Joint Power Control and Clustering in NOMA-enabled Cell-free Massive MIMO

Muralikrishnan Srinivasan[1] and Arsenia Chorti[2]

Abstract—**We investigate the downlink of a cell-free massive multiple-input multiple-output (CF-mMIMO) system with non-orthogonal multiple access (NOMA), where two sets of user equipment (UE) with distinct quality of service (QoS) requirements are served. The cUEs are required to achieve a sum ergodic rate maximization goal, whereas the pUEs need to fulfill a minimum rate threshold. Since there is no feasible optimal solution available, we propose computationally feasible power-control and clustering algorithms that maximize the ergodic rate of cUEs without compromising the rate requirements of pUEs.**

Nous étudions la liaison descendante d'un système à entrées multiples et sorties multiples massif (CF-mMIMO) sans cellule avec accès multiple non-orthogonal (NOMA), où deux ensembles d'équipements utilisateurs (UE) avec des exigences de qualité de service (QoS) distinctes sont servis. Les UE de type cUE doivent atteindre un objectif de maximisation du débit ergodique total, tandis que les UE de type pUE doivent respecter un seuil de débit minimum. Comme il n'existe pas de solution optimale réalisable, nous proposons des algorithmes de contrôle de puissance et de regroupement calculatoirement réalisables qui maximisent le débit ergodique des cUEs sans compromettre les exigences de débit des pUEs.

Index Terms—**Cell-free massive MIMO, NOMA, Optimization, Power-control, QoS.**

I. Introduction

Cell-free massive multiple-input multiple-output (CF-mMIMO) systems have been extensively studied under different scenarios, such as downlink pilot training, channel hardening, hardware impairments, non-orthogonal multiple access (NOMA), etc., by several treatises [1]–[3]. Concerning NOMA, in particular, its intersection with CF-mMIMO offers a significant advantage in terms of spectral efficiency (SE) gains [4]. There has been some early literature on the study of the performance of NOMA in CF-mMIMO systems [2], [5]. In a very recent and essential contribution [4], the authors have extensively studied the performance of three different precoders in the context of power-domain NOMA in CF-mMIMO systems.

It has been shown in [4] that NOMA-enabled CF-mMIMO systems allow more users to be simultaneously supported when compared to OMA. However, the optimal power allocation and clustering of users in the NOMA-enabled CF-mMIMO system is still largely unexplored, especially when the user equipment (UEs) have different quality of service (QoS) requirements. Very recently, an algorithm to maximize the total uplink SE subject to SIC and QoS constraints was proposed in [6], while the use of unsupervised machine learning algorithms to cluster users was discussed only recently in [7].

This paper considers a downlink NOMA-enabled CF-mMIMO system without pilot contamination and devises power control and clustering strategies that maximize the SE of a set of cellular UEs with moderate throughput requirements. Simultaneously, the algorithm ensures that a minimum SE is guaranteed for a group of bandwidth-hungry users, which we refer to as priority UEs.[1] Furthermore, we study the performance of our allocation strategies with extensive simulations. The following notation is used in the paper: $\mathcal{C}$ denotes a complex number, $\mathcal{Z}$ denotes an integer, $(\cdot)^H$ denotes the Hermitian of a vector/matrix, $\mathbb{E}(\cdot)$ denotes the expectation of a random variable, $\mathbf{I}$ denotes the identity matrix, $\mathcal{CN}(\cdot)$ denotes the complex normal random variable and $\mathbb{1}_{(\cdot)}$ denotes the indicator function.

