Potentials and Limits of Using Preconfigured Spatial Beams as Bandwidth Resources: Beam Selection vs Beam Aggregation

Zhiguo Ding, *Fellow, IEEE*

Abstract—This letter studies how to use spatial beams pre-configured in a legacy spatial division multiple access (SDMA) network as bandwidth resources via the implementation of non-orthogonal multiple access (NOMA). Two different beam management schemes, namely beam selection and beam aggregation, are developed to improve the overall system throughput without consuming extra spectrum or changing the performance of the legacy network. Analytical and simulation results are presented to show that the two schemes realize different tradeoffs between system performance and complexity.

Index Terms—NOMA, SDMA, beam management, power allocation, outage probability.

I. INTRODUCTION

Recently, using non-orthogonal multiple access (NOMA) as a type of add-ons in communication networks has received a lot of attention, because it makes the implementation of NOMA more flexible in practice and also improves the performance of various communication networks in a spectrally efficient manner [1], [2]. For example, consider a legacy spatial division multiple access (SDMA) network, where spatial beams have already been configured to serve legacy primary users [3]. Conventional multiple-input multiple-output NOMA (MIMO-NOMA) approaches need to redesign the legacy network in order to serve new secondary users [4]–[6]. An alternative is to directly use those preconfigured beams for serving the new users, which makes the implementation of NOMA and the admission of the new users transparent to the legacy network. As shown in [7], these preconfigured spatial beams indeed have the potential to be used as bandwidth resources, similar to orthogonal frequency-division multiplexing (OFDM) subcarriers, but the benefits of using these beams can be severely limited by inter-beam interference.

The aim of this letter is to provide more detailed analysis for the potentials and limits of using these preconfigured spatial beams, by focusing on a simple case with multiple preconfigured beams and a single secondary user. For the scenario where the secondary user's signals on different beams are independently encoded, [7] shows that the use of beam selection, i.e., selecting a single beam to serve the secondary user, is optimal, and the outage probability achieved by beam selection is analyzed in this letter. Despite its simplicity, beam selection suffers a drawback that its performance is degraded with more beams available, because of inter-beam interference. To mitigate this inter-beam interference, a new scheme, termed beam aggregation, is proposed in the letter, where the secondary user is encouraged to use multiple beams and its signals on these beams are jointly designed. Simulation results are presented to show that beam aggregation can effectively suppress inter-beam interference and always outperform beam selection, but at a price of more system complexity.

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

II. SYSTEM MODEL

Consider a legacy SDMA network with one N-antenna base station and M single-antenna primary users, denoted by U_m^P, whose channel vectors are denoted by $\mathbf{g}_m$. Assume that M beamforming vectors, denoted by $\mathbf{f}_m$, have been configured for the primary users. Furthermore, assume that zero-forcing beamforming is used, i.e., $\begin{bmatrix} \mathbf{f}_1 & \cdots & \mathbf{f}_M \end{bmatrix} = \mathbf{G}(\mathbf{G}^H\mathbf{G})^{-1}\mathbf{D}$, where $\mathbf{G} = \begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_M \end{bmatrix}$, $\mathbf{D}_{i,i} = \left(M[(\mathbf{G}^H\mathbf{G})^{-1}]_{i,i}\right)^{-\frac{1}{2}}$ and $\mathbf{A}_{i,j}$ denotes an element of $\mathbf{A}$ on its i-th row and j-th column [3]. As a result, $\mathbf{g}_m^H\mathbf{f}_i = 0$, for $m \neq i$.

As shown in [7], these preconfigured beams, $\mathbf{f}_m$, can be exploited as a type of bandwidth resources and used to serve additional secondary users, similarly to OFDM subcarriers. In this paper, a simple case with a single secondary user, denoted by U^S, is focused. On each of the M beams, the base station superimposes U_m^P's signal, denoted by s_m^P, with the secondary user's signal, denoted by s_m^S, where the transmit powers of s_m^P and s_m^S are denoted by $\alpha_m^P\rho$ and $\alpha_m^S\rho$, respectively, ρ denotes the transmit power budget, α_m^P and α_m^S denote the power allocation coefficients, and $\alpha_m^P + \alpha_m^S \leq 1$. Therefore, the system model at U^S is given by

$$y^S = \mathbf{h}^H \sum_{i=1}^M \mathbf{f}_i \left(\sqrt{\alpha_i^P\rho}s_i^P + \sqrt{\alpha_i^S\rho}s_i^S \right) + n^S, \quad (1)$$

where $\mathbf{h}$ denotes the secondary user's channel vector and n^S denotes the Gaussian noise with its power normalized.

