IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 June 2024, accepted 24 June 2024, date of publication 2 July 2024, date of current version 10 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3421906

RESEARCH ARTICLE

Coverage, Throughput, and Energy Efficiency Enhancement in Beamspace Massive MIMO System Using Rate-Splitting and Orthogonal Precoding

DAVID ALIMO[1], (Student Member, IEEE), MASANORI HAMAMURA[1,2], (Member, IEEE), AND SAIFUR RAHMAN SABUJ[3], (Senior Member, IEEE)

[1]Graduate School of Engineering, Kochi University of Technology, Tosayamada, Kami, Kochi 782-8502, Japan
[2]School of Information, Kochi University of Technology, Tosayamada, Kami, Kochi 782-8502, Japan
[3]Department of Electrical and Electronic Engineering, Brac University, Dhaka 1212, Bangladesh

Corresponding authors: David Alimo (david.lagu2012@gmail.com) and Masanori Hamamura (hamamura.masanori@kochi-tech.ac.jp)

ABSTRACT Millimeter-wave beamspace massive multiple-input multiple-output system with a lens antenna array can minimize transceiver hardware complexity without compromising performance. However, the number of supported portable user terminals cannot exceed the number of radio frequency blocks accessible at the same time, frequency, and coding resources. In this paper, we propose the integration of rate-splitting multiple access and orthogonal random precoding into the downlink of beamspace massive multiple-input multiple-output system to support a larger number of portable user terminals than the number of available radio frequency blocks while minimizing both inter- and intra-beam interferences and extending the cell coverage percentage. Then, we formulate an optimization problem to optimize the system's overall throughput while keeping the minimum needed throughput and power budget in consideration. The nonconvex optimization issue is then approximated into a convex optimization problem using the successive convex approximation approach. Following that, we offer an alternating method to solve the approximate optimization issue and select an optimal solution. Furthermore, we deduce an analytic expression for the downlink cell coverage percentage and evaluate the effectiveness of the suggested method in terms of total throughput, energy efficiency, and cell coverage percentage. Finally, we compare the proposed method with benchmark techniques for perfect and imperfect channel state information, and numerical results confirm the superior performance of the proposed method over benchmark techniques in terms of sum throughput, energy efficiency, and cell coverage percentage.

INDEX TERMS Beamspace, cell coverage percentage, energy efficiency, lens antenna array, massive multiple-input multiple-output, orthogonal random precoding, rate-splitting multiple access, throughput.

I. INTRODUCTION

Millimeter-Wave (mmWave) communications have been a notable technology in the Fifth Generation (5G) of mobile communication networks and will continue to be a notable technology in the subsequent generations of wireless networks because of its capability to enable Gbps throughput (data rates) [1], [2], [3]. This is due to the fact that there are a large number of unused frequencies in the mmWave band, and increasing the bandwidth is a successful strategy to increase system capacity [4]. Nevertheless, owing to the extremely high carrier frequencies in the mmWave spectrum, there is a significant path loss compared with lower frequency bands [3]. Fortunately, a large number of antennas can be squeezed into a very small space owing to the small wavelength of mmWave signals that can be utilized to attain

The associate editor coordinating the review of this manuscript and approving it for publication was Ding Xu.

beamforming enhancements to offset the significant path loss at mmWave frequencies [1], [2], [3], [5], [6], [7], [8]. Despite the advantages brought about by the large antenna arrays, full digital beamforming introduces high energy consumption and computation complexity owing to the fact that each antenna needs its own Radio Frequency (RF) block that includes a Power Amplifier (PA), Low-Noise Amplifier (LNA), converter, mixer, filter, oscillator, and others in a traditional massive Multiple-Input Multiple-Output (mMIMO) communication system [1], [2], [3], [4], [7], [9], [10], [11], [12]. The use of a lens antenna array based on the beamspace mMIMO communication system can simplify the complexity of mmWave communication systems [3]. By changing the direction of the electromagnetic rays, one can realize a lens antenna array with an angle-dependent energy focus [3]. The number of required RF blocks can be decreased by rigorously choosing the dominant beams depending on the sparse beamspace channel [2]. However, since each RF block can only attend to one Portable User Terminal (PUT) at the same time and frequency, the maximum number of PUTs supported can be limited to the number of available RF blocks [2]. To overcome this limitation, Non-Orthogonal Multiple Access (NOMA) was proposed [2]. NOMA can serve multiple PUTs using the same time and frequency resources simultaneously by multiplexing PUTs in the power domain and utilizing Successive Interference Cancellation (SIC) to mitigate inter-PUT interference [2], [13], [14], [15]. Nevertheless, when using NOMA, the PUT with the lowest allocated power will have to decode all the signals intended for other PUTs, which increases computational complexity [13], [16]. Therefore, Rate-Splitting Multiple Access (RSMA) can be used in the downlink of a beamspace mMIMO system to reduce the complexity brought about by employing NOMA [17]. RSMA is a type of power-domain non-orthogonal multiple access technique that is based on the multi-antenna rate-splitting technique, where each message is divided into a common segment and a private segment, and the common segments of all the PUTs are encoded into a common message and the private segment of each PUT is encoded into a private message. Then the common message is overlaid on top of each PUT's private message [18], [19], [20], [21], [22], [23], [24], [25].

Moreover, with mmWave mMIMO systems, the coverage area can be increased by using precoding techniques in the downlink [26]. Linear precoding techniques such as Zero-Forcing (ZF) and Minimum Mean Square Error (MMSE) can achieve near optimal performance with perfect Channel State Information at the Transmitter (CSIT). However, these techniques requires channel inversion which introduces computation complexity into the system design. In [27], a method for constructing M random beams to transmit information to the PUTs that have the highest M Signal-to-Interference plus Noise Ratios (SINRs) was proposed. It was revealed that the throughput scales as $N \log K$, which is comparable to perfect Channel State Information (CSI) utilizing dirty paper coding when the number of PUTs (K) increases but the number of beams N stays the same [27]. Low-complexity precoding techniques are therefore essential for designing future wireless networks in order to enhance the system throughput and coverage area of mMIMO systems in the downlink scenario.

A. RELATED WORKS

Several studies have already explored RSMA-related issues, including [5], [16], [18], [19], [22], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42]. In [5], the asymptotic sum throughput of Rate-Splitting (RS) and Hierarchical Rate-Splitting (HRS) was analyzed and the common messages precoder was optimized. Clerckx et al. [16] compared NOMA with traditional Multiple-User Linear Precoding (MU-LP) and RSMA schemes and demonstrated that despite an increased receiver complexity, NOMA suffers from a marked multiplexing gain loss because of the inadequate use of SIC receivers. In contrast to more conventional techniques such as single-PUT mode Time Division Multiple Access (TDMA) and multiple-PUT mode Zero-Forcing Beamforming (ZFBF), Clerckx et al. [18], [43] showed that RS can significantly increase mMIMO wireless networks' spectral and energy efficiencies, dependability, and CSI feedback overhead reduction. To increase the ergodic sum throughput, the investigations in [19] provided an RS strategy that was combined with a linear precoder design and optimization methods. In another study [44], Joudeh et al. proposed an RS multigroup multicast beamforming strategy where an alternating optimization algorithm depending on the Weighted Minimum Mean Square Error (WMMSE) technique was utilized to obtain RS precoders.

In [29], Yang et al. formulated a sum throughput maximization problem for wireless networks that used downlink RSMA, whose goal was to maximize the sum throughput for all PUTs. To resolve the nonconvex maximization problem, they first determined the best power for transmitting the private message in closed form for a specific rate allocation and a common message, and they then derived the ideal rate allocation in accordance with the highest private message transmit power under a constant common message transmit power. Similarly, Li et al. investigated resource allocation in multi-carrier RSMA network [28]. In [30], a cell-free mMIMO-based joint multiservice transmission strategy was suggested, with which access point selection is performed and broadcast, multicast, and unicast messages are created at each access point in accordance with the RSMA method. To counteract the impacts of quantization noise innate in low-resolution Analog-to-Digital Converters (ADCs) and Digital-to-Analog Converters (DACs), Ahiadormey and Choi [45] proposed an RSMA method. They then demonstrated that as the ADC/DAC resolution increased, their method outperformed non-RSMA methods in terms of spectral efficiency at a high SNR. When all of the PUTs used the same pilot sequence for channel estimation as described in [31], RS was investigated in the downlink of a single-cell mMIMO

system, and Thomas et al. found that RS outperformed traditional mMIMO with maximum ratio precoding in terms of spectral efficiency. In a similar study [32], it has been demonstrated that RS was robust for multiple-PUT MIMO in the presence of phase and magnified thermal noises, but the gains decreased under both ideal and practical conditions as the number of PUTs increased. Papazafeiropoulos and Ratnarajah [33] considered a multipair decode-and-forward full-duplex relay channel with a relay station that was equipped with many antennas, and they showed how RS enabled the extension of the range of Self-Interference (SI), over which full-duplex relays outperformed half-duplex relays. Dizdar et al. [34] provided a general overview of RSMA and highlighted its potential to meet 6G requirements. Similarly, Mishra et al. [22] utilized RSMA as a pilot contamination reduction method in the downlink of machine-type communications with random access and revealed that RSMA successfully reduces the impact of pilot contamination. In a related work, pilot-sharing PUTs were modeled as an interference channel [35], and the effectiveness of the technique that decoded the interference partly using RS and the technique that decoded the interference completely was analyzed. The results revealed that RS had higher spectral efficiency. In another study, low-complexity RSMA algorithms using hierarchical streams was proposed in [38] and analyzed under an interference nulling scenario. Arora et al. [36] proposed quadrature-RSMA that eliminates the requirement of performing SIC prior to decoding the private messages. Moreover, RSMA can achieve higher achievable rate than the conventional wireless networks in vortex wave communications [37]. Tong et al. [40] proposed two RS schemes with different SIC layers and analyzed their model in terms of outage probability in the presence of untrusted near user. In another study, Lee et al. [42] considered a sum secrecy spectral efficiency problem in a downlink RSMA system with multiple antennas. Recently, RSMA has gain attention in Unmanned Aerial Vehicle (UAV)-assisted communication. Singh et al. [41] investigated the effectiveness of RSMA in a multiuser downlink wireless network consisting of UAV-assisted base station that serves multiple ground users using infinite block length and finite block length transmission schemes under imperfect SIC and CSI. Similarly, the integration of RSMA and Reconfigurable Intelligent Surfaces (RIS) into wireless networks has been shown to maximize the achievable rate and energy efficiency [39]. Therefore, RSMA is a promising technique for future wireless networks owing to its robust performance in many scenarios.

The coverage extension issues in wireless communication systems have been investigated in a number of studies such as those in [26], [46], [47] and [48]. In the studies in [26], and [46], it was investigated how Orthogonal Random Precoding (ORP) could be used to increase coverage in the downlink of mMIMO systems. The locations of Next-generation Node Bs (gNBs) and PUTs were modeled using Poisson point processes in [47], and the coverage performance of the multiple-PUT MIMO downlink cellular network was analyzed. The effectiveness of location-aware rank transmission in MIMO cellular networks was examined by Lone et al. [48]. In [26] and [46], the ORP method was not investigated in conjunction with RSMA in a hybrid mmWave mMIMO architecture.

B. CONTRIBUTIONS AND ORGANIZATION

In [5], [16], [18], [19], [22], [29], [30], [31], [32], [33], [34], [35], [38], [40], [41], and [42], it has been shown that RSMA can improve the sum throughput performance in mMIMO systems. On the other hand, it has been revealed that by utilizing ORP, the cell coverage can be extended for the cell-edge PUTs in the downlink of mMIMO cellular systems [26], [46], [47], [48]. In addition, ORP is a low-complexity precoding technique, which makes it preferable over conventional linear precoding methods that rely on channel inversion [46]. As such, we propose the use of ORP and RSMA to extend the coverage, sum throughput, and energy efficiency in the downlink of beamspace mmWave mMIMO systems. In an ORP method, a gNB transmits training signals and receives a real number from each PUT indicating the best SINR and its index. Specifically, each PUT feeds back its largest SINR value and the index of the orthonormal precoding vector corresponding to that SINR to the gNB through a feedback channel. The key contributions of this study are outlined as follows:

- We propose the adoption of ORP and one-layer[1] RSMA techniques to enhance the sum throughput, energy efficiency, and cell coverage percentage performance of beamspace mmWave mMIMO systems in the downlink with reduced complexity. We utilize ORP precoding to reduce inter-beam interference, while RSMA suppresses intra-beam interference by partially decoding interference and partially treating interference as noise [20].
- Using the proposed system model, we use SIC and throughput constraints to model the sum throughput maximization problem with perfect CSI. The nonconvex optimization problem is then solved using a Successive Convex Approximation (SCA)-based technique to produce a locally optimal result. Furthermore, we deduce the analytical expressions for the downlink cell coverage percentage of the beamspace mmWave mMIMO system that utilizes ORP and RSMA.
- The performance of the proposed ORP-RSMA method is evaluated through simulations. The convergence of the iterative optimization algorithm for joint power allocation and rate-splitting is validated. Moreover, numerical findings show that the proposed ORP-RSMA method can surpass ZF Space Division Multiple Access (ZF-SDMA), ZF NOMA (ZF-NOMA) [2], and ZF Orthogonal Multiple Access (ZF-OMA) [14] in terms of

[1] One-layer RSMA refers to a system where each PUT performs SIC once to decode the common message [18]. Without loss of generality, we use RSMA to refer to one-layer RSMA in the remainder of this study.

cell coverage percentage, sum throughput, and energy efficiency. Furthermore, we compare the proposed ORP-RSMA method for perfect and imperfect CSI scenarios.

