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

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

DOI: 10.1109/TWC.2020.3000192

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, *Fellow*, *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 American College of Greece, Greece (email: cpapadias@acg.edu).

Part of this paper has been submitted in IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2020), London, UK.

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

Advances in technologies and the rise of new applications, such as unmanned vehicles, smart homes, smart grid, and massive sensor networks, are triggering an accelerated growth in the number of devices connected to communication systems. As an attempt to support this explosive trend, the 5th generation of wireless networks (5G) is being developed, and the first commercial systems have been deployed worldwide. 5G and beyond networks are expected to support a variety of demanding requisites, going from massive connectivity and ultra-low latency to improved user fairness [1]. Massive multiple-input-multiple-output (MIMO) is being credited as one of the key enabling components of 5G [2]. In particular, by employing a very large number of antennas and exploiting the space domain to multiplex different users, the massive MIMO technology has the potential to reduce system latency and to provide remarkable connectivity gains. Power-domain non-orthogonal multiple access (NOMA) is another promising technology for the future-generation wireless systems that allow multiple users to be served in parallel within the same frequency and time slot. The relying 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. With such features, NOMA can also provide massive connectivity capabilities and a reduction in latency to the network.

If the NOMA technique is applied to massive MIMO, the achievable spectral and connectivity improvements are shown to be even greater [3]–[5]. 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 [6]. To improve the average performance of the weaker users, one could decrease as much as possible the power allocated to the strong ones so that a certain degree of fairness could be achieved [7]. However, such a strategy can severely impact the system sum-rate, and, due to the random behavior of the wireless channels, it will also result in unequal data rates. This characteristic can be detrimental to certain emerging applications. For example, in the upcoming industrial internet of things, it can be very important that all terminals experience similar data rates [8]. In such scenarios, the employment of fixed allocation policies can make some of the terminals not able to meet their minimum requirements and result in a poor network performance. Besides, 5G and beyond are expected to support the concept of network slicing, which is to create isolated logical networks, i.e., slices, each dedicated to a subset of terminals with specific requirements [9]. Since within each slice users are expected to share identical requisites, it will be crucial to perform a fair distribution of resources, which is a feature that fixed power allocation is not capable of providing. Therefore, more sophisticated and adaptive power allocation strategies are important and necessary to guarantee fairness in future MIMO-NOMA networks. 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 [10], [11]. 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 [10], 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 [12] 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. In [13], the authors analyzed the outage probability and minimized the total transmit power of a multi-carrier NOMA system by modeling the SIC error propagation as a complex Gaussian random variable. Complementary geometric programming and arithmetic geometric mean approximation techniques were used to transform the non-convex formulated problem into a convex one. The heterogeneous networks case was considered in [14]. By taking into consideration various sources of interference, such as inter-cell interference, power disparity, and imperfect SIC, the work proposed user clustering and power-bandwidth allocation algorithms. The impact of imperfect SIC in the uplink of MIMO-NOMA systems employing the concept of virtualized wireless networks (VWN) was studied in [15], 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 [16]. In this contribution, a closed-form expression for the ergodic sum-rate was derived. In [17], 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 α-μ fading channels in cooperative NOMA networks with hardware impairments was considered in [18], and the work in [19] 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 [20] 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 [21], 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 [22]. 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 [23] and [24] proposed user clustering algorithms based on proportional fairness to balance between throughput and fairness. [23] 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 [6], the authors developed dynamic resource allocation policies, which are optimally obtained via Lagrangian dual decomposition. The millimeter-wave MIMO-NOMA case was addressed in [25], 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, and 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 ith element of a vector $\mathbf{a}$ is denoted by $[\mathbf{a}]_i$ and the (ij) 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 tr$\{\mathbf{A}\}$. In 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 [26, eq. (8.310.1)], $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function [26, eq. (8.350.1)], and Ei$(\cdot)$ corresponds to the exponential integral [26, 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 a 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 [27]. Within each cluster, the BS subdivides the users into G smaller sub-groups, each

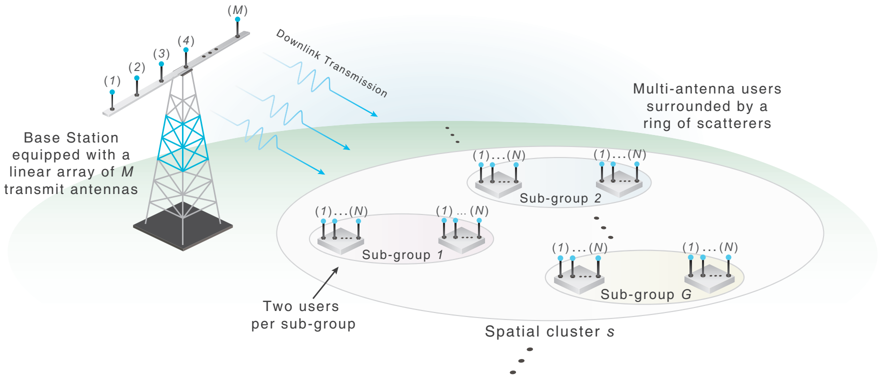

Fig. 1: System model. Users within each spatial cluster are organized into multiple sub-groups.

one containing 2 users[1], 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 [27], the channel matrix for the kth user in the gth sub-group in the sth cluster, can be expressed by

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

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 the first r_s^* dominant eigenvalues of $\mathbf{R}_s$, sorted out in decreasing order, in which $r_s^* \leq r_s$. $\mathbf{U}_s \in \mathbb{C}^{M \times r_s^*}$ 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 gth sub-group from the BS, η is the path-loss exponent, and Φ is a gain parameter that is adjusted based on the desired performance of the receivers [28]. Moreover, all users confined within the sth cluster are assumed to share identical[2] covariance matrices $\mathbf{R}_s$, whose entries can be generated by [27]

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

where δ_s and φ_s are, respectively, the angular spread and the azimuth angle of the sth spatial cluster, ϕ 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 $1 \leq m, m' \leq M$.

[1] Given that SIC is an interference-limited technique, the consideration of a large number of users per sub-group can lead to performance degradation (due to decoding error propagation), increase in detection and hardware complexities, and higher energy consumption. Therefore, small sub-groups (usually of two users) are preferable in practical downlink NOMA systems [7].

[2] The assumption of users sharing identical covariance matrices cannot be exactly satisfied in real-world scenarios. However, as stated in [27], if users within the same cluster of scatterers are grouped properly, this condition can be efficiently approximated. Even though user grouping is an important topic and an active area of research [14], it goes beyond the scope of this work.