Assuming that s_m^S, $1 \leq m \leq M$, are independently encoded, on beam $\mathbf{f}_m$, the secondary user U^S can decode s_m^P with the following data rate:

$$\tilde{R}_m = \log\left(1 + \frac{h_m\alpha_m^P}{h_m\alpha_m^S + \sum_{i\neq m} h_i\left(\alpha_i^P + \alpha_i^S\right) + \frac{1}{\rho}} \right), \quad (2)$$

where $h_m = |\mathbf{h}^H\mathbf{f}_m|^2$. Assume that all the primary users have the same target data rate, denoted by R^P. If $\tilde{R}_m \geq R^P$, U^S can carry out successive interference cancellation (SIC) successfully and decode s_m^S with the following data rate:

$$R_m = \log\left(1 + \frac{h_m\alpha_m^S}{\sum_{i\neq m} h_i\left(\alpha_i^P + \alpha_i^S\right) + \frac{1}{\rho}} \right). \quad (3)$$

As shown in [7], with s_m^S, $1 \leq m \leq M$, independently encoded, the optimal resource allocation strategy is to select a single beam to serve the secondary user, and the performance of beam selection is analyzed in the following section.

III. BEAM SELECTION

The performance analysis of beam selection requires the explicit expressions for the power allocation coefficients, α_m^P and α_m^S. Depending on whether U^S is active on the beams,

they can be expressed differently. For the case that U^S is not active on beam $\mathbf{f}_m$, denote the corresponding power allocation coefficients by $\alpha_{m,I}^P$ and $\alpha_{m,I}^S$. Otherwise, they are denoted by $\alpha_{m,II}^P$ and $\alpha_{m,II}^S$.

If U^S is not active on beam $\mathbf{f}_m$, $\alpha_{m,I}^S = 0$, and hence the use of $\alpha_{m,I}^P = \min\left\{1, \frac{\epsilon_P}{\rho g_m}\right\}$ is sufficient to ensure that U_m^P decodes s_m^P, i.e., $\log(1 + \rho\alpha_{m,I}^P g_m) \geq R^P$, where $g_m = |\mathbf{g}_m^H \mathbf{f}_m|^2$ and $\epsilon_P = 2^{R^P} - 1$.

If U^S is active on beam $\mathbf{f}_m$, U_m^P's data rate on beam $\mathbf{f}_m$ is given by $R_m^P = \log\left(1 + \frac{g_m \alpha_{m,II}^P}{g_m \alpha_{m,II}^S + \frac{1}{\rho}}\right)$. To ensure that $R_m^P \geq R^P$, α_m^S should be chosen as follows:

$$\alpha_{m,II}^S \leq \max\left\{0, \frac{g_m - \frac{\epsilon_P}{\rho}}{(\epsilon_P + 1)g_m}\right\}, \tag{4}$$

and $\alpha_{m,II}^P = 1 - \alpha_{m,II}^S$. To ensure that U^S decodes s_m^P, $\tilde{R}_m \geq \bar{R}_m^P$ is required, which leads to the following constraint:

$$\alpha_{m,II}^S \leq \max\left\{0, \frac{h_m - \epsilon_P \sum_{i \neq m} h_i \alpha_{i,I}^P - \frac{\epsilon_P}{\rho}}{(1 + \epsilon_P)h_m}\right\}, \tag{5}$$

where $\alpha_{i,I}^P$ is used instead of $\alpha_{i,II}^P$ because the secondary user is not active on beam $\mathbf{f}_i$, $i \neq m$. Therefore, $\alpha_{m,II}^S$ can be chosen as follows:

$$\alpha_{m,II}^S = \min\left\{\max\left\{0, \frac{g_m - \frac{\epsilon_P}{\rho}}{(\epsilon_P + 1)g_m}\right\}\right. \tag{6}$$
$$\left.\max\left\{0, \frac{h_m - \epsilon_P \sum_{i \neq m} h_i \alpha_{i,I}^P - \frac{\epsilon_P}{\rho}}{(1 + \epsilon_P)h_m}\right\}\right\}.$$

Assume that beam $\mathbf{f}_{m^*}$ is selected, which means that the outage probability of interest can be expressed as follows: $\mathrm{P}^o \triangleq 1 - \mathrm{P}\left(\tilde{R}_{m^*} \geq R^P, R_{m^*} \geq R^S,\right)$, where R^S denotes the secondary user's target data rate. Analyzing P^o is challenging since h_m, g_m and h_i, $m \neq i$, are correlated. The following lemma shows an interesting property of the NOMA transmission protocol.

