

# Massive MIMO-NOMA Networks with Imperfect SIC: Design and Fairness Enhancement

Arthur S. de Sena, *Student Member, IEEE*, F. Rafael M. Lima, *Member, IEEE*,  
Daniel B. da Costa, *Senior Member, IEEE*, Zhiguo Ding, *Senior Member, IEEE*,  
Pedro H. J. Nardelli, *Senior Member, IEEE*, Ugo S. Dias, *Senior Member, IEEE*,  
Constantinos B. Papadias, *Fellow, IEEE*

## Abstract

This paper addresses multi-user multi-cluster massive multiple-input-multiple-output (MIMO) systems with non-orthogonal multiple access (NOMA). Assuming the downlink mode, and taking into consideration the impact of imperfect successive interference cancellation (SIC), an in-depth analytical analysis is carried out, in which closed-form expressions for the outage probability and ergodic rates are derived. Subsequently, the power allocation coefficients of users within each sub-group are optimized to maximize fairness. The considered power optimization is simplified to a convex problem, which makes it possible to obtain the optimal solution via Karush-Kuhn-Tucker (KKT) conditions. Based on the achieved solution, we propose an iterative algorithm to provide fairness also among different sub-groups. Simulation results alongside with insightful discussions are provided to investigate the impact of imperfect SIC and demonstrate the fairness superiority of the proposed dynamic power allocation policies. For example, our results show that if the residual error propagation levels are high,

A. S. de Sena and P. H. J. Nardelli are with the Lappeenranta University of Technology, Finland (email: arthurssena@ieee.org, pedro.nardelli@lut.fi).

F. R. M. Lima and D. B. da Costa are with the Federal University of Ceará, Brazil (email: rafaelm@gtel.ufc.br, danielbcosta@ieee.org).

Z. Ding is with the University of Manchester, UK (email: zhiguo.ding@manchester.ac.uk).

U. S. Dias is with the University of Brasília, Brazil (email: ugodias@ieee.org).

C. B. Papadias is with the Athens Information Technology, Greece (email: papadias@ait.edu.gr).

Part of this paper has been submitted in IEEE International Conference on Communications (ICC 2020), Dublin, Ireland.

the employment of orthogonal multiple access (OMA) is always preferable than NOMA. It is also shown that the proposed power allocation outperforms conventional massive MIMO-NOMA setups operating with fixed power allocation strategies in terms of outage probability.

### Index Terms

Fairness maximization, imperfect SIC, massive MIMO, NOMA.

## I. INTRODUCTION

The 5th generation of wireless networks (5G) is expected to incorporate various technologies and to support a variety of services with diverse and demanding requisites, going from massive connectivity and ultra-low latency to improved user fairness [1]. Power-domain non-orthogonal multiple access (NOMA) is being credited as one of the key components of 5G and beyond since it allows multiple users to be served in parallel within the same resource block, i.e., the same frequency and time slot. The underlying concept of NOMA consists of superposing the data symbols of different users in the power domain at the base station (BS) and employing successive interference cancellation (SIC) at the receivers. This featured technique has the potential to reduce system latency and to provide remarkable connectivity gains. Massive multiple-input-multiple-output (MIMO) is another important technology confirmed to be part of 5G [2]. In particular, by employing a very large number of antennas, massive MIMO technology also provides massive connectivity capabilities and reduction in latency to the network, but in contrast to NOMA, it exploits the space domain to multiplex the different users.

If the NOMA technique is applied to massive MIMO, the achievable spectral and connectivity improvements are shown to be even greater [3], [4]. However, if the transmission power is not well allocated within the MIMO-NOMA network, the performance of some users can be severely compromised. For instance, the adoption of fixed power allocation policies in NOMA can be very beneficial to users with good channel conditions, however, it can be extremely disadvantageous to users that suffer from strong channel attenuation [5]. Therefore, more sophisticated and adaptive power allocation strategies are important and necessary to guarantee fairness in networks employing NOMA. In addition, the majority of existing works make the strong assumption that SIC can be carried out perfectly, which is idealistic and difficult to hold in practice. In real-world deployments, various impairments such as fast varying channels, atmospheric absorption, strong

channel correlation, and hardware issues, can degrade the signal reception and introduce errors during the detection of transmitted symbols [6], [7]. As a result, since the recovery of each symbol with SIC depends on previous decodings, errors will inevitably propagate and impact the system performance. This makes SIC residual error propagation an important parameter that must be considered while designing realistic massive MIMO-NOMA systems.

#### A. Related Works

A few NOMA-related works have considered the impact of imperfect SIC. For instance, in [6], a massive MIMO-NOMA system with non-orthogonal channel estimation and SIC error propagation was investigated. The work considered a single-cell downlink scenario, where a single multi-antenna BS communicates with multiple single-antenna users. In particular, the authors derived a lower bound expression for the spectral efficiency and developed iterative optimization algorithms for maximizing the weighted sum spectral efficiency. The provided numerical results validated the analytical approximation, although the gap between the bounds and simulation curves were not very tight. The work in [8] proposed a sub-optimum iterative algorithm for maximizing the sum-rate of a downlink MIMO-NOMA network. By equipping both BS and users with two antennas, the paper provided results for two very specific scenarios, in which a high and a low value of SIC residual error was considered. However, intermediate error levels were not investigated. The impact of imperfect SIC in the uplink of MIMO-NOMA systems employing the concept of virtualized wireless networks (VWN) was studied in [9], in which algorithms based on successive convex approximation and complementary geometric programming were proposed for power and sub-carrier allocation. A massive MIMO-NOMA system with distributed antenna arrays was investigated in [10]. In this contribution, a closed-form expression for the ergodic sum-rate was derived. In [11], by taking into consideration SIC error propagation and in-phase and quadrature-phase imbalance, the performance of a full-duplex relaying system was investigated. The study of  $\alpha$ - $\mu$  fading channels in cooperative NOMA networks with hardware impairments was considered in [12], and the work in [13] addressed the application of deep learning techniques to MIMO-NOMA systems with imperfect SIC decoding.

Some contributions have addressed fairness in NOMA systems, although none of them have considered the effects of imperfect SIC. For example, by assuming that SIC can be carried out

perfectly, the work in [14] investigated the impact of power allocation on the fairness of a simple system where a single-antenna BS serves multiple single-antenna users. The authors developed low-complexity bisection-based iterative algorithms to optimally solve the optimization problem. In [15], the fairness of user clustering in a multi-user MIMO-NOMA setup was considered. Bisection algorithms were also adopted to optimize the power of users within each cluster. In addition, three sub-optimum clustering algorithms have been proposed. A fair NOMA protocol, in which the user capacity is always at least equal to the capacity achieved with orthogonal multiple access (OMA), was proposed in [16]. The referred work pairs near and cell-edge single-antenna users to form NOMA groups, based on which the power allocation coefficients are determined. The outage probability was also investigated. The authors of [17] and [18] proposed user clustering algorithms based on proportional fairness to balance between throughput and fairness. [17] also presented an optimal power allocation for maximizing the system sum-rate such that the rate of weak users is guaranteed to be equal to that achieved with OMA. In [5], the authors developed dynamic resource allocation policies, which are optimally obtained via Lagrangian dual decomposition. The millimeter-wave MIMO-NOMA case was addressed in [19], in which spatial sparsity was exploited to propose sub-optimum power allocation solutions.

### *B. Motivation and Contributions*

Even though there are numerous contributions showing that massive MIMO-NOMA systems can provide remarkable spectral gains and outperform the massive MIMO-OMA counterpart, the majority of these works do not consider the impact of imperfect SIC decoding. In addition, to the best of the authors' knowledge, only a very limited number of works investigates fairness in massive MIMO-NOMA networks, and none of them have considered SIC error propagation. Given the aforementioned research gap, this paper aims to design, analyze, and optimize the performance of a massive MIMO-NOMA network under the impact of residual error propagation from imperfect SIC. More details and the main original contributions provided in this work are summarized as follows.

- Inspired by the works in [3], [4], and assuming that the users are confined within multiple clusters of scatterers, we employ at the BS a two-stage precoder. Specifically, the first-stage precoder, which is intended to eliminate inter-cluster interference, is designed based only

on the slowly varying covariance matrices of interfering clusters. By its turn, the second-stage precoder is responsible for directing the superposed symbols to the corresponding NOMA sub-groups, where each sub-group is formed by two users so that the computational complexity of SIC is reduced. This strategy provides attractive advantages to massive MIMO-NOMA setups, such as less processing overload and reduced feedback overhead.

- Assuming first a fixed power allocation policy, a novel analytical framework for the proposed massive MIMO-NOMA network is developed. In particular, by considering the impact of residual error from imperfect SIC, we derive the system signal-to-interference-plus-noise ratio (SINR) expression and carry out a statistical characterization of the effective channel gains. Then, based on this initial analysis, exact closed-form expressions for the outage probability and for the users' ergodic rates are derived, whose accuracies are validated through numerical and simulation examples. The obtained analytical results provide a practical alternative for designing massive MIMO-NOMA systems with imperfect SIC decoding.
- Next, we develop a more sophisticated dynamic power allocation that maximizes the achievable rates of users with worst channel conditions within each NOMA sub-group. More specifically, the optimization problem is formulated to guarantee that weaker users never experience a rate less than what is achieved by the stronger ones, in which the optimal solution is obtained via Karush-Kuhn-Tucker (KKT) conditions. Then, to balance the data rates also among different sub-groups, we propose an iterative algorithm that extends the fairness concept to all users within the spatial clusters so that all terminals can reach identical performance levels, i.e., maximum fairness is provided.
- Simulation results alongside with insightful discussions are provided to investigate the impact of imperfect SIC and demonstrate the fairness superiority of the proposed dynamic power allocation policies. For example, our results show that if the residual error propagation levels are high, the employment of OMA is always preferable than NOMA. It is also shown that the proposed power allocation outperforms conventional massive MIMO-NOMA setups operating with fixed power allocation strategies in terms of outage probability.

**Notation and Special Functions:** Bold-faced lower-case letters denotes vectors and upper-case represent matrices. The  $i$ th element of a vector  $\mathbf{a}$  is denoted by  $[\mathbf{a}]_i$  and the  $(i,j)$  entry of a matrix  $\mathbf{A}$  by  $[\mathbf{A}]_{ij}$ . The Hermitian transposition of  $\mathbf{A}$  is represented by  $\mathbf{A}^H$  and the trace by  $\text{tr}\{\mathbf{A}\}$ . In

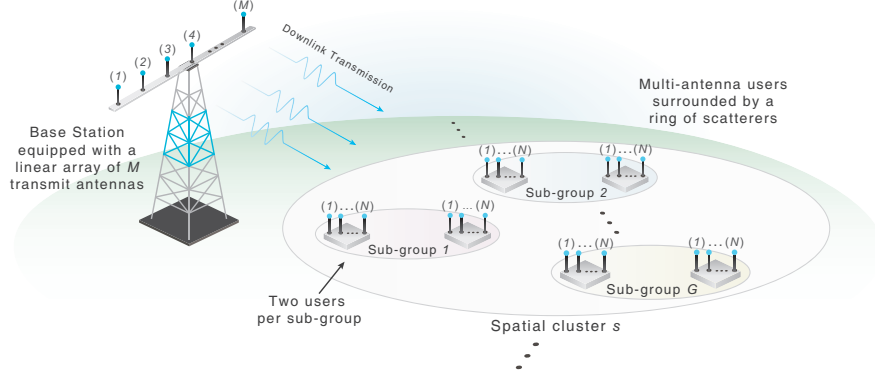


Fig. 1: System model. Users within each spatial cluster are organized into multiple sub-groups.

addition,  $\mathbf{0}_{M \times N}$  denotes the  $M \times N$  matrix with all zero entries,  $\mathbb{E}[\cdot]$  denotes expectation,  $\Gamma(\cdot)$  is the Gamma function [20, eq. (8.310.1)],  $\gamma(\cdot, \cdot)$  is the lower incomplete Gamma function [20, eq. (8.350.1)], and  $\text{Ei}(\cdot)$  corresponds to the exponential integral [20, eq. (8.211.1)].

