Massive MIMO-NOMA Networks with Successive Sub-Array Activation

Arthur Sousa de Sena, *Member*, *IEEE*, Daniel Benevides da Costa, *Senior Member*, *IEEE*, Zhiguo Ding, *Senior Member*, *IEEE*, Pedro H. J. Nardelli, *Member*, *IEEE*, Ugo S. Dias, *Senior Member*, *IEEE*, and Constantinos B. Papadias, *Fellow*, *IEEE*

Abstract

In this paper, we propose a novel successive sub-array activation (SSAA) diversity scheme for a massive multiple-input multiple-output (MIMO) system in combination with non-orthogonal multiple access (NOMA). Considering a single-cell multi-cluster downlink scenario, where the base station (BS) sends redundant symbols through multiple transmit sub-arrays to multi-antenna receivers, a low-complexity two-stage beamformer, that is constructed based only on the long-term channel statistical information, is proposed. An in-depth analytical analysis is carried out, in which an exact closed-form expression for the outage probability is derived. A high signal-to-noise ratio (SNR) outage approximation is obtained and the system diversity order is determined. The ergodic sum-rate is also investigated, in which a closed-form solution is evaluated considering a particular case. Numerical and simulation results are provided to validate the analytical analysis and to demonstrate the performance superiority of the proposed SSAA scheme. For example, our results show that the proposed system operating with SSAA outperforms conventional full array massive MIMO setups.

Index Terms

A. S. de Sena was with the Federal University of Ceará, Brazil. He is now with the Lappeenranta University of Technology, Finland (email: arthurssena@ieee.org).

D. B. da Costa is with the Federal University of Ceará, Brazil (email: danielbcosta@ieee.org).

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

P. H. J. Nardelli is with the Lappeenranta University of Technology, Finland (email: pedro.nardelli@lut.fi).

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).

Non-orthogonal multiple access (NOMA), massive MIMO, successive sub-array activation.

I. Introduction

Together with the 5th generation of wireless communication networks (5G), a new world of possibilities will emerge. In this new era, unforeseen new services and applications will arise, from smart-homes and self-driving cars to interconnected low-power sensors in agriculture and factories. These new deployments will impose the most diverse requirements to the wireless systems, such as massive connectivity, low latency, and reliable communications [1]–[4]. Non-orthogonal multiple access (NOMA) and massive multiple-input multiple-output (MIMO) are considered as key enabling technologies for meeting these requisites. Specifically, massive MIMO explores the space domain through a massive number of antennas to serve multiple users in parallel, while NOMA can also serve multiple users simultaneously, but differently from MIMO, the parallel transmission is performed by multiplexing the users in the power domain, in which successive interference cancellation (SIC) is employed for reception. Furthermore, the combination of MIMO and NOMA can provide remarkable improvements and outperform conventional orthogonal multiple access (OMA) systems [5], [6]. Due to the mentioned features, both technologies were included in the 3rd generation partnership project (3GPP) long-term evolution advanced (LTE-A) Release 13, 14 and in the specification for 5G New Radio standard, introduced in Release 15 [7], [8].

In addition to the significant performance improvements, massive MIMO-NOMA systems can reduce the overall latency and increase connectivity capacity. However, a reliable communication cannot always be guaranteed. Due to various reasons, the system performance can be degraded and the communication quality impacted. Especially in 5G networks, due to the use of higher frequency bands (above 6 GHz), the penetration loss and atmospheric absorption can strongly affect the system reliability [9]. The high mobility scenario is another critical application, in which the fast-varying channels and double-selective fading make very difficult to establish a good communication [10]. Employing a massive number of antennas also introduces some problems. The strong channel correlation caused by the closely spaced antennas can severely deteriorate the performance of MIMO systems [11]. Furthermore, in real NOMA deployments, if not well projected, the SIC process can lead to error propagation [12]. In all mentioned situations, the introduction of redundancy through diversity techniques is crucial for enhancing the system performance and improve reliability, which is a very important requisite for many applications.

A. Related Works

Even though diversity strategies in classical communication systems have been exhaustively investigated for decades, the great majority of existing MIMO-NOMA contributions does not explore the refereed subject. For instance, in [5] and [6], the performance of multi-user multi-cluster MIMO-NOMA systems was studied. In both works, an in-depth analytical analysis was carried out and closed-form outage expressions were derived, but diversity was not taken into consideration. In [13], NOMA was applied to a millimeter wave (mmWave) massive MIMO system, where only the ergodic capacity was investigated. The authors derived exact and asymptotic capacity expressions and demonstrated through simulations that NOMA can provide significant improvements to mmWave setups. The work in [14] addressed the design and optimization of linear beamformers in MIMO-NOMA multi-cell multi-user systems. The objective was to maximize the sum throughput of the network while guaranteeing the users' quality of service (QoS). To solve the optimization problem, iterative path-following algorithms were proposed, and it was shown to improve the performance of cell-edge users. The works in [15] and [16] also considered multi-cell MIMO-NOMA networks. Specifically, in [15], a sub-optimal power allocation algorithm based on successive convex approximation was developed. The authors demonstrated that their proposal outperforms conventional NOMA and OMA schemes. In [16], the application of improper Gaussian signaling (IGS) to both MIMO-OMA and MIMO-NOMA systems was considered. The paper developed an optimization technique based on linear matrix inequality to maximize the users' minimum throughput subject to power constraints. It was shown that MIMO-NOMA with IGS achieves significant throughput gains over the OMA system. The performance of a full-duplex multiple antenna relay-assisted NOMA system was studied in [17], in which bounds for the outage probability were achieved. In [18], a cooperative relaying system using spatial modulation and NOMA was proposed, and approximations for the bit error probability were provided. Transmit antenna selection and simultaneous wireless information and power transfer (SWIPT) were applied to a two-user relaying scenario in [19]. In this work, by proposing three different transmission strategies, the authors improved the performance of the cell-edge user while providing energy efficiency to the cell-center user. Exact and asymptotic closed-form expressions for the outage probability were also derived. The application of spatial modulation, MIMO, and NOMA to vehicle-to-vehicle communication was considered in [20], in which an ergodic capacity analysis was carried out. In [21], by applying a stochastic geometrical model, the performance of a MIMO-NOMA assisted unmanned aerial vehicle network was evaluated, and the cell-free massive MIMO-NOMA case was analyzed in [12].

A few works have addressed diversity in massive MIMO-NOMA setups. In [22], antenna diversity was employed to enhance the outage performance of a downlink MIMO-NOMA system with users equipped with only two receive antennas, and the expansion to a scenario with multiple receive antennas was proposed in [23]. In both works, the authors provided closed-form expressions for the outage probability and also analyzed the system ergodic rates. However, in the proposed designs, only one NOMA group is served at a time, which is a limitation. Considering a single base station (BS) with only two transmit antennas, the work in [24] achieves diversity through the combination of Alamouti space-time block coding (STBC) and MIMO-NOMA, in which a high signal-to-noise ratio (SNR) approximation and an exact expression for the outage probability was derived. In [25], a multi-dimensional STBC was proposed to a MIMO sparse code multiple access system, for two and four transmit antennas. The application of STBC to cooperative NOMA networks was investigated in [26] and [27]. Coordinated multi-point (CoMP) is employed to MIMO-NOMA systems in [28] and [29] to improve the bit error rate and energy efficiency of cell-edge users. Performance improvements in MIMO-NOMA systems can also be achieved through antenna selection schemes [30], [31]. However, it was shown that the optimal solution is very complex and, since NOMA sends a unique superposed symbol, the antenna selection can not maximize the signal-to-interference-plus-noise ratio (SINR) of all users at the same time.

B. Motivation and Contributions

The exploration of all forms of diversity, in frequency, code or time domains, will be essential for satisfying the high quality of service requirements of 5G networks. However, there is a lack of related contributions, and only a very limited number of works investigates diversity techniques in massive MIMO-NOMA systems. This motivates further studies in this field of research. Owing to this fact, in this paper, by combining concepts of time diversity and antenna sub-array selection, we propose a novel low-complexity scheme with the potential of improving the outage performance of each user in a massive MIMO-NOMA deployment. More details and main contributions of this treatise are summarized as follows.

- We design a novel open-loop diversity technique that enables each of the users to improve their reception performance without the need for transmitting back their fast varying channel matrices. This is achieved by successively activating antenna sub-arrays at the base station and exploring space diversity in different instants of time at the users' side. With this strategy, the system latency can still be maintained low.
- We adopt two-stage beamformers for each of the antenna sub-arrays. The outer beamforming matrices are constructed based only on the channel statistical information, i.e. the channel covariance matrices, and the inner beamformers are only intended for assigning the messages for each of the NOMA groups. These choices also do not require knowledge of the fast varying channels. Therefore, in our proposal, either for providing diversity or for constructing the beamformers, only the slowly varying covariance matrices are needed, which is attractive for applications where feedback is expensive or very limited.
- An in-depth analytical analysis in the proposed design is performed, in which an exact closed-form expression for the outage probability is derived. The system behavior at high SNR regime is also investigated, where an asymptotic outage approximation is attained, enabling us to determine the system diversity order. Furthermore, we investigate and obtain a closed-form expression for the ergodic sum-rate considering a particular case.
- Numerical and simulation results are presented to corroborate the theoretical development, and insightful discussions are presented. In particular, our results show that the proposed strategy outperforms conventional full array massive MIMO-NOMA and MIMO-OMA systems operating in time diversity mode in all considered scenarios.

C. Structure of the Paper

The rest of the paper is organized as follows. In Section II, the massive MIMO-NOMA system model with multiple antenna sub-arrays at the BS is introduced. A detailed description of the proposed diversity scheme protocol and the design of the beamforming and detection matrices are also presented. In Section III, the analytical analysis of the system performance is carried out, in which exact closed-form and high SNR approximation expressions for the outage probability are derived. The ergodic sum-rate is also investigated. Details about conventional MIMO-OMA and MIMO-NOMA schemes, used for performance comparison, are described in Section IV. Numerical and simulation results along with comprehensive discussions are presented in Section V, while conclusions and perspectives for future works are drawn in Section VI.

