NOMA-Based Coexistence of Near-Field and Far-Field Massive MIMO Communications

Zhiguo Ding, *Fellow, IEEE*, Robert Schober, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—This letter considers a legacy massive multiple-input multiple-output (MIMO) network, in which spatial beams have been preconfigured for near-field users, and proposes to use the non-orthogonal multiple access (NOMA) principle to serve additional far-field users by exploiting the spatial beams preconfigured for the legacy near-field users. Our results reveal that the coexistence between near-field and far-field communications can be effectively supported via NOMA, and that the performance of NOMA-assisted massive MIMO can be efficiently improved by increasing the number of antennas at the base station.**

I. Introduction

Non-orthogonal multiple access (NOMA) has been recognized as a promising component of the sixth-generation (6G) wireless network due to its superior spectral efficiency [1], [2]. One recent advance in NOMA is its use as an add-on in a massive multiple-input multiple-output (MIMO) based legacy space division multiple access (SDMA) network, where spatial beams preconfigured for legacy users are used to serve additional users [3]. As a result, the connectivity and the overall system throughput of SDMA can be improved in a low-complexity and spectrally efficient manner. For conventional SDMA networks based on far-field communications, where the transceiver distance is larger than the Rayleigh distance [4], this application of NOMA is intuitive, as explained in the following. Far-field beamforming is based on steering vectors, i.e., the spatial beams are cone-shaped [5]. In practice, each of these cone-shaped beams can cover a large area, and it is intuitive to encourage multiple users which are inside of one cone-shaped area to exploit the same beam via NOMA.

Near-field communications have recently received a lot of attention because, for high carrier frequencies and large numbers of antennas, the Rayleigh distance becomes significantly large [6], [7]. Unlike far-field communications, the spherical-wave channel model has to be used for near-field communications, which motivates the use of beam-focusing, i.e., a beam is focused on not only a spatial direction but also a specific location [8]. As a result, a naturally arising question is whether the preconfigured spatial beams in near-field communication networks can still be used to admit additional users in the same manner as far-field beams, which is the motivation for this work.

This letter considers a legacy near-field SDMA network, in which spatial beams have been preconfigured for legacy near-field users, and proposes to apply the principle of NOMA to serve additional far-field users by exploiting these preconfigured spatial beams. A resource allocation optimization problem is formulated to maximize the far-field users' sum data rate while guaranteeing the legacy near-field users' quality-of-service (QoS) requirements. First, a suboptimal low-complexity algorithm based on successive convex approximation (SCA) is proposed to solve the problem, and then the optimal performance is obtained for two special cases by applying the branch-and-bound (BB) algorithm [9], [10]. Simulation results are presented to demonstrate that the use of NOMA can effectively support the coexistence of near-field and far-field communications, and the performance of NOMA assisted massive MIMO can be efficiently improved by increasing the number of antennas at the base station.

Z. Ding is with Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE, and Department of Electrical and Electronic Engineering, University of Manchester, Manchester, UK. R. Schober is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg (FAU), Germany. H. V. Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA.

II. System Model

Consider a legacy downlink near-field SDMA network, in which a base station employs an N-antenna uniform linear array (ULA) and serves M single-antenna near-field users, where $M \leq N$. In this letter, it is assumed that M near-field beamforming vectors, denoted by $\mathbf{p}_m$, have already been configured to serve the legacy users individually. The aim of this letter is to admit K additional far-field users based on these preconfigured spatial beams. Denote the 2-dimensional coordinates of the m-th near-field user, the k-th far-field user, the center of the array, and the n-th element of the array by $\boldsymbol{\psi}_m^{\rm NF}$, $\boldsymbol{\psi}_k^{\rm FF}$, $\boldsymbol{\psi}_0$, and $\boldsymbol{\psi}_n$, respectively. According to the principle of near-field communications, $|\boldsymbol{\psi}_m^{\rm NF}-\boldsymbol{\psi}_0| < d_R(N)$, and $|\boldsymbol{\psi}_k^{\rm FF}-\boldsymbol{\psi}_0| > d_R(N)$, where $d_R(N) = \frac{2((N-1)d)^2}{\lambda}$, λ, and d denote the Rayleigh distance, the wavelength, and the antenna spacing of the ULA, respectively [8], [11].