## II. SYSTEM MODEL

We consider a single-cell scenario where one elevated BS is communicating in downlink mode with  $L$  multi-antenna users. The BS is equipped with a uniform linear array of  $M$  transmit antenna elements, which are separated by half of the wavelength, i.e.,  $\lambda/2$ . Moreover, each user is equipped with  $N$  receive antennas, in which we assume that  $M$  is much greater than  $N$ , i.e.,  $M \gg N$ , which characterizes a typical massive MIMO setup. The users are considered to be uniformly distributed within  $S$  spatial clusters of scatterers, modeled by the one-ring scattering model [21]. Within each cluster, the BS subdivides the users into  $G$  smaller sub-groups, each one containing 2 users, as illustrated in Fig. 1. Then, power-domain NOMA is employed within each sub-group. Given the described scenario, and applying the Karhunen-Loeve transform [21], the channel matrix for the  $k$ th user in the  $g$ th sub-group in the  $s$ th cluster, can be expressed by

$$\mathbf{H}_{sgk} = \sqrt{\Phi d_{sg}^{-\eta}} \mathbf{U}_s \mathbf{\Lambda}_s^{\frac{1}{2}} \mathbf{G}_{sgk} \in \mathbb{C}^{M \times N}, \quad (1)$$

which has covariance matrix given by  $\mathbb{E}\{\mathbf{H}_{sgk} \mathbf{H}_{sgk}^H\} = \Phi d_{sg}^{-\eta} \mathbf{R}_s \in \mathbb{C}^{M \times M}$ , with rank denoted by  $r_s$ .  $\mathbf{\Lambda}_s \in \mathbb{R}_{>0}^{r_s^* \times r_s^*}$  represents a diagonal matrix formed by  $r_s^*$  dominant eigenvalues of  $\mathbf{R}_s$ , with  $r_s^* \leq r_s$ .  $\mathbf{U}_s \in \mathbb{C}^{M \times r_s^*}$  is a matrix of eigenvectors corresponding to the dominant eigenvalues of  $\mathbf{R}_s$ , and  $\mathbf{G}_{sgk} \in \mathbb{C}^{r_s^* \times N}$  is the fast varying channel matrix, which has complex Gaussian

distributed entries with zero-mean and unit-variance.  $d_{sg}$  is the distance of the  $g$ th sub-group from the BS,  $\eta$  is the path-loss exponent, and  $\Phi$  is a gain parameter that is adjusted based on the desired performance of the receivers [22]. Moreover, all users confined within the  $s$ th cluster are assumed to share identical covariance matrices  $\mathbf{R}_s$ , whose entries can be generated by [21]

$$[\mathbf{R}_s]_{mm'} = \frac{1}{2\delta_s} \int_{-\delta_s}^{\delta_s} e^{-j\frac{2\pi}{\lambda}[\cos(\phi+\varphi_s), \sin(\phi+\varphi_s)](\mathbf{a}_m - \mathbf{a}_{m'})} d\phi, \quad (2)$$

where  $\delta_s$  and  $\varphi_s$  are, respectively, the angular spread and the azimuth angle of the  $s$ th spatial cluster,  $\phi$  corresponds to the angle of arrival of incident planar waves at the BS, and  $\mathbf{a}_m, \mathbf{a}_{m'}$  are the Cartesian coordinates of the antenna elements  $m$  and  $m'$ , for  $1 \leq m, m' \leq M$ .

At the BS, the symbols for users within each sub-group are superposed and transmitted through the wireless channels. Then, the  $k$ th user in the  $g$ th sub-group receives the following signal

$$\mathbf{y}_{sgk} = \mathbf{H}_{sgk}^H \sum_{n=1}^S \mathbf{B}_n \sum_{i=1}^G \mathbf{v}_{ni} \sum_{j=1}^2 \sqrt{\alpha_{nij}} x_{nij} + \mathbf{n}_{sgk}, \quad (3)$$

where  $\mathbf{n}_{gk} \in \mathbb{C}^{N \times 1}$  is a noise vector with entries following the complex Gaussian distribution with zero-mean and variance  $\sigma_n^2$ . The variable  $\alpha_{nij}$  denotes the power coefficient, and  $x_{nij}$  is the symbol intended for the  $j$ th user in the  $i$ th sub-group at the  $n$ th cluster.  $\mathbf{B}_n \in \mathbb{C}^{M \times V}$  is the beamforming matrix responsible for eliminating the interference from other clusters, where  $V$  is a parameter that defines the number of parallel effective transmissions, and  $\mathbf{v}_{ni} \in \mathbb{C}^{V \times 1}$  is the precoding vector designed to assign the superposed symbols to the corresponding sub-groups.

### A. Beamforming Design

The beamforming matrix  $\mathbf{B}_s$  is designed to focus the signal transmission to a desired spatial cluster  $s$ , such that everywhere else outside the area of interest the propagation is nulled out. To achieve this spatial directivity,  $\mathbf{B}_s$  is constructed based on the null space spanned by the nonzero eigenmodes of the covariance matrices of interfering clusters [21]. To this end, we define  $\mathbf{U}_s^- = [\mathbf{U}_1, \dots, \mathbf{U}_{s-1}, \mathbf{U}_{s+1}, \dots, \mathbf{U}_S]$  and denote the last  $M - (S - 1)r_s^*$  left eigenvectors of  $\mathbf{U}_s^-$  by  $\mathbf{E}_s^0 \in \mathbb{C}^{M \times M - (S-1)r_s^*}$ , which corresponds to its vanishing eigenmodes. Given that  $(\mathbf{E}_s^0)^H \mathbf{U}_s^- = 0$ , the matrix  $\tilde{\mathbf{H}}_{sgk} = (\mathbf{E}_s^0)^H \mathbf{U}_s \mathbf{\Lambda}_s^{\frac{1}{2}} \mathbf{G}_{sgk}$  is orthogonal to the  $r_s^*$  dominant eigenmodes of interfering clusters, and it has covariance matrix given by  $\tilde{\mathbf{R}}_s = \tilde{\mathbf{H}}_{sgk} \tilde{\mathbf{H}}_{sgk}^H = (\mathbf{E}_s^0)^H \mathbf{R}_s \mathbf{E}_s^0 = \mathbf{F}_s \mathbf{R}_s \mathbf{F}_s^H$ ,

where  $\mathbf{F}_s$  represents the left eigenvectors of  $\tilde{\mathbf{R}}_s$ . Then, by defining the first  $V$  eigenvectors of  $\mathbf{F}_s$  by  $\mathbf{F}_s^{(1)} \in \mathbb{C}^{M-(S-1)r_s^* \times V}$ , the beamforming matrix  $\mathbf{B}_s$  is finally obtained, as follows

$$\mathbf{B}_s = \mathbf{E}_s^0 \mathbf{F}_s^{(1)} \in \mathbb{C}^{M \times V}, \quad (4)$$

where  $S \leq V \leq (M - (S - 1)r_s^*)$  and  $V \leq r_s^* \leq r_s$ .

More clarifications for the precoding vector  $\mathbf{v}_{sg}$  are now provided. We design  $\mathbf{v}_{sg}$  to assign the superimposed symbols to each corresponding NOMA sub-group, such that it should not introduce additional power, i.e.,  $\|\mathbf{v}_{sg}\|^2 = 1$ . This can be accomplished by defining  $\mathbf{v}_{sg}$  as

$$\mathbf{v}_{sg} = [\mathbf{0}_{1 \times (g-1)}, 1, \mathbf{0}_{1 \times (V-g)}]^T, \quad \forall g = 1, \dots, G. \quad (5)$$

Note that the above design associates the  $g$ th effective data stream to the  $g$ th NOMA sub-group, and it does not modify by any form the data messages. Besides, since  $\mathbf{B}_s$  only addresses the inter-cluster interference, the users still need to employ some reception technique for canceling the remaining intra-cluster interference. Details for signal reception are provided next.

### B. Signal Reception

For simplicity, from this point forward, we drop the subscript corresponding to the spatial cluster, e.g., we represent  $\mathbf{y}_{sgk}$  as  $\mathbf{y}_{gk}$ . Accordingly, assuming that the beamforming matrix  $\mathbf{B}_s$  successfully suppresses all inter-cluster interferences, the superposed signal observed by the  $k$ th user,  $k \in \{1, 2\}$ , in the  $g$ th sub-group can be rewritten as

$$\mathbf{y}_{gk} = \mathbf{H}_{gk}^H \mathbf{B} \sum_{i=1}^G \mathbf{v}_i \sum_{j=1}^2 \sqrt{\alpha_{ij}} x_{ij} + \mathbf{n}_{gk}. \quad (6)$$

To eliminate the remaining inter-group interference, the users employ a zero-forcing receiver. Therefore, the detection matrix can be defined by  $\mathbf{H}_{gk}^\dagger = [(\mathbf{H}_{gk}^H \mathbf{B})^H \mathbf{H}_{gk}^H \mathbf{B}]^{-1} (\mathbf{H}_{gk}^H \mathbf{B})^H$ , which corresponds to the pseudo-inverse of the virtual channel. After the received signal has been filtered through  $\mathbf{H}_{gk}^\dagger$ , the  $k$ th user in the  $g$ th group achieves the following data vector

$$\hat{\mathbf{x}}_{gk} = \begin{bmatrix} \sqrt{\alpha_{11}} x_{11} + \sqrt{\alpha_{12}} x_{12} \\ \vdots \\ \sqrt{\alpha_{G1}} x_{G1} + \sqrt{\alpha_{G2}} x_{G2} \end{bmatrix} + \mathbf{H}_{gk}^\dagger \mathbf{n}_{gk}. \quad (7)$$



Note that the zero-forcing receiver has decoupled the signal in (6) into  $G$  parallel symbols, each one belonging to a different sub-group. This enables the users within the  $g$ th NOMA sub-group to apply SIC to their corresponding superposed symbol, i.e., the  $g$ th element of  $\hat{\mathbf{x}}_{gk}$ .

### III. PERFORMANCE ANALYSIS FOR FIXED POWER ALLOCATION

In this section, the performance of the proposed massive MIMO-NOMA system operating under fixed power allocation policy is investigated. Specifically, by considering the impact of residual error from imperfect SIC decoding, we derive the SINR experienced by the users and identify the statistical distributions of the effective channel gains, based on which closed-form analytical expressions for the outage probability and for the users' ergodic rates are obtained.