Notation and Special Functions: Bold-faced lower-case letters represent vectors and upper-case letters denote matrices. The norm and the ith element of a vector $\mathbf{a}$ are represented by $||\mathbf{a}||$ and $[\mathbf{a}]_i$, respectively. The notations $[\mathbf{A}]_{ij}$ and $[\mathbf{A}]_{i*}$ correspond the (ij) entry and the ith row of the matrix $\mathbf{A}$, respectively. The Hermitian transposition of a matrix $\mathbf{A}$ is donated by $\mathbf{A}^H$ and the trace by tr$\{\mathbf{A}\}$. $\mathbf{I}_M$ represents the identity matrix of dimension $M \times M$, and $\mathbf{0}_{M\times N}$ denotes the $M \times N$ matrix with all zero entries. In addition, $\otimes$ represents the Kronecker product, $\mathbb{E}[\cdot]$ denotes expectation, $\Gamma(\cdot)$ is the Gamma function, $\gamma(\cdot,\cdot)$ corresponds to the lower incomplete Gamma function, and $\text{Ei}(\cdot)$ is the exponential integral.

II. System Model and Protocol Description

Consider a scenario where a single BS equipped with a linear array of M antennas is transmitting to multiple users. Each user employs N receive antennas, with $M \gg N$. Besides, the users are assumed to be confined within S rich scattering clusters, following the one ring geometrical model [32], [33]. In each spatial cluster, there are G sub-groups, each one containing K users that are multiplexed through power-domain NOMA. It is also assumed that the users require a reliable data reception. To fulfill this requirement, we propose a novel diversity strategy by exploring space and time dimensions. Firstly, at the base station, the transmit antennas are equally divided into L sub-arrays, i.e., we create L partitions of M/L antenna elements, as shown in Fig. 1. Due to this structure, M must be a multiple of L. In addition, we adjust the separation distance between two adjacent sub-arrays to be greater than half of the wavelength, i.e. greater than $\lambda/2$, so that the channels among sub-arrays become uncorrelated. Within each sub-array, we set the inter-antenna space separation to be exactly $\lambda/2$ and we consider correlation between antenna elements. Then, in order to improve reliability, we configure the system to send L replicas of the same symbol. More specifically, each symbol replica is transmitted by

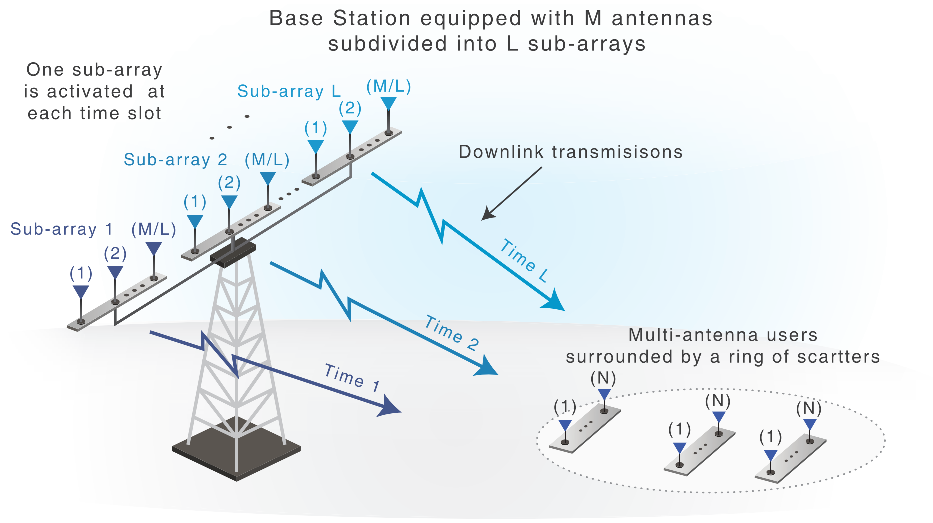

Fig. 1. System model. The BS is equipped with multiple antennas subdivided into multiple sub-arrays.

sequentially activating each sub-array. Consequently, the transmission of one symbol is performed within L instants of time. The proposed strategy will be called as successive sub-array activation (SSAA). Note that, since the sub-arrays are uncorrelated, each transmission will propagate through different paths, regardless of the separation time between two retransmissions. Therefore, diversity is achieved through the space dimension.

As one can realize, differently from conventional time diversity schemes, that need to wait for a whole coherent time to retransmit the data, our proposed system can operate with very fast transmission rates. The time needed to retransmit the symbol replicas must be just the enough to the receiver distinguishes the signals from each sub-array[1]. As a result, a very low latency can still be achieved, while guaranteeing an enhanced performance. This capability can be very important, for example, for short data packet transmissions in 5G, in which the time duration for transmitting one symbol is much smaller than the channel coherence time [34]. Our diversity strategy is also energy efficient in terms of power consumption, since, regardless of the number of transmitted replicas, the total number of antennas activated during all retransmissions remains constant, i.e. a total of M antennas is used to transmit L symbol replicas. In addition to these advantages, since only one sub-array is activated at a time, it is possible to reduce the number

[1] In this work, we assume that the time duration needed to retransmit each data symbol is enough to avoid any collision so that no errors are introduced from retransmission collisions.

of dedicated electronic components that are connected to the antenna elements, known as radio frequency (RF) chains. More specifically, by using simple switches [35], the number of dedicated RF chains could be reduced to M/L, i.e. the same number of antennas in a sub-array, without degrading the system performance, what would lead to a decrease in hardware cost and to further improvements in power requirements.

A. Channel Model

We consider that all users within the sth spatial cluster share the same channel covariance matrix $\bar{\mathbf{R}}_s = \mathbf{I}_L \otimes \mathbf{R}_s \in \mathbb{C}^{M \times M}$, where $\mathbf{R}_s \in \mathbb{C}^{\frac{M}{L} \times \frac{M}{L}}$ corresponds to the covariance matrix of each sub-array, which has rank denoted by r_s. Note that, for simplicity, we assume identical covariance matrices among sub-arrays. In addition, the (mp) entry of each $\mathbf{R}_s$, representing the correlation between antenna elements m and p, for $m, p = 1, \cdots, \frac{M}{L}$, can be obtained by [33]

$$[\mathbf{R}_s]_{mp} = \frac{1}{2\Delta_s} \int_{-\Delta_s}^{\Delta_s} e^{j\mathbf{k}^T(\alpha+\theta_s)(\mathbf{u}_m - \mathbf{u}_p)} d\alpha, \tag{1}$$

where $\Delta_s = \tan^{-1}(\frac{s_s}{d_s})$ is the angular spread for the sth cluster, which is located at a distance d_s and azimuth angle θ_s, surrounded by a ring of scatterers of radios s_s. $\mathbf{k}(\alpha) = -\frac{2\pi}{\lambda}[\cos(\alpha), \sin(\alpha)]^T$ is a vector that describes the incident planar waves arriving at the BS with angle of arrival α, and $\mathbf{u}_m, \mathbf{u}_p \in \mathbb{R}^{2 \times 1}$ are coordinate vectors that specify the positions of transmit antennas m and p, respectively. For convenience, in this paper we suppose that the spatial clusters share identical radios and are located at equal distances from the BS[2], i.e., $d_1 = d_2 = \cdots = d_S$ and $s_1 = s_2 = \cdots = s_S$.

Considering the proposed design, the BS transmits each symbol in L instants of time, where each replica propagates independently and experiences different fast fading channels. Besides, it is assumed a perfect downlink channel estimation at the users' side. Under these considerations, each user will acquire L distinct channel matrices, each one corresponding to a different transmission. For mathematical convenience, the channel matrices belonging to the kth user in

[2] Even though in this work we consider the simplified scenario in which $d_1 = d_2 = \cdots = d_S$ and $s_1 = s_2 = \cdots = s_S$, our design can be easily extended to the case where the clusters have different dimensions and randomly located, i.e., $d_1 \neq d_2 \neq \cdots \neq d_S$ and $s_1 \neq s_2 \neq \cdots \neq s_S$.

the gth sub-group of the sth cluster are organized in the following block diagonal arrangement

$$\bar{\mathbf{H}}_{sgk} = \begin{bmatrix} \mathbf{H}_{sgk}^{1} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_{sgk}^{L} \end{bmatrix} \in \mathbb{C}^{M \times LN}, \tag{2}$$

where $\mathbf{H}_{sgk}^{l} \in \mathbb{C}^{\frac{M}{L} \times N}$ represents the channel matrix obtained during reception of the message transmitted by the lth antenna sub-array. By invoking the Karhunen-Loeve transformation [33], we can decompose the sub-matrices of $\bar{\mathbf{H}}_{sgk}$ as follows

$$\bar{\mathbf{H}}_{sgk} = \begin{bmatrix} \mathbf{U}_s \mathbf{\Lambda}_s^{\frac{1}{2}} \mathbf{G}_{sgk}^{1} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{U}_s \mathbf{\Lambda}_s^{\frac{1}{2}} \mathbf{G}_{sgk}^{L} \end{bmatrix}, \tag{3}$$

where $\mathbf{\Lambda}_s$ stands for a diagonal matrix of dimension $r_s^* \times r_s^*$ composed by r_s^* decreasing nonzero eigenvalues of $\mathbf{R}_s$, $\mathbf{U}_s \in \mathbb{C}^{\frac{M}{L} \times r_s^*}$ represents a tall unitary matrix formed by eigenvectors of $\mathbf{R}_s$, and $\mathbf{G}_{sgk}^{l} \in \mathbb{C}^{r_s^* \times N}$ is the fast varying channel matrix corresponding to the lth sub-array. Considering a non-line of sight communication, the entries of $\mathbf{G}_{sgk}^{l}$ are modeled as i.i.d. complex Gaussian distributed random variables with zero mean and unit variance.

Then, after the BS superposes the messages of all users within each sub-group and transmit the L successive replicas, over each sub-array, through the fast fading channels, the users observe the following signal

$$\mathbf{y} = \bar{\mathbf{H}}_{sgk}^{H} \sum_{n=1}^{S} \bar{\mathbf{B}}_n \sum_{i=1}^{G} \bar{\mathbf{v}}_{ni} \sum_{j=1}^{K} \alpha_{nij} x_{nij} \in \mathbb{C}^{LN \times 1}, \tag{4}$$

where $\bar{\mathbf{B}}_n \in \mathbb{C}^{M \times LV}$ is the beamforming matrix designed to remove inter-cluster interference, with V being a parameter that defines the virtual channel dimension, $\bar{\mathbf{v}}_{ni} \in \mathbb{C}^{LV \times 1}$ is the precoding vector responsible for assigning the superposed messages to its respective sub-groups, α_{nij} is the power coefficient allocation, and x_{nij} is the message intended for the user j in the ith sub-group of the nth cluster.

