

Millimeter Wave Massive MIMO Downlink Per-Group Communications with Hybrid Linear Precoding

T. Ketseoglou, *Senior Member, IEEE*, M.C. Valenti, *Fellow, IEEE*, and
E. Ayanoglu, *Fellow, IEEE*

Abstract

We address the problem of analyzing and classifying in groups the downlink channel environment in a millimeter-wavelength cell, accounting for path loss, multipath fading, and User Equipment (UE) blocking, by employing a hybrid propagation and multipath fading model, thus using accurate inter-group interference modeling. The base station (BS) employs a large Uniform Planar Array (UPA) to facilitate massive Multiple-Input, Multiple-Output (MIMO) communications with high efficiency. UEs are equipped with a single antenna and are distributed uniformly within the cell. The key problem is analyzing and defining groups toward precoding. Because balanced throughput is desired between groups, Combined Frequency and Spatial Division and Multiplexing (CFSDM) is found to be necessary. We show that by employing three or four subcarrier frequencies, depending on the number of UEs in the cell and the cell range, the UEs can be efficiently separated into high throughput groups, with each group employing Virtual Channel Model Beams (VCMB) based inner precoding, followed by efficient Multi-User Multiple-Input Multiple-Output (MU-MIMO) outer precoders. For each group, we study three different sub-grouping methods offering different advantages. We show that the improvement offered by Zero-Forcing Per-Group Precoding (ZF-PGP) over Zero-Forcing Precoding (ZFP) is very high. In addition, for medium-correlation channels, it is shown that ZF-PGP performance is near to the ideal one offered by a Virtual Additional Antenna Concept PGP (VAAC-PGP). Finally, a new technique for power allocation among different Per-Group Precoding within Groups (PGP-WG) groups is proposed, called Optimized Per-Group Power Allocation (OPGPA), which allows for high power efficiency with equal throughput among all UEs.

T. Ketseoglou is with the Electrical and Computer Engineering Department, California State Polytechnic University, Pomona, USA (e-mail: tketseoglou@cpp.edu). M. Valenti is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA (email: mvalenti@wvu.edu). E. Ayanoglu is with the Electrical Engineering and Computer Science Department, University of California, Irvine, USA (e-mail: ayanoglu@uci.edu). This work was partially supported by NSF grant 1547155.

I. INTRODUCTION

Because it offers a wide spectrum that can support short-range high-rate wireless connectivity [1], millimeter-wavelength (mmWave) communication is an attractive solution for future wireless applications, including massive Multiple-Input Multiple-Output (MIMO) applications [2]–[5]. Downlink input-output mutual information maximizing (IOMIM) linear precoding with finite-alphabet inputs, e.g., by employing Quadrature Amplitude Modulation (QAM), has been extensively studied [6]–[13] due to its potential to offer high data rate collectively. However, all existing studies have focused on multipath fading without considering essential propagation effects in mmWave communications such as User Equipment (UE) blocking, path loss, and varying fading scaling factors. The latter effects have been modeled in [14], [15] for mmWave device-to-device communications in a flexible way with success.

In this paper, we apply the model proposed in [14], [15] for capturing mmWave propagation effects and combine it with UE grouping techniques. This combination allows us to analyze the potential of separating users in quasi-orthogonal groups, when the number of total Uniform Planar Array (UPA) elements at the Base Station (BS) is large, i.e., in massive MIMO, in order to improve performance and simplify complexity. Furthermore, IOMIM linear precoding techniques tend to offer varying throughput to different UEs depending on their received signal-to-noise ratio (SNR); i.e., a varying quality-of-service (QoS) among UEs. This problem has not been studied in detail before, although it has the potential to support future wireless communication applications. Due to channel correlation and received power variation among different UEs, we propose assigning different subcarrier frequencies through Orthogonal Division Multiplexing (OFDM) to different UE groups. This method was developed in [9], where it is called Combined Frequency and Spatial Division Multiplexing (CFSDM), and is critical to achieving a balanced QoS to all UEs in a cell. Furthermore, in order to improve the power efficiency in each CFSDM group, we also propose Optimized Per-Group Power Allocation (OPGPA), a new method that achieves significant power savings at the BS while simultaneously achieving equal QoS to all UEs in a group. We furthermore combine IOMIM with Joint Spatial Division and Multiplexing with Finite Alphabets (JSDM-FA), which imposes orthogonality between groups, but has not been studied in realistic deployments, e.g., a mmWave cell.

This paper addresses all these open issues while providing a comparison of various downlink precoding techniques. We first study different sub-grouping (SBG) forming techniques within

each CFSDM subcarrier, in order to achieve high throughput with low complexity. These SBG techniques form the inner precoders of each group. To meet the goal of forming subgroups, we exploit the Virtual Channel Model Beam (VCMB), which was originally presented in [9], [16]. VCMBs are created by projecting the actual UE channels to a Discrete Fourier Transform (DFT) type basis that helps exploit the channel's spatial domain characteristics. We propose and analyze three techniques for SBG, including JSMD-FA together with a careful inter-sub-group interference analysis. We then compare the performance of two types of outer precoding: a) Zero-Forcing Precoding (ZFP) [17]–[19], and b) Zero-Forcing Per-Group Precoding (ZF-PGP). Due to high channel correlation between different UEs, ZFP gains evaporate rapidly and a Per-Group Precoding within Groups (PGP-WG) precoder performs much better as we demonstrate.

The contributions of this paper can be summarized as follows:

- 1) It employs a realistic mmWave communications model for the massive MIMO downlink, that includes random UE blocking, with the path-loss exponent and multipath-fading distribution dependent on the blocking state.
- 2) It presents a comprehensive approach to dividing UEs in CFSDM groups based on their spatial and power features, then subdividing groups in sub-groups by SBG in order to improve performance and lower the system complexity.
- 3) It presents results for three types of SBG, including a detailed inter-sub-group interference analysis in the case of JSMD-FA.
- 4) It shows that due to the debilitating impact of mmWave channels, the outer precoder in each group faces high channel correlation that leads to very poor performance of ZFP.
- 5) It demonstrates that in general JSMD-FA suffers a very high performance loss over Total Grouping (TG) and Simple Grouping (SG), due to its inherent significant inter-sub-group interference and high inter-UE channel correlation.
- 6) It shows the very high gains of ZF-PGP over ZFP for a wide range of SNR.
- 7) It develops OPGPA, which is a new power-allocation strategy that allows for very high power efficiency.

Notation: We use small bold letters for vectors and capital bold letters for matrices. \mathbf{A}^T , \mathbf{A}^H , \mathbf{A}^* , $\mathbf{A}_{:,i}$, $\mathbf{A}_{i,:}$, and $\mathbf{A}_{k,l}$ denote the transpose, Hermitian, conjugate, complex conjugate, column i , row i , and row k , column l element of matrix \mathbf{A} , respectively. Further, $\text{tr}(\mathbf{A})$ denotes the trace of a (square) matrix \mathbf{A} . \mathbf{S}^T denotes a selection matrix, i.e., of size $k \times n$ with $k < n$ consists of

rows equal to different unit row vectors \mathbf{e}_i where the row vector element i is equal to 1 in the i th position and is equal to 0 in all other positions, the specific \mathbf{e}_i vectors used are defined by the desired selection. \mathbf{F}_N denotes the DFT matrix of order N , $\text{diag}[x_1, \dots, x_k]$ is the diagonal matrix with main diagonal equal to vector $[x_1, \dots, x_k]^T$, and \mathbf{I}_N denotes an identity matrix of dimension $N \times N$. We use $\mathbf{h}_{d,g,k,n}$ for the downlink channel of user k 's antenna n in group g . \mathbf{H}_g is the downlink channel of group g , while $\tilde{\mathbf{H}}_{g,v}$ is its projection to the Virtual Channel Model (VCM) basis.

II. SYSTEM MODEL AND PROBLEM STATEMENT

A. MmWave Channel Model Employing Random UE Blocking

We assume a dense population of UEs that are uniformly distributed within a cell [15]. The BS UPA has height h and employs an x (horizontal) and z (vertical) orientation.¹ The total number of UE in the cell is N_{UE} . Each UE employs a single, uniformly radiating antenna. By employing Time Division Duplexing (TDD), the downlink channels will be reciprocal to the uplink ones. The channel between UE n ($1 \leq n \leq N_{UE}$) and the BS is denoted by \mathbf{h}_n . With $P = 1$ multipath components [9], due to mmWave conditions [14], [15], we get

$$\mathbf{h}_n = \tilde{g}_n (\mathbf{a}_z(\theta_n) \otimes \mathbf{a}_x(\theta_n, \phi_n)), \quad (1)$$

where \otimes denotes Kronecker matrix product, $\tilde{g}_n = g_n \exp(j2\pi b_n)$ is the multipath fading complex coefficient of amplitude $g_n = |\tilde{g}_n|$ and phase b_n , uniformly distributed in $[0, 2\pi]$, θ_n, ϕ_n represent UE n 's ($1 \leq n \leq N_{UE}$) elevation and azimuth angle, respectively,

$$\begin{aligned} \mathbf{a}_x(\theta_n, \phi_n) &= [1, \exp(-j2\pi D \sin(\theta_n) \cos(\phi_n)), \dots, \exp(-j2\pi D(N_{u,x} - 1) \sin(\theta_n) \cos(\phi_n))]^T, \\ \mathbf{a}_z(\theta_n) &= [1, \exp(-j2\pi D \cos(\theta_n)), \dots, \exp(-j2\pi D(N_{u,z} - 1) \cos(\theta_n))]^T, \end{aligned} \quad (2)$$

with $D = \frac{d}{\lambda}$, d being the distance between adjacent antenna elements, λ the wavelength, and $N_{u,x}, N_{u,z}$ representing the number of elements of the UPA in the x and z direction, respectively.

¹Any UPA or Uniform Linear Array (ULA) orientation would result in similar results and could be used with success in our model.

The total antenna elements at the BS equal to $N_T = N_{u,x}N_{u,z}$ (the number of rows in \mathbf{h}_n). The instantaneous received SNR at UE n ($1 \leq n \leq N_{UE}$) under the breakpoint model [20] is

$$\text{SNR}_n = g_n^2 \text{SNR}_0 \left(\frac{R_{break}}{R_n} \right)^k, \quad (3)$$

for $R_n \geq R_{break}$, where R_n is the distance between the UPA and the UE, R_{break} is the break distance [20], $\text{SNR}_0 = \frac{E_{s,0}}{N_0}$ is the SNR at $R_n = R_{break}$, where $E_{s,0}$ is the symbol energy and N_0 is the one-sided noise Power Spectral Density (PSD), k is the path-loss exponent, and g_n^2 is the power gain of the fading. As in [14], [15], if UE n ($1 \leq n \leq N_{UE}$) is blocked, it is non-line-of-sight (NLOS) and we use $k = k_{NLOS}$ and g_n is Nakagami with $m = m_{NLOS}$, while when the UE is not blocked, it is line-of-sight (LOS) and we apply $k = k_{LOS}$ and g_n is Nakagami with $m = m_{LOS}$, with $k_{NLOS} > k_{LOS}$ and $m_{NLOS} < m_{LOS}$.