#### A. SINR Analysis

In our design, the users within each sub-group are organized by the BS in ascending order based on their effective channel gains, that is, the first user has the lowest gain and the second user the highest one. Following the SIC protocol, the weak user retrieves its data symbol directly from (7) and treats the message intended for the second user as interference, so no further processing is required. On the other hand, the second user, which has the best channel condition, first decodes the message intended for the first user and, subsequently, recovers its own data symbol. Ideally, the strong user can recover its information without interference, but, as previously discussed, this is difficult to happen in practice. In real deployments, due to many impairments, the strong users may still suffer from residual interference. On these grounds, we model the effects of imperfect SIC in the proposed design. More specifically, the  $k$ th user in the  $g$ th sub-group will recover the following symbol

$$\hat{x}_{gk} = \begin{cases} \sqrt{\alpha_{g1}}x_{g1} + \sqrt{\alpha_{g2}}x_{g2} + [\mathbf{H}_{g1}^\dagger \mathbf{n}_{g1}]_g, & \text{if } k = 1, \\ \sqrt{\alpha_{g2}}x_{g2} + \sqrt{\mu\alpha_{g1}}x_{g1} + [\mathbf{H}_{g2}^\dagger \mathbf{n}_{g2}]_g, & \text{if } k = 2. \end{cases} \quad (8)$$

$\uparrow$  symbol of interest       $\uparrow$  interference       $\uparrow$  noise  
 $\uparrow$  symbol of interest       $\uparrow$  interference       $\uparrow$  noise

in which  $\mu \in [0, 1]$  is the error propagation factor that models the impact of imperfect SIC, where  $\mu = 1$  represents the scenario of maximum interference, and  $\mu = 0$  corresponds to the ideal case of perfect SIC. Considering the signal model in (8), the SINR achieved by each user during each NOMA decoding is defined in the following Lemma.

*Lemma I:* Supposing that the users within each sub-group are sorted out in increasing order based on their effective channel gains, the SINR achieved at the current  $k$ th user,  $1 \leq k \leq 2$ , when decoding the symbol that belongs to the  $i$ th user,  $1 \leq i \leq 2$ , is obtained by

$$\gamma_{gk}^i = \frac{\rho \varrho_{gk} \alpha_{gi}}{\rho \varrho_{gk} \alpha_{gi}^* + 1}, \quad \text{for } 1 \leq i \leq k \leq 2, \quad (9)$$

where  $\varrho_{gk} = \frac{1}{[\mathbf{H}_{gk}^\dagger \mathbf{H}_{gk}^{\dagger H}]_{gg}}$  is the effective channel gain,  $\rho = 1/\sigma_n^2$  denotes the signal-to-noise ratio (SNR), and  $\alpha_{gi}^*$  corresponds to the interference power, which is given by

$$\alpha_{gi}^* = \begin{cases} \alpha_{g2}, & \text{for } i = 1, \\ \mu \alpha_{g1}, & \text{for } i = k = 2. \end{cases} \quad (10)$$

*Proof:* From (8), it is straightforward to see that the current  $k$ th user,  $1 \leq k \leq 2$ , decodes the first message, i.e., the message intended to the first user, with the following SINR

$$\gamma_{gk}^1 = \frac{\mathbb{E}\{|\sqrt{\alpha_{g1}}x_{g1}|^2\}}{\mathbb{E}\{|\sqrt{\alpha_{g2}}x_{g2}|^2\} + \mathbb{E}\{|\mathbf{H}_{gk}^\dagger \mathbf{n}_{gk}|_g|^2\}} = \frac{\frac{1}{[\mathbf{H}_{gk}^\dagger \mathbf{H}_{gk}^{\dagger H}]_{gg}} \alpha_{g1}}{\frac{1}{[\mathbf{H}_{gk}^\dagger \mathbf{H}_{gk}^{\dagger H}]_{gg}} \alpha_{g2} + \sigma_n^2}. \quad (11)$$

For convenience, let  $\varrho_{gk} = \frac{1}{[\mathbf{H}_{gk}^\dagger \mathbf{H}_{gk}^{\dagger H}]_{gg}}$  be the effective channel gain, and let  $\rho = 1/\sigma_n^2$  represent the transmit SNR. Given these definitions, (11) can be rewritten as

$$\gamma_{gk}^1 = \frac{\rho \varrho_{gk} \alpha_{g1}}{\rho \varrho_{gk} \alpha_{g2} + 1}, \quad \text{for } 1 \leq k \leq 2. \quad (12)$$

Note that, since user 2 is the strongest one, it will decode its own message with some residual interference, resulting in the following SINR

$$\gamma_{g2}^2 = \frac{\rho \varrho_{g2} \alpha_{g2}}{\rho \varrho_{g2} \mu \alpha_{g1} + 1}. \quad (13)$$

Then, for achieving a general SINR expression valid for both users, the following is defined

$$\alpha_{gi}^* = \begin{cases} \alpha_{g2}, & \text{for } i = 1, \\ \mu \alpha_{g1}, & \text{for } i = k = 2, \end{cases} \quad (14)$$

Lastly, by combining (12), (13), and (14), the final SINR expression is obtained, as follows

$$\gamma_{gk}^i = \frac{\rho \varrho_{gk} \alpha_{gi}}{\rho \varrho_{gk} \alpha_{gi}^* + 1}, \quad \text{for } 1 \leq i \leq k \leq 2, \quad (15)$$

which completes the proof.  $\blacksquare$

### B. Statistical Characterization of the Effective Channel Gains

Before obtaining the desired outage probability and ergodic rate expressions, we need to statistically characterize the effective channel gains. As one can observe in (11),  $\varrho_{gk}$  is the inverse of the  $g$ th element on the main diagonal of the following matrix

$$\begin{aligned} \mathbf{H}_{gk}^\dagger \mathbf{H}_{gk}^{\dagger H} &= [(\mathbf{H}_{gk}^H \mathbf{B})^H \mathbf{H}_{gk}^H \mathbf{B}]^{-1} (\mathbf{H}_{gk}^H \mathbf{B})^H \mathbf{H}_{gk}^H \mathbf{B} [( \mathbf{H}_{gk}^H \mathbf{B})^H \mathbf{H}_{gk}^H \mathbf{B}]^{-1 H} = (\mathbf{B}^H \mathbf{H}_{gk} \mathbf{H}_{gk}^H \mathbf{B})^{-1} \\ &= \left( \mathbf{B}^H \Phi d_g^{-\eta} (\mathbf{U} \Lambda^{\frac{1}{2}} \mathbf{G}_{gk}) (\mathbf{U} \Lambda^{\frac{1}{2}} \mathbf{G}_{gk})^H \mathbf{B} \right)^{-1} = \frac{1}{\Phi d_g^{-\eta}} (\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1} \in \mathbb{C}^{V \times V}. \end{aligned} \quad (16)$$

As demonstrated in [3], [4], since  $\mathbf{G}_{gk}$  consists of a complex Gaussian distributed matrix, the resulting matrix  $(\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1}$  is inverse Wishart distributed with  $N \geq V - 1$  degrees of freedom. Consequently, the unordered effective channel gain  $\varrho_{gk} = \frac{1}{[\mathbf{H}_{gk}^\dagger \mathbf{H}_{gk}^{\dagger H}]_{gg}}$  follows a Gamma distribution with shape parameter  $N - V + 1$  and scale parameter given by  $\Phi d_g^{-\eta} [(\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1}]_{gg}$ . However, since the BS sorts the users out in ascending order, we need to find the probability density function (PDF) of the ordered effective channel gains. To this end, we use the theory of order statistics, which allow us to achieve the desired PDF in the following way [23]

$$f_{\varrho_{gk}}(x) = K \binom{K-1}{k-1} \sum_{i=0}^{K-k} (-1)^i \binom{K-k}{i} \tilde{f}_{\varrho_{gk}}(x) \tilde{F}_{\varrho_{gk}}(x)^{k-1+i}, \quad (17)$$

where  $\tilde{f}_{\varrho_{gk}}(x)$  and  $\tilde{F}_{\varrho_{gk}}(x)$  are, respectively, the PDF and the cumulative distribution function (CDF) of unordered gains, which are provided in [3], [4]. Then, by using the fact that in our considered scenario  $K = 2$ , the PDF for the ordered gain of user 1 can be achieved as

$$f_{\varrho_{g1}}(x) = 2[\tilde{f}_{\varrho_{gk}}(x) - \tilde{f}_{\varrho_{gk}}(x) \tilde{F}_{\varrho_{gk}}(x)] = \frac{2\beta_g^\vartheta}{\Gamma(\vartheta)} \left[ x^{\vartheta-1} e^{-\beta_g x} - x^{\vartheta-1} e^{-\beta_g x} \frac{\gamma(\vartheta, \beta_g x)}{\Gamma(\vartheta)} \right], \quad (18)$$

and for user 2 as

$$f_{\varrho_{g2}}(x) = 2\tilde{f}_{\varrho_{gk}}(x) \tilde{F}_{\varrho_{gk}}(x) = \frac{2\beta_g^\vartheta}{\Gamma(\vartheta)} x^{\vartheta-1} e^{-\beta_g x} \frac{\gamma(\vartheta, \beta_g x)}{\Gamma(\vartheta)}, \quad (19)$$

in which, for notation simplicity, we have defined  $\vartheta = N - V + 1$  and  $\beta_g = \Phi d_g^{-\eta} [(\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1}]_{gg}$ .

### C. Outage Probability

The outage probability for the  $k$ th user in the  $g$ th sub-group, represented by  $P_{gk}$ , is the probability of achieving a data rate less than the target rate,  $T_{gi}$ , required to decode the message intended to the  $i$ th user,  $i \leq k \in \{1, 2\}$ , and it can be defined as

$$P_{gk} = P[\log_2(1 + \gamma_{gk}^i) < T_{gi}], \quad \forall i = 1, \dots, k. \quad (20)$$

Closed-form expressions for the outage probability are provided in the following proposition.

*Proposition 1:* Assuming that  $\varrho_{g1} < \varrho_{g2}$ , and considering imperfect SIC, the exact closed-form expressions for the outage probability achieved by users 1 and 2, can be derived as follows:

- For user 1:

$$P_{g1} = \begin{cases} \frac{2\gamma(\vartheta, \rho^{-1}\beta_g \mathcal{L}_{g1})}{\Gamma(\vartheta)} - \left[ \frac{\gamma(\vartheta, \rho^{-1}\beta_g \mathcal{L}_{g1})}{\Gamma(\vartheta)} \right]^2, & \text{if } \mathcal{L}_{g1} \geq 0, \\ 1, & \text{otherwise.} \end{cases} \quad (21)$$

- For user 2:

$$P_{g2} = \begin{cases} \left[ \frac{\gamma(\vartheta, \rho^{-1}\beta_g \max\{\mathcal{L}_{g1}, \mathcal{L}_{g2}\})}{\Gamma(\vartheta)} \right]^2, & \text{if } \min\{\mathcal{L}_{g1}, \mathcal{L}_{g2}\} \geq 0, \\ 1, & \text{otherwise,} \end{cases} \quad (22)$$

where  $\mathcal{L}_{g1} = \frac{2^{T_{g1}} - 1}{\alpha_{g1} - \alpha_{g2}(2^{T_{g1}} - 1)}$ , and  $\mathcal{L}_{g2} = \frac{2^{T_{g2}} - 1}{\alpha_{g2} - \mu\alpha_{g1}(2^{T_{g2}} - 1)}$ .