A. Near-Field and Far-Field Channel Models

The m-th legacy near-field user's observation is given by

$$y_m = \mathbf{h}_m^H \mathbf{x} + n_m, \tag{1}$$

where $(\cdot)^H$ denotes Hermitian transpose, $\mathbf{x}$ denotes the signal vector sent by the base station, n_m denotes the additive Gaussian noise with its power denoted by σ^2, the spherical-wave propagation model is used to describe the near-field user's channel vector as follows [7], [8], [12]:

$$\mathbf{h}_m = \alpha_m \left[e^{-j\frac{2\pi}{\lambda}|\boldsymbol{\psi}_m^{\rm NF}-\boldsymbol{\psi}_1|} \quad \ldots \quad e^{-j\frac{2\pi}{\lambda}|\boldsymbol{\psi}_m^{\rm NF}-\boldsymbol{\psi}_N|} \right]^T, \tag{2}$$

where $\alpha_m = \frac{c}{4\pi f_c |\boldsymbol{\psi}_m^{\rm NF}-\boldsymbol{\psi}_0|}$, c, and f_c denote the free-space path loss, the speed of light, and the carrier frequency, respectively. We note that, as is common in the near-field communication literature [8], [13], only the line of sight (LoS) path is considered for illustrative purposes.

The k-th far-field user receives the following signal:

$$z_k = \mathbf{g}_k^H \mathbf{x} + w_k, \tag{3}$$

where w_k denotes the additive Gaussian noise having the same power as n_m, the conventional beamsteering vector is used to model the far-field user's channel vector, $\mathbf{g}_k$, as follows: [4]

$$\mathbf{g}_k = \alpha_k e^{-j\frac{2\pi}{\lambda}|\boldsymbol{\psi}_k^{\rm FF}-\boldsymbol{\psi}_1|} \times \left[1 \quad e^{-j\frac{2\pi d}{\lambda}\sin\theta_k} \quad \ldots \quad e^{-j\frac{2\pi d}{\lambda}(N-1)\sin\theta_k} \right]^T, \tag{4}$$

and θ_k denotes the conventional angle of departure.

Remark 1: Comparing (2) to (4), one can observe that the near-field channel model is fundamentally different from the far-field one. In particular, the channel vector in (4) is mainly parameterized by the angle of departure, θ_k, but the elements of the vector in (2) depend on the near-field user's specific location.

B. Near-Field Beamforming and NOMA Data Rates

For illustrative purposes, full-digital near-field beamforming based on the zero-forcing principle is adopted in this letter:

$$\mathbf{P} \triangleq \begin{bmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_M\end{bmatrix} = \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{Q}, \tag{5}$$

where $\mathbf{H} = \begin{bmatrix}\mathbf{h}_1 & \cdots & \mathbf{h}_M\end{bmatrix}$, and $\mathbf{Q}$ is an $M \times M$ diagonal matrix to ensure power normalization. In particular, the i-th element on the main diagonal of $\mathbf{Q}$ is given by $[\mathbf{Q}]_{i,i} = \left[\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\right]_{i,i}^{-\frac{1}{2}}$, which ensures that the beamforming vectors are normalized, i.e., $\mathbf{p}_m^H\mathbf{p}_m = 1$, for all $m \in \{1, \cdots, M\}$.

The NOMA principle is applied to ensure that each preconfigured spatial beam is used as a type of bandwidth resource for serving additional far-field users, which means that the signal vector sent by the base station is given by

$$\mathbf{x} = \sum_{m=1}^{M}\mathbf{p}_m\left(\sqrt{P_m}s_m^{\mathrm{NF}} + \sum_{k=1}^{K}f_{m,k}s_k^{\mathrm{FF}}\right), \tag{6}$$

where P_m is the transmit power allocated to the m-th near-field user's signal, $f_{m,k}$ denotes the coefficient assigned to the k-th far-field user on beam $\mathbf{p}_m$, and s_m^{NF} and s_k^{FF} denote the signals for the near-field and far-field users, respectively. If the k-th far-field user uses only a single beam, $f_{m,k}$ can be viewed as a power allocation coefficient. If multiple beams are used, $f_{m,k}$ can be interpreted as a beamforming coefficient.

Therefore, the observation at the m-th near-field user can be written as follows:

$$y_m = \mathbf{h}_m^H\mathbf{p}_m\left(\sqrt{P_m}s_m^{\mathrm{NF}} + \sum_{k=1}^{K}f_{m,k}s_k^{\mathrm{FF}}\right) + n_m. \tag{7}$$

The features of the near-field channels make the design of successive interference cancellation (SIC) different from the far-field case considered in [3]. In particular, the near-field users suffer less path loss than the far-field users, and the $\mathbf{p}_m$'s have been configured for the near-field users' channel vectors. Therefore, the near-field users have better channel conditions than the far-field users, and hence have the capability to carry out SIC. To reduce the system complexity, assume that at most a single far-field user is scheduled on a near-field user's beam, and each of the K far-field users utilize D_x beams, where $KD_x \leq M$. Denote the subset collecting the indices of the beams used by the k-th far-field user by $\mathcal{S}_k$, where $|\mathcal{S}_k| = D_x$.