Lemma 1. *Assume that the users' channels are independent and identically complex Gaussian distributed with zero mean and unit variance. $\mathrm{P}^o \rightarrow 0$ for $\rho \rightarrow \infty$.*

Proof. See Appendix A. □

Remark 1: Lemma 1 shows that there is no error floor for the considered outage probability, despite strong co-channel interference. Or in other words, by using NOMA, an additional user can be supported with an arbitrarily high data rate, without consuming extra spectrum but simply increasing the transmit power. This property is valuable given the scarcity of the spectrum available to communications.

Remark 2: Our conducted simulation results show that increasing M degrades the outage probability, which can be explained in the following. Each beam can be viewed as an OFDM subcarrier. By increasing M is similar to increase the number of subcarriers. If the overall bandwidth is kept the same, increasing the number of subcarriers reduces the bandwidth of each subcarrier, and hence degrades the user's performance if a user can use one subcarrier only. Similarly, for the considered NOMA system, increasing M increases inter-beam interference, which causes the performance degradation. In the next section, two schemes based beam aggregation is introduced to mitigate this inter-beam interference.

IV. Beam Aggregation

Unlike beam selection, beam aggregation encourages the secondary user to use multiple beams. Denote $\mathcal{D}$ by the set including the beams used by the secondary user, which means that the system model in (1) can be rewritten as follows:

$$y^S = \mathbf{h}^H \sum_{i \in \mathcal{D}} \mathbf{f}_i \left(\sqrt{\alpha_i^P \rho} s_i^P + \sqrt{\alpha_i^S \rho}\beta_i s^S\right) \tag{7}$$
$$+ \mathbf{h}^H \sum_{j \in \mathcal{D}^c} \mathbf{f}_j \sqrt{\alpha_j^P \rho} s_j^P + n^S,$$

where s^S denotes the secondary user's signal, $\beta_i = \frac{(\mathbf{h}^H \mathbf{f}_i)^*}{\sqrt{|\mathbf{h}^H \mathbf{f}_i|^2}}$ is introduced in order to ensure coherent combining and also avoid changing the overall transmit power, and $\mathcal{D}^c$ denotes the complementary set of $\mathcal{D}$.

A. Scheme I

A low-complexity and straightforward scheme is to ask the secondary user to directly decode its own message by treating the primary users' signals as noise. By using (7), it is straightforward to show that the following data rate is achievable to the secondary user:

$$R_S = \log\left(1 + \frac{\left(\sum_{i \in \mathcal{D}} \sqrt{h_i \alpha_i^S}\right)^2}{\sum_{j=1}^M h_j \alpha_j^P + \frac{1}{\rho}}\right). \tag{8}$$

To decide the power allocation coefficients, α_m^P and α_m^P, recall that the system model at the primary users can be expressed as follows:

$$y_m^P = \mathbf{g}_m^H \mathbf{f}_m \left(\sqrt{\alpha_m^P \rho} s_m^P + \sqrt{\alpha_m^S \rho}\beta_m s^S\right) + n_m^P, \quad m \in \mathcal{D}$$
$$y_m^P = \mathbf{g}_m^H \mathbf{f}_m \sqrt{\alpha_m^P \rho} s_m^P + n_m^P, \quad m \in \mathcal{D}^c, \tag{9}$$

which means the following data rates are achievable

$$R_m^P = \log\left(1 + \frac{g_m \alpha_m^P}{g_m \alpha_m^S \beta_m^2 + \frac{1}{\rho}}\right), \quad m \in \mathcal{D} \tag{10}$$
$$R_m^P = \log\left(1 + g_m \alpha_m^P \rho\right), \quad m \in \mathcal{D}^c,$$

where n_m^P denotes the noise at U_m^P.

In order to ensure that $R_m^P \geq R^P$, the following choices for the power allocation coefficients can be used:

$$\alpha_m^S = \max\left\{0, \frac{g_m - \frac{\epsilon_P}{\rho}}{(\epsilon_P + 1)g_m}\right\}, \alpha_m^P = \min\left\{1, \frac{\epsilon_P\left(g_m + \frac{1}{\rho}\right)}{g_m(1 + \epsilon_P)}\right\},$$

for $m \in \mathcal{D}$. Otherwise, $\alpha_m^P = \min\left\{1, \frac{\epsilon_P}{\rho g_m}\right\}$ and $\alpha_m^S = 0$.

As shown in the next section, the performance of Scheme I can ensure that the secondary user's data rate grows by increasing M; however, it suffers some performance loss compared to beam selection, particularly at high SNR.