II. System Model

We consider the downlink of a CF-mMIMO system with M access points (APs), N cellular user equipment (cUE) and K priority user equipment (pUE). We also assume that $M \gg N + K$. Each AP has L antennas and each user has only one 1 antenna. The channel coefficient, denoted by $\mathbf{g}_{mn}^{(c)} \in \mathcal{C}^{L\times 1}$, between the m-th AP and the nth cUE, is modeled as follows:

$$\mathbf{g}_{mn}^{(c)} = \sqrt{\beta_{mn}^{(c)}}\mathbf{h}_{mn}^{(c)},\ \forall m = 1,\dots,M,\ \forall n = 1,\dots,N, \quad (1)$$

where $\beta_{mn}^{(c)}$ represents the large-scale fading coefficient and $\mathbf{h}_{mn}^{(c)}$ represent small-scale fading coefficients that $\forall m = 1,\dots,M,\ n = 1,\dots,N$ are independent and identically distributed (i.i.d.) random variables. Similarly, the channel coefficient, $\mathbf{g}_{mk}^{(p)} \in \mathcal{C}^{L\times 1}$, between the m-th AP and the kth pUE, is modeled as $\mathbf{g}_{mk}^{(p)} = \sqrt{\beta_{mk}^{(p)}}\mathbf{h}_{mk}^{(p)},\ \forall m = 1,\dots,M,\ \forall k = 1,\dots,K$, where $\beta_{mk}^{(p)}$ represents the large-scale fading coefficient and $\mathbf{h}_{mk}^{(p)}$ represents the small-scale fading coefficient, with the same set of assumptions as $\mathbf{g}_{mn}^{(c)}$.

A. Training phase

Let τ denote the uplink training duration in samples per coherence interval. Let $\sqrt{\tau}\boldsymbol{\phi}_n^{(c)} \in \mathcal{C}^{\tau\times 1}$ be the pilot sequence transmitted by the nth cUE. Similarly, $\sqrt{\tau}\boldsymbol{\phi}_k^{(p)} \in \mathcal{C}^{\tau\times 1}$ be

1. Muralikrishnan Srinivasan is with Electrical Engineering, Chalmers University of Chalmers University of Technology. (Email:mursri@chalmers.se)

2. Arsenia Chorti is with ETIS UMR8051, CY University, ENSEA, CNRS, Cergy, France. Arsenia Chorti is also a Wireless Connectivity Research Group Leader at Barkhausen Institut gGmbH, Dresden DE. (Email: arsenia.chorti@ensea.fr)

[1]Typically, priority users might also have stringent delay constraints, an aspect that will be investigated in future works.

the pilot sequence transmitted by the kth pUE. Note that, if $\tau > N+K$, then we can choose pairwise orthonormal pilot sequences and hence avoid pilot contamination [8]. The signal received at m-th AP $\forall m = 1,\ldots,M$ during the training phase can be expressed as $\mathbf{Y}_m = \sqrt{\tau\rho}\sum_{n=1}^{N}\mathbf{g}_{mn}^{(c)}\left(\boldsymbol{\phi}_n^{(c)}\right)^H + \sqrt{\tau\rho}\sum_{k=1}^{K}\mathbf{g}_{mk}^{(p)}\left(\boldsymbol{\phi}_k^{(p)}\right)^H + \mathbf{Z}_m$, where ρ is the normalized transmit SNR of each pilot symbol, $\mathbf{Z}_m \in \mathcal{C}^{L\times\tau}$ is the noise vector whose entries are i.i.d. zero-mean unit-variance complex Gaussian. To estimate the channel coefficients, define $\bar{\mathbf{G}}_m \triangleq \mathbf{Y}_m\boldsymbol{\Phi}$, where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1^{(c)},\ldots,\boldsymbol{\phi}_N^{(c)},\boldsymbol{\phi}_1^{(p)},\ldots,\boldsymbol{\phi}_K^{(p)}]$. $\bar{\mathbf{G}}_m$ can be then written as

$$\bar{\mathbf{G}}_m = \left[\sqrt{\tau\rho}\mathbf{g}_{m1}^{(c)} + \tilde{\mathbf{n}}_{m1}^{(c)},\ldots,\sqrt{\tau\rho}\mathbf{g}_{mK}^{(p)} + \tilde{\mathbf{n}}_{mK}^{(p)}\right], \tag{2}$$