*Proof:* Please, see Appendix A. ■

### D. Ergodic Rates

In this subsection, it is assumed that the strong user can successfully decode the symbols intended for the weak user so that its achievable rate will depend only on the observed SINR while decoding its own data symbol. Under this consideration, the instantaneous data rate achieved by the first user can be written as

$$R_{g1} = \log_2(1 + \gamma_{g1}^1) = \log_2 \left( 1 + \frac{\rho \varrho_{g1} \alpha_{g1}}{\rho \varrho_{g1} \alpha_{g2} + 1} \right), \quad (23)$$

and, for the second user, the data rate is given by

$$R_{g2} = \log_2(1 + \gamma_{g2}^2) = \log_2 \left( 1 + \frac{\rho \varrho_{g2} \alpha_{g2}}{\rho \varrho_{g2} \mu \alpha_{g1} + 1} \right). \quad (24)$$

From (23) and (24), exact closed-form expressions for the users' ergodic rates will be derived, which are presented in Proposition II.

*Proposition II:* Assuming that the rate of strong users depends only on their own SINR, exact closed-form expressions for the ergodic rates with imperfect SIC can be obtained as follows:

- For user 1:

$$\bar{R}_{g1} = \xi_1(\kappa_{g1}) - \xi_1(\tilde{\kappa}_{g1}), \quad (25)$$

where  $\kappa_{g1} = \rho(\alpha_{g1} + \alpha_{g2})$ ,  $\tilde{\kappa}_{g1} = \rho\alpha_{g2}$ , and

$$\xi_1(\kappa) = \begin{cases} \sum_{i=0}^{\vartheta-1} \frac{1}{2^{\vartheta+i-1} \ln(2) \Gamma(\vartheta) i!} \sum_{m=0}^{\vartheta+i-1} \frac{(\vartheta+i-1)!}{(\vartheta+i-m-1)!} \left[ \frac{(-1)^{\vartheta+i-m-2}}{\left(\frac{\kappa}{2\beta_g}\right)^{\vartheta+i-m-1}} e^{\frac{2\beta_g}{\kappa}} \text{Ei} \left( -\frac{2\beta_g}{\kappa} \right) \right. \\ \left. + \sum_{n=1}^{\vartheta+i-m-1} \frac{(n-1)!}{\left(-\frac{\kappa}{2\beta_g}\right)^{\vartheta+i-m-n-1}} \right], & \text{if } \vartheta > 1, \\ -\frac{1}{\ln(2)} e^{\frac{2\beta_g}{\kappa}} \text{Ei} \left( -\frac{2\beta_g}{\kappa} \right), & \text{if } \vartheta = 1. \end{cases}$$

- For user 2:

$$\bar{R}_{g2} = \xi_2(\kappa_{g2}) - \xi_2(\tilde{\kappa}_{g2}), \quad (26)$$

where  $\kappa_{g2} = \rho(\mu\alpha_{g1} + \alpha_{g2})$ ,  $\tilde{\kappa}_{g2} = \rho\mu\alpha_{g1}$ , and

$$\xi_2(\kappa) = \begin{cases} \frac{2}{\ln(2)} \sum_{m=0}^{\vartheta-1} \frac{1}{(\vartheta-m-1)!} \left[ \frac{(-1)^{\vartheta-m-2}}{\left(\frac{\kappa}{\beta_g}\right)^{\vartheta-m-1}} e^{\frac{\beta_g}{\kappa}} \text{Ei} \left( -\frac{\beta_g}{\kappa} \right) \right. \\ \left. + \sum_{n=1}^{\vartheta-m-1} \frac{(n-1)!}{\left(-\frac{\kappa}{\beta_g}\right)^{\vartheta-m-n-1}} \right] - \xi_1(\kappa), & \text{if } \vartheta > 1, \\ -\frac{2}{\ln(2)} e^{\frac{\beta_g}{\kappa}} \text{Ei} \left( -\frac{\beta_g}{\kappa} \right) - \xi_1(\kappa), & \text{if } \vartheta = 1. \end{cases}$$

*Proof:* Please, see Appendix B. ■

#### IV. ENHANCING USER FAIRNESS THROUGH DYNAMIC POWER ALLOCATION

Even though fixed power allocation policy, which was considered in the previous section, is simpler to employ and has been widely adopted in several previous works [3], [4], [24], it

cannot guarantee fairness to users within the system, mainly to weaker users. The fact is that the performance achieved in NOMA networks is highly dependent on the power allocation strategy. Therefore, in this section, we develop dynamic power allocation protocols for enhancing user fairness within the proposed massive MIMO-NOMA network. More details are provided next.

#### A. Power Allocation within the NOMA Sub-Groups

First, we focus on enhancing user fairness only within each NOMA sub-group. Our goal is to develop an efficient adaptive power allocation policy to balance to similar levels the data rates of users within each specific sub-group. To achieve this objective, the rates of weaker users should be greater than or equal to that achieved by the stronger ones, i.e.,  $\log_2 \left( 1 + \frac{\rho \varrho_{g1} \alpha_{g1}}{\rho \varrho_{g1} \alpha_{g2} + 1} \right) \geq \log_2 \left( 1 + \frac{\rho \varrho_{g2} \alpha_{g2}}{\rho \varrho_{g2} \mu \alpha_{g1} + 1} \right)$ . This can be accomplished with the following optimization problem

$$\max_{\alpha_{g1}, \alpha_{g2}} \{R_{g2}\} \quad (27a)$$

$$\text{s.t. } R_{g1} \geq R_{g2}, \quad (27b)$$

$$\alpha_{g1} + \alpha_{g2} = \bar{\alpha}_g, \quad (27c)$$

$$\alpha_{g1} \geq 0, \alpha_{g2} \geq 0, \quad (27d)$$

where  $\bar{\alpha}_g$  denotes the total transmit power available for the  $g$ th sub-group.

Given that  $\log_2(\cdot)$  is a monotonic increasing function, from the constraint (27b), it follows that  $\frac{\rho \varrho_{g1} \alpha_{g1}}{\rho \varrho_{g1} \alpha_{g2} + 1} \geq \frac{\rho \varrho_{g2} \alpha_{g2}}{\rho \varrho_{g2} \mu \alpha_{g1} + 1}$ . As a result, the problem (27) can be simplified to

$$\max_{\alpha_{g1}, \alpha_{g2}} \left\{ \log_2 \left( 1 + \frac{\rho \varrho_{g2} \alpha_{g2}}{\rho \varrho_{g2} \mu \alpha_{g1} + 1} \right) \right\} \quad (28a)$$

$$\text{s.t. } \varrho_{g1} \varrho_{g2} \alpha_{g2}^2 - \mu \varrho_{g1} \varrho_{g2} \alpha_{g1}^2 + \varrho_{g2} \alpha_{g2} \rho^{-1} \leq \varrho_{g1} \alpha_{g1} \rho^{-1}, \quad (28b)$$

$$\alpha_{g1} + \alpha_{g2} = \bar{\alpha}_g, \quad (28c)$$

$$\alpha_{g1} \geq 0, \alpha_{g2} \geq 0. \quad (28d)$$

Then, by letting  $\alpha_{g1} = \bar{\alpha}_g - \alpha_{g2}$ , the constraint in (28b) becomes  $\alpha_{g2}^2 (\varrho_{g1} \varrho_{g2} - \mu \varrho_{g1} \varrho_{g2}) +$

$\alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) \leq 0$ , which results in

$$\max_{\alpha_{g2}} \left\{ \log_2 \left( 1 + \frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu\alpha_{g1} + 1} \right) \right\} \quad (29a)$$

$$\begin{aligned} \text{s.t. } & \alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) \\ & - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) \leq 0, \end{aligned} \quad (29b)$$

$$\alpha_{g2} \geq 0. \quad (29c)$$

By using the fact that  $\log_2 \left( 1 + \frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu\alpha_{g1} + 1} \right)$  increases monotonically with  $\alpha_{g2}$ , the problem (29) can be reduced to the optimization of only  $\alpha_{g2}$ , as follows

$$\min_{\alpha_{g2}} - \{ \alpha_{g2} \} \quad (30a)$$

$$\begin{aligned} \text{s.t. } & \alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) \\ & - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) \leq 0, \end{aligned} \quad (30b)$$

$$- \alpha_{g2} \leq 0, \quad (30c)$$

The optimal solution for (30) is given in the following proposition.

*Proposition III:* The optimization problem in (30) is convex and, consequently, has a global optimal solution, which is given in closed-form by

$$\begin{aligned} \alpha_{g2}^* = & \frac{1}{2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2})} \left[ -(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) \right. \\ & \left. + \sqrt{(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)^2 + 4\varrho_{g1}\varrho_{g2}(1 - \mu)(\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)} \right], \end{aligned} \quad (31)$$

and

$$\alpha_{g1}^* = \bar{\alpha}_g - \alpha_{g2}^*. \quad (32)$$

*Proof:* Please, see Appendix C. ■

## B. Providing Fairness Among Sub-Groups

In the last sub-section, we have developed a dynamic power allocation policy for providing fairness to the users within each NOMA sub-group. However, users located at different sub-

---

**Algorithm 1:** Iterative Algorithm for Fairness Among Sub-Groups
 

---

**Input:**  $\epsilon, \rho, \varrho_{g1}, \varrho_{g2}$ .

- 1 Set the initial available power to  $\bar{\alpha}_g = 1, \forall g$ ;
- 2 **do**
- 3     **for**  $g = 1$  to  $G$  **do**
- 4         Calculate  $\alpha_{g2}^*$  and  $\alpha_{g1}^*$  using (31) and (32), respectively;
- 5         Calculate the sub-group's sum-rate by  $\mathcal{R}_g = R_{g1} + R_{g2}$ ;
- 6     **end**
- 7      $\hat{g} = \mathbf{argmax}(\mathcal{R}_g : \forall g \in \{1, 2, \dots, G\})$ ;
- 8      $\check{g} = \mathbf{argmin}(\mathcal{R}_g : \forall g \in \{1, 2, \dots, G\})$ ;
- 9     Calculate  $\Delta_\alpha$  using (33);
- 10     Update  $\bar{\alpha}_{\hat{g}} = \bar{\alpha}_{\hat{g}} - \Delta_\alpha$ ;
- 11     Update  $\bar{\alpha}_{\check{g}} = \bar{\alpha}_{\check{g}} + \Delta_\alpha$ ;
- 12      $\epsilon^* = \mathcal{R}_{\hat{g}} - \mathcal{R}_{\check{g}}$ ;
- 13 **while**  $\epsilon^* > \epsilon$ ;

---

groups can still experience different performance levels. This represents an unfair distribution of resources since some groups can achieve high data rates while others can be almost in a state of outage. In view of this, in this sub-section, we develop a strategy for improving fairness also among different sub-groups. For achieving this goal, we propose an iterative algorithm that provides a fair power allocation for all users in all sub-groups within each cluster so that all terminals can reach identical data rates. The basic idea is to, at each iteration, transfer a certain amount of power, denoted by  $\Delta_\alpha$ , from the best to the worst sub-group, and to use the dynamic power allocation derived in the last sub-section to iteratively rebalance each user's individual rate. This iterative solution is shown in Algorithm 1. As one can observe, in the first stage of the algorithm, we calculate the users' power allocation coefficients by using (31) and (32) and compute the resulting sum-rate,  $\mathcal{R}_g$ , for each sub-group. Then, in lines 7 and 8, the indexes of sub-groups with the highest and lowest sum-rate are selected, which are represented by  $\hat{g}$  and  $\check{g}$ , respectively. After that, we calculate the amount of power  $\Delta_\alpha$  that needs to be reallocated from the group  $\hat{g}$  to the group  $\check{g}$ . This process repeats until the sum-rate difference between the best and worst sub-group reaches a value lower than a predefined threshold  $\epsilon$ .