The remaining sections of this study are structured as follows. In section II, we provide the system model, which includes the path loss, channel model, signal model, ORP method, and RSMA method. In section III, we present the optimization problem, solution, and convergence analysis. In section IV, we outline the performance evaluation and comparison with benchmark schemes. Finally, in section V, we provide the conclusion.

C. NOTATION

Boldface uppercase characters ($\mathbf{A}$) denote matrices, boldface lowercase characters ($\mathbf{a}$) denote vectors, and standard characters (a) denote scalars. The superscript $(\cdot)^H$ denotes the conjugate transpose operator, $\mathbf{I}_K$ is a $K \times K$ identity matrix, $\text{diag}(a_1, a_2, \ldots, a_K)$ is a $K \times K$ diagonal matrix that has diagonal elements of $a_1, a_1, \ldots, a_K$, and $\mathbf{A}(i,:)_{i\in\mathcal{B}}$ is the submatrix of $\mathbf{A}$ that consists of the ith row of $\mathbf{A}$ for all $i \in \mathcal{B}$. $|a|$ denotes the modulus of scalar a, $\|\mathbf{a}\|_2$ denotes the ℓ_2-norm of vector $\mathbf{a}$, and $\mathbb{E}(\cdot)$ is the expectation operator. The notation $\mathcal{CN}(\mathbf{u}, \mathbf{V})$ denotes the complex Gaussian distribution with the mean vector $\mathbf{u}$ and the covariance matrix $\mathbf{V}$.

II. SYSTEM MODEL

A. PATH LOSS MODEL

In this work, we use the close-in free space reference distance model for its simplicity and accuracy across several propagation environments and frequency bands [46]. Therefore, the average path loss due to large-scale and shadow fading of a PUT at a distance of r meters from gNB is given as

$$P_L(r)\,[\text{dB}] = P_L(r_0)\,[\text{dB}] + 10\bar{n}\log_{10}\left(\frac{r}{r_0}\right) + X_\xi, \quad (1)$$

where $P_L(r_0)\,[\text{dB}] = 20\log_{10}\left(\frac{4\pi r_0}{\lambda}\right)$ is the close-in free space path loss, with $r_0 = 1$ m, $\bar{n}$ is the average path loss exponent depending on the propagation environment, X_ξ is a Gaussian random variable with zero mean and a standard deviation ξ accounting for the shadowing effect, and $r \geq r_0$ [46], [49].

B. CHANNEL MODEL

We consider a single cell consisting of a gNB deployed with N antennas and N_{RF} RF blocks [1], [2], [7] as shown in Fig. 1 where the gNB is able to obtain the channel information perfectly from all PUTs. The gNB supports K PUTs concurrently, and each PUT has a single antenna [1], [2], [7]. We implemented beamspace mMIMO at the gNB because it has the potential to decrease the energy consumption and hardware complexity of the mmWave mMIMO system [1], [2], [7]. The spatial channel is transformed into a sparse beamspace channel at the gNB by the lens antenna array [1], [2], [7]. As a result, only few beams are required to serve PUTs without clearly degrading performance, which reduces the quantity of RF blocks needed [1], [2], [7].

The lens antenna array can be expressed mathematically by an $N \times N$ Discrete Fourier Transform (DFT) matrix $\mathbf{F}$ as [1], [2], [7], [50]

$$\mathbf{F} = \left[\mathbf{a}\left(\tilde{\theta}_1\right), \mathbf{a}\left(\tilde{\theta}_2\right), \ldots, \mathbf{a}\left(\tilde{\theta}_N\right)\right]^H, \quad (2)$$

where $\tilde{\theta}_n = \frac{1}{N}\left(n - \frac{(N-1)}{2}\right)$ for $n = 1, 2, \ldots, N$ are the predetermined spatial directions [1], [2], [7], [50]. For a conventional Uniform Linear Array (ULA), it is possible to express the steering vector as [1], [2], [7], [50]

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N}}\left[e^{-j\,2\pi\theta q}\right]_{q\in\jmath(N)}, \quad (3)$$

where $\jmath(N) = \{n - (N-1)/2, n = 0, 1, 2, \ldots, N-1\}$ is a symmetric collection of indices that is centered at the origin [1], [2], [7], [50]. The spatial direction of the channel is given by $\theta = \frac{d\sin\phi}{\lambda}$, where ϕ is the corresponding path's actual direction such that $-\frac{\pi}{2} \leq \phi \leq \frac{\pi}{2}$, $d = \lambda/2$ stands for the spacing between the antenna elements, and λ is the signal wavelength [1], [2], [7], [50].

The most commonly used Saleh-Valenzuela channel model for mmWave communications is considered in this study, and it is written as [1], [2], [7], [50]

$$\mathbf{h}_k = \sum_{l=0}^{N_p} \beta_k^{(l)} \mathbf{a}\left(\theta_k^{(l)}\right), \quad (4)$$

where $\beta_k^{(l)}$ and $\mathbf{a}\left(\theta_k^{(l)}\right)$ are the complex gain and steering vector of the lth multipath component, respectively [1], [2], [7], [50]. Furthermore, the Line-Of-Sight (LOS) path is denoted by $\beta_k^{(0)}\mathbf{a}\left(\theta_k^{(0)}\right)$, and the Non-Line-Of-Sight (NLOS) paths are denoted by $\beta_k^{(l)}\mathbf{a}\left(\theta_k^{(l)}\right)$, with $l = 1, 2, \ldots, N_p$ [1], [2], [7], [50]. Given the spatial direction θ of the channel, (3) can be used to determine the steering vectors for the kth PUT's LOS path and NLOS paths.

By using the DFT matrix $\mathbf{F}$, we can convert the spatial channel matrix $\mathbf{H}$ into the beamspace channel matrix $\hat{\mathbf{H}}$ as

$$\hat{\mathbf{H}} = \left[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \ldots, \hat{\mathbf{h}}_K\right] = [\mathbf{F}\mathbf{h}_1, \mathbf{F}\mathbf{h}_2, \ldots, \mathbf{F}\mathbf{h}_K], \quad (5)$$

where $\hat{\mathbf{h}}_k = \mathbf{F}\mathbf{h}_k$ is the beamspace channel vector between the kth PUT and the gNB [1], [2], [7], [50]. Each row of the beamspace channel matrix $\hat{\mathbf{H}}$ in (5) represents the gain of one beam, and all N rows represent the gains of N orthogonal beams with predetermined spatial directions $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_N$ [1], [2]. Moreover, the quantity of NLOS paths N_p in the mmWave channel is typically less than the quantity N of gNB antennas [2]. As a result, a relatively small number of dominant elements exist in the beamspace channel vector $\hat{\mathbf{h}}_k$ of the kth PUT [2]. Taking advantage of this sparse property, we choose a very small portion of the beams for all PUT accommodations using the conventional maximum-magnitude-based beam selection method without

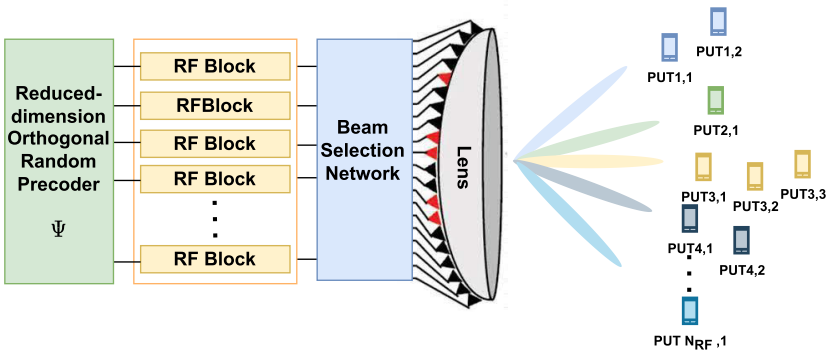

FIGURE 1. System model of downlink beamspace mMIMO system with RSMA and ORP.

compromising the system performance [2]. Precisely, in the beamspace channel $\hat{\mathbf{h}}_k$ between the gNB and the kth PUT, the components are sorted in descending order [7]. Furthermore, for each PUT, the beam chosen corresponds to the channel gain with the largest magnitude [7]. Because one RF block typically yields a single beam, the above-mentioned beam selection technique can significantly lower the number of RF blocks, thereby increasing energy efficiency and cost-effectiveness [2].

C. SIGNAL MODEL

In accordance with the beam selection approach, the signal received at the PUTs can be stated as

$$\hat{\mathbf{y}} = \sqrt{\frac{G_m}{P_L(r)}}\hat{\mathbf{H}}_r^H \boldsymbol{\Psi}_r \mathbf{P}\mathbf{s} + \mathbf{n}, \tag{6}$$

where $G_m = G_{TX}G_{RX}$ is the proper beam alignment gain, such that G_{TX} and G_{RX} are the antenna gains at the gNB and PUT, respectively [26]. Furthermore, $P_L(r)$ is the average path loss at a distance of r m, $\hat{\mathbf{H}}_r = \hat{\mathbf{H}}(b,:)_{b\in\mathcal{B}}$ is the beamspace channel matrix with the reduced dimensions of size $|\mathcal{B}| \times K$, and $\mathcal{B}$ is the collection of indices of the selected beams; $\boldsymbol{\Psi}_r$ is the digital precoding matrix with a reduced dimension of size $|\mathcal{B}| \times K$ whose row dimension is equal to $|\mathcal{B}| = N_{RF} < N$; $\mathbf{P} = \text{diag}\left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K}\right)$ is the transmit power allocated for all K PUTs that satisfies $\sum_{k=1}^{K} p_k \leq P_T$; where P_T is the total transmit power at the gNB; $\mathbf{s}$ is the transmitted signal vector of size $K \times 1$ with normalized power such that $\mathbb{E}\left(\mathbf{s}\mathbf{s}^H\right) = \mathbf{I}_K$; and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_K)$ is the additive thermal noise vector [2], [51].

However, reducing the number of RF blocks, on the other hand, creates an issue of limited connections [2]. Therefore, in conventional beamspace MIMO, the Degree-of-Freedom (DoF) is limited to the number of available RF blocks [2]. Consequently, the number of PUTs served concurrently can be confined to N_{RF}; nevertheless, in available beamspace mMIMO systems, a single beam can only accommodate one PUT [2]. However, it is very likely that several PUTs will share the same beam as their strongest beam [9]. Subsequently, in this study, we consider PUTs sharing the same beam as Interfering PUTs (I-PUTs), whereas those that do not share the same beam are regarded as Non Interfering PUTs (NI-PUTs) [1], [2], [7]. To overcome the limitation of the DoF, RSMA is utilized to allow each beam to support more than one PUT at the same time, frequency, and code resources. Therefore, the number of PUTs served concurrently can be larger than N_{RF}. Let S_n designate the collection of indices of PUTs in the nth beam for $n = 1, 2, \ldots, N_{RF}$ such that $S_i \cap S_j = \varnothing$ for $i \neq j$ and $\sum_{n=1}^{N_{RF}} |S_n| = K$ [2], [7].

D. ORP METHOD

In the ORP method, the downlink signals are precoded by the gNB using orthogonal random vectors before transmission [26], [46]. To explain the ORP technique succinctly, the ORP method is divided into two stages [26], [46] as shown in Fig. 2.