B. Beamforming Design

In this subsection, we present details for the construction of the beamforming matrices adopted in our design. As previously stated, our proposed system employs a two-stage beamformer, which is composed by an outer pre-processing matrix $\bar{\mathbf{B}}_s$ and an inner beamforming vector $\bar{\mathbf{v}}_{sg}$. This choice provides some attractive advantages to massive MIMO setups, such as less processing overload and reduced feedback overhead. Due to these benefits, the two-stage beamforming has been vastly explored in the literature and adopted in many relevant and recent works [5], [11], [33]. In our system, we choose to project the beamformers not to depend on the fast varying channel matrices. Consequently, the BS will only need to estimate the covariance matrices of the lower dimension antenna sub-arrays, which provides an even further reduction in the overall feedback overhead.

Considering the proposed antenna structure, the outer beamformer $\bar{\mathbf{B}}_s$, belonging to the sth cluster, can be arranged in the following block diagonal matrix

$$\bar{\mathbf{B}}_s = \begin{bmatrix} \mathbf{B}_s & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_s \end{bmatrix}, \tag{5}$$

where $\mathbf{B}_s \in \mathbb{C}^{\frac{M}{L} \times V}$ denotes the beamforming sub-matrix that is designed based on the slowly varying channel covariance matrix of each sub-array. Note that the beamforming matrices $\mathbf{B}_s$ have the role of suppressing the interference generated by other clusters, which means that $(\mathbf{H}_{sgk}^{l})^H \mathbf{B}_{s'} \approx 0$, for all $s' \neq s$, must be accomplished. Perfect orthogonality is obtained when the value of r_s^* is equal to the rank r_s of $\mathbf{R}_s$. However, increasing r_s^* too much may not be advantageous, since it may yield in fewer clusters being served in parallel, without real performance improvements [33].

In order to achieve the desired objective, we explore the null space of dominant eigenmodes from interfering clusters to build $\mathbf{B}_s$. Assuming that all clusters share equal values of r_s^* and r_s, the index s can be omitted, i.e., $r_1^* = \cdots = r_S^* = r^*$ and $r_1 = \cdots = r_S = r$. Then, we concatenate the left eigenvectors of interfering clusters to form the following matrix

$$\mathbf{U}_s^{-} = [\mathbf{U}_1, \cdots, \mathbf{U}_{s-1}, \mathbf{U}_{s+1}, \cdots, \mathbf{U}_S] \in \mathbb{C}^{\frac{M}{L} \times (S-1)r^*}. \tag{6}$$

Next, we denote the left eigenvectors of $\mathbf{U}_s^-$ by $\mathbf{E}_s = [\mathbf{E}_s^1, \mathbf{E}_s^0]$. The matrix $\mathbf{E}_s^0 \in \mathbb{C}^{\frac{M}{L} \times \frac{M}{L} - (S-1)r^*}$ collects the last $\frac{M}{L} - (S-1)r^*$ columns of $\mathbf{E}_s$, which corresponds to the eigenvectors associated with the vanishing eigenvalues of $\mathbf{U}_s^-$. Then, we define the projected channel $\tilde{\mathbf{H}}_{sgk}^l = (\mathbf{E}_s^0)^H \mathbf{U}_s \mathbf{\Lambda}_s^{\frac{1}{2}} \mathbf{G}_{sgk}^l$, which is orthogonal to the eigen-space spanned by interfering clusters and has the following covariance matrix

$$\tilde{\mathbf{R}}_s = (\mathbf{E}_s^0)^H \mathbf{R}_s \mathbf{E}_s^0. \tag{7}$$

Subsequently, by applying the singular value decomposition (SVD) in (7), we rewrite the matrix $\tilde{\mathbf{R}}_s$ as

$$\tilde{\mathbf{R}}_s = \mathbf{F}_s \mathbf{R}_s \mathbf{F}_s^H. \tag{8}$$

Let now $\mathbf{F}_s^{(1)} \in \mathbb{C}^{\frac{M}{L} - (S-1)r^* \times V}$ be the first V columns of the eigenvector matrix $\mathbf{F}_s$, corresponding to the dominant eigenvalues of $\tilde{\mathbf{R}}_s$. Then, finally, the outer beamforming matrix for each antenna sub-array can be obtained as

$$\mathbf{B}_s = \mathbf{E}_s^0 \mathbf{F}_s^1 \in \mathbb{C}^{\frac{M}{L} \times V}, \tag{9}$$

in which, due to the matrix dimension in (6) and the rank of $\mathbf{R}_s$, the constraints $S \leq V \leq \left(\frac{M}{L} - (S-1)r^*\right)$ and $V \leq r^* \leq r$ must be fulfilled. In addition, note that the parameter V defines the number of parallel effective transmissions that are performed by each sub-array to each spatial cluster.

Now, we focus on the design of the inner beamforming vector. Due to the multi-array architecture, $\bar{\mathbf{v}}_{sg}$ is composed by L sub-vectors, which can be represented as follows

$$\bar{\mathbf{v}}_{sg} = [(\mathbf{v}_{sg}^1)^T, \cdots, (\mathbf{v}_{sg}^L)^T]^T, \tag{10}$$

where $\mathbf{v}_{sg}^l$ is the beamforming sub-vector corresponding to the lth sub-array, and it is responsible for assigning the data message for the gth sub-group in the sth cluster. Since each sub-array contains the same number of antenna elements, we have that $\mathbf{v}_{sg}^1 = \mathbf{v}_{sg}^2 = \cdots = \mathbf{v}_{sg}^L$. Therefore, we define the sub-beamformers as

$$\mathbf{v}_{sg}^{l} = [\, \mathbf{0}_{1\times(g-1)} \, , \, 1 \, , \, \mathbf{0}_{1\times(V-g)} \,]^{T}, \quad \forall l = 1, \cdots, L. \tag{11}$$

With this choice, note that the gth effective data stream transmitted by each sub-array is associated with the gth sub-group. As a consequence, the parameter V also defines the maximum number of sub-groups that can be simultaneously served within a given cluster. This approach enables each sub-group to receive L copies of the same superposed symbol. Moreover, since we design $\bar{\mathbf{v}}_{sg}$ only to assign the superposed data messages to the respective sub-groups, its construction does not depend on any channel state information (CSI), that is, neither on the covariance matrices nor on the fast varying channels. Therefore, the inner beamformer construction does not require any additional feedback overhead. Besides, it imposes an extremity low computational complexity, which contributes to a further decrease in the overall system latency.

C. Signal Reception

Hereafter, we will direct our attention to the users located in the first spatial cluster. Thus, for the sake of brevity, the cluster subscript is omitted, e.g. $\mathbf{y}_{sgk}$ is expressed as $\mathbf{y}_{gk}$.

Considering that the outer beamformer, designed in the last subsection, perfectly cancels the inter-cluster interferences, and after all L successive transmissions have been received, the data signal at the kth user in the gth sub-group can be structured as

$$\mathbf{y}_{gk} = \bar{\mathbf{H}}_{gk}^{H} \bar{\mathbf{B}} \sum_{i=1}^{G} \bar{\mathbf{v}}_i \sum_{j=1}^{K} \alpha_{ij} x_{ij} + \left[\mathbf{n}_{gk}^{1}, \cdots, \mathbf{n}_{gk}^{L}\right]^{T}, \tag{12}$$

where $\mathbf{n}_{gk}^{l} \in \mathbb{C}^{N\times 1}$ is a complex Gaussian noise vector obtained during reception of the signal transmitted by the lth sub-array, with entries having zero-mean and variance σ_n^2.

One can notice that, even though the inter-cluster interference has been eliminated, the users still need to separate the superposed data symbols intended for each sub-group, so that they can recover their messages. In order to accomplish this task, zero-forcing detectors are employed in the users' terminals. Mathematically, the signal obtained in (12) is filtered through the following detection matrix

$$\bar{\mathbf{H}}^{\dagger}_{gk} = \begin{bmatrix} \mathbf{H}^{1\dagger}_{gk} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{H}^{L\dagger}_{gk} \end{bmatrix}, \tag{13}$$

in which $\mathbf{H}^{l\dagger}_{gk} = (((\mathbf{H}^{l}_{gk})^{H}\mathbf{B})^{H}(\mathbf{H}^{l}_{gk})^{H}\mathbf{B})^{-1}((\mathbf{H}^{l}_{gk})^{H}\mathbf{B})^{H}$ is the pseudoinverse of the virtual channel observed during the lth reception, where we suppose that $V \leq N$. After zero-forcing equalization, the channel distortion is completely removed and the users obtain a noise corrupted version of the signals transmitted by each antenna sub-array. With the proposed inner precoder in (11), the detected symbol vector can be represented by

$$\hat{\mathbf{x}}_{gk} = \left[\mathbf{x}^{1}, \cdots, \mathbf{x}^{L}\right]^{T} + \bar{\mathbf{H}}^{\dagger}_{gk}\left[\mathbf{n}^{1}_{gk}, \cdots, \mathbf{n}^{L}_{gk}\right]^{T} = \left[\begin{bmatrix} \sum\limits_{j=1}^{K} \alpha_{1j}x_{1j} \\ \vdots \\ \sum\limits_{j=1}^{K} \alpha_{Gj}x_{Gj} \end{bmatrix}, \dots, \begin{bmatrix} \sum\limits_{j=1}^{K} \alpha_{1j}x_{1j} \\ \vdots \\ \sum\limits_{j=1}^{K} \alpha_{Gj}x_{Gj} \end{bmatrix}\right]^{T} + \bar{\mathbf{H}}^{\dagger}_{gk}\begin{bmatrix} \mathbf{n}^{1}_{gk} \\ \vdots \\ \mathbf{n}^{L}_{gk} \end{bmatrix}, \tag{14}$$

where $\mathbf{x}^{l}$ is the vector of superposed data symbols transmitted through the lth sub-array, in which $\mathbf{x}^{1} = \cdots = \mathbf{x}^{L}$.

Note that, the users within the gth sub-group can recover their data messages through the gth element of any of the L sub-vectors in (14), which can be accomplished by employing any combining diversity technique. In our design, due to low complexity, we simply select the symbol from the sub-vector that delivers the best effective channel gain.

III. Performance Analysis

In this section, the performance of the proposed massive MIMO-NOMA design operating with the SSAA scheme is investigated. First, we analyze the SINR that is experienced by the users during each NOMA decoding. Then, based on the SINR result, the system outage probability is evaluated, in which an exact closed-form expression is obtained. In addition, a high SNR asymptotic analysis is also performed, where we identify the diversity order obtained with the proposed SSAA strategy. Aiming to achieve further insights about the system behavior, the ergodic sum-rate is also studied. In particular, considering a special case, we derive an exact ergodic sum-rate expression.