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

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

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 σ_n^2. The variable α_{nij} denotes the power coefficient, and x_{nij} is the symbol intended for the jth user in the ith sub-group at the nth 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 [27]. To this end, we define $\mathbf{U}_s^- = [\mathbf{U}_1, \cdots, \mathbf{U}_{s-1}, \mathbf{U}_{s+1}, \cdots, \mathbf{U}_S]$ and denote its last $M - (S-1)r_s^*$ left eigenvectors by $\mathbf{E}_s^0 \in \mathbb{C}^{M \times M-(S-1)r_s^*}$, which corresponds to the vanishing eigenmodes of $\mathbf{U}_s^-$. As a result, due to the dimension of $\mathbf{E}_s^0$, the constraint $M > (S-1)r_s^*$ must be satisfied.

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

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

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

It is noteworthy that the number of dominant eigenvalues r_s^* will determine the amount of interference that will leak from other clusters[3]. Specifically, the inter-cluster interference will

[3] As discussed in [4] and [27], finding the optimum value of r_s^* depends on the parameters and requisites of each specific system, such as the number of antennas, desired number of clusters, and maximum interference level, which is not a trivial task and goes beyond the scope of this work. We configure this parameter not aiming its optimally, but to achieve a good system performance and to satisfy the beamforming design constraints.

approach zero as r_s^* approaches to the rank r_s, such that, when $r_s^* = r_s$, $\mathbf{H}_{sgk}^H \mathbf{B}_{s'} = \mathbf{0}, \forall s \neq s'$. However, since part of the eigenvalues can be very small, i.e., ≈ 0, the extreme choice $r_s^* = r_s$ is usually not efficient, and it does not result in significant performance improvements [4], [27]. In addition, given that $S < M/r_s^* + 1$, increasing r_s^* will reduce the maximum number of clusters S. In particular, due to the beamforming constraints, we set the dominant eigenvalues parameter to $r_s^* = \min\{r_s, \lfloor (M-V)/(S-1) \rfloor\}$.

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

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

Note that the above design associates the gth effective data stream to the gth NOMA sub-group, and it does not modify by any form the data messages. Besides, to construct the beamformers, it is not necessary to acquire the fast fading channel matrices $\mathbf{G}_{sgk}$ at the BS, that is, only the channel covariance matrices are required. And given that $\mathbf{R}_s$ varies slowly, the users only need to measure it once after several coherence intervals. Once this statistical information is obtained, users can feed it back to the BS with low feedback overhead [27]. As a result, it is reasonable to assume that all the channel information can be accurately estimated at the users' side through downlink training techniques[4]. Therefore, as in [3]–[5], we consider that $\mathbf{R}_s$ is known in the system (at both BS and users), and that $\mathbf{G}_{sgk}$ can be estimated perfectly[5] by the users' terminals. Also, 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 $\mathbf{B}_s$

[4] Channel estimation and acquisition are critical in massive MIMO and are topics of ongoing interest in the literature (see [29] and references therein, for example). However, the investigation of such topics goes beyond the scope of this work.

[5] In practice, the estimation of the fast varying channel matrices, $\mathbf{G}_{sgk}$, is usually not perfect. Therefore, the investigation of the impact of channel estimation errors on our proposed massive MIMO-NOMA design arises as an interesting future direction.

successfully suppresses all inter-cluster interferences, the superposed signal observed by the kth user, $k \in \{1, 2\}$, in the gth sub-group can be rewritten as

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

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. Note that, in order to construct the zero-forcing receiver, the users need to have access also to the beamforming matrices. Since, in our design, $\mathbf{B}$ is built based only on the channel covariance matrices, which vary slowly, such beamforming information can be efficiently informed back to the users, imposing low overhead.

After the received signal has been filtered through $\mathbf{H}_{gk}^{\dagger}$, the kth user in the gth group achieves the following data vector

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

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 gth NOMA sub-group to apply SIC to their corresponding superposed symbol, i.e., the gth 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 [3]–[5]. 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 achieve an imperfect estimation of the symbols intended to the weak users and suffer from residual interference. On these grounds, as in [10], [14], [15], we model the effects of imperfect SIC as a function of the interfering power. More specifically, the kth user in the gth sub-group will recover the following symbol

$$\hat{x}_{gk} = \begin{cases} \underset{\substack{\uparrow \\ \text{symbol of interest}}}{\sqrt{\alpha_{g1}}x_{g1}} + \underset{\substack{\uparrow \\ \text{interference}}}{\sqrt{\alpha_{g2}}x_{g2}} + \underset{\substack{\uparrow \\ \text{noise}}}{[\mathbf{H}_{g1}^{\dagger}\mathbf{n}_{g1}]_g}, & \text{if } k = 1, \\ \underset{\substack{\uparrow \\ \text{symbol of interest}}}{\sqrt{\alpha_{g2}}x_{g2}} + \underset{\substack{\uparrow \\ \text{residual SIC interference}}}{\sqrt{\mu\alpha_{g1}}x_{g1}} + \underset{\substack{\uparrow \\ \text{noise}}}{[\mathbf{H}_{g2}^{\dagger}\mathbf{n}_{g2}]_g}, & \text{if } k = 2, \end{cases} \tag{8}$$

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. As demonstrated in [30], [31], the error factor μ can be easily calculated at the receivers by dividing the variance of the interference term, which can be obtained by averaging a large number of samples, by the power allocated to the interfering user, e.g., supposing that $\tilde{x}_{g1}$ is the symbol intended for the user 1 estimated at user 2 during the SIC process, the error factor can be computed as[6] $\mu = \mathbb{E}\{|\sqrt{\alpha_{g1}}(x_1 - \tilde{x}_1)|^2\}/\alpha_{g1}$. In practice, the value of μ will depend on factors such as the type of the receiver, channel characteristics, and hardware sensibility [14], [15]. 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 kth user, $1 \leq k \leq 2$, when decoding the symbol that belongs to the ith user, $1 \leq i \leq 2$, is obtained by

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

[6] Note that in our analysis, we model μ as a deterministic parameter. However, since the residual SIC interference term in (8) can be approximated by a Gaussian distribution [30], μ can also be modeled as a chi-squared random variable, as in [13]. This possibility arises as a potential extension of this work.

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 $\alpha_{gi}^{\star}$ corresponds to the interference power, which is given by

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

Proof: Please, see Appendix A. ■

Observe that, since users are ordered based on their effective channel gains, to enable NOMA, they are required to feed the gains ϱ_{gk} back to the BS at each coherence interval. However, since ϱ_{gk} is just a scalar parameter, such a task will result in low additional overhead only [32].