B. Scheme II

To facilitate the description of the scheme, assume that the secondary user's channel gains are ordered as follows: $h_1 \geq \cdots \geq h_M$, which means that a reasonable choice for $\mathcal{D}$ is given by $\mathcal{D} = \{1, \cdots, |\mathcal{D}|\}$, where $|\mathcal{D}|$ denotes the size of $\mathcal{D}$. Scheme II is to carry out SIC on the beams included in $\mathcal{D}$ successively, i.e., s_i^P is decoded and removed before decoding s_j^P, for $i < j$ and $i, j \in \mathcal{D}$. After s_m^P, $m \in \mathcal{D}$, are decoded, the secondary user can decode its own signal, s^S.

Therefore, on beam $\mathbf{f}_m$, $m \in \mathcal{D}$, the data rate for the secondary user to decode s_m^P is given by

$$\tilde{R}_m = \log\left(1 + \frac{h_m \alpha_m^P}{\sum_{j=m+1}^{M} h_j \alpha_j^P + |\sum_{i \in \mathcal{D}} \sqrt{h_i \alpha_i^S}|^2 + \frac{1}{\rho}}\right).$$

Conditioned on $\tilde{R}_m \geq R^P$, $\forall m \in \mathcal{D}$, the secondary user can carry out SIC successfully and realize the following data rate for its own signal:

$$R^S = \log\left(1 + \frac{|\sum_{i \in \mathcal{D}} \sqrt{h_i \alpha_i^S}|^2}{\sum_{j \in \mathcal{D}^c} h_j \alpha_j^P + \frac{1}{\rho}}\right). \tag{11}$$

Compared to Scheme I, for Scheme II, the choices for the power allocation coefficients, α_m^P and α_m^S, $m \in \mathcal{D}$, are more difficult to find, because they need to ensure that s_m^P, $m \in \mathcal{D}$, can be decoded by both U_m^P and U^S. Compared to beam selection which uses a single beam only, the design of Scheme II is also more challenging since multiple beams are used and hence a choice of α_m^S affects that of α_i^S, $m \neq i$. Note that designing the power allocation coefficients to maximize the the secondary user's data rate can be formulated as the following optimization problem:

$$\max_{\alpha_m^P, \alpha_m^S} \quad \log\left(1 + \frac{|\sum_{i \in \mathcal{D}} \sqrt{h_i \alpha_i^S}|^2}{\sum_{j \in \mathcal{D}^c} h_j \alpha_j^P + \frac{1}{\rho}}\right) \tag{P1a}$$

$$\text{s.t.} \quad \tilde{R}_m \geq R_m^P, \forall m \in \mathcal{D} \tag{P1b}$$

$$\alpha_m^P \geq \frac{\epsilon_P\left(g_m + \frac{1}{\rho}\right)}{g_m(1 + \epsilon_P)}, \forall m \in \mathcal{D} \tag{P1c}$$

$$\alpha_m^S + \alpha_m^P \leq 1, \forall m \in \mathcal{D} \tag{P1d}$$

$$\alpha_m^S \geq 0, \alpha_m^P \geq 0, \forall m \in \mathcal{D}. \tag{P1e}$$

Note that constraint (P1b) ensures that U^S can decode s_m^P, and constraint (P1c) ensures that serving U^S on beam $\mathbf{f}_m$ does not degrade U_m^P's performance, $m \in \mathcal{D}$.

It is straightforward to show that problem P1 is not concave, but it can be recast to an equivalent concave form as follows. First, note that the following power allocation coefficients are fixed: $\alpha_m^P = \min\left\{1, \frac{\epsilon_P}{\rho g_m}\right\}$ and $\alpha_m^S = 0$, $m \in \mathcal{D}^c$, as shown in the previous section. By using the fact that α_m^P and α_m^S,

$m \in \mathcal{D}^c$, are constants, problem P1 can be equivalently recast as follows:

$$\max_{\alpha_m^P, \alpha_m^S} \quad \sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i \alpha_i^S} \tag{P2a}$$

$$\text{s.t.} \quad \frac{h_m \alpha_m^P}{\tau_D + \sum_{j=m+1}^{|\mathcal{D}|} h_j \alpha_j^P + \left(\sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i \alpha_i^S}\right)^2} \geq \epsilon_P,$$
$$1 \leq m \leq |\mathcal{D}| \tag{P2b}$$

$$\alpha_m^P \geq \eta_m, \forall m \in \mathcal{D} \tag{P2c}$$

$$(\text{P1d}), (\text{P1e})$$

where $\eta_m = \frac{\epsilon_P\left(g_m + \frac{1}{\rho}\right)}{g_m(1 + \epsilon_P)}$ and $\tau_D = \sum_{j=|\mathcal{D}|+1}^{M} h_j \alpha_j^P + \frac{1}{\rho}$. Problem P2 can be further equivalently recast as follows:

$$\max_{\alpha_m^P, \alpha_m^S} \quad \sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i \alpha_i^S} \tag{P3a}$$

$$\text{s.t.} \quad \epsilon_P \sum_{j=m+1}^{|\mathcal{D}|} h_j \alpha_j^P + \epsilon_P \left(\sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i \alpha_i^S}\right)^2$$
$$- h_m \alpha_m^P + \epsilon_P \tau_D \leq 0, 1 \leq m \leq |\mathcal{D}| \tag{P3b}$$

$$(\text{P2b}), (\text{P1d}), (\text{P1e}).$$

Problem P3 is not a concave problem mainly due to the fact that $\left(\sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i \alpha_i^S}\right)^2$ is a concave function of α_i^S. Fortunately, problem P3 can be still recast to a concave problem as shown in the following. Define $x_m = \sqrt{\alpha_m^S}$, and problem P3 can be reformulated as follows:

$$\max_{\alpha_m^P, x_m} \quad \sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i} x_i \tag{P4a}$$

$$\text{s.t.} \quad \epsilon_P \sum_{j=m+1}^{|\mathcal{D}|} h_j \alpha_j^P + \epsilon_P \left(\sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i} x_i\right)^2$$
$$- h_m \alpha_m^P + \epsilon_P \tau_D \leq 0, 1 \leq m \leq |\mathcal{D}| \tag{P4b}$$

$$\alpha_m^P \geq \eta_m, 1 \leq m \leq |\mathcal{D}| \tag{P4c}$$

$$x_m^2 + \alpha_m^P \leq 1, 1 \leq m \leq |\mathcal{D}| \tag{P4d}$$

$$x_m \geq 0, \alpha_m^P \geq 0, 1 \leq m \leq |\mathcal{D}|. \tag{P4e}$$

Note that $\left(\sum_{i=1}^{|\mathcal{D}|} \sqrt{h_i} x_i\right)^2$ is a convex function, which means that the constraint function in (P4b) is convex. Similarly, constraint (P4d) is also convex, and all the other functions in problem P4 are affine functions, which means that problem P4 is a concave problem and can be easily solved by using those optimization solvers [8].

Remark 3: The performance of beam aggregation is depending on the choice of $\mathcal{D}$. For the simulations conducted for this paper, an exhaustive search is used to find the optimal choice of $\mathcal{D}$, where designing a more computationally efficient way to optimize $\mathcal{D}$ is an important direction for future research.

V. SIMULATION RESULTS

In this section, the performance of the considered beam management schemes is investigated and compared by using computer simulation results.
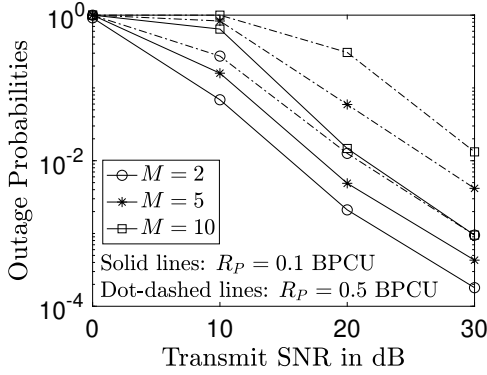
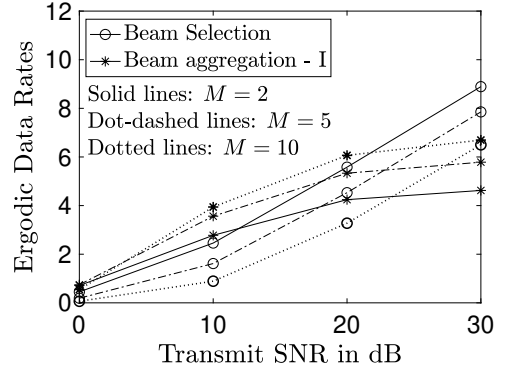(a) Outage Performance $R^S = 1 BPCU$
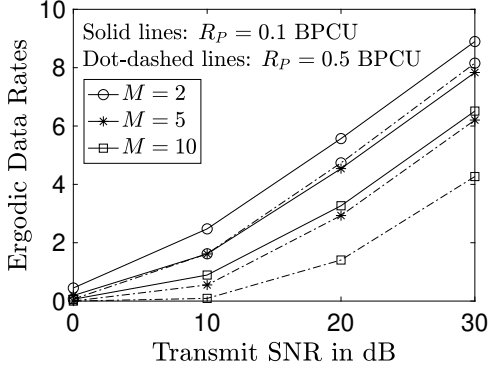
(b) Ergodic Data Rates

Fig. 1. Performance of beam selection based NOMA transmission. $N = M$ and BPCU denotes bits per channel use.

