Partial Non-Orthogonal Multiple Access (NOMA) in Downlink Poisson Networks

Konpal Shaukat Ali[†], Ekram Hossain[†], and Md. Jahangir Hossain[+]

Abstract—**Non-orthogonal multiple access (NOMA) allows users sharing a resource-block to efficiently reuse spectrum and improve cell sum rate $\mathcal{R}_{\rm tot}$ at the expense of increased interference. Orthogonal multiple access (OMA), on the other hand, guarantees higher coverage. We introduce partial-NOMA in a large two-user downlink network to provide both throughput and reliability. The associated partial overlap controls interference while still offering spectrum reuse. The nature of the partial overlap also allows us to employ receive-filtering to further suppress interference. For signal decoding in our partial-NOMA setup, we propose a new technique called flexible successive interference cancellation (FSIC) decoding. We plot the rate region abstraction and compare with OMA and NOMA. We formulate a problem to maximize $\mathcal{R}_{\rm tot}$ constrained to a minimum throughput requirement for each user and propose an algorithm to find a feasible resource allocation efficiently. Our results show that partial-NOMA allows greater flexibility in terms of performance. Partial-NOMA can also serve users that NOMA cannot. We also show that with appropriate parameter selection and resource allocation, partial-NOMA can outperform NOMA.**

Index Terms—**Partial non-orthogonal multiple access (NOMA), flexible successive interference cancellation, stochastic geometry, resource allocation**

I. INTRODUCTION

Traditionally, users (UEs) avoid interference from other UEs being served by the same base station (BS) by a multiple access technique known as orthogonal multiple access (OMA) which allocates orthogonal resources to these UEs. This is done by allotting UEs different time slots or different frequency channels; in fact, the the available time and frequency resources are split into a grid of what are referred to as time-frequency resource-blocks (also referred to as resource-blocks from hereon). This way a UE in one resource-block has orthogonal resources to a UE in any other resource-block. In contrast to OMA, in non-orthogonal multiple access (NOMA), multiple UEs share a time-frequency resource-block for transmissions by superposing their messages in the power domain, i.e., multiple UEs transmit their messages in the same time slot over the same frequency channel using different power levels for their messages. Hence, NOMA UEs improve spectral reuse by sharing a resource-block with other UEs but the price paid is the introduction of interference with UEs being served by the BS on the same resource-block, i.e., intracell interference.

NOMA employs successive interference cancellation (SIC) for the decoding of these superposed messages. SIC requires ordering of NOMA UEs based on some measure of channel strength. Most of the existing works on NOMA in the literature order UEs based on the mean signal power received [1]–[6] or on the quality of the transmission channel such as the fading coefficient [7]–[10], the fading-to-noise ratio [11], the instantaneous received signal-to-intercell-interference-and-noise ratio [12], and the instantaneous received signal-to-intercell-interference ratio [13].[1] Such UE ordering and appropriate resource allocation allows a UE to decode messages of UEs weaker than itself and treat the messages of UEs stronger than itself as noise. It has been shown extensively in the literature that NOMA is superior to OMA in terms of the sum throughput that can be achieved [2], [4], [5], [9]–[16].

While NOMA allows complete sharing of a resource-block, thereby resulting in better throughput, the introduction of intracell interference deteriorates the coverage of NOMA UEs. OMA, on the other hand, has no intracell interference, and therefore, provides superior coverage. However, OMA limits only one UE to a resource-block thereby not making efficient use of the scarce spectrum which leads to lower throughput. To cater to better coverage requirements than NOMA as well as better throughput requirements than OMA, we introduce the concept of partial-NOMA. In partial-NOMA, UEs share only a fraction of the resource-block, this way we can reduce the intracell interference but still allow some spectrum reuse. The motivation behind this concept is to introduce flexibility into the system by having control over how much of the resource-block overlaps and does not overlap for each UE. This way, partial-NOMA is a general technique which offers flexibility between the two extremes of traditional OMA and NOMA.

To the best of our knowledge, only the work in [17] studies a partial-NOMA like setup in a two-user downlink scenario. Different from our work, the authors study a single-cell setup where the non-overlapping areas of the resource-block allow each UE's message to be given full power while the overlapping regions share power between the two UEs so that it sums to the full power ([17, Fig. 1]). Such an analysis is equivalent to studying the average performance of a system with OMA in some resource-blocks and NOMA in other resource-blocks. In our work, we study the performance of shared power in one resource-block that sums up to the total power, with a partial area of overlap (c.f. Fig. 1). Additionally, we study a large network since a single-cell setup does not account for

[†] The authors are with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada (Email: {konpal.ali, ekram.hossain}@umanitoba.ca). E. Hossain is the corresponding author.
[+] The author is with the School of Engineering, the University of British Columbia (Okanagan Campus), Canada (Email: jahangir.hossain@ubc.ca).

[1]Note that UE ordering in NOMA is based on a measure of channel strength and does not imply that the message of a strong (weak) UE has high (low) transmission power.

intercell interference. Accounting for intercell interference is particularly important for setups where multiple UEs share a resource-block (such as NOMA and partial-NOMA) as it has a drastic negative impact on both performance and resource allocation, as was shown in [18]. Stochastic geometry has succeeded in providing a unified mathematical paradigm for modeling large wireless cellular networks and characterizing their operation while taking into account intercell interference [19]–[22]. Works on NOMA such as [4]–[6], [13], [16], [23] use stochastic geometry based modeling for analyzing large networks. A number of works on NOMA have focused on resource allocation as well [2], [5], [10]–[12], [15], [24], [25]. We also study resource allocation in this paper for a partial-NOMA setup.

In this work, we analytically study a large multi-cell downlink system that employs partial-NOMA for the two-user setup. Partial sharing of a resource-block in our work is accomplished by having the two signals overlap only with a fraction of each other in the frequency domain while having complete access to the entire time slot. An overlap in the frequency domain allows us to employ matched filtering at the receiver side to reduce the impact of the intracell interference[2]. Traditional NOMA relies on SIC, which involves a strong UE decoding and removing the message of the weak UE before it decodes its own message. In partial-NOMA, however, after matched filtering at the receiver side, the message of the weak UE may be too weak for the strong UE to be able to decode. To combat this issue and improve performance, we propose a new decoding scheme, which we call flexible-SIC (FSIC) decoding. Using stochastic geometry tools, we mathematically analyze the performance of a large network with partial-NOMA employing FSIC decoding. The rate region for NOMA and OMA has been studied in the literature. To benchmark the performance of a partial-NOMA setup, we study the rate region abstraction for different values of the overlap[3]. The partial-NOMA setup introduces two new parameters that can be varied, the overlap and the amount of non-overlap given to each UE. Accordingly, the rate region for different values of the overlap is not enough to quantify performance. A more practical problem is that of maximizing cell sum rate, defined as the sum of the throughput of the two UEs sharing a resource-block, subject to a *threshold minimum throughput* (TMT) constraint on the individual UEs. In this context, we formulate an optimization problem and propose an algorithm for resource allocation. We show a significant reduction in complexity between our proposed algorithm and an exhaustive search.

The contributions of this paper can be summarized as follows:

- We show that our partial-NOMA setup can result in lower intercell interference than both traditional OMA and NOMA thanks to the received filtering.

---

[2]Note that the intracell interference is only a fraction of the intracell interference experienced in the case of traditional NOMA. However, with matched filtering at the receiver we can further reduce its impact.

[3]The rate region will be abstracted into two figures to make it easier to read for different values of the overlap.

- While it is well known that the impact of bandwidth is more significant on throughput than the impact of interference, we show that partial-NOMA is able to outperform traditional NOMA in terms of the cell sum rate in the rate region. This superiority is due to the associated received filtering and proposed FSIC decoding.
- We show that in terms of individual UE throughput, NOMA that has complete overlap is closer to OMA that has no overlap, while partial-NOMA allows the individual rates to stray farther away. This highlights the greater flexibility in individual UE performance that a partial-NOMA setup introduces.
- We show that while traditional NOMA cannot support UEs with high transmission rate requirements, partial-NOMA can. Thus instead of allocating an entire resource-block to such UEs, they can be served in a partial-NOMA fashion to efficiently reuse the spectrum.
- We show that with careful resource allocation, partial-NOMA with a range of overlap values outperforms traditional NOMA, in terms of cell sum rate. We also show that an optimum overlap exists that maximizes the cell sum rate given a threshold minimum throughput constraint.

The rest of the paper is organized as follows. The system model is described Section II. The SINR analysis, FSIC decoding and relevant statistics are in Section III. In Section IV, the rate region for partial-NOMA is studied, an optimization problem is formulated, and an algorithm is proposed to solve the problem. The results are presented in Section V and the paper is concluded in Section VI.

Notation: Vectors are denoted using bold text, $\|\mathbf{x}\|$ denotes the Euclidean norm of the vector $\mathbf{x}$, $b(\mathbf{x}, r)$ denotes a disk centred at $\mathbf{x}$ with radius r. $\mathcal{L}_X(s) = \mathbb{E}[e^{-sX}]$ denotes the Laplace transform (LT) of the PDF of the random variable X. We use the indicator function, denoted as $\mathbb{1}_A$, to have value 1 when event A occurs and to be 0 otherwise. The ordinary hypergeometric function is denoted by ${}_2F_1$. We use $\text{Sinc}(x) = \sin(\pi x)/(\pi x)$ when $x \neq 0$, and $\text{Sinc}(x) = 1$ when $x = 0$.

II. SYSTEM MODEL

A. Network Model for Partial-NOMA

We consider a downlink cellular network where each BS serves two UEs in each time-frequency resource-block. In traditional NOMA, the two UEs share the entire resource-block (i.e., each UE has access to the full time slot and frequency channel associated with the resource-block), having their messages multiplexed in the power domain. In contrast to this, in partial-NOMA, the messages of the two UEs only overlap over a fraction α of the resource-block. This overlap of the resource-block can be achieved in two ways:

1) Both UEs transmit over the entire time slot but there is only an overlap of the fraction α of the frequency channel shared by the two UEs.
2) Both UEs occupy the entire frequency channel but simultaneous transmission only happens for a fraction α of the time slot.

As mentioned in Section I, in this work, we consider the first approach, i.e., the two UEs only share a fraction α of the frequency resources while transmitting in the entire time slot of a resource-block. This allows us to take advantage of matched filtering explained in Section II-B which would not have been possible with the second approach. With a slight abuse of notation, in the remainder of the manuscript, an overlap of α of the resource-block refers to an overlap of α in the frequency domain. Since the entire time slot is available to both UEs, we disregard this aspect when referencing the partial overlap of the resource-block. Note that we stick to the notation of resource-block to remain consistent with works on traditional NOMA.

As shown in Fig. 1, we refer to the fraction of the resource-block designated to only UE$_1$ by β; consequently, the fraction designated to only UE$_2$ is $1 - \alpha - \beta$. Thus, the effective bandwidth of UE$_1$ is BW$_1 = \alpha + \beta$, and that of UE$_2$ is BW$_2 = 1 - \beta$. The BSs use fixed-rate transmissions where the transmission rate of each UE can be different. Such transmissions result in effective rates, referred to as the throughput of the UEs, that are lower than the transmission rate because of outage.
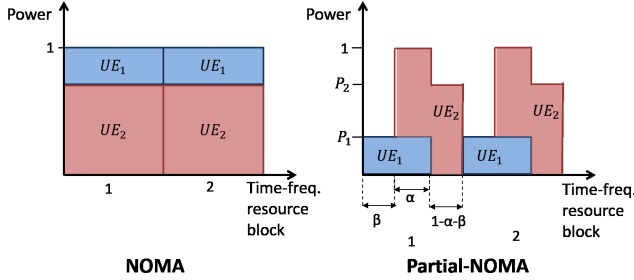

Fig. 1: A contrast between resource allocation in traditional NOMA and partial-NOMA.