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 (A-1), ϱ_{gk} is the inverse of the gth element on the main diagonal of the following matrix

$$\begin{aligned}\mathbf{H}_{gk}^{\dagger}\mathbf{H}_{gk}^{\dagger H} &= [(\mathbf{H}_{gk}^{H}\mathbf{B})^{H}\mathbf{H}_{gk}^{H}\mathbf{B}]^{-1}(\mathbf{H}_{gk}^{H}\mathbf{B})^{H}\mathbf{H}_{gk}^{H}\mathbf{B}([(\mathbf{H}_{gk}^{H}\mathbf{B})^{H}\mathbf{H}_{gk}^{H}\mathbf{B}]^{-1})^{H} = (\mathbf{B}^{H}\mathbf{H}_{gk}\mathbf{H}_{gk}^{H}\mathbf{B})^{-1} \\ &= \left(\mathbf{B}^{H}\Phi d_g^{-\eta}(\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{G}_{gk})(\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{G}_{gk})^{H}\mathbf{B}\right)^{-1} = \frac{1}{\Phi d_g^{-\eta}}\left(\mathbf{B}^{H}\mathbf{R}\mathbf{B}\right)^{-1} \in \mathbb{C}^{V\times V}.\end{aligned} \tag{11}$$

As demonstrated in [3], [4], since $\mathbf{G}_{gk}$ consists of a complex Gaussian distributed matrix, the resulting matrix $\left(\mathbf{B}^{H}\mathbf{R}\mathbf{B}\right)^{-1}$ 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 [33]

$$f_{\varrho_{gk}}(x) = K\binom{K-1}{k-1}\sum_{i=0}^{K-k}(-1)^{i}\binom{K-k}{i}\tilde{f}_{\varrho_{gk}}(x)\tilde{F}_{\varrho_{gk}}(x)^{k-1+i}, \tag{12}$$

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

$$f_{\varrho_{g1}}(x) = 2[\tilde{f}_{\varrho_{gk}}(x) - \tilde{f}_{\varrho_{gk}}(x)\tilde{F}_{\varrho_{gk}}(x)] = \frac{2\beta_g^{\vartheta}}{\Gamma(\vartheta)}\left[x^{\vartheta-1}e^{-\beta_g x} - x^{\vartheta-1}e^{-\beta_g x}\frac{\gamma(\vartheta,\beta_g x)}{\Gamma(\vartheta)}\right], \tag{13}$$

and for user 2 as

$$f_{\varrho_{g2}}(x) = 2\tilde{f}_{\varrho_{gk}}(x)\tilde{F}_{\varrho_{gk}}(x) = \frac{2\beta_g^{\vartheta}}{\Gamma(\vartheta)} x^{\vartheta-1} e^{-\beta_g x} \frac{\gamma(\vartheta, \beta_g x)}{\Gamma(\vartheta)}, \tag{14}$$

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 kth user, in the gth 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 ith user, $i \leq k \in \{1, 2\}$, and it can be defined as [3], [4]

$$P_{gk} = P[\log_2(1 + \gamma_{gk}^i) < T_{gi}], \quad \forall i = 1, \cdots, k. \tag{15}$$

Note from (15) that user 1 (weak user) will face an outage event only when its achieved data rate is not enough to satisfy its own target rate, i.e., when $\log_2(1 + \gamma_{g1}^1) < T_{g1}$. While user 2 (strong user) will experience outage either if $\log_2(1+\gamma_{g2}^1) < T_{g1}$ or $\log_2(1+\gamma_{g2}^2) < T_{g2}$. Also see that, due to (10), the outage probability of the strong user will be also impacted by residual SIC interference. Closed-form expressions for the outage probability are provided in the following proposition.

Proposition I: 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:

$$P_{g1} = \begin{cases} \frac{2\gamma(\vartheta, \rho^{-1}\beta_g \mathcal{L}_{g1})}{\Gamma(\vartheta)} - \left[\frac{\gamma(\vartheta, \rho^{-1}\beta_g \mathcal{L}_{g1})}{\Gamma(\vartheta)}\right]^2, & \text{if} \quad \mathcal{L}_{g1} \geq 0, \\ 1, & \text{otherwise.} \end{cases} \tag{16}$$

- For user 2:

$$P_{g2} = \begin{cases} \left[\frac{\gamma(\vartheta, \rho^{-1}\beta_g \max\{\mathcal{L}_{g1}, \mathcal{L}_{g2}\})}{\Gamma(\vartheta)}\right]^2, & \text{if} \quad \min\{\mathcal{L}_{g1}, \mathcal{L}_{g2}\} \geq 0, \\ 1, & \text{otherwise,} \end{cases} \tag{17}$$

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 B. ■

D. Ergodic Rates

In this subsection, we analyze the ergodic rates experienced by each user within the sub-groups. In particular, it is considered that the strong user cannot decode perfectly the symbol intended to the weak user. As a consequence, its achievable rate, which is resulted from the

SINR observed while decoding its own data symbol, will be impacted by residual interference. Under this consideration, the instantaneous data rate achieved by the first user can be written as

$$R_{g1} = \log_2(1+\gamma_{g1}^{1}) = \log_2\left(1+\frac{\rho\varrho_{g1}\alpha_{g1}}{\rho\varrho_{g1}\alpha_{g2}+1}\right), \tag{18}$$

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

$$R_{g2} = \log_2(1+\gamma_{g2}^{2}) = \log_2\left(1+\frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu\alpha_{g1}+1}\right). \tag{19}$$

From (18) and (19), exact closed-form expressions for the users' ergodic rates will be derived, which are presented in Proposition II.

Proposition II: In presence of residual error propagation from imperfect SIC, exact closed-form expressions for the ergodic rates of users 1 and 2 can be obtained as follows:

- For user 1:

$$\bar{R}_{g1} = \xi_1(\kappa_{g1}) - \xi_1(\tilde{\kappa}_{g1}), \tag{20}$$

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

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

- For user 2:

$$\bar{R}_{g2} = \begin{cases} \xi_2(\kappa_{g2}) - \xi_2(\tilde{\kappa}_{g2}), & \textit{if} \quad \mu > 0, \\ \xi_2(\kappa_{g2}), & \textit{if} \quad \mu = 0, \end{cases} \tag{21}$$

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

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

Proof: Please, see Appendix C. ■

Note that as long as we have SIC error propagation, there will be a negative term in (21), i.e., $-\xi_2(\tilde{\kappa}_{g2})$, which indicates degradation in the rate of the strong user. In fact, numerical results show that $\xi_2(\kappa)$ is an increasing function of the SNR ρ. However, when $\rho \to \infty$, the negative term in (21) will make the expression to converge to a saturation point, which means that the achievable data rate at the strong user will be always capped if $\mu \neq 0$. Similar behavior can be observed in the expression for the weak user, in (20). Nevertheless, since it does not perform SIC, its rate ceiling will be independent of μ. More details are provided in Section V, where we perform an insightful numerical analysis.