(a) Beam Selection vs Beam Aggregation Scheme I

(b) Beam Selection vs Beam Aggregation Scheme II

Fig. 2. Comparison between the beam selection and beam aggregation based schemes. $N = M$ and $R^P = 0.1$ BPCU.

In Fig. 1, the performance of beam selection based NOMA transmission is studied by using two metrics, namely the outage probability and ergodic data rates. In particular, Fig. 1(a) demonstrates that there is no outage probability error floor for the beam selection scheme, which confirms Lemma 1. In addition, Fig. 1(a) shows that the performance of beam selection is degraded by increasing M, as discussed in Remark 2. Fig. 1(b) shows that the use of NOMA transmission yields a considerably large ergodic data rate for the secondary user, and it is important to point out that such a large data rate is supported without consuming extra spectrum or degrading the performance of the legacy network.
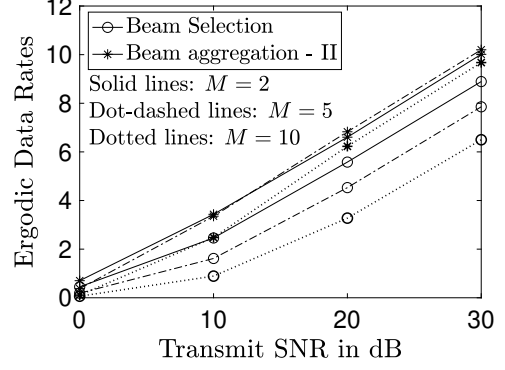
In Fig. 2, the performance of beam aggregation based NOMA transmission is studied by using beam selection as the benchmarking scheme. Fig. 2(a) shows that the use of the first type of beam aggregation can result in a moderate performance gain over beam selection, particularly with a large M and at low SNR. Fig. 2(b) demonstrates that the second type of beam aggregation can always outperform beam selection and the beam aggregation scheme I. More importantly, the use of the second type of beam aggregation can ensure robust performance regardless of the choices of the beam number. But it is important to point out that the use of the second type of beam aggregation results in more system complexity.

VI. CONCLUSIONS

This letter has demonstrated how those spatial beams preconfigured in a legacy SDMA network can be used as bandwidth resources via the implementation of NOMA. Two different NOMA schemes have been developed to ensure an additional secondary user served without consuming extra spectrum or changing the performance of the legacy network. Analytical and simulation results have been presented to show that the two schemes realize different tradeoffs between system performance and complexity.

APPENDIX A
PROOF FOR LEMMA 1

To facilitate the performance analysis, the outage probability is rewritten as follows:

$$\mathrm{P}^o = \mathrm{P}\left(\alpha_{m,II}^S = 0, \forall m\right) \tag{12}$$
$$+ \sum_{\mathcal{S}_i \subseteq \mathcal{S}} \mathrm{P}\left(\alpha_{m,II}^S > 0, \forall m \in \mathcal{S}_i, \alpha_{j,II}^S = 0, \forall j \in \mathcal{S}_i^c\right.$$
$$\max\left\{\gamma_m, m \in \mathcal{S}_i\right\} \leq \epsilon_s,\left.\right),$$

where $\gamma_m = \frac{h_m \alpha_{m,II}^S}{\sum_{i \neq m} h_i \alpha_{i,I}^P + \frac{1}{\rho}}$, $\mathcal{S} = \{1, \cdots, M\}$, $\mathcal{S}_i$ is a subset of $\mathcal{S}$, and $\mathcal{S}_i^c$ is a complementary set of $\mathcal{S}_i$.

P^o can be first upper bounded as follows:

$$\mathrm{P}^o \leq \mathrm{P}\left(\alpha_{1,II}^S = 0\right) + \sum_{\mathcal{S}_i \subset \mathcal{S}} \mathrm{P}\left(\alpha_{j,II}^S = 0, \forall j \in \mathcal{S}_i^c\right) \tag{13}$$
$$+ \mathrm{P}\left(\alpha_{m,II}^S > 0, \forall m \in \mathcal{S}, \max\left\{\gamma_m, m \in \mathcal{S}\right\} \leq \epsilon_s,\right).$$

Note that the probability $\mathrm{P}\left(\alpha_{1,II}^S = 0\right)$ can be further bounded as follows:

$$\mathrm{P}\left(\alpha_{1,II}^S = 0\right) \leq \mathrm{P}\left(g_1 \leq \frac{\epsilon_P}{\rho}\right) \tag{14}$$
$$+ \mathrm{P}\left(h_1 \leq \epsilon_P \sum_{i \neq 1} h_i \alpha_{i,I}^P + \frac{\epsilon_P}{\rho}\right).$$

Denote the two terms at the right hand side of (14) by Q_1 and Q_2, respectively.