The BSs are distributed according to a homogeneous Poisson point process (PPP) Φ with intensity λ. To the network we add a BS at the origin $\mathbf{o}$, which under expectation over Φ, becomes the typical BS serving UEs in the typical cell. We study the typical cell in the remainder of this work. Note that as Φ does not include the BS at $\mathbf{o}$, the set of interfering BSs for the UEs in the typical cell is denoted by Φ. We denote by ρ the distance between the typical BS at $\mathbf{o}$ and its nearest neighboring BS. Since Φ is a PPP, the distribution of the distance ρ thus follows

$$f_\rho(x) = 2\pi\lambda x e^{-\pi\lambda x^2}, \quad x \geq 0. \quad (1)$$

An important challenge associated with NOMA is the selection of the UEs that share a resource-block, which is referred to as UE clustering. There is a common misconception that clustering NOMA UEs with high channel disparity is beneficial. However, works such as [26] have shown that this not necessarily the case and in [6], [27] it is explicitly shown that in fact clustering UEs with lower channel disparity but overall better channel conditions is superior. Hence, we use the cell-center model for the selection of NOMA UEs employed in [12, Model 1]. As such, we consider a disk centred at $\mathbf{o}$ with radius $\rho/2$, $b(\mathbf{o}, \rho/2)$, referred to as the in-disk. The

in-disk is the largest disk centred at the serving BS that fits inside the Voronoi cell. The rationale behind employing this model is that UEs outside of this disk are relatively far from their BS, have weaker channels and thus are better served in their own resource-block (without sharing) or even using coordinated multipoint (CoMP) transmission if they are near the cell edge [28], [29]. These UEs are not discussed further in this work. We focus on UEs inside the in-disk since they have good channel conditions, yet enough disparity among themselves, and thus can effectively be served while (partially) sharing a resource-block. The two UEs are located uniformly at random in the in-disk, independent of one another. Note that this way the UEs form a Poisson cluster process where two daughter points are placed independently and uniformly at random on disks of varying random radii. Such a model is superior to using a Matern cluster process where the radius of the disks is fixed and thus: 1) risks a disk of one BS going into the cell of a neighboring BS, 2) risks overlap between disks.

A BS transmits with a power budget P, where the power for the signal intended for UE$_i$ is denoted by P_i and $P = \sum_{i=1}^{2} P_i$; without loss of generality we set $P = 1$. A Rayleigh fading environment is assumed such that the fading coefficients are i.i.d. with a unit mean exponential distribution. A power law path-loss model is considered where the signal power decays with distance r at the rate $r^{-\eta}$, where $\eta > 2$ denotes the path-loss exponent.

As has been mentioned, NOMA requires ordering the UEs based on some measure of channel strength. In this work, we order the UEs based on the link distance, R, between the typical BS at $\mathbf{o}$ and its UE uniformly distributed in the in-disk is conditioned on ρ. Note that ordering based on increasing link distance is equivalent to ordering based on the decreasing mean signal power received, i.e., $R^{-\eta}$. We thus refer to the strong UE, with the shorter link distance, as UE$_1$ and the weak UE as UE$_2$. As the order of the UEs is known at the BS, we use ordered statistics for the pdf of R_i, the ordered link distance of UE$_i$, where $i \in \{1, 2\}$. Hence, using the theory of order statistics [30], in the typical cell

$$f_{R_i|\rho}(r \mid \rho) = \frac{16r}{\rho^2}\left(\frac{4r^2}{\rho^2}\right)^{i-1}\left(1 - \frac{4r^2}{\rho^2}\right)^{2-i} \quad 0 \leq r \leq \frac{\rho}{2}. \quad (2)$$

Note that as there is no interfering BS inside $b(\mathbf{o}, \rho)$, the nearest interfering BS from UE$_i$ is at least $\rho - R_i$ away. Thus the in-disk model allows a larger lower bound on the distance from the nearest interferer than the usual lower bound of link distance for UEs in a downlink Poisson network [12].

B. Filtering at the Receiver

Since partial-NOMA is studied in this work, the signal of interest at a UE only overlaps with a fraction α in the resource-block with the other UE's signal. Accordingly, it is important to consider the impact of matched filtering at the receiver, along the lines of the works in [31], [32]. The message of the signal for UE$_i$ is transmitted using the pulse shape $s_i(t) \leftrightarrow S_i(f)$. The signal is passed through a matched filter $H_i(f) = S_i^*(f)$ at the receiver before decoding, where

$S_i^*(f)$ denotes the conjugate of $S_i(f)$. Accordingly, the impact of the interference caused at UE$_i$ by a message that has an α-overlap, is referred to as the effective interference factor $\mathcal{I}_i(\alpha, \beta)$. Note that $\mathcal{I}_i(\alpha, \beta)$ is a function of both α and β as both are required to determine the center frequencies of the signals. In this work, we assume that square pulses are used by both UEs for transmissions. Because of the choice of $H_i(f)$ and since the same pulse shape is transmitted by both UEs, $\mathcal{I}_i(\alpha, \beta) = \mathcal{I}_j(\alpha, \beta)$ when $i \neq j$; hence, in the remainder of the manuscript, we drop the subscript. Along the lines of [31], the effective interference factor as a function of β and the overlap α is calculated as follows:

$$\mathcal{I}(\alpha, \beta) = \left(\int_{\beta}^{\beta + \alpha} \frac{\text{Sinc}\left(\frac{2(f - f_a)}{BW_1}\right)}{E_1} \frac{\text{Sinc}\left(\frac{2(f - f_b)}{BW_2}\right)}{E_2} df \right)^2, \quad (3)$$

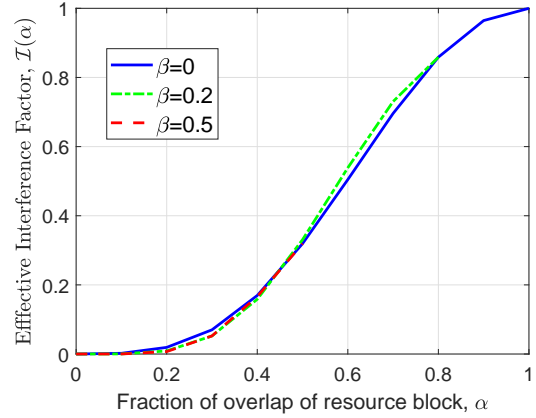where the center frequency of UE$_1$'s message is $f_a = \frac{\alpha + \beta}{2}$ and UE$_2$'s message is $f_b = \frac{1 + \beta}{2}$. The factors E_i for $i \in \{1, 2\}$ are used to scale the energy to 1 and are calculated as $E_i^2 = \int_{-BW_i/2}^{BW_i/2} \text{Sinc}^2\left(\frac{2f}{BW_i}\right) df$.

Note that we define $\mathcal{I}(\alpha, \beta)$ scaled to unit power from the interfering message as this is the fraction of the interference that actually impacts the received message. The actual interfering power, which is the transmit power of the message scaled by $\mathcal{I}(\alpha, \beta)$, will be accounted for in the SINR analysis. Additionally, it is important to emphasize that any interfering message that has an α-overlap with the message of UE$_i$, whether it is from inside the typical cell or outside, will have its power scaled by $\mathcal{I}(\alpha, \beta)$.
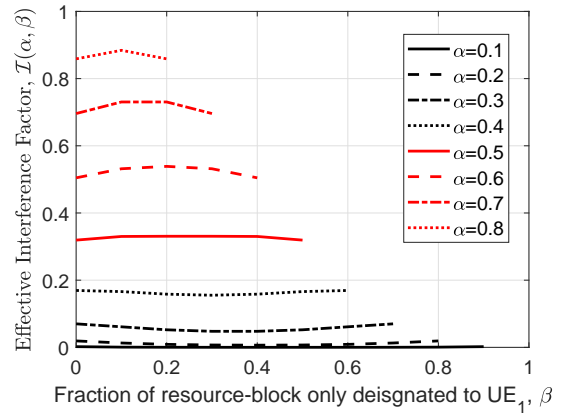
Fig. 2 plots $\mathcal{I}(\alpha, \beta)$ for the case of square pulses used in this work. It should be noted from Fig. 2a that when the overlap α is 0 (1), as in the case of OMA (traditional NOMA), the effective interference factor is 0 (1). Also note that for a given β, α is restricted to a maximum value of $1 - \beta$; hence, when $\beta > 0$, $\alpha \leq 1 - \beta < 1$. Accordingly, $\mathcal{I}(\alpha, \beta)$ in Fig. 2a is plotted up to the maximum value of the overlap α possible. For a given α, on the other hand, β can take on values $0 \leq \beta \leq \beta_{\max}$, where $\beta_{\max} = 1 - \alpha$. We observe in Fig. 2b that for each value of α, the curve for $\mathcal{I}(\alpha, \beta)$ vs. β is symmetric about the midpoint which occurs at $\beta = \beta_{\max}/2$. Additionally, it is important to note that for smaller (larger) values of α, $\mathcal{I}(\alpha, \beta)$ decreases (increases) from $\beta = 0$ to $(1 - \alpha)/2$ (i.e., $\beta_{\max}/2$ the point of symmetry) and then increases (decreases) from $\beta = (1 - \alpha)/2$ to $1 - \alpha$.

C. Decoding of Partial-NOMA

Traditional NOMA employs successive interference cancellation (SIC) for decoding. Hence, in the downlink, the strong UE first decodes the weak UE's message and subtracts it from the signal before decoding its own message. The weak UE, on the other hand, only decodes its own message, treating the signal of the strong UE as noise. Accordingly, the power allocation and transmission rate are designed so that decoding the weak UE's message is easier; this is done by allocating it higher power than the strong UE and/or low transmission rate. Unlike traditional NOMA, however, the intracell interference

(a) As a function of the overlap α for different values of β.

(b) As a function of β for different values of the overlap α.

Fig. 2: Effective interference factor for square pulses.

is scaled by the effective interference factor $\mathcal{I}(\alpha, \beta)$ which can be much lower than 1. While this is always beneficial for the weak UE which treats intracell interference as noise, it can make traditional SIC decoding difficult for the strong UE which has to decode the weak UE's message as the effective power of the weak message at the strong UE could be significantly lower. In this context, we propose FSIC in the next section which is a more effective decoding technique for the partial-NOMA setup.

III. SINR ANALYSIS

A. The SINRs of Interest

In the two-user downlink partial-NOMA setup, we are interested in the following SINRs associated with decoding:

- the message of UE$_2$ at UE$_2$: SINR_2^2,
- the message of UE$_2$ at UE$_1$: SINR_2^1,
- the message of UE$_1$ at UE$_1$ after the message of UE$_2$ has been removed: SINR_1^1,
- the message of UE$_1$ at UE$_1$ if the message of UE$_2$ is not removed: $\widetilde{\text{SINR}}_1^1$.

Accordingly, the SINR of decoding the jth message at UE$_i$, where $j \geq i$, assuming that messages of all UEs weaker than

UE_j have been removed is

$$\text{SINR}_j^i = \frac{h_i R_i^{-\eta} \left(P_j \mathcal{I}(\alpha, \beta) \mathbb{1}_{i \neq j} + P_i \mathbb{1}_{i=j}\right)}{h_i R_i^{-\eta} \left(\left(1 - \sum_{k=j}^2 P_k\right) \mathcal{I}(\alpha, \beta) \mathbb{1}_{i=j} + P_i \mathbb{1}_{i \neq j}\right) + \tilde{I}_i^\emptyset + \sigma^2}.$$
(4)

The intercell interference experienced at UE_i is $\tilde{I}_i^\emptyset = (P_i + (1 - P_i)\mathcal{I}(\alpha, \beta)) \sum_{\mathbf{x} \in \Phi} g_{\mathbf{y}_i} \|\mathbf{y}_i\|^{-\eta}$, where $\mathbf{y}_i = \mathbf{x} - \mathbf{u}_i$, $\mathbf{u}_i$ is the location of UE_i. The fading coefficient from the serving BS (interfering BS) located at $\mathbf{o}$ ($\mathbf{x}$) to UE_i is h_i ($g_{\mathbf{y}_i}$). For notational convenience, we define the intercell interference scaled to unit transmission power by each interferer as $I_i^\emptyset$; hence, $\tilde{I}_i^\emptyset = (P_i + (1 - P_i)\mathcal{I}(\alpha, \beta)) I_i^\emptyset$. The noise power is denoted by σ^2. Using these notations, $\widetilde{\text{SINR}}_1^1$ is

$$\widetilde{\text{SINR}}_1^1 = \frac{h_1 R_1^{-\eta} P_1}{h_1 R_1^{-\eta} P_2 \mathcal{I}(\alpha, \beta) + \tilde{I}_1^\emptyset + \sigma^2}.$$
(5)

Remark 1: It ought to be mentioned that because of the received filtering and associated interference factor $\mathcal{I}(\alpha, \beta)$, the partial-NOMA setup experiences lower intercell interference than both OMA and traditional NOMA.

