Successive Sub-Array Activation for Massive MIMO-NOMA Networks

Arthur S. de Sena*, Daniel B. da Costa†, Zhiguo Ding‡, Pedro H. J. Nardelli*,
Ugo S. Dias§ and Constantinos B. Papadias¶
* Lappeenranta University of Technology, Finland † Federal University of Ceará, Brazil
‡ The University of Manchester, UK § University of Brasília, Brazil ¶ Athens Information Technology, Greece
Emails: arthurssena@ieee.org, danielbcosta@ieee.org, zhiguo.ding@manchester.ac.uk,
pedro.nardelli@lut.fi, ugodias@ieee.org, papadias@ait.edu.gr

***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). A single-cell multi-cluster downlink scenario is considered, where the base station (BS) sends redundant symbols through multiple transmit sub-arrays to multi-antenna receivers. An in-depth analytical analysis is carried out, in which an exact closed-form expression for the outage probability is derived. Also, a high signal-to-noise ratio (SNR) outage approximation is obtained and the system diversity order is determined. Our results show that the proposed scheme outperforms conventional full array massive MIMO setups.**

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

I. Introduction

The 5th generation of wireless communication networks (5G) will enable unforeseen new services and applications with the most diverse requirements, such as massive connectivity, low latency, and high reliability [1]. 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 power domain, in which successive interference cancellation (SIC) is employed for reception. Both technologies can reduce the system latency and increase the connectivity capacity. In addition, the combination of MIMO and NOMA can provide remarkable improvements and outperform conventional orthogonal multiple access (OMA) systems [2], [3].

Even though massive MIMO-NOMA can provide enormous advantages over classical MIMO-OMA deployments, in some scenarios, the fast fading channels can still degrade the system performance and reduce the communication quality. In such environments, diversity techniques are crucial for enhancing the system performance and improve reliability. The application of diversity strategies in classical communication systems have been exhaustively investigated for decades. However, only few works have addressed the refereed subject in massive MIMO-NOMA setups. In [4], antenna diversity was employed to improve the outage performance of a MIMO-NOMA system assuming a simple scenario with only one NOMA group. In [5], the combination of Alamouti space-time block coding and MIMO-NOMA was investigated, in which a closed-form outage probability expression was derived. Performance improvements in MIMO-NOMA systems can also be achieved through antenna selection schemes [6]. 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.

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 potential of improving the outage performance of each user in a massive MIMO-NOMA deployment. This is achieved by successively activating antenna sub-arrays at the base station (BS) and exploring space diversity in different instants of time at the users' side. With this strategy, the system latency can still be maintained low. Furthermore, we perform an in-depth analytical analysis of the proposed design, in which an exact closed-form expression for the outage probability is derived. The system behavior at high signal-to-noise ratio (SNR) regime is also investigated, where an asymptotic outage approximation is attained, enabling us to determine the system diversity order. Moreover, numerical 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.

Notation and Special Functions: Bold-faced lower-case letters represent vectors and upper-case letters denote matrices. The norm and the i-th 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 i-th 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 $\mathrm{tr}\{\mathbf{A}\}$. $\mathbf{I}_M$ represents the identity matrix of dimension $M \times M$, and $\mathbf{0}_{M\times N}$ denotes

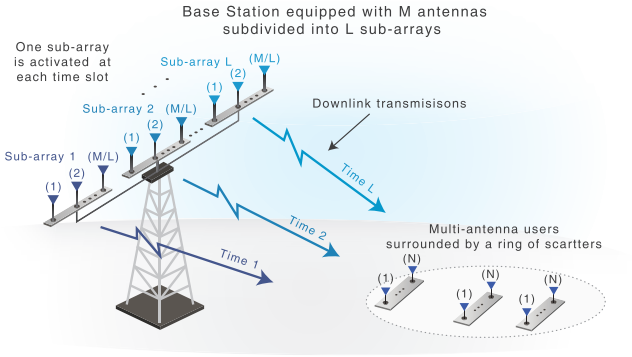

Fig. 1. System model. The BS is equipped with multiple antennas subdivided into multiple sub-arrays.

the $M \times N$ matrix with all zero entries. In addition, $\otimes$ represents the Kronecker product, $E[\cdot]$ denotes expectation, $\Gamma(\cdot)$ is the Gamma function, and $\gamma(\cdot,\cdot)$ corresponds to the lower incomplete Gamma function.

