# Zero-Forcing Per-Group Precoding for Robust Optimized Downlink Massive MIMO Performance

Thomas Ketseoglou, *Senior Member, IEEE*, and Ender Ayanoglu, *Fellow, IEEE*

***Abstract*— In this paper, we propose a new, robust near-optimal mutual information combined zero-forcing per-group precoding (ZF-PGP) method for finite alphabets with imperfect channel state information (CSI) at the receiver. ZF-PGP achieves very high gains over ZF precoding techniques, while it simultaneously offers individually separate streams to reach individual user equipment (UE), i.e., it obliterates the need for coordinated, joint decoding by the group's UEs. The proposed robust ZF-PGP scheme comprises an inner linear beamformer (BF), which relies on second-order CSI and is based on the massive MIMO-based joint division and multiplexing for finite alphabets (JSDM-FA) to form orthogonal user groups, and an outer linear precoder that relies on instantaneous group CSI to form a per-group precoding within groups (PGP-WG) with separate data streams to each user. As we show by multiple examples, for a uniform linear array (ULA), the proposed ZF-PGP technique is very efficient, near-optimal design with reduced complexity at the UEs. Due to JSDM-FA second-order CSI inner beamforming, the groups are easily identified. We also demonstrate the robustness of the proposed precoding techniques to channel estimation errors, showing minimal performance loss.**

***Index Terms*— Beamforming, inner linear beamformer, outer precoder, channel state information.**

## I. Introduction

In massive Multiuser (MU) Multiple-Input Multiple-Output (MIMO), a large number of base station (BS) antennas is used in order to facilitate provision of high data rates to many UEs and thus achieve very high spectral efficiency [1]–[4]. Due to the very high data rate requirements in massive MIMO, downlink precoding prevails as essential for increased downlink data rates. In addition, in order for massive MIMO downlink precoding to be capable of offering its full benefits, quite accurate Channel State Information (CSI) is required at the BS, however at the UEs the CSI becomes more difficult to derive. Thus, CSI errors need to be taken into account at the UEs. Furthermore, complexity-reducing techniques such as dividing users in approximately orthogonal groups, e.g., Joint Spatial Division and Multiplexing (JSDM) [5], [6], dramatically simplify the downlink precoding problem, resulting in a hybrid downlink precoding scheme comprising: a) A prebeamforming matrix that divides users into different, semi-orthogonal groups (the inner precoder), and b) A Multi-User MIMO (MU-MIMO) precoder that aims at delivering a high data rate to each user within a group (the outer precoder). An efficient simplification to user grouping for JSDM with Finite Alphabet inputs (JSDM-FA) is proposed in [7], where the common sparse support of spatially similar users is employed to implement JSDM and a PGP-WG [8] near-optimal linear precoder is employed for each group to offer outer precoding with high performance and with significantly lower complexity. JSDM-FA achieves this much lower complexity due to its inherently sparse representation that allows for a much smaller dimension channel representation for each group. This is achieved by exploiting the sparse DFT matrix projection behavior of a BS antenna array response vector with many antenna elements $N_u \gg 1$, as is the case with massive MIMO. In addition, the JSDM-FA sparsity-unleashing methodology resorts to second order, i.e., slower varying CSI to derive the user groups, i.e., its inner precoder is slowly varying with time. However, [7] shows that without CFSDM, the users in a group need to jointly demodulate their information in order to achieve the benefits of PGP-WG. In this paper, we mitigate this essential joint demodulation issue, by proposing a ZF-PGP outer precoder.