where $\tilde{\mathbf{n}}_{mn}^{(c)} \sim \mathcal{CN}(0,\mathbf{I}_L)$ and $\tilde{\mathbf{n}}_{mk}^{(p)} \sim \mathcal{CN}(0,\mathbf{I}_L)$, since the entries of $\mathbf{Z}_m\boldsymbol{\Phi} \in \mathcal{C}^{L\times(N+K)}$ are i.i.d. zero-mean and unit-variance complex Gaussian random variables. Therefore, the columns of $\bar{\mathbf{G}}_m$ are given by $\bar{\mathbf{g}}_{mn}^{(c)} \sim \mathcal{CN}(\mathbf{0},(1+\tau\rho\beta_{mn}^{(c)})\mathbf{I}_L)$ and $\bar{\mathbf{g}}_{mk}^{(p)} \sim \mathcal{CN}(\mathbf{0},(1+\tau\rho\beta_{mk}^{(p)})\mathbf{I}_L)$ $\forall m = 1,\ldots,M$, $\forall n = 1,\ldots,N$ and $\forall k = 1,\ldots,K$. The MMSE estimator of the channel coefficients $\forall m,n,k$ is hence given by $\hat{\mathbf{g}}_{mn}^{(c)} = \frac{\sqrt{\tau\rho}\beta_{mn}^{(c)}}{\tau\rho\beta_{mn}^{(c)}+1}\bar{\mathbf{g}}_{mn}^{(c)}$ and $\hat{\mathbf{g}}_{mk}^{(p)} = \frac{\sqrt{\tau\rho}\beta_{mk}^{(p)}}{\tau\rho\beta_{mk}^{(p)}+1}\bar{\mathbf{g}}_{mk}^{(p)}$. $\hat{\mathbf{g}}_{mn}^{(c)}$ and $\hat{\mathbf{g}}_{mn}^{(c)}$ $\forall m,n,k$ can also be written as, $\hat{\mathbf{g}}_{mn}^{(c)} = \sqrt{\theta_{mn}^{(c)}}\nu_{mn}^{(c)}$ and $\hat{\mathbf{g}}_{mk}^{(p)} = \sqrt{\theta_{mk}^{(p)}}\nu_{mk}^{(p)}$, respectively, where $\nu_{mn}^{(c)}$ and $\nu_{mk}^{(p)}$ $\forall m,n,k$ are $\mathcal{CN}(\mathbf{0},\mathbf{I}_L)$ and $\theta_{mn}^{(c)} = \frac{\tau\rho\left(\beta_{mn}^{(c)}\right)^2}{1+\tau\rho\beta_{mn}^{(c)}}$ and $\theta_{mk}^{(p)} = \frac{\tau\rho\left(\beta_{mk}^{(p)}\right)^2}{1+\tau\rho\beta_{mk}^{(p)}}$ $\forall m,n,k$. The channel estimation error can then be defined as $\boldsymbol{\epsilon}_{mn}^{(c)} = \mathbf{g}_{mn}^{(c)} - \hat{\mathbf{g}}_{mn}^{(c)}$ where $\boldsymbol{\epsilon}_{mn}^{(c)} \sim \mathcal{CN}(\mathbf{0},(\beta_{mn}^{(c)} - \theta_{mn}^{(c)})\mathbf{I}_L)$. Similarly, $\boldsymbol{\epsilon}_{mk}^{(p)} = \mathbf{g}_{mk}^{(p)} - \hat{\mathbf{g}}_{mk}^{(p)}$ where $\boldsymbol{\epsilon}_{mk}^{(p)} \sim \mathcal{CN}(\mathbf{0},(\beta_{mk}^{(p)} - \theta_{mk}^{(p)})\mathbf{I}_L)$, $\forall m,n,k$. Also, for maximum ratio transmission (MRT) precoding, the precoding vectors from the mth AP to the users in the nth cUE and kth pUE $\forall m = 1,\ldots,M$, $\forall n = 1,\ldots,N$ and $\forall k = 1,\ldots,K$ are, respectively, expressed as $\mathbf{w}_{mn}^{(c)} = \frac{\hat{\mathbf{g}}_{mn}^{(c)}}{\mathbb{E}\left[\left\|\hat{\mathbf{g}}_{mn}^{(c)}\right\|\right]}$ and $\mathbf{w}_{mk}^{(p)} = \frac{\hat{\mathbf{g}}_{mk}^{(p)}}{\mathbb{E}\left[\left\|\hat{\mathbf{g}}_{mk}^{(p)}\right\|\right]}$.