A. *SINR Analysis for the SSAA Diversity Scheme*

Aiming to enable the implementation of NOMA, the BS sorts out the users in ascending order based on the magnitude of their effective channel gains. Moreover, in this paper, we adopt a fixed power allocation policy, in which more power is assigned to users with poorer channel conditions, i.e. the power allocation coefficients are adjusted to satisfy $\alpha_{g1} > \alpha_{g2} > \cdots > \alpha_{gK}$. In order to perform such allocation, the effective channel gains are considered to be perfectly known at the BS. In addition, we assume that all the users can carry out SIC without introducing any error in the recovered message. On these terms, after all the SIC decodings are completed, the kth user in the gth sub-group will observe the following data symbol

$$\hat{x}_{gk} = \underset{\substack{\uparrow\\ \text{symbol of interest}}}{\alpha_{gk}x_{gk}} + \underset{\substack{\uparrow\\ \text{interference}}}{\sum_{j=k+1}^{K} \alpha_{gj}x_{gj}} + \underset{\substack{\uparrow\\ \text{noise}}}{[\mathbf{H}_{gk}^{m\dagger}\mathbf{n}_{gk}^{m}]_g}. \qquad (15)$$

where $m \in \{1, 2, \cdots, L\}$ corresponds to the sub-array that achieves the maximum effective channel gain among all L transmissions. From (15), the SINR obtained at the kth user while recovering its message is defined in Lemma 1.

Lemma I: Supposing that the users recover its desired message from the sub-array that delivers the best effective gain, the SINR of the kth user in the gth sub-group while decoding the message intended for the ith user, $1 \leq i \leq k \leq K$, is given by

$$\text{SINR}_{gk}^{i} = \frac{\rho\gamma_{gk}\alpha_{gi}^{2}}{\rho\gamma_{gk}\mathcal{P}_i + 1}, \qquad \text{for} \quad 1 \leq i \leq k \leq K, \qquad (16)$$

where $\gamma_{gk} = \max\left\{\varsigma_{gk}^{1}, \cdots, \varsigma_{gk}^{L}\right\}$ denotes the effective channel gain, with $\varsigma_{gk}^{l} = \frac{1}{\|[\mathbf{H}_{gk}^{l\dagger}]_{g*}\|^2}$, $1 \leq l \leq L$. $\rho = \frac{1}{\sigma_n^2}$ represents the transmit SNR, and $\mathcal{P}_i$ corresponds to the power of interfering users, which is defined by

$$\mathcal{P}_i = \begin{cases} \sum_{j=i+1}^{U} \alpha_{gj}^{2}, & \text{for} \quad 1 \leq i \leq k < K, \\ 0, & \text{for} \quad i = k = K, \end{cases} \qquad (17)$$

Proof: Please, see Appendix A.

By analyzing the expressions in (16) and (17), we can observe that, for a fixed SNR value, the interference term increases with the decrease of the user order, which eventually leads to an

SINR degradation. In contrast, when the user order increases the interference is decreased, which improves the SINR. In fact, this is the expected behavior in a NOMA deployment, in which due to the SIC protocol, the lower order users (the ones with worse channel conditions) experience more interference, and the higher order ones (the users with the best channels) experience less.

B. Outage Probablity

In NOMA, before the kth user at the gth sub-group can recover its own message, it needs first to employ SIC to decode and remove every symbol intended for the ith weaker user, $\forall$ $i = 1, \cdots, k$. Therefore, if at least one of the SIC decodings does not succeed, e.g., if the achieved data rate while decoding the ith message is less than the required rate R_{gi}, this user will not be able to retrieve its message and an outage event occurs. Thus, the outage probability of the kth user at the gth sub-group can be formulated as

$$P_{gk} = P[\log_2(1 + \text{SINR}_{gk}^{i}) < R_{gi}], \quad \forall i = 1, \cdots k. \tag{18}$$

A closed-form expression for the outage probability achieved in the massive MIMO-NOMA setup with the proposed diversity strategy is provided in the following proposition.

Proposition I: Supposing that the effective channel gains are ordered as $\gamma_{g1} < \gamma_{g2} < \cdots < \gamma_{gK}$, the outage probability for the massive MIMO-NOMA system operating with the proposed SSAA scheme can be derived as

$$P_{gk} = \sum_{j=0}^{K-k} \frac{\mathcal{K}_{kj}}{k+j} \left[\frac{\gamma\left(N - V + 1, \mathcal{M}_{gk}[(\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1}]_{gg}\right)}{\Gamma(N - V + 1)} \right]^{L(k+j)}, \tag{19}$$

where $\mathcal{K}_{kj} = K\binom{K-1}{k-1}\binom{K-k}{j}(-1)^j$ and $\mathcal{M}_{gk} = \max\limits_{1 \leq i \leq k} \left\{ \frac{2^{R_{gi}}-1}{\rho[\alpha_{gi}^2 - \mathcal{P}_i(2^{R_{gi}}-1)]} \right\}$.

Proof: Please, see Appendix B.

From Proposition I, we can obtain some insights about the performance of our proposed diversity scheme. It can be noticed that the expression in (19) is a monotonically decreasing function of the transmit SNR ρ, in which the exponent determines how fast the outage probability decreases, i.e. greater exponents lead to a faster decrease. This suggests that by increasing the number of sub-arrays L and, consequently, increasing the exponent in (19), the system performance might be improved.

C. Asymptotic Analysis

Even though we have derived an exact expression for the outage probability in Proposition I, it is still complex to analyze some important performance features. Therefore, a simpler expression is desirable to investigate further the behavior of the proposed system. In view of this, an asymptotic outage analysis is carried out in Proposition II, in which we derive a simpler outage probability expression and determine the system diversity order, as follows.

Proposition II: When the transmit SNR approaches infinity, (19) can be approximated by

$$P_{gk} \approx \frac{K}{k}\binom{K-1}{k-1}\frac{[\rho\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}]^{(N-V+1)Lk}}{\rho^{(N-V+1)Lk}[(N-V+1)!]^{Lk}}. \tag{20}$$

As a result, the kth user experiences the following diversity order

$$D_k = (N-V+1)\,Lk. \tag{21}$$

Proof: Please, see Appendix C.

From (21), we see that, for a fixed number of N and V, the diversity order achieved with our proposed system design scales with the increase in the number of sub-arrays. This means that, when the SNR is high, i.e. $\rho \to \infty$, our multi-array system achieves superior performance than the conventional MIMO-NOMA counterpart, whose diversity order can be obtained in [5]. Another detail that can be observed in (21) is that, the diversity order increases with the order of the user, what is indeed expected.

D. Ergodic Sum-Rate

In this subsection, we investigate the ergodic sum-rate for the proposed massive MIMO-NOMA system operating with the SSAA strategy. Here, we assume that the kth user can successfully decode all symbols intended for the ith weaker user, $\forall i = 1, \cdots, k$, so that, the instantaneous capacity of each user will depend only on the SINR achieved while decoding its own message. Under this assumption, the ergodic sum-rate for the gth sub-group, after the L successive receptions, can be evaluated by

$$\bar{R}_g = \mathbb{E}\left[\sum_{k=1}^{K}\log_2(1+\text{SINR}_{gk}^{k})\right]. \tag{22}$$

A general solution for the expectation in (22) is presented in Proposition III.

Proposition III: Under the assumption of perfect SIC decoding, the ergodic sum-rate for the gth sub-group achieved with the SSAA strategy can be attained by

$$\begin{aligned}\bar{R}_g =& \sum_{k=1}^{K}\sum_{j=0}^{K-k}\mathcal{K}_{kj}\frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^{L(k+j)}} \\ &\times\int_0^\infty \log_2\left(\frac{1+x\varepsilon_k}{1+x\tilde{\varepsilon}_k}\right)x^{N-V}e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}\gamma(N-V+1,x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L(k+j)-1}dx.\end{aligned} \tag{23}$$

where $\varepsilon_k = \rho(\alpha_{gk}^2 + \mathcal{P}_k)$ and $\tilde{\varepsilon}_k = \rho\mathcal{P}_k$.

Proof: Please, see Appendix D.

As one can notice, the ergodic sum-rate expression of Proposition III is not a closed-form solution. The reason for this is that the integral in (23) is quite challenging to solve. Even so, it is still useful, since (23) can be evaluated very efficiently through numerical methods. An exact and simpler analytical solution for (22) can be obtained if we consider the case in which the number of receive antennas is equal to the number of effective data streams, i.e. $N = V$, as derived in Proposition IV.

Proposition IV: For the particular case when $N = V$, a closed form solution for the ergodic sum-rate can be obtained as

$$\begin{aligned}\bar{R}_g =& \sum_{k=1}^{K}\sum_{j=0}^{K-k}\sum_{i=0}^{L(k+j)-1}\frac{\mathcal{K}_{kj}L}{(i+1)\ln(2)}\tbinom{L(k+j)-1}{i}(-1)^i\left[\mathrm{Ei}\left(-\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\tilde{\varepsilon}_k}\right)\right. \\ &\left.\times\, e^{\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\tilde{\varepsilon}_k}} - \mathrm{Ei}\left(-\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\varepsilon_k}\right)e^{\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\varepsilon_k}}\right].\end{aligned} \tag{24}$$

where ε_k and $\tilde{\varepsilon}_k$ are defined in the same way as in Proposition III.

Proof: Please, see Appendix E.

IV. Schemes for Performance Comparison

In order to show the advantages of our proposed multi-array strategy, we compare its performance with some conventional schemes, with and without the exploration of time diversity. For the implementation of these schemes, we consider the same geometrical scenario as the adopted in the SSAA design, in which we assume the existence of S scattering spatial clusters with G sub-groups of K users. In addition, for a fair comparison, the clusters are considered to be located within the same azimuth angles and have the same angular spread as in SSAA. The main difference between the conventional schemes and the system operating with the SSAA approach is that all antennas are activated at each transmission, i.e. the full transmit array is activated. Besides, for the systems that employ time diversity, in order to obtain different channel responses, the time separation between retransmissions must be greater than the channel coherent time. More details about each particular scheme are provided next.

a) MIMO-NOMA with time diversity: This first scheme consists of a conventional massive MIMO-NOMA system, similar as in [5], but with the difference that each superposed symbol intended for each NOMA sub-group is redundantly transmitted over T instants of time. In this implementation, the users recover their symbols by selecting one of the T receptions, the one that achieves the highest effective channel gain magnitude.

b) MIMO-OMA with time diversity: For this conventional OMA time diversity scheme, we suppose that there is only one user per sub-group, what means that $G = K$. Aiming to provide fairness in the performance comparisons, the beamforming and detection matrices are constructed identically as in the MIMO-NOMA time diversity system, where each user selects its desired symbol from the best of the T receptions. In addition, the total transmit power available for each sub-group is entirely allocated to the only existing user, that is, the power allocation coefficient for each user in the MIMO-OMA system is adjusted to unity.

c) MIMO-NOMA and MIMO-OMA without diversity: These two schemes consist of conventional full array implementations, the MIMO-NOMA system from [5] and the MIMO-OMA counterpart. In contrast to the time diversity schemes, for these implementations, each distinct symbol is transmitted only one time, so no diversity is explored.