1) TRAINING STAGE

A training signal vector $\mathbf{s}$ such that $\mathbb{E}\left(\mathbf{s}\mathbf{s}^H\right) = \mathbf{I}_K$ and a precoding matrix $\boldsymbol{\Psi}$ consisting of K orthonormal vectors $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_K)$ are generated by the gNB where the size of $\boldsymbol{\psi}_i$ is $N_{RF} \times 1$, $i = 1, 2, \ldots, K$ [26], [46]. The training signals are then precoded using the generated precoding matrix before transmission to PUTs [26], [46].

On the receiver side, each PUT computes K SINR values $(\text{SINR}_1, \text{SINR}_2, \ldots, \text{SINR}_K)$ corresponding to K orthonormal vectors, selects the largest among them, and sends this largest scalar value back to the gNB through a feedback channel together with the appropriate index of the precoding vector [26], [46]. In this study, the ORP scheme is implemented per beam[2] to mitigate the inter-beam interferences. Specifically, the gNB selects the orthonormal vector corresponding to the index of the highest SINR from among the maximum SINR values obtained at all the PUT positions in the nth beam as the precoding vector for that

[2]Different orthonormal precoding vectors are selected for each beam such that $\boldsymbol{\psi}_i^H \boldsymbol{\psi}_j = \begin{cases} 0, & i \neq j \\ 1, & \text{otherwise} \end{cases}$ and is used to precode the private streams of all PUTs in the same beam.

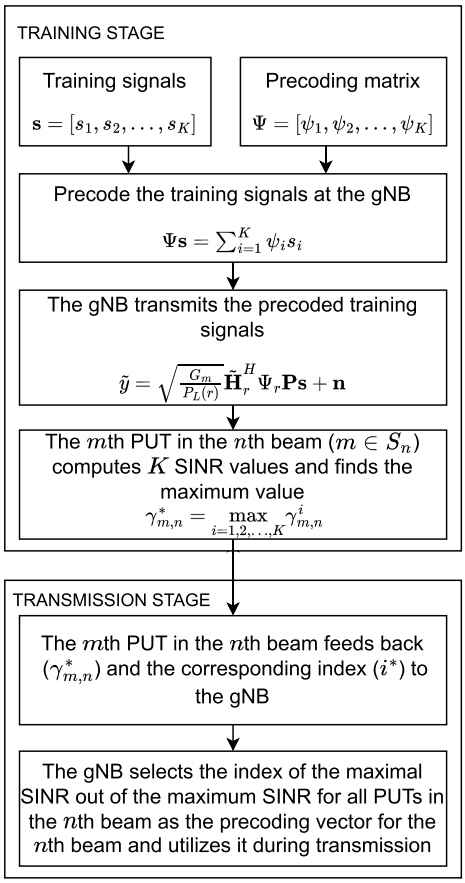

FIGURE 2. Illustration of ORP method training and transmission stages.

beam. During the training stage, the signal received at the mth PUT in the nth beam is

$$y_{m,n} = \sum_{i=1}^{K} \sqrt{\frac{\rho G_m}{N_{RF} P_L(r_{m,n})}} \hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_i s_i + n_{m,n}, \tag{7}$$

where ρ is the SNR, $\hat{\mathbf{h}}_{m,n}$ is the corresponding downlink beamspace channel[3] vector of size $N_{RF} \times 1$ between the gNB and the mth PUT of the nth beam after beam selection, and $n_{m,n} \sim \mathcal{CN}(0, \sigma^2)$ is the AWGN noise at the mth PUT in the nth beam [46].

Each PUT calculates the SINR corresponding to the kth orthonormal vector $\boldsymbol{\psi}_k$ by considering s_k as the desired signal, whereas other SINRs are interferences from $K-1$ orthonormal vectors (i.e., $i \neq k, i = 1, 2, \ldots, K$) [26]. Therefore, the SINR for the signal spanned by $\boldsymbol{\psi}_k$ at the mth PUT in the nth beam can be expressed as

$$\gamma_{m,n}^k = \frac{\eta_{m,n} |\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_k|^2}{\eta_{m,n} \sum_{i \neq k}^{K} |\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_i|^2 + \sigma^2}, \tag{8}$$

[3]We assumed that each PUT can estimate $\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_i, i = 1, 2, \ldots, K$ and feed it back to the gNB by utilizing training procedures [26] and reliable channel estimation algorithms. The estimation of $\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_i$ requires less overhead than the uplink channel estimation for $K < N$ [26].

where

$$\eta_{m,n} = \frac{\rho G_m}{N_{RF} P_L(r_{m,n})}. \tag{9}$$

2) TRANSMISSION STAGE

In this stage, the gNB knows the index of the selected orthonormal vector for each beam and utilizes it to precode the transmit signal for all PUTs within each beam [26], [46]. The information about the orthonormal precoding vectors employed for PUTs from other beams is not important at all PUTs [46].

E. RSMA METHOD

In RSMA, the messages $(x_1, x_2, \ldots, x_K)$ intended for K PUTs are divided into common $(x_{c,1}, x_{c,2}, \ldots, x_{c,K})$ and private $(x_{p,1}, x_{p,2}, \ldots, x_{p,K})$ segments [5], [16], [18], [29]. The common segments of all PUTs are merged into a common message X_c and encoded into a common stream s_c using a codebook known to all PUTs, whereas the private segments are encoded separately into private streams $(s_1, s_2, \ldots, s_K)$ [5], [16], [18], [29]. Hence, the common stream s_c can be decoded by all PUTs in all beams with negligible error probability, and it contains segments of messages $x_1, x_2, \ldots, x_K$ [5], [16], [18], [29]. Moreover, the private streams $(s_{1,n}, s_{2,n}, \ldots, s_{|S_n|,n})$ intended for all PUTs in the nth beam are superimposed over the common message and then linearly precoded by beam-specific ORP precoding. Therefore, the superimposed and precoded transmit signal s at the gNB can be expressed as

$$\mathbf{s} = \sqrt{P_c} \boldsymbol{\psi}_c s_c + \sum_{n=1}^{N_{RF}} \sum_{m=1}^{|S_n|} \sqrt{P_{m,n}} \boldsymbol{\psi}_n s_{m,n}, \tag{10}$$

where P_c and $\boldsymbol{\psi}_c$ are the transmit power and the precoding vector of the common stream s_c, respectively, such that $\boldsymbol{\psi}_c$ is randomly chosen from the unselected columns of the ORP matrix $\boldsymbol{\Psi}$, $P_{m,n}$ is the transmit power of the private stream $s_{m,n}$ transmitted to the mth PUT in the nth beam, $\boldsymbol{\psi}_n$ is the precoding vector of the private streams of PUTs in the nth beam, and $s_{m,n}$ is the private stream of the mth PUT in the nth beam [5], [16], [18], [29].

The transmit power constraint can be expressed as $P_c + \sum_{n=1}^{N_{RF}} \sum_{m=1}^{|S_n|} P_{m,n} \leq P_T$, where P_T is the total transmit power, by defining $\tilde{\mathbf{s}} = [s_c, s_{1,n}, s_{2,n}, \ldots, s_{|S_n|,n}]$ and assuming that $\mathbb{E}\left(\tilde{\mathbf{s}}\tilde{\mathbf{s}}^H\right) = 1$ [5], [16], [18], [29].

At each PUT in the nth beam, the common stream s_c is decoded first into $\hat{X}_c$ by considering the interference from the private streams as noise [5], [16], [18], [29], [36]. Employing SIC, each PUT in the nth beam will re-encode $\hat{X}_c$, precode it, and then eliminate it from the signal that was received. The PUT will then decode its private stream $s_{m,n}$ into $\hat{x}_{p,m}$ by treating the residual interference from the other private stream as noise [5], [16], [18], [29], [36]. Then the mth PUT in the nth beam restores the original message by removing $\hat{x}_{c,m}$ from $\hat{X}_c$, and combining $\hat{x}_{c,m}$ with $\hat{x}_{p,m}$ resulting in $\hat{x}_m$ [5], [16], [18], [29], [36].

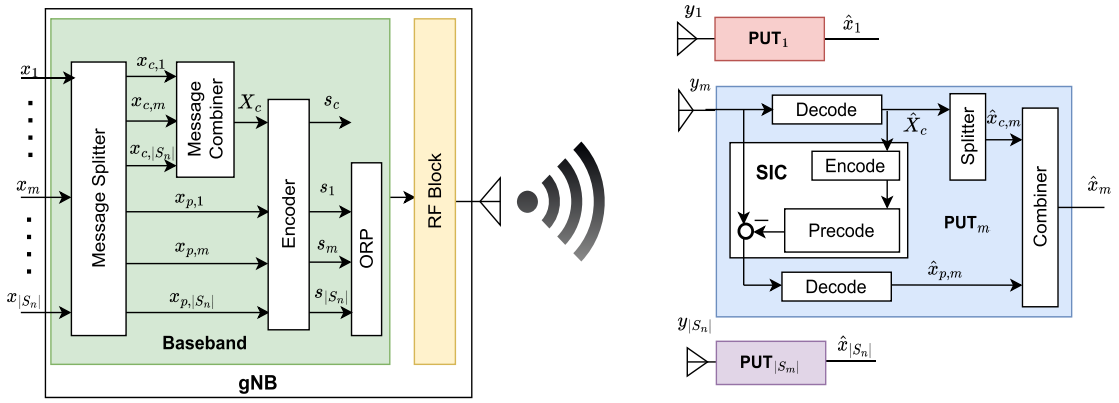

FIGURE 3. Illustration of RSMA architecture for PUTs in the nth beam.

This operation is referred to as 1-layer rate splitting since it only depends on a single common message and a single SIC layer at each PUT [5], [16] as shown in Fig. 3.

The received signal at the mth PUT in the nth beam can be expressed as

$$y_{m,n} = \sqrt{\frac{G_m P_c}{P_L\left(r_{m,n}\right)}}\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_c s_c + \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\sqrt{\frac{G_m P_{m,n}}{P_L\left(r_{m,n}\right)}}\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n s_{m,n} + n_{m,n}. \quad (11)$$

Consequently, the attainable throughput of the mth PUT in the nth beam for decoding the common stream is given as

$$R_{m,n}^c = \log_2\left(1 + \frac{\upsilon_{m,n} P_c |\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_c|^2}{\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n} P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2 + \sigma^2}\right), \quad (12)$$

where

$$\upsilon_{m,n} = \frac{G_m}{P_L\left(r_{m,n}\right)}. \quad (13)$$

The possible throughput of the common stream should be selected as $\min_{m,n} R_{m,n}^c, \forall n, m$ for the common stream s_c to be successfully decoded by all PUTs [29].

Moreover, for SIC operation to be implemented successfully at the mth PUT in the nth beam, the transmit power must satisfy the following constraint:

$$\upsilon_{m,n} P_c |\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_c|^2 - \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n} P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2 - \sigma^2 \geq \vartheta, \forall n, m, \quad (14)$$

where ϑ is the difference between the desired signal power and the undesired interference signal power plus noise power [29], [52].

After perfectly decoding the common stream, the attainable throughput of the mth PUT in the nth beam for decoding its private stream is

$$R_{m,n}^p = \log_2\left(1 + \frac{\upsilon_{m,n} P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2}{I_{m'} + \sigma^2}\right), \quad (15)$$

where

$$I_{m'} = \sum_{m' \neq m}\upsilon_{m,n} P_{m',n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2 + \sum_{i \neq n}^{N_{RF}}\sum_{j=1}^{|S_n|}\upsilon_{m,n} P_{j,i}|\hat{\mathbf{h}}_{m,i}^H \boldsymbol{\psi}_i|^2. \quad (16)$$

Hence, the attainable sum throughput of the mth PUT in the nth beam is given as

$$R_{m,n} = R_{m,n}^c + R_{m,n}^p. \quad (17)$$

The total system attainable throughput can be written as

$$R_{\text{sum}} = \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\left(R_{m,n}^c + R_{m,n}^p\right). \quad (18)$$

F. COVERAGE ANALYSIS

In terms of network planning and deployment of communication systems, cell coverage percentage is a key Quality of Service (QoS) performance metrics [47]. Therefore, it is very important to analyze the performance of this model in terms of cell coverage percentage. Thus, the downlink cell coverage percentage is defined as the expected percentage of locations within a cell where the SINR of the received signal at PUT exceeds the threshold $T_{\min}$ [48], [49], [53]. We denote the probability that the power of the signal received at the mth PUT in the nth beam exceeds the SINR threshold $T_{\min}$ at a distance of $r_{m,n}$ meters from the gNB as $P_A\left(r_{m,n}\right)$.