B. Downlink phase

Assuming each cUE is paired with a pUE to form an nth NOMA pair/cluster, where only one cUE and pUE can be present in each cluster/pair, we define $\omega_{n,k} \in 0,1$ to indicate cluster sharing. If the kth pUE is paired with the nth cUE, then $\omega_{n,k} = 1$, else $\omega_{n,k} = 0$. Users are ordered in each cluster based on the mean of their effective channel gains. In every cluster, we assume that the users are ordered based on the mean of the effective channel gains, similar to [4]. The effective channel gains of the nth cUE and kth pUE are given by $\chi_n^{(c)} = \mathbb{E}\left\{\left|\sum_{m=1}^{M}\left(\hat{\mathbf{g}}_{mn}^{(c)}\right)^{\mathrm{H}}\mathbf{w}_{mn}^{(c)}\right|^2\right\}$ and $\chi_k^{(p)} = \mathbb{E}\left\{\left|\sum_{m=1}^{M}\left(\hat{\mathbf{g}}_{mk}^{(p)}\right)^{\mathrm{H}}\mathbf{w}_{mk}^{(p)}\right|^2\right\}$ respectively, if the kth pUE is paired with the nth cUE. Without loss of generality, in each cluster, we assume that user 1 denotes the user with higher effective channel gain. Therefore, define the triplets $\{\mathbf{g}_{mn1},\hat{\mathbf{g}}_{mn1},\mathbf{w}_{mn1}\}$ and $\{\mathbf{g}_{mn2},\hat{\mathbf{g}}_{mn2},\mathbf{w}_{mn2}\}$ to denote the actual channel, the estimate and the precoding vector of user 1 and user 2 in each cluster. For example if $\chi_n^{(c)} > \chi_k^{(p)}$, then $\{\mathbf{g}_{mn1},\hat{\mathbf{g}}_{mn1},\mathbf{w}_{mn1}\} \triangleq \{\mathbf{g}_{mn}^{(c)},\hat{\mathbf{g}}_{mn}^{(c)},\mathbf{w}_{mn}^{(c)}\}$ and $\{\mathbf{g}_{mn2},\hat{\mathbf{g}}_{mn2},\mathbf{w}_{mn2}\} \triangleq \{\mathbf{g}_{mk}^{(p)},\hat{\mathbf{g}}_{mk}^{(p)},\mathbf{w}_{mk}^{(p)}\}$ and vice-versa. Higher power is allocated to the user with lower effective channel gain. i.e., $P_{n1} \le P_{n2}$, where $P_{nr} = P_u\delta_{nr}$ is the transmit power and q_{nr} denotes the data signal to the rth user in the nth cluster $\forall n = 1,\ldots N$. Note that, the total power allocated to each cluster from each AP is P_u and $P_u = \frac{\text{Total power budget}}{NM}$.