Assume that the k-th far-field user is active on $\mathbf{p}_m$. On the one hand, this far-field user's signal can be decoded by the m-th near-field user with the following data rate:

$$R_{m,k}^{\mathrm{FF-NF}} = \log\left(1 + \frac{|f_{m,k}|^2 h_m}{\sigma^2 + P_m h_m}\right), \tag{8}$$

where $h_m = |\mathbf{h}_m^H\mathbf{p}_m|^2$. If the first stage of SIC is successful, the near-field user can remove the far-field user's signal and decode its own signal with the following data rate: $R_m^{\mathrm{NF}} = \log\left(1 + \frac{P_m}{\sigma^2}h_m\right)$.

On the other hand, the k-th far-field user directly decodes its own signal with the following data rate:

$$R_k^{\mathrm{FF}} = \log\left(1 + \frac{|\tilde{\mathbf{g}}_k^H\mathbf{f}_k|^2}{\sigma^2 + \sum_{m=1}^{M}P_m g_{m,k} + \sum\limits_{i\neq k}|\tilde{\mathbf{g}}_k^H\mathbf{f}_i|^2}\right),$$

where $\tilde{\mathbf{g}}_k = \mathbf{P}^H\mathbf{g}_k$, $g_{m,k} = |\mathbf{g}_k^H\mathbf{p}_m|^2$, and $\mathbf{f}_k = \begin{bmatrix}f_{1,k} & \cdots f_{M,k}\end{bmatrix}^T$.

The aim of this letter is to maximize the far-field users' sum data rate while guaranteeing the near-field users' QoS requirements, based on the following optimization problem:

$$\max_{P_m, f_{m,k}} \quad \sum_{k=1}^{K}\min\left\{R_k^{\mathrm{FF}}, R_{m,k}^{\mathrm{FF-NF}}, m \in \mathcal{S}_k\right\} \tag{P1a}$$

$$s.t. \quad R_m^{\mathrm{NF}} \geq R, \sum_{k=1}^{K}\mathbf{1}_{f_{m,k}\neq 0} \leq 1, 1 \leq m \leq M \tag{P1b}$$

$$P_m + \sum_{k=1}^{K}|f_{m,k}|^2 \leq P, 1 \leq m \leq M \tag{P1c}$$

$$P_m \geq 0, 1 \leq m \leq M, \tag{P1d}$$

where $\mathbf{1}_{x\neq 0}$ denotes an indicator function, i.e., $\mathbf{1}_{x\neq 0} = 1$ if $x \neq 0$, otherwise $\mathbf{1}_{x\neq 0} = 0$, R denotes the near-field users' target data rate, and P denotes the transmit budget per beam.

Remark 2: We note that, unlike far-fielding beamforming, near-field beamforming is location-specific, as illustrated by the following example. Assume that all the users and the center of the ULA are on the same straight-line, i.e., the users have the same angle of departure. In the far-field case, one user's channel vector is simply a scaled version of another user's, which means that the channel matrix is not invertible and it is reasonable to ask the users to share the same beam via NOMA. However, the near-field beamforming matrix $\mathbf{P} = \mathbf{H}\left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{Q}$ may still exist even for this extreme example. This is because near-field beamforming is tailored to users' specific locations. This unique feature leads to the question whether a near-field user's beam can benefit another user, which motivates problem (P1). In particular, if the solution of problem (P1) offers a significant sum data rate, the feasibility to use the preconfigured beams to support additional users can be demonstrated.

Remark 3: The objective function in (P1a) indicates that the k-th far-field user prefers to use a beam on which both h_m and $g_{m,k}$ are strong. Therefore, to find $\mathcal{S}_k$ and remove the indicator function in (P1a), a simple sub-optimal scheduling scheme can be used first, where the far-field users are successively asked to select the best D_x beams based on the following criterion: $\arg\max\limits_{m}\ \min\left\{\frac{h_m}{\max\{h_1,\cdots,h_M\}}, \frac{g_{m,k}}{\max\{g_{1,k},\cdots,g_{M,k}\}}\right\}$. Here, the channel gains, h_m and $g_{m,k}$, are normalized to ensure that they are in the same order of magnitude. As a result, $f_{m,k} = 0$ for $m \notin \mathcal{S}_k$, and only $f_{m,k}$, $m \in \mathcal{S}_k$, need to be optimized. We note that optimal scheduling is possible by applying integer programming as in [10]. However, this is out of the scope of this paper due to the space limitations.