B. FSIC Decoding

FSIC decoding for UE_2 is the same as that of SIC decoding, i.e., UE_2 decodes its message while treating the interference from the message of UE_1 as noise. Accordingly, the event of successful decoding at UE_2 is defined as $C_2 = \left\{\text{SINR}_2^2 > \theta_2\right\}$, where θ_2 is the SINR threshold corresponding to the transmission rate of UE_2's message $\log(1 + \theta_2)$. Note that the (intracell) interference in the case of partial-NOMA is lower than in the case of traditional NOMA as it is scaled by $\mathcal{I}(\alpha, \beta)$. For UE_1, on the other hand, the message of interest can be decoded successfully if either: 1) the message of UE_2 can be decoded (treating the message of UE_1 as noise) and removed, followed by decoding of the message of UE_1[4], or 2) the message of UE_1 can be decoded while treating the interference from the message of UE_2 as noise. Accordingly, it is defined by the following joint event $C_1 = \left\{\left(\text{SINR}_2^1 > \theta_2 \cap \text{SINR}_1^1 > \theta_1\right) \cup \widetilde{\text{SINR}}_1^1 > \theta_1\right\}$, where θ_1 is the SINR threshold corresponding to the transmission rate of UE_1's message $\log(1 + \theta_1)$.

C. Coverage Analysis and Throughput

Lemma 1: The event C_2 can be rewritten as follows:

$$C_2 = h_2 > R_2^\eta (\tilde{I}_2^\emptyset + \sigma^2) \bar{M}_2$$
(6)

where we have used

$$\tilde{P}_2 = P_2 - \theta_2 P_1 \mathcal{I}(\alpha, \beta),$$

such that

$$\bar{M}_2 = \frac{\theta_2}{\tilde{P}_2}.$$
(7)

[4]We assume perfect SIC in this work. As a result, there is no leakage of the canceled message of the weak UE which would add as interference when decoding the message of the strong UE.

Proof: Using (4) and the definition of C_2, (6) is obtained as follows:

$$
\begin{aligned}
C_2 &= \left\{\text{SINR}_2^2 > \theta_2\right\} \\
&= \left\{\frac{h_2 R_2^{-\eta} P_2}{h_2 R_2^{-\eta} P_1 \mathcal{I}(\alpha, \beta) + \tilde{I}_2^\emptyset + \sigma^2} > \theta_2\right\} \\
&= \left\{h_2 > R_2^\eta \left(\tilde{I}_2^\emptyset + \sigma^2\right) \frac{\theta_2}{P_2 - \theta_2 P_1 \mathcal{I}(\alpha, \beta)}\right\}.
\end{aligned}
$$

∎

Lemma 2: The event C_1 can be rewritten as follows:

$$C_1 = h_1 > R_1^\eta (\tilde{I}_1^\emptyset + \sigma^2) \bar{M}_1.$$
(8)

where we have used

$$M_0 = \frac{\theta_1}{\tilde{P}_1},$$

$$M_1 = \max\left\{\frac{\theta_2}{\tilde{P}_2^1}, \frac{\theta_1}{P_1}\right\},$$

$$\tilde{P}_1 = P_1 - \theta_1 P_2 \mathcal{I}(\alpha, \beta),$$

and $$\tilde{P}_2^1 = P_2 \mathcal{I}(\alpha, \beta) - \theta_2 P_1,$$

such that

$$\bar{M}_1 = \min\{M_0, M_1\} \mathbb{1}_{\tilde{P}_1 > 0} \mathbb{1}_{\tilde{P}_2^1 > 0} \mathbb{1}_{P_1 > 0} +$$
$$M_0 \mathbb{1}_{\tilde{P}_1 > 0} \mathbb{1}_{\tilde{P}_2^1 \leq 0 \cup P_1 \leq 0} + M_1 \mathbb{1}_{\tilde{P}_1 \leq 0} \mathbb{1}_{\tilde{P}_2^1 > 0} \mathbb{1}_{P_1 > 0}.$$
(9)

Proof: Using (4), (5) and the definition of C_1:

$$
\begin{aligned}
C_1 &= \left\{\left(\text{SINR}_2^1 > \theta_2 \cap \text{SINR}_1^1 > \theta_1\right) \cup \widetilde{\text{SINR}}_1^1 > \theta_1\right\} \\
&= \left\{\left(\frac{h_1 R_1^{-\eta} P_2 \mathcal{I}(\alpha, \beta)}{h_1 R_1^{-\eta} P_1 + \tilde{I}_1^\emptyset + \sigma^2} > \theta_2 \bigcap \frac{h_1 R_1^{-\eta} P_1}{\tilde{I}_1^\emptyset + \sigma^2} > \theta_1\right) \bigcup \right. \\
&\qquad \left. \frac{h_1 R_1^{-\eta} P_1}{h_1 R_1^{-\eta} P_2 \mathcal{I}(\alpha, \beta) + \tilde{I}_1^\emptyset + \sigma^2} > \theta_1\right\} \\
&= \left\{h_1 > R_1^\eta \left(\tilde{I}_1^\emptyset + \sigma^2\right) M_1 \bigcup h_1 > R_1^\eta \left(\tilde{I}_1^\emptyset + \sigma^2\right) M_0\right\}.
\end{aligned}
$$

For the event $\left\{\left(\text{SINR}_2^1 > \theta_2 \cap \text{SINR}_1^1 > \theta_1\right)\right\}$, $\tilde{P}_2^1 > 0$ and $P_1 > 0$ are required. For the event $\left\{\widetilde{\text{SINR}}_1^1 > \theta_1\right\}$, $\tilde{P}_1 > 0$ is required. Thus, (8) is obtained by rewriting C_1 in terms of these conditions as follows:

$$
\begin{aligned}
C_1 &= \left\{h_1 > R_1^\eta \left(\tilde{I}_1^\emptyset + \sigma^2\right) \left(\min\{M_0, M_1\} \mathbb{1}_{\tilde{P}_1 > 0} \mathbb{1}_{\tilde{P}_2^1 > 0} \mathbb{1}_{P_1 > 0}\right.\right. \\
&\qquad \left.\left. + M_0 \mathbb{1}_{\tilde{P}_1 > 0} \mathbb{1}_{\tilde{P}_2^1 \leq 0 \cup P_1 \leq 0} + M_1 \mathbb{1}_{\tilde{P}_1 \leq 0} \mathbb{1}_{\tilde{P}_2^1 > 0} \mathbb{1}_{P_1 > 0}\right)\right\}.
\end{aligned}
$$

∎

Remark 2: Note that the introduction of intracell interference impacts the transmit power of the message being decoded by deteriorating it to what is referred to as the effective transmit power [12]. The amount of deterioration is the intracell interference experienced scaled by the SINR threshold corresponding to the transmission rate of the message to be decoded. Thus, $\tilde{P}_2$, $\tilde{P}_1$, and $\tilde{P}_2^1$ are the effective transmit powers that have experienced a reduction from the power of the messages to be decoded (P_2, P_1 and $P_2 \mathcal{I}(\alpha, \beta)$, respectively).

Using [12, Lemma 1], the LT of I_i^ϕ at the typical UE_i conditioned on R_i and ρ, where $u_i = \rho - R_i$, is approximated as

$$\mathcal{L}_{I_i^\phi | R_i, \rho}(s) \approx \exp\left(\frac{-2\pi\lambda s}{(\eta-2)u_i^{\eta-2}} {}_2F_1\left(1, 1-\delta; 2-\delta; \frac{-s}{u_i^\eta}\right)\right) \frac{1}{1+s\rho^{-\eta}} \tag{10}$$

$$\overset{\eta=4}{=} e^{-\pi\lambda\sqrt{s}\tan^{-1}\left(\frac{\sqrt{s}}{u_i^2}\right)} \frac{1}{1+s\rho^{-4}}. \tag{11}$$

Theorem 1: If the effective transmit power is positive, the coverage probability of the typical UE_i is approximated as

$$\mathbb{P}(C_i) \approx \int_0^\infty \int_0^{x/2} e^{-r^\eta \sigma^2 \bar{M}_i} \mathcal{L}_{I_i^\phi | R_i, \rho}\left(r^\eta \bar{M}_i \left(P_i + (1-P_i)\mathcal{I}(\alpha, \beta)\right)\right)$$
$$\times f_{R_i | \rho}(r \mid x) \, dr \, f_\rho(x) \, dx, \tag{12}$$

where the LT of I_i^ϕ conditioned on R_i and ρ is approximated in (10). For UE_1, the effective transmit power is positive if $\{\tilde{P}_2^1 > 0, P_1 > 0\}$ and/or $\tilde{P}_1 > 0$, while for UE_2, it is positive when $\tilde{P}_2 > 0$. If effective transmit power is not positive, $\mathbb{P}(C_i) = 0$.

Proof: Using C_i in Lemma 1 or Lemma 2, if the effective transmit power is positive, we can write

$$\mathbb{P}(C_i) = \mathbb{P}\left(h_i > R_i^\eta(\tilde{I}_i^\phi + \sigma^2)\bar{M}_i\right)$$
$$\overset{(a)}{=} \mathbb{E}_\rho\left[\mathbb{E}_{R_i | \rho}\left[e^{-R_i^\eta \sigma^2 \bar{M}_i} \mathbb{E}_{I_i^\phi | R_i, \rho}\left[e^{-R_i^\eta \bar{M}_i(P_i + (1-P_i)\mathcal{I}(\alpha,\beta))I_i^\phi}\right]\right]\right]$$
$$\approx \mathbb{E}_\rho\left[\mathbb{E}_{R_i | \rho}\left[e^{-R_i^\eta \sigma^2 \bar{M}_i} \mathcal{L}_{I_i^\phi | R_i, \rho}\left(R_i^\eta \bar{M}_i \left(P_i + (1-P_i)\mathcal{I}(\alpha,\beta)\right)\right)\right]\right],$$

where (a) follows from $h_i \sim \exp(1)$ and using $\tilde{I}_i^\phi = (P_i + (1-P_i)\mathcal{I}(\alpha,\beta))I_i^\phi$. The coverage probability becomes an approximation when the approximate LT of I_i^ϕ (given R_i and ρ), $\mathcal{L}_{I_i^\phi | R_i, \rho}$, is used. From this, we obtain (12). Since $h_i \geq 0$, if the effective transmit power is not positive, $\mathbb{P}(C_i) = 0$. ∎

The throughput of UE_i, $i \in \{1, 2\}$, for a given SINR threshold of θ_i corresponding to a transmission rate of $\log(1 + \theta_i)$ is given by

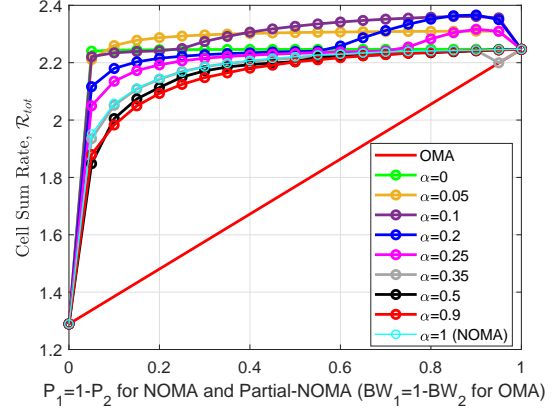$$\mathcal{R}_i = \text{BW}_i \, \mathbb{P}(C_i) \log(1 + \theta_i). \tag{13}$$

We define the cell sum rate $\mathcal{R}_{\text{tot}}$ as the sum of the throughput of the UEs in the typical cell; thus, $\mathcal{R}_{\text{tot}} = \mathcal{R}_1 + \mathcal{R}_2$.