Proposition: Given $P_c > P_{m,n}$, the downlink cell coverage percentage of mmWave beamspace mMIMO gNB utilizing ORP and RSMA is

$$C_{\text{cov}} = \mathrm{Q}\left(-x\right) + \exp\left[2\left(\frac{1-xy}{y^2}\right)\right]\mathrm{Q}\left(\frac{2}{y} - x\right), \quad (19)$$

where $\mathrm{Q}(\cdot)$ is the Q-function defined as $\mathrm{Q}(x) = \frac{1}{\sqrt{2\pi}}\int_x^t e^{\frac{-t^2}{2}}dt$, $x = \frac{P_L(r_0)+10\bar{n}\log_{10}\left(\frac{r_{m,n}}{r_0}\right)-a}{\xi}$, $y = \frac{10\bar{n}\log_{10}(e)}{\xi}$, $a = \frac{G_m P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2}{\sigma^2 T_{\min}} - \frac{I_n}{\sigma^2}$, and $I_n = \sum_{m'\neq m} G_m P_{m',n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2 + \sum_{i\neq n}^{N_{RF}}\sum_{j=1}^{|S_n|} G_m P_{j,i}|\hat{\mathbf{h}}_{m,i}^H \boldsymbol{\psi}_i|^2$.

Proof: See Appendix A.

III. OPTIMIZATION AND CONVERGENCE ANALYSIS

To enhance the overall feasible throughput performance of this system, it is necessary to optimize the common stream throughput and the power allotted to the private stream of each PUT and the common stream. Therefore, the feasible throughput optimization can be formulated mathematically as

$$\begin{aligned}
&\max_{\{\mathbf{a},\mathbf{P}\}} \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\left(a_{m,n}+R_{m,n}^p\right),\\
&\text{s.t. C1}: \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} a_{m,n} \le R_{m,n}^c,\ \forall n,m,\\
&\quad \text{C2}: a_{m,n}+R_{m,n}^p \ge R_{\min},\ \forall n,m,\\
&\quad \text{C3}: \upsilon_{m,n}P_c|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_c|^2 - \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2\\
&\quad -\sigma^2 \ge \vartheta,\quad \forall n,m,\\
&\quad \text{C4}: P_c + \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} P_{m,n} \le P_T,\\
&\quad \text{C5}: a_{m,n}, P_c, P_{m,n} \ge 0,\quad \forall n,m,
\end{aligned} \tag{20}$$

where $R_{\min}$ is the minimum throughput guaranteed for all PUTs. Constraint C1 ensures that all PUTs can decode the common stream. This means that the sum of the throughputs of all PUTs that are receiving the common stream must be lower than the common stream throughput $R_{m,n}^c$ [29]. C2 guarantees the minimum sum throughput for all PUTs, Whereas C3 requires more power to be allocated to the common stream to ensure successful SIC operation. C4 represents the maximum power constraint and C5 guarantees non-negative throughput $a_{m,n}$, common stream power P_c, and private stream power $P_{m,n}$. Owing to the nonconvex structure of the objective function and constraints, the aforementioned optimization problem (20) is infeasible.

A. THROUGHPUT AND POWER ALLOCATION

To approximate the optimization problem (20) into a convex optimization problem, we introduce slack variables $\gamma_{m,n}^p$ and $\gamma_{m,n}^c$. Therefore, the feasible sum throughput optimization problem (20) can be reframed mathematically as

$$\begin{aligned}
&\max_{\{\mathbf{a},\mathbf{P},\boldsymbol{\gamma}^c,\boldsymbol{\gamma}^p\}} \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\left(a_{m,n}+\log_2\left(1+\gamma_{m,n}^p\right)\right),\\
&\text{s.t. C1}: \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} a_{m,n} \le \log_2(1+\gamma_{m,n}^c),\ \forall n,m,\\
&\quad \text{C2}: a_{m,n}+\log_2(1+\gamma_{m,n}^p) \ge R_{\min},\ \forall n,m,\\
&\quad \text{C3}: \gamma_{m,n}^c \le \frac{\upsilon_{m,n}P_c|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_c|^2}{\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2+\sigma^2},\\
&\quad \forall n,m,\\
&\quad \text{C4}: \gamma_{m,n}^p \le \frac{\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2}{I_{m'}+\sigma^2},\ \forall n,m,\\
&\quad \text{C5}: \upsilon_{m,n}P_c|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_c|^2 - \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2\\
&\quad -\sigma^2 \ge \vartheta,\ \forall n,m,\\
&\quad \text{C6}: P_c + \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} P_{m,n} \le P_T,\\
&\quad \text{C7}: a_{m,n}, P_c, P_{m,n} \ge 0,\ \forall n,m.
\end{aligned} \tag{21}$$

The problem in (21) is nonconvex owing to constraints C3 and C4, as shown in Appendix B. Therefore, it is necessary to introduce the variables $\alpha_{m,n}$ and $\beta_{m,n}$ to handle constraints C3 and C4, respectively. Let us rewrite constraint C3 such that

$$\frac{\upsilon_{m,n}P_c|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_c|^2}{\alpha_{m,n}} \ge \gamma_{m,n}^c, \tag{22}$$

where

$$\alpha_{m,n} = \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2+\sigma^2. \tag{23}$$

By using the first-order Taylor series approximation, we can transform the constraint referred to as (22) into the convex constraint shown here as

$$\upsilon_{m,n}P_c\frac{|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_c|^2}{\alpha_{m,n}^{(l)}} - \frac{\upsilon_{m,n}P_c^{(l)}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_c|^2}{\left(\alpha_{m,n}^{(l)}\right)^2}\alpha_{m,n} \ge \gamma_{m,n}^c, \tag{24}$$

where

$$\alpha_{m,n}^{(l)} = \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\upsilon_{m,n}P_{m,n}^{(l)}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2+\sigma^2, \tag{25}$$

and the superscript (l) denotes the variable's value at the lth iteration.

To transform C4 into a convex constraint, first, we rewrite it as

$$\frac{\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2}{\beta_{m,n}} \ge \gamma_{m,n}^p, \tag{26}$$

where

$$\begin{aligned}
\beta_{m,n} &= I_{m'}+\sigma^2\\
&= \sum_{m'\neq m}\upsilon_{m,n}P_{m',n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2\\
&\quad + \sum_{i\neq n}^{N_{RF}}\sum_{j=1}^{|S_n|}\upsilon_{m,n}P_{j,n}|\hat{\mathbf{h}}_{m,i}^H\boldsymbol{\psi}_i|^2+\sigma^2.
\end{aligned} \tag{27}$$

We then utilize the first-order Taylor series approximation to transform (26) into the following convex constraint:

$$\frac{\upsilon_{m,n}P_{m,n}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2}{\beta_{m,n}^{(l)}} - \frac{\upsilon_{m,n}P_{m,n}^{(l)}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2}{\left(\beta_{m,n}^{(l)}\right)^2}\beta_{m,n} \geq \gamma_{m,n}^p, \tag{28}$$

where

$$\beta_{m,n}^{(l)} = \sum_{m'\neq m}\upsilon_{m,n}P_{m',n}^{(l)}|\hat{\mathbf{h}}_{m,n}^H\boldsymbol{\psi}_n|^2 + \sum_{i\neq n}^{N_{RF}}\sum_{j=1}^{|S_n|}\upsilon_{m,n}P_{j,n}^{(l)}|\hat{\mathbf{h}}_{m,i}^H\boldsymbol{\psi}_i|^2 + \sigma^2. \tag{29}$$

$\beta_{m,n}^{(l)}$ is the variable's value at the lth iteration.

To this end, the nonconvex feasible sum throughput optimization problem (21) can be reframed into the following approximated convex problem:

$$\begin{aligned}&\max_{\{\mathbf{a},\mathbf{P},\boldsymbol{\gamma}^c,\boldsymbol{\gamma}^p,\boldsymbol{\alpha},\boldsymbol{\beta}\}} \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}R_{m,n},\\ &\text{s.t. C1, C2, C5, C6, C7, (24), (28),}\\ &\text{C10}: \alpha_{m,n}, \beta_{m,n} \geq 0, \quad \forall n, m,\end{aligned} \tag{30}$$

where

$$R_{m,n} = \left(a_{m,n} + \log_2\left(1+\gamma_{m,n}^p\right)\right). \tag{31}$$

The approximated optimization problem (30) can be solved iteratively using the SCA-based algorithm summarized in Algorithm 1.

Algorithm 1 Feasible Sum Throughput Maximization Algorithm

1: Initialize $a_{m,n}^{(0)}, P_{m,n}^{(0)}, P_c^{(0)}, \gamma_{m,n}^{c(0)}, \gamma_{m,n}^{p(0)}, \alpha_{m,n}^{(0)}, \beta_{m,n}^{(0)}$. Set $l = 1$ as the iteration number, and a constant ϵ.
2: **repeat**
3: Solve the approximated convex optimization problem (30), and obtain the optimal solution of (30) denoted by $a_{m,n}^{(l)}, P_{m,n}^{(l)}, P_c^{(l)}, \gamma_{m,n}^{c(l)}, \gamma_{m,n}^{p(l)}, \alpha_{m,n}^{(l)}, \beta_{m,n}^{(l)}$.
4: Update $l := l + 1$
5: **until** $\|\Phi^{(l+1)} - \Phi^{(l)}\| \leq \epsilon$; (i.e., the convergence of the objective function (30) in the lth iteration);
6: where $\Phi^{(l)} = \{a_{m,n}^{(l)}, P_{m,n}^{(l)}, P_c^{(l)}, \gamma_{m,n}^{c(l)}, \gamma_{m,n}^{p(l)}, \alpha_{m,n}^{(l)}, \beta_{m,n}^{(l)}\}$ is the set of the optimal solution of problem (30) at the lth iteration.

B. COMPLEXITY ANALYSIS

The SCA-based approach is used to solve the approximated convex optimization problem (30), where the approximated problem (30) is solved and the solutions found are optimum at each iteration. As a result, iteratively updating the variables would enhance or maintain the viable total throughput (i.e., nondecreasing monotonically) and converge to a position that fulfills the Karush-Kuhn-Tucker (KKT) optimum conditions [29]. The number of constraints in problem (21) is $(6K + 1)$. Hence, for Algorithm 1, the necessary number of iterations for solving the problem by using SCA method is $\mathcal{O}\left(\sqrt{6K+1}\log_2(1/\varepsilon)\right)$, where $\varepsilon > 0$ is the accuracy of the SCA-based algorithm [54]. The difficulty of solving the problem (30) at each iteration is given as $\mathcal{O}\left(T_1^2T_2\right)$, where $T_1 = 7K$ is the overall number of variables and $T_2 = (8K + 1)$ is the total number of constraints [29], [55]. Therefore, the total complexity to solve problem (30) using SCA method in Algorithm 1 is $\mathcal{O}\left(T_{\max}K^{3.5}\log_2(1/\varepsilon)\right)$, where $T_{\max}$ is the total number of iterations. Table 1 provides the complexity comparison between the proposed method ORP-RSMA and the iterative power allocation algorithm (i.e., ZF-NOMA) proposed in [2], which requires $\mathcal{O}\left(T_{\max}K^2\log_2(\varepsilon)\right)$ number of iterations.

TABLE 1. Computational complexity comparison.

Method	Proposed ORP-RSMA	ZF-NOMA [2]
Multiplications	$\mathcal{O}\left(T_{\max}K^{3.5}\log_2(1/\varepsilon)\right)$	$\mathcal{O}\left(T_{\max}K^2\log_2(\varepsilon)\right)$

IV. PERFORMANCE EVALUATION

Simulations are carried out to evaluate the efficacy of the proposed method in this section. We considered the mmWave mMIMO system in a downlink context in which the gNB is deployed with a ULA of $N = 64$ antennas and $N_{RF} = K$ that serves K PUTs concurrently [1], [7]. At each instance, the number of dominant beams can be less than or equal to N_{RF}. To generate a precoding matrix, an orthonormal basis is calculated for the column space of a matrix that is randomly generated, which results in a precoding matrix consisting of orthonormal precoding vectors [26], [46]. One LOS component and $N_p = 2$ NLOS components are considered for the channels between the gNB and all PUTs [2]. For the channel parameters of the kth PUT, it is assumed that $\beta_k^{(0)} \sim \mathcal{CN}(0,1)$ and $\beta_k^{(l)} \sim \mathcal{CN}\left(0, 10^{-1}\right)$ for $1 \leq l \leq N_p$ [1], [2]. In addition, $\theta_k^{(0)}$ and $\theta_k^{(l)}$ for $1 \leq l \leq N_p$ are random variables uniformly distributed within $[-\frac{1}{2}, \frac{1}{2}]$ [1], [2]. Furthermore, in this study, the SNR is expressed as $\rho = P_T/\sigma^2$.