V. Numerical Results and Discussions

In this section, some illustrative numerical examples of the proposed system operating with the SSAA technique are presented and compared with the conventional massive MIMO-NOMA and OMA systems described in Section IV. The total number of transmit antennas at the BS is set to $M = 90$, in which, for comparison purposes, we adjust the number of sub-arrays in MIMO-NOMA with SSAA to be equal to the number of time slots in the full array time diversity

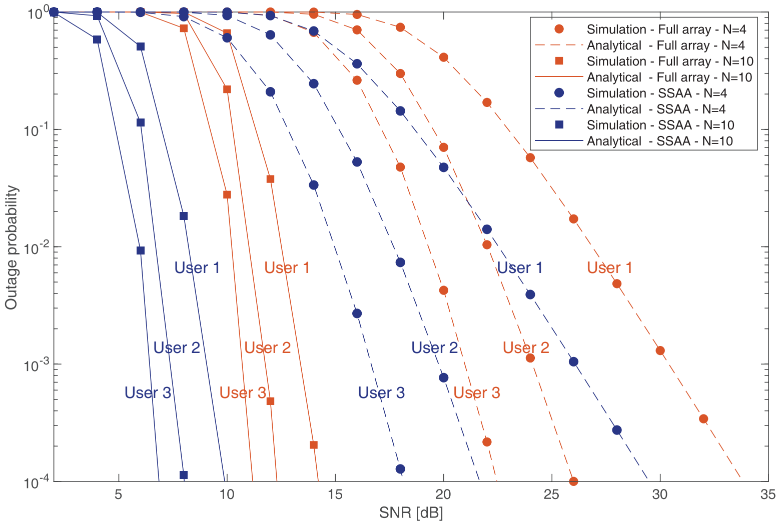

Fig. 2. Outage probability versus transmit SNR for massive MIMO-NOMA system with SSAA technique and conventional full array system operating in time diversity mode. $L = T = 3$; $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

implementation, i.e. $L = T$. In addition, we consider a scenario with $S = 4$ scattering clusters, where, in each cluster, we assume the existence of 12 users that are further subdivided into $G = 4$ sub-groups with $K = 3$ users each, and we configure the beamformers to deliver $V = 4$ effective data streams at each transmission for each of the clusters. The angular spread is set to $\Delta_s = 10^\circ$, and the azimuth angle for the sth cluster is chosen as $\theta_s = \frac{\pi}{4} + \frac{\pi}{2}\left(\frac{s-1}{S-1}\right)$, for $s = 1, \cdots, S$. Furthermore, in order to maximize the array gain, the azimuth angle of the BS is directed to the cluster of interest.

Fig. 2 shows the outage probability in terms of transmit SNR. As can be seen, a perfect agreement among simulated and analytical curve is observed. In addition, it can be noticed that the proposed scheme provides remarkable outage performance improvements to massive MIMO-NOMA systems. For example, when employing either $N = 4$ or $N = 10$, all users adopting the SSAA strategy requires roughly 5dB less SNR to achieve the same outage level of that achieved with the full array time diversity scheme, and when $N = 10$, the worst user in SSAA can outperform even the best user in the time diversity counterpart. These performance gains can be explained as follows. Full array massive MIMO systems operating in strongly correlated scenarios, as considered in this work, experience bad conditioned channel matrices, i.e., the channel covariance matrices are rank-deficient. This characteristic is detrimental to the performance of the zero-forcing receivers employed at the users. On its turn, the array structure of

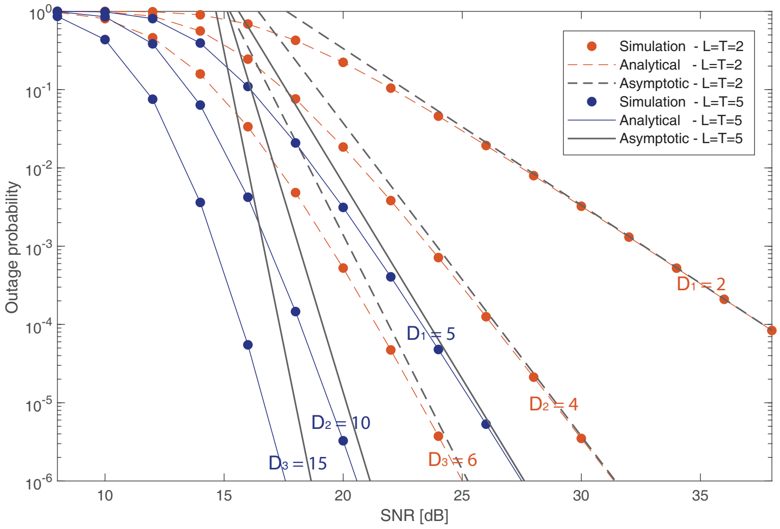

Fig. 3. Exact and asymptotic outage probability curves for massive MIMO-NOMA system operating with the proposed SSAA technique. $N = 4$; $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

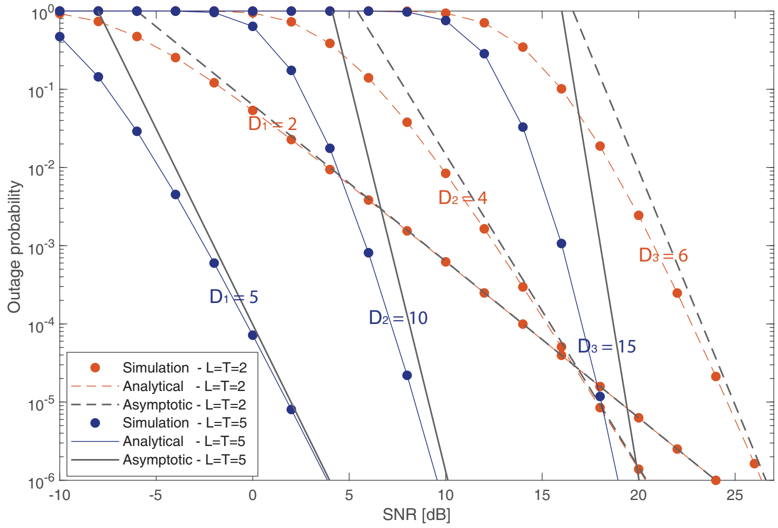

Fig. 4. Exact and asymptotic outage probability curves for massive MIMO-NOMA system operating with the proposed SSAA technique. $N = 4$; $\alpha_1^2 = 0.72, \alpha_2^2 = 0.18$, and $\alpha_3^2 = 0.1$; $R_1 = 0.5, R_2 = 1$ and $R_3 = 4.5$ BPCU.

the proposed SSAA scheme alleviates the mentioned correlation, i.e., offers better conditioned channel matrices. This explains the performance superiority observed in the SSAA diversity scheme. Figs. 3 and 4 bring the validation of the high-SNR analysis derived in Section III-C. One can observe that, for a fixed number of transmit and receive antennas, the system diversity order increases as the number of sub-arrays gets higher, which is in total concordance with the diversity order expression in (21). Moreover, we see that changing the values of target rates and

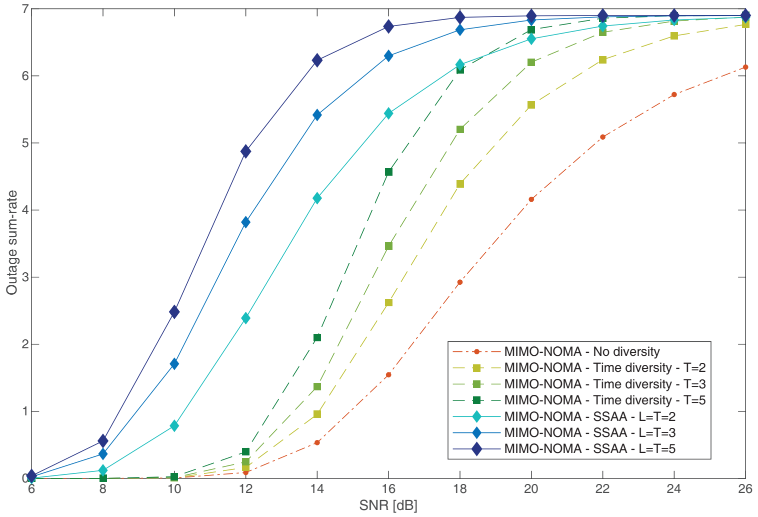

Fig. 5. Outage sum-rate for the proposed SSAA technique and the conventional full array time diversity approach in massive MIMO-NOMA systems. $N = 4$; $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

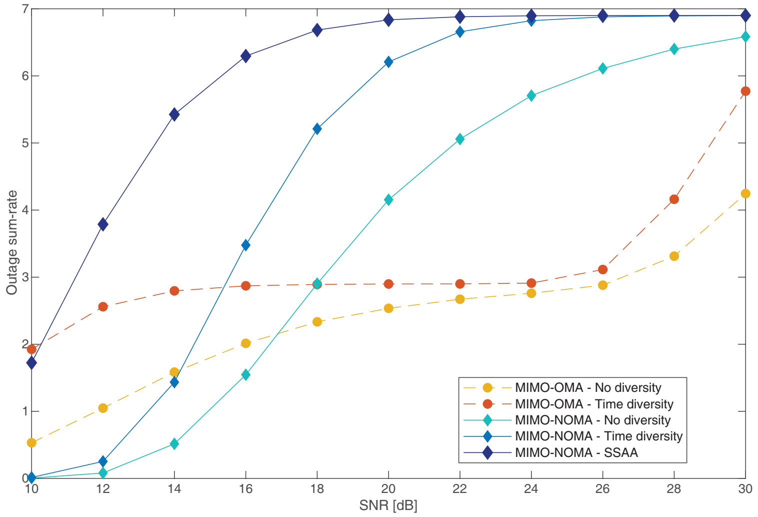

Fig. 6. Outage sum-rate for the proposed SSAA technique and various conventional schemes in massive MIMO-NOMA systems. $N = 4$; $L = T = 3$; $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

power allocation coefficients does not affect the diversity order of the users, what validates again our analysis.