III. Proposed Resource Allocation Algorithms

A. SCA-based Resource Allocation

In this section, the general case with $K \geq 1$ and $D_x \geq 1$ is considered first. Problem (P1) can be first recast as follows:

$$\max_{P_m, x_k, f_{m,k}} \quad \sum_{k=1}^{K} \log(1+x_k) \tag{P2a}$$

$$s.t. \quad \frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{\sigma^2 + \sum_{m=1}^{M} P_m g_{m,k} + \sum\limits_{i \neq k} |\tilde{\mathbf{g}}_k^H \mathbf{f}_i|^2} \geq x_k \geq 0, \quad 1 \leq k \leq K \tag{P2b}$$

$$\frac{|f_{m,k}|^2 h_m}{\sigma^2 + P_m h_m} \geq x_k, m \in \mathcal{S}_k, 1 \leq k \leq K \tag{P2c}$$

$$f_{m,k} = 0, m \notin \mathcal{S}_k, 1 \leq k \leq K, \tag{P2d}$$

$$P_m \geq \frac{\sigma^2 \epsilon}{h_m}, 1 \leq m \leq M, \tag{P2e}$$

$$\text{(P1c), (P1d)}, \tag{P2f}$$

where $\epsilon = 2^R - 1$, and constraint (P2d) is due to the use of the scheduling scheme discussed in Remark 3. Because (P2b) and (P2c) are decreasing functions of P_m, it is straightforward to show that the optimal solution of P_m is $P_m^* = \min\left\{\frac{\sigma^2 \epsilon}{h_m}, P\right\}$. We note that the main challenges involving problem (P2) are the non-convex constraints in (P2b) and (P2c), which motivates the use of SCA. To facilitate the application of SCA, problem (P2) can be first recast as follows:

$$\max_{x_k, f_{m,k}} \quad \sum_{k=1}^{K} \log(1+x_k) \tag{P3a}$$

$$s.t. \quad \eta_k + \sum_{i \neq k} |\tilde{\mathbf{g}}_k^H \mathbf{f}_i|^2 \leq \frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{x_k}, 1 \leq k \leq K \tag{P3b}$$

$$x_k \mu_m \leq |f_{m,k}|^2, m \in \mathcal{S}_k, 1 \leq k \leq K \tag{P3c}$$

$$\sum_{k=1}^{K} |f_{m,k}|^2 \leq P - P_m^*, 1 \leq m \leq M, x_k \geq 0 \tag{P3d}$$

$$\text{(P2d)}, \tag{P3d}$$

where $\eta_k = \sigma^2 + \sum_{m=1}^{M} P_m^* g_{m,k}$ and $\mu_m = \frac{\sigma^2 + P_m^* h_m}{h_m}, m \in \mathcal{S}_k$. The term $\frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{x_k}$ can be approximated as an affine function via the Taylor expansion. In particular, first express $|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2$ as follows:

$$|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2 = \bar{\mathbf{f}}_k^T \left(\hat{\mathbf{g}}_k \hat{\mathbf{g}}_k^T + \check{\mathbf{g}}_k \check{\mathbf{g}}_k^T\right) \bar{\mathbf{f}}_k, \tag{9}$$

where $(\cdot)^T$ denote the transpose, $\bar{\mathbf{f}}_k = \left[\mathrm{Re}(\mathbf{f}_k)^T \quad \mathrm{Im}(\mathbf{f}_k)^T\right]^T$, $\hat{\mathbf{g}}_k = \left[\mathrm{Re}(\hat{\mathbf{g}}_k)^T \quad \mathrm{Im}(\hat{\mathbf{g}}_k)^T\right]^T$, $\check{\mathbf{g}}_k = \left[-\mathrm{Im}(\hat{\mathbf{g}}_k)^T \quad \mathrm{Re}(\hat{\mathbf{g}}_k)^T\right]^T$. By building the real-valued vector, $\tilde{\mathbf{f}}_k = \left[\mathbf{f}_k^T \quad x_k\right]^T$, and applying the first order Taylor expansion, $\frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{x_k}$ can be approximated as follows:[1]

$$\frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{x_k} \approx \frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k^0|^2}{x_k^0} + \nabla_k^T|_{\tilde{\mathbf{f}}_k = \tilde{\mathbf{f}}_k^0} \left(\tilde{\mathbf{f}}_k - \tilde{\mathbf{f}}_k^0\right), \tag{10}$$

[1] An alternative way is to apply complex-valued Taylor expansion directly, i.e, $\frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{x_k} \approx \frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k^0|^2}{x_k^0} + 2\mathrm{Re}\left(\frac{\partial |\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2 x_k^{-1}}{\partial \hat{\mathbf{f}}_k}\Big|_{\hat{\mathbf{f}}_k = \hat{\mathbf{f}}_k^0}^H \left(\hat{\mathbf{f}}_k - \hat{\mathbf{f}}_k^0\right)\right)$, where $\hat{\mathbf{f}}_k = \left[\mathbf{f}_k^T \quad x_k\right]^T$ and $\hat{\mathbf{f}}_k^0$ denotes the initial value for the Taylor expansion [14]. However, our simulations show that this alternative approach causes the SCA to not converge. This could be due to the fact that $\mathbf{f}_k$ is complex-valued but x_k is not.

where $\tilde{\mathbf{f}}_k^0 = \left[\mathrm{Re}(\mathbf{f}_k^0)^T \quad \mathrm{Im}(\mathbf{f}_k^0)^T \quad x_k^0\right]^T$ denotes the initial value of $\tilde{\mathbf{f}}_k$, and ∇_k is given by

$$\nabla_k = \left[\frac{2\bar{\mathbf{f}}_k^T \left(\hat{\mathbf{g}}_k \hat{\mathbf{g}}_k^T + \check{\mathbf{g}}_k \check{\mathbf{g}}_k^T\right)^T}{x_k} \quad -\frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k|^2}{x_k^2}\right]^T. \tag{11}$$

Constraint (P3c) can be similarly approximated by using an initial value $f_{m,k}^0$.