B. Problem Statement

From [16], an equivalent cell downlink channel receiving equation, after normalization and encompassing both large-scale and small-scale effects [20], i.e., propagation loss and multipath fading, respectively, together with noise effects (including Additive White Gaussian Noise (AWGN) and Multiple-Access Interference (MAI)) can be written in the virtual domain.² Toward this end, let's start by defining \mathbf{y}_d to be the downlink received vector over all users and antennas of size $N_{UE} \times 1$, \mathbf{G} , and \mathbf{x}_d to be the $N_{UE} \times 1$ vector of transmitted symbols³ drawn independently from a QAM constellation. Also define the unit SNR downlink virtual channel matrix to be $\mathbf{H}_{u,v}$, of size $N_{UE} \times N_T$ downlink for all N_{UE} UEs with its rows being the corresponding UE Hermitian of the uplink channel vector, employing N_T receiving antennas at the BS, [9], [16]. Then, the downlink receiving equation is as follows

$$\mathbf{y}_d = \sqrt{\text{SNR}_0} \mathbf{H}_{d,v} \mathbf{G} \mathbf{x}_d + \mathbf{n}_{d,AWGN} + \mathbf{n}_{d,MAI}, \quad (4)$$

where $\mathbf{n}_{d,AWGN}$ represents the complex circularly symmetric Gaussian noise of mean zero and variance per component $\sigma_d^2 = 1$ (after normalization by dividing the original receiving equation by the standard deviation of the AWGN noise), and $\mathbf{n}_{d,MAI}$ represents the multiple-

²This is the channel representation in the VCM basis, also called the beam-domain representation in the literature .

³We assume that there is one symbol per receiving antenna in (4), for simplicity.

access interference (MAI) between sub-groups, present only in the JSDM-FA case. We focus on the input-output mutual information $I(\mathbf{x}_d; \mathbf{y}_d)$ maximizing downlink precoding problem, where we assume that the channel is known at both the transmitter and the receiver(s),⁴ which can be cast as

$$\begin{aligned} & \underset{\mathbf{G}}{\text{maximize}} && I(\mathbf{x}_d; \mathbf{y}_d) \\ & \text{subject to} && \text{tr}(\mathbf{G}\mathbf{G}^H) = N_{UE}, \end{aligned} \quad (5)$$

where the constraint is due to keeping the total power transmitted from the BS to all downlink users equal to the total power without precoding. It is well-known that this problem is complexity-burdened and thus grouping UEs offers a solution to this [16], [21]–[23]. However, in this paper we aim at offering balanced throughput to UEs, thus additional methods are needed to achieve this goal, as described below. Furthermore, employing data symbols from a finite-alphabet constellation, e.g., QAM in (4), makes the problem more realistic, but at the same time more complex [11]. For example, in order to solve the problem the mutual information $I(\mathbf{x}_d; \mathbf{y}_d)$ needs to be computed and this can be performed using the Gauss-Hermite (GH) quadrature approximation. For the MIMO channel model presented in (4), the GH approximation toward evaluating $I(\mathbf{x}_d; \mathbf{y}_d)$ is presented in Lemma 1 of [6] and is also described for completeness in Appendix A.

C. UE Grouping and Sub-grouping

a) Pre-selection of VCMBs

By projecting the uplink UPA response vector to the complete orthonormal basis $\mathbf{B}_{VCM} = (\mathbf{F}_{N_{u,z}} \otimes \mathbf{F}_{N_{u,x}})$ where \mathbf{F}_N represents the Discrete Fourier Transform (DFT) matrix of size N , [9] showed that with a large number of array elements, e.g., $N_T \sim 100$ and with equal elements per dimension ($N_{u,x} = N_{u,z}$), this projection achieves a sparse representation of the UE channels with only a few components from the columns of $\mathbf{B} = (\mathbf{F}_{N_{u,z}} \otimes \mathbf{F}_{N_{u,x}})$ needed. Since only a few columns (VCMBs) of the orthonormal matrix $(\mathbf{F}_{N_{u,z}} \otimes \mathbf{F}_{N_{u,x}})$ are needed to characterize each channel, a significant dimensionality reduction is available by employing the VCMBs. Furthermore, for spatially distant UEs, different users form quasi-orthogonal groups of

⁴In [16] we show that estimated channels can be used successfully instead of the perfect channel knowledge assumed here.

non-intersecting VCMBs. In addition, VCMBs can be used to derive many useful spatial-domain features for the entirety of downlink channels in the cell. Here, we use the VCMB in the cell due to its spatial-feature-revealing capabilities. First, we extract the most “loaded” VCMBs in the cell, by determining the $N_{V,INIT}$ VCMBs that carry the most instantaneous power to UEs. The parameter $N_{V,INIT}$ is determined by the percentage of overall power in the cell we aim at capturing. Due to the nature of the VCMB structure, only a fraction of the total N_T VCMBs are needed to guarantee that more than, e.g., 90%, of the total power is captured. This VCMB selection phase is called pre-selection (PS). The pseudo code for the PS algorithm is shown in Algorithm 1.

b) CFSDM-based Grouping for Balanced QoS

In mmWave communications, the existence of NLOS UEs which suffer a significant additional propagation power loss requires placing NLOS UEs in separate frequency sub-carriers, then employing higher power to improve the NLOS UE throughput, better balancing the throughput with that of the LOS UEs. Furthermore, due to high spatial correlation of UE channels, some UEs, although in the LOS class, will also experience low QoS. This QoS imbalance can be mitigated by adding another separate set of sub-carriers to accommodate these UEs, thus offering a solution to balancing the cell QoS. Thus, by employing a total of three or four sets of subcarriers in the form of CFSDM [9], depending on the number of UEs and the size of the cell, we can achieve a balanced QoS. Thus, there will be N_G CFSDM groups in the cell, with $N_G = 3$ or 4. The first group G_1 is the NLOS group formed by aggregating all the NLOS UEs, while the other groups, G_2, \dots, G_{N_G} denote LOS groups. The corresponding number of UEs in each CFSDM group is denoted by N_1, N_2, \dots, N_{N_G} , respectively. The selection of UEs for each CFSDM group is based on maximum statistical decorrelation per group, i.e., each group presents relatively low correlation among its member UE channels.

c) SBG Techniques

After CFSDM grouping of UEs, there is additional opportunity with SBG for improved performance or lower complexity. It is important to mention that sub-grouping employs the same frequency for all subgroups in a group, i.e., spatial multiplexing takes place to improve performance in each group. After PS takes place, in the reduced dimension VCMB space comprising $N_{V,INIT}$ VCMBs, there are many alternatives one can use for further SG of different UEs in a

group. In this paper we consider the following three:

- 1) Employ all pre-selected VCMBs in a group, which is TG.
- 2) Select only the strongest VCMBs in the group, which is SG, resulting in a final number of VCMBs per group, $N_{V,FINAL} = N_{N_g}$, $g = 1, 2, \dots, N_{N_G}$, i.e., significantly smaller than $N_{V,INIT}$.
- 3) JSDM-FA for semi-orthogonal sub-groups [9], [16], which offers an $N_{V,FINAL}$ even smaller than SG, but in general it suffers intra-sub-group interference. Due to the intra-group MAI issue, JSDM-FA can be applied by dividing each of the CFSDM groups in sub-groups. In other words, for each CFSDM group G_k , $k = 1, 2, \dots, N_G$, we create N_{S_k} sub-groups denoted as $G_{k,l}$ with $l = 1, \dots, N_{S_k}$, which apply JSDM-FA, where the number of subgroups per group, N_{S_k} , depends on the VCMB strength in each group. We use $N_{G_{k,l}}$ to denote the number of UEs in sub-group $G_{k,l}$, $k = 1, 2, \dots, N_{N_G}$, $l = 1, \dots, N_{S_k}$. The selection of the UE members of each sub-group is based on the first few strongest VCMBs.

It is important to stress that during SG and JSDM-FA, VCMBs that are unused by some sub-groups are set off, i.e., no power is transmitted over these VCMBs by these sub-groups. If other subgroups employ these VCMBs, then the potential for MAI between subgroups arises. Thus, although in the SB scenario, there is no MAI between groups, in the JSDM-FA case, there might be some inter-sub-group MAI to the other subgroups that do have the affecting VCMBs unused is due to the lack of full orthogonality between JSDM-FA sub-groups. Thus, a careful calculation of the MAI between sub-groups in the JSDM-FA case is required. The details of this MAI calculation are presented in Appendix B.

Algorithm 1 PS algorithm

- 1: **for** $i = 1$ to N_T **do**
 - 2: calculate power of VCMB i , $P_i = ||H_{d,v}[:, i]||^2$;
 - 3: **end for**
 - 4: sort in descending order the vector $\mathbf{P} = [P_1 \ P_2 \ \dots \ P_{N_T}]$, resulting in a new sorted power list vector $\mathbf{P}_s = [P_{s_1} \ P_{s_2} \ \dots \ P_{s_{N_{V,INIT}}}]$ and descending-order sorted VCMB list vector \mathbf{v}_s
 - 5: select the first $N_{V,INIT}$ entries of \mathbf{P}_s , \mathbf{v}_s and denote them \mathbf{P}_{PS} , \mathbf{v}_{PS} , respectively and calculate the total power in \mathbf{P}_{PS} as percentage of the original total power, i.e.,

$$P_F = \frac{\sum_{i=1}^{N_{V,INIT}} P_{s_i}}{\sum_{i=1}^{N_T} P_i};$$
 - 6: re-arrange the selected \mathbf{v}_{PS} VCMBs per UE in descending channel power order, resulting in a $N_{UE} \times N_{V,INIT}$ matrix $\mathbf{M}_{V,PS}$ used in further processing
-

D. Efficient MU-MIMO Outer Precoding through ZF-PGP for Spatially Correlated Groups

After groups and sub-groups are selected and the pre-beamformer (inner precoder) is constructed, an MU-MIMO linear precoder (outer precoder) is deployed to offer individual and high data rate streams to each UE in a sub-group. References [21], [24] envisaged this type of hybrid precoding in order to achieve high throughput to each user in a group by employing a Zero-Forcing Precoder (ZFP). However, since all the UEs in a formed group possess spatial similarity, the channels within a group are highly correlated. Thus, ZFP results in low data rates and low spectral efficiency. Here we apply ZF-PGP [16] which combines the benefits of ZFP and Per-Group Precoding within Groups (PGP-WG) [25] in order to improve the performance, as illustrated in Fig. 1. ZF-PGP also employs the Virtual Additional Antenna Concept (VAAC) [16] that delivers two symbols to each UE, thus doubling the high SNR throughput of ZF-PGP over ZFP.

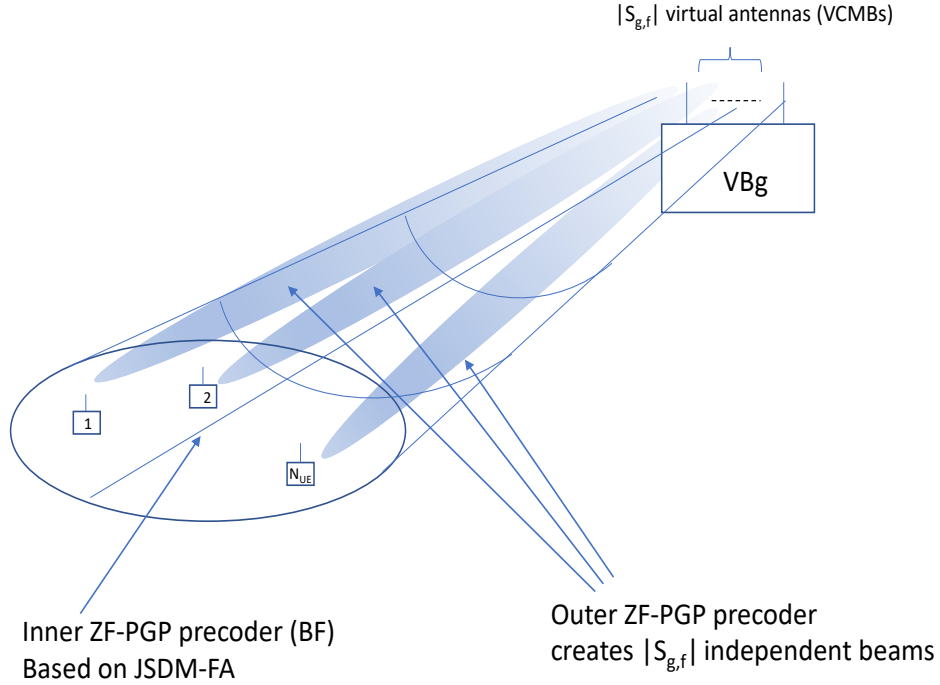


Fig. 1. Outer group precoder. After inner precoding, the outer ZF-PGP linear precoder creates $S_{g,f}$ independent data streams, one for each UE. $S_{g,f}$ denotes the final number of VCMBs in the sub-group.