It ought to be mentioned that in a partial-NOMA setup, the resources to be allocated are the powers, the transmission rates, and the overlapping as well as non-overlapping fractions of the resource-block; this means allocating $P_1(= 1 - P_2)$, θ_1, θ_2, α and β.
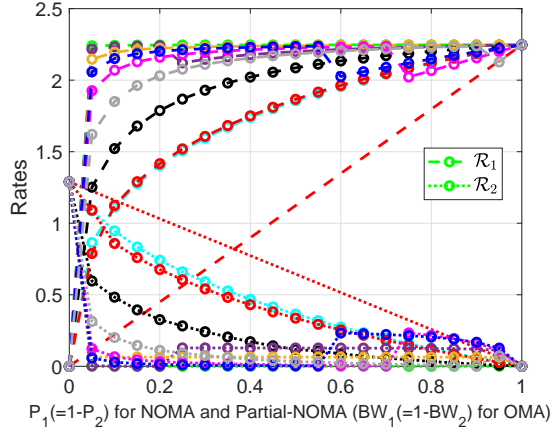
IV. RATE REGION AND OPTIMIZATION

A. Rate Region Abstraction

Fig. 3 is an abstraction of the rate region typically plotted for NOMA and OMA. For NOMA and partial-NOMA, the rate region plots the throughput of UE_1 against the throughput of UE_2 for every $P_1(= 1 - P_2)$ from 0 to 1. For OMA, on the other hand, the rates are plotted for every $\text{BW}_1(= 1 - \text{BW}_2)$; of course OMA transmissions enjoy full power. It should be noted that, although traditionally OMA is defined as one

(a) Cell Sum Rate

(b) Individual UE Throughput. Dashed (dotted) lines are used for $\mathcal{R}_1$ ($\mathcal{R}_2$).

Fig. 3: Rates vs. $P_1(= 1 - P_2)$ for different α with corresponding optimum β, θ_1, θ_2 to maximize $\mathcal{R}_{\text{tot}}$.

UE having access to the entire resource-block, in the context of rate regions for the sake of comparison, two UEs share non-overlapping fractions of the same resource-block. In our work, since partial-NOMA UEs have their overlap in the frequency domain, OMA UEs share non-overlapping fractions of the frequency channel, i.e., non-overlapping fractions of the bandwidth BW_1; thus, $\text{BW}_2 = 1 - \text{BW}_1$. This way the rate region compares the performance of the following: 1) OMA: where two UEs split the bandwidth resource in a non-overlapping fashion while each utilizing the full power (i.e., $P = 1$) for their transmissions, 2) NOMA: where two UEs fully share the bandwidth resource while splitting the power among themselves for the transmission of their messages, i.e., $P_1 \leq 1$, $P_2 \leq 1$ and $P_1 + P_2 = 1$. Thus, in the literature, rate regions are used to show the superiority of NOMA due to the full spectrum reuse despite the lower transmit powers of the individual UE messages and introduction of intracell interference, over OMA where there are higher (full) transmit powers of individual UE messages and no intracell interference but no spectrum reuse. Here, we aim to shed light on the case of: 3) partial-NOMA: where two UEs split the power among their messages like in the case of traditional NOMA; however, they share only an overlap α of the bandwidth, resulting in both spectrum reuse and intracell interference lower than traditional NOMA.

It should be noted that as there is no overlap of bandwidth in OMA, the OMA case may seem similar to partial-NOMA with $\alpha = 0$. However, the two are different as each UE in OMA has messages with full power, i.e., $P = 1$ while in partial-NOMA with $\alpha = 0$, due to the definition of the setup which is for general α, $P_1 \leq 1$, $P_2 \leq 1$ and $P_1 + P_2 = 1$. Additionally, the rate region curves for OMA are plotted against BW_1, while the rate region curves for partial-NOMA (including $\alpha = 0$), similar to those for traditional NOMA, are plotted against P_1.

Since the rate region for different values of α can be difficult to read in the partial-NOMA setup, we abstract the rate region by plotting the cell sum rate and the individual UE throughput against increasing P_1 for NOMA and partial-NOMA (against increasing BW_1 for OMA) in Figs. 3a and 3b, respectively. It ought to be mentioned that in addition to using the optimum θ_1 and θ_2 that maximize $\mathcal{R}_{\text{tot}}$, as is typical for obtaining the boundary of the rate region, the partial-NOMA setup also involves using the optimum β for each value of α and P_1. We consider BS intensity $\lambda = 10$, noise power $\sigma^2 = -90$ dB and $\eta = 4$.

From Fig. 3a, we observe that partial-NOMA with lower values of α outperforms traditional NOMA in terms of $\mathcal{R}_{tot}$. It is, however, important to mention that partial-NOMA is able to do so because it has the advantage of modified-SIC decoding courtesy of the received filtering. We also observe that all values of α still outperform OMA significantly in terms of $\mathcal{R}_{\text{tot}}$. Additionally, by increasing α from 0, we observe that $\mathcal{R}_{\text{tot}}$ first increases with α, followed by a decrease in $\mathcal{R}_{\text{tot}}$ with α, and then an increase to $\alpha = 1$. Note that the value of α until which $\mathcal{R}_{\text{tot}}$ increases initially, grows with P_1. This trend of an increase in $\mathcal{R}_{\text{tot}}$ with α at first followed by a decrease can be attributed to the trade off between spectrum reuse and interference. In the low α regime, increasing α does not increase $\mathcal{I}(\alpha, \beta)$ significantly and the impact from the resulting increase in interference is lower than the impact of the gains from the increased spectrum reuse with α. This results in an increase in $\mathcal{R}_{\text{tot}}$ with α. After a certain α, the impact of $\mathcal{I}(\alpha, \beta)$ becomes more significant and the impact of the increasing interference with α is more dominant than the impact of the increasing spectrum reuse[5]. We thus observe a decrease in $\mathcal{R}_{\text{tot}}$ with α. At $\alpha = 1$, although interference is maximum, the impact of full spectrum reuse between the two UEs is more significant, resulting in an increase in $\mathcal{R}_{\text{tot}}$. These trends shed light on the existence of a range of smaller values of α which are superior to traditional NOMA in terms of $\mathcal{R}_{\text{tot}}$ because of its spectrum reuse and interference trade off, followed by a range of larger α values that are inferior. This also highlights the importance of the careful choice of α required for different network goals.

Fig. 3b shows that in the case of OMA, the throughput of UE_1 is inferior to any partial-NOMA or NOMA setup while its throughput for UE_2 is superior to any partial-NOMA or NOMA. As α increases from 0 to 1, the throughput of UE_1 (UE_2) decreases (increases). This highlights the unexpected observation that, in terms of the individual UE throughput,

traditional NOMA ($\alpha = 1$) is closer to OMA than partial-NOMA with an overlap $\alpha < 1$. Since the rate region reflects the boundaries of achievable throughput, these results show the ability of partial-NOMA to achieve more disparate performance than the other two schemes, highlighting the potential for greater flexibility. Also, note that for values of α such as $\{0.1, 0.2, 0.25, 0.35\}$ $\mathcal{R}_1$ ($\mathcal{R}_2$) does not increase (decrease) monotonically with P_1. A significant change is seen at these values which corresponds to a switch in the decoding technique from $\bar{M}_1 = M_1$ to $\bar{M}_1 = M_0$ (i.e., from using traditional SIC to when UE_1 does not decode the message of UE_2). This is because at higher P_1, with these relatively smaller α values, decoding the message of the weak UE becomes inefficient so UE_1 starts treating the message of UE_2 as noise. Note that smaller (larger) α values have $\bar{M}_1 = M_0$ ($\bar{M}_1 = M_1$) for all P_1.

It is important to mention that Fig. 3 is plotted to maximize $\mathcal{R}_{\text{tot}}$; in the case of traditional NOMA, β can only take on the value 0 and so the rate region can be used to identify the maximum achievable throughput of a TMT constrained setup. However, when $\alpha < 1$, the selected β impacts performance and so the results in Fig. 3 cannot be used to see the gains that would be achievable from a TMT constrained setup.

B. Problem Formulation − Constrained Cell Sum Rate Maximization

As the abstraction of the rate region plotted in the previous subsection aims to maximize the unconstrained cell sum rate for a given α, in this subsection we formulate a problem for a more practical setup where a TMT is required to be achieved by each UE. Accordingly, we formally state the problem as follows:

- $\mathcal{P}1$ - Maximum cell sum rate, given α, subject to the TMT $\mathcal{T}$:

$$\max_{(P_1, \theta_1, \theta_2, \beta)} \mathcal{R}_{\text{tot}}$$

$$\text{subject to: } \sum_{i=1}^{2} P_i = 1$$

$$0 \leq \beta \leq \beta_{\max}$$

$$\mathcal{R}_i \geq \mathcal{T}, \ i \in \{1, 2\},$$

where $\beta_{\max} = 1 - \alpha$. It is evident that the constraints in $\mathcal{P}1$ are not affine, and therefore, the problem is non-convex. Thus, an optimal solution, i.e., choice of β, $P_1 = (1 - P_2)$ and θ_i for $i \in \{1, 2\}$ that results in the maximum constrained $\mathcal{R}_{\text{tot}}$, can only be obtained by using an exhaustive search.

C. Efficient Algorithm

As has been mentioned, only an exhaustive search over all combinations of $P_1 = (1 - P_2)$, θ_1, θ_2 and β can guarantee the optimum resource allocation for the above problem. In this subsection, we propose an algorithm based on intuition. While we cannot guarantee our algorithm to be optimum, it provides a feasible solution to meet the constraints of the problem.

The following is known:

---

[5]Note that FSIC plays an important role in this; without it, the impact of spectrum reuse would always be more significant.

1) From the rate region for static channels in traditional NOMA, a resource allocation (RA) that results in the weak UE achieving the TMT $\mathcal{T}$, while all of the remaining power being allocated to the strong UE to maximize its throughput, is the optimum solution for that problem. An example of this is presented in [15]. Extending this to our large-scale partial-NOMA setup, for a given α and β, i.e., fixed bandwidths for the two UEs, the optimum RA would require achieving TMT for the weak UE and using the remaining power to maximize $\mathcal{R}_1$.

2) The impact of bandwidth is generally more significant on throughput than the impact of power, as throughput grows linearly with bandwidth but only as the logarithm with power.

Remark 3: In regard to 1), for traditional NOMA, achieving TMT for UE_2 and maximizing $\mathcal{R}_{tot}$ would require allocating the smallest P_2 required to achieve $\mathcal{R}_2 = \mathcal{T}$ [5], [18] so that the largest P_1 possible would be left for UE_1 to maximize its throughput with. This is because a strong UE with its superior channel can obtain more from any power than the weak UE, and hence the least required power should be spent on the weak UE. However, in the partial-NOMA setup, the minimum P_2 that achieves TMT for the weak UE may result in UE_1 being in outage because of the impact of received filtering and corresponding $\mathcal{I}(\alpha, \beta)$. While we still want to allocate least P_2, it may need to be increased so that UE_1 can be in coverage as will be explained.

Remark 4: In light of 2), it may be tempting to think that the largest β, i.e., β_{max} (corresponding to the smallest BW_2), with the smallest P_2 that can achieve TMT for the weak UE will result in the optimum solution. However, if α is small, the largest β may result in insufficient bandwidth for UE_2. This may cause it to either not be able to meet TMT at all or to compensate for the small BW_2 by using a very large P_2 to achieve TMT. The latter would result in very little P_1 being left for UE_1 resulting in very low $\mathcal{R}_1$ because of low coverage despite having a large BW_1.

Hence, we propose opting for an RA strategy that aims to find the lowest P_2 required to meet TMT for UE_2 and obtain the maximum $\mathcal{R}_1$ (and therefore $\mathcal{R}_{tot}$) for each value of β, starting from β_{max} and decreasing it. Starting from the largest β, β_{max}, this is done until $\mathcal{R}_{tot}$ starts decreasing. At this point we have found the optimum β because further decreasing β will only deteriorate $\mathcal{R}_{tot}$. The optimum RA is then selected by choosing the β and its corresponding $P_1 (= 1 - P_2)$, θ_1 and θ_2 that result in the largest $\mathcal{R}_{tot}$.