Therefore, problem (P3) can be approximated as follows:

$$\max_{x_k, f_{m,k}} \quad \sum_{k=1}^{K} \log(1+x_k) \tag{P4a}$$

$$s.t. \quad \eta_k + \sum_{i \neq k} |\tilde{\mathbf{g}}_k^H \mathbf{f}_i|^2 \leq \frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k^0|^2}{x_k^0} + \nabla_k^T|_{\tilde{\mathbf{f}}_k = \tilde{\mathbf{f}}_k^0} \left(\tilde{\mathbf{f}}_k - \tilde{\mathbf{f}}_k^0\right), \quad 1 \leq k \leq K \tag{P4b}$$

$$x_k \mu_m \leq |f_{m,k}^0|^2 + 4\mathrm{Re}\left\{(f_{m,k}^0)^H (f_{m,k} - f_{m,k}^0)\right\}, \quad m \in \mathcal{S}_k, 1 \leq k \leq K \tag{P4c}$$

$$\text{(P2d), (P3d)}, \tag{P4c}$$

which is a convex optimization problem and can be straightforwardly solved by convex solvers.

The implementation of SCA requires that $\tilde{\mathbf{f}}_k^0$ is a feasible solution of problem (P3), and $\tilde{\mathbf{f}}_k^0$ can be obtained as follows. First, by using (P3d), we choose $f_{m,k}^0 = 0$ for $m \notin \mathcal{S}_k$, and $f_{m,k}^0 = (P - P_m^*)$ for $m \in \mathcal{S}_k$, which means that the following choices of x_k^0, $1 \leq k \leq K$, are feasible:

$$x_k^0 = \min\left\{\frac{|\tilde{\mathbf{g}}_k^H \mathbf{f}_k^0|^2}{\eta_k + \sum\limits_{i \neq k} |\tilde{\mathbf{g}}_k^H \mathbf{f}_i^0|^2}, \frac{|f_{m,k}^0|^2}{\mu_m}, m \in \mathcal{S}_k\right\}. \tag{12}$$

Based on $\tilde{\mathbf{f}}_k^0$, SCA can be applied in an iterative manner, i.e., by using $\tilde{\mathbf{f}}_k^0$ and solving problem (P4), a new estimate of $\tilde{\mathbf{f}}_k$ can be generated and used to replace $\tilde{\mathbf{f}}_k^0$ in the next iteration. In general, SCA can only converge to a stationary point, which cannot be guaranteed to be the optimal solution. Motivated by this, the optimal performance of NOMA transmission is studied for the two special cases in the following subsection.

B. Optimal Performance in Two Special Cases

1) Special case with $K = 1$: If there is a single far-field user, i.e., $K = 1$, problem (P1) can be simplified as follows:[2]

$$\max_{x, f_m} \quad x \tag{P5a}$$

$$s.t. \quad R^{\mathrm{FF}} \geq x, x \geq 0, R_m^{\mathrm{NF}} \geq R, 1 \leq m \leq M \tag{P5b}$$

$$R_m^{\mathrm{FF-NF}} \geq x, m \in \mathcal{S}, f_m = 0, m \notin \mathcal{S} \tag{P5c}$$

$$P_m + |f_m|^2 \leq P, P_m \geq 0, 1 \leq m \leq M. \tag{P5d}$$

By following steps similar to those in the previous subsection, problem (P5) can be recast as follows:

$$\max_{y, f_m} \quad y \tag{P6a}$$

$$s.t. \quad |\mathbf{g}^H \mathbf{P} \mathbf{f}|^2 \geq \eta_0 y, y \geq 0 \tag{P6b}$$

$$f_m^2 h_m \geq \eta_m y, m \in \mathcal{S}, f_m = 0, m \notin \mathcal{S} \tag{P6c}$$

$$|f_m|^2 \leq P - P_m^*, 1 \leq m \leq M, \tag{P6d}$$

[2] The subscript, k, is omitted since there is a single far-field user.