E. Theoretical MU-MIMO Outer Precoding through VAAC-PGP for Comparison of Performance

When the cell radius grows, then there is additional power propagation loss incurred as well as additional correlation present to the channels of more distant UEs. Due to this, ZF-PGP can

perform poorly for these UEs. A modified, improved, higher-performance precoder is studied for these UEs. VAAC-PGP is a PGP-WG [25] in conjunction with the application of the Virtual Additional Antenna Concept (VAAC) and it can offer significant gains over ZF-PGP, albeit with higher complexity. Its higher complexity emanates from the fact that it requires knowledge of the overall group receiving vector and one of the downlink group channel's left singular vectors at each UE. Because of these mostly unrealistic demands, we study VAAC-PGP with the goal of providing a more fundamental understanding of the underpinnings of the channel correlation in the mmWave cell.

Let LOS group G_k , with $k = 2, 3, \dots, N_G$ employ the VAAC-PGP concept. The VAAC-PGP precoder adds one additional virtual antenna [16] per UE, similar to the ZF-PGP precoder, but it does not include a ZF component, i.e., it does not offer the benefits of ZF precoding. This means that the UEs involved in VAAC-PGP need to know (through side information sent by the BS), or estimate the channel's right singular values. An additional benefit of VAAC-PGP, besides higher throughput to stressed UEs, is that one can rotate the use of the channel's right singular values and then by-reassigning data, a throughput-balancing effect takes place. In other words, all UEs involved in the VAAC-PGP precoding have the same throughput.

The VAAC-PGP is derived as follows. Assume that a group g ($g = G_2, G_3, \dots, G_{N_G}$) applies VAAC-PGP. From [21], the equation for group g is

$$\tilde{\mathbf{y}}_g = \mathbf{H}_{g,v} \mathbf{P}_g \mathbf{x}_g + \mathbf{n}_g, \quad (6)$$

where $\mathbf{H}_{g,v}$ is the VCM group's downlink matrix of size $N_g \times N_g$, \mathbf{y}_g is the group's size N_g reception vector, \mathbf{P}_g is the $N_g \times 2 \cdot N_g$ precoding matrix, \mathbf{x}_g is the $2 \cdot N_g \times 1$ data symbol vector, and \mathbf{n}_g is the corresponding AWGN noise. For the data vector \mathbf{x}_g , we assume without loss of generality, that $\mathbf{x}_g = [x_{g,1,1} \ x_{g,1,2} \ \dots \ x_{g,N_g,1} \ x_{g,N_g,2}]^T$, where $x_{g,i,k}$ with $i = 1, 2, \dots, N_g$ and $k = 1, 2$ is the i th UE k th data symbol.

Then, the VAAC-PGP of the group solves the following optimization problem

$$\begin{aligned} & \underset{\mathbf{P}_g}{\text{maximize}} \quad I(\mathbf{x}_g; \mathbf{y}_g) \\ & \text{subject to} \quad \text{tr}(\mathbf{P}_g \mathbf{P}_g^H) = 2 \cdot N_{UE}, \end{aligned} \quad (7)$$

under the constraint that $\mathbf{P}_g = \mathbf{U}_{g,v} \mathbf{S}_P \mathbf{V}_P^H$, with $\mathbf{U}_{g,v}$ being the matrix of left singular vectors

of $\mathbf{H}_{g,v}$ of size $N_g \times N_g$, \mathbf{S}_P being the singular value matrix of the VAAC-PGP precoder of size $N_g \times N_g$, and \mathbf{V}_P of size $2N_g \times N_g$ is the matrix of right singular vectors of \mathbf{P}_g . The constraint of (7) is placed in order to guarantee that the total transmitted power before and after recoding stays the same. An equivalent way of expressing (7) is through the singular value decomposition (SVD) of $\mathbf{P}_g = \mathbf{U}_{g,v} \mathbf{S}_P \mathbf{V}_P^H$, as $\sum_{i=1}^{N_g} s_i^2 = 2 \cdot N_g$, with s_i being the i th singular value of \mathbf{S}_P . Due to applying PGP-WG, the matrix of the right singular vectors of the precoding matrix \mathbf{V}_P observes a block-diagonal structure [25] as follows

$$\mathbf{V}_P^H = \begin{bmatrix} \mathbf{v}_{p,1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_{p,2} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{v}_{p,3} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{v}_{p,N_g-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{v}_{N_g} \end{bmatrix}, \quad (8)$$

where $\mathbf{v}_{p,i}$, $i = 1, \dots, N_g$ are 1×2 row vectors of unit norm. As it is shown in Appendix C, the VAAC-PGP precoder has all its singular values equal to $\sqrt{2}$, i.e. $\mathbf{S}_P = \sqrt{2}\mathbf{I}_{N_g}$, due to the VAAC applied in the precoding process, i.e., the fact that the IOMIM optimal precoder with VAAC applied always needs all the possible power set to the useful antenna (see Appendix C). The UEs can decode their data by forming the inner product of the proper left singular vectors with the received vector of the group, \mathbf{y}_g .

F. Optimized Per-Group Power Allocation (OPGPA) for Equal QoS and Reduced Average Power

Due to the nature of PGP and ZF-PGP, a previously uninvestigated possibility is available toward reducing the average power of a group's precoding, by allocating different powers to different effective singular values employed in the ZF-PGP. We propose OPGPA in order to dramatically reduce the average power employed in a group, under medium channel correlation, to achieve equal throughput to all UEs in a group. OPGPA's premise is that for a specific group-wide pre-set throughput goal of I_S , with $(0 \leq I_S \leq 2\log_2(M))^5$ delivered to each UE in a group, based on ZF-PGP, the larger effective channel singular values defined in [16] will need lower SNR than the smaller ones in order to achieve this goal. In addition, due to the PGP part of ZF-PGP, the

⁵The reason for the 2 in the equation is due to the VAAC concept that sends two symbols per receiving antenna. Please see Appendix B.

power sent over one effective singular value, e.g., $s_{g,v,eff,m} \doteq \frac{1}{w_m} = \left(\sum_{m'=1}^{N_{d,g}} \frac{|(\mathbf{V}_{g,v})_{m,m'}|^2}{s_{g,v,m'}^2} \right)^{-1/2}$, where $m = 1, 2, \dots, N_g$ [16]. Assume that PGP group m employs an SNR at $r = 1$ m equal to SNR_0 , as per (4), then an equivalent reception model with ZF-PGP for the m th PGP group incorporating the VAAC model [16] is

$$\begin{bmatrix} y_{m,1} \\ y_{m,a} \end{bmatrix} = \begin{bmatrix} s_{g,v,eff,m}^{(\text{SNR}_0)} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{V}_{PGP,m}^H \begin{bmatrix} x_{m,1} \\ x_{m,a} \end{bmatrix} + \begin{bmatrix} n_{m,1} \\ n_{m,a} \end{bmatrix}, \quad (9)$$

where $s_{g,v,eff,m}^{(\text{SNR}_0)}$ is the m th channel's effective singular value under ZF-PGP for the value of SNR_0 employed (see (4)), the vector of the noise has variance per component equal to 1, and $\mathbf{V}_{PGP,m}$ represents the 2×2 PGP part right singular vector matrix of the ZF-PGP.

Under OPGPA, we allow weaker PGP groups to apply higher transmit power in order to achieve higher throughput, equal to the set acceptable one, I_S . This means that each input to the ZF-PGP in (10) is employing an appropriate power gain factor denoted as k_m , thus achieving an overall SNR for the m th PGP group of $k_m \cdot \text{SNR}_0$. This has no effect to the other PGP groups in the precoder, because in PGP-WG each group is orthogonal to the other groups by construction. When OPGPA applies, we call SNR_0 as the initial SNR to avoid confusion. The corresponding receiving equation under OPGPA is then

$$\begin{bmatrix} y_{m,1} \\ y_{m,a} \end{bmatrix} = \begin{bmatrix} s_{g,v,eff,m}^{(\text{SNR}_0)} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{V}_{PGP,m}^H \sqrt{k_m} \begin{bmatrix} x_{m,1} \\ x_{m,a} \end{bmatrix} + \begin{bmatrix} n_{m,1} \\ n_{m,a} \end{bmatrix}. \quad (10)$$

It is worth mentioning that the total transmitted SNR at $R = R_{break}$ for the two symbols sent in each group is equal to $2k_m \cdot \text{SNR}_0$ in OPGPA. Then, the following theorem and corollaries hold, as proved in Appendix D.

Theorem 1. *Under the model of ZF-PGP in (10), the problem of maximizing the input-output mutual information over the matrix \mathbf{V}_{PGP} with SNR_0 employed is equivalent to using the following reception model, then maximizing the input, output mutual information over \mathbf{V}_{PGP}*

$$\begin{bmatrix} y \\ y_a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{V}_{PGP}^H \begin{bmatrix} x \\ x_a \end{bmatrix} + \begin{bmatrix} n'_1 \\ n'_a \end{bmatrix}, \quad (11)$$

where the noise vector is complex, cyclically symmetric Gaussian with mean zero and covariance

matrix equal to $\left(\sqrt{2k_m} s_{g,v,eff,m}^{(SNR_0)}\right)^{-2} \mathbf{I}_2$. Thus, the effective channel SNR is equal to $SNR_{eff,m} = \left(\sqrt{2k_m} s_{g,v,eff,m}^{(SNR_0)}\right)^2$. Furthermore, in the reception model of (11), the IOMIM precoder is the same for all $m = 1, 2, \dots, N_g$ under a constant $SNR_{eff,m}$.