The parameters of the simulation are described in Table 2.

TABLE 2. Simulation parameters.

Parameter	Value
The maximum transmit power P_T	32 mW [2]
Minimum required throughput for each PUT R_{min}	0.5 bps/Hz [56]
SIC detection threshold ϑ	2 dBm [29]
The number of PUTs K	$10 \sim 30$
Average path loss exponent $\bar{n}$	3.5
Standard deviation accounting for shadowing effect ξ	8.7 dB
The downlink receiver SINR threshold	$-3 \sim 9$ dB
The downlink receiver SNR	$0 \sim 20$ dB
Cell radius r	200 m
Main-lobe antenna gain at gNB G_{TX}	25 dBi [57]
Main-lobe antenna gain at PUT G_{RX}	13.3 dBi [57]
Signal wavelength λ	10 mm

The proposed beamspace mMIMO with the lens antenna array using ORP and RSMA (ORP-RSMA) is compared with three benchmark schemes: (*i*) the beamspace mMIMO system that utilizes ZF precoding and SDMA, where the equivalent throughput maximization problem is solved, is referred to as (ZF-SDMA). Moreover, SDMA is a multiple access technique that is a special case of the RSMA method when rate-splitting is not performed [20], [39]; (*ii*) the beamspace mMIMO-NOMA system that groups PUTs into clusters using maximal magnitude and utilizes ZF precoding to suppress inter-beam interference and NOMA to mitigate intra-beam interference and serves $K \geq N_{RF}$ PUTs is referred to as (ZF-NOMA) [2], where sum throughput maximization and power allocation are performed as in [2], and (*iii*) the beamspace mMIMO-OMA system with $N_{RF} \leq K$ that utilizes ZF precoding to mitigate inter-beam interference and implements orthogonal frequency resource allocation to mitigate intra-beam interference for PUTs within the same beam and equal power allocation is referred to as (ZF-OMA). These benchmark schemes are selected for evaluation as SDMA and NOMA are potential multiple access techniques, while OMA is a conventional multiple access technique.

A. THROUGHPUT

Figure 4 shows the sum throughput vs SNR when the number of PUTs served by the gNB is $K = 32$. As can be seen from the figure, the proposed ORP-RSMA method achieves higher performance than the benchmark methods. ZF-SDMA shows better performance compared to ZF-NOMA [2] and ZF-OMA at high SNR regions; however, its performance remains inferior to our proposed method. The reason behind the poor performance of ZF-NOMA [2] is that more power is assigned to the PUTs with low channel gains. Moreover, ZF-NOMA treats multi-PUT interference as pure noise and forces a PUT to decode all the messages intended for all other PUTs, which results in poor sum throughput performance. On the other hand, the ORP-RSMA method

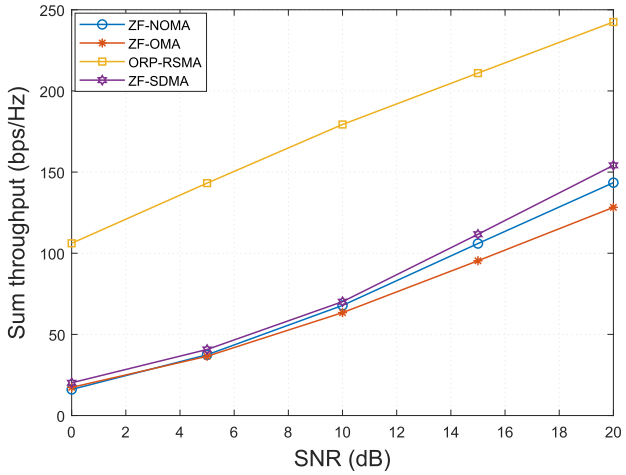

FIGURE 4. Total throughput vs SNR for $K = 32$ PUTs.

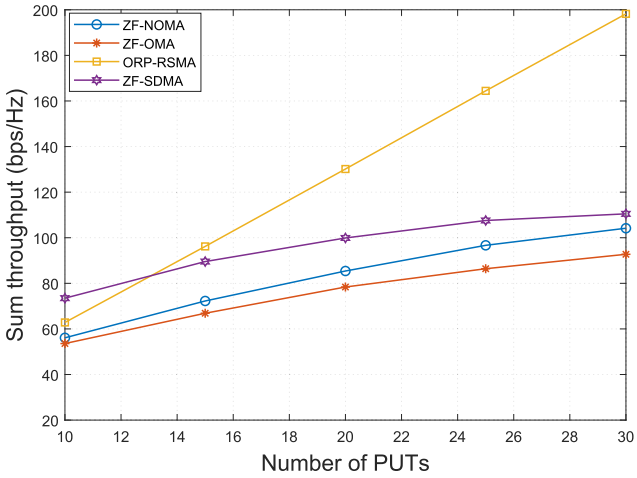

FIGURE 5. Total throughput vs various numbers of PUTs when SNR = 15 dB.

aligns each beam with a precoding vector that nullifies inter-beam interference and mitigates intra-beam interference by treating multi-PUT interference as noise or interference, which in-turn maximizes SINR, and allocates power to the common stream based on constraint C5 in (30), resulting in superior sum throughput performance when K is large. Furthermore, in the ORP-RSMA method, one-layer of SIC is performed as opposed to $|S_n|$-layer[4] of SIC in ZF-NOMA [2], giving rise to higher DoFs and lower complexity at the receivers. Compared with ZF-OMA, ORP-RSMA achieves better performance owing to the fact that all the PUTs in the same beam are served with the entire bandwidth, whereas ZF-OMA splits the bandwidth among the PUTs [9].

Figure 5 shows the sum throughput vs the number of PUTs when SNR = 15 dB. It is clear from this figure that ZF-SDMA outperforms the proposed ORP-RSMA method when the number of PUTs is small owing to higher DoF despite the sparsity of beamspace massive MIMO. However, as the system becomes overloaded (i.e., reduction in DoF), the proposed method outperforms ZF-SDMA. On the other hand, the proposed method slightly outperforms the ZF-NOMA [2] and ZF-OMA methods by 10 bps/Hz and 12 bps/Hz, respectively, when the number of PUTs is 10. As more PUTs are served, the performance difference between the proposed approach and the benchmark methods increases. This is because when the number of PUTs is small, our proposed ORP-RSMA method allocates portion of the downlink power to the common stream, which is decoded by fewer available PUTs, resulting in poor total throughput performance.

B. ENERGY EFFICIENCY

The energy efficiency η is described as the ratio of the maximized sum throughput to the overall power consumed

[4] $|S_n|$-layer refers to the number of PUTs within the same beam.

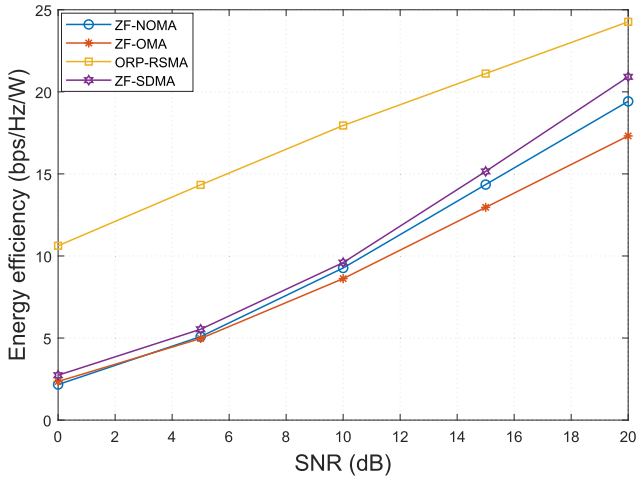

FIGURE 6. Energy efficiency vs SNR for $K = 32$ PUTs.

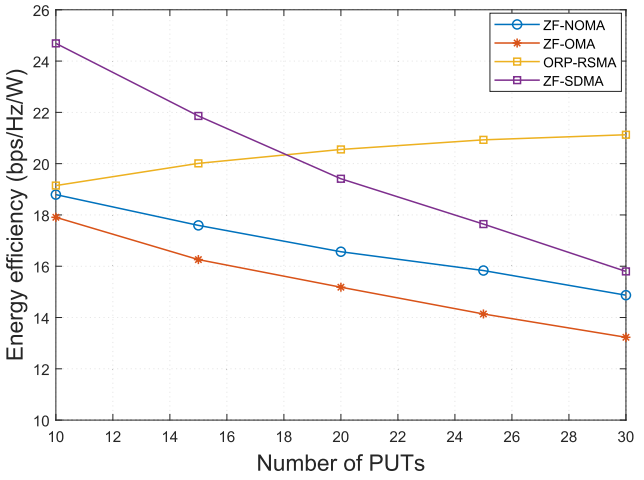

FIGURE 7. Energy efficiency vs various numbers of PUTs when SNR = 15 dB.

by the system and can be expressed as

$$\eta = \frac{R_T}{P_T + N_{RF}P_{RF} + N_{RF}P_S + P_B} \text{ (bps/Hz/W)}, \quad (32)$$

where $R_T = \sum_{n=1}^{N_{RF}} \sum_{m=1}^{|S_n|} R_{m,n}$ is the maximum sum throughput of the system, P_T is the maximum power transmitted, P_{RF} is the power dissipated at each RF block, P_S is the power consumed by each switch, and P_B is the power consumed at the baseband [2], [7]. Specifically, we consider that $P_{RF} = 300$ mW, $P_S = 5$ mW, and $P_B = 200$ mW [2].

In Fig. 6, the energy efficiency vs the SNR of the suggested ORP-RSMA method is compared with those of the benchmark methods for $K = 32$ PUTs. Clearly, the proposed ORP-RSMA method attains superior performance compared with the ZF-SDMA, ZF-NOMA [2], and ZF-OMA methods in both low- and high-SNR regions. Specifically, the proposed method shows a gain of approximately 11 bps/Hz/W at SNR = 0 dB and gains of approximately 13 bps/Hz/W and 15 bps/Hz/W compared with ZF-NOMA [2] and ZF-OMA [14], respectively, at SNR = 20 dB. The superior performance of the proposed method stems from the fact that more power is allocated to the common stream while performing only one SIC layer.

Figure 7 depicts the relationship between energy efficiency and the number of PUTs at an SNR of 15 dB. We observe that ZF-SDMA achieves better performance than the proposed ORP-RSMA, ZF-NOMA [2], and ZF-OMA methods when the number of PUTs is small. However, the proposed ORP-RSMA method achieves gains of approximately 2.5 bps/Hz/W, 2 bps/Hz/W, and 3 bps/Hz/W as compared to ZF-SDMA, ZF-NOMA, and ZF-OMA [2], [14], respectively, as the number of PUTs approaches 20. As the number of PUTs continue to increase, the proposed method become more energy-efficient than all the benchmark methods. Conversely, the ZF-SDMA, ZF-NOMA [2], and ZF-OMA methods become more energy-inefficient as the number of PUTs continue to increase. The superior performance of the proposed ORP-RSMA method stems from the allocation of more power to the common stream than the individual private streams, resulting in higher energy efficiency when the number of PUTs increases as more PUTs decode the common stream.

C. CELL COVERAGE PERCENTAGE

Figure 8 shows the cell coverage percentage vs the SINR threshold requirement. It is clear from the figure that ORP-RSMA method achieves a higher coverage percentage than ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] methods at both low and high SINR thresholds. In particular, ORP-RSMA method can achieve approximately 2%, 14%, and 20% gains in terms of cell coverage percentage compared with ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] methods at low SINR thresholds, and up to 100%, 26%, and 38% gains compared with ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] at high SINR thresholds, respectively. This is

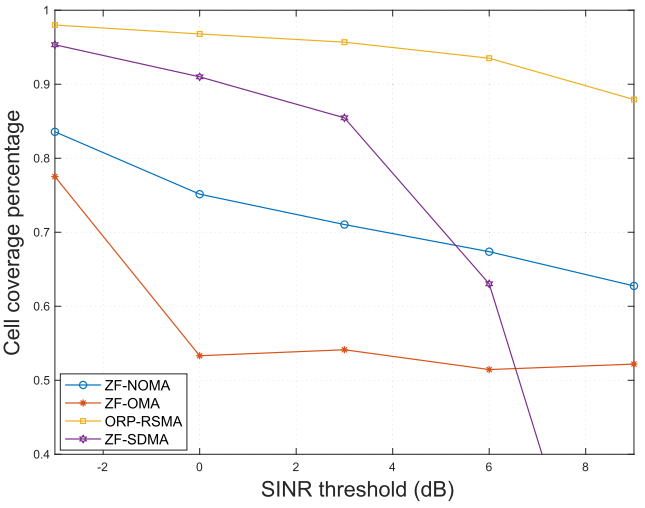

FIGURE 8. Cell coverage percentage vs the SINR threshold when SNR = 15 dB.