Algorithm 1 Branch and Bound Algorithm

1: Set $\bar{\mathcal{S}}_0 = \{\mathcal{B}_0\}$ and tolerance ϵ, $i = 0$, $\beta_0^u = \phi^{\rm up}(\mathcal{B}_0)$, $\beta_0^l = \phi^{\rm lb}(\mathcal{B}_0)$, and $\delta = \beta_0^u - \beta_0^l$
2: **while** $\delta \geq \epsilon$ **do**
3: $i = i + 1$
4: Find $\mathcal{B} \in \bar{\mathcal{S}}_{i-1}$ with the criterion: $\min \phi^{\rm lb}(\mathcal{B})$
5: Split $\mathcal{B}$ along its longest edge into $\mathcal{B}_1$ and $\mathcal{B}_2$
6: Construct $\bar{\mathcal{S}}_i = \{\mathcal{B}_1 \cup \mathcal{B}_2 \cup (\bar{\mathcal{S}}_{i-1} \backslash \mathcal{B})\}$
7: Update the upper and lower bounds $\beta_i^u = \max \phi^{\rm up}(\mathcal{B})$ and $\beta_i^l = \max \phi^{\rm lb}(\mathcal{D})$, $\forall \mathcal{B} \in \bar{\mathcal{S}}_i$.
8: $\delta = \beta_i^u - \beta_i^l$
9: Prune $\mathcal{B}$ with upper bounds smaller than β_i^l.
10: **end**

where $y = 2^x - 1$, $P_m^* = \frac{\sigma^2 \epsilon}{h_m}$, $\eta_0 = \sigma^2 + \sum_{m=1}^M P_m^* g_m$, and $\eta_m = \sigma^2 + P_m^* h_m$. While problem (P6) is not convex, it can be solved by using the following lemma.

Lemma 1. *The optimal value of problem* (P6) *is the same as that of the following optimization problem:*

$$\max_{y, f_m} \quad y \tag{P7a}$$

$$s.t. \quad \left(\sum_{m=1}^M g_m^{\frac{1}{2}} |f_m| \right)^2 \geq \eta_0 y, y \geq 0, \text{(P6c)}, \text{(P6d)}, \tag{P7b}$$

where $g_m = |\mathbf{g}^H \mathbf{p}_m|^2$.

Proof. Denote a feasible solution of problem (P6) by $(y^0, \mathbf{f}^0 \triangleq \begin{bmatrix} f_1^0 & \cdots & f_M^0 \end{bmatrix}^T)$. Constraint (P6b) ensures that $|\mathbf{g}^H \mathbf{P} \mathbf{f}^0|^2 \geq \eta_0 y^0$. Because $\sum_{m=1}^M g_m^{\frac{1}{2}} |f_m^0| \geq |\mathbf{g}^H \mathbf{P} \mathbf{f}^0|$, $\left(\sum_{m=1}^M g_m^{\frac{1}{2}} |f_m^0| \right)^2 \geq \eta_0 y^0$, which means that any feasible solution of problem (P6) is also feasible for problem (P7). In other words, the feasible set of problem (P6) is a subset of that of problem (P7), and hence the optimal value of problem (P7) is no less than that of problem of (P6). We also note that the optimal solution of problem (P7) leads to a feasible solution of problem (P6). For example, assume that $(y^*, \mathbf{f}^* \triangleq \begin{bmatrix} f_1^* & \cdots & f_M^* \end{bmatrix}^T)$ is the optimal solution of problem (P7). $(y^*, \tilde{\mathbf{f}}^* \triangleq \begin{bmatrix} |f_1^*| e^{-j\bar{\theta}_1} & \cdots & |f_M^*| e^{-j\bar{\theta}_M} \end{bmatrix}^T)$ must be feasible to problem (P6), where $\bar{\theta}_m$ is the argument of the complex-valued number $\mathbf{g}^H \mathbf{p}_m$. By using the fact that the optimal value of problem (P7) is no less than that of problem of (P6), y^* must be the optimal value of problem (P6) as well. Therefore, problems (P6) and (P7) have the same optimal value, and the proof is complete. □

By using Lemma 1, the optimal performance of NOMA transmission in the special case with $K = 1$ can be obtained by defining $z_m = |f_m|^2$ and transferring problem (P7) into the following equivalent convex form:

$$\max_{y, z_m} \quad y \tag{P8a}$$

$$s.t. \quad \left(\sum_{m=1}^M g_m^{\frac{1}{2}} z_m^{\frac{1}{2}} \right)^2 \geq \eta_0 y, y \geq 0, z_m h_m \geq \eta_m y, m \in \mathcal{S}$$
$$0 \leq z_m \leq P - P_m^*, 1 \leq m \leq M, z_m = 0, m \notin \mathcal{S}.$$

2) Special case with $D_x = 1$: If each far-field user uses a single beam, problem (P1) can be simplified as follows:

$$\max_{x_k, f_k} \quad \sum_{k=1}^K \log(1 + x_k) \tag{P9a}$$

$$s.t. \quad \eta_k + \sum_{i \neq k} |\tilde{g}_k|^2 |f_i|^2 \leq \frac{|\tilde{g}_k|^2 |f_k|^2}{x_k}, 1 \leq k \leq K \tag{P9b}$$

$$|f_k|^2 \geq x_k \frac{\sigma^2 + P_{m_k}^* h_{m_k}}{h_{m_k}}, 1 \leq k \leq K \tag{P9c}$$

$$|f_k|^2 \leq P - P_{m_k}^*, 1 \leq m \leq M, \tag{P9d}$$

where m_k denotes the index of the beam used by the k-th far-field user, and $f_{m,k}$ is simplified to f_k for this special case.