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], [32], it can lead to low data rates at the weaker users. As mentioned before, such an unbalanced performance can be very detrimental to certain 5G applications with strict fairness requirements. 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. Specifically, the BS needs to distribute the power resources between the two users within the sub-groups in such a way that their rates become balanced. Given that the weak users face the worst channel conditions, we must ensure that their rates are greater than or equal to that achieved by the stronger ones, i.e., $\log_2\left(1+\frac{\rho\varrho_{g1}\alpha_{g1}}{\rho\varrho_{g1}\alpha_{g2}+1}\right) \geq \log_2\left(1+\frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu\alpha_{g1}+1}\right)$. With this in mind, our objective can be accomplished with the following optimization problem

$$\max_{\alpha_{g1},\alpha_{g2}} \{R_{g2}\} \tag{22a}$$

$$\text{s.t. } R_{g1} \geq R_{g2}, \tag{22b}$$

$$\alpha_{g1}+\alpha_{g2}=\bar{\alpha}_g, \tag{22c}$$

$$\alpha_{g1} \geq 0, \alpha_{g2} \geq 0, \tag{22d}$$

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

Given that $\log_2(\cdot)$ is a monotonic increasing function, from the constraint (22b), 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 (22) can be simplified to

$$\max_{\alpha_{g1},\alpha_{g2}} \left\{ \log_2\left(1+\frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu\alpha_{g1}+1}\right)\right\} \tag{23a}$$

$$\text{s.t.}\ \ \varrho_{g1}\varrho_{g2}\alpha_{g2}^2 - \mu\varrho_{g1}\varrho_{g2}\alpha_{g1}^2 + \varrho_{g2}\alpha_{g2}\rho^{-1} \leq \varrho_{g1}\alpha_{g1}\rho^{-1}, \tag{23b}$$

$$\alpha_{g1}+\alpha_{g2} = \bar{\alpha}_g, \tag{23c}$$

$$\alpha_{g1} \geq 0, \alpha_{g2} \geq 0. \tag{23d}$$

Then, by letting $\alpha_{g1} = \bar{\alpha}_g - \alpha_{g2}$, the constraint in (23b) becomes $\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) \leq 0$, and (23) can be rewritten as

$$\max_{\alpha_{g2}} \left\{ \log_2\left(1+\frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu(\bar{\alpha}_g-\alpha_{g2})+1}\right)\right\} \tag{24a}$$

$$\begin{aligned}\text{s.t.}\ \ &\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)\\ &-(\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) \leq 0,\end{aligned} \tag{24b}$$

$$\alpha_{g2} \geq 0. \tag{24c}$$

By using the fact that $\log_2\left(1+\frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu(\bar{\alpha}_g-\alpha_{g2})+1}\right)$ increases monotonically with α_{g2}, the problem (24) can be reduced to the optimization of only α_{g2}, as follows

$$\min_{\alpha_{g2}} -\{\alpha_{g2}\} \tag{25a}$$

$$\begin{aligned}\text{s.t.}\ \ &\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)\\ &-(\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) \leq 0,\end{aligned} \tag{25b}$$

$$-\alpha_{g2} \leq 0, \tag{25c}$$

The optimal solution for (25) is given in the following proposition.

Proposition III: The optimization problem in (25) is convex and, consequently, has a global optimal solution, which is given in closed-form by

$$\alpha_{g2}^* = \begin{cases} \frac{1}{2(\varrho_{g1}\varrho_{g2}-\mu\varrho_{g1}\varrho_{g2})}\left[-(\varrho_{g1}\rho^{-1}+\varrho_{g2}\rho^{-1}+2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)\right. \\ \left. +\sqrt{(\varrho_{g1}\rho^{-1}+\varrho_{g2}\rho^{-1}+2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)^2+4\varrho_{g1}\varrho_{g2}(1-\mu)(\varrho_{g1}\rho^{-1}\bar{\alpha}_g+\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)}\,\right], & \textit{if} \quad 0 \leq \mu < 1, \\ (\varrho_{g1}\rho^{-1}\bar{\alpha}_g+\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)/(\varrho_{g1}\rho^{-1}+\varrho_{g2}\rho^{-1}+2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g), & \textit{if} \quad \mu = 1, \end{cases} \tag{26}$$

and

$$\alpha_{g1}^* = \bar{\alpha}_g - \alpha_{g2}^*. \tag{27}$$

Proof: Please, see Appendix D. ■

Note that the calculation of the solution above requires knowledge of the error propagation factor μ and the effective channel gains $\varrho_{gk}, \forall g, k$. Since the gains ϱ_{gk} are already needed at the BS for enabling NOMA, they can be directly used in power allocation. On the other hand, as clarified in Subsection III-A, μ can be estimated at the receiver's side through long-term measurements. Therefore, this scalar parameter can be fed back to the BS with little impact on the feedback overhead, and the optimum power allocation in (26) and (27) can be computed.

B. Providing Fairness Among Sub-Groups

In the last subsection, 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-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 subsection, we develop a strategy for improving fairness also among different sub-groups. For achieving this goal, we propose an iterative algorithm that enables the BS to provide 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 Δ_α, from the best to the worst sub-group, and to use the dynamic power allocation derived in the last subsection 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 (26) and (27) 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 Δ_α 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 ϵ. Observe that, since this iterative approach is computed based on the effective channel gains, $\varrho_{gk}, \forall g, k$, the BS will be required to execute Algorithm 1 at each coherence interval.

The value of Δ_α 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. Δ_α is calculated in the following proposition.

Algorithm 1: Iterative Algorithm for Fairness Among Sub-Groups

```
Input: $\epsilon, \rho, \varrho_{g1}, \varrho_{g2}$.
Set the initial available power to $\bar{\alpha}_g = 1, \forall g$;
do
    for $g = 1$ to $G$ do
        Calculate $\alpha^*_{g2}$ and $\alpha^*_{g1}$ using (26) and (27), respectively;
        Calculate the sub-group's sum-rate by $\mathcal{R}_g = R_{g1} + R_{g2}$;
    end
    $\hat{g} = \textbf{argmax}(\mathcal{R}_g : \forall g \in \{1, 2, \cdots, G\})$;
    $\check{g} = \textbf{argmin}(\mathcal{R}_g : \forall g \in \{1, 2, \cdots, G\})$;
    Calculate $\Delta_\alpha$ using (28);
    Update $\bar{\alpha}_{\hat{g}} = \bar{\alpha}_{\hat{g}} - \Delta_\alpha$;
    Update $\bar{\alpha}_{\check{g}} = \bar{\alpha}_{\check{g}} + \Delta_\alpha$;
    $\epsilon^* = \mathcal{R}_{\hat{g}} - \mathcal{R}_{\check{g}}$;
while $\epsilon^* > \epsilon$;
```

Proposition IV: The amount of power Δ_α needed to balance the rates of the strongest users from the sub-groups with the highest and lowest sum-rate, assuming perfect SIC[7], is given by

$$\Delta_\alpha = \frac{-A_2 \pm \sqrt{A_2^2 - 4A_1A_3}}{2A_1}, \tag{28}$$

where

$$A_1 = 4\varrho_{\check{g}1}^2 \varrho_{\check{g}2} \rho^{-1} + 4K_1, \qquad A_2 = 2A_1K_3 + 16K_2K_1,$$

$$A_3 = K_3^2 - 4K_2 \frac{\varrho_{\check{g}1}^2}{\varrho_{\hat{g}1}^2} (\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1})^2 - 16K_2K_1\bar{\alpha}_{\hat{g}},$$

and

$$K_1 = \frac{\varrho_{\check{g}1}^2}{\varrho_{\hat{g}1}^2} \varrho_{\hat{g}1}^2 \varrho_{\hat{g}2} \rho^{-1}, \qquad K_2 = (\varrho_{\check{g}1}\rho^{-1} + \varrho_{\check{g}2}\rho^{-1} - \frac{\varrho_{\check{g}1}}{\varrho_{\hat{g}1}}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1}))^2,$$