Because the users' channels are independent and ideally complex Gaussian distributed, $g_1 = \left(\frac{[(\mathbf{G}^H \mathbf{G})^{-1}]_{i,i}}{M} \right)^{-1}$ follows the inverse-Wishart distribution, i.e., [9], [10]

$$f_{[(\mathbf{G}^H \mathbf{G})^{-1}]_{i,i}}(x) = W^{-1}(I_M, N) \tag{15}$$
$$= \frac{1}{\Gamma(N-M+1)} x^{-(N-M+2)} e^{-\frac{1}{x}},$$

where $\Gamma(x)$ denotes the gamma function and $\gamma(n,x)$ denotes the incomplete gamma function [11]. By using the distribution of g_1, Q_1 can be evaluated as follows:

$$Q_1 = \int_{\frac{\rho}{M\epsilon_P}}^{\infty} \frac{1}{\Gamma(N-M+1)} x^{-(N-M+2)} e^{-\frac{1}{x}} dx \tag{16}$$
$$= \frac{1}{\Gamma(N-M+1)} \gamma\left(N-M+1, \frac{M\epsilon_P}{\rho}\right).$$

At high SNR, Q_1 can be approximated as follows:

$$Q_1 = \frac{\sum_{n=0}^{\infty} \frac{(-1)^n}{n!(N-M+1+n)} \left(\frac{M\epsilon_P}{\rho}\right)^{N-M+1+n}}{\Gamma(N-M+1)} \tag{17}$$
$$\approx \frac{1}{(N-M+1)!} \left(\frac{M\epsilon_P}{\rho}\right)^{N-M+1} \to 0,$$

for $\rho \to \infty$.

Recall that $\alpha_{i,I}^P = \min\left\{1, \frac{\epsilon_P}{\rho g_i}\right\}$. By using the definition $\mathcal{B}_1 = \{2, \cdots, M\}$, Q_2 can be evaluated as follows:

$$Q_2 = \sum_{\mathcal{B} \subseteq \mathcal{B}_1} P\left(h_1 \leq \epsilon_P \sum_{i \neq 1} h_i \alpha_{i,I}^P + \frac{\epsilon_P}{\rho}, \right. \tag{18}$$
$$\left. \alpha_{j,I}^P = 1, j \in \mathcal{B}^c, \alpha_{k,I}^P = \frac{\epsilon_P}{\rho g_k}, k \in \mathcal{B}\right),$$

where $\mathcal{B}^c$ denotes the complementary set of $\mathcal{B}$. Therefore, the probability can be upper bounded as follows:

$$Q_2 \leq \sum_{\mathcal{B} \subset \mathcal{B}_1} P\left(\alpha_{j,I}^P = 1, j \in \mathcal{B}^c\right) \tag{19}$$
$$+ P\left(h_1 \leq \epsilon_P \sum_{i \neq 1} h_i \alpha_{i,I}^P + \frac{\epsilon_P}{\rho}, \alpha_{i,I}^P = \frac{\epsilon_P}{\rho g_i}, i \in \mathcal{B}_1\right).$$

Denote the last term in (19) by Q_3 which can be expressed as follows:

$$Q_3 = P\left(\xi_1 \leq \frac{\epsilon_P}{\rho}, g_i \geq \frac{\epsilon_P}{\rho}, i \in \mathcal{B}_1\right) \leq P\left(\xi_1 \leq \frac{\epsilon_P}{\rho}\right),$$

where $\xi_m = \frac{h_m}{\sum_{i \neq m} h_i \frac{\epsilon_P}{g_i} + 1}$. Denote the CDF of ξ_m by $F_{\xi_m}(x)$. Therefore, at high SNR, Q_3 can be approximated as follows:

$$Q_3 \leq F_{\xi_1}\left(\frac{\epsilon_P}{\rho}\right) \to F_{\xi_1}(0) = 0, \tag{20}$$