In other past related work, [9] has presented an iterative algorithm for precoder optimization for sum rate maximization of Multiple Access Channels (MAC) using Kronecker MIMO channels. Furthermore, more recent work has shown that when only statistical CSI (SCSI)[1] is available at the transmitter, and in asymptotic conditions when the number of transmitting and receiving antennas grows large, but with a constant transmitting to receiving antenna number ratio, one can design the optimal precoder by looking at an equivalent constant channel and its corresponding adjustments as per the pertinent theory [12], and applying a modified expression for the corresponding ergodic mutual information evaluation over all channel realizations. This development allows for a precoder optimization under SCSI in a much easier way [12]. Finally, [8], [13] presented for the first time results for mutual

Manuscript received October 21, 2018; revised February 26, 2019 and May 13, 2019; accepted June 26, 2019. Date of publication July 5, 2019; date of current version October 16, 2019. This work was partially supported by NSF grant 1547155. The associate editor coordinating the review of this paper and approving it for publication was T. Q. Duong. *(Corresponding author: Thomas Ketseoglou.)*

T. Ketseoglou is with the Electrical and Computer Engineering Department, California State Polytechnic University, Pomona, CA 91768 USA (e-mail: tketseoglou@cpp.edu).

E. Ayanoglu is with the Center for Pervasive Communications and Computing, Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA 92697 USA (e-mail: ayanoglu@uci.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2019.2927205

[1]SCSI pertains to the case in which the transmitter has knowledge of only the MIMO channel correlation matrices, or in general slow-varying parameters [10], [11] and the thermal noise variance.

information maximizing linear precoding with large size MIMO configurations and QAM constellations. Such systems are particularly difficult to analyze and design when the inputs are from a finite alphabet, especially with QAM constellation sizes of $M \geq 16$. On the other hand, for non-linear precoding, recently [14], [15] have presented a robust, two-stage precoding approach, with channel estimation errors and for a hierarchical two-stage (hybrid) precoding design and with an outer Tomlinson-Harashima (TH) precoder per group that aims at minimizing the maximum mean square error (MSE) per user for each group. These non-linear outer precoders achieve near Dirty Paper Coding (DPC) results [16], i.e., they almost attain the channel capacity of a Gaussian Broadcast Channel (GBC),[2] due to the two-stage JSDM-type approach. Another line of work represents beamforming with security and energy harvesting employing Gaussian inputs [17], [18]. New robust path-following algorithms in the presence of imperfect channel estimation and multiantenna eavesdroppers that jointly optimize the energy and information beamformers are presented, which involve one simple convex quadratic program at each iteration. These algorithms are proposed in [17], while maximization methods for the minimum secrecy user rate under BS transmit power and UE minimum harvested energy constraints are presented in [18].

For finite alphabet inputs (FAI) and large antenna or modulation sizes, the mutual information optimizing linear precoder is impossible to derive due to the overall complexity involved that grows exponentially with the number of transmitter antennas [8]. However, near-optimal[3] FAI mutual information linear precoding approaches [7], [13], [19], [20] offer the possibility of achieving the practically highest possible spectral efficiency under perfect CSI or CSI with errors, in many cases significantly higher than the (non-precoded) group channel mutual information (GCMI) with full cooperation at the receiver.[4] The latter is a strict upper bound for many techniques, including Zero-Forcing (ZF), Regularized Zero-Forcing (RZF), and TH non-linear precoding [14], [16]. On the other hand, as [7] has shown, the linear precoding approach requires cooperation at the demodulation of each group and perfect CSI. It is thus still an open and important problem to address the issue of near-optimal downlink precoding in conjunction with user-independent decoding, i.e., without Multiple Access Interference (MAI) present at the UEs and under mismatched decoding at the UE. A survey paper on linear precoding techniques can be found in [19].

In this paper, we present near-optimal[5] linear precoding techniques for downlink massive MIMO, suitable for QAM with constellation size $M \geq 16$ and CSI at the BS with either a Uniform Linear Array (ULA) or a Uniform Planar Array (UPA) configuration, without the requirement for co-ordinated decoding at the UE. The proposed ZF-PGP method for finite alphabets with imperfect channel state information (CSI) at the receiver achieves very high gains over ZF Precoding techniques, while it simultaneously offers individually separate streams to reach individual User Equipment (UE). The proposed robust ZF-PGP scheme comprises an inner linear beamformer (BF), which relies on second order CSI and is based on the massive MIMO based Joint Division and Multiplexing for Finite Alphabets (JSDM-FA) to form orthogonal user groups, and an outer linear precoder that relies on instantaneous group CSI to form a Per-Group Precoding within groups (PGP-WG) with separate data streams to each user. This high performance remains valid, even with channel estimation errors at the UEs, i.e., the ZF-PGP precoder is robust. Thus, ZF-PGP has many advantages over other linear precoding techniques for massive MU MIMO.