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

Proof: Please, see Appendix E. ■

C. Computational Complexity of Algorithm 1

In this subsection, we provide the worst-case computational complexity of the proposed power allocation solution shown in Algorithm 1. As in [34], we consider summations, multiplications, comparisons, and square-roots as the most relevant and time-consuming operations. The proposed algorithm is iterative, and the number of iterations, denoted here by I, depends on the

[7] Due to the complicated expression in (26), obtaining a closed-form solution for Δ_α considering imperfect SIC becomes a very challenging task. Thus, a different approach for computing Δ_α is necessary when $\mu \neq 0$, but this is left for future works.

predefined threshold ϵ. This shows a clear trade-off between the accuracy of the solutions and the computational complexity. Between lines 3 and 6 of Algorithm 1, at each iteration, the power allocation coefficients for strong and weak users are computed for each sub-group according to equations (26) and (27) and, then, the sum-rate of each sub-group is calculated. Note that, in (26), the variables ϱ_{g_1}, ϱ_{g_2}, μ and ρ do not change along the iterations. In fact, only the sub-group's power budget, $\bar{\alpha}$, changes. As a result, the number of summations, multiplications, and square-root operations performed at each algorithm iteration are $6G$, $8G$, and G, respectively. In line 5, we have G summations per iteration. In lines 7 and 8, we have the search for the maximum and minimum sub-group's data rate, respectively. Thereby, in each line, the algorithm performs $G-1$ comparisons. In the calculation of (28), in line 9, the number of summations, multiplications, and square-root operations are 7, 8, and 1 per iteration, respectively. Lastly, in lines from 10 to 12, we have 3 summations per iteration. To sum up, the total number of operations for a given number of iterations, I, is $17IG+17I$. Consequently, we can conclude that the worst-case computational complexity of Algorithm 1 is $\mathcal{O}\left(IG\right)$.

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 $\delta=10°$, 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 $\varphi=7°$, 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 $\Phi=4\times10^4$. Moreover, when fixed power allocation is considered, the power coefficients of 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.

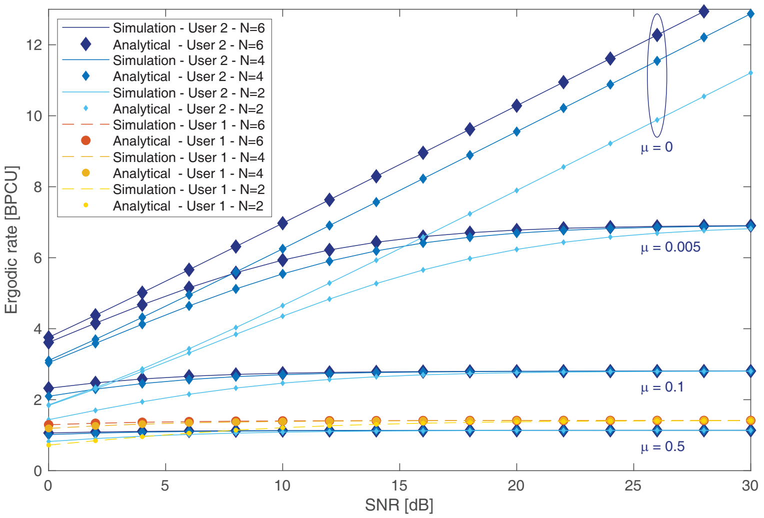

Fig. 2: Impact of imperfect SIC on the users' ergodic rates for different number of receive antennas.

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. 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 μ 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 $\rho \to \infty$, if $\mu > 0$, $R_{g2} \to \log_2\left(1 + \frac{\alpha_{g2}}{\mu\alpha_{g1}}\right)$. Therefore, if there is some residual SIC error and $\alpha_{g1} > 0$, there will be always a rate ceiling for the strong user, as anticipated in the Subsection III-D.

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

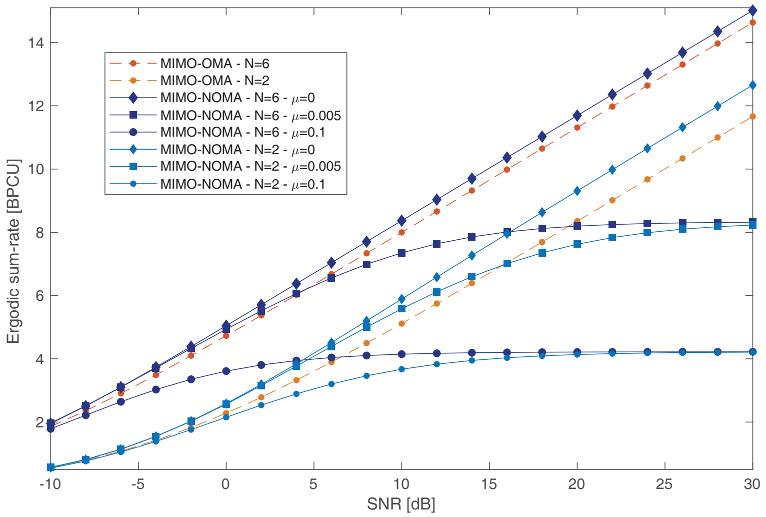

Fig. 3: Simulated ergodic sum-rate curves for massive MIMO-NOMA system with imperfect SIC and conventional massive MIMO-OMA counterpart.

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 μ 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. This fast degradation happens because the maximum achievable rate of the strong user shifts very close to its target data rate when μ reaches values above 0.36, i.e., $R_{g2} \rightarrow \log_2\left(1 + \frac{\alpha_{g2}}{0.36\alpha_{g1}}\right) \approx 1.4$ BPCU when $\rho \rightarrow \infty$. As a result, from low to moderate SNR ranges, the strong user will face an increased probability of achieving a throughput lower than its target rate, which explains the observed behavior. 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

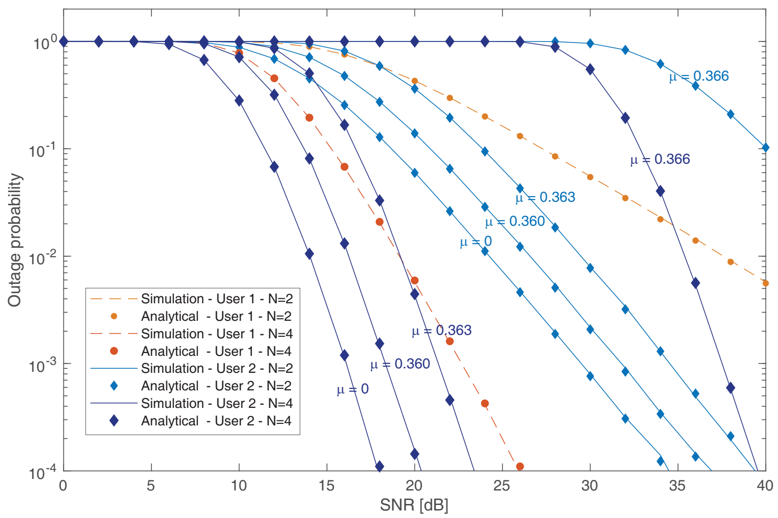

Fig. 4: Impact of imperfect SIC on the outage probability for different number of receive antennas ($T_1 = T_2 = 1.4$ BPCU).