The 1st user decodes the messages of the 2nd user using SIC. Also assume that $\sum_{r=1}^{2}\delta_{nr} = 1, \forall n$. The mth AP therefore transmits $\mathbf{x}_m = \sum_{n=1}^{N}\sum_{r=1}^{2}\mathbf{w}_{mnr}\sqrt{P_{nr}}q_{nr}$. The received signal at the rth user in the nth cluster can be expressed, using the statistical knowledge of the effective channel gains at the users, as

$$\begin{aligned}
y_{nr} &= \underbrace{\sum_{m=1}^{M}\sqrt{P_{nr}}\mathbb{E}\left\{\mathbf{g}_{mnr}^{\mathrm{H}}\mathbf{w}_{mnr}\right\}q_{nr}}_{T_0:\text{desired signal}} \\
&\quad + \underbrace{\sum_{m=1}^{M}\sqrt{P_{nr}}\left(\mathbf{g}_{mnr}^{\mathrm{H}}\mathbf{w}_{mnr} - \mathbb{E}\left\{\mathbf{g}_{mnr}^{\mathrm{H}}\mathbf{w}_{mnr}\right\}\right)q_{nr}}_{T_1:\text{beamforming gain uncertainty}} \\
&\quad + \underbrace{\sum_{m=1}^{M}\mathbf{g}_{mnk}^{\mathrm{H}}\sum_{r'<r}\mathbf{w}_{mnr'}\sqrt{P_{nr'}}q_{nr'}}_{T_2:\text{intra-cluster interference after SIC}} \\
&\quad + \underbrace{\sum_{m=1}^{M}\mathbf{g}_{mnr}^{\mathrm{H}}\sum_{\substack{n'=1\\n'\neq n}}^{N}\sum_{r'=1}^{2}\mathbf{w}_{mn'r'}\sqrt{P_{n'r'}}q_{n'r'} + n_{nk},}_{T_3:\text{Inter-cluster interference}}
\end{aligned} \tag{3}$$

where the desired signal is defined by T_0, the beamforming gain uncertainty is defined by T_1. The third term T_2 represents the intra-cluster interference after SIC and T_3 represents the inter-cluster interference.

In the NOMA-enabled CF-mMIMO system with MRT precoding, the approximate instantaneous signal-to-interference-noise ratio (SINR) of the rth user in the nth cluster is γ_{nr}, for finite values of M, L, N and K, is given by (4), at the top of the next page, where $\eta_{nr} \triangleq \sum_{m=1}^{M}\mathbf{g}_{mnr}^{\mathrm{H}}\mathbf{w}_{mnr}$, $\eta_{nr'} \triangleq \sum_{m=1}^{M}\mathbf{g}_{mnr}^{\mathrm{H}}\mathbf{w}_{mnr'}$ and $\eta_{n'r'} \triangleq \sum_{m=1}^{M}\mathbf{g}_{mnr}^{\mathrm{H}}\mathbf{w}_{mn'r'}$.

The instantaneous rate for the rth user in the nth cluster can then be computed as $R_{nr} = \log_2(1+\gamma_{nr})$, and the ergodic rate by $\bar{R}_{nr} = \mathbb{E}[\log_2(1+\gamma_{nr})]$. Note that obtaining a closed-form expression for the ergodic rate is mathematically intractable. However, we can utilize the use-and-then-forget bound to obtain an approximation [4]

$$\bar{R}_{nr} \approx \log_2(1+\bar{\gamma}_{nr}), \tag{5}$$

$$\gamma_{nr} = \frac{P_{nr}\left|\mathbb{E}\left\{\eta_{nr}\right\}\right|^2}{P_{nr}\mathbb{E}\left\{\left|\eta_{nr}-\mathbb{E}\left\{\eta_{nr}\right\}\right|^2\right\} + \sum\limits_{r'<r} P_{nr'}\mathbb{E}\left\{|\eta_{nr'}|^2\right\} + \sum\limits_{\substack{n'=1\\ n'\neq n}}^{N}\sum\limits_{r'=1}^{2} P_{n'r'}\mathbb{E}\left\{\left|\eta_{n'r'}\right|^2\right\}+1}. \tag{4}$$

where $\bar{\gamma}_{nr}$ is given by (6), at the top of the next page, and can be derived trivially following the steps in [4].