The value of  $\Delta_\alpha$  is determined by the amount of power that is required to balance the data rates of the strongest users from the best and worst sub-groups, in which, in this section, we assume that  $\mu = 0$ , i.e., SIC is carried out perfectly.  $\Delta_\alpha$  is calculated in the following proposition.



*Proposition IV:* The amount of power  $\Delta_\alpha$  needed to balance the rates of the strongest users from the sub-groups with the highest and lowest sum-rate, assuming perfect SIC, is given by

$$\Delta_\alpha = \frac{-A_2 \pm \sqrt{A_2^2 - 4A_1A_3}}{2A_1}, \quad (33)$$

where

$$\begin{aligned} A_1 &= 4\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1} + 4K_1, & A_2 &= 2A_1K_3 + 16K_2K_1, \\ A_3 &= K_3^2 - 4K_2\frac{\varrho_{\hat{g}1}^2}{\varrho_{\hat{g}1}^2}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 - 16K_2K_1\bar{\alpha}_{\hat{g}}, \end{aligned}$$

and

$$\begin{aligned} K_1 &= \frac{\varrho_{\hat{g}1}^2}{\varrho_{\hat{g}1}^2}\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1}, & K_2 &= (\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1} - \frac{\varrho_{\hat{g}1}}{\varrho_{\hat{g}1}}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1}))^2, \\ K_3 &= (\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 + 4\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1}\bar{\alpha}_{\hat{g}} - K_2 - \frac{\varrho_{\hat{g}1}^2}{\varrho_{\hat{g}1}^2}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 - 4K_1\bar{\alpha}_{\hat{g}}. \end{aligned}$$

*Proof:* Please, see Appendix D. ■

## V. SIMULATION RESULTS AND DISCUSSIONS

In this section, we investigate the performance of the proposed massive MIMO-NOMA system under the impact of imperfect SIC employing both fixed and dynamic power allocation policies. We also present performance comparisons with conventional massive MIMO-OMA scheme, whose implementation details can be found in [4]. We configure the BS with a uniform linear array of  $M = 90$  antennas, which is transmitting information to users that are distributed among  $S = 4$  spatial clusters, each one having a diameter of  $D = 50$  m and an angular spread of  $\delta = 10^\circ$ , which corresponds to a distance of  $L = \frac{D}{2\tan(\delta)} \cong 141$  m from the BS to the center of the cluster. In addition, we configure the direction of the antenna array to the cluster that is being analyzed, i.e., the first cluster, which is located at the azimuth angle of  $\varphi = 7^\circ$ , so that the array gain is maximized. Within each cluster, if not stated otherwise, there are  $G = V = 2$  NOMA sub-groups with  $K = 2$  users each, and we focus on the first sub-group, which is located at 115 m from the BS. The path-loss exponent is set to  $\eta = 2$ , and the array gain parameter to  $\Phi = 4 \times 10^4$ . Moreover, when fixed power allocation is considered, the power coefficients of

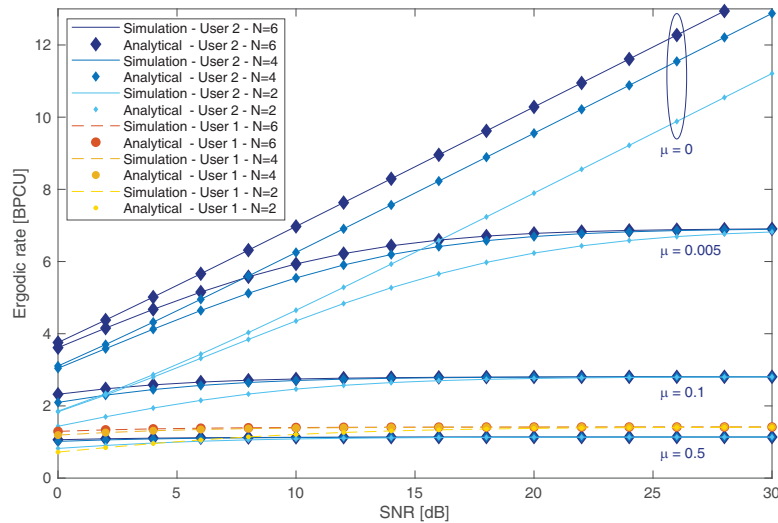


Fig. 2: Impact of imperfect SIC on the users' ergodic rates for different number of receive antennas.

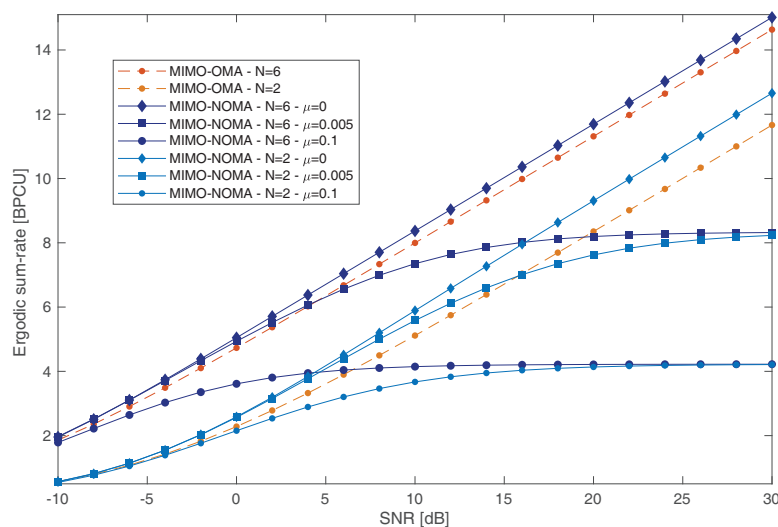


Fig. 3: Simulated ergodic sum-rate curves for massive MIMO-NOMA system with imperfect SIC and conventional massive MIMO-OMA counterpart.

users 1 and 2 are configured as  $\alpha_1 = 5/8$  and  $\alpha_2 = 3/8$ , respectively. All provided simulation results are generated by averaging extensive random channel realizations.

#### A. Fixed Power Allocation Results

In this subsection, the performance analysis derived in Section III is validated, in which, in all figures, a perfect agreement between analytical and simulated curves can be visualized. Besides, all results provided in this subsection are generated by employing a fixed power allocation policy.

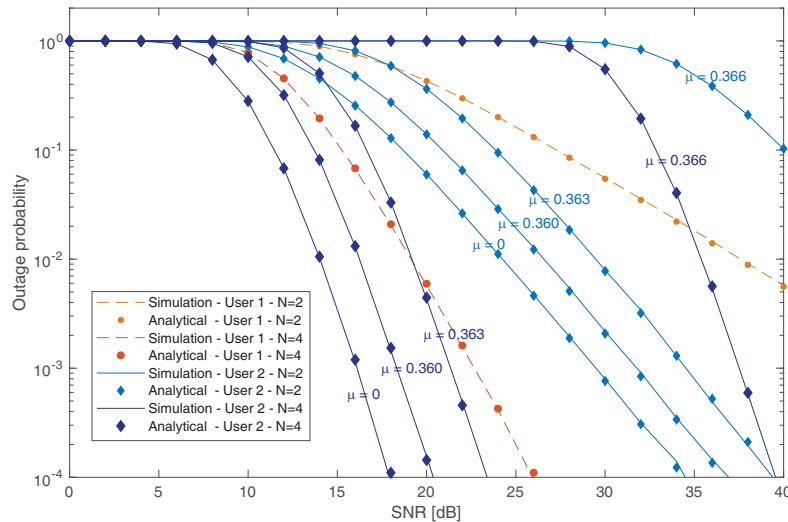


Fig. 4: Impact of imperfect SIC on the outage probability for different number of receive antennas ( $T_1 = T_2 = 1.4$  BPCU).

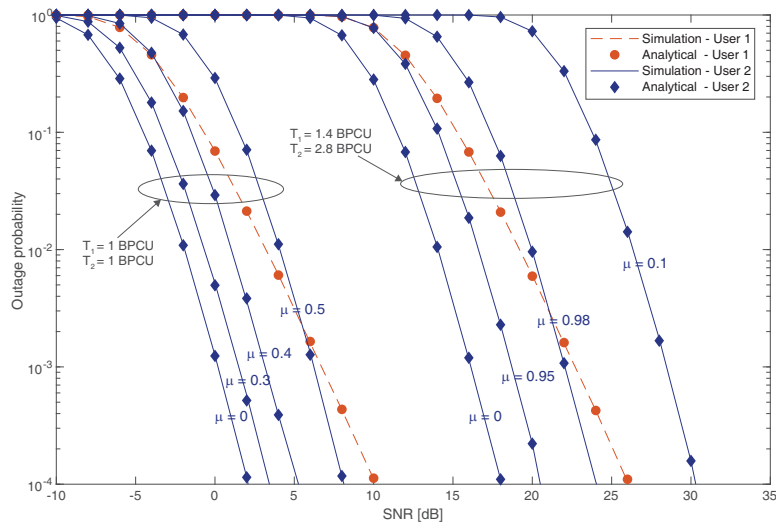


Fig. 5: Outage probability versus transmit SNR for different SIC interference levels and different target rates ( $N = 4$ ).

Fig. 2 shows the ergodic rates in terms of transmit SNR for different levels of SIC error propagation and various numbers of receive antennas. As one can notice, when perfect SIC is considered, by increasing the number of receive antennas, the performance of the strong user is improved for all considered SNR range. However, when there is some residual error from imperfect SIC, the maximum achievable rate decreases as  $\mu$  gets higher. For instance, for an error factor of  $\mu = 0.005$ , when  $\rho = 30$  dB and  $N = 2$ , the strong user's rate reaches a limit

of 6.82 bits per channel use (BPCU), which represents a reduction of 4.4 BPCU if compared with the perfect SIC case considering the same value of transmit SNR. When  $\mu = 0.5$ , the impact on the performance of the strong user is even more severe, where regardless of how many antennas are employed, a rate of only 1.14 BPCU can be reached, which is lower than that achievable by the weak user. This behavior is justified by the fact that when  $\rho \rightarrow \infty$ , if  $\mu > 0$ ,  $R_{g2} \rightarrow \log_2 \left( 1 + \frac{\alpha_{g2}}{\mu\alpha_{g1}} \right)$ . Therefore, if there is some residual SIC error and  $\alpha_{g1} > 0$ , there will be always a rate ceiling for the strong user.