Corollary 1. *The effect of multiplying the m input vector by $\sqrt{k_m}$ is to increase the effective channel reception SNR by k_m .*

Corollary 2. *The IOMIM precoder of (10) is a function of $SNR_{eff,m}$, only. Therefore, for the same IOMIM-achievable I_S value, the condition $SNR_{eff,m} = SNR_{req}(I_S)$ needs to be valid for all $m = 1, 2, \dots, N_g$, where $SNR_{req}(I_S)$ represents the effective SNR needed for achieving I_S by the IOMIM PGP precoder.*

Corollary 3. *For a QoS pre-set I_S , each PGP group can attain I_S , if the required k_m in the model of (11) is set by $\sqrt{k_m} = \sqrt{\frac{SNR_{req}(I_S)}{2}} \frac{1}{s_{g,v,eff,m}^{(SNR_0)}}$.*

We would like to stress that due to Corollary 2, since all sub-group IOMIM precoders under OPGPA are the same, the BS needs to determine a single IOMIM precoder for the pre-set I_S , resulting in a major simplification in system design. Furthermore, with OPGPA the system achieves the same QoS for all UEs in a group. Finally, OPGPA achieves very high gains in power efficiency. Thus, OPGPA represents a major improvement in downlink precoding for future wireless networks. Note that once the value of I_S is set, one needs to only determine the corresponding value of $SNR_{req}(I_S)$, then set the value of the gains $\sqrt{k_m}$ by $\sqrt{k_m} = \sqrt{\frac{SNR_{req}(I_S)}{2}} \frac{1}{s_{g,v,eff,m}^{(SNR_0)}}$, for $m = 1, 2, \dots, N_g$. Then, the corresponding average SNR employed by the OPGPA system will be

$$SNR_{OPGPA} = \frac{SNR_0 \cdot SNR_{req}(I_S)}{N_g} \sum_{m=1}^{N_g} \frac{1}{(s_{g,v,eff,m}^{(SNR_0)})^2}, \quad (12)$$

while the corresponding average required SNR without OPGPA will be

$$SNR_{NOPGPA} = SNR_0 \cdot SNR_{req}(I_S) \left(\frac{1}{\min_{m'=1, \dots, N_g} \left\{ s_{g,v,eff,m'}^{(SNR_0)} \right\}} \right)^2, \quad (13)$$

which shows that because of the high channel correlation, the approach without OPGPA will require much higher average SNR in order to achieve the requirement of all UEs meeting the set I_S requirement. Also, a more prudent approach can be to set SNR_0 as the minimum SNR

and also set a maximum SNR equal to SNR_1 , where $\text{SNR}_1 > \text{SNR}_0$. Then, depending on the scenario, some I_S might not be possible to be met, even without OPGPA. Finally, notice that when $I_S < 2 \cdot \log_2(M)$, without OPGPA some UEs could experience much higher throughput than I_S , i.e., the group experiences non-balanced performance among the different UEs, but under OPGPA the performance and thus the QoS is equal among all the UEs in the group, provided that $\min_{m'=1,\dots,N_g} \left\{ s_{g,v,eff,m'}^{(\text{SNR}_0)} \right\} \geq \sqrt{\frac{\text{SNR}_0 \cdot \text{SNR}_{req}(I_S)}{\text{SNR}_1}}$, as it can be seen easily.

III. NUMERICAL RESULTS

In this section, we present spectral efficiency results based on achievable input-output mutual information for an annular mmWave cell with internal radius r_i and external radius r_o . The BS antenna array is at a height $h = 3$ m. Two cases are considered: a) $N_{UE} = 10$, $r_i = 1$ m, $r_o = 5$ m, and b) $N_{UE} = 20$, $r_i = 1$ m, $r_o = 20$ m. For the NLOS UEs we use $k_{NLOS} = 4$ and $m_{NLOS} = 2$ (Nakagami fading), while for the LOS UEs we use $k_{LOS} = 2$ and $m_{LOS} = 4$ (Nakagami fading). We employ an annular cell of interior radius r_i and exterior radius r_o [15]. For both cases, we consider multiple scenarios, including Total Grouping ZF-PGP (TGZF-PGP), Sub-group ZF-PGP (SGZF-PGP), JSDM-FA ZF-PGP, and corresponding results for Zero-Forcing Precoding (ZFP). In order to offer high QoS to the NLOS UEs, we employ CFSDM [9], so that the LOS UEs employ a different subcarrier frequency than the NLOS ones. Furthermore, the LOS UEs employ $\text{SNR}_0 = 20$ dB, while the NLOS UEs employ a 13 dB higher SNR_0 than the LOS UEs, in order to balance the QoS between the NLOS group and LOS groups. Transmitted symbols are drawn from an $M = 16$ QAM constellation with two LOS groups selected from the \mathbf{S}_{EFF} matrix of the virtual LOS channel, after the pre-selection VCMB phase. The BS antenna has $N_T = 100$, with equal elements in the x and z dimensions. We use the SNR_0 at distance $r_i = 1$ m from the BS array base. Also, in order to stress the differences in precoding between finite alphabet inputs and Gaussian ones, we present results for Gaussian inputs in some cases in addition to the QAM ones.

A. Results with TG or SG without OPGPA

For $N_{UE} = 10$, $r_i = 1$ m, $r_o = 5$ m, i.e., a short-range cell deployment, with TG and $N_{V,INIT} = 20$ (98 % of total cell power captured), we get 2 NLOS users and the rest of the UEs are put in one group, G_2 , initially. NLOS G_1 results are depicted in Fig. 2. We observe that due to employing CFSDM with higher power, the NLOS group is able to attain high throughput.

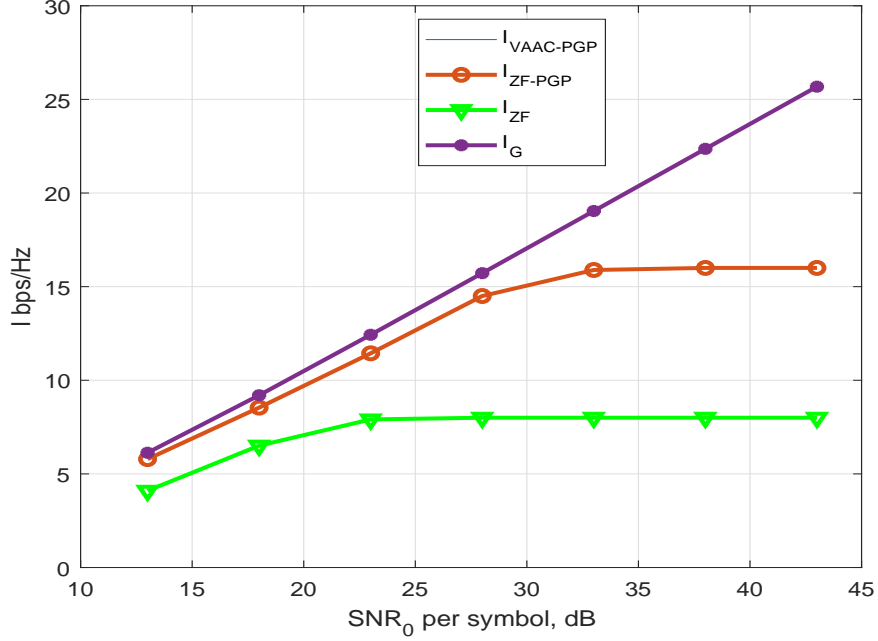


Fig. 2. Spectral efficiency for NLOS Group G_1 with TG employing 20 VCMBs for the first deployment scenario.

In addition, we see that the Gaussian input performance is close to the finite-alphabet one until the finite alphabet reaches saturation. The corresponding TG results for LOS G_2 are shown in Fig. 3. We see that due to the high correlation in G_2 , there is a large difference in performance between VAAC-PGP and ZF-PGP.

Next, we show results for G_2 with SG. We split G_2 in two frequency groups, based on similarities in their VCMBs, by allocating maximally distant spatial signatures in each group. Each of the two new groups G_2 and G_3 are allocated 4 VCMBs, then we apply SG with $N_{V,FINAL} = 4$ in both groups. The results for G_2 and G_3 are shown in Fig. 4 and 5, respectively. We observe that because G_2 and G_3 are structured with minimum correlation, the difference in performance between VAACP and ZF-PGP diminishes.

For the second deployment scenario, $N_{UE} = 20$, $r_i = 1$ m, $r_0 = 20$ m, i.e., a longer range cell is considered. For NLOS group G_1 with TG, we get the results shown in Fig. 6. We observe the same behavior as in the first deployment scenario. For LOS group G_2 with TG, we get the results shown in Fig. 7. Next, we split the LOS UEs in two (distant) groups with 8 UEs each, and apply TG or SG. Fig. 7 shows results for G_2 in conjunction with TG. Since using $N_{V,FINAL} = 20$ might be unrealistic, in Fig. 8, we show results with SG using $N_{V,FINAL} = 8$.

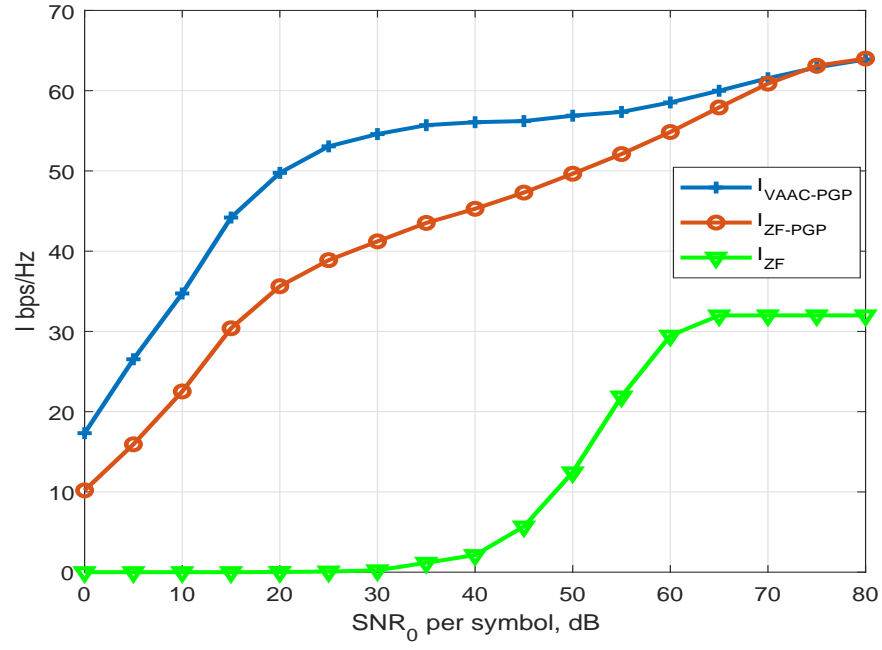


Fig. 3. Spectral efficiency for LOS UEs group G_2 for the first deployment scenario TG (20 VCMBs).

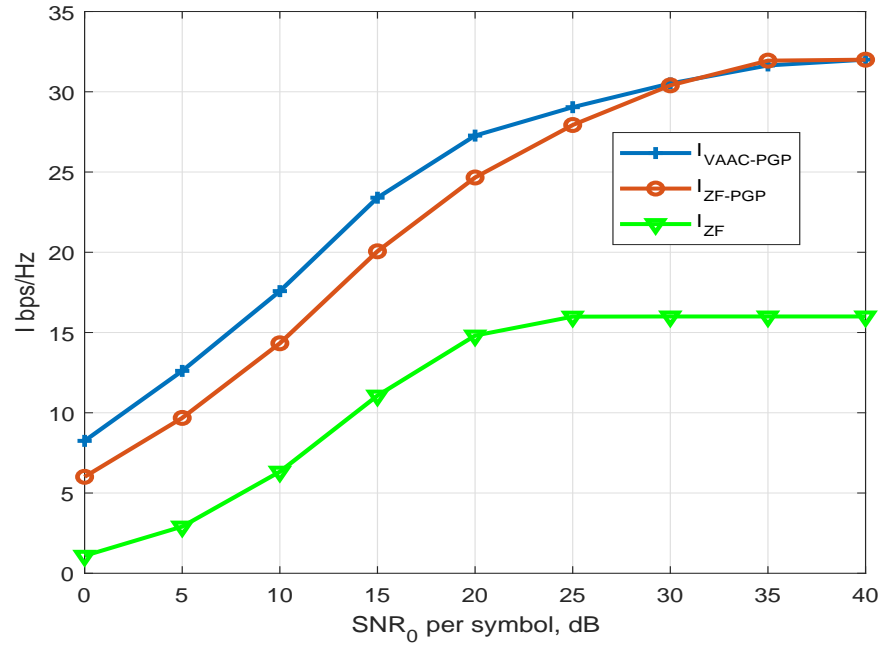


Fig. 4. Spectral efficiency for LOS Group G_2 for the first deployment scenario with SG, using 4 VCMBs.