since $F_{\xi_m}(x)$ is not related to ρ. It is straightforward to show that $P\left(\alpha_{j,I}^P = 1, j \in \mathcal{B}^c\right) \to 0$ for $\rho \to \infty$, since $P\left(\alpha_{j,I}^P = 1, j \in \mathcal{B}^c\right) = P\left(\frac{\epsilon_P}{\rho g_j} \geq 1, j \in \mathcal{B}^c\right)$. Therefore, both the terms in (19) goes to zero at high SNR, which means

that $Q_2 \to 0$ for $\rho \to \infty$. Since $Q_i \to 0$, $i \in \{1, 2\}$, $P\left(\alpha_{1,II}^S = 0\right) \to 0$ for $\rho \to \infty$. Similarly, one can also prove that $P\left(\alpha_{j,II}^S = 0, \forall j \in \mathcal{S}_i^c\right)$ in (13) goes to zero for $\rho \to \infty$.

Denote the last term in (13) by Q_4. Similar to Q_2 in (19), Q_4 can be upper bounded as follows:

$$Q_4 \leq \sum_{\mathcal{B} \subset \mathcal{B}_1} P\left(\alpha_{j,I}^P = 1, j \in \mathcal{B}^c\right) \tag{21}$$
$$+ P\left(\max\left\{\xi_m \alpha_{m,II}^S, m \in \mathcal{S}\right\} \leq \frac{\epsilon_s}{\rho}, \alpha_{i,I}^P = \frac{\epsilon_P}{\rho g_i}, i \in \mathcal{B}_1\right).$$

By using the two choices for $\alpha_{m,II}^S$ shown in (6), Q_4 can be further upper bounded as follows:

$$Q_4 \leq \sum_{\mathcal{B} \subset \mathcal{B}_1} P\left(\alpha_{j,I}^P = 1, j \in \mathcal{B}^c\right) + P\left(\xi_1 \alpha_{1,II}^S \leq \frac{\epsilon_s}{\rho}, \alpha_{1,I}^P > 0\right)$$
$$= \sum_{\mathcal{B} \subset \mathcal{B}_1} P\left(\alpha_{j,I}^P = 1, j \in \mathcal{B}^c\right) + P\left(\frac{h_1 \frac{g_1 - \frac{\epsilon_P}{\rho}}{(\epsilon_P + 1)g_1}}{\epsilon_P \sum_{i \neq 1} \frac{h_i}{g_i} + 1} \leq \frac{\epsilon_s}{\rho}\right)$$
$$+ P\left(\xi_1 \leq \frac{\epsilon_P + \epsilon_s(1 + \epsilon_P)}{\rho}\right). \tag{22}$$

Following the steps to bound Q_1 and Q_2, it is straightforward to show that the first and third terms in (22) go to zero at high SNR. Furthermore, note that $P\left(\frac{h_1 \frac{g_1 - \frac{\epsilon_P}{\rho}}{(\epsilon_P + 1)g_1}}{\epsilon_P \sum_{i \neq 1} \frac{h_i}{g_i} + 1} \leq \frac{\epsilon_s}{\rho}\right) \to \tilde{F}_1(0) = 0$ for $\rho \to \infty$, where $\tilde{F}_1(x)$ denotes the CDF of $\frac{h_1 \frac{g_1}{(\epsilon_P + 1)g_1}}{\epsilon_P \sum_{i \neq 1} \frac{h_i}{g_i} + 1}$ and $\tilde{F}_1(x)$ is not related to ρ. Therefore, $Q_4 \to 0$ for $\rho \to \infty$. Since all the three terms in (13) approach zero at high SNR, the proof is complete.

REFERENCES

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

[2] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of NOMA toward next generation multiple access (NGMA) for 6G," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1037–1071, Jan. 2022.

[3] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[4] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32–38, Apr. 2018.

[5] G. Li, Z. Ming, D. Mishra, L. Hao, Z. Ma, and O. A. Dobre, "Energy-efficient design for IRS-empowered uplink MIMO-NOMA systems," *IEEE Trans. Veh. Tech.*, pp. 1–1, to appear in 2022.

[6] D. N. Amudala, B. Kumar, and R. Budhiraja, "Spatially-correlated Rician-faded multi-relay multi-cell massive MIMO NOMA systems," *IEEE Trans. Commun.*, pp. 1–1, to appear in 2022.

[7] Z. Ding and H. V. Poor, "Joint beam management and power allocation in THz-NOMA networks," *IEEE Trans. Commun.*, Available on-line at arXiv:2205.12938, 2022.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2003.

[9] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse Wishart distributed matrices," *IEE Proc. Radar, Sonar and Navigation*, vol. 147, no. 4, pp. 162–168, Aug. 2000.

[10] Z. Ding and H. V. Poor, "Design of Massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. New York: Academic Press, 2000.