Using the definitions of $\mathcal{R}_1$ and $\mathcal{R}_2$ based on the developed analysis in Section III, our algorithm for $\mathcal{P}1$ thus solves the problem

$$\max_{(P_1, \theta_1, \theta_2, \beta)} \mathcal{R}_1$$

$$\text{subject to: } \sum_{i=1}^{2} P_i = 1, \ 0 \leq \beta \leq 1 - \alpha, \ \text{and } \mathcal{R}_2 = \mathcal{T}.$$

Given α and β, we first search for the minimum P_2 that allows UE_2 to attain a throughput equal to the TMT; this leaves the largest possible P_1 for UE_1. Corresponding to this P_2, UE_1

can be in the following three states:

- State I ($\tilde{P}_2^1 > 0$): If $\tilde{P}_1 \leq 0$, increasing P_2 makes $\tilde{P}_1$ more negative and can therefore not impact $\mathcal{R}_1$. If $\tilde{P}_1 > 0$, increasing P_2 will make $\tilde{P}_1$ smaller and consequently M_0 larger which will not result in its selection for a potentially larger $\mathcal{R}_1$. Hence, if $\tilde{P}_2^1 > 0$ for the minimum P_2 that can achieve $\mathcal{R}_2 = \mathcal{T}$, the optimum P_2 and θ_2 have been found, the corresponding optimum θ_1 that maximizes $\mathcal{R}_1$ should be selected to maximize $\mathcal{R}_{tot}$.

- State II ($\tilde{P}_2^1 < 0$ and $\tilde{P}_1 > 0$):
  1) Optimize θ_1 to maximize $\mathcal{R}_1$ and store as $\theta_{1,II}$ and $\mathcal{R}_{1,II}$, respectively. Note that here M_0 is selected as $\tilde{P}_2^1 < 0$ and so M_1 is not possible.
  2) Increase P_2 until $\tilde{P}_2^1 > 0$ and M_1 is selected. Calculate the corresponding $\mathcal{R}_1$ using the optimum θ_1 that maximizes it and store as $\mathcal{R}_{1,I}$ and $\theta_{1,I}$, respectively. If M_1 is never selected, $\mathcal{R}_{1,I} = 0$.
  3) Store the larger of the two throughputs $\mathcal{R}_{1,I}$ and $\mathcal{R}_{1,II}$, and its corresponding transmission rate as $\mathcal{R}_1$ and θ_1, respectively.

- State III ($\tilde{P}_2^1 < 0, \tilde{P}_1 \leq 0$): As UE_1 is in outage in this state, increase P_2 until state I or II is achieved and follow the corresponding steps.

We formally state the working in **Algorithm 1**.

In **Algorithm 1**, $\text{flag}_1 = 0$ denotes that State II has not been achieved as of yet for the current value of β while $\text{flag}_1 = 1$ denotes that State II has been achieved at least once. If the power budget has been expended but UE_2 cannot meet the TMT, $\text{flag}_2 = 1$; otherwise, $\text{flag}_2 = 0$. Thus if $\text{flag}_2 = 1$, β needs to be decreased to give UE_2 a larger bandwidth to achieve TMT; if β is decreased to 0 but UE_2 can still not achieve TMT, the TMT is too high to be met by the system and ought to be decreased. Since the range of possible β changes with α, we standardize Δ_β to be a function of β_{max} in line 1 so that we select from a fixed number of β values irrespective of α for fairness between different values of α. Similarly, since the range of transmission rates is $\theta_i \geq 0$ $i \in \{1, 2\}$, we make our search finite by searching in the range $\theta_{LB} \leq \theta_i \leq \theta_{UB}$, increasing in steps of Δ_θ; P_2 is also increased in steps of Δ_P.

Given α and $\mathcal{T}$, **Algorithm 1** starts with β_{max}, the largest value of β, in line 2. For a β, it searches for the smallest P_2 and the corresponding lowest θ_2 that can attain the TMT. If UE_1 can be in State I with the selected P_2 and θ_2, θ_1 is optimized to maximize $\mathcal{R}_1$ and the optimum parameters that maximize $\mathcal{R}_{tot}$ have been found for this β. However, if State I is not achieved, but State II is achieved, we optimize θ_1 to maximize $\mathcal{R}_1$ and store it as $\mathcal{R}_{1,II}$. P_2 is then increased until State I can be achieved. If it is achieved before exhausting the power budget, the corresponding θ_1 is optimized to maximize $\mathcal{R}_1$ and stored as $\mathcal{R}_{1,I}$; $\mathcal{R}_{1,I}$ and $\mathcal{R}_{1,II}$ are compared to see which is larger and the corresponding parameters $P_1 = (1 - P_2)$, θ_1 and θ_2 are stored as the optimum for this β. If State I cannot be achieved, $\mathcal{R}_{1,II}$ and its corresponding parameters are stored. If we are in State III, P_2 is increased until State I or II is achieved and the corresponding steps are followed. However, if the TMT cannot be met for UE_2 using full power, i.e., we are in State III even when P_2 is increased to 1 in line 57, BW_2 is not

---

Algorithm 1 RA for a feasible solution to $\mathcal{P}1$

1: $\beta_{\max} = 1 - \alpha$, $\Delta_\beta = \beta_{\max}/10$, $\mathcal{R}_1^{\beta,\text{vec}} = [\ ]$, $\theta_1^{\beta,\text{vec}} = [\ ]$,
     $P_1^{\beta,\text{vec}} = [\ ]$, $\mathcal{R}_2^{\beta,\text{vec}} = [\ ]$, $\theta_2^{\beta,\text{vec}} = [\ ]$
2: **for** $\beta = \beta_{\max} : -\Delta_\beta : 0$ **do**
3:    $\text{flag}_1 = 0$, State=$[\ ]$
4:    **for** $P_2 = 0 : \Delta_P : 1$ **do**
5:      **for** $\theta_2 = \theta_{LB} : \Delta_\theta : \theta_{UB}$ **do**
6:        Calculate $\mathcal{R}_2$ using (13) with (12) and (7)
7:        **if** $\mathcal{R}_2 \geq \mathcal{T}$ **then**
8:          **if** $\tilde{P}_2^1 > 0$ **then**
9:            State=I
10:          **end if**
11:          Go to 22
12:        **end if**
13:      **end for**
14:      $\text{flag}_2 = 0$
15:      **if** $\mathcal{R}_2 < \mathcal{T}$ **then**
16:        **if** $P_2 = 1$ **then**
17:          $\text{flag}_2 = 1$
18:        **else**
19:          Go to 4
20:        **end if**
21:      **end if**
22:      **if** $\text{flag}_2 = 0$ **then**
23:        **if** State=I **then**
24:          **for** $\theta_1 = \theta_{LB} : \Delta_\theta : \theta_{UB}$ **do**
25:            Calculate $\mathcal{R}_1$ using (13) with (12) and (9)
26:            Update $\mathcal{R}_{1,I}^{\text{vec}} = [\mathcal{R}_{1,I}^{\text{vec}}; \mathcal{R}_1]$
27:          **end for**
28:          Store $\mathcal{R}_{1,I} = \max(\mathcal{R}_{1,I}^{\text{vec}})$ and corresponding $\theta_{1,I}$,
           $P_{1,I}$, $\mathcal{R}_{2,I}$, $\theta_{2,I}$
29:        **else**
30:          **for** $\theta_1 = \theta_{LB} : \Delta_\theta : \theta_{UB}$ **do**
31:            **if** $\tilde{P}_1 > 0$ **then**
32:              State=II
33:              **if** $\text{flag}_1 = 0$ **then**
34:                Calculate $\mathcal{R}_1$ using (13) with (12) and (9)
35:                Update $\mathcal{R}_{1,II}^{\text{vec}} = [\mathcal{R}_{1,II}^{\text{vec}}; \mathcal{R}_1]$
36:              **end if**
37:            **end if**
38:          **end for**
39:          Store $\mathcal{R}_{1,II} = \max(\mathcal{R}_{1,II}^{\text{vec}})$ and corresponding
           $\theta_{1,II}$, $P_{1,II}$, $\mathcal{R}_{2,II}$, $\theta_{2,II}$
40:          **if** $\mathcal{R}_{1,II} = 0$ **then**
41:            State=III
42:            Go to 4
43:          **else**
44:            $\text{flag}_1 = 1$
45:            **if** $P_2 < 1$ **then**
46:              Go to 4
47:            **end if**
48:          **end if**
49:        **end if**
50:      **end if**
51:      **if** $\text{flag}_1 = 1$ **then**
52:        Store $\mathcal{R}_1^{\beta,\text{vec}} = [\max(\mathcal{R}_{1,I}, \mathcal{R}_{1,II}); \mathcal{R}_1^{\beta,\text{vec}}]$ and corresponding $\theta_1^{\beta,\text{vec}}$, $P_1^{\beta,\text{vec}}$, $\mathcal{R}_2^{\beta,\text{vec}}$, $\theta_2^{\beta,\text{vec}}$
53:        Go to 64
54:      **else**
55:        **if** $\text{flag}_2 = 1$ **then**
56:          TMT cannot be met by UE$_2$
57:          Go to 2
58:        **else**
59:          Store $\mathcal{R}_1^{\beta,\text{vec}} = [\mathcal{R}_{1,I}; \mathcal{R}_1^{\beta,\text{vec}}]$ and corresponding $\theta_1^{\beta,\text{vec}}$, $P_1^{\beta,\text{vec}}$, $\mathcal{R}_2^{\beta,\text{vec}}$, $\theta_2^{\beta,\text{vec}}$
60:          Go to 64
61:        **end if**
62:      **end if**
63:    **end for**
64:    **if** $\beta > \beta_{\max}$ **then**
65:      **if** $\mathcal{R}_1^{\beta,\text{vec}}(\text{end}) < \mathcal{R}_1^{\beta,\text{vec}}(\text{end} - 1)$ **then**
66:        Store $\mathcal{R}_1 = \max(\mathcal{R}_1^{\beta,\text{vec}})$ and corresponding θ_1, P_1, $\mathcal{R}_2$, θ_2, β.
67:        **if** $\mathcal{R}_1 < \mathcal{T}$ **then**
68:          TMT cannot be met by UE$_1$
69:        **end if**
70:        β that maximizes $\mathcal{R}_{\text{tot}}$ found; **exit**
71:      **end if**
72:    **end if**
73: **end for**

---

sufficient, $\text{flag}_2 = 1$, and we go to line 2 to reduce β. If the TMT is met by UE$_2$ and $\mathcal{R}_1^{\beta,\text{vec}}$ is stored for the iteration of β, we go to line 64. If this is the first iteration of the β loop, we go to the next iteration. If it is not the first iteration and if the throughput of UE$_1$ calculated is larger than that in the last iteration of the β loop, we again go to the next iteration of the β loop. However, if it is not the first iteration of the β loop, and the throughput of UE$_1$ calculated is smaller than that in the last iteration, the optimum throughput of UE$_1$ has already been found and we store it as $\mathcal{R}_1$ along with its corresponding $\mathcal{R}_2$ and parameters $P_1 = (1 - P_2)$, θ_1, θ_2, and β. We check to see if UE$_1$ is able to meet the TMT; however, it should

be noted that increasing β will not improve the throughput of UE$_1$ further irrespective of whether the TMT has been met or not. We thus exit the algorithm in line 70.

Since the purpose of **Algorithm 1** is to provide an efficient alternative to an exhaustive search in terms of complexity, it is important to define a measure of complexity to compare the two. We measure complexity in terms of the sum of the number of times a UE's throughput $\mathcal{R}_i$, $i \in \{1, 2\}$, is calculated. The algorithm iterates over the number of β, power, and transmission rate combinations. As has been mentioned, our algorithm searches in $\theta_{LB} \leq \theta_i \leq \theta_{UB}$ in steps of Δ_θ to make the search finite. Similarly, the algorithm searches

for $0 \leq P_2 \leq 1$ in steps of Δ_P. When $\alpha < 1$, the algorithm searches for $\beta_{\max} \leq \beta \leq 0$ in steps of $\Delta_\beta = \beta_{\max}/10$; for $\alpha = 1$, there is only one choice of β which is 0. For a fair comparison, we use the same search space for the exhaustive search. In particular, there are $\hat{\Delta}_\theta = (\theta_{UB}-\theta_{LB})/\Delta_\theta+1$ choices of θ_i, $i \in \{1,2\}$ and $\hat{\Delta}_P = 1/\Delta_P+1$ choices of P_2. When $\alpha = 1$, there is one choice of β and so $\hat{\Delta}_\beta = 1$. However, when $\alpha < 1$, there are $\hat{\Delta}_\beta = \beta_{\max}/\Delta_\beta + 1$ choices of β. Note that since we have fixed Δ_β to $\beta_{\max}/10$, $\hat{\Delta}_\beta = 11$ when $\alpha < 1$. While **Algorithm 1** does not go through all combinations of these choices, an exhaustive search does. Hence, the complexity of an exhaustive search for this setup is $\hat{\Delta}_\beta \hat{\Delta}_P \hat{\Delta}_\theta^2$.