III. Spectrum and Power Allocation Formulation

Let $\bar{R}_n^{(c)}$ $\forall n$ and $\bar{R}_k^{(p)}$ $\forall k$ denote the ergodic rate of cUEs and pUEs respectively. They can be expressed, in terms of (5), as $\bar{R}_n^{(c)} = \sum_{k=1}^{K} \bar{R}_{n1}\mathbb{1}_{\chi_n^{(c)}>\chi_k^{(p)}}\mathbb{1}_{\omega_{n,k}} + \sum_{k=1}^{K} \bar{R}_{n2}\mathbb{1}_{\chi_n^{(c)}<\chi_k^{(p)}}\mathbb{1}_{\omega_{n,k}}$ and $\bar{R}_k^{(p)} = \sum_{n=1}^{N} \bar{R}_{n1}\mathbb{1}_{\chi_k^{(p)}>\chi_n^{(c)}}\mathbb{1}_{\omega_{n,k}} + \sum_{n=1}^{N} \bar{R}_{n2}\mathbb{1}_{\chi_n^{(c)}<\chi_k^{(p)}}\mathbb{1}_{\omega_{n,k}}$ The spectrum and power allocation problem is formulated in this work as:

$$\max_{\{\omega_{n,k}\},\{\delta_{nr}\}} \sum_{n=1}^{N} \bar{R}_n^{(c)}, \tag{7}$$

$$\text{s.t.} \quad \bar{R}_k^{(p)} \geq R_{th}, \forall k \tag{7a}$$

$$\sum_{n} \omega_{n,k} = 1,\ \omega_{n,k} \in \{0,1\}, \forall k, \tag{7b}$$

$$\sum_{k} \omega_{n,k} = 1,\ \omega_{n,k} \in \{0,1\}, \forall n, \tag{7c}$$

$$0 \leq \delta_{n1} \leq \delta_{n2} \leq 1, \forall n, \tag{7d}$$

$$\delta_{n1} + \delta_{n2} = 1, \forall n, \tag{7e}$$

We have constraints for minimum rate requirements for pUE links (Eq. (7a)), one pUE link per cUE cluster (Eq. (7b)), one cUE cluster per pUE link (Eq. (7c)), and NOMA power constraints for cUE and pUE links (Eqs. (7d) and (7e)). Due to integer constraints, obtaining an optimal solution to the power and spectrum allocation problem is computationally exhaustive. Therefore, we discuss sub-optimal yet effective power allocation and clustering procedures in the following subsections.

A. Power allocation procedure

To circumvent the power-allocation search over 2^N possibilities, we assume that all the $2(N-1)$ inter-cluster interference terms are present, which is a much stricter condition to be satisfied. Now, for a cluster n with a cUE and pUE, note that the objective function monotonically increases with the power allocated to the cUE. However, the presence of the minimum rate constraints of the priority users should also be satisfied. If $\chi_n^{(c)} > \chi_k^{(p)}$, then $\frac{LP_u\delta_{n2}\left(\sum\limits_{m=1}^{M}\sqrt{\theta_{mnr}}\right)^2}{P_u\sum\limits_{m=1}^{M}\beta_{mnr}+2(N-1)P_u\sum\limits_{m=1}^{M}\beta_{mnr}+1} \geq 2^{R_{th}}-1$ and if $\chi_n^{(c)} < \chi_k^{(p)}$, $\frac{LP_u\delta_{n1}\left(\sum\limits_{m=1}^{M}\sqrt{\theta_{mnr}}\right)^2}{\delta_{n1}P_u\sum\limits_{m=1}^{M}\beta_{mnr}+2(N-1)P_u\sum\limits_{m=1}^{M}\beta_{mnr}+1} \geq 2^{R_{th}}-1$. The above relations have to be satisfied with equality, otherwise the power allocated to cUEs can be further increased while satisfying the above relations and maximizing the objective at the same time. Therefore, solving the above equations under the equality sign will give us the optimal $(\delta_{n1}^*, \delta_{n2}^*)$. If $\delta_{n1}^* > \delta_{n2}^*$, then the pUE is not served in that cluster and only the cUE is served.