Fig. 5 plots the outage sum-rate of the system, which is defined by $\sum_{k=1}^{K}(1 - P_{gk})R_{gk}$. The outage sum-rate informs the sum of the users' average throughput achieved when the BS is transmitting at a constant data rate. Once again, the benefits that the SSAA scheme can provide to massive MIMO-NOMA networks are noticeable. As one can observe, for all values of T,

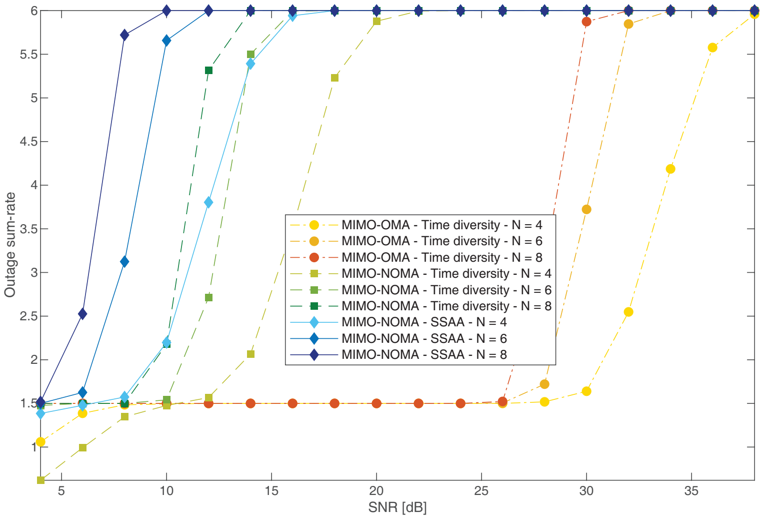

Fig. 7. Outage sum-rate for the proposed SSAA technique and the conventional full array time diversity approach in massive MIMO-NOMA systems. $L = 3$; $\alpha_1^2 = 0.72, \alpha_2^2 = 0.18$, and $\alpha_3^2 = 0.1$; $R_1 = 0.5, R_2 = 1$ and $R_3 = 4.5$ BPCU.

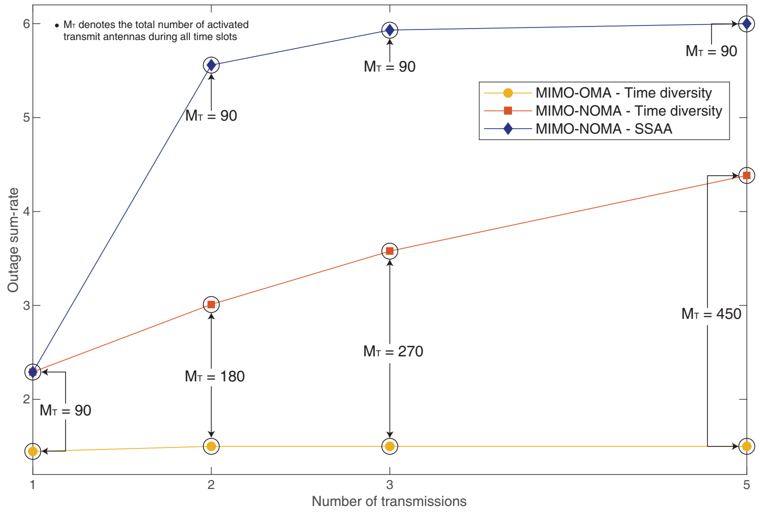

Fig. 8. Outage sum-rate versus number of redundant transmissions for a fixed transmit SNR of 16dB. $N = 4$; $\alpha_1^2 = 0.72, \alpha_2^2 = 0.18$, and $\alpha_3^2 = 0.1$; $R_1 = 0.5, R_2 = 1$ and $R_3 = 4.5$ BPCU.

the proposed strategy outperforms the conventional time diversity full array setup. For example, when $T = 3$ and SNR = 12dB, the SSAA scheme achieves an outage sum-rate of 3.76 BPCU against only 0.25 BPCU from the full array system, which represents a spectral gain of almost 15 times. The expressive gain that SSAA can achieve over the system without diversity becomes also evident in this figure, in which a gap of almost 8dB can be observed. In Fig. 6, for $L = T = 3$, the outage sum-rate of the SSAA scheme is compared with conventional full array systems,

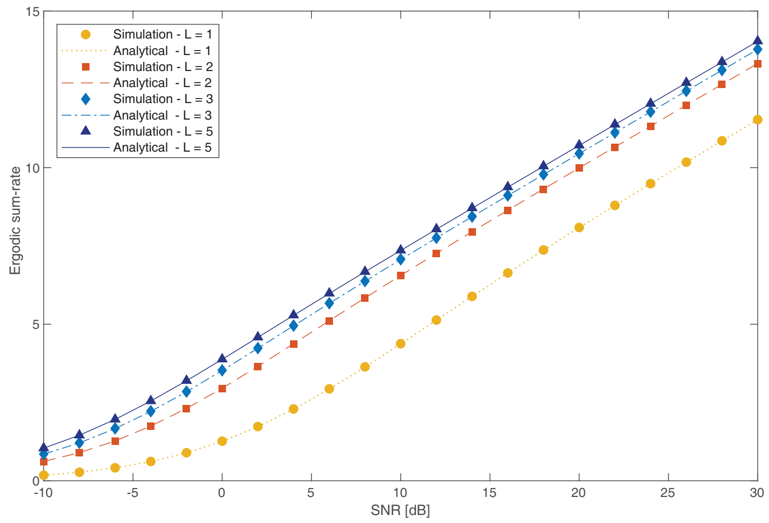

Fig. 9. Simulated and analytical (generated with (24)) ergodic sum-rate curves versus transmit SNR for massive MIMO-NOMA system operating with the proposed SSAA technique. $N = V = 4$; $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

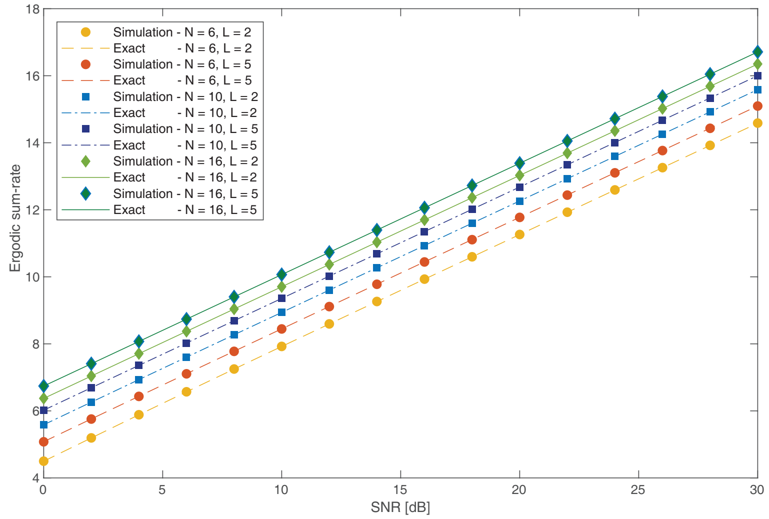

Fig. 10. Simulated and exact (generated with (23)) ergodic sum-rate versus transmit SNR for massive MIMO-NOMA system operating with the proposed SSAA technique. $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

with and without time diversity. It can be observed that our MIMO-NOMA design with SSAA outperforms all the other systems. In particular, considering a transmit SNR of 18dB, the SSAA scheme achieves an outage sum-rate of 6.68 BPCU against 5.21 BPCU of the time diversity MIMO-NOMA system and 2.91 BPCU of the MIMO-NOMA system without diversity. The performance gains over the OMA implementations are even more expressive, reaching up a

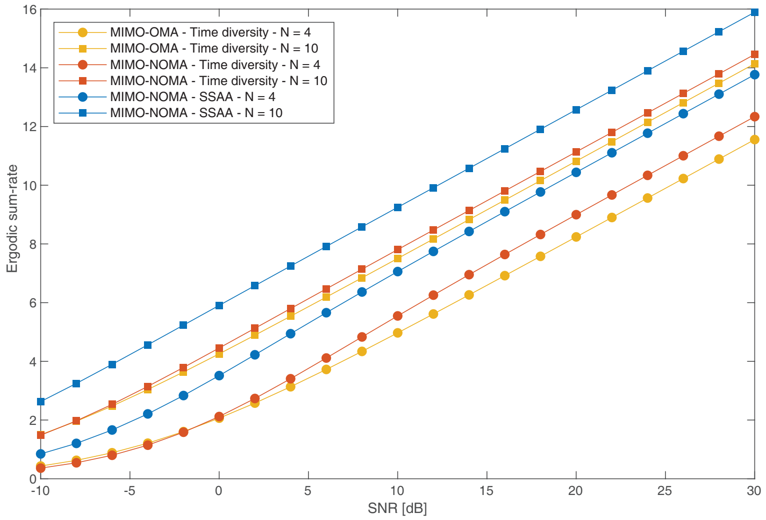

Fig. 11. Ergodic sum-rate versus transmit SNR for massive MIMO-NOMA with SSAA and the conventional time diversity counterparts. $L = 3$; $\alpha_1^2 = 0.625, \alpha_2^2 = 0.25$ and $\alpha_3^2 = 0.125$; $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ BPCU.

performance gap of 4.35 BPCU over the MIMO-OMA system without diversity.

Fig. 7 brings the outage sum-rate curves for various values of receive antennas considering a different set of power allocation and target rates. As one can notice, the gains achieved by the SSAA scheme over the full array time diversity MIMO-OMA setups are enormous, in which, for all values of N, the curves of the two schemes are separated by more than 20dB. Fig. 8 shows the outage sum-rate versus the number of redundant transmissions for a fixed SNR of 16dB. It can be seen again the superior outage sum-rate performance of SSAA for a fixed SNR setup. For instance, with 3 redundant transmissions, the massive MIMO-NOMA system with SSAA achieves a rate of 5.93 BPCU, which is 2.3 BPCU higher than that of the full array MIMO-NOMA system and almost 4 times higher than what is achieved by the MIMO-OMA counterpart. For 5 redundant transmissions, the performance gains of SSAA becomes even more prominent. In addition, as it can be observed, another advantage of our proposed strategy is that, independently of how many retransmissions are performed, the total number of activated antenna elements remains constant, i.e. $M = 90$, what does not happen for the full array schemes.