Remark 5: Each value of α requires an exhaustive search over all combinations of β, $P_1 = (1 - P_2)$, θ_1 and θ_2 to obtain the contour plots against β and P_1 that maximize $\mathcal{R}_{\mathrm{tot}}$. It is thus not possible to obtain these plots from an exhaustive search for many α values. However, we would like to mention that the results obtained from **Algorithm 1** matched those of the exhaustive search for the α values that we did conduct them for. While we can still not guarantee that our algorithm finds the optimum solution for all values of α and the TMT constraint, this highlights the accuracy of the feasible solution found by our algorithm.
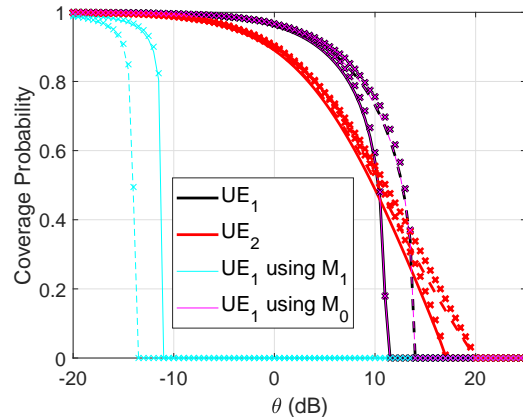
V. RESULTS

We consider BS intensity $\lambda = 10$, noise power $\sigma^2 = -90$ dB and $\eta = 4$. In the results in Section V-A, resource allocation is fixed and unless stated otherwise, we transmit using $P_1 = 1 - P_2 = 1/3$ and use identical transmission rates for clarity of presentation. The SINR thresholds (θ_1 and θ_2) corresponding to the transmission rates are thus represented using θ. The results in Section V-B use resource allocation obtained from **Algorithm 1** for solving the optimization problem $\mathcal{P}1$.
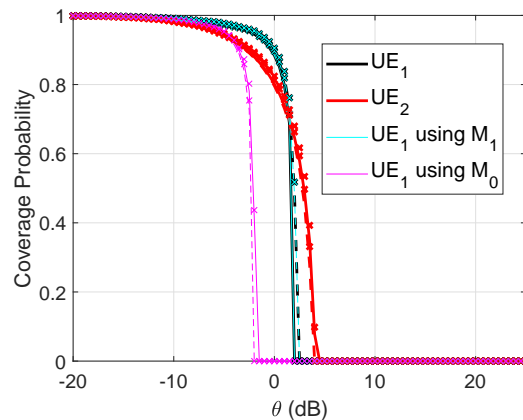
A. Fixed Resource Allocation

Fig. 4 is a plot of coverage probabilities against SINR threshold. Fig. 4a uses $\alpha = 0.25$ while Fig. 4b uses $\alpha = 0.75$; each plots the probabilities for $\beta = 0$ and $\beta = (1 - \alpha)/2$. The figure validates our analysis by using Monte Carlo simulations to show that the approximation in **Theorem 1** is tight. In addition to the coverage probabilities of the two UEs, the figure also plots the coverage probability of UE$_1$ using traditional SIC decoding (cyan curves), i.e., $\bar{M}_1 = M_1$, and when UE$_1$ does not decode the message of UE$_2$ and treats the intracell interference from UE$_2$ as noise (magenta curves), i.e., $\bar{M}_1 = M_0$. Since identical transmission rates are used for both UEs in this figure, the coverage of UE$_1$ for a given α is one or the other; however, if this was not the case, the coverage could have been equal to different decoding techniques at different θ depending on the selected (superior) technique in that case.

We observe that when α is small in Fig. 4a, increasing β from 0 to $(1 - \alpha)/2$ increases the coverage probability as $\mathcal{I}(\alpha,\beta)$ decreases from 0 to $\beta_{\max}/2$ ($= (1-\alpha)/2$) as shown in Fig. 2b. With the larger α used in Fig. 4b, $\mathcal{I}(\alpha,\beta)$ increases from 0 to $\beta_{\max}/2$. Corresponding to this increase in interference, the coverage probability of UE$_2$, which treats the message of UE$_1$ as noise, decreases with β from 0 to

(a) $\alpha = 0.25$

(b) $\alpha = 0.75$

Fig. 4: Coverage probabilities vs. θ. Solid (dashed) lines represent $\beta = 0$ ($\beta = (1 - \alpha)/2$). Markers represent Monte Carlo simulations.

$\beta_{\max}/2$. UE$_1$, on the other hand, decodes the message of UE$_2$ as $\bar{M}_1 = M_1$ in the case of Fig. 4b; thus, as $\mathcal{I}(\alpha,\beta)$ increases from 0 to $\beta_{\max}/2$ decoding the message of UE$_2$ becomes easier for UE$_1$ and the coverage of UE$_1$ improves. Note that the case of UE$_1$ using M_0 follows a similar trend to UE$_2$ as it treats the message of the other UE as noise. It should also be noted that the coverage probability for the UEs given an α is the same when $\beta = 0$ and when $\beta = (1-\alpha)$. This is due to the symmetry of $\mathcal{I}(\alpha,\beta)$ about $\beta_{\max}/2$ which results in identical coverage for β values of the form $(1 - \alpha)x$ and $(1 - \alpha)(1 - x)$, where $x \in [0,1]$, due to identical values of $\mathcal{I}(\alpha,\beta)$ for such values of β. It is important to note that the throughput of the UEs will not be the same for such pairs of β values. This is because, the β value directly impacts bandwidth and therefore throughput; this will be observed in Figs. 9 and 10 where different rates are observed for $\beta = 0$ and $\beta = 1 - \alpha$ which have identical coverage.

Fig. 5 plots the coverage probability of the UEs with increasing $P_1(= 1-P_2)$ using $\theta = 0$ dB and different β values for $\alpha = 0.35$ and $\alpha = 0.75$. As anticipated, the coverage of UE$_2$ decreases with P_1 because of deteriorating SINR for UE$_2$ as P_1 increases (i.e., P_2 decreases). Additionally, the coverage of UE$_2$ decreases with α due to the higher interference encountered as $\mathcal{I}(\alpha,\beta)$ increases with α. In Fig.
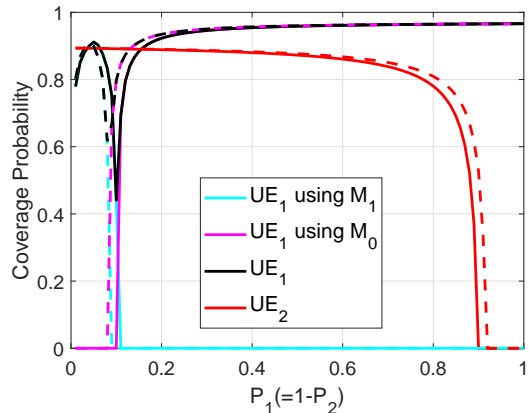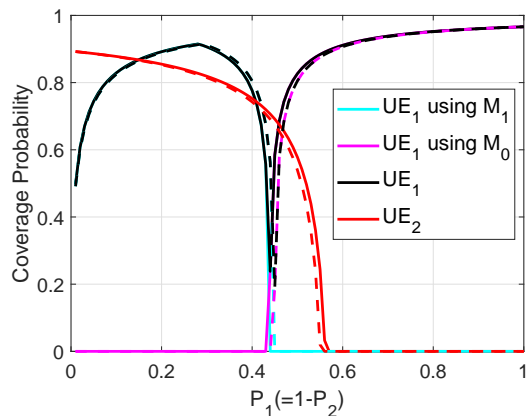
(a) $\alpha = 0.35$

(b) $\alpha = 0.75$

Fig. 5: Coverage probabilities vs. $P_1(= 1 - P_2)$ using $\theta = 0$ dB. Solid (dashed) lines represent $\beta = 0$ ($\beta = (1-\alpha)/2$).

Fig. 6: Coverage probabilities vs. θ using $\beta = (1 - \alpha)/2$. Black (red) lines represent UE_1 (UE_2).

Fig. 7: Coverage probabilities vs. α using $\beta = (1 - \alpha)/2$. Solid lines represent $\theta = -1$ dB, dotted represent $\theta = 0$ dB and dashed represent $\theta = 1$ dB.

5a, we observe that the coverage of UE_2 increases with β, while in Fig. 5b, we observe a slight decrease in coverage with β. This occurs because for lower (higher) values of α, $\mathcal{I}(\alpha, \beta)$ decreases (increases) from $\beta = 0$ to $\beta = (1 - \alpha)/2$, thereby decreasing (increasing) interference. The coverage for UE_1 is more complex as the coverage first increases at low P_1 as the message of UE_2 is easily decoded due to high P_2 and then decreases as increasing P_1 makes decoding the message of UE_2 hard; $\bar{M}_1 = M_1$ in this regime. After this, we observe a sharp increase in the coverage of UE_1 as the message of UE_2 is treated as noise with growing P_1 (and therefore, decreasing P_2) since $\bar{M}_1 = M_0$. Note that in the region where UE_2's message is being decoded by UE_1 (i.e., $\bar{M}_1 = M_1$), when $\alpha = 0.35$, having small $\mathcal{I}(\alpha, \beta)$ is a disadvantage as it reduces the power of the message of UE_2 being decoded; hence, we observe that $\beta = 0$ outperforms $\beta = (1 - \alpha)/2$ in this region. At a higher P_1, where the message of UE_2 is treated as noise, $\beta = (1 - \alpha)/2$, with the smaller $\mathcal{I}(\alpha, \beta)$, outperforms $\beta = 0$ as it experiences lower interference. The opposite trends hold in Fig. 5b with the higher α value of 0.75; this is because here $\beta = (1 - \alpha)/2$ has a higher interference factor $\mathcal{I}(\alpha, \beta)$ than $\beta = 0$.

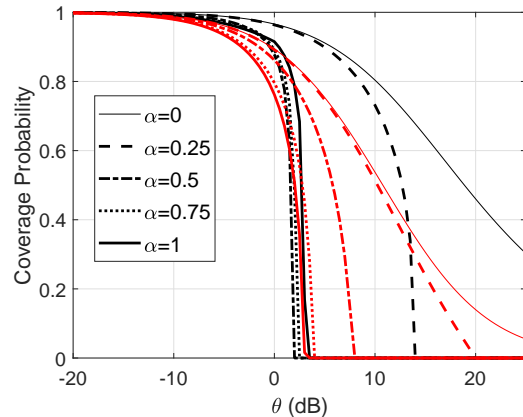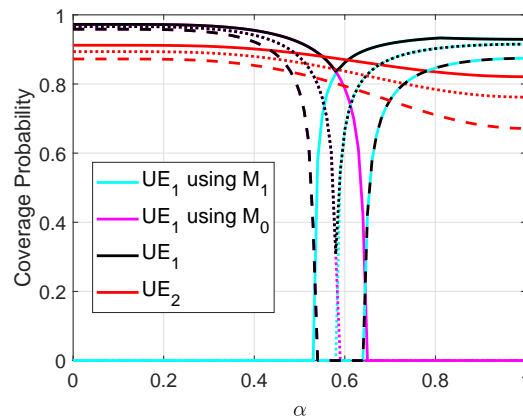Fig. 6 is a plot of the coverage probabilities of the UEs vs. θ using $\beta = (1 - \alpha)/2$ for different values of α. We observe that

the coverage of UE_2 for any θ decreases as α increases. This is anticipated because $\mathcal{I}(\alpha, \beta)$ increases with α; consequently, both intracell and intercell interference increase with α thereby reducing coverage. A different trend is observed for UE_1, on the other hand, where the coverage does not decrease monotonously with α.