II. System Model and Protocol Description

Consider a scenario where a single BS equipped with an 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 [7]. 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 BS, 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 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. As a result, a very low latency can still be achieved, while guaranteeing an enhanced performance. 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 of dedicated electronic components that are connected to the antenna elements, known as radio frequency (RF) chains. More specifically, by using simple switches [8], 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 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. With the proposed design, the BS transmits each symbol in L instants of time, where each replica experiences different fast fading channels. Besides, it is assumed a perfect downlink channel estimation. Under these considerations, the L channel matrices, belonging to the kth user in the gth group of the sth cluster, can be organized in the following block diagonal arrangement

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

where $\mathbf{H}^l_{sgk} \in \mathbb{C}^{\frac{M}{L}\times N}$ denotes the channel matrix corresponding to the lth transmit sub-array. By applying the Karhunen-Loeve transformation [9], the channel matrices in (1) can be decomposed as

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

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 $\mathbf{G}^l_{sgk} \in \mathbb{C}^{r_s^*\times N}$ is the fast varying channel matrix corresponding to the lth sub-array, whose entries are 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 L successive replicas over each sub-array, the users observe the following signal

$$\mathbf{y}_{sgk} = \bar{\mathbf{H}}^H_{sgk} \sum_{n=1}^{S} \bar{\mathbf{B}}_n \sum_{i=1}^{G} \bar{\mathbf{v}}_{ni} \sum_{j=1}^{K} \alpha_{nij} x_{nij} \in \mathbb{C}^{LN\times 1}, \tag{3}$$

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 group of the nth cluster.

B. Beamforming Design

Given the proposed antenna structure, the beamformer for the sth cluster can be arranged as

$$\bar{\mathbf{B}}_s = \begin{bmatrix} \mathbf{B}_s & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_s \end{bmatrix}, \tag{4}$$

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 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 $(\mathbf{H}^l_{sgk})^H\mathbf{B}_{s'} \approx 0$, 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 $\mathbf{R}_s$ [9]. In order to achieve the desired objective, we explore next 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. Then, we concatenate the left eigenvectors of interfering clusters to form the following matrix

$$\mathbf{U}^-_s = [\mathbf{U}_1, \cdots, \mathbf{U}_{s-1}, \mathbf{U}_{s+1}, \cdots, \mathbf{U}_S] \in \mathbb{C}^{\frac{M}{L}\times(S-1)r^*}. \tag{5}$$

Next, the left eigenvectors of $\mathbf{U}^-_s$ are arranged as $\mathbf{E}_s = [\mathbf{E}^1_s, \mathbf{E}^0_s]$. The matrix $\mathbf{E}^0_s \in \mathbb{C}^{\frac{M}{L}\times\frac{M}{L}-(S-1)r^*}$ 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 $\tilde{\mathbf{H}}^l_{sgk} = (\mathbf{E}^0_s)^H\mathbf{U}_s\mathbf{\Lambda}^{\frac{1}{2}}_s\mathbf{G}^l_{sgk}$, which is orthogonal to the eigen-space spanned by interfering clusters and has covariance matrix given by $\tilde{\mathbf{R}}_s = (\mathbf{E}^0_s)^H\mathbf{R}_s\mathbf{E}^0_s$.