Algorithm 1 Greedy Algorithm for cUE-pUE clustering

Arbitrarily assign one pUE link to each of the N clusters.
Compute power allocation as in Section III-A and compute the objective obj $= \sum_{n=1}^{N} \bar{R}_n^{(c)}$.
for $n = 1 : N-1$ **do**
  pUE in the nth cluster is $k = \arg\max\limits_{k} \omega_{n,k}$
  for $n' = n : N$ **do**
    pUE in the n'th cluster is $k' = \arg\max\limits_{k} \omega_{n',k}$
    Exchange the pUEs between the nth and n'th cluster, i.e., $\omega_{n,k} = 0$, $\omega_{n,k'} = 1$, $\omega_{n',k} = 1$ and $\omega_{n',k'} = 0$
    Denote the exchange as $\zeta_{n,n'}$
    Compute the new power allocation and compute the new objective obj$^*(\zeta_{n,n'})$.
  end for
  Exchange the pUEs between the nth and n'th cluster if $n' = \arg\max\limits_{n'}$ obj$^*(\zeta_{n,n'})$ and obj$^* >$ obj.
end for
Return the clustering result.

B. Clustering

One sub-optimal solution is to search sequentially for the clustering that maximizes the objective by considering all $N!$ different clustering arrangements. An alternative is a greedy heuristic that takes $\frac{N^2-N}{2}$ steps to obtain a sub-optimal solution. At each step, the algorithm exchanges the pUEs of the nth cluster with each of the subsequent clusters, computes the power allocation, and the new objective function. If the new objective function is greater than the existing one, the corresponding exchange is chosen. This process is repeated until $n = N-1$. The steps of the clustering algorithm are detailed in Algorithm 1.

IV. Simulation Results

We examine a CF-mMIMO system with varying M, N, and K in a $D \times D$ km^2 square area with M APs and $N+K$ users. The total power budget is 100 mW across all APs, and $D = 1$. Our simulation setup is similar to that in [8]. We plot the sum ergodic rate of cUEs obtained by the greedy clustering algorithm against M, for different N, K, and R_{th} in Figs. 1 and 2. Table I shows the corresponding average percentage of

$$\bar{\gamma}_{nr} = \frac{LP_{nr}\left(\sum_{m=1}^{M}\sqrt{\theta_{mnr}}\right)^2}{P_{nr}\sum_{m=1}^{M}\beta_{mnr} + \sum_{r'<r}P_{nr'}\sum_{m=1}^{M}\beta_{mnr} + 2(N-1)P_u\sum_{m=1}^{M}\beta_{mnr} + 1}. \quad (6)$$

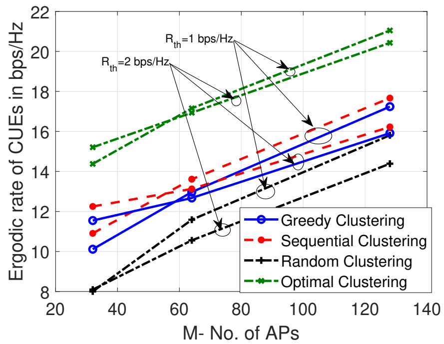

Fig. 1: Sum of ergodic rate of cUEs vs No of APs M for $N = K = 5$, $L = 8$ for perfect SIC

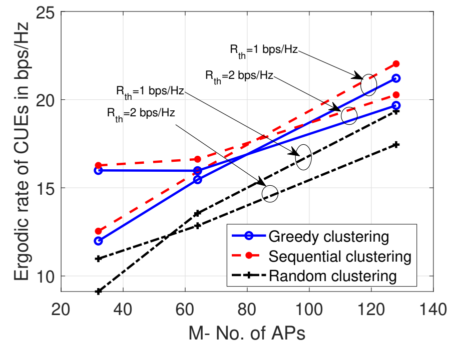

Fig. 2: Sum of ergodic rate of cUEs vs No of APs M for $N = K = 8$, $L = 8$ for perfect SIC