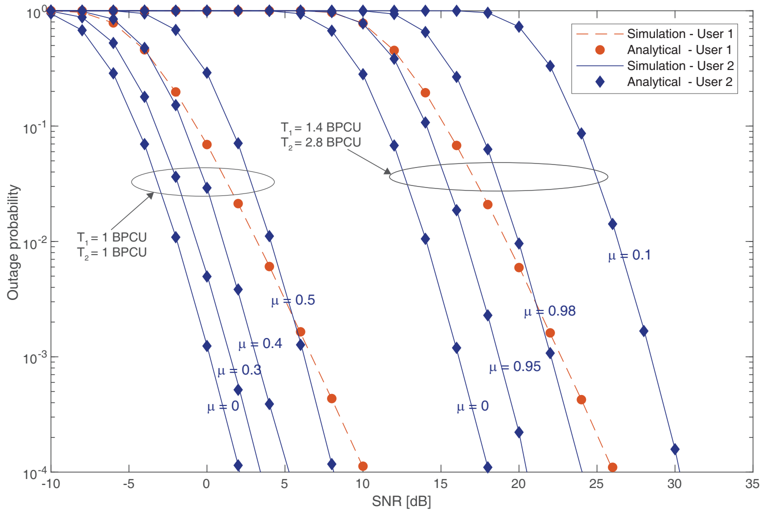

Fig. 5: Outage probability versus transmit SNR for different SIC interference levels and different target rates ($N = 4$).

setting the target rates of both users to 1 BPCU, when μ 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.

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

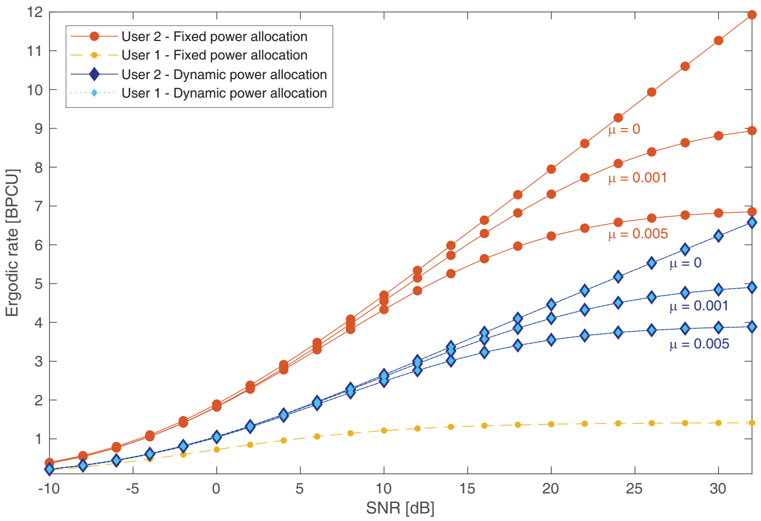

Fig. 6: Ergodic rates for strong and weak users in massive MIMO-NOMA system with dynamic and fixed power allocation policies ($N = 2$).

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 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.

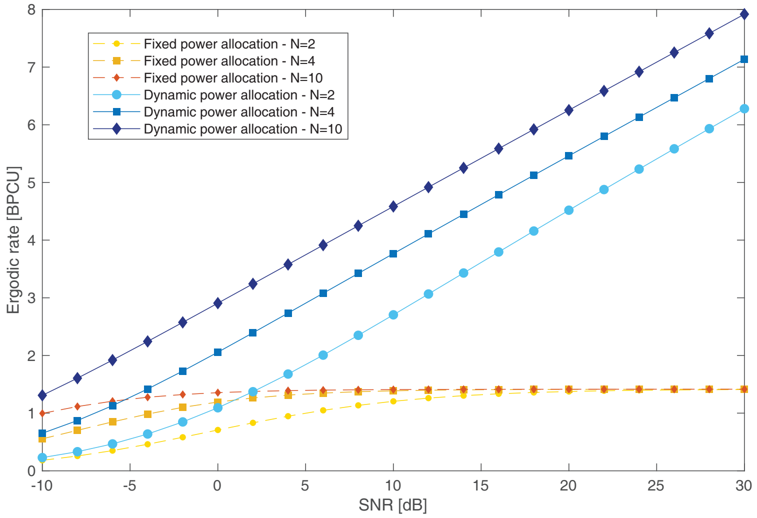

Fig. 7: Ergodic rates for the weak user in massive MIMO-NOMA system with dynamic and fixed power allocation policies ($\mu = 0$).

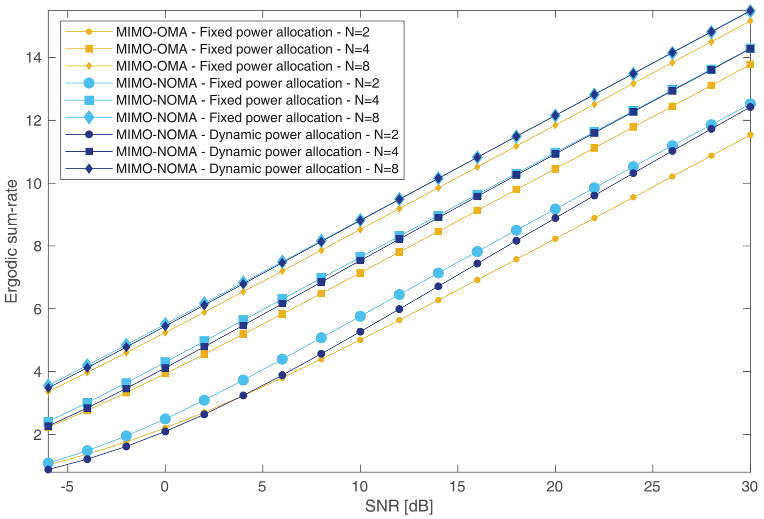

Fig. 8: Ergodic sum-rate curves for massive MIMO-NOMA and MIMO-OMA systems with dynamic and fixed power allocation policies ($\mu = 0$).