The contributions of this paper can be summarized as follows:

1) It presents near-optimal linear hybrid (i.e., comprising an inner and outer precoder part) linear precoding on the downlink that includes: a) An inner linear beamformer (BF), which relies on second order CSI and is based on the massive MIMO based Joint Division and Multiplexing for Finite Alphabets (JSDM-FA) to form orthogonal user groups, and b) An outer linear ZF-PGP precoder that relies on instantaneous group CSI to form a Per-Group Precoding within groups (PGP-WG) with separate data streams to each user while it achieves fully separated data streams for each user.
2) It shows that the presented approach can achieve the full benefit of PGP-WG in high $\mathrm{SNR}$, i.e., it achieves the same power efficiency, which was shown to be near-optimal in the past.
3) It studies the problem of errors between the PGP-WG ideal (perfect CSI) optimal precoder employed for each group at the BS and the one resulting from estimating the channel with errors, i.e., with imperfect CSI, and shows only marginal loss for power efficiency.
4) For the type of channels that are present in massive MIMO, it shows that ZF-PGP significantly outperforms other commonly employed linear precoding techniques such as ZF, RZF, and other suboptimal precoding techniques, even with imperfect CSI, and it even achieves significantly higher performance than GCMI in medium to high $\mathrm{SNR}$.[6]

The paper is organized as follows: Section II presents the system model and problem statement together with a short introduction of the JSDM-FA channel representation [7]. Section III presents the ZF-PGP precoder concept that allows for optimized mutual information precoding in each group, while it distributes the information to each user independently, i.e., without any cross-user interference. We also include a short description of PGP-WG in order to enhance readability. Then, the ZF-PGP concept is generalized to cases where the ZF-PGP precoder is estimated at the UE, i.e., under a downlink

[2]This GBC capacity (system throughput) [16] precludes receiver cooperation.

[3]Please see [19] for more information on the meaning of near-optimal linear precoding.

[4]We use the term GCMI to mean the input-output mutual information with equally likely input symbols and without any type of precoding applying to the input of the MIMO channel.

[5]The reason we use this term is due to the fact that the ZF-PGP method is shown herein to converge to the near-optimal PGP-WG [13], [19] one in high $\mathrm{SNR}$.

[6]This is an additional reason to call ZF-PGP near-optimal.

channel estimation with errors model. In Section IV, numerical results are presented for ZF-PGP with and without CFSDM and for ideal CSI as well as CSI with errors. Comparisons with many other precoding techniques are also made. Finally, our conclusions are presented in Section V.

Notation: We use small bold letters for vectors and capital bold letters for matrices. $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathbf{A}^*$, $\mathbf{A}_{\cdot,i}$, $\mathbf{A}_{i,\cdot}$, 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, $\mathrm{tr}(\mathbf{A})$ 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 $N$, $\mathrm{diag}[x_1, \cdots, x_k]$ is the diagonal matrix with main diagonal equal to vector $[x_1, \cdots, x_k]^T$, and $\mathbf{I}$ denotes a generic identity matrix. We use $\mathbf{h}_{u,g,k,n}$ for the uplink channel of user $k$'s antenna $n$ in group $g$. $\mathbf{H}_g$ is the uplink channel of group $g$, while $\tilde{\mathbf{H}}_{g,v}$ is its projection to the Virtual Channel Model (VCM) basis. The overall virtual channel, reduced size representation of $\mathbf{H}_g$ is denoted by $\mathbf{H}_{g,v}$. $\mathbf{H}_{d,g}$ represents the overall downlink channel for group $g$, i.e., due to time division duplex reciprocity $\mathbf{H}_{d,g} = \mathbf{H}_g^H$. For a ULA or a UPA, $\mathbf{H}_{u,g,k,n}$ represents the uplink channel of user $k$'s antenna $n$ in group $g$, then $\tilde{\mathbf{H}}_{u,g,k,n}$ is its projection to the two DFT matrices (VCM bases), $\tilde{\mathbf{h}}_{u,g,k,n}$ is its corresponding vectorized form.