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

$$\mathbf{B}_s = \mathbf{E}^0_s\mathbf{F}^1_s \in \mathbb{C}^{\frac{M}{L}\times V}, \tag{6}$$

where $S \leq V \leq \left(\frac{M}{L}-(S-1)r^*\right)$ and $V \leq r^* \leq r$ must be satisfied. Now, regarding to the inner precoding vector, due to the multi-array arrangement, $\bar{\mathbf{v}}_{sg}$ is composed by L sub-vectors, that is, $\bar{\mathbf{v}}_{sg} = [(\mathbf{v}^1_{sg})^T, \cdots, (\mathbf{v}^L_{sg})^T]^T$, where $\mathbf{v}^l_{sg}$ is the precoding vector corresponding to the lth sub-array responsible for assigning the data for the gth group in the sth cluster. Since each sub-array contains the same number of antenna elements, we have that $\mathbf{v}^1_{sg} = \mathbf{v}^2_{sg} = \cdots = \mathbf{v}^L_{sg}$. Therefore, we define the sub-precoders as

$$\mathbf{v}^l_{sg} = [\ \mathbf{0}_{1\times(g-1)}\ ,\ 1\ ,\ \mathbf{0}_{1\times(V-g)}\ ]^T, \quad \forall l = 1, \cdots, L. \tag{7}$$

With this choice, note that the gth effective data stream transmitted by each sub-array is associated with the gth sub-group, which enables each group to receive L copies of the same superposed symbol.

C. Signal Reception

Hereafter, for the sake of brevity, the cluster subscript is omitted, e.g. $\mathbf{y}_{sgk}$ is expressed as $\mathbf{y}_{gk}$. Then, considering perfect cancellation of inter-cluster interferences, and after all L transmissions have been received, the signal at the kth user in the gth group can be structured as

$$\mathbf{y}_{gk} = \bar{\mathbf{H}}^H_{gk}\bar{\mathbf{B}}\sum_{i=1}^{G}\bar{\mathbf{v}}_i\sum_{j=1}^{K}\alpha_{ij}x_{ij} + \left[\mathbf{n}^1_{gk}, \cdots, \mathbf{n}^L_{gk}\right]^T, \tag{8}$$

where $\mathbf{n}^l_{gk} \in \mathbb{C}^{N\times 1}$ 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^2_n.

Then, the users separate the superposed data symbols intended for each sub-group by adopting a zero-forcing detector, that is, the signal obtained in (8) is filtered through the following detection matrix

$$\bar{\mathbf{H}}^\dagger_{gk} = \begin{bmatrix} \mathbf{H}^{1\dagger}_{gk} & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{H}^{L\dagger}_{gk} \end{bmatrix}, \tag{9}$$

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, which can be represented by

$$\begin{aligned}\hat{\mathbf{x}}_{gk} &= \left[\mathbf{x}^1, \cdots, \mathbf{x}^L\right]^T + \bar{\mathbf{H}}^\dagger_{gk}\left[\mathbf{n}^1_{gk}, \cdots, \mathbf{n}^L_{gk}\right]^T \\ &= \left[\begin{bmatrix} \sum_{j=1}^{K}\alpha_{1j}x_{1j} \\ \vdots \\ \sum_{j=1}^{K}\alpha_{Gj}x_{Gj} \end{bmatrix}, \cdots, \begin{bmatrix} \sum_{j=1}^{K}\alpha_{1j}x_{1j} \\ \vdots \\ \sum_{j=1}^{K}\alpha_{Gj}x_{Gj} \end{bmatrix}\right]^T + \bar{\mathbf{H}}^\dagger_{gk}\begin{bmatrix} \mathbf{n}^1_{gk} \\ \vdots \\ \mathbf{n}^L_{gk} \end{bmatrix},\end{aligned} \tag{10}$$

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 (10), 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 in terms of the outage probability, in which an exact closed-form expression is derived, followed by an asymptotic analysis.