This is because, in order to enhance fairness, the optimization problem in (22) decreases the 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

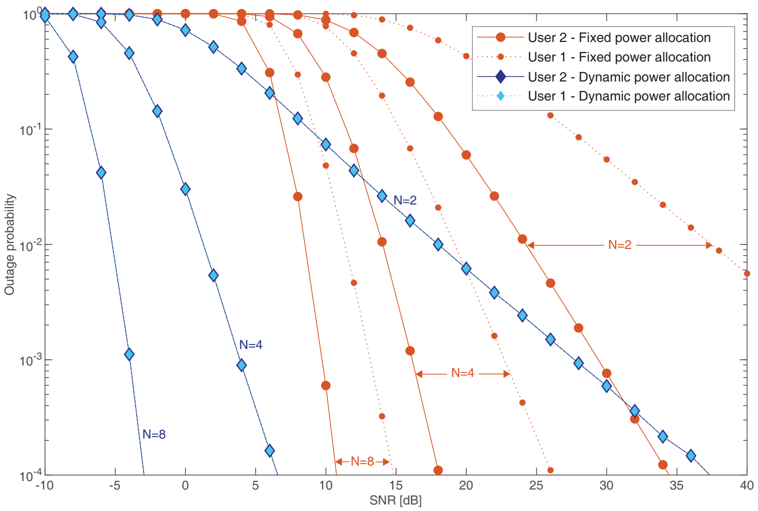

Fig. 9: Outage probabilities achieved with dynamic and fixed power allocation policies in massive MIMO-NOMA systems ($T_1 = T_2 = 1.4$ BPCU; $\mu = 0$).

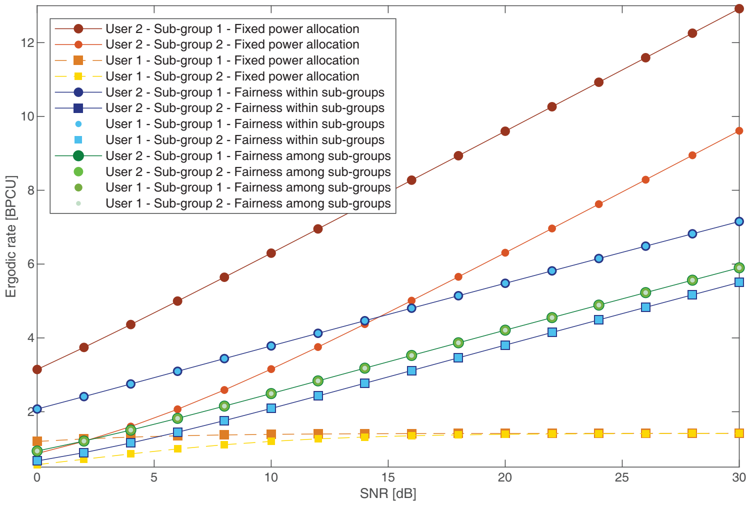

Fig. 10: Ergodic rates for users in different sub-groups employing different power allocation policies ($d_1 = 115$m; $d_2 = 150$m; $N = 4$; $\mu = 0$).

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 remarkable gain of 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 (22) 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 Lemma I

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

$$\gamma_{gk}^{1} = \frac{|\sqrt{\alpha_{g1}}x_{g1}|^2}{|\sqrt{\alpha_{g2}}x_{g2}|^2 + |[\mathbf{H}_{gk}^{\dagger}\mathbf{n}_{gk}]_g|^2} = \frac{\frac{1}{[\mathbf{H}_{gk}^{\dagger}\mathbf{H}_{gk}^{\dagger H}]_{gg}}\alpha_{g1}}{\frac{1}{[\mathbf{H}_{gk}^{\dagger}\mathbf{H}_{gk}^{\dagger H}]_{gg}}\alpha_{g2} + \sigma_n^2}. \tag{A-1}$$

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, (A-1) can be rewritten as

$$\gamma_{gk}^{1} = \frac{\rho\varrho_{gk}\alpha_{g1}}{\rho\varrho_{gk}\alpha_{g2} + 1}, \qquad \text{for} \quad 1 \leq k \leq 2. \tag{A-2}$$

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

$$\gamma_{g2}^{2} = \frac{\rho\varrho_{g2}\alpha_{g2}}{\rho\varrho_{g2}\mu\alpha_{g1} + 1}. \tag{A-3}$$

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

$$\alpha_{gi}^{\star} = \begin{cases} \alpha_{g2}, & \text{for} \quad i = 1, \\ \mu\alpha_{g1}, & \text{for} \quad i = k = 2. \end{cases} \tag{A-4}$$

Lastly, by combining (A-2), (A-3), and (A-4), the final SINR expression is obtained, as follows

$$\gamma_{gk}^{i} = \frac{\rho\varrho_{gk}\alpha_{gi}}{\rho\varrho_{gk}\alpha_{gi}^{\star} + 1}, \qquad \text{for} \quad 1 \leq i \leq k \leq 2, \tag{A-5}$$

which completes the proof. ■

Appendix B

Proof of Proposition I

The outage probability expression in (15) can be rewritten as follows

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

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

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

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

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

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, (B-2) and (B-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 (13) and (14), in which, for user 1, it results in

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

while, for user 2, the following is obtained

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

which completes the proof. ■

APPENDIX C
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 (18), as follows

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

and, for user 2, as

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

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 (C-1), the ergodic rate for user 1 can be expressed as

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

Then, by invoking the PDF of ϱ_{g1}, provided in (13), $\xi_1(\kappa)$ can be calculated as follows

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

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

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

Lastly, after some algebraic manipulation and applying results from [26], we achieve the desired solution, as follows

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

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

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

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

$$\xi_2(\kappa) = \frac{2\beta_g^{\vartheta}}{\ln(2)\Gamma(\vartheta)} \int_0^{\infty} \ln(1+\kappa x)\, x^{\vartheta-1} e^{-\beta_g x} dx - \sum_{i=0}^{\vartheta-1} \frac{2\beta_g^{\vartheta+i}}{\ln(2)\Gamma(\vartheta) i!} \int_0^{\infty} \ln(1+\kappa x)\, x^{\vartheta+i-1} e^{-2\beta_g x} dx$$
$$= \frac{2\beta_g^{\vartheta}}{\ln(2)\Gamma(\vartheta)} \int_0^{\infty} \ln(1+\kappa x)\, x^{\vartheta-1} e^{-\beta_g x} dx - \xi_1(\kappa). \tag{C-7}$$

Finally, by doing some manipulations in (C-7), and also using results from [26], we obtain the following solution