## II. System Model and Problem Statement

In this section, we present the system model, the problem we aim at solving, and the methodology of the solution. The notation and model employed is that of the sparsity-unleashing JSDM-FA methodology originally proposed in [7]. We extend previously presented results by including large-scale channel effects [21], including path loss and shadowing in our model.

### A. Downlink Precoding Without Grouping

Consider the downlink precoding equation on a narrowband, i.e., small-scale flat-fading[7] massive MIMO system that also includes large-scale channel effects such as propagation loss and shadowing, and with a single cell of $K$ users and without grouping by slightly extending results in [5], [7]

$$\mathbf{y}_d = \mathrm{diag}[\sqrt{a_{p,1}a_{s,1}}, \cdots, \sqrt{a_{p,K_{eff}}a_{s,K_{eff}}}]\mathbf{H}_u^H\mathbf{P}\sqrt{2E_{s,u}}\mathbf{x}_d + \mathbf{n}_d, \quad (1)$$

where $\mathbf{y}_d$ is the downlink received vector over all users and antennas of size $K_{eff} \times 1$, with $K_{eff} = \sum_k N_{d,k}$, the total number of UE antennas over all users, $\mathbf{P}$ is the linear precoding matrix of size $N_u \times K_{eff}$, $\mathbf{x}_d$ is the $K_{eff} \times 1$[8] vector of transmitted symbols drawn independently from a QAM constellation, the downlink channel matrix $\mathbf{H}_d = \mathbf{H}_u^H$, with $\mathbf{H}_u$ is the $N_u \times K_{eff}$ uplink (small-scale) fading channel matrix from all $K$ users, employing $N_u$ receiving antennas at the BS, $N_{d,k}$ is the total number of antennas of users $k$ with $1 \le k \le K$, $E_{s,u}$ is the transmitted energy per symbol and $\mathbf{n}_d$ represents the independent, identically distributed (i.i.d.) complex circularly symmetric Gaussian noise of mean zero and variance per component $\sigma_d^2 = 2N_0$, where $N_0$ is the one-sided noise power spectral density (PSD) on the downlink. The large-scale channel effects for the receiving antenna $i$ ($1 \le i \le K_{eff}$) is $\sqrt{a_{p,i}a_{s,i}}$ with $a_{p,i}$, and $a_{s,i}$ being the $ith$ receiving antenna path loss factor, and shadowing factor, respectively. We see that by setting $\mathrm{diag}[\sqrt{a_{p,1}a_{s,1}}, \cdots, \sqrt{a_{p,K_{eff}}a_{s,K_{eff}}}]\mathbf{H}_u^H \doteq \tilde{\mathbf{H}}_u^H$, an equivalent channel is defined, and after normalizing (1) by $\sqrt{2E_{s,u}}$, we get

$$\tilde{\mathbf{y}}_d = \tilde{\mathbf{H}}_u^H\mathbf{P}\mathbf{x}_d + \tilde{\mathbf{n}}_d, \quad (2)$$

where $\tilde{\mathbf{n}}_d$ is i.i.d. complex circular symmetric Gaussian noise with variance per component $\tilde{\sigma}_d^2 = \frac{1}{\mathrm{SNR}_u}$, with the transmitted symbol $\mathrm{SNR}_u = \frac{E_{s,u}}{N_0}$. We thus see that large-scale channel effects can be accommodated in the existing small-scale oriented past precoding works [7], [8], [12]–[15], [20]. The input-output mutual information $I(\mathbf{x}_d; \mathbf{y}_d|\hat{\tilde{\mathbf{H}}}_u,\ \tilde{\mathbf{H}}_u)$ maximizing downlink precoding problem, where the conditioning by $\hat{\tilde{\mathbf{H}}}_u,\ \tilde{\mathbf{H}}_u$ means that the UEs have only access to the estimated channel, while the transmitter has access to the actual channel, can now be cast as