In Fig. 3, the ergodic sum-rate performance achieved with the proposed massive MIMO-NOMA system is compared with conventional massive MIMO-OMA counterpart, in which the impact of imperfect SIC is investigated. One can see that when the error factor is greater than zero, at some point the MIMO-OMA system outperforms the MIMO-NOMA design. This behavior show us that employing NOMA is not always advantageous when SIC imperfection is significant. For example, when  $\mu = 0.005$  and  $N = 2$ , from 16dB onward, the OMA sum-rate performance becomes superior to that achieved by the NOMA scheme, which saturates at 8.33 BPCU. When  $\mu = 0.1$ , either for  $N = 2$  and  $N = 6$ , the MIMO-OMA system always achieves higher performance than the massive MIMO-NOMA system, meaning that when error propagation is high, the employment of OMA is always preferable.

Figs. 4 and 5 show the outage probability curves for different numbers of receive antennas, target rates, and error propagation factors. In Fig. 4, by fixing the target rates of weak and strong users at 1.4 BPCU, one can see that when the error factor gets higher than 0.36, with just a tiny error increase, the performance of the strong user is severely degraded. In particular, when  $N = 4$  and  $\mu$  is increased from 0.363 to 0.366, the outage probability of the strong user becomes worse even than that achieved by the weak user employing  $N = 2$  receive antennas for SNR values lower than 36dB. In Fig. 5, we can observe the impact of SIC error propagation for different sets of target rates. One can realize that for higher target values, the outage probability performance becomes more sensible to imperfect SIC. For example, by setting the target rates of both users to 1 BPCU, when  $\mu$  is increased from 0 to 0.5, the outage curve of the strong user shifted only 6dB to the right. On the other hand, when  $T_1 = 1.4$  and  $T_2 = 2.8$  BPCU, for an error factor of only  $\mu = 0.1$ , the strong user requires approximately 12dB more SNR to reach the same performance of that achieved when perfect SIC is considered.

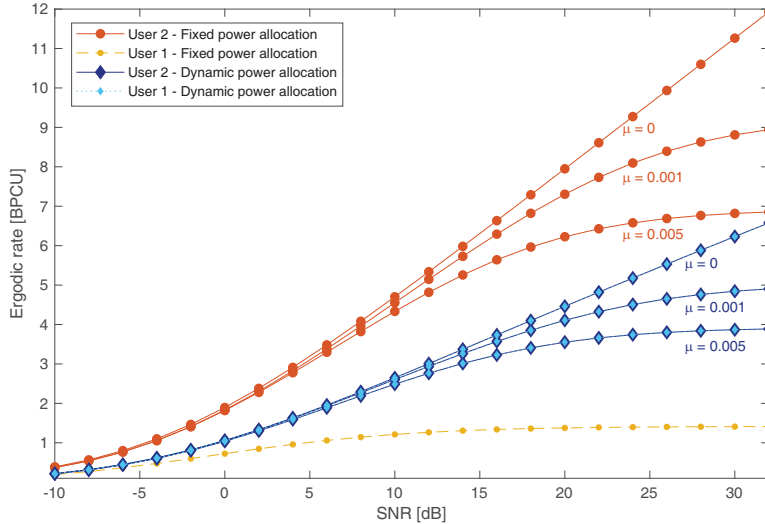


Fig. 6: Ergodic rates for strong and weak users in massive MIMO-NOMA system with dynamic and fixed power allocation policies ( $N = 2$ ).

### B. Dynamic Power Allocation Results

Now, the dynamic power allocation policies achieved in Section IV are investigated. Fig. 6 demonstrates the effectiveness of the optimum solution obtained in Proposition III, in which the ergodic rates of weak and strong users employing fixed and dynamic power allocation are shown. One can see that, with fixed power allocation, the performance of the weak user is strongly impacted, such that its ergodic rate reaches a very low limit for higher values of transmit SNR. In contrast, the strong user experiences high data rates even when SIC error propagation is present. This illustrates an unfair resource allocation. On the other hand, the dynamic policy provides great benefits to the weak user, improving fairness within the sub-group. As one can observe, the rates of the two users are balanced so that both achieve an acceptable performance. For instance, for an SNR of 22dB when perfect SIC is considered, both users can reach a rate of 4.82 BPCU with dynamic power allocation, what represents an improvement of 3.43 BPCU to the weak user if compared with that achieved with fixed policy of only 1.39 BPCU.

Considering perfect SIC, Fig. 7 brings the ergodic rate curves for various values of receive antennas, exclusively for the weak user within the considered sub-group. As one can see, in the fixed power allocation, regardless of how many receive antennas are employed or how much the transmit SNR is improved, the achievable rate approaches a common limiting value. This does not happen with the proposed dynamic allocation. As it can be observed, the performance continues

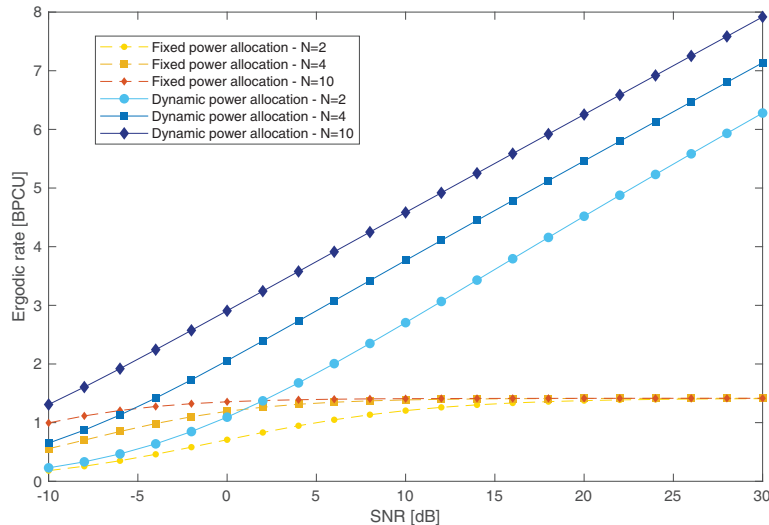


Fig. 7: Ergodic rates for the weak user in massive MIMO-NOMA system with dynamic and fixed power allocation policies ( $\mu = 0$ ).

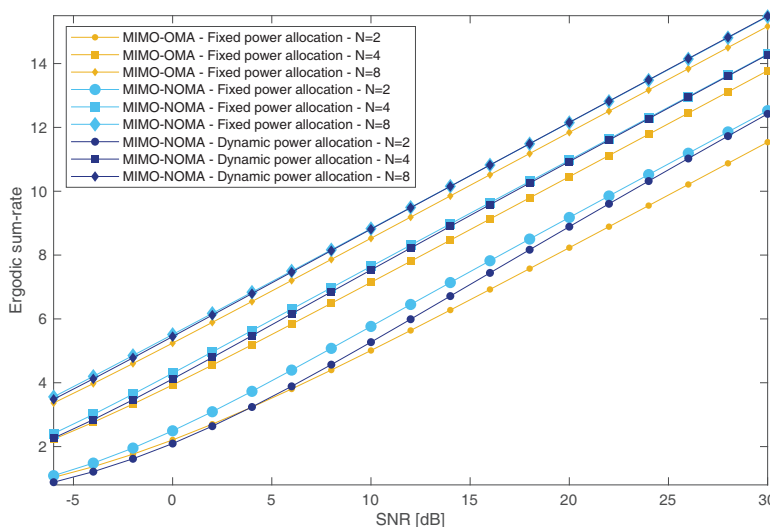


Fig. 8: Ergodic sum-rate curves for massive MIMO-NOMA and MIMO-OMA systems with dynamic and fixed power allocation policies ( $\mu = 0$ ).

to increase even for higher SNR values. For example, considering a transmit SNR of 24dB and  $N = 10$  receive antennas, the dynamic allocation can achieve a rate of 6.92 BPCU, which is almost 5 times greater than the achieved with the fixed policy. Fig. 8 compares the ergodic sum-rate curves achieved in MIMO-NOMA and MIMO-OMA systems. One can realize that dynamic allocation causes a slight decrease in the performance of the MIMO-NOMA scheme. This is because, in order to enhance fairness, the optimization problem in (27) decreases the

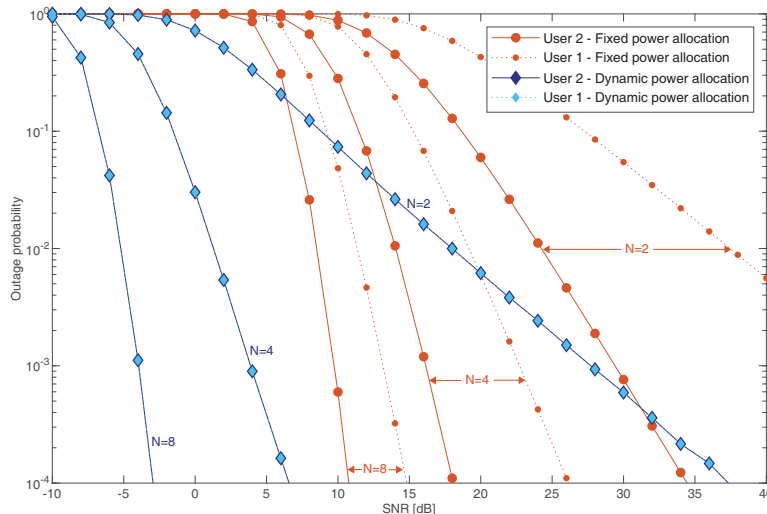


Fig. 9: Outage probabilities achieved with dynamic and fixed power allocation policies in massive MIMO-NOMA systems ( $T_1 = T_2 = 1.4$  BPCU;  $\mu = 0$ ).

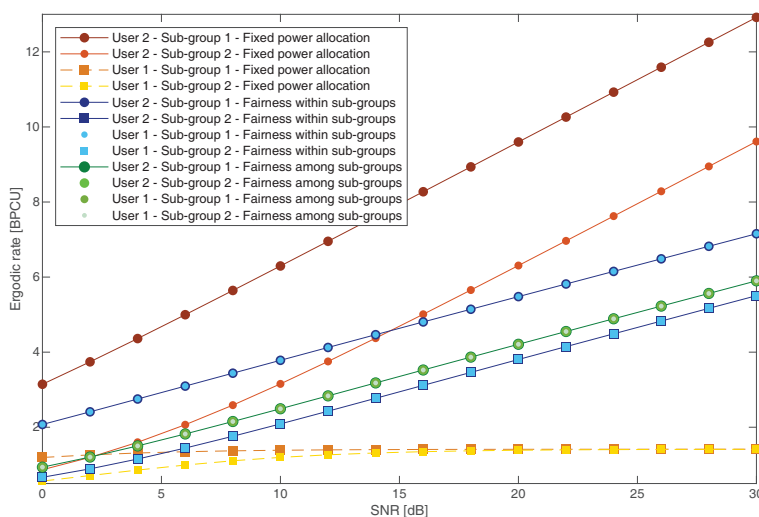


Fig. 10: Ergodic rates for users in different sub-groups employing different power allocation policies ( $d_1 = 115\text{m}$ ;  $d_2 = 150\text{m}$ ;  $N = 4$ ;  $\mu = 0$ ).

strong users' rates, which impacts the system sum-rate. However, it is noteworthy that, for all values of receive antennas, the performance achieved in the MIMO-NOMA system employing dynamic power allocation can still outperform the conventional MIMO-OMA counterpart.