A. Preliminaries

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. Thus, after SIC process, the kth user in the gth group observes the following data symbol

$$\hat{x}_{gk} = \underbrace{\alpha_{gk}x_{gk}}_{\text{symbol of interest}} + \underbrace{\sum_{j=k+1}^{K}\alpha_{gj}x_{gj}}_{\text{interference}} + \underbrace{[\mathbf{H}^{m\dagger}_{gk}\mathbf{n}^m_{gk}]_g}_{\text{noise}}. \tag{11}$$

where $m \in \{1, 2, \cdots, L\}$ corresponds to the sub-array that achieves the maximum effective channel gain among all L transmissions. From (11), 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

$$\text{SINR}_{gk}^{i} = \frac{\rho \gamma_{gk} \alpha_{gi}^2}{\rho \gamma_{gk} \mathcal{P}_i + 1}, \qquad \text{for} \quad 1 \leq i \leq k \leq K, \tag{12}$$

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

$$\mathcal{P}_i = \begin{cases} \sum_{j=i+1}^{U} \alpha_{gj}^2, & \text{for} \quad 1 \leq i \leq k < K, \\ 0, & \text{for} \quad i = k = K, \end{cases} \tag{13}$$

Proof: Please, see Appendix A.

B. Outage Probablity

In NOMA, before the kth user at the gth sub-group recovers 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 group can be formulated as

$$P_{gk} = P[\log_2(1 + \text{SINR}_{gk}^{i}) < R_{gi}], \quad \forall i = 1, \cdots k. \tag{14}$$

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

$$P_{gk} = \sum_{j=0}^{K-k} \frac{\mathcal{K}_{kj}}{(k+j)\Gamma(N-V+1)^{L(k+j)}} \times \gamma\left(N - V + 1, \mathcal{M}_{gk}[(\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1}]_{gg}\right)^{L(k+j)}, \tag{15}$$

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 V.

C. Asymptotic Analysis

To investigate further the behavior of the proposed system, an asymptotic outage analysis is now carried out.

Proposition II: When the transmit SNR approaches infinity, (15) can be approximated by

$$P_{gk} \approx \frac{K}{k}\binom{K-1}{k-1} \frac{[\rho \mathcal{M}_{gk}[(\mathbf{B}^H \mathbf{R} \mathbf{B})^{-1}]_{gg}]^{(N-V+1)Lk}}{\rho^{(N-V+1)Lk}[(N-V+1)!]^{Lk}}. \tag{16}$$

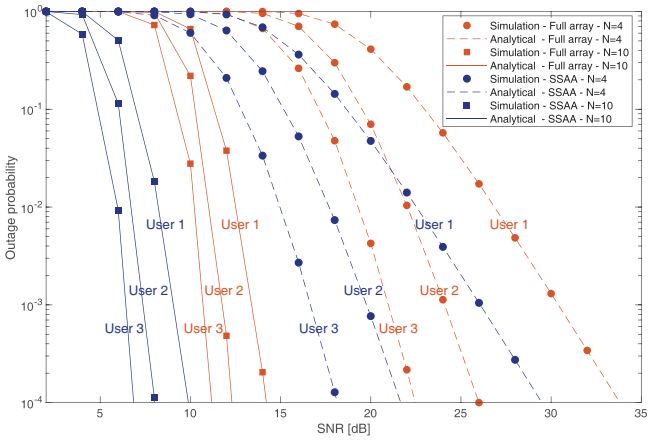

Fig. 2. Outage probability versus transmit SNR for massive MIMO-NOMA system with SSAA technique and conventional full array system operating in time diversity mode ($L = T = 3$).

As a result, it can be concluded that the kth user experiences the following diversity order

$$D_k = (N - V + 1)\, Lk. \tag{17}$$

Proof: Please, see Appendix B.

IV. Numerical Results and Discussions

In this section, some illustrative numerical examples of the proposed SSAA technique are presented and compared with conventional massive MIMO-NOMA and OMA systems operating in time diversity mode using the full antenna array. 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 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. For users within each sub-group, the power allocation coefficients are set to $\alpha_1 = 0.625, \alpha_2 = 0.250$ and $\alpha_3 = 0.125$, and fixed target rates, $R_1 = 1.4, R_2 = 1.5$ and $R_3 = 4$ bits per channel use (BPCU). 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 scheme. Fig. 3 brings the validation of the high-SNR analysis. One can observe that, for 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 (17).

