439, Aug. 2011.
[41] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[42] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.
[43] A. M. Tulino and S. Verdu, *Random Matrix Theory and Wireless Communications*. Boston, MA, USA: Now, 2004.
[44] J. Hoydis, S. T. Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
[45] A. Tang, J. Sun, and K. Gongand, "Mobile propagation loss with a low base station antenna for NLOS street microcells in urban area," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, May 2001, pp. 333–336.
[46] J. Zhu, R. Schober, and V. K. Bhargava, "Linear precoding of data and artificial noise in secure massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2245–2261, Mar. 2016.
[47] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*, 1st ed. New York, NY, USA: Cambridge Univ. Press, 2011.

Fatemeh Rezaei (Student Member, IEEE) received the M.Sc. degree in electrical engineering from Yazd University, Yazd, Iran, where she is currently pursuing the Ph.D. degree in electrical engineering. She is currently a visiting Ph.D. Student at the University of Alberta, Edmonton, AB, Canada. Her research interests are in the areas of wireless communications and signal processing, including MIMO and massive MIMO, NOMA, interference alignment, and cognitive radio networks.

Chintha Tellambura (Fellow, IEEE) received the B.Sc. degree in electronics and telecommunications from the University of Moratuwa, Sri Lanka, the M.Sc. degree in electronics from Kings College, University of London, and the Ph.D. degree in electrical engineering from the University of Victoria, Canada. He was with Monash University, Australia, from 1997 to 2002. Since 2002, he has been with the Department of Electrical and Computer Engineering, University of Alberta, where he is currently a Full Professor. He has authored or coauthored over 560 journal and conference papers, with an H-index of 75 (Google Scholar). He has supervised or co-supervised 66 M.Sc., Ph.D., and PDF trainees. His current research interests include cognitive radio, heterogeneous cellular networks, fifth-generation wireless networks, and machine learning algorithms. He was elected as a fellow of The Canadian Academy of Engineering in 2017. He received the Best Paper Award from the IEEE International Conference on Communications (ICC) in 2012 and 2017. He is the winner of the prestigious McCalla Professorship and the Killam Annual Professorship from the University of Alberta. He has served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS from 1999 to 2012, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2001 to 2007. He was an Area Editor of *Wireless Communications Systems and Theory* from 2007 to 2012.

Ali Akbar Tadaion (Senior Member, IEEE) was born in Iran in 1976. He received the B.Sc. degree in electronics, and the M.Sc. and Ph.D. degrees in communication systems from the Sharif University of Technology, Tehran, Iran, in 1998, 2000, and 2006, respectively. From August 2004 to June 2005, he was with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada, as a Visiting Researcher. He is currently an Associate Professor with the Department of Electrical Engineering, Yazd University, Yazd, Iran. His main research interests are statistical signal processing, detection theory, array signal processing, and wireless communications.

Ali Reza Heidarpour (Student Member, IEEE) received the B.Sc. degree from the University of Isfahan, Isfahan, Iran, in 2013, and the M.Sc. degree from Ozyegin University, Istanbul, Turkey, in 2016. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Alberta, Edmonton, AB, Canada. His research interests include cooperative communications, orthogonal frequency division multiplexing, multiple-input multiple-output (MIMO) systems, and cell-free massive MIMO systems.