For this special case, we note that the optimization problem is with respect to $|f_k|^2$, i.e., the phase of f_k has no impact. Therefore, $|f_k|^2$ can be treated as the optimization variable, and problem (P9) is similar to conventional power allocation in interference channels, where the BB algorithm can be used to obtain the optimal solution [9], [10].

Due to the space limitations, the principle of the BB algorithm is described only briefly in the following. As shown in Algorithm 1, the initilization of the algorithm builds an initial box, denoted by $\mathcal{B}_0$, by using an upper bound on x_k as follows:

$$x_k \leq \min \left\{ \frac{|\tilde{g}_k|^2 (P - P_m^*)}{\eta_k}, \frac{P - P_m^*}{\mu_m}, m \in \mathcal{S}_k \right\}. \tag{13}$$

In each iteration of the BB algorithm, the key step is to calculate the upper and lower bounds for a box, $\mathcal{B}$, which are denoted by $\phi^{\rm up}(\mathcal{B})$ and $\phi^{\rm lb}(\mathcal{B})$, respectively. Further denote the minimum and maximum vertices of $\mathcal{B}$ by $\mathbf{x}_{\max}$ and $\mathbf{x}_{\min}$, respectively. $\phi^{\rm up}(\mathcal{B}) = \sum_{k=1}^K \log(1 + x_{k,\max})$ and $\phi^{\rm lb}(\mathcal{B}) \sum_{k=1}^K \log(1 + x_{k,\min})$, if $\mathbf{x}_{\min}$ is a solution of the following feasibility optimization problem:

$$\max_{f_k} \quad 1 \tag{P10a}$$

$$s.t. \quad \eta_k + \sum_{i \neq k} |\tilde{g}_k|^2 |f_i|^2 \leq \frac{|\tilde{g}_k|^2 |f_k|^2}{x_{k,\min}}, 1 \leq k \leq K \tag{P10b}$$

$$|f_k|^2 \geq x_{k,\min} \frac{\sigma^2 + P_{m_k}^* h_{m_k}}{h_{m_k}}, 1 \leq k \leq K \tag{P10c}$$

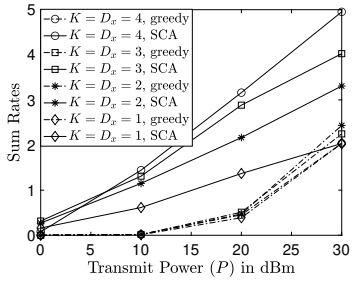
$$|f_k|^2 \leq P - P_{m_k}^*, 1 \leq m \leq M, \tag{P10d}$$

where $x_{k,\max}$ and $x_{k,\min}$ are the k-th elements of $\mathbf{x}_{\max}$ and $\mathbf{x}_{\min}$, respectively. If $\mathbf{x}_{\min}$ is not feasible, the upper and lower bounds are set as 0. The details for implementing the BB algorithm can be found in [10].

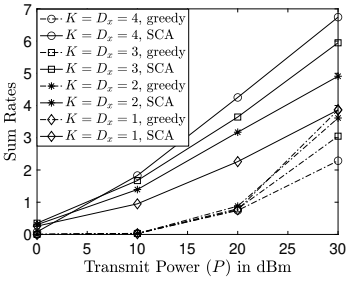
IV. Simulation Results

In this section, simulation results are presented to evaluate the performance of the proposed NOMA scheme. For all simulations we used, $f_c = 28$ GHz, $d = \frac{\lambda}{2}$, $R = 0.1$ bits per channel use, $\sigma^2 = -80$ dBm, and $M = 36$ [8]. The ULA is placed on the vertical coordinate axis and $\boldsymbol{\psi}_0 = (0, 0)$.

In Fig. 1, the performance of NOMA transmission is evaluated with randomly located users. In particular, the near-field users are uniformly located inside of a half-ring with its inner and outer radii being 5 m and $d_R(64)$ m, respectively. We

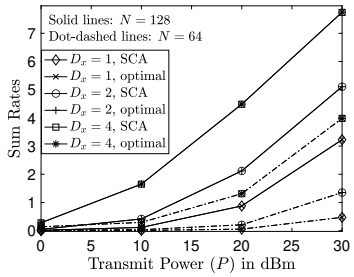


(a) $N = 64$

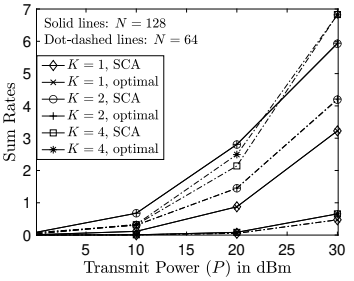


(b) $N = 128$

Fig. 1. Far-field users' sum data rates achieved by NOMA with randomly located users. The greedy benchmarking scheme is based on (12).