TABLE I: Table: Average percentage of pUEs satisfying the rate threshold for Greedy allocation

N, R_{th}	$M = 32$	$M = 64$	$M = 128$
$N = 5, R_{th} = 1$	69.5	87.5	96.5
$N = 5, R_{th} = 2$	39.5	58.5	81.0
$N = 8, R_{th} = 1$	58.0	76.8	91.8
$N = 8, R_{th} = 2$	24.2	43.2	66.8

TABLE II: Table: Average percentage of pUEs satisfying the rate threshold for Random allocation

N, R_{th}	$M = 32$	$M = 64$	$M = 128$
$N = 5, R_{th} = 1$	85.7	94.6	98.5
$N = 5, R_{th} = 2$	52.6	74.1	88.9
$N = 8, R_{th} = 1$	73.1	89.0	96.6
$N = 8, R_{th} = 2$	27.0	55.1	77.5

pUEs meeting the rate threshold. The greedy clustering algorithm performs similarly to the sequential clustering algorithm with fewer steps. The rate achieved by the greedy algorithm is better than random clustering for almost all M, N, and K. We also show the optimal clustering result for $N = 5$. The greedy algorithm bridges the gap between random and optimal clustering by nearly 50%, and for $N = 8$, the optimal algorithm is not shown due to computational exhaustion.

As R_{th} increases, fewer pUEs meet the threshold, resulting in a higher cUE rate for lower M ($M = 32$). However, for higher M values ($M = 64$ and $M = 128$), more pUEs meet the threshold, leading to lower cUE rates due to power allocation. Nevertheless, the cUE rate for $M = 128$ increases. Increasing N and K increases the cUE rate for $M = 32$ at the cost of fewer pUEs meeting the threshold, as shown in Table I. For lower M values, random allocation performs better than greedy allocation, which prioritizes objective maximization over pUE satisfaction. However, an algorithm that equally prioritizes pUEs is not addressed in this letter and can be a topic for future work.

V. Conclusions

We studied a NOMA-enabled CF-mMIMO system with two sets of UEs with different QoS rate requirements in the downlink. We proposed power-control and clustering solutions that balance the ergodic rate of cUE with the minimum rate requirements of pUEs. Our solutions were sub-optimal but computationally tractable. We conducted extensive simulations and found that our greedy clustering heuristic achieved similar performance to the sequential clustering with a significant reduction in complexity. The cUEs also achieved a higher ergodic rate compared to random allocation.

References

[1] Z. Chen and E. Björnson, “Channel Hardening and Favorable Propagation in Cell-Free Massive MIMO With Stochastic Geometry,” *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, 2018.

[2] 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, 2018.

[3] S. Shekhar, M. Srinivasan, and S. Kalyani, “Outage Probability of Uplink Cell-Free Massive MIMO Network with Imperfect CSI Using Dimension-Reduction Method,” *arXiv preprint arXiv:2101.07737*, 2021.

[4] F. Rezaei, C. Tellambura, A. A. Tadaion, and A. R. Heidarpour, “Rate Analysis of Cell-Free Massive MIMO-NOMA with Three Linear Precoders,” *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3480–3494, 2020.

[5] 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.* IEEE, 2019, pp. 1–6.

[6] Y. Zhang, H. Cao, M. Zhou, and L. Yang, “Spectral efficiency maximization for uplink Cell-Free Massive MIMO-NOMA networks,” in *Proc. IEEE Int. Conf. Commun. Workshops*. IEEE, 2019, pp. 1–6.

[7] Q. N. Le, V.-D. Nguyen, N.-P. Nguyen, S. Chatzinotas, O. A. Dobre, and R. Zhao, “Learning-Assisted User Clustering in Cell-Free Massive MIMO-NOMA Networks,” *arXiv preprint arXiv:2011.07549*, 2020.

[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, 2017.