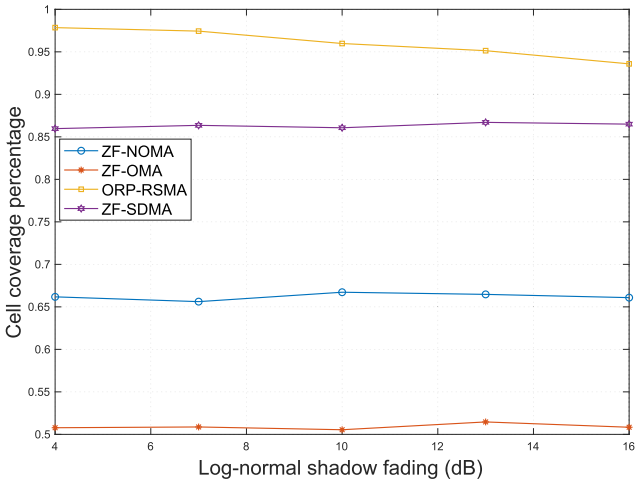

FIGURE 9. Cell coverage percentage vs log-normal shadow fading when SNR = 15 dB and SINR threshold $T_{\text{min}} = 2$ dB.

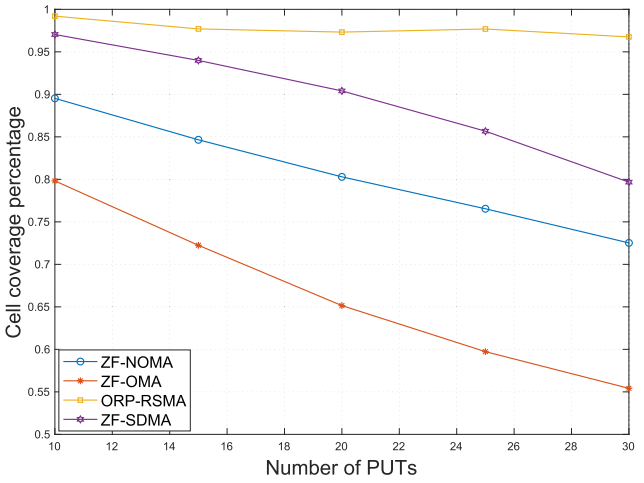

FIGURE 10. Cell coverage percentage vs number of PUTs when SNR = 15 dB and SINR threshold $T_{\text{min}} = 2$ dB.

because the ORP-RSMA method selects the precoding vector that produces the maximum SINR. On the other hand, at higher SINR threshold requirements, ORP-RSMA, ZF-NOMA [2], and ZF-OMA [14] outperform the ZF-SDMA method with ORP-RSMA and ZF-NOMA [2] outperforming ZF-OMA [14]. This is because both ORP-RSMA and ZF-NOMA [2] methods utilize non-orthogonal multiple access techniques, which can lead to higher SINRs than the ZF-OMA [14] method.

Figure 9 show result of assessment of performance in terms of cell coverage percentage vs log-normal shadowing, when SNR = 15 dB and the SINR threshold $T_{\text{min}} = 2$ dB. It can be inferred from Fig. 9 that the proposed ORP-RSMA method attains superior performance compared with the other three benchmark methods even when the value of log-normal shadow fading is high. We can clearly see that the ZF-OMA [14] method achieves a worse performance than ORP-RSMA, ZF-SDMA, and ZF-NOMA [2]. In particular, the performance difference between ZF-OMA [14] and the three methods continues to persist as the degree of the shadowing effect increases.

Figure 10 shows the cell coverage percentage vs the number of PUTs when SNR = 15 dB and the SINR threshold $T_{\text{min}} = 2$ dB. From the figure, it is clear that the cell coverage percentage decreases for ORP-RSMA, ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] methods as the number of PUTs increases within the cell while the total transmit power is fixed. However, the cell coverage percentage for ORP-RSMA decreases gradually and the method achieves an acceptable performance at both small and large numbers of PUTs in terms of cell coverage percentage, whereas those for ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] decrease sharply with respect to ORP-RSMA. Note that ORP-RSMA achieves superior performance compared with the ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] methods. This is due to the robustness of RSMA against interferences and the capability of ORP to select a precoder that maximizes SINR for the PUTs served by the nth beam. On the other hand, ZF-SDMA outperforms both ZF-NOMA [2] and ZF-OMA [14], while ZF-NOMA [2] outperforms the ZF-OMA [14] method owing to its interference management mechanism, which depends on SIC at the PUTs; nonetheless, the complexity of the SIC increases with an increasing number of PUTs.

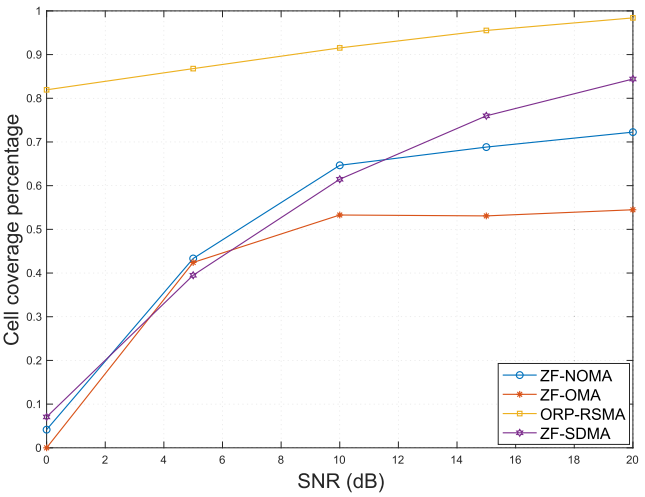

FIGURE 11. Cell coverage percentage vs SNR for $K = 32$ PUTs when SINR threshold $T_{\text{min}} = 2$ dB.

Figure 11 shows the cell coverage percentage vs SNR when the SINR threshold $T_{\text{min}} = 2$ dB. From Fig. 11 we can observe that the coverage percentage for all the methods increases with the SNR. However, we can recognize that ORP-RSMA outperforms ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] methods at low- and high-SNR regions. ZF-SDMA outperforms both ZF-NOMA [2] and ZF-OMA [14] at high SNR regions, while the coverage percentage of ZF-NOMA [2] is slightly larger than that of the ZF-OMA [14] method at low SNRs. However, the performance difference

between the ZF-NOMA [2] and ZF-OMA [14] methods widens as the SNR increases. Moreover, as SNR approaches 15 dB, the coverage percentages for ZF-NOMA [2] and ZF-OMA [14] remain constant despite the increase in SNR.

D. CONVERGENCE

The convergence of the proposed algorithm in Section III is evaluated when the number of PUTs is $K = 32$, and SNR = 5 dB for SIC detection thresholds $\vartheta =$ 2dBm and $\vartheta = 4$ dBm. As shown in Fig. 12, the sum throughput of the proposed method becomes stable after two iterations regardless of the initialization, which confirms the convergence of the algorithm as discussed in Section III-B.

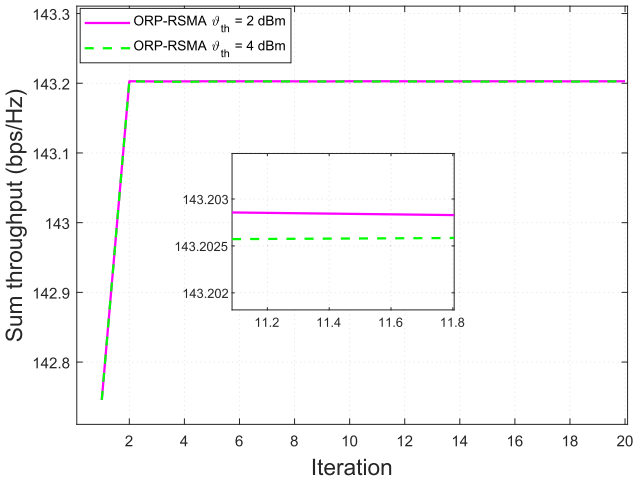

FIGURE 12. Total throughput vs number of iterations when SNR = 5 dB.

E. IMPERFECT CSI

In this subsection, we evaluate the impact of imperfect CSI on the proposed ORP-RSMA method as the residual multiuser interference can degrade the performance of wireless networks [58]. The imperfect estimated channel matrix is modeled as

$$\tilde{\mathbf{H}}_r = \tau \hat{\mathbf{H}}_r + \sqrt{1-\tau^2}\mathbf{E}, \tag{33}$$

where $\tau \in (0, 1)$ is the error parameter that represents the accuracy of the CSI, $\hat{\mathbf{H}}_r$ is the actual reduced-dimension beamspace channel matrix, and $\mathbf{E}$ is the error matrix whose entries are identically and independently distributed (iid) and follows the distribution $\mathcal{CN}(0, 1)$ [59].

In Fig. 13, the sum throughput vs SNR of the proposed model is evaluated when $K = 32$ PUTs for perfect and imperfect CSI. It is clear that the proposed method with imperfect CSI ($\tau = 0.6$) achieves superior performance compared to ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] schemes with perfect CSI.

Figure 14 presents the sum throughput vs SNR for imperfect CSI when $K = 32$ PUTs. The proposed ORP-RSMA method achieves better performance in terms of sum throughput compared to ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] methods in both low and high SNR regions when CSI imperfection exists. It is obvious that ZF-NOMA [2] achieves the worst performance by treating multi-PUT interference as pure noise and forcing a PUT to decode the messages of all other PUTs, and hence, imperfect CSI is detrimental to ZF-NOMA [2] more than ZF-SDMA and ZF-OMA [14] in terms of sum throughput.

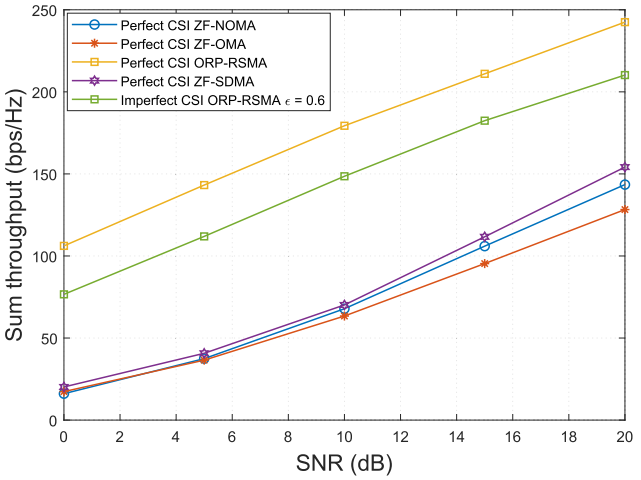

FIGURE 13. Total throughput vs SNR for $K = 32$ PUTs.

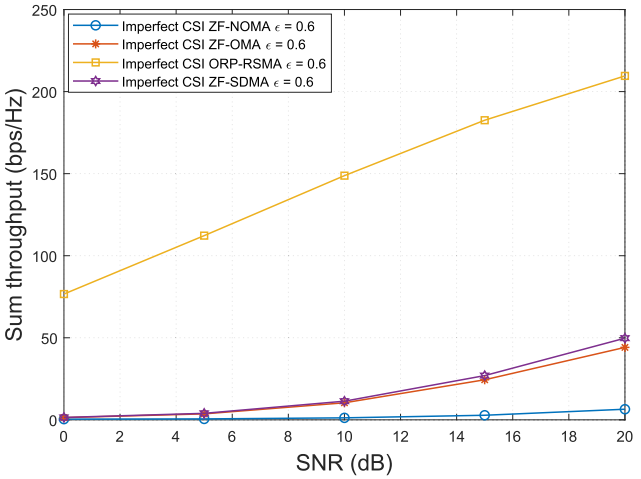

FIGURE 14. Total throughput vs SNR for $K = 32$ PUTs.