Figs. 9 and 10 validates the analytical development for the ergodic sum-rate in Section III-D. In particular, Fig. 9 shows results for various values of L when $N = V$. For this case, the analytical curves are generated with (24), in which a perfect agreement with the simulation results is verified. The cases when $N > V$ are shown in Fig. 10. In this latter figure, the expression in (23) is used to plot the exact curves. Again, for all values of N and L, we can observe perfect correspondence among exact and simulated results. Moreover, through both figures, it can be noticed that the system ergodic sum-rate is improved with the increase in the number of sub-arrays. For instance, in Fig. 9, for a transmit SNR of 10dB, the system operating with $L = 5$ sub-arrays achieves an ergodic sum-rate of 7.35 BPCU, which is 2.98 BPCU superior than the achieved with $L = 1$.

Finally, in Fig. 11 the ergodic sum-rate performance of MIMO-NOMA and MIMO-OMA diversity systems are compared, for $N = 4$ and $N = 10$ receive antennas. As it can be seen, the superiority of the proposed SSAA technique is demonstrated again in terms of ergodic sum-rate. For example, when $N = 10$ and the SNR is 18dB, the MIMO-NOMA system operating with SSAA achieves 1.44 BPCU more sum-rate than the full array time diversity MIMO-NOMA setup and, for $N = 4$, the gap reaches up 1.45 BPCU. If we take into consideration the MIMO-OMA full array system, the gains are even more remarkable.

VI. Conclusions

In this paper, by partitioning the transmit antennas into multiple sub-arrays at the BS, we have proposed and investigated a novel low-complexity diversity scheme for massive MIMO-NOMA deployments. Detailed design of beamformers and detection matrices were presented, and a full in-depth analytical analysis was carried out. In particular, closed-form expressions for the outage probability and ergodic sum-rate were derived. A high SNR asymptotic outage analysis was also conducted, in which the diversity order achieved with the proposed protocol was determined. Furthermore, representative numerical examples were presented to corroborate the analytical analysis. In all results, the proposed scheme outperformed conventional full array massive MIMO-OMA and MIMO-NOMA systems operating with and without the exploration of time diversity. Besides, it became clear the superiority of SSAA in terms of energy consumption, implementation complexity, latency, and feedback overhead. This makes our proposal attractive for applications that require an enhanced performance but have limited resources and restricted computational capabilities.

Note that, even though we have chosen to transmit the symbol replicas in different instants of time, we could have separated the transmissions by exploring different domains, such as frequency or code. Also, concepts of STBC could be combined together with SSAA to achieve full spatial diversity. However, these possibilities arise as potential future works.

Appendix A
Proof of Lemma I

From (15), one can see that the kth user decodes the ith weaker message, for $1 \leq i \leq k < K$, with the following SINR

$$\begin{aligned}\text{SINR}_{gk}^{i} &= \frac{\mathbb{E}[|\alpha_{gi}x_{gi}|^2]}{\mathbb{E}\left[\sum_{j=i+1}^{K}|\alpha_{gj}x_{gj}|^2\right] + \mathbb{E}[|[\mathbf{H}_{gk}^{m\dagger}\mathbf{n}_{gk}^{m}]_g|^2]} \\ &= \frac{\alpha_{gi}^2}{\sum_{j=i+1}^{K}\alpha_{gj}^2 + \sigma_n^2\mathbb{E}[\text{tr}\{[\mathbf{H}_{gk}^{m\dagger}(\mathbf{H}_{gk}^{m\dagger})^H]_{gg}\}]} = \frac{\frac{1}{\|[\mathbf{H}_{gk}^{m\dagger}]_{g*}\|^2}\alpha_{gi}^2}{\frac{1}{\|[\mathbf{H}_{gk}^{m\dagger}]_{g*}\|^2}\sum_{j=i+1}^{K}\alpha_{gj}^2 + \sigma_n^2}. \qquad \text{(A-1)}\end{aligned}$$

Since $m \in \{1, \cdots, L\}$ corresponds to the signal reception with the highest effective channel gain magnitude, we define

$$\gamma_{gk} = \max\left\{\varsigma_{gk}^{1}, \cdots, \varsigma_{gk}^{L}\right\}. \qquad \text{(A-2)}$$

where $\varsigma_{gk}^{l} = \frac{1}{\|[\mathbf{H}_{gk}^{l\dagger}]_{g*}\|^2}$, for $1 \leq l \leq L$. Now, replacing (A-2) in (A-1) and denoting the transmit SNR by $\rho = \frac{1}{\sigma_n^2}$, we obtain

$$\text{SINR}_{gk}^{i} = \frac{\gamma_{gk}\alpha_{gi}^2}{\gamma_{gk}\sum_{j=i+1}^{K}\alpha_{gj}^2 + \frac{1}{\rho}}. \qquad \text{(A-3)}$$

Note that, since the user K is the strongest one, when $i = k = K$, the ith message will be recovered without any interference. Then, we can represent the term corresponding to the power of interfering users in (A-3) as

$$\mathcal{P}_i = \begin{cases}\sum_{j=i+1}^{K}\alpha_{gj}^2, & \text{for} \quad 1 \leq i \leq k < K, \\ 0, & \text{for} \quad i = k = K.\end{cases} \qquad \text{(A-4)}$$

Finally, by substituting (A-4) in (A-3), the SINR expression can be attained as

$$\text{SINR}_{gk}^{i} = \frac{\rho\gamma_{gk}\alpha_{gi}^2}{\rho\gamma_{gk}\mathcal{P}_i + 1}, \qquad 1 \leq i \leq k \leq K, \qquad \text{(A-5)}$$

which completes the proof.

APPENDIX B

PROOF OF PROPOSITION II

By replacing (16) in (18) and performing some algebraic manipulations, we get the following

$$P_{gk} = \Pr\left[\log_2\left(1 + \frac{\rho\gamma_{gk}\alpha_{gi}^2}{\rho\gamma_{gk}\mathcal{P}_i + 1}\right) < R_{gi}\right] = \Pr\left[\gamma_{gk} < \frac{2^{R_{gi}} - 1}{\rho[\alpha_{gi}^2 - \mathcal{P}_i(2^{R_{gi}} - 1)]}\right] = P\left[\gamma_{gk} < \mathcal{M}_{gk}\right], \tag{B-1}$$

where

$$\mathcal{M}_{gk} = \max_{1\leq i\leq k}\left\{\frac{2^{R_{gi}} - 1}{\rho[\alpha_{gi}^2 - \mathcal{P}_i(2^{R_{gi}} - 1)]}\right\}. \tag{B-2}$$

The expression (B-1) corresponds to the cumulative distribution function (CDF) of γ_{gk}. By analyzing (A-1), one can verify that the effective channel gain obtained at each reception l, for $l = 1, \cdots, L$, is equivalent to the inverse of the gth main diagonal element of the matrix $\hat{\mathbf{R}} = \mathbf{H}_{gk}^{m\dagger}(\mathbf{H}_{gk}^{m\dagger})^H \in \mathbb{C}^{N\times N}$, which can be expanded as

$$\hat{\mathbf{R}} = (\mathbf{B}^H\mathbf{H}_{gk}^l(\mathbf{H}_{gk}^l)^H\mathbf{B})^{-1}\mathbf{B}^H\mathbf{H}_{gk}^l(\mathbf{H}_{gk}^l)^H\mathbf{B}(\mathbf{B}^H\mathbf{H}_{gk}^l(\mathbf{H}_{gk}^l)^H\mathbf{B})^{-1} = (\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}. \tag{B-3}$$

As stated in [5], the matrix in (B-3) follows the inverse Wishart distribution and, consequently, the inverse of its main diagonal elements follows the Gamma distribution [36]. As a result, the effective channel gains delivered by the L sub-arrays can be seen as L independent and identically distributed Gamma random variables. Therefore, first considering unordered gains, the CDF of $\max\left\{\varsigma_{gk}^1, \cdots, \varsigma_{gk}^L\right\}$ is given by

$$F_{\max}(x) = \left[\frac{\gamma\left(N - V + 1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}\right)}{\Gamma(N - V + 1)}\right]^L, \tag{B-4}$$

and its probability density function (PDF) can be derived as

$$f_{\max}(x) = \frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^L}x^{N-V}e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}\gamma(N - V + 1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L-1}. \tag{B-5}$$

Consequently, the PDF for the ordered effective channel gains γ_{gk} can be obtained as

$$
\begin{aligned}
f_{\gamma_{gk}}(x) &= \sum_{j=0}^{K-k} K\binom{K-1}{k-1}\binom{K-k}{j}(-1)^j f_{\max}(x) F_{\max}(x)^{k-1+j} \\
&= \sum_{j=0}^{K-k} \mathcal{K}_{kj} \frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^{L(k+j)}} x^{N-V} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} \gamma(N-V+1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L(k+j)-1},
\end{aligned}
\tag{B-6}
$$

where, for easy of notation, we have defined $\mathcal{K}_{kj} = K\binom{K-1}{k-1}\binom{K-k}{j}(-1)^j$.

At last, the closed-form expression for the general outage probability of the proposed system is obtained by integrating (B-6), as follows

$$
\begin{aligned}
P_{gk} &= \sum_{j=0}^{K-k} \mathcal{K}_{kj} \frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^{L(k+j)}} \int_0^{\mathcal{M}_{gk}} x^{N-V} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} \gamma(N-V+1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L(k+j)-1} dx \\
&= \sum_{j=0}^{K-k} \frac{\mathcal{K}_{kj}}{k+j} \left[\frac{\gamma\left(N-V+1, \mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}\right)}{\Gamma(N-V+1)}\right]^{L(k+j)},
\end{aligned}
\tag{B-7}
$$

which completes the proof.

Appendix C
Proof of Proposition II

By applying the series representation of the Gamma function in (19), we have [37]

$$
\begin{aligned}
P_{gk} &= \sum_{j=0}^{K-k} \frac{\mathcal{K}_{kj}}{(k+j)[(N-V)!]^{L(k+j)}} [(N-V)!]^{L(k+j)} \\
&\quad \times \left(1 - e^{-\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} \sum_{n=0}^{N-V} \frac{(\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^n}{n!}\right)^{L(k+j)} \\
&= \sum_{j=0}^{K-k} \frac{\mathcal{K}_{kj}}{k+j} \left(e^{-\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} \sum_{n=N-V+1}^{\infty} \frac{(\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^n}{n!}\right)^{L(k+j)}.
\end{aligned}
\tag{C-2}
$$

Then, by exploring properties of the Taylor's series and performing some algebraic manipulations in (C-2), we obtain the desired high-SNR approximation as

$$
P_{gk} \approx \frac{K}{k}\binom{K-1}{k-1} \frac{[\rho\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}]^{(N-V+1)Lk}}{\rho^{(N-V+1)Lk}[(N-V+1)!]^{Lk}}.
\tag{C-3}
$$

Consequently, the diversity order experienced by the kth user is determined by

$$D_k = (N - V + 1)\, Lk, \tag{C-4}$$

which completes the proof.