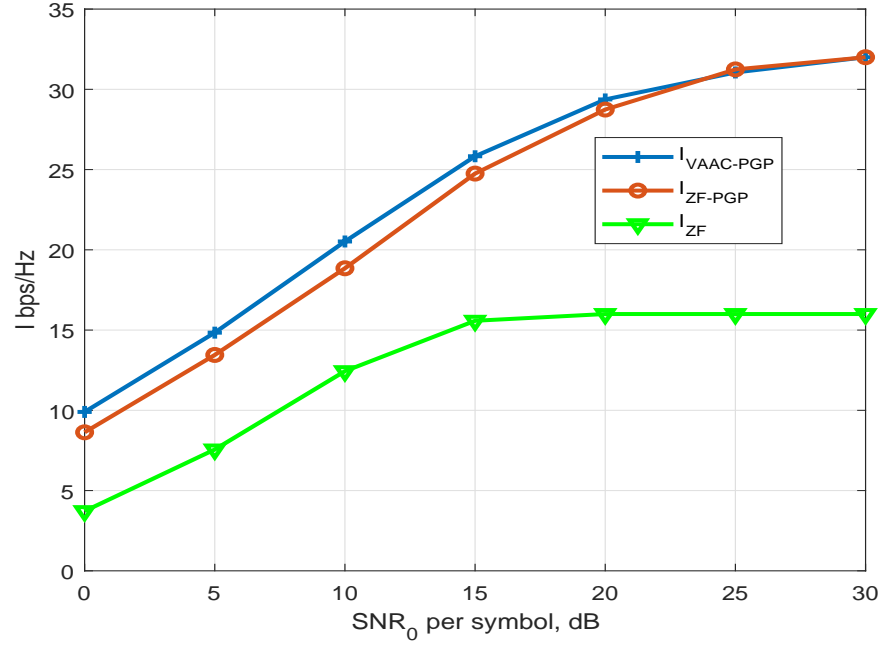


Fig. 5. Spectral efficiency for LOS Group G_3 for the first deployment scenario with SG, using 4 VCMBs.

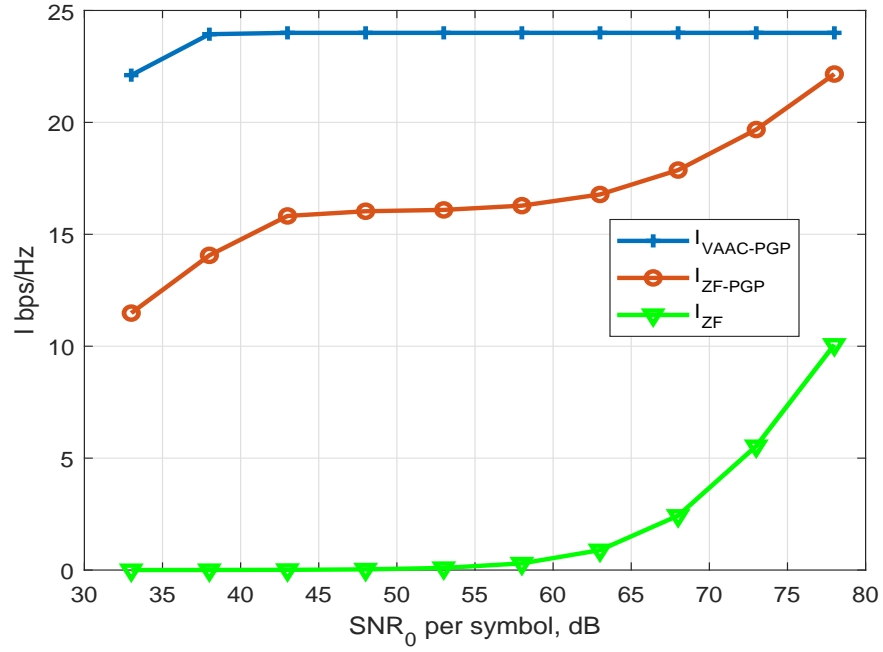


Fig. 6. Spectral efficiency for NLOS Group G_1 for the second deployment scenario with TG, using 10 VCMBs.

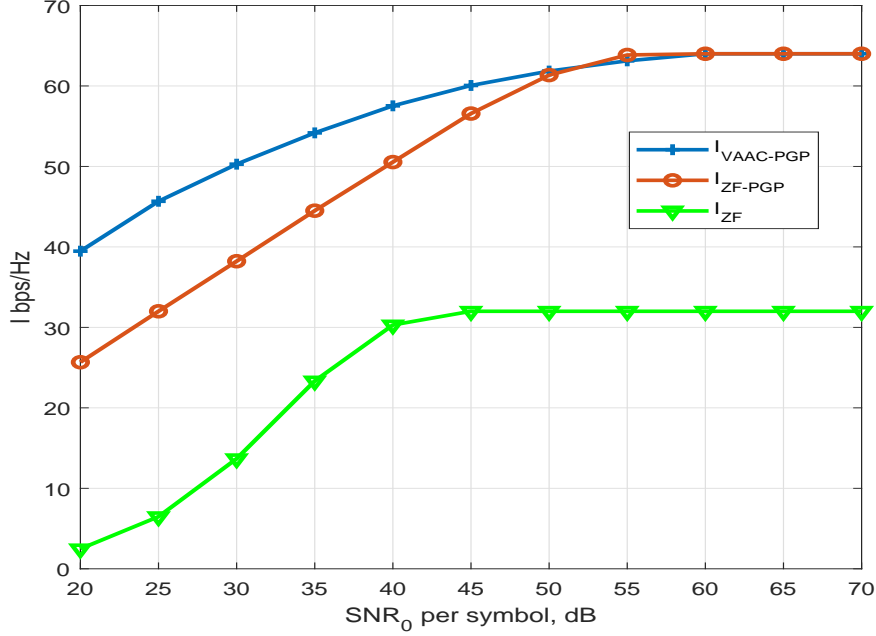


Fig. 7. Spectral efficiency for LOS Groups G_2 for the second deployment scenario with TG, using 20 VCMBs.

Corresponding results for G_3 are shown in Fig. 9, both SG and TG achieve almost identical results. We see that G_3 exhibits high correlation, thus its performance is really low. In order to remedy the situation, further splitting of G_3 in two frequency groups is required, resulting in two new sub-groups of 4 UEs each, G_3 , G_4 . In Fig. 10 and 11 we show G_3 and G_4 results with $N_{V,FINAL} = 4$.

For the Gaussian inputs, we observe in this case that their performance deviates significantly, even at lower SNR, due to the additional correlation present in the current scenario. This is a quite important observation in this regard. It is worth noting that the spectral efficiency achieved for the first scenario employing $\text{SNR}_0 = 30$ dB for all three subcarriers in the SG case, is $\text{SE} = 26.33$ bps/Hz, or SE per unit area, $\text{SEUA} = 0.3647$ bps/Hz/m². For the second scenario, the corresponding numbers are at $\text{SNR}_0 = 60$ dB, $\text{SE} = 31.50$ bps/Hz, and $\text{SEUA} = 0.0252$ bps/Hz/m², respectively.

B. Results for TG or SG with OPGPA

Here we present results with TG or SG employing the OPGPA concept, in order to offer equal QoS under lower average power consumption by the BS. We focus on the second scenario and

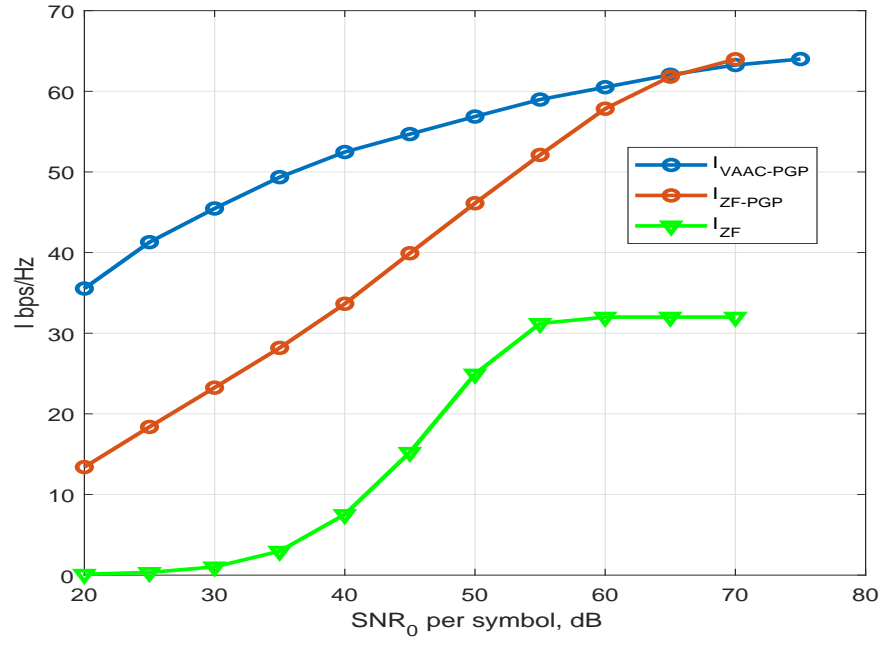


Fig. 8. Spectral efficiency for LOS Group G_2 for the second deployment scenario with SG, using 8 VCMBs.

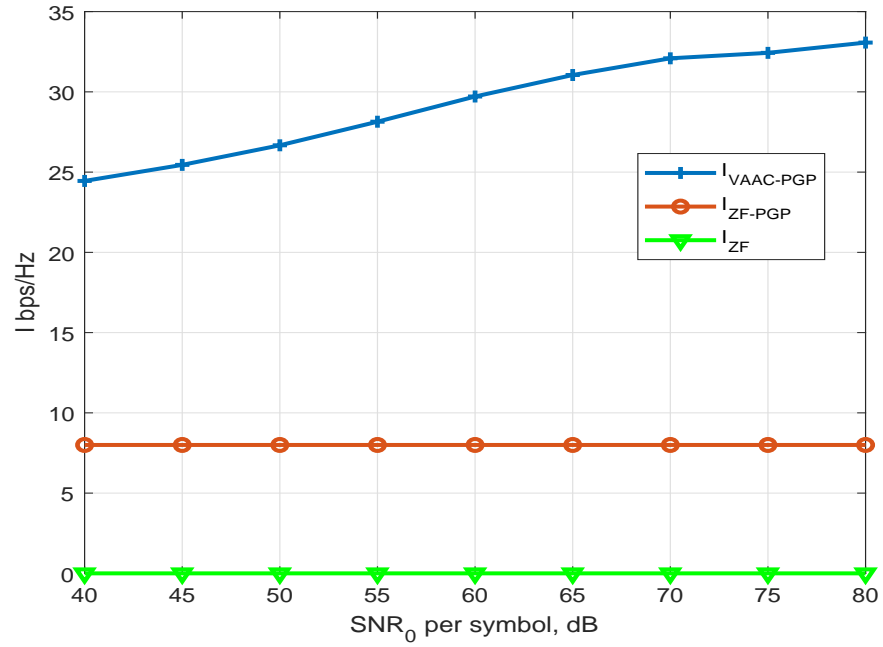


Fig. 9. Spectral efficiency for LOS Group G_3 for the second deployment scenario with TG, using 20 VCMBs (TG) or 8 VCMBs (SG), both having the same performance.