Figure 15 shows the energy efficiency vs SNR for imperfect CSI scenario when $K = 32$ PUTs. The proposed ORP-RSMA method has shown the best performance compared to ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] schemes in terms of energy efficiency when SNR increases at the PUTs. The superior performance is due to the fact that ORP-RSMA dynamically treats multi-PUT interference as noise or interference. While ZF-NOMA [2] scheme achieves the worst performance when imperfect CSI exists.

Figure 16 presents the cell coverage area vs SNR for an imperfect CSI ($\epsilon = 0.6$) scenario. We can see from Fig. 16 that the proposed ORP-RSMA method outperforms ZF-SDMA, ZF-NOMA [2], and ZF-OMA [14] schemes

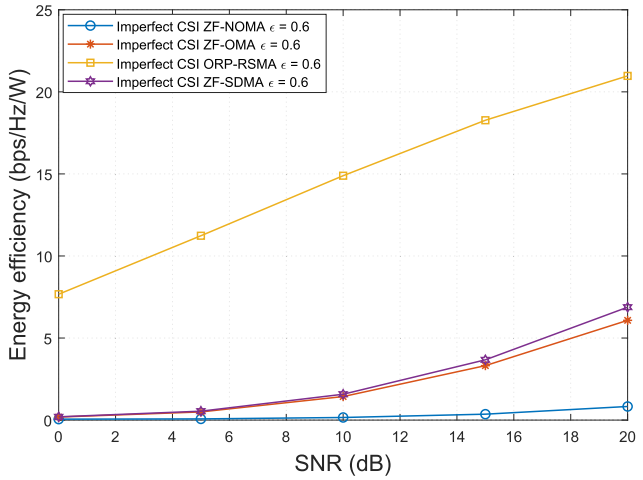

FIGURE 15. Energy efficiency vs SNR for $K = 32$ PUTs.

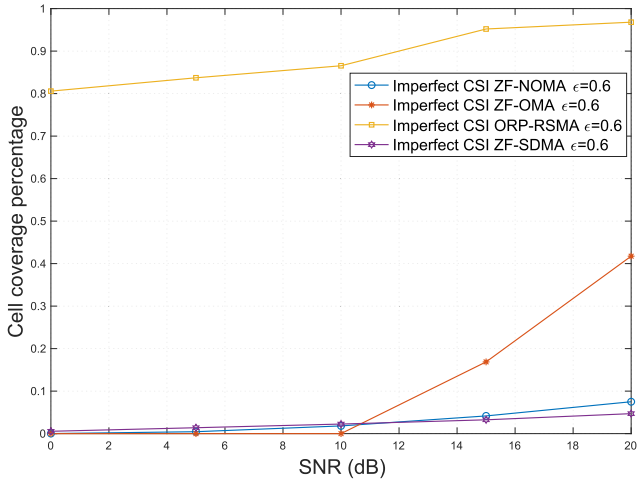

FIGURE 16. Cell coverage area vs SNR for $K = 32$ PUTs.

in terms of cell coverage area owing to it's flexibility in treating multi-PUT interference as noise or interference. The ZF-OMA [14] scheme outperforms both the ZF-SDMA and ZF-NOMA [2] schemes at high SNR regions, while the ZF-SDMA scheme shows the worst performance at high SNR regions.

V. CONCLUSION

In this study, we investigated the sum throughput maximization problem in downlink mmWave beamspace mMIMO that utilizes ORP and RSMA. ORP is employed to mitigate inter-beam interferences and extend the cell coverage percentage, whereas RSMA is used to divide the power and rate between common and private streams to minimize intra-beam interferences. In addition, the SCA approach is implemented to address the specific issue of maximizing the total of the throughputs. The results of the simulation show that the proposed ORP-RSMA approach can achieve greater total throughput and energy efficiency at both low- and high-SNR regions, as well as when the number of PUTs in the cell coverage percentage of the gNB is considerably large. This is the case regardless of whether the SNR is high or low. Furthermore, the proposed method achieves the best performance and shows acceptable cell coverage percentage compared with the benchmark methods in both low- and high-SINR threshold regions. Moreover, the proposed method can be utilized in both low- and high-log-normal shadow fading environments. We intend to expand on this research in subsequent work by examining coverage expansion for multi-cell scenarios involving inter-cell interferences and multiple antenna PUTs.

APPENDIX A
PROOF OF PROPOSITION

Let $T_{\min}$ be the required minimum SINR threshold and $P_A\left(r_{m,n}\right)$ be the probability that the power of the signal received at the mth PUT in the nth beam exceeds the required minimum SINR threshold at a distance of $r_{m,n}$ meters from the gNB.

Then, given $P_c > P_{m,n}$, we assume that the mth PUT in the nth beam is within the coverage area if the SINR of a private stream exceeds the required minimum SINR threshold.

Therefore,

$$\begin{aligned} P_A\left(r_{m,n}\right) &= \mathrm{P}_r\left(\gamma_{m,n}^{p} > T_{\min}\right) \\ &= \mathrm{P}_r\left(\frac{G_m P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2}{I_n + \sigma^2 P_L\left(r_{m,n}\right)} > T_{\min}\right), \end{aligned} \tag{34}$$

where

$$P_L\left(r_{m,n}\right) = P_L\left(r_0\right) + 10\bar{n}\log_{10}\left(\frac{r_{m,n}}{r_0}\right) + X_\xi, \tag{35}$$

and

$$I_n = \sum_{m' \neq m} G_m P_{m',n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2 + \sum_{i \neq n}^{N_{RF}} \sum_{j=1}^{|S_n|} G_m P_{j,i}|\hat{\mathbf{h}}_{m,i}^H \boldsymbol{\psi}_i|^2. \tag{36}$$

Suppose $z = cY + d$, where Y is a Gaussian random variable with mean μ_Y and variance σ_Y^2, and c and d are both nonrandom constants. Then z is a random variable with mean $\mu_z = c\mu_Y + d$ and variance $\sigma_z^2 = c^2\sigma_Y^2$ [60]. Therefore, $P_L\left(r_{m,n}\right)$ becomes a Gaussian random variable with mean $P_L\left(r_0\right) + 10\bar{n}\log_{10}\left(\frac{r_{m,n}}{r_0}\right)$ and variance ξ^2.

As such,

$$\begin{aligned} P_A\left(r_{m,n}\right) &= \mathrm{P}_r\left(\frac{G_m P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2}{\sigma^2 T_{\min}} - \frac{I_n}{\sigma^2} > P_L\left(r_{m,n}\right)\right) \\ &= \mathrm{P}_r\left(P_L\left(r_{m,n}\right) < a\right), \end{aligned} \tag{37}$$

where

$$a = \frac{G_m P_{m,n}|\hat{\mathbf{h}}_{m,n}^H \boldsymbol{\psi}_n|^2}{\sigma^2 T_{\min}} - \frac{I_n}{\sigma^2}. \tag{38}$$

Subsequently,

$$P_A\left(r_{m,n}\right) = \mathrm{Q}\left(\frac{P_L\left(r_0\right) + 10\bar{n}\log_{10}\left(\frac{r_{m,n}}{r_0}\right) - a}{\xi}\right), \quad (39)$$

where $\mathrm{Q}(\cdot)$ is the Q-function [60].

Therefore, the cell coverage percentage of a cell of radius R is given as

$$\begin{aligned} C_{\mathrm{cov}} &= \frac{1}{\pi R^2}\int_0^{2\pi}\int_0^{R} P_A\left(r_{m,n}\right) r_{m,n}\, dr_{m,n}\, d\theta \\ &= \frac{1}{\pi R^2}\int_0^{2\pi}\int_0^{R} \mathrm{Q}\left(\frac{L-a}{\xi}\right) r_{m,n}\, dr_{m,n}\, d\theta, \end{aligned} \quad (40)$$

where

$$L = P_L\left(r_0\right) + 10\bar{n}\log_{10}\left(\frac{r_{m,n}}{r_0}\right). \quad (41)$$

As such,

$$\begin{aligned} C_{\mathrm{cov}} &= \frac{2}{R^2}\int_0^{R} \mathrm{Q}\left(\frac{L-a}{\xi}\right) r_{m,n}\, dr_{m,n} \\ &= \frac{2}{R^2}\int_0^{R} r_{m,n}\mathrm{Q}\left(\frac{L-a}{\xi}\right) dr_{m,n} \\ &= \frac{2}{R^2}\int_0^{R} r_{m,n}\mathrm{Q}\left(x + y\ln\left(\frac{r_{m,n}}{R}\right)\right) dr_{m,n}, \end{aligned} \quad (42)$$

where

$$x = \frac{P_L\left(r_0\right) + 10\bar{n}\log_{10}\left(\frac{R}{r_0}\right) - a}{\xi}, \quad (43)$$

when the mth PUT in the nth beam is at the cell edge and

$$y = \frac{10\bar{n}\log_{10}(e)}{\xi}. \quad (44)$$

Applying integration by parts yields a closed form:

$$C_{\mathrm{cov}} = \mathrm{Q}(-x) + \exp\left[2\left(\frac{1-xy}{y^2}\right)\right]\mathrm{Q}\left(\frac{2}{y} - x\right). \quad (45)$$

Hence, the proof of the proposition is completed.

APPENDIX B
PROOF OF NONCONVEXITY OF CONSTRAINTS C3 AND C4

From (21), constraints C3 and C4 are inequality functions. Therefore, to obtain the gradient and hessian of the objective function and these constraints, we used the Lagrangian method [61]. Hence, the Lagrangian function is given as

$$\begin{aligned} &L\left(P, \gamma^c_{m,n}, \gamma^p_{m,n}\right) \\ &= \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\left(a_{m,n} + \log_2\left(1+\gamma^p_{m,n}\right)\right) \\ &\quad + \lambda_1\left(\gamma^c_{m,n} - \frac{v_{m,n}P_c|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_c|^2}{\left(\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} v_{m,n}P_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2 + \sigma^2\right)}\right) \\ &\quad + \lambda_2\left(\gamma^p_{m,n} - \frac{v_{m,n}P_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2}{\left(I_{m'} + \sigma^2\right)}\right), \end{aligned} \quad (46)$$

where λ_1 and λ_2 are Lagrangian multipliers corresponding to constraints C3 and C4, respectively [62]. Taking the partial derivatives of the Lagrangian function with respect to $\gamma^c_{m,n}$ and $\gamma^p_{m,n}$, the gradient is given as

$$\nabla L\left(P, \gamma^c_{m,n}, \gamma^p_{m,n}\right) = \begin{bmatrix} -\lambda_1 v_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_c|^2 \\ P'_{m,n} - \lambda_2 v_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2 \\ \gamma^{c'}_{m,n} + \lambda_1\sigma^2 \\ \gamma^{p'}_{m,n} + \lambda_2 I_{m'} + \lambda_2\sigma^2 \end{bmatrix} \quad (47)$$

where $P'_{m,n} = \lambda_1\gamma^c_{m,n}\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} v_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2 + \lambda_2\gamma^p_{m,n}I'_{m'}$, $\gamma^{c'}_{m,n} = \lambda_1\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} v_{m,n}P_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2$, and $\gamma^{p'}_{m,n} = \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}\left(\frac{1}{(1+\gamma^p_{m,n})\ln 2}\right)$.

Taking the partial derivatives of the gradient, the hessian is given as

$$\begin{aligned} &\nabla^2 L\left(P, \gamma^c_{m,n}, \gamma^p_{m,n}\right) \\ &= \mathbf{H}\left(P, \gamma^c_{m,n}, \gamma^p_{m,n}\right) \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\partial^2 L}{\partial P_{m,n}\partial\gamma^c_{m,n}} & \lambda_2 I'_{m'} \\ 0 & \frac{\partial^2 L}{\partial\gamma^c_{m,n}\partial P_{m,n}} & 0 & 0 \\ 0 & \lambda_2 I'_{m'} & 0 & \frac{-\ln 2}{\left(\left(1+\gamma^p_{m,n}\right)\ln 2\right)^2} \end{bmatrix} \end{aligned} \quad (48)$$

where $\partial^2 L/\partial P_{m,n}\partial\gamma^c_{m,n} = \lambda_1\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} v_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2$ and $\partial^2 L/\partial\gamma^c_{m,n}\partial P_{m,n} = \lambda_1\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} v_{m,n}|\hat{\mathbf{h}}^H_{m,n}\boldsymbol{\psi}_n|^2$.