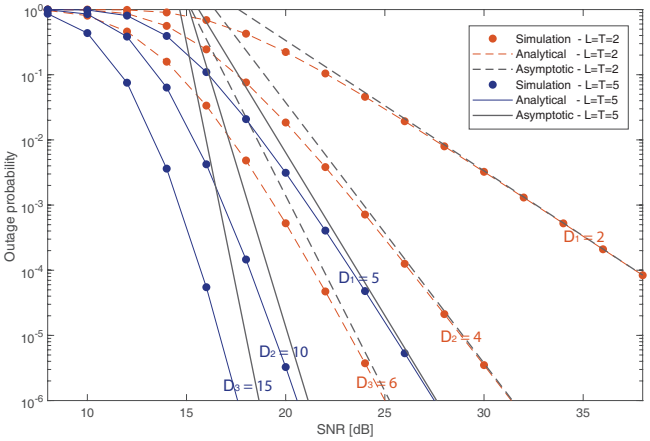

Fig. 3. Exact and asymptotic outage probability curves for massive MIMO-NOMA system operating with the proposed SSAA technique ($N = 4$).

Fig. 4 plots the outage sum-rate, $\sum_{k=1}^{K}(1 - P_{gk})R_{gk}$, versus transmit SNR. 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, 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.

Fig. 5 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.

V. Proof of Proposition I

Appendix A
Proof of Lemma I

From (11), one can see that the kth user decodes the ith weaker message with the following SINR

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

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

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

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

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

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

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

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

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

which completes the proof.

By replacing (12) in (14) and performing some algebraic manipulations, we get the following

$$\begin{aligned} P_{gk} &= \Pr\left[\log_2\left(1 + \frac{\rho\gamma_{gk}\alpha_{gi}^2}{\rho\gamma_{gk}\mathcal{P}_i + 1}\right) < R_{gi}\right] \\ &= \Pr\left[\gamma_{gk} < \frac{2^{R_{gi}} - 1}{\rho[\alpha_{gi}^2 - \mathcal{P}_i(2^{R_{gi}} - 1)]}\right] \\ &= P\left[\gamma_{gk} < \mathcal{M}_{gk}\right] \end{aligned} \tag{A-6}$$

where

$$\mathcal{M}_{gk} = \max_{1 \le i \le k}\left\{\frac{2^{R_{gi}} - 1}{\rho[\alpha_{gi}^2 - \mathcal{P}_i(2^{R_{gi}} - 1)]}\right\}. \tag{A-7}$$

The expression (A-6) 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

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

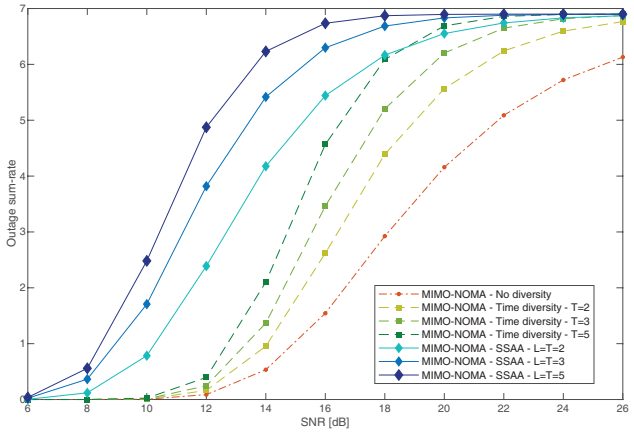

Fig. 4. Outage sum-rate for the proposed SSAA technique and the conventional full array time diversity approach in massive MIMO-NOMA systems ($N = 4$).