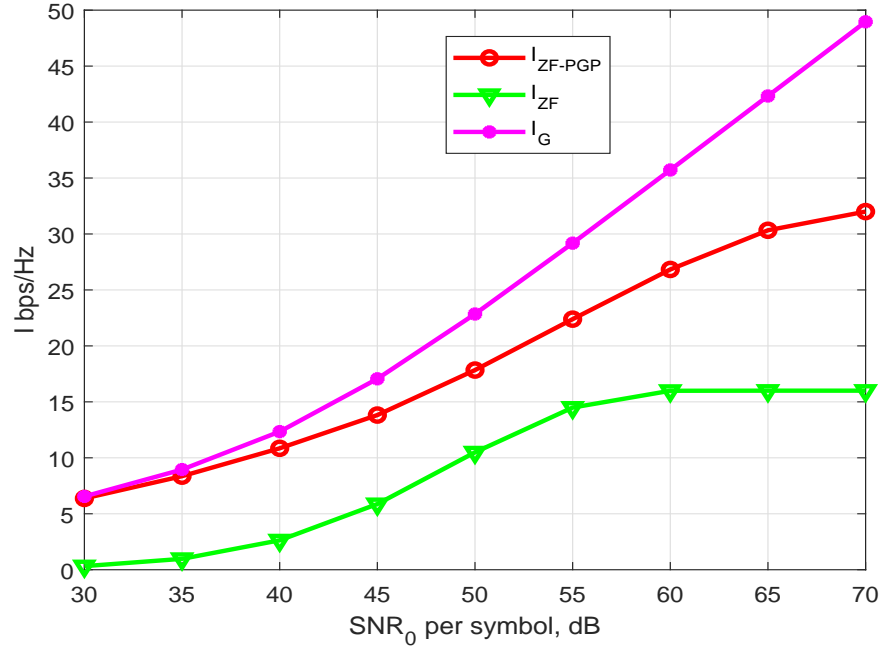


Fig. 10. Spectral efficiency for LOS Group G_3 for the second deployment scenario with SG, using 4 VCMBs.

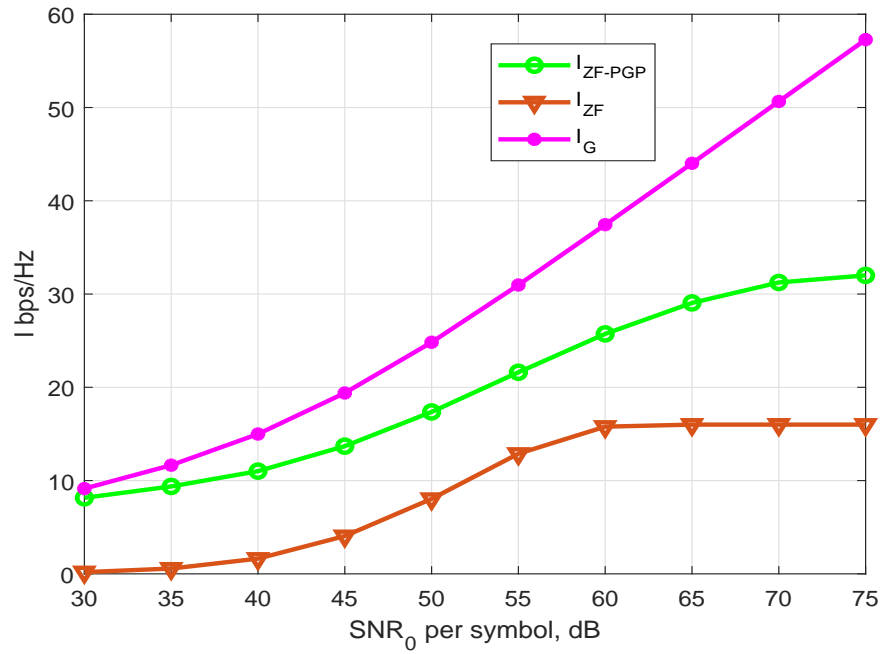


Fig. 11. Spectral efficiency for LOS Group G_4 for the second deployment scenario with SG, using 4 VCMBs.

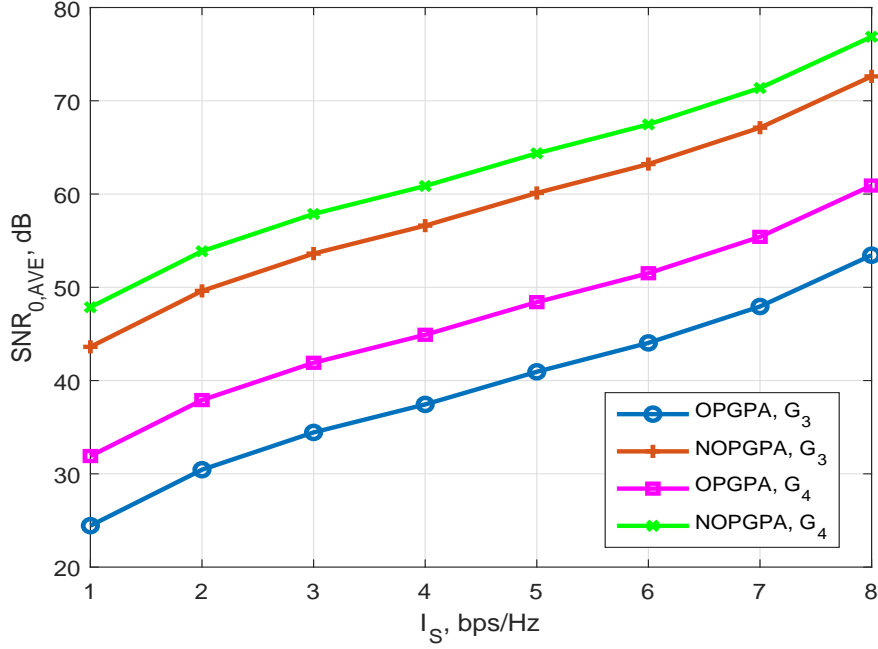


Fig. 12. OPGPA and NOPGPA $SNR_{0,AVE}$ for LOS Group G_3 and G_4 for the second deployment scenario with SG, using 4 VCMBs.

we apply OPGPA and without OPGPA (NOPGPA) with $SNR_0 = 20$ dB, in the case of groups G_3 and G_4 presented above, with four UEs each and SG. In Fig. 12 we present the corresponding results with respect to the pre-set QoS, I_S . For both cases, we see that by applying OPGPA, the system achieves more than 15 dB power reduction in the average SNR_0 required. This is a very big improvement in the power requirement, while simultaneously OPGPA offers equal throughput to all UEs, i.e., it mitigates the imbalance in the QoS of different UEs.

C. Results for JSDM-FA

For JSDM-FA, we focus on the first scenario without OPGPA. In general, since full orthogonality between JSDMA-FA sub-groups is not possible, there is always MAI entering each UE's receiver, reducing its effective SNR. The effective SNR for UE i of sub-group l of group k , $G_{k,l}$, $SNR_{eff,G_{k,l},i} = \left(\frac{1}{SNR_0} + \frac{1}{\sum_{l' \in G_{k,l'}, l' \neq l} p_{MAI,k,l,l'}(i)} \right)^{-1}$, where $p_{MAI,k,l,l'}(i)$ is the MAI power to the i th UE in $G_{k,l}$ from $G_{k,l'}$, evaluated in Appendix B for ZFP and ZF-PGP.

In Fig. 13 we present results for $G_{2,1}$ and $G_{2,2}$, comprising 3 and 5 UEs, respectively. We see that the performance of both sub-groups becomes MAI limited and it is significantly lower than the one achievable by, e.g., SG ZF-PGP. In addition, due to the high correlation in the

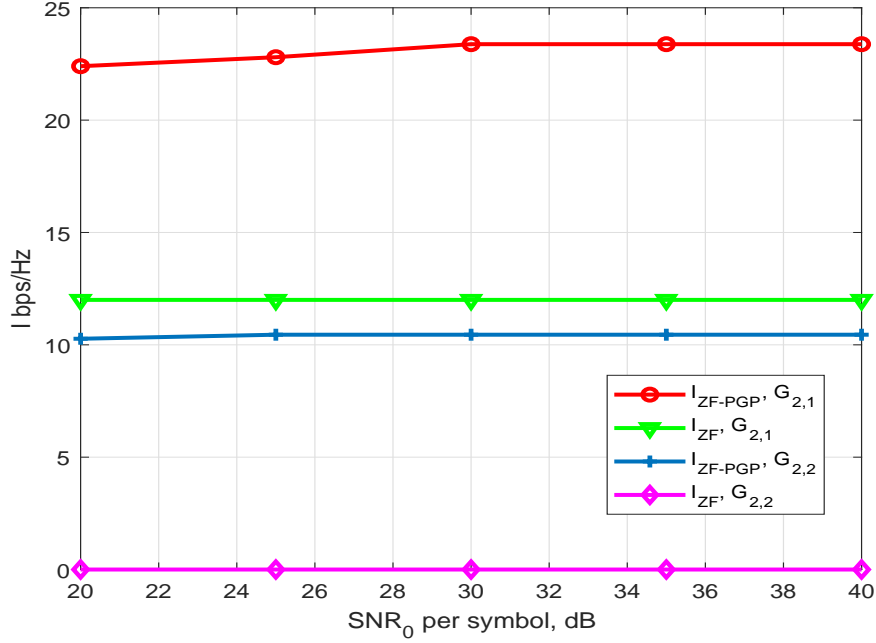


Fig. 13. JSDM-FA spectral efficiency for LOS Groups $G_{2,1}$ and $G_{2,2}$ for the first deployment scenario with JSDM-FA.

sub-group UE channels, the performance of ZFP is negligible. For other scenarios with more UEs or longer cell range, the effects of correlation and MAI create an even harsher environment for JSDM-FA.

IV. CONCLUSIONS

We have applied the concept of ZF-PGP in a mmWave massive MIMO cell, carefully modeling all channel intricacies, including UE blocking, path propagation, and multipath fading. It is shown that VCMBs offer many advantages in simplifying the channel representation, analyzing the system, and improving group and UE throughput. For the scenarios presented, it is shown that ZF-PGP offers significant throughput improvements, especially when the channel presents a medium-to-high degree of correlation. When such spatial correlation is present, ZF-PGP offers a 100% improvement in throughput over ZFP. When even higher spatial correlation is present or the cell range becomes longer, ZF-PGP can offer more than 300% throughput improvement over ZFP, albeit in the lower SNR region. We compare three different VCMB-based group-forming techniques which offer different advantages, depending on the application scenario. Finally, our work has demonstrated that when $N_{UE} < 30$, only four subcarrier frequencies suffice for CFSDM

to be used in the cell, in order to guarantee a balanced QoS for all UEs. We also introduced a new method, OPGPA, that significantly reduces the average power transmitted per group, while it meets a specific QoS requirement for all UEs in a group. Finally, from our presented results, it becomes more evident that under the correlated channel conditions in a mmWave cell, the widely used Gaussian approximation falls short in accurately predicting the optimal precoder performance.