From (48), it is clear that $\nabla^2 L\left(P, \gamma^c_{m,n}, \gamma^p_{m,n}\right) \not\succeq 0$ [63]. Hence, the problem in (21) is nonconvex due to constraints C3 and C4.

Hence, the proof of the nonconvexity of constraints C3 and C4 is completed.

REFERENCES

[1] R. Jiao and L. Dai, "On the max-min fairness of beamspace MIMO-NOMA," *IEEE Trans. Signal Process.*, vol. 68, pp. 4919–4932, 2020.

[2] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[3] Y. Wang, H. Lu, D. Zhao, Y. Deng, and A. Nallanathan, "Intelligent reflecting surface-assisted mmWave communication with lens antenna array," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 1, pp. 202–215, Mar. 2022.

[4] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[5] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611–4624, Jul. 2016.

[6] D. Alimo and M. Saito, "Beam selection for mm-wave massive MIMO systems using ACO & combined digital precoding under hybrid transceiver architecture," *IEICE Commun. Exp.*, vol. 9, no. 6, pp. 170–175, 2020.

[7] D. Alimo, M. Hamamura, and S. R. Sabuj, "Threshold-based user-assisted cooperative relaying in beamspace massive MIMO NOMA systems," *Sensors*, vol. 22, no. 19, p. 7445, Sep. 2022.

[8] W. Hao, G. Sun, Z. Chu, P. Xiao, Z. Zhu, S. Yang, and R. Tafazolli, "Beamforming design in SWIPT-based joint multicast-unicast mmWave massive MIMO with lens-antenna array," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1124–1128, Aug. 2019.

[9] X. Gao, L. Dai, Z. Chen, Z. Wang, and Z. Zhang, "Near-optimal beam selection for beamspace mmWave massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1054–1057, May 2016.

[10] R. Pal, A. K. Chaitanya, and K. V. Srinivas, "Low-complexity beam selection algorithms for millimeter wave beamspace MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 768–771, Apr. 2019.

[11] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[12] L. Zhu, Z. Xiao, X.-G. Xia, and D. O. Wu, "Millimeter-wave communications with non-orthogonal multiple access for B5G/6G," *IEEE Access*, vol. 7, pp. 116123–116132, 2019.

[13] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu, and M. Shikh-Bahaei, "Energy efficient rate splitting multiple access (RSMA) with reconfigurable intelligent surface," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[14] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, pp. 2123–2129, 2016.

[15] D. Xu, "Device scheduling and computation offloading in mobile edge computing networks: A novel NOMA scheme," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 9071–9076, Jun. 2024.

[16] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, and G. Caire, "Is NOMA efficient in multi-antenna networks? A critical look at next generation multiple access techniques," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1310–1343, 2021.

[17] X. Ou, X. Xie, H. Lu, and H. Yang, "Resource allocation in MU-MISO rate-splitting multiple access with SIC errors for URLLC services," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 229–243, Jan. 2023.

[18] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[19] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.

[20] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2073–2126, 4th Quart., 2022.

[21] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[22] A. Mishra, Y. Mao, L. Sanguinetti, and B. Clerckx, "Rate-splitting assisted massive machine-type communications in cell-free massive MIMO," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1358–1362, Jun. 2022.

[23] M. Katwe, K. Singh, B. Clerckx, and C.-P. Li, "Rate-splitting multiple access and dynamic user clustering for sum-rate maximization in multiple RISs-aided uplink mmWave system," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7365–7383, Nov. 2022.

[24] H. Cho, B. Ko, B. Clerckx, and J. Choi, "Cooperative rate-splitting for enhanced THz frequency coverage," in *Proc. IEEE 34th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2023, pp. 1–6.

[25] C. Kong and H. Lu, "Cooperative rate-splitting multiple access in heterogeneous networks," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2807–2811, Oct. 2023.

[26] N. T. Nguyen and K. Lee, "Cell coverage extension with orthogonal random precoding for massive MIMO systems," *IEEE Access*, vol. 5, pp. 5410–5424, 2017.

[27] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[28] L. Li, K. Chai, J. Li, and X. Li, "Resource allocation for multi-carrier rate-splitting multiple access system," *IEEE Access*, vol. 8, pp. 174222–174232, 2020.

[29] Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5988–6002, Sep. 2021.

[30] W. Wang, L. Li, G. Deng, and J. Li, "A joint multiservice transmission scheme for RSMA-aided cell-free mMIMO system," *IEEE Commun. Lett.*, vol. 27, no. 2, pp. 591–594, Feb. 2023.

[31] C. K. Thomas, B. Clerckx, L. Sanguinetti, and D. Slock, "A rate splitting strategy for mitigating intra-cell pilot contamination in massive MIMO," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[32] A. Papazafeiropoulos, B. Clerckx, and T. Ratnarajah, "Mitigation of phase noise in massive MIMO systems: A rate-splitting approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[33] A. Papazafeiropoulos and T. Ratnarajah, "Rate-splitting robustness in multi-pair massive MIMO relay systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5623–5636, Aug. 2018.

[34] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access: A new frontier for the PHY layer of 6G," in *Proc. IEEE 92nd Veh. Technol. Conf. (VTC-Fall)*, Nov. 2020, pp. 1–7.

[35] M. Shahrbaf Motlagh, S. Majhi, P. Mitran, and H. Ochiai, "On rate-splitting with non-unique decoding in multi-cell massive MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 5929–5945, Sep. 2022.

[36] G. Arora and A. Jaiswal, "Zero SIC based rate splitting multiple access technique," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2430–2434, Oct. 2022.

[37] R. Chen, F. Cheng, J. Lin, L. Liang, and Y. Sun, "Performance analysis of rate splitting multiple access based vortex wave communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1570–1574, Aug. 2022.

[38] E. Sadeghabadi and S. D. Blostein, "RSMA precoding design based on interference nulling and sum rate upper bound," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 4091–4104, Jul. 2023.

[39] H. Niu, Z. Lin, K. An, J. Wang, G. Zheng, N. Al-Dhahir, and K.-K. Wong, "Active RIS assisted rate-splitting multiple access network: Spectral and energy efficiency tradeoff," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1452–1467, May 2023.

[40] Y. Tong, D. Li, Z. Yang, Z. Xiong, N. Zhao, and Y. Li, "Outage analysis of rate splitting networks with an untrusted user," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2626–2631, Feb. 2023.

[41] S. K. Singh, K. Agrawal, K. Singh, Y.-M. Chen, and C.-P. Li, "Performance analysis and optimization of RSMA enabled UAV-aided IBL and FBL communication with imperfect SIC and CSI," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3714–3732, Jun. 2022.

[42] S. Lee, S. Park, J. Park, and J. Choi, "Rate-splitting multiple access precoding for selective security," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2023, pp. 1–5.

[43] B. Clerckx, Y. Mao, E. A. Jorswieck, J. Yuan, D. J. Love, E. Erkip, and D. Niyato, "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1265–1308, May 2023.

[44] H. Joudeh and B. Clerckx, "A rate-splitting strategy for max-min fair multigroup multicasting," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2016, pp. 1–5.

[45] R. K. Ahiadormey and K. Choi, "Performance analysis of rate splitting in massive MIMO systems with low resolution ADCs/DACs," *Appl. Sci.*, vol. 11, no. 20, p. 9409, Oct. 2021.

[46] N. T. Nguyen and K. Lee, "Coverage and cell-edge sum-rate analysis of mmWave massive MIMO systems with ORP schemes and MMSE receivers," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5349–5363, Oct. 2018.

[47] A. Arsal, M. R. Civanlar, and M. Uysal, "Coverage analysis of downlink MU-MIMO cellular networks," *IEEE Commun. Lett.*, vol. 25, no. 9, pp. 2859–2863, Sep. 2021.

[48] A. A. Lone, A. K. Gupta, H. S. Dhillon, and S. Sharma, "Coverage and rate in MIMO cellular networks with location-aware transmission rank selection," *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2026–2030, Oct. 2022.

[49] D. G. Cileo, N. Sharma, and M. Magarini, "Coverage, capacity and interference analysis for an aerial base station in different environments," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2017, pp. 281–286.

[50] S. Tang, Z. Ma, M. Xiao, and L. Hao, "Hybrid transceiver design for beamspace MIMO-NOMA in code-domain for mmWave communication using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2118–2127, Sep. 2020.
[51] S. Zhong and X. Wang, "Wireless power transfer by beamspace large-scale MIMO with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1390–1403, Feb. 2019.
[52] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.
[53] G. Zhao and Q. Liang, "Outage analysis of massive MIMO system," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 354–359.
[54] M. Grant and S. Boyd, "CVX: MATLAB software for disciplined convex programming, version 2.1," Rep., 2014.
[55] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, nos. 1–3, pp. 193–228, Nov. 1998.
[56] X. Xie, J. Liu, J. Huang, and S. Zhao, "Ergodic capacity and outage performance analysis of uplink full-duplex cooperative NOMA system," *IEEE Access*, vol. 8, pp. 164786–164794, 2020.
[57] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.
[58] A. R. Flores and R. C. de Lamare, "Robust and adaptive power allocation techniques for rate splitting based MU-MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4656–4670, Jul. 2022.
[59] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath Jr., "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
[60] A. Grami, *Probability, Random Variables, Statistics, and Random Processes: Fundamentals & Applications*. Hoboken, NJ, USA: Wiley, 2019.
[61] J. A. Snyman and D. N. Wilke, *Practical Mathematical Optimization*. Springer, 2018.
[62] S. R. Sabuj, D. K. P. Asiedu, K.-J. Lee, and H.-S. Jo, "Delay optimization in mobile edge computing: Cognitive UAV-assisted eMBB and mMTC services," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1019–1033, Jun. 2022.
[63] S. R. Sabuj, S. Ahmed, and H.-S. Jo, "Multiple CUAV-enabled mMTC and URLLC services: Review of energy efficiency and latency performance," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 3, pp. 1369–1382, Sep. 2023.

DAVID ALIMO (Student Member, IEEE) received the B.Sc. degree (Hons.) in engineering from Future University, Sudan, in 2006, and the M.Eng. degree in electrical and electronics engineering from the University of the Ryukyus, Japan, in 2020. He is currently pursuing the Ph.D. degree in engineering with the Graduate School of Engineering, Kochi University of Technology, Japan. From 2007 to 2010, he was a Faculty Member with the Telecommunication Engineering Department, Future University. From 2010 to 2012, he was a Field Operations and Maintenance Engineer with ZTE Corporation, Sudan. He was a PS Core and MPBN Engineer with MTN South Sudan, from 2013 to 2017. His research interests include wireless communications, with an emphasis on mmWave communications, cooperative communications, intelligent reflecting surfaces, and adaptive signal processing for B5G/6G communication systems.

MASANORI HAMAMURA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagaoka University of Technology, Japan, in 1993, 1995, and 1998, respectively. He is a Professor with the School of Informatics, Kochi University of Technology, Japan. From 1998 to 2000, he was a Research Fellow with Japan Society for the Promotion of Science. From 1998 to 1999, he was a Visiting Researcher with the Centre for Telecommunications Research, King's College London, U.K. From 2017 to 2021, he was the Dean of the School of Informatics and Graduate School of Informatics, Kochi University of Technology. From 2020 to 2021, he was the Chair of IEEE Shikoku Section, and the IEICE Technical Committee on Wideband Systems (WBS), from 2020 to 2022. He has been serving as an associate editor for the regular section; and a Guest Associate Editor, a Guest Editor, and the Guest Editor-in-Chief for the Special Section on Wideband Systems of *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*.

SAIFUR RAHMAN SABUJ (Senior Member, IEEE) was born in Bangladesh. He received the B.Sc. degree in electrical, electronic and communication engineering from Dhaka University, Bangladesh, in 2007, the M.Sc. degree in information and communication technology from the Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology, Bangladesh, in 2011, and the Ph.D. degree in communication engineering from the Graduate School of Engineering, Kochi University of Technology, Japan, in 2017. He is currently with the Electrical and Electronic Engineering Department, Brac University, Bangladesh, as an Associate Professor. From 2020 to 2022, he was a Postdoctoral Research Fellow with the Electronic Engineering Department, Hanbat National University, South Korea. From 2008 to 2013, he was a Faculty Member of the Green University of Bangladesh; Metropolitan University, Sylhet; and Bangladesh University. His research interests include MIMO-OFDM/NOMA, cognitive radio, the Internet of Things, relay networks, unmanned aerial vehicle, and machine-to-machine for wireless communications.