Fig. 9 demonstrates the benefits of the dynamic power allocation on the outage probability. It is interesting to observe that, in addition to the fairness improvements, the outage performance of both weak and strong users is remarkably improved. For instance, with  $N = 4$  receive antennas,

when employing the dynamic policy, the strong user requires roughly 12dB less SNR to reach the same outage level of that achieved with fixed power allocation. The performance gains obtained by the weak user with dynamic allocation are even more impressive, in which a gain of incredible 20dB can be achieved.

At last, by considering different power allocation protocols, Fig. 10 plots the ergodic rates for users within two different sub-groups, one located at 115m and other at 150m from the BS. It becomes clear that, even though the optimization problem in (27) is capable of providing fairness to users within the same sub-group, users from other sub-groups can still experience different performance levels, which, in some applications, might not be desirable. This figure also illustrates the performance of the iterative algorithm proposed in Section IV-B, which provides fairness also among different sub-groups. We see that, when the referred algorithm is adopted, the rates of users from the worst sub-group are improved at the cost of reducing the performance of users from the best one. However, if we compare with fixed power allocation, Algorithm 1 is very beneficial to the weak users independently of the group. For example, for an SNR of 24dB, all users employing the iterative algorithm can reach a rate of 4.9 BPCU, which represents a gain of 3.5 BPCU for all weak users when adopting the fixed policy.

## VI. CONCLUSIONS

In this paper, by modeling residual error propagation from imperfect SIC, the performance of a massive MIMO-NOMA network was investigated. In particular, the detailed design of beamformers and detection matrices were presented, an in-depth analytical analysis was carried out, and optimum power allocation for maximizing the rates of weak users within each sub-group was derived. An iterative algorithm for providing fairness among different sub-groups was also proposed. The simulation results demonstrated that the developed dynamic power allocation provides remarkable fairness enhancements and, at the same time, enormous performance gains in terms of outage probability. It also became evident that when SIC error propagation is present, the employment of NOMA is not always advantageous.



## APPENDIX A

## PROOF OF PROPOSITION I

The outage probability expression in (20) can be rewritten as follows

$$P_{gk} = P \left[ \log_2 \left( 1 + \frac{\rho \varrho_{gk} \alpha_{gi}}{\rho \varrho_{gk} \alpha_{gi}^* + 1} \right) < T_{gi} \right] = P \left[ \varrho_{gk} < \frac{2^{T_{gi}} - 1}{\rho(\alpha_{gi} - \alpha_{gi}^*(2^{T_{gi}} - 1))} \right], \quad (\text{A-1})$$

in which, for user 1 (the weak user), (A-1) can be simplified as

$$P_{g1} = P \left[ \varrho_{g1} < \frac{\rho^{-1}(2^{T_{g1}} - 1)}{\alpha_{g1} - \alpha_{g2}(2^{T_{g1}} - 1)} \right] = P [\varrho_{g1} < \rho^{-1} \mathcal{L}_{g1}], \quad (\text{A-2})$$

while, for user 2 (the strong user), (A-1) becomes

$$\begin{aligned} P_{g2} &= P \left[ \varrho_{g2} < \rho^{-1} \max \left\{ \frac{2^{T_{g1}} - 1}{\alpha_{g1} - \alpha_{g2}(2^{T_{g1}} - 1)}, \frac{2^{T_{g2}} - 1}{\alpha_{g2} - \mu \alpha_{g1}(2^{T_{g2}} - 1)} \right\} \right] \\ &= P [\varrho_{g2} < \rho^{-1} \max \{\mathcal{L}_{g1}, \mathcal{L}_{g2}\}], \end{aligned} \quad (\text{A-3})$$

in which, for simplicity, we have defined  $\mathcal{L}_{g1} = \frac{2^{T_{g1}} - 1}{\alpha_{g1} - \alpha_{g2}(2^{T_{g1}} - 1)}$  and  $\mathcal{L}_{g2} = \frac{2^{T_{g2}} - 1}{\alpha_{g2} - \mu \alpha_{g1}(2^{T_{g2}} - 1)}$ .

As one can observe, (A-2) and (A-3) are equivalent to the CDF of the effective channel gains of users 1 and 2, respectively. Consequently, the outage probability expressions can be obtained by integrating the PDFs in (18) and (19), in which, for user 1, it results in

$$\begin{aligned} P_{g1} &= \frac{2\beta_g^\vartheta}{\Gamma(\vartheta)} \left[ \int_0^{\rho^{-1}\mathcal{L}_{g1}} x^{\vartheta-1} e^{-\beta_g x} dx - \int_0^{\rho^{-1}\mathcal{L}_{g1}} x^{\vartheta-1} e^{-\beta_g x} \frac{\gamma(\vartheta, \beta_g x)}{\Gamma(\vartheta)} dx \right] \\ &= \begin{cases} \frac{2\gamma(\vartheta, \rho^{-1}\beta_g \mathcal{L}_{g1})}{\Gamma(\vartheta)} - \left[ \frac{\gamma(\vartheta, \rho^{-1}\beta_g \mathcal{L}_{g1})}{\Gamma(\vartheta)} \right]^2, & \text{if } \mathcal{L}_{g1} \geq 0, \\ 1, & \text{otherwise,} \end{cases} \end{aligned} \quad (\text{A-4})$$

while, for user 2, the following is obtained

$$P_{g2} = \begin{cases} \left[ \frac{\gamma(\vartheta, \rho^{-1}\beta_g \max\{\mathcal{L}_{g1}, \mathcal{L}_{g2}\})}{\Gamma(\vartheta)} \right]^2, & \text{if } \min\{\mathcal{L}_{g1}, \mathcal{L}_{g2}\} \geq 0, \\ 1, & \text{otherwise,} \end{cases} \quad (\text{A-5})$$

which completes the proof. ■

## APPENDIX B

## PROOF OF PROPOSITION II

The ergodic rates for users 1 and 2, can be obtained by calculating the expected value of their instantaneous rates. Then, firstly, let us rewrite the rate expression of user 1, in (23), as follows

$$R_{g1} = \log_2 \left( \frac{1 + \rho(\alpha_{g1} + \alpha_{g2})\varrho_{g1}}{1 + \rho\alpha_{g2}\varrho_{g1}} \right) = \log_2 (1 + \kappa_{g1}\varrho_{g1}) - \log_2 (1 + \tilde{\kappa}_{g1}\varrho_{g1}), \quad (\text{B-1})$$

and, for user 2, as

$$R_{g2} = \log_2 \left( \frac{1 + \rho(\mu\alpha_{g1} + \alpha_{g2})\varrho_{g2}}{1 + \rho\mu\alpha_{g1}\varrho_{g2}} \right) = \log_2 (1 + \kappa_{g2}\varrho_{g2}) - \log_2 (1 + \tilde{\kappa}_{g2}\varrho_{g2}), \quad (\text{B-2})$$

in which, for notation convenience, it has been defined  $\kappa_{g1} = \rho(\alpha_{g1} + \alpha_{g2})$ ,  $\tilde{\kappa}_{g1} = \rho\alpha_{g2}$ ,  $\kappa_{g2} = \rho(\mu\alpha_{g1} + \alpha_{g2})$ , and  $\tilde{\kappa}_{g2} = \rho\mu\alpha_{g1}$ .

Given the expression in (B-1), the ergodic rate for user 1 can be expressed as

$$\bar{R}_{g1} = \int_0^\infty \log_2 (1 + \kappa_{g1}x) f_{\varrho_{g1}}(x) dx - \int_0^\infty \log_2 (1 + \tilde{\kappa}_{g1}x) f_{\varrho_{g1}}(x) dx = \xi_1(\kappa_{g1}) - \xi_1(\tilde{\kappa}_{g1}). \quad (\text{B-3})$$

Then, by invoking the PDF of  $\varrho_{g1}$ , provided in (18),  $\xi_1(\kappa)$  can be calculated as follows

$$\xi_1(\kappa) = \frac{2\beta_g^\vartheta}{\Gamma(\vartheta)} \left[ \int_0^\infty \log_2 (1 + \kappa x) x^{\vartheta-1} e^{-\beta_g x} dx - \int_0^\infty \log_2 (1 + \kappa x) x^{\vartheta-1} e^{-\beta_g x} \frac{\gamma(\vartheta, \beta_g x)}{\Gamma(\vartheta)} dx \right]. \quad (\text{B-4})$$

Next, by applying the series representation of the incomplete gamma function to the second integral in (B-4), we obtain

$$\xi_1(\kappa) = \sum_{i=0}^{\vartheta-1} \frac{2\beta_g^{\vartheta+i}}{\Gamma(\vartheta)i!} \int_0^\infty \log_2 (1 + \kappa x) x^{\vartheta+i-1} e^{-2\beta_g x} dx. \quad (\text{B-5})$$

Lastly, after some algebraic manipulation and applying results from [20], we achieve the

desired solution, as follows

$$\xi_1(\kappa) = \begin{cases} \sum_{i=0}^{\vartheta-1} \frac{1}{2^{\vartheta+i-1} \ln(2) \Gamma(\vartheta) i!} \sum_{m=0}^{\vartheta+i-1} \frac{(\vartheta+i-1)!}{(\vartheta+i-m-1)!} \left[ \frac{(-1)^{\vartheta+i-m-2}}{\left(\frac{\kappa}{2\beta_g}\right)^{\vartheta+i-m-1}} e^{\frac{2\beta_g}{\kappa}} \text{Ei}\left(-\frac{2\beta_g}{\kappa}\right) \right. \\ \left. + \sum_{n=1}^{\vartheta+i-m-1} \frac{(n-1)!}{\left(-\frac{\kappa}{2\beta_g}\right)^{\vartheta+i-m-n-1}} \right], & \text{if } \vartheta > 1, \\ -\frac{1}{\ln(2)} e^{\frac{2\beta_g}{\kappa}} \text{Ei}\left(-\frac{2\beta_g}{\kappa}\right), & \text{if } \vartheta = 1. \end{cases}$$