$$\begin{aligned} \underset{\mathbf{P}}{\text{maximize}} \quad & I(\mathbf{x}_d; \mathbf{y}_d|\hat{\tilde{\mathbf{H}}}_u,\ \tilde{\mathbf{H}}_u) \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{P}\mathbf{P}^H) = \sum_{k=1}^{K} N_{d,k} = K_{eff}, \end{aligned} \quad (3)$$

where the constraint is due to keeping the total power transmitted from the BS to all downlink users with precoding equal to the one without precoding, due to keeping the total transmitted power, over all antennas and symbols, constant. This is easy to see since the normalized transmitted symbol power without precoding is 1 for each downlink transmitting antenna and due to the fact that the total power transmitted at the BS with precoding is from (1), equal to $\mathbb{E}\left\{||\mathbf{x}_d||^2\right\} = \mathbb{E}\left\{\mathrm{tr}(\mathbf{x}_d^H\mathbf{P}^H\mathbf{P}\mathbf{x}_d)\right\} = \mathrm{tr}(\mathbf{P}\mathbf{P}^H)$, by using trace properties and the fact that $\mathbb{E}\left\{\mathbf{x}_d{\mathbf{x}_d}^H\right\} = \mathbf{I}$. When FA modulation is employed, e.g., QAM, the complexity involved in solving the problem in (3) is on the order of $M^{2N_u}$ [8], with $M$ being the modulation size. We thus see that the problem is of very high complexity at the transmitter. Thus, user grouping as originally proposed in [5] and in [7] for finite inputs with the sparsity-unleashing JSDM-FA methodology can dramatically reduce the complexity of downlink precoding and also increase performance, as further explained below.

### B. Downlink With JSDM-FA Grouping

When JSDM-FA is applied, users are divided in $G$ approximately orthogonal groups, each group $g$ ($1 \le g \le G$) comprising $|\mathcal{S}_g| \ll N_u$ VCMBs which are the projections of the actual channel to the VCM[9] basis vectors. This basis comprises

[7]Note that using the framework of [7] and employing Orthogonal Frequency Division Multiplexing (OFDM) in conjunction with CFSDM, frequency selective channels can be easily employed in our model, however this is beyond the scope of this paper.

[8]We assume that there is one symbol per receiving antenna here.

[9]The VCM as used here, is not a propagation channel model. It is the exact representation of the original channel in a different basis.

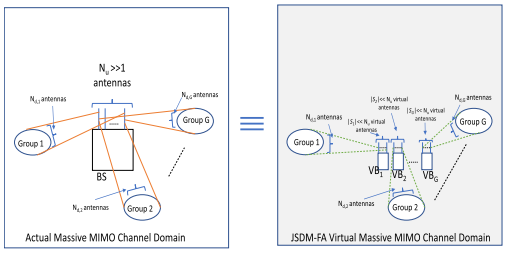



Fig. 1. Equivalency between the original channel model for a BS with $N_u \gg 1$ and the VCM in JSDM-FA. The original BS with all transmitting antennas communicating to all groups is equivalent to $G$ Virtual BS (VB), each employing significantly fewer antennas and each only affecting communication in one group, due to inter-group VCMB orthogonality.