Appendix D
Proof of Proposition III

The expression in (22) can be rearranged as

$$\bar{R}_g = \mathbb{E}\left[\sum_{k=1}^{K} \log_2\left(\frac{1 + \gamma_{gk}\rho(\alpha_{gk}^2 + \mathcal{P}_k)}{1 + \gamma_{gk}\rho\mathcal{P}_k}\right)\right]. \tag{D-1}$$

By invoking the PDF of ordered channel gains in (B-6), and defining $\varepsilon_k = \rho(\alpha_{gk}^2 + \mathcal{P}_k)$ and $\tilde{\varepsilon}_k = \rho\mathcal{P}_k$, the exact ergodic sum-rate can be evaluated by

$$\begin{aligned}\bar{R}_g = \sum_{k=1}^{K} \mathbb{E}\left[\log_2\left(\frac{1 + \gamma_{gk}\varepsilon_k}{1 + \gamma_{gk}\tilde{\varepsilon}_k}\right)\right] &= \sum_{k=1}^{K}\sum_{j=0}^{K-k} \mathcal{K}_{kj} \frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^{L(k+j)}} \\ \times \int_0^{\infty} \log_2\left(\frac{1 + x\varepsilon_k}{1 + x\tilde{\varepsilon}_k}\right) & x^{N-V} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} \gamma(N - V + 1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L(k+j)-1} dx.\end{aligned} \tag{D-2}$$

which completes the proof.

Appendix E
Proof of Proposition IV

When $N = V$, the PDF for the effective channel gains in (B-6) can be simplified to

$$f_{\gamma_{gk}}(x) = \sum_{j=0}^{K-k} \mathcal{K}_{kj} L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} \gamma(1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L(k+j)-1}. \tag{E-1}$$

Then, by expanding the incomplete gamma function in (E-1) by its series representation [37] and performing some manipulations, the PDF becomes

$$
\begin{aligned}
f_{\gamma_{gk}}(x) &= \sum_{j=0}^{K-k} \sum_{i=0}^{L(k+j)-1} \mathcal{K}_{kj} L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} (-e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}})^i \binom{L(k+j)-1}{i} \\
&= \sum_{j=0}^{K-k} \sum_{i=0}^{L(k+j)-1} \mathcal{K}_{kj} L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg} \binom{L(k+j)-1}{i} (-1)^i e^{-x(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}. \qquad \text{(E-2)}
\end{aligned}
$$

Given the above simplification, the ergodic sum-rate expression obtained in (D-2) can be rewritten as

$$
\begin{aligned}
\bar{R}_g &= \sum_{k=1}^{K} \sum_{j=0}^{K-k} \sum_{i=0}^{L(k+j)-1} \mathcal{K}_{kj} L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg} \binom{L(k+j)-1}{i} (-1)^i \int_0^\infty \log_2\left(\frac{1+x\varepsilon_k}{1+x\tilde{\varepsilon}_k}\right) e^{-x(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} dx \\
&= \sum_{k=1}^{K} \sum_{j=0}^{K-k} \sum_{i=0}^{L(k+j)-1} \frac{\mathcal{K}_{kj} L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\ln(2)} \binom{L(k+j)-1}{i} (-1)^i \\
&\times \left[\int_0^\infty \ln\left(1+x\varepsilon_k\right) e^{-x(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} dx - \int_0^\infty \ln\left(1+x\tilde{\varepsilon}_k\right) e^{-x(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}} dx\right]. \qquad \text{(E-3)}
\end{aligned}
$$

The integrals in (E-3) are of the form $\int_0^\infty \ln(1+\beta x)e^{-\mu x}dx$, which solution is given in [37]. Then, by applying the refereed result, we can finally obtain the desired expression for the ergodic sum-rate, as follows

$$
\begin{aligned}
\bar{R}_g = &\sum_{k=1}^{K} \sum_{j=0}^{K-k} \sum_{i=0}^{L(k+j)-1} \frac{\mathcal{K}_{kj} L}{(i+1)\ln(2)} \binom{L(k+j)-1}{i} (-1)^i \left[\mathrm{Ei}\left(-\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\tilde{\varepsilon}_k}\right)\right. \\
&\times \left. e^{\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\tilde{\varepsilon}_k}} - \mathrm{Ei}\left(-\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\varepsilon_k}\right) e^{\frac{(i+1)[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}{\varepsilon_k}}\right], \qquad \text{(E-4)}
\end{aligned}
$$

which completes the proof.

References

[1] J. Stryjak and M. Sivakumaran, "The mobile economy 2019." *GSMA Intelligence*, Feb. 2019.

[2] G. A. Akpakwu, B. J. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A survey on 5g networks for the internet of things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.

[3] I. B. F. de Almeida, L. L. Mendes, J. J. P. C. Rodrigues, and M. A. A. da Cruz, "5g waveforms for iot applications," *IEEE Commun. Surv. Tuts.*, pp. 1–1, 2019.

[4] G. Liu, Y. Huang, F. Wang, J. Liu, and Q. Wang, "5G features from operation perspective and fundamental performance validation by field trial," *China Commun.*, vol. 15, no. 11, pp. 33–50, Nov. 2018.

[5] Z. Ding and V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, May 2016.

[6] 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.

[7] W. A. Al-Hussaibi and F. H. Ali, “Efficient user clustering, receive antenna selection, and power allocation algorithms for massive MIMO-NOMA systems,” *IEEE Access*, vol. 7, pp. 31 865–31 882, Feb. 2019.

[8] 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 Commu. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.

[9] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, “Millimeter wave communication: A comprehensive survey,” *IEEE Commun. Surv. Tuts.*, vol. 20, no. 3, pp. 1616–1653, Jun. 2018.

[10] X. Chen, J. Lu, P. Fan, and K. B. Letaief, “Massive mimo beamforming with transmit diversity for high mobility wireless communications,” *IEEE Access*, vol. 5, pp. 23 032–23 045, Oct. 2017.

[11] J. Park and B. Clerckx, “Multi-user linear precoding for multi-polarized massive mimo system under imperfect csit,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 05, May 2015.

[12] Y. Li and G. A. Aruma Baduge, “Noma-aided cell-free massive mimo systems,” *IEEE Wireless Commun. Let.*, vol. 7, no. 6, pp. 950–953, Dec. 2018.

[13] D. Zhang, Z. Zhou, C. Xu, Y. Zhang, J. Rodriguez, and T. Sato, “Capacity analysis of noma with mmwave massive mimo systems,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1606–1618, Jul. 2017.

[14] V. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O. Shin, “Precoder design for signal superposition in MIMO-NOMA multicell networks,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2681–2695, Dec. 2017.

[15] X. Sun, N. Yang, S. Yan, Z. Ding, D. W. K. Ng, C. Shen, and Z. Zhong, “Joint beamforming and power allocation in downlink noma multiuser mimo networks,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5367–5381, Aug. 2018.

[16] H. D. Tuan, A. A. Nasir, H. H. Nguyen, T. Q. Duong, and H. V. Poor, “Non-orthogonal multiple access with improper gaussian signaling,” *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 496–507, Jun. 2019.

[17] Z. Mobini, M. Mohammadi, B. K. Chalise, H. A. Suraweera, and Z. Ding, “Beamforming design and performance analysis of full-duplex cooperative noma systems,” *IEEE Trans. Wireless Commun.*, pp. 1–1, 2019.

[18] Q. Li, M. Wen, E. Basar, H. V. Poor, and F. Chen, “Spatial modulation-aided cooperative noma: Performance analysis and comparative study,” *IEEE J. Sel. Topics Signal Process.*, pp. 1–1, 2019.

[19] T. N. Do, D. B. da Costa, T. Q. Duong, and B. An, “Improving the performance of cell-edge users in miso-noma systems using tas and swipt-based cooperative transmissions,” *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 49–62, Mar. 2018.

[20] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, “Performance analysis of noma-sm in vehicle-to-vehicle massive mimo channels,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2653–2666, Dec. 2017.

[21] T. Hou, Y. Liu, Z. Song, X. Sun, and Y. Chen, “Multiple antenna aided noma in uav networks: A stochastic geometry approach,” *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1031–1044, Feb 2019.

[22] M. Gong, Z. Yang, and B. Lyu, “Antenna diversity for downlink mimo-noma systems with partial channel state information,” *IEEE Commun. Let.*, vol. 22, no. 10, pp. 2172–2175, Oct. 2018.

[23] M. Gong and Z. Yang, “The application of antenna diversity to NOMA with statistical channel state information,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3755–3765, Apr. 2019.

[24] M. Toka and O. Kucur, “Non-orthogonal multiple access with Alamouti space-time block coding,” *IEEE Commun. Let.*, vol. 22, no. 9, pp. 1954–1957, Sep. 2018.

[25] Z. Pan, W. Liu, J. Lei, J. Luo, L. Wen, and C. Tang, "Multi-dimensional space time block coding aided downlink mimo-scma," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2019.

[26] M. F. Kader and S. Y. Shin, "Cooperative relaying using space-time block coded non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 5894–5903, Jul. 2017.

[27] J. Zhao, Z. Ding, P. Fan, Z. Yang, and G. K. Karagiannidis, "Dual relay selection for cooperative noma with distributed space time coding," *IEEE Access*, vol. 6, pp. 20 440–20 450, 2018.

[28] H. R. Ahmed, E. Sourour, and H. M. Elkamchouchi, "Analysis for noma-comp-jt global precoding matrix and irc receiver for lte-a," in *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, Apr. 2016, pp. 1–6.

[29] H. Wang, S. Leung, and R. Song, "Precoding design for two-cell mimo-noma uplink with comp reception," *IEEE Commun. Let.*, vol. 22, no. 12, pp. 2607–2610, Dec. 2018.

[30] Y. Yu, H. Chen, Y. Li, Z. Ding, L. Song, and B. Vucetic, "Antenna selection for MIMO nonorthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3158–3171, Apr. 2018.

[31] A. S. Khan, I. Chatzigeorgiou, S. Lambotharan, and G. Zheng, "Network-coded noma with antenna selection for the support of two heterogeneous groups of users," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1332–1345, Feb. 2019.

[32] D. Shiu, G. Foschini, M. Gans, and J. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.

[33] 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.

[34] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[35] Y. Gao, H. Vinck, and T. Kaiser, "Massive mimo antenna selection: Switching architectures, capacity bounds, and optimal antenna selection algorithms," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1346–1360, Mar. 2018.

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

[37] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.