APPENDIX A

GH QUADRATURE APPROXIMATION IN MIMO INPUT-OUTPUT MUTUAL INFORMATION

Let us consider a generic N_t transmit antenna, N_r receive antenna MIMO model as described by the following equation

$$\mathbf{y} = \mathbf{H}\mathbf{G}\mathbf{x} + \mathbf{n}, \quad (14)$$

where \mathbf{y} is the $N_r \times 1$ received vector, \mathbf{H} is the $N_r \times N_t$ MIMO channel matrix, \mathbf{G} is the precoder matrix of size $N_t \times N_t$, \mathbf{x} is the $N_t \times 1$ data vector with independent components each of which is drawn from the QAM constellation of size M , \mathbf{n} represents the circularly symmetric AWGN vector of size $N_r \times 1$, with mean zero and covariance matrix $\mathbf{K}_n = \sigma_n^2 \mathbf{I}_{N_r}$, where \mathbf{I}_{N_r} is the $N_r \times N_r$ identity matrix, and $\sigma^2 = \frac{1}{\text{SNR}}$. $I(\mathbf{x}; \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) = N_t \log_2(M) - H(\mathbf{x}|\mathbf{y})$, where the conditional entropy, $H(\mathbf{x}|\mathbf{y})$ can be written as [12]

$$\begin{aligned} H(\mathbf{x}|\mathbf{y}) &= \frac{N_r}{\log(2)} + \frac{1}{M^{N_t}} \sum_k \mathbb{E}_{\mathbf{n}} \left(\log_2 \left(\sum_m \exp\left(-\frac{1}{\sigma^2} \|\mathbf{n} - \mathbf{H}\mathbf{G}(\mathbf{x}_k - \mathbf{x}_m)\|^2\right) \right) \right) \\ &= \frac{N_r}{\log(2)} + \frac{1}{M^{N_t}} \sum_k \int_{-\infty}^{+\infty} \mathcal{N}_c(\mathbf{n}|\mathbf{0}, \sigma^2 \mathbf{I}) \log_2 \left(\sum_m \exp\left(-\frac{1}{\sigma^2} \|\mathbf{n} - \mathbf{H}\mathbf{G}(\mathbf{x}_k - \mathbf{x}_m)\|^2\right) \right) d\mathbf{n}, \end{aligned} \quad (15)$$

where $\mathcal{N}_c(\mathbf{n}|\mathbf{0}, \sigma^2 \mathbf{I})$ represents the probability density function (pdf) of the circularly symmetric complex random vector due to AWGN. Let us define

$$f_k \doteq \int_{-\infty}^{+\infty} \mathcal{N}_c(\mathbf{n}|\mathbf{0}, \sigma^2 \mathbf{I}) \log_2 \left(\sum_{\mathbf{x}_m} \exp\left(-\frac{1}{\sigma^2} \|\mathbf{n} - \mathbf{H}\mathbf{G}(\mathbf{x}_k - \mathbf{x}_m)\|^2\right) \right) d\mathbf{n}. \quad (16)$$

There is an strong connection between f_k and the parameter $H_k(\mathbf{y}) \doteq \mathbb{E}_{\mathbf{y}|\mathbf{x}_k} \{-\log_2(p(\mathbf{y}))|\mathbf{x} = \mathbf{x}_k\}$ called the Input-Dependent Output Entropy (IDOE) herein. Note that IDOE represents the entropy at the receiver output when the input is \mathbf{x}_k . This parameter is different that the conditional entropy.

By using standard entropic identities, we can easily see that

$$f_k = -H_k(\mathbf{y}) + 2N_r \log_2(\sigma) + N_r \log_2(\pi) + N_t \log_2(M). \quad (17)$$

Thus, since our Gauss-Hermite approximation focuses on finding approximations to each of the f_k terms, one per input symbol, it equivalently offers estimates of the IDOE terms $H_k(\mathbf{y})$. This gives a physical meaning to the estimated terms \hat{f}_k presented below.

Since \mathbf{n} has independent components over the different receiving antennas, and over the real and imaginary dimensions, the integral above can be partitioned into $2N_r$ real integrals in tandem, in the following manner: Define by n_{rv}, n_{iv} , with $v = 1, \dots, N_r$, the v th receiving antenna real and imaginary noise component, respectively. Also define by $(\mathbf{HG}(\mathbf{x}_k - \mathbf{x}_m))_{rv}$ and $(\mathbf{HG}(\mathbf{x}_k - \mathbf{x}_m))_{iv}$, the v th receiving antenna real and imaginary component of $(\mathbf{HG}(\mathbf{x}_k - \mathbf{x}_m))$, respectively. We then have

$$\mathcal{N}_c(\mathbf{n}|\mathbf{0}, \sigma^2 \mathbf{I}) = \frac{1}{\pi^{N_r} \sigma^{2N_r}} \exp\left(-\frac{\sum_l n_{rv}^2 + n_{iv}^2}{\sigma^2}\right), \quad (18)$$

$$d\mathbf{n} = \prod_{v=1}^{N_r} dn_{rv} dn_{iv}, \quad (19)$$

and

$$\begin{aligned} & \sum_m \exp\left(-\frac{1}{\sigma^2} \|\mathbf{n} - \mathbf{HG}(\mathbf{x}_k - \mathbf{x}_m)\|^2\right) \\ &= \sum_m \exp\left(-\frac{1}{\sigma^2} \left(\sum_v (n_{rv} - (\mathbf{HG}(\mathbf{x}_k - \mathbf{x}_m))_{rv})^2 \right. \right. \\ & \quad \left. \left. + \sum_v (n_{iv} - (\mathbf{HG}(\mathbf{x}_k - \mathbf{x}_m))_{iv})^2\right)\right). \end{aligned} \quad (20)$$

Based on the above equations, the N_r -size complex integral of (16) can be written as a $2N_r$ -size real integral to which the GH quadrature approximation can be applied easily.

By applying the GH quadrature theory to the integral of a Gaussian function multiplied with an arbitrary real function $f(x)$, i.e.,

$$F \doteq \int_{-\infty}^{+\infty} \exp(-x^2) f(x) dx, \quad (21)$$

one gets the following approximation upon employing with L GH weights and GH nodes as

$$F \approx \sum_{l=1}^L w(l)f(v_l) = \mathbf{w}^T \mathbf{f}, \quad (22)$$

with $\mathbf{w} = [w(1) \cdots w(L)]^t$, $\{v_l\}_{l=1}^L$, and $\mathbf{f} = [f(v_1) \cdots f(v_L)]^t$, being the vector of the GH weights, the GH nodes, and the function GH node values, respectively. For the GH weights and nodes we have [26]

$$w(l) = \frac{2^{L-1} L! \sqrt{2\pi}}{L^2 (H_{L-1}(v_l))^2} \quad (23)$$

where $H_L(x) = (-1)^L \exp(x^2) \frac{d^L}{dx^L} (\exp(-x^2))$ is the (L) -th order Hermitian polynomial, and the value of the node v_l equals the l th root of $H_L(x)$ for $l = 1, 2, \dots, L$. Let us first introduce some notations that make the overall understanding easier. Let \mathbf{n}_e denote the equivalent to \mathbf{n} , complex vector of length N_r derived from \mathbf{n} as follows

$$\mathbf{n}_e = [n_{r1} + jn_{i1} \cdots n_{rN_r} + jn_{iN_r}]^T, \quad (24)$$

with $n_{rv} + jn_{iv}$ being the values of the of the v th ($1 \leq v \leq N_r$) element of \mathbf{n} , respectively. Let us also define the length N_r complex vector defined as follows

$$\mathbf{v} = [v_{kr1} + jv_{ki1}, \dots, v_{krN_r} + jv_{kiN_r}]^T, \quad (25)$$

with k_{rv} , k_{iv} being in the set $\{1, 2, \dots, L\}$, i.e., \mathbf{v} is a function of the complex vector $\mathbf{k}_c = [k_{r1} + jK_{i1} \ k_{r2} + jk_{i2} \cdots \ k_{rN_r} + jk_{iN_r}]^T$.

Finally, the following lemma is proven in [6] concerning the Gauss-Hermite approximation for $I(\mathbf{x}; \mathbf{y})$ in (14).

Lemma. *For the MIMO channel model presented in (14), the Gauss-Hermite approximation for the input $I(\mathbf{x}; \mathbf{y})$ with L nodes per receiving antenna is given as*

$$I(\mathbf{x}; \mathbf{y}) \approx N_t \log_2(M) - \frac{N_r}{\log(2)} - \frac{1}{M^{N_t}} \sum_{k=1}^{M^{N_t}} \hat{f}_k, \quad (26)$$

where

$$\hat{f}_k = \left(\frac{1}{\pi}\right)^{N_r} \sum_{k_{r1}=1}^L \sum_{k_{i1}=1}^L \cdots \sum_{k_{rN_r}=1}^L \sum_{k_{iN_r}=1}^L w(k_{r1})w(k_{i1}) \cdots w(k_{rN_r}) \times w(k_{iN_r})z(\sigma\mathbf{v}(\mathbf{k}_c)), \quad (27)$$

with

$$z(\sigma\mathbf{v}(\mathbf{k}_c)) \quad (28)$$

being the value of the function

$$\log_2 \left(\sum_m \exp\left(-\frac{1}{\sigma^2} \|\mathbf{n} - \mathbf{H}\mathbf{G}(\mathbf{x}_k - \mathbf{x}_m)\|^2\right) \right) \quad (29)$$

evaluated at $\mathbf{n}_e = \sigma\mathbf{v}(\mathbf{k}_c)$.

APPENDIX B

DERIVATION OF THE MAI INTERFERENCE POWERS FOR JSMD-FA

Here we derive the MAI for different precoding scenarios, including ZFP, ZF-PGP, and VAAC-PGP. In the following analysis we use the following generic notation for brevity. $\mathbf{H}_{v,k,l,l'}$ denotes the effective virtual domain downlink channel for group's k, l sub-group, containing the VCMBs employed by $G_{k,l'}$, and $\mathbf{P}_{v,k,l}$ is the precoder employed for the $G_{k,l}$ sub-group, based on a specific, but otherwise arbitrary technique employed for downlink precoding to sub-group $G_{k,l}$. Assume that we need to determine sub-group $G_{k,l'}$ interfering power to sub-group $G_{k,l}$, under JSMD-FA. Let the receiving equation for sub-group $G_{k,l}$ with interference from all sub-groups $l' \neq l$ be

$$\mathbf{y}_{k,l} = \mathbf{H}_{v,k,l,l'}^H \mathbf{P}_{v,k,l} \mathbf{x}_{k,l} + \mathbf{n}_{k,l,AWGN} + \sum_{l' \text{ in } G_{k,l}, l' \neq l} \mathbf{n}_{MAI,k,l,l'}, \quad (30)$$

where $\mathbf{y}_{k,l}$, $\mathbf{x}_{k,l}$, $\mathbf{n}_{k,l,AWGN}$, and $\mathbf{n}_{MAI,k,l,l'}$ represent the $G_{k,l}$ JSMD-FA sub-group received vector, data vector, AWGN noise vector, and MAI vector from $G_{k,l'}$ to $G_{k,l}$, respectively. We thus see that the covariance matrix of the interference from $G_{k,l'}$ to $G_{k,l}$, $\mathbf{K}_{k,l,l'}$ becomes

$$\mathbf{K}_{k,l,l'} = \mathbf{H}_{v,k,l,l'} \mathbf{P}_{v,k,l'} \mathbf{P}_{v,k,l'}^H \mathbf{H}_{v,k,l,l'}^H, \quad (31)$$

where the input symbols are uncorrelated and of unit power. Applying this result using the generic $\mathbf{P}_{v,k,l'}$ and focusing on the diagonal elements of $\mathbf{K}_{k,l,l'}$ only, we get the following expressions

for the MAI power at the input of $G_{k,l}$ i th UE receiver from $G_{k,l'}$

$$\mathbf{p}_{MAI,k,l,l'}(i) = \mathbf{K}_{k,l,l'}[i, i], \quad (32)$$

where $i = 1, \dots, N_{S_k}$. The total MAI to $G_{k,l}$ i th UE receiver is then

$$\mathbf{p}_{MAI,k,l}(i) = \sum_{l' \text{ in } G_k, l' \neq l} \mathbf{p}_{MAI,k,l,l'}(i). \quad (33)$$

Then, for the two main precoder types considered in this paper, we use

$$\mathbf{P}_{v,k,l'}^{(ZFP)} = w_{ZF,k,l'}^2 \mathbf{H}_{v,k,l'}^H (\mathbf{H}_{v,k,l'} \cdot \mathbf{H}_{v,k,l'}^H)^{-1}, \quad (34)$$

where $w_{ZF,k,l'}$ is the ZFP SNR normalizing weight of the $G_{k,l}$ group [16], while for the ZF-PGP one, we get

$$\mathbf{P}_{v,k,l'}^{(ZF-PGP)} = \mathbf{S}_{EFF} \tilde{\mathbf{S}}_{v,k,l'}^{(ZF-PGP)} \mathbf{V}_{v,k,l'}^{(ZF-PGP)}, \quad (35)$$

where \mathbf{S}_{EFF} is the effective channel singular value matrix in ZF-PGP [16], and $\tilde{\mathbf{S}}_{v,k,l'}^{(ZF-PGP)} = \sqrt{2} \mathbf{I}_{N_g}$ is the PGP singular values of the ZF-PGP precoder, due to VAAC, and $\mathbf{V}_{v,k,l'}^{(ZF-PGP)}$ is the right singular matrix of the ZF-PGP precoder.