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

which completes the proof. ■

APPENDIX D
PROOF OF PROPOSITION III

Clearly, the objective function in (25a) is linear and the function on the left-hand-side of constraint (25b) consists in a quadratic polynomial. As $\varrho_{gk} \geq 0$, $\forall\ g, k$, the constraint in (25b) is convex. This makes (25) a convex optimization problem. Therefore, the KKT conditions are necessary and sufficient to determine the global optimal solution of the considered problem [35]. The Lagrangian function of (25) can be written as

$$\mathcal{L}(\alpha_{g2}, \omega, \nu) = -\alpha_{g2} + \omega[\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)$$
$$- (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)] - \nu\alpha_{g2}, \tag{D-1}$$

where $\omega \geq 0$ and $\nu \geq 0$ are, respectively, the Lagrangian multipliers associated with the constraints (25b) and (25c). The KKT conditions are summarized as follows

$$\nabla\mathcal{L}(\alpha_{g2}, \omega, \nu) = 0, \tag{D-2a}$$

$$\omega[\alpha_{g2}^2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2}) + \alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)$$
$$- (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)] = 0, \tag{D-2b}$$

$$-\nu\alpha_{g2} = 0. \tag{D-2c}$$

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

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

Considering that the expression in (D-3) never reaches zero, and that $0 \leq \mu < 1$, the solution for (D-2b) can be obtained from the following quadratic equation

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

If $\mu = 1$, (D-4) becomes the following linear equation

$$\alpha_{g2}(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) - (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2) = 0. \tag{D-5}$$

Therefore, the optimal power allocation for user 2 can be obtained by calculating the zeros of (D-4), if $0 \leq \mu < 1$, or solving (D-5), if $\mu = 1$, as follows

$$\alpha_{g2}^* = \begin{cases} \frac{1}{2(\varrho_{g1}\varrho_{g2} - \mu\varrho_{g1}\varrho_{g2})}\Big[-(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g) & \\ + \sqrt{(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g)^2 + 4\varrho_{g1}\varrho_{g2}(1-\mu)(\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)}\,\Big], & \textit{if} \quad 0 \leq \mu < 1, \\ (\varrho_{g1}\rho^{-1}\bar{\alpha}_g + \mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g^2)/(\varrho_{g1}\rho^{-1} + \varrho_{g2}\rho^{-1} + 2\mu\varrho_{g1}\varrho_{g2}\bar{\alpha}_g), & \textit{if} \quad \mu = 1, \end{cases} \tag{D-6}$$

which completes the proof. ■

Appendix E
Proof of Proposition IV

The amount of power Δ_α 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

$$R_{\hat{g}2} = R_{\check{g}2} \implies \log_2\left(1 + \rho\varrho_{\hat{g}2}\alpha_{\hat{g}2}^*\right) = \log_2\left(1 + \rho\varrho_{\check{g}2}\alpha_{\check{g}2}^*\right) \implies \alpha_{\hat{g}2}^*\varrho_{\hat{g}2} = \alpha_{\check{g}2}^*\varrho_{\check{g}2}. \tag{E-1}$$

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

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

Then, after some algebraic manipulation, and defining $K_1 = \frac{\varrho_{\check{g}1}^2}{\varrho_{\hat{g}1}^2}\varrho_{\hat{g}1}^2\varrho_{\hat{g}2}\rho^{-1}$, $K_2 = (\varrho_{\check{g}1}\rho^{-1} + \varrho_{\check{g}2}\rho^{-1} - \frac{\varrho_{\check{g}1}}{\varrho_{\hat{g}1}}(\varrho_{\hat{g}1}\rho^{-1} + \varrho_{\hat{g}2}\rho^{-1}))^2$, $K_3 = ((\varrho_{\check{g}1}\rho^{-1} + \varrho_{\check{g}2}\rho^{-1})^2 + 4\varrho_{\check{g}1}^2\varrho_{\check{g}2}\rho^{-1}\bar{\alpha}_{\check{g}} - K_2 - \frac{\varrho_{\check{g}1}^2}{\varrho_{\hat{g}1}^2}(\varrho_{\hat{g}1}\rho^{-1} +$

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

$$A_1\Delta_\alpha^2 + A_2\Delta_\alpha + A_3 = 0. \tag{E-3}$$

The final result is obtained by calculating the zeros of (E-3). This completes the proof. ■

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.*, vol. 18, no. 12, pp. 5630–5642, Dec. 2019.

[5] A. S. de Sena, D. B. da Costa, Z. Ding, P. H. J. Nardelli, U. S. Dias, and C. B. Papadias, "Massive MIMO-NOMA networks with successive sub-array activation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1622–1635, Mar. 2020.

[6] 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, pp. 3556–3571, Jun. 2018.

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

[8] A. S. de Sena, D. Carrillo, F. Fang, P. H. J. Nardelli, D. B. da Costa, U. S. Dias, Z. Ding, C. B. Papadias, and W. Saad, "What role do intelligent reflecting surfaces play in non-orthogonal multiple access?" *TechRxiv*, Feb. 2020. [Online]. Available: 10.36227/techrxiv.11791050.v1.

[9] H. D. R. Albonda and J. Pérez-Romero, "An efficient RAN slicing strategy for a heterogeneous network with eMBB and V2X services," *IEEE Access*, vol. 7, pp. 44 771–44 782, Mar. 2019.

[10] 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.

[11] 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.

[12] 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.

[13] S. Li, M. Derakhshani, and S. Lambotharan, "Outage-constrained robust power allocation for downlink MC-NOMA with imperfect SIC," in *IEEE Int. Conf. Commun.*, 2018, pp. 1–7.

[14] A. Celik, M. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M. Alouini, "Distributed cluster formation and power-bandwidth allocation for imperfect NOMA in DL-HetNets," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1677–1692, 2019.

[15] 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.

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

[17] 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.*, vol. 9, no. 1, pp. 17–20, 2020.

[18] X. Li, J. Li, Y. Liu, Z. Ding, and A. Nallanathan, "Residual transceiver hardware impairments on cooperative NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 680–695, 2020.

[19] J. Kang, I. Kim, and C. Chun, "Deep learning-based MIMO-NOMA with imperfect SIC decoding," *IEEE Syst. J.*, pp. 1–4, 2019.

[20] 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.

[21] 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.

[22] 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.

[23] 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, Aug. 2019.

[24] 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.

[25] 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, pp. 3411–3423, Jul. 2019.

[26] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed. Academic Press, Amsterdam, 2007.

[27] 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.

[28] 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.

[29] S. Bazzi and W. Xu, "Downlink training sequence design for FDD multiuser massive MIMO systems," *IEEE Trans. Signal Proces.*, vol. 65, no. 18, pp. 4732–4744, Sep. 2017.

[30] A. Hasan and J. Andrews, "Cancellation error statistics in a power-controlled CDMA system using successive interference cancellation," in *IEEE Int. Symp. on Spread Spectrum Tech. Appli.*, 2004, pp. 419–423.

[31] A. Agrawal, J. G. Andrews, J. M. Cioffi, and Teresa Meng, "Iterative power control for imperfect successive interference cancellation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 878–884, 2005.

[32] 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.

[33] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Wiley Series in Probability and Statistics, Aug. 2003.

[34] F. R. M. Lima, T. F. Maciel, W. C. Freitas, and F. R. P. Cavalcanti, "Improved spectral efficiency with acceptable service provision in multiuser MIMO scenarios," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2697–2711, Jul. 2014.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.