Fig. 7, a plot of coverage probability vs. α, explains the above phenomenon better. Different θ values and $\beta = (1-\alpha)/2$ are used. As before, the coverage probability of UE_2 decreases monotonically with α. For UE_1, however, because of the employed modified-SIC decoding, there is a switch between not decoding the message of UE_2 (the curves corresponding to using $\bar{M}_1 = M_0$) and employing traditional SIC decoding (the curves corresponding to $\bar{M}_1 = M_1$). This results in a non-monotonic decrease in the coverage as α increases because the impact of increasing $\mathcal{I}(\alpha, \beta)$ is not as trivial as in the case of UE_2. We also observe that for larger values of θ (see $\theta = 1$ dB), for some choices of power and transmission rate allocation, certain values of α will result in guaranteed outage ($0.54 \leq \alpha \leq 0.64$ for $\theta = 1$ dB). This highlights the importance of careful resource allocation as well as parameter selection in a partial-NOMA setup to avoid guaranteed outage.

It should also be mentioned that with appropriate resource

allocation, partial-NOMA can result in better coverage than traditional NOMA ($\alpha = 1$) for both UEs. Additionally, the curves for UE_1 using M_1 highlight the traditional SIC decoding scheme's inadequacy in the low α regime where, due to small $\mathcal{I}(\alpha, \beta)$, decoding the message of the weak UE becomes the bottleneck for coverage. Similarly, the curves for UE_1 using M_0 show that for higher values of α the intracell interference becomes significant and treating it as noise results in outage.

Fig. 8 is a plot of the cell sum rate against θ using $\beta = (1-\alpha)/2$. The figure highlights that reducing α from the traditional NOMA setup ($\alpha = 1$) increases the tolerance of the system to outage for high transmission rates. While traditional NOMA cannot support UEs that have messages with high transmission rates, instead of opting for such UEs to be designated an entire resource-block for the transmission of their messages, a more efficient approach is to use partial-NOMA where multiple UEs still share a resource-block and can transmit a message with high transmission rate. Essentially, the partial-NOMA setup is less restrictive in terms of the transmission rates that can be supported. We also observe that the peak $\mathcal{R}_{tot}$ first increases from $\alpha = 0$, followed by a decrease and then an increase again to $\alpha = 1$. Additionally, the peaks of the lower α values outperform that of traditional NOMA. This trend again highlights the existence of a range of α values which provide superior performance compared to traditional NOMA in terms of $\mathcal{R}_{tot}$ followed by another range inferior to it.

Fig. 9 is a plot of the cell sum rate against α. Corresponding to the outage regions for UE_1 in Fig. 7, we observe dips in the cell sum rate. It is interesting to observe that θ values that support larger rates overall, such as $\theta = 1$ dB have larger dips than lower θ. This occurs because while the transmission rates being used make both M_1 and M_0 result in superior coverage conditions, since they are identical, they put a larger gap between the two conditions resulting in a larger region of outage for UE_1. This gap, and the consequent dip in rate, reduces as θ increases but the price paid is lower overall rate. Other than the dip caused by the outage region of UE_1, we observe that cell sum rate increases roughly linearly with α. Fig. 10 is plotted to gain better insight of why this occurs

Fig. 10 is a plot of rates with increasing α using $\theta = 0$ dB. For a given α, the throughput of UE_2 (UE_1) decreases (increases) as β increases because its bandwidth decreases (increases). For $\beta = 0$ we observe that UE_2's throughput decreases with α. This is because, it has a fixed bandwidth of 1 in this case and its throughput is only impacted by coverage which decreases as $\mathcal{I}(\alpha, \beta)$ increases with α. For the other two β values, the throughput of UE_2 increases with α as the impact of the increasing bandwidth is greater than the increased interference resulting from higher $\mathcal{I}(\alpha, \beta)$. For the lower β values, we observe that the throughput of UE_1 increases with α (other than the dips occurring from the outage region) as the impact of increasing bandwidth is larger than the increasing interference. When $\beta = 1 - \alpha$, however, the bandwidth is 1 and the throughput decreases with α because of the impact of increased interference occurring from $\mathcal{I}(\alpha, \beta)$. It should be noted that the rate of this decrease is much lower
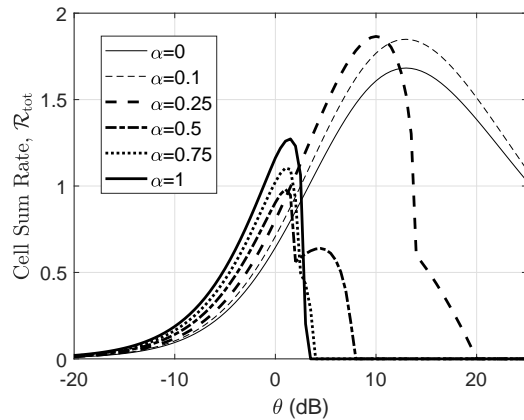

Fig. 8: Cell sum rate vs. θ using $\beta = (1-\alpha)/2$.
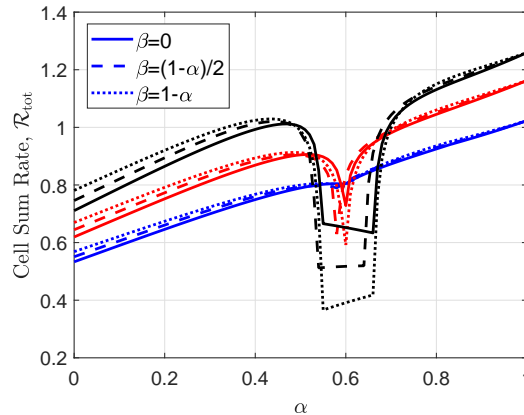

Fig. 9: Cell sum rate vs. α. Blue lines represent $\theta = -1$ dB, red represent $\theta = 0$ dB, and black represent $\theta = 1$ dB.
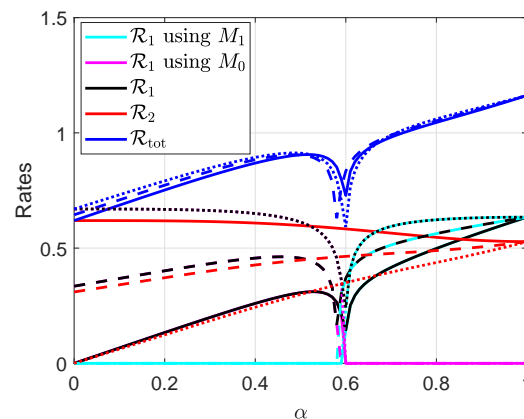

Fig. 10: Rates vs. α using $\theta = 0$ dB. Solid lines represent $\beta = 0$, dashed represent $\beta = (1-\alpha)/2$, and dotted represent $\beta = 1 - \alpha$.

than the rate of increase from bandwidth gains for the other curves which is anticipated. This explains why an optimum α, which is not 1, that maximizes cell sum rate is not observed in Fig. 9 or 10.

B. Resource Allocation Using Algorithm 1

Fig. 11 plots the impact of increasing α on different elements when the cell sum rate is maximized subject to a TMT constraint, i.e., $\mathcal{P}1$. We consider two different values of TMT, $\mathcal{T} = 0.05$ and $\mathcal{T} = 0.25$. Fig. 11a plots the individual UE and cell sum rates. As has been mentioned, **Algorithm 1** is used to obtain the resource allocation; hence $\mathcal{R}_2$ attains the TMT. For a given $\mathcal{T}$, we observe in Fig. 11b that upto $\alpha = 0.3$ for $\mathcal{T} = 0.25$ ($\alpha = 0.35$ for $\mathcal{T} = 0.05$), P_2 slowly increases because of the increasing $\mathcal{I}(\alpha, \beta)$ and consequent intracell interference requiring higher power by UE$_2$ to achieve TMT. Corresponding to this range of α, there is first an increase and then a decrease in $\mathcal{R}_1$ (and therefore $\mathcal{R}_{\text{tot}}$) although P_1 decreases slowly. The initial increase in rate is attributed to the more significant impact of BW$_1$ at first; however, after the local optimum, the impact of lower P_1 and higher $\mathcal{I}(\alpha, \beta)$, contributing to lower power of the message of interest and higher intracell interference, takes over and causes a degradation in rate as α increases.

After $\alpha = 0.3$ for $\mathcal{T} = 0.25$ ($\alpha = 0.35$ for $\mathcal{T} = 0.05$), we observe from Fig. 11b that, there is a switch in the decoding technique from $\bar{M}_1 = M_0$ to $\bar{M}_1 = M_1$, i.e., from UE$_1$ treating the message of UE$_2$ as noise to decoding it. This switch also corresponds to a sudden increase in P_2 leaving behind less power for UE$_1$'s message. However, we still observe an increase in $\mathcal{R}_1$ because decoding and removing UE$_2$'s message before decoding its own improves UE$_1$'s performance. This need for UE$_1$ to decode UE$_2$'s message scaled by $\mathcal{I}(\alpha, \beta)$ is also why there is a spike in P_2 when the decoding technique switches. As α grows, P_2 decreases as $\mathcal{I}(\alpha, \beta)$ grows so it is easier for UE$_1$ to decode the weak UE's message and because lower P_2 is required by UE$_2$ which also has larger bandwidth as the overlap α grows. Hence, as α grows after the switch in $\bar{M}_1$, $\mathcal{R}_{\text{tot}}$ grows with α.

We also observe that lower $\mathcal{T}$ corresponds to higher $\mathcal{R}_{\text{tot}}$ as there are more resources available for UE$_1$ to maximize its throughput with. Additionally, we observe a range of α that outperforms traditional NOMA in terms of $\mathcal{R}_{\text{tot}}$ and thus there exists an optimum $\alpha \neq 1$ that maximizes cell sum rate subject to a TMT constraint. Note that this is in line with partial-NOMA outperforming traditional NOMA in the rate region abstraction as unconstrained cell sum rate maximization is equivalent to $\mathcal{P}1$ with $\mathcal{T} = 0$. It ought to be noted that the values of α (including the optimum) that outperform traditional NOMA in terms of $\mathcal{R}_{\text{tot}}$ occur in the region where $\bar{M}_1 = M_0$ (i.e., the message of the weak UE is treated as noise by the strong UE). This highlights the important role that FSIC plays in the superiority of partial-NOMA.

In Fig. 11c, we plot the number of iterations of β required by the algorithm. Note that the algorithm actually goes through one more iteration than the iteration at which the optimum β is found for $\alpha < 1$ as we terminate the algorithm once $\mathcal{R}_{\text{tot}}$
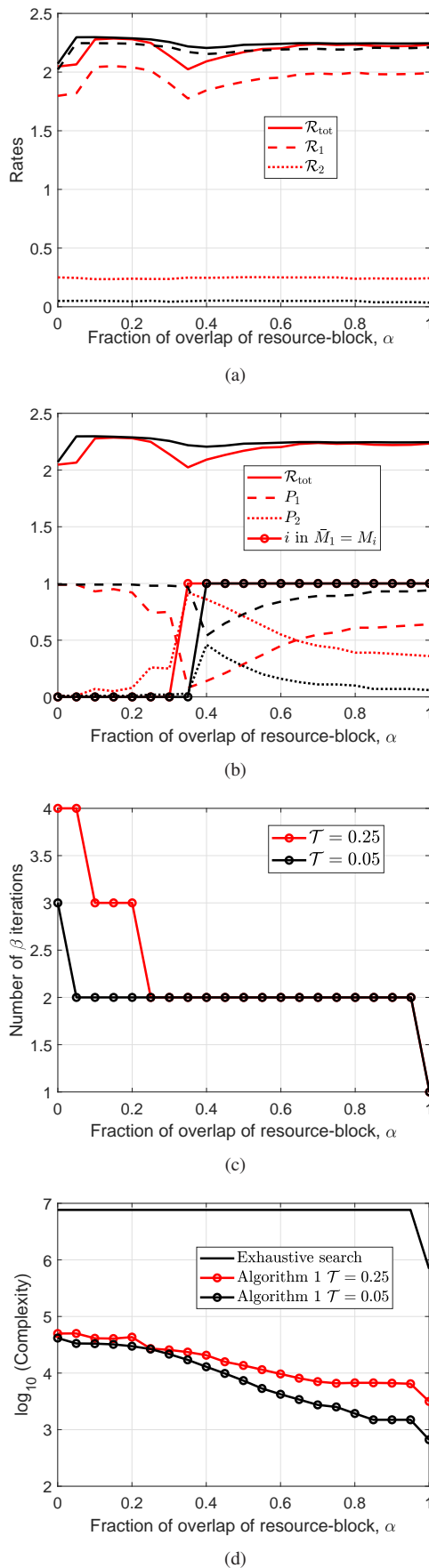
(a)

(b)

(c)

(d)

Fig. 11: Using **Algorithm 1** to solve $\mathcal{P}1$ with increasing α. Red curves are for $\mathcal{T} = 0.25$ and black curves are for $\mathcal{T} = 0.05$.

starts decreasing with decreasing β. For $\alpha = 1$, there is only one possible value of β, and therefore, only one iteration. We observe that a higher $\mathcal{T}$ naturally requires more iterations since a lower BW_2 may not be sufficient to meet the TMT or may be consuming too much P_2, thereby requiring us to increase BW_2 by reducing β. Note that the number of iterations required decreases monotonically with α for a given $\mathcal{T}$. This is because, as α increases, BW_2 increases and so the need to decrease β is less.