Now, we focus on the second user, in which, from (B-2), its ergodic rate can be obtained as

$$\bar{R}_{g2} = \int_0^\infty \log_2(1 + \kappa_{g2}x) f_{\varrho_{g2}}(x) dx - \int_0^\infty \log_2(1 + \tilde{\kappa}_{g2}x) f_{\varrho_{g2}}(x) dx = \xi_2(\kappa_{g2}) - \xi_2(\tilde{\kappa}_{g2}), \quad (\text{B-6})$$

where  $\xi_2(\kappa)$  can be derived as

$$\begin{aligned} \xi_2(\kappa) &= \frac{2\beta_g^\vartheta}{\ln(2)\Gamma(\vartheta)} \int_0^\infty \ln(1 + \kappa x) x^{\vartheta-1} e^{-\beta_g x} dx - \sum_{i=0}^{\vartheta-1} \frac{2\beta_g^{\vartheta+i}}{\ln(2)\Gamma(\vartheta) i!} \int_0^\infty \ln(1 + \kappa x) x^{\vartheta+i-1} e^{-2\beta_g x} dx \\ &= \frac{2\beta_g^\vartheta}{\ln(2)\Gamma(\vartheta)} \int_0^\infty \ln(1 + \kappa x) x^{\vartheta-1} e^{-\beta_g x} dx - \xi_1(\kappa). \end{aligned} \quad (\text{B-7})$$

Finally, by doing some manipulations in (B-7), and also using results from [20], we obtain the following solution

$$\xi_2(\kappa) = \begin{cases} \frac{2}{\ln(2)} \sum_{m=0}^{\vartheta-1} \frac{1}{(\vartheta-m-1)!} \left[ \frac{(-1)^{\vartheta-m-2}}{\left(\frac{\kappa}{\beta_g}\right)^{\vartheta-m-1}} e^{\frac{\beta_g}{\kappa}} \text{Ei}\left(-\frac{\beta_g}{\kappa}\right) \right. \\ \left. + \sum_{n=1}^{\vartheta-m-1} \frac{(n-1)!}{\left(-\frac{\kappa}{\beta_g}\right)^{\vartheta-m-n-1}} \right] - \xi_1(\kappa), & \text{if } \vartheta > 1, \\ -\frac{2}{\ln(2)} e^{\frac{\beta_g}{\kappa}} \text{Ei}\left(-\frac{\beta_g}{\kappa}\right) - \xi_1(\kappa), & \text{if } \vartheta = 1, \end{cases}$$

which completes the proof. ■

## APPENDIX C

### PROOF OF PROPOSITION III

Clearly, the objective function in (30a) is linear and the function on the left-hand-side of constraint (30b) consists in a quadratic polynomial. As  $\varrho_{gk} \geq 0, \forall g, k$ , the constraint in (30b) is convex. This makes (30) a convex optimization problem. Therefore, the KKT conditions are

necessary and sufficient to determine the global optimal solution of the considered problem [25].

The Lagrangian function of (30) can be written as

$$\begin{aligned} \mathcal{L}(\alpha_{g2}, \omega, \nu) = & -\alpha_{g2} + \omega[\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) \\ & - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)] - \nu\alpha_{g2}, \end{aligned} \quad (\text{C-1})$$

where  $\omega \geq 0$  and  $\nu \geq 0$  are, respectively, the Lagrangian multipliers associated with the constraints (30b) and (30c). The KKT conditions are summarized as follows

$$\nabla \mathcal{L}(\alpha_{g2}, \omega, \nu) = 0, \quad (\text{C-2a})$$

$$\begin{aligned} \omega[\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) \\ - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)] = 0, \end{aligned} \quad (\text{C-2b})$$

$$- \nu\alpha_{g2} = 0. \quad (\text{C-2c})$$

Assuming that  $\alpha_{g2} > 0$ , from (C-2c) it can be concluded that  $\nu = 0$ . Then, from (C-2a), the value of  $\omega$  is easily determined as follows

$$\omega = (2\alpha_{g2}(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)^{-1}. \quad (\text{C-3})$$

Considering that the expression in (C-3) never reaches zero, the only possible solution for (C-2b) can be obtained from the following quadratic equation

$$\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) = 0. \quad (\text{C-4})$$

Finally, the optimal power allocation for user 2 can be obtained by calculating the zeros of (C-4), as follows

$$\begin{aligned} \alpha_{g2}^* = & \frac{1}{2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2})} \left[ -(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) \right. \\ & \left. + \sqrt{(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)^2 + 4\varrho_{g1}\varrho_{g2}(1 - \mu)(\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)} \right], \end{aligned} \quad (\text{C-5})$$

which completes the proof. ■

## APPENDIX D

## PROOF OF PROPOSITION IV

The amount of power  $\Delta_\alpha$  can be calculated by equalizing the rate expressions of the two strongest users of interest, in which, by considering that  $\mu = 0$ , the following is obtained

$$R_{\hat{g}2} = R_{\bar{g}2} \implies \log_2(1 + \rho \varrho_{\hat{g}2} \alpha_{\hat{g}2}^*) = \log_2(1 + \rho \varrho_{\bar{g}2} \alpha_{\bar{g}2}^*) \implies \alpha_{\hat{g}2}^* \varrho_{\hat{g}2} = \alpha_{\bar{g}2}^* \varrho_{\bar{g}2}. \quad (\text{D-1})$$

Next, by replacing  $\alpha_{\hat{g}2}^*$  and  $\alpha_{\bar{g}2}^*$  by their corresponding closed-form expressions, (D-1) becomes

$$\begin{aligned} & -2\varrho_{\hat{g}1}\varrho_{\hat{g}1}\rho^{-1} - 2\varrho_{\hat{g}1}\varrho_{\hat{g}2}\rho^{-1} + 2\varrho_{\hat{g}1}\sqrt{(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 + 4\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1}(\bar{\alpha}_{\hat{g}} - \Delta_\alpha)} \\ & = -2\varrho_{\hat{g}1}\varrho_{\hat{g}1}\rho^{-1} - 2\varrho_{\hat{g}1}\varrho_{\hat{g}2}\rho^{-1} + 2\varrho_{\hat{g}1}\sqrt{(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 + 4\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1}(\bar{\alpha}_{\hat{g}} + \Delta_\alpha)}. \end{aligned} \quad (\text{D-2})$$

Then, after some algebraic manipulation, and defining  $K_1 = \frac{\varrho_{\hat{g}1}^2}{\varrho_{\bar{g}1}^2}\varrho_{\hat{g}1}\varrho_{\hat{g}2}\rho^{-1}$ ,  $K_2 = (\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1} - \frac{\varrho_{\hat{g}1}}{\varrho_{\bar{g}1}}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1}))^2$ ,  $K_3 = ((\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 + 4\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1}\bar{\alpha}_{\hat{g}} - K_2 - \frac{\varrho_{\hat{g}1}^2}{\varrho_{\bar{g}1}^2}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 - 4K_1\bar{\alpha}_{\hat{g}})$ ,  $A_1 = 4\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1} + 4K_1$ ,  $A_2 = 2A_1K_3 + 16K_2K_1$ , and  $A_3 = (K_3^2 - 4K_2\frac{\varrho_{\hat{g}1}^2}{\varrho_{\bar{g}1}^2}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 - 16K_2K_1\bar{\alpha}_{\hat{g}})$ , we achieve the following quadratic equation

$$A_1\Delta_\alpha^2 + A_2\Delta_\alpha + A_3 = 0. \quad (\text{D-3})$$

The final result is obtained by calculating the zeros of (D-3). This completes the proof.  $\blacksquare$

## REFERENCES

- [1] L. Lv, J. Chen, Q. Ni, Z. Ding, and H. Jiang, "Cognitive non-orthogonal multiple access with cooperative relaying: A new wireless frontier for 5G spectrum sharing," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 188–195, Apr. 2018.
- [2] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.
- [3] Z. Ding and V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Proces. Lett.*, vol. 23, no. 5, May 2016.
- [4] A. S. de Sena, D. B. da Costa, Z. Ding, and P. H. J. Nardelli, "Massive MIMO-NOMA networks with multi-polarized antennas," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2019.
- [5] H. Xing, Y. Liu, A. Nallanathan, Z. Ding, and H. V. Poor, "Optimal throughput fairness tradeoffs for downlink non-orthogonal multiple access over fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3556–3571, Jun. 2018.
- [6] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Trans. Wireless Commun.*, vol. 66, no. 4, pp. 1717–1731, Apr. 2018.

- [7] M. Zeng, N. Nguyen, O. A. Dobre, and H. V. Poor, "Securing downlink massive MIMO-NOMA networks with artificial noise," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 685–699, Jun. 2019.
- [8] H. Sun, B. Xie, R. Q. Hu, and G. Wu, "Non-orthogonal multiple access with SIC error propagation in downlink wireless MIMO networks," in *IEEE Veh. Technol. Conf.*, Sep. 2016, pp. 1–5.
- [9] D. Tweed and T. Le-Ngoc, "Dynamic resource allocation for uplink MIMO NOMA VWN with imperfect SIC," in *IEEE Int. Conf. Commun.*, May 2018, pp. 1–6.
- [10] Y. Li and G. Amarasuriya, "NOMA-aided massive MIMO downlink with distributed antenna arrays," in *IEEE Int. Conf. Commun.*, May 2019, pp. 1–7.
- [11] X. Li, M. Liu, C. Deng, P. T. Mathiopoulos, Z. Ding, and Y. Liu, "Full-duplex cooperative NOMA relaying systems with I/Q imbalance and imperfect SIC," *IEEE Wireless Commun. Lett.*, pp. 1–1, 2019.
- [12] X. Li, J. Li, Y. Liu, Z. Ding, and A. Nallanathan, "Residual transceiver hardware impairments on cooperative NOMA networks," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2019.
- [13] J. Kang, I. Kim, and C. Chun, "Deep learning-based MIMO-NOMA with imperfect SIC decoding," *IEEE Syst. J.*, pp. 1–4, 2019.
- [14] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Proces. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.
- [15] Y. Liu, M. ElKashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Wireless Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Jul. 2016.
- [16] J. A. Oviedo and H. R. Sadjadpour, "A fair power allocation approach to NOMA in multiuser SISO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7974–7985, Sep. 2017.
- [17] M. M. Al-Wani, A. Sali, N. K. Noordin, S. J. Hashim, C. Y. Leow, and I. Krikidis, "Robust beamforming and user clustering for guaranteed fairness in downlink NOMA with partial feedback," *IEEE Access*, vol. 7, pp. 121 599–121 611, Aug. 2019.
- [18] M. M. Al-Wani, A. Sali, B. M. Ali, A. A. Salah, K. Navaie, C. Y. Leow, N. K. Noordin, and S. J. Hashim, "On short term fairness and throughput of user clustering for downlink non-orthogonal multiple access system," in *IEEE Veh. Technol. Conf.*, Apr. 2019, pp. 1–6.
- [19] Z. Xiao, L. Zhu, Z. Gao, D. O. Wu, and X. Xia, "User fairness non-orthogonal multiple access (NOMA) for millimeter-wave communications with analog beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3411–3423, Jul. 2019.
- [20] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Elsevier/Academic Press, Amsterdam, 2007.
- [21] A. Adhikary, J. Nam, J. Ahn, and G. Caire, "Joint spatial division and multiplexing - The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.
- [22] H. Yin, D. Gesbert, and L. Cottatellucci, "Dealing with interference in distributed large-Scale MIMO systems: A statistical approach," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 942–953, Oct. 2014.
- [23] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Wiley Series in Probability and Statistics, Aug. s2003.
- [24] 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.
- [25] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.