(a) $K = 1$



(b) $D_x = 1$

Fig. 2. Deterministic studies for the optimal performance achieved by NOMA transmission.

use 5 m for the inner radius to ensure that the ring is outside of the reactive near field. The far-field users are uniformly located inside of a half-ring with its inner and outer radii being $d_R(128)$ m and $(d_R(128) + 10)$ m, respectively. A greedy resource allocation scheme based on (12) is used as a benchmarking scheme in the figure. As can be seen from Fig. 1, the use of NOMA ensures that the spatial beams preconfigured for the near-field users are efficiently utilized to support additional far-field users. In addition, Fig. 1 shows that the developed SCA algorithm yields a significant performance gain over the benchmark scheme. Comparing Fig. 1(a) to Fig. 1(b), one can also observe that the use of more antennas at the base station can further improve the performance of NOMA, which indicates the importance of NOMA for massive MIMO.

As shown in Sections III-B1 and III-B2, the optimal performance of NOMA transmission can be obtained for the two special cases, which are studied in Fig. 2. The following deterministic scenario is focused on in order to avoid Monte Carlo simulations due to the high complexity of the BB algorithm. On the one hand, assume that the $M = 36$ near-field users are equally spaced with $\frac{10}{\sqrt{M}}$ m distance, and located within a square with its center located at $(0, 9)$ m. On the other hand, assume that the far-field users are also equally spaced and located on a half-circle with radius 90 meters. Fig. 2(a) focuses on the scenario, where a single far-field user is scheduled. The figure shows that the far-field user's performance can be improved by using more beams. Fig. 2(b) focuses on the scenario, where each far-field user uses a single beam. For the scenario of $N = 128$, the sum rate for the case of $K = 2$ is significantly larger than those for $K = 1$ and $K = 4$, which indicates that increasing K does not always improve the sum rate, and there is an optimal choice of K for sum-rate maximization. Fig. 2 also shows that SCA provides a reasonable estimate for the optimal performance.

V. Conclusions

This letter considered a legacy network, where spatial beams have been preconfigured for legacy near-field users, and showed that via NOMA, additional far-field users can be efficiently served by using these preconfigured beams.

References

[1] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple Access Techniques for 5G Wireless Networks and Beyond*. Springer International Publishing, 2019.

[2] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 155–169, Jan. 2023.

[3] Z. Ding, "NOMA beamforming in SDMA networks: Riding on existing beams or forming new ones?" *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 868–871, Apr. 2022.

[4] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[5] Y. Zou, W. Rave, and G. Fettweis, "Analog beamsteering for flexible hybrid beamforming design in mmWave communications," in *Proc. European Conference on Networks and Communications (EuCNC)*, Athens, Greece, Jun. 2016.

[6] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality - what is next?: Five promising research directions for antenna arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, Oct. 2019.

[7] J. Zhu, Z. Wan, L. Dai, M. Debbah, and H. V. Poor, "Electromagnetic information theory: Fundamentals, modeling, applications, and open problems," Available on-line at arXiv:2209.09562, 2022.

[8] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, M. F. Imani, and Y. C. Eldar, "Beam focusing for near-field multiuser MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7476–7490, Sept. 2022.

[9] P. C. Weeraddana, M. Codreanu, M. Latva-aho, and A. Ephremides, "Weighted sum-rate maximization for a set of interfering links via branch and bound," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3977–3996, Aug. 2011.

[10] Z. Ding and H. V. Poor, "Joint beam management and power allocation in THz-NOMA Networks," *IEEE Trans. Commun.*, to appear in 2023.

[11] K. Chen, C. Qi, and C.-X. Wang, "Two-stage hybrid-field beam training for ultra-massive MIMO systems," in *2022 IEEE/CIC Int. Conf. on Commun. in China (ICCC)*, Foshan, China, Sept. 2022.

[12] X. Zhang, H. Zhang, and Y. C. Eldar, "Near-field sparse channel representation and estimation in 6G wireless communications," Available on-line at arXiv:2212.13527, 2022.

[13] X. Zhang, H. Zhang, , and Y. C. Eldar, "Near-field sparse channel representation and estimation in 6G wireless communications," Available on-line at arXiv:2212.13527, 2022.

[14] L. Cantos, M. Awais, and Y. H. Kim, "Max-min rate optimization for uplink IRS-NOMA with receive beamforming," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2512–2516, Dec. 2022.