a DFT matrix of order $N_u$. JSDM-FA [7] facilitates group determination and simultaneously reduces the dimensionality of the outer (intra-group) MU-MIMO precoding problem, based on a projection of the actual uplink channels to the DFT orthogonal base of size $N_u$. Then, the resulting groups under $N_u \gg 1$ have a sparse structure and domain of just a few VCMBs (Fig. 1). Note that JSDM-FA exploits the DFT-projected array response vector sparse behavior of a BS array, when $N_u \gg 1$ [7], which is the typical massive MIMO regime. We would like to stress that this sparse behavior is due to the Vandermonde structure of the antenna array response vector and the DFT matrix (see Appendix A and [7]) under massive MIMO. This is different than, e.g., the Central Limit Theorem (CLT) exploitation in deriving some of the early results on the advantages of massive MIMO [4]. The overall properties and advantages of JSDM-FA were extensively discussed in [7]. We shortly review the JSDM-FA concept here, then we employ it for the purpose of this paper. We denote the downlink channel matrix $\mathbf{H}_d = \mathbf{H}_u^H$, where $\mathbf{H}_u = [\mathbf{H}_1, \cdots, \mathbf{H}_G]$ is the $N_u \times K_{eff}$ uplink channel matrix from all $K$ users, employing $N_u$ receiving antennas at the BS, with $K_{eff} = \sum_g N_{d,g}$, where $N_{d,g}$ is the total number of antennas of all users in group $g$. Users have been divided into $G$ groups with $K_g$ users in group $g$ ($1 \leq g \leq G$), with user $k$ of group $g$ denoted as $k^{(g)}$. In this paper, we assume that each UE employs (without loss of generality) $N_{d,k^{(g)}} = 2$ antennas, with ($\sum_{g=1}^{G} K_g = K$), $\mathbf{H}_g = [\mathbf{H}_{g^{(1)}} \cdots \mathbf{H}_{g^{(K_g)}}]$ being group $g$'s uplink channel matrix of size $N_u \times N_{d,g}$, with $N_{d,g}$ comprising the total number of antennas in the group. As before, we assume that Time Division Duplexing (TDD) is employed in the system, to be able to exploit the reciprocity between the uplink and downlink channels. Let us denote by $C_{d,g,s}$ the total number of symbols destined to group $g$[10] for each transmission and that large-scale channel effects [21] are included as described above. The received data equation in the VCM domain for group $g$ ($1 \leq g \leq G$), due to inter group orthogonality, now becomes

$$\tilde{\mathbf{y}}_{d,g} = \mathbf{H}_{g,v}^H \mathbf{P}_g \mathbf{x}_{d,g} + \mathbf{n}_{d,g} \tag{4}$$

[10]By employing the Virtual additional Antenna Concept (VAAC) (Appendix F), up to 2 symbols per receiving antenna can be transmitted, resulting in $C_{d,g,s} = 2N_{d,g}$.

where due to UE geographic closeness in each group, we have made the reasonable assumption that a common path loss factor $a_{p,g}$, and a common shadowing factor $a_{s,g}$ apply to all members of each group. In (26), $\tilde{\mathbf{y}}_{d,g}$ has size $|\mathcal{S}_g| \times 1$, $\mathbf{H}_{g,v}^H$ is the size $N_{d,g} \times |\mathcal{S}_g|$ VCM representation downlink channel matrix for group $g$ ($1 \leq g \leq G$) (see Appendix A), $\mathbf{P}_g$ is the size $|\mathcal{S}_g| \times C_{d,g,s}$ outer precoding matrix, $\mathbf{x}_{d,g}$ is the size $C_{d,g,s} \times 1$ unit power, finite-alphabet QAM symbols, and $\mathbf{n}_{d,g}$ is the size $N_{d,g} \times 1$ (i.i.d.) complex circularly symmetric Gaussian noise of mean zero and variance per component $\sigma_d^2 = \frac{1}{\mathrm{SNR}_{s,d,g}}$, with $\mathrm{SNR}_{s,d,g} = \frac{a_{p,g} a_{s,g} E_{s,u}}{N_0} = a_{p,g} a_{s,g} \mathrm{SNR}_{s,u}$ being the received SNR of the group (see Appendix B).

Then, the optimal outer linear precoder $\mathbf{P}_g$ for group $g$ ($1 \leq g \leq G$) needs to satisfy

$$\begin{aligned} \underset{\mathbf{P}}{\text{maximize}} \quad & I(\mathbf{x}_{d,g}; \tilde{\mathbf{y}}_{d,g} | \hat{\mathbf{H}}_{g,v}, \ \mathbf{H}_{g,v}) \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{P}_g \mathbf{P}_g^H) = C_{d,g,s}, \end{aligned} \tag{5}$$