APPENDIX C

VAAC-PGP DERIVATION DETAILS

Here the concept of adding virtual antennas, i.e., additional data streams to the same antennas employed by a MIMO system jointly with PGP-WG is explained in detail. Without a loss of generality, we consider a MIMO system with equal number of transmitting and receiving antennas, i.e., $N_t = N_r = N_g$, where N_t and N_r represent the number of transmitting and receiving antennas, respectively. The channel model under consideration then becomes

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \quad (36)$$

where \mathbf{y} , \mathbf{H} , \mathbf{x} , and \mathbf{n} represent the received data, the MIMO channel, the transmitted data, and the AWGN noise, respectively, and where matrices are of size $N \times N$ and vectors are of size $N \times 1$. The equivalent singular value decomposition based model for the MIMO channel is

$$\mathbf{y} = \mathbf{U}_H \Sigma_H \mathbf{V}_H^H \mathbf{x} + \mathbf{n}, \quad (37)$$

with \mathbf{U}_H , $\mathbf{\Sigma}_H$, \mathbf{V}_H representing the size $N \times N$ matrices of left singular vectors, singular values, and right singular vectors, respectively. Consider adding N virtual antennas of zero singular values, i.e., useless, noise-only channels. This can be added to the previous model as follows

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_a \end{bmatrix} = \begin{bmatrix} \mathbf{U}_H & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_H & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_H^H & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_a \end{bmatrix} + \begin{bmatrix} \mathbf{n} \\ \mathbf{n}_a \end{bmatrix}, \quad (38)$$

where the subscript a is used to indicate the N added, fictitious antennas. In the above equation, the vector \mathbf{x}_a represents the N added QAM inputs to the MIMO system. Note that in (38) the inputs represented by \mathbf{x}_a cannot be transmitted, due to their corresponding zero singular values (noise-only channel) which result in zero input-output mutual information. However, one can still apply the virtual model of (38) with PGP-WG. The PGP-WG algorithm will optimize and assign an amplitude diagonal matrix as per PGP-WG $\mathbf{\Sigma}_{P_i} = \text{diag}[\sqrt{2}, 0]$, $i = 1, 2, \dots, K_g$ for each sub-group of the PGP-WG, i.e., no power sent to the noise-only antenna. This results in $\mathbf{S}_P = \sqrt{2}\mathbf{I}_{N_g}$ in VAAC-PGP. On the other hand, PGP-WG will also determine the optimal unitary precoder matrix to each sub-group in PGP-WG, thus it will be multiplexing optimally two QAM symbols to each actual transmitting antenna of the original MIMO system in (37). Now, VAAC-PGP proceeds as follows: for group g , $g = 1, 2, 3, 4$, the precoder employs the singular values of the downlink channel $\mathbf{H}_{g,v}$ as the matrix $\mathbf{\Sigma}_H$ of equation (38). Then, it applies PGP-WG as explained above. VAAC-PGP is basically similar to ZF-PGP, but without the ZF part, i.e., no left diagonalization of the channel matrix takes place in the BS precoding process, thus the need for the UEs to do that arises.

APPENDIX D

PROOF OF THEOREM 1 AND THE COROLLARIES OF OPGPA

First, we prove Theorem 1. By multiplying (10) by $\frac{1}{\sqrt{2k_m s_{g,v,eff,m}^{(\text{SNR}_0)}}}$, we get (11). Since the original circular complex Gaussian noise vector in (10) has a variance 1 per component (due to normalization) with correlation 0 (independent components), the new noise vector has a covariance matrix equal to

$$\frac{1}{2k_m \left(s_{g,v,eff,m}^{(\text{SNR}_0)} \right)^2} \mathbf{I}_2. \quad (39)$$

Thus, the effective SNR of the PGP group is $\text{SNR}_{eff,m} = 2k_m \left(s_{g,v,eff,m}^{(\text{SNR}_0)} \right)^2$. To prove the rest of Theorem 1 claims, we first notice that the IOMIM precoder operating on (10) depends only on the effective SNR. Thus, the IOMIM precoder for the m PGP group is the same, under the requirement that $\text{SNR}_{eff,m} = \text{constant}$ for all $m = 1, 2, \dots, N_g$.

For Corollary 1, we notice that by setting $k_m = 1$ in (10), we get $\text{SNR}_{eff,m} = \frac{1}{2 \left(s_{g,v,eff,m}^{(\text{SNR}_0)} \right)^2}$, thus the claim becomes obvious.

For Corollary 2, we use the model in (11) and observe that for a set $\text{SNR}_{eff,m}$ the precoder is constant, in other words, the IOMIM precoder is only a function of $\text{SNR}_{eff,m}$. Since an increase of $\text{SNR}_{eff,m}$ results in higher throughput by the IOMIM precoder, we see that there is a unique required $\text{SNR}_{req}(I_S)$ for each I_S . This means that by setting $\text{SNR}_{eff,m} = \text{SNR}_{req}(I_S)$, the desired I_S is attained by all PGP groups.

Finally, for Corollary 3, we need to substitute into $\text{SNR}_{eff,m} = \text{SNR}_{req}(I_S)$ the expression $\text{SNR}_{eff,m} = 2k_m \left(s_{g,v,eff,m}^{(\text{SNR}_0)} \right)^2$ from Theorem 1. Then, we get directly that the m th ($1 \leq m \leq N_g$) gain is determined as follows $\sqrt{k_m} = \sqrt{\frac{\text{SNR}_{req}(I_S)}{2} \frac{1}{s_{g,v,eff,m}^{(\text{SNR}_0)}}}$.

REFERENCES

- [1] C. Park and T. S. Rappaport, "Short-Range Wireless Communications for Next-Generation Networks: UWB, 60 GHz Millimeter-Wave WPAN, and ZigBee," *IEEE Wireless Communications*, vol. 14, no. 4, pp. 70–78, August 2007.
- [2] T. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Transactions on Wireless Communications*, vol. 9, pp. 3590–3600, November 2010.
- [3] J. Jose, A. Ashikhmin, T. Marzetta, and S. Vishwanath, "Pilot Contamination and Precoding in Multi-Cell TDD Systems," *IEEE Transactions on Wireless Communications*, vol. 10, pp. 2640–2651, August 2011.
- [4] H. Ngo, E. Larsson, and T. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," *IEEE Transactions on Communications*, vol. 61, pp. 1436–1449, April 2013.
- [5] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
- [6] T. Ketseoglou and E. Ayanoglu, "Linear Precoding Gain for Large MIMO Configurations with QAM and Reduced Complexity," *IEEE Transactions on Communications*, vol. 64, pp. 4196–4208, October 2016.
- [7] S. Zarei, W. H. Gerstacker, R. Weigel, M. Vossiek, and R. Schober, "Robust MSE-Balancing Hierarchical Linear/Tomlinson-Harashima Precoding for Downlink Massive MU-MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7309–7324, November 2018.
- [8] C. Xiao, Y. Zheng, and Z. Ding, "Globally Optimal Linear Precoders for Finite-Alphabet Signals Over Complex Vector Gaussian Channels," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3301–3314, July 2011.
- [9] T. Ketseoglou and E. Ayanoglu, "Downlink Precoding for Massive MIMO Systems Exploiting Virtual Channel Model Sparsity," *IEEE Transactions on Communications*, vol. 66, pp. 1925–1939, May 2018.
- [10] Y. Wu, C.-K. Wen, D. Wing Kwan Ng, R. Schober, and A. Lozano, "Low-Complexity MIMO Precoding for Finite-Alphabet Signals," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4571–4584, July 2017.

- [11] Y. Wu, C. Xiao, Z. Ding, X. Gao, and S. Jin, "A Survey on MIMO Transmission With Finite Input Signals: Technical Challenges, Advances, and Future Trends," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1779–1833, October 2018.
- [12] W. Zeng, C. Xiao, and J. Lu, "A Low Complexity Design of Linear Precoding for MIMO Channels with Finite-Alphabet Inputs," *IEEE Wireless Communications Letters*, vol. 1, pp. 38–42, February 2012.
- [13] Y. Wu, C.-K. Wen, C. Xiao, X. Gao, and R. Schober, "Linear Precoding for the MIMO Multiple Access Channel With Finite-Alphabet Inputs and Statistical CSI," *IEEE Transactions on Wireless Communications*, pp. 983–997, February 2015.
- [14] K. Venugopal, M. C. Valenti, and R. W. Heath, "Device-to-Device Millimeter Wave Communications: Interference, Coverage, Rate, and Finite Topologies," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6175–6188, September 2016.
- [15] E. Hriba, M. C. Valenti, K. Venugopal, and R. W. Heath, "Accurately Accounting for Random Blockage in Device-to-Device mmWave Networks," in *GLOBECOM 2017*, December 2017, pp. 1–6.
- [16] T. Ketseoglou and E. Ayanoglu, "Zero-Forcing Per-Group Precoding (ZF-PGP) for Robust Optimized Downlink Massive MIMO Performance," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 6816–6828, October 2019.
- [17] C. Peel, B. Hockwald, and A. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multiantenna Multiuser Communication - Part I: Channel Inversion and Regularization," *IEEE Transactions on Communications*, vol. 53, pp. 195–202, January 2005.
- [18] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge: Cambridge University Press, 2005.
- [19] A. Wiesel, Y. C. Eldar, and S. S. (Shitz), "Zero-Forcing Precoding and Generalized Inverses," *IEEE Transactions on Signal Processing*, vol. 56, pp. 4409–4418, September 2008.
- [20] A. Molisch, *Wireless Communications*. New York: Wiley-IEEE Press, 2011.
- [21] A. Adhikary, J. Nam, J. Y. Ahn, and G. Caire, "Joint Spatial Division and Multiplexing: The Large-Scale Array Regime," *IEEE Trans. Inf. Theory*, vol. 59, pp. 6441–6463, October 2013.
- [22] J. Nam, A. Adhikary, J. Y. Ahn, and G. Caire, "Joint Spatial Division and Multiplexing: Opportunistic Beamforming, User Grouping and Simplified Downlink Scheduling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 876–890, October 2014.
- [23] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "Secure and Energy-Efficient Beamforming for Simultaneous Information and Energy Transfer," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7523–7537, November 2017.
- [24] A. Adhikary, E. Al Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint Spatial Division and Multiplexing for mm-Wave Channels," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1239–1255, June 2014.
- [25] T. Ketseoglou and E. Ayanoglu, "Linear Precoding for MIMO with LDPC Coding and Reduced Complexity," *IEEE Transactions on Wireless Communications*, pp. 2192–2204, April 2015.
- [26] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington D.C.: U.S. Government Printing Office, 1972.