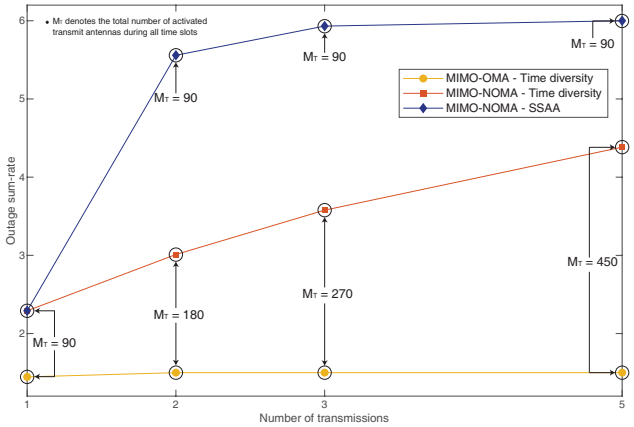

Fig. 5. Outage sum-rate versus number of redundant transmissions for a fixed transmit SNR of 16dB ($N = 4$).

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

$$\hat{\mathbf{R}} = (\mathbf{B}^H\mathbf{H}_{gk}^l(\mathbf{H}_{gk}^l)^H\mathbf{B})^{-1}\mathbf{B}^H\mathbf{H}_{gk}^l(\mathbf{H}_{gk}^l)^H\mathbf{B}$$
$$\times (\mathbf{B}^H\mathbf{H}_{gk}^l(\mathbf{H}_{gk}^l)^H\mathbf{B})^{-1} = (\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1} \quad \text{(A-8)}$$

As stated in [2], the matrix in (A-8) follows the inverse Wishart distribution and, consequently, the inverse of its main diagonal elements follows the Gamma distribution [10]. 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, considering first unordered gains, the probability density function (PDF) of $\max\left\{\varsigma_{gk}^1, \cdots, \varsigma_{gk}^L\right\}$ can be derived as

$$f_{\max}(x) = \frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^L} x^{N-V} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}$$
$$\times \gamma(N-V+1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L-1} \quad \text{(A-9)}$$

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

$$f_{\gamma_{gk}}(x) = \sum_{j=0}^{K-k} K\binom{K-1}{k-1}\binom{K-k}{j}(-1)^j f_{\max}(x) F_{\max}(x)^{k-1+j}$$
$$= \sum_{j=0}^{K-k} \mathcal{K}_{kj} \frac{L[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}^{N-V+1}}{\Gamma(N-V+1)^{L(k+j)}} x^{N-V} e^{-x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}}$$
$$\times \gamma(N-V+1, x[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg})^{L(k+j)-1}, \quad \text{(A-10)}$$

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 (A-10). This completes the proof.

Appendix B
Proof of Proposition II

By invoking the series representation of the incomplete Gamma function [11], the outage expression in (15) can be simplified as in (C-1), shown on the top of this page. Then, by exploring properties of the Taylor's series and performing some algebraic manipulations in (C-1), we obtain the desired high-SNR approximation as

$$P_{gk} \approx \frac{K}{k}\binom{K-1}{k-1} \frac{[\mathcal{M}_{gk}[(\mathbf{B}^H\mathbf{R}\mathbf{B})^{-1}]_{gg}]^{(N-V+1)Lk}}{\rho^{(N-V+1)Lk}[(N-V+1)!]^{Lk}}, \quad \text{(C-2)}$$

which completes the proof.

Acknowledgements

This project is partly supported by Academy of Finland via: (a) ee-IoT project under Grant n.319009, (b) FIREMAN consortium under Grant CHIST-ERA/n.326270, and (c) EnergyNet research fellowship under Grants n.321265 and n.328869.

References

[1] 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.

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

[3] A. S. de Sena, D. B. da Costa, Z. Ding, and P. H. J. Nardelli, "Massive MIMO-NOMA networks with multi-polarized antennas," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2019.

[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Wiley Series in Probability and Statistics, Aug. 2003.
[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.