As mentioned in Section IV, it is impractical to carry out an exhaustive search for each value of α due to the high complexity involved. However, since we have data from the exhaustive searches in Fig. 3 for α values of $\{0, 0.05, 0.1, 0.2, 0.25, 0.35, 0.5, 0.9, 1\}$, we use these to benchmark the performance of the results in Fig. 11a. For these values of α with both TMT values used in Fig. 11 we find matches with the data from the exhaustive searches. As mentioned in Remark 5, we would like to emphasize that while this is not proof for the optimality of our algorithm, it sheds light on the accuracy of our feasible solution.

In Fig. 11, we search for the transmission rates in -20 dB $\leq \theta_i \leq 21$ dB and use step size $\Delta_\theta = 0.5$. As a result, there are $\hat{\Delta}_\theta = 83$ choices of θ_i, $i \in \{1, 2\}$. For P_2, we use step size $\Delta_P = 0.01$; hence, there are $\hat{\Delta}_P = 101$ choices of P_2. Δ_β is already defined in Section IV-C so that there are $\hat{\Delta}_\beta = 11$ choices of β when $\alpha < 1$ and there is only $\hat{\Delta}_\beta = 1$ choice of β when $\alpha = 1$. While we do not carry out an exhaustive search in Fig. 11, we still compare the complexity of our algorithm with that of an exhaustive search as it is constant for the latter. As has been mentioned, for a fair comparison, the same search space is considered for the exhaustive search.

Fig. 11d shows that the proposed **Algorithm 1** requires significantly lower complexity than an exhaustive search. Additionally, the complexity of **Algorithm 1** increases as $\mathcal{T}$ increases due to the larger number of iterations of both P_2 and β required to achieve TMT and find the optimum. We also observe that overall as α increases, the complexity of **Algorithm 1** decreases[6]. This is in line with the fact that a higher value of α requires a fewer number of iterations of β for the algorithm to find the optimum solution. It should also be noted that for both the exhaustive search and **Algorithm 1**, there is a decrease in complexity at $\alpha = 1$ because there is only one value of β possible making the search space smaller.

C. Summary of Main Results

We summarize the main results as follows:

- The coverage of UE$_2$ decreases monotonically with the overlap α. The coverage of UE$_1$, on the other hand, does not. This is because the impact of increasing $\mathcal{I}(\alpha, \beta)$ is not as trivial due to FSIC decoding.

---

[6]Our complexity curves for **Algorithm 1** are not very smooth as the grid for θ_i is not very fine. This sometimes results in a longer search for the resources that allow UE$_2$ to attain TMT and therefore the decrease in complexity with α is not monotonic (by small amounts). A finer θ_i grid would result in smoother curves but the price paid would be much higher complexity. As the difference in resource allocation and performance would be marginal, we do not do this to avoid longer computation times.

- Some choices of resource allocation will result in guaranteed outage for certain α, emphasizing the importance of careful resource allocation.
- With appropriate resource allocation, partial-NOMA results in better coverage than traditional NOMA for both UEs.
- Traditional SIC is inadequate in the low α regime, while treating the message of the weak UE as noise at the strong UE is inadequate when α is higher.
- Reducing the overlap α allows the system to support UEs with higher transmission rate requirements. This allows partial-NOMA to serve UEs that traditional NOMA cannot, thereby preventing inefficient spectrum reuse.
- As anticipated, for a given power and transmission rate allocation, the impact of increased interference with α is lower than that of increased bandwidth.
- Using **Algorithm 1**, when $\bar{M}_1 = M_0$, we observe a local optimum for $\mathcal{R}_1$ although P_1 is decreasing in this range because of the trade off between increasing bandwidth and decreasing SINR (due to lower signal power and higher intracell interference).
- Using **Algorithm 1**, after the switch from $\bar{M}_1 = M_0$ to $\bar{M}_1 = M_1$, $\mathcal{R}_1$ increases with α.
- We observe partial-NOMA to outperform traditional NOMA in terms of $\mathcal{R}_{\text{tot}}$ both in the rate region and using **Algorithm 1**. This occurs in the low α regime where $\bar{M}_1 = M_0$ highlighting the important role that FSIC plays in the superiority of partial-NOMA.
- An optimum $\alpha < 1$ exists given a TMT constraint that maximizes $\mathcal{R}_{\text{tot}}$.
- **Algorithm 1** is shown to have much lower complexity than an exhaustive search. Its complexity grows with TMT and decreases with α.

VI. CONCLUSION

Partial-NOMA is proposed as a technique to strike a balance between the high interference associated with NOMA resulting in low coverage and no spectrum reuse in OMA resulting in low rates. A large downlink two-user network employing partial-NOMA is studied. The nature of the partial overlap allows us to employ receive-filtering to further suppress the interference in a partial-NOMA setup. The received filtering not only suppresses intracell interference but also results in a suppression of intercell interference allowing our setup to have lower intercell interference than both NOMA and OMA. A technique called FSIC decoding is proposed for decoding the partial-NOMA setup. An abstraction of the rate region is studied and compared to that of NOMA and OMA. It is observed that for some values of the overlap α, partial-NOMA can outperform NOMA. It is also shown that partial-NOMA allows more flexibility in terms of the achievable individual UE throughput. A problem of maximizing cell sum rate subject to a TMT constraint is formulated. Since the problem is non-convex, the only known solution requires an exhaustive search. An efficient algorithm that finds a feasible solution to the problem is proposed. The complexity of the algorithm is shown to be much lower than an exhaustive search. It is shown

that the partial-NOMA setup outperforms NOMA in terms of cell sum rate for a range of values of α. Additionally, in this range, there exists an optimum value of α that maximizes the cell sum rate for a given TMT constraint. Furthermore, it is observed that the range of α which results in superior cell sum rate to traditional NOMA corresponds to $\bar{M}_1 = M_0$, highlighting the role of FSIC in the superiority of partial-NOMA. It is also shown that while NOMA cannot support UEs with high transmission rate requirements, partial-NOMA can. Instead of allocating an entire resource-block to such UEs via OMA, these UEs ought to be served via partial-NOMA to reuse the spectrum efficiently. The work in this paper studies a two-user setup. An important direction for future work is to study partial-NOMA in an N-user setup, where N is general. In this paper, we have considered downlink transmissions; investigating partial-NOMA in the uplink is also an important and interesting direction.

REFERENCES

[1] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and A. Nallanathan, "Non-orthogonal multiple access in massive MIMO aided heterogeneous networks," in *Proc. of IEEE Global Communications Conference (GLOBE-COM16)*, Dec. 2016.

[2] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Comm. Letters*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.

[3] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Select. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[4] H. Tabassum, E. Hossain, and M. J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access (NOMA) in large-scale cellular networks using poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[5] K. S. Ali, H. ElSawy, A. Chaaban, M. Haenggi, and M. Alouini, "Analyzing non-orthogonal multiple access (NOMA) in downlink Poisson cellular networks," in *Proc. of IEEE International Conference on Communications (ICC18)*, May 2018, pp. 1–6.

[6] K. S. Ali, H. E. Sawy, and M. Alouini, "Meta distribution of downlink non-orthogonal multiple access (NOMA) in Poisson networks," *IEEE Wireless Comm. Letters*, vol. 8, no. 2, pp. 572–575, Apr. 2019.

[7] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Vehicular Tech.*, vol. 65, no. 12, pp. 10 152–10 157, Dec. 2016.

[8] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Comm. Letters*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[9] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Proc. Letters*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[10] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Proc. Letters*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[11] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Selec. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[12] K. S. Ali, M. Haenggi, H. E. Sawy, A. Chaaban, and M. Alouini, "Downlink non-orthogonal multiple access (NOMA) in Poisson networks," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1613–1628, Feb. 2019.

[13] Z. Zhang, H. Sun, R. Q. Hu, and Y. Qian, "Stochastic geometry based performance study on 5G non-orthogonal multiple access scheme," in *Proc. of IEEE Global Communications Conference (GLOBECOM16)*, Dec. 2016, pp. 1–6.

[14] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. of IEEE 77th Vehicular Technology Conference (VTC13)*, Jun. 2013, pp. 1–5.

[15] C. L. Wang, J. Y. Chen, and Y. J. Chen, "Power allocation for a downlink non-orthogonal multiple access system," *IEEE Wireless Comm. Letters*, vol. 5, no. 5, pp. 532–535, Oct. 2016.

[16] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Selec. Areas Commun.*, vol. 35, no. 12, pp. 2771–2784, Dec. 2017.

[17] B. Kim, Y. Park, and D. Hong, "Partial non-orthogonal multiple access (P-NOMA)," *IEEE Wireless Comm. Letters*, pp. 1–1, 2019.

[18] K. S. Ali, H. Elsawy, A. Chaaban, and M. S. Alouini, "Non-orthogonal multiple access for large-scale 5G networks: Interference aware design," *IEEE Access*, vol. 5, pp. 21 204–21 216, 2017.

[19] B. Blaszczyszyn, M. Haenggi, P. Keeler, and S. Mukherjee, *Stochastic Geometry Analysis of Cellular Networks*. Cambridge University Press, 2018.

[20] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[21] H. ElSawy, A. Sultan-Salem, M. S. Alouini, and M. Z. Win, "Modeling and analysis of cellular networks using stochastic geometry: A tutorial," *IEEE Commun. Surveys and Tutorials*, vol. 19, no. 1, pp. 167–203, Firstquarter 2017.

[22] W. Lu and M. D. Renzo, "Stochastic geometry modeling of cellular networks: Analysis, simulation and experimental validation," *CoRR*, vol. abs/1506.03857, 2015. [Online]. Available: http://arxiv.org/abs/1506.03857

[23] M. Salehi, H. Tabassum, and E. Hossain, "Accuracy of distance-based ranking of users in the analysis of noma systems," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5069–5083, Jul. 2019.

[24] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[25] P. D. Mankar and H. S. Dhillon, "Downlink analysis of noma-enabled cellular networks with 3gpp-inspired user ranking," 2019.

[26] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Vehicular Tech.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[27] K. S. Ali, M. Alouini, E. Hossain, and M. J. Hossain, "On clustering and channel disparity in non-orthogonal multiple access (NOMA)," *CoRR*, vol. abs/1905.02337, 2019. [Online]. Available: http://arxiv.org/abs/1905.02337

[28] A. H. Sakr and E. Hossain, "Location-aware cross-tier coordinated multipoint transmission in two-tier cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6311–6325, Nov. 2014.

[29] A. H. Sakr, H. ElSawy, and E. Hossain, "Location-aware coordinated multipoint transmission in OFDMA networks," in *Proc. of IEEE International Conference on Communications (ICC14)*, June 2014, pp. 5166–5171.

[30] H. A. David, *Order statistics*. NJ: John Wiley, 1970.

[31] A. AlAmmouri, H. ElSawy, O. Amin, and M. Alouini, "In-band α-duplex scheme for cellular networks: A stochastic geometry approach," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6797–6812, Oct. 2016.

[32] I. Randrianantenaina, H. Dahrouj, H. Elsawy, and M. Alouini, "Interference management in full-duplex cellular networks with partial spectrum overlap," *IEEE Access*, vol. 5, pp. 7567–7583, 2017.