where the conditioning again indicates that accurate group channel CSI is employed at the transmitter, but only the estimated group channel CSI is employed at the receiver, and the constraint is due to keeping the total power emitted to the totality of downlink antennas and symbols with group precoding equal to the one without precoding, as it is easy to see that since the normalized received symbol energy is 1, the total power transmitted is kept constant and equal to $C_{d,g,s}$ for each group. Note that the original high size, high complexity problem in (3) is converted to $G$ much smaller size problems, with an outer precoder size of $|\mathcal{S}_g| \times C_{d,g,s}$, for $g = 1, \cdots, G$. As shown in [7], the performance of the system is very good, however the MU-MIMO PGP-WG precoder needs cooperation among all receivers in the group in order to achieve this near-optimal performance. More explicitly, PGP-WG requires that each user in a JSDM-FA group has knowledge of the received data of other users in the group in order to decode his data. An alternative we considered in [7] is to employ multiple carriers for each user, i.e., Combined Frequency and Space Division and Multiplexing (CFSDM), but at a significant cost of power efficiency. Thus, the problem of designing a high efficiency outer linear precoder that delivers independent data streams to each user and with FA inputs is very essential. Another alternative way is to investigate how the left singular vector matrix on the downlink channel of a group can be eliminated in the transmission phase, i.e., an integration of a ZFP into the PGP-WG framework needs to be introduced and studied. This integration of ZFP and PGP-WG in the outer precoder design results in the following equation for the combined ZF-PGP outer precoder of group $g$ ($1 \leq g \leq G$), $\mathbf{P}_g$,

$$\mathbf{P}_g = \mathbf{P}_{g,ZF} \mathbf{P}_{g,PGP}, \tag{6}$$

where $\mathbf{P}_{g,ZF}$, $\mathbf{P}_{g,PGP}$ represent the ZF, and the PGP-WG precoder, respectively. In the next section, we show that by employing this combined ZF-PGP approach to the outer precoder, we can achieve the same power efficiency as the original PGP-WG of JSDM-FA per group at high SNR, albeit without requiring any receiver cooperation, i.e., with much

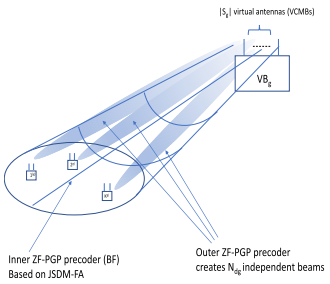



Fig. 2. Problem statement and proposed solution for the outer group precoder. After the second order CSI JSDM-FA based inner DFT precoder (BF) focuses the group's information to the appropriate group $g$ ($1 \leq g \leq G$), the outer ZF-PGP linear precoder creates $N_{d,g}$ independent FA data streams, one for each group receiving antenna, assuming that $C_{d,g,s} = 2N_{d,g}$ (see Appendix F).

smaller complexity. The rationale for this is the fact that ZF precoding creates independent data streams to each receiving antenna. Fig. 2 presents the fundamental idea of the problem tackled in this paper.

## III. Combined Outer ZF-PGP Robust Downlink Precoder

### A. Outer ZF-PGP for Near Optimal Independent Data Stream Distribution

For this section, we assume without loss of generality that each user's UE comprises two receiving antennas, $C_{d,g,s} = N_{d,g}$, and that the number of VCMBs available to a generic group $g$ ($1 \leq g \leq G$) is an even number. These assumptions are made in order to facilitate the description of the concept in conjunction with PGP-WG [22]. However, more general scenarios, e.g., with odd number of antennas at the UEs can be easily accommodated by appropriate subgrouping in PGP-WG [22]. As explained in the previous section, the focus of the ZF-PGP precoding in this paper is on groups, as the overall attributes and advantages of JSDM-FA were discussed extensively in [7] and only the solution to the intra-group MU-MIMO precoding is of relevance here. Focusing on single groups is widely used in the literature, e.g., [23], [24]. From (25), the receiving equation on the downlink of group $g$ is

$$\tilde{\mathbf{y}}_{d,g} = \mathbf{H}_{g,v}^H \mathbf{P}_g \mathbf{x}_{d,g} + \mathbf{n}_{d,g}, \tag{7}$$

$\mathbf{H}_{g,v}^H$ is the VCM group's downlink matrix of size $N_{d,g} \times |\mathcal{S}_g|$, $\mathbf{y}_{d,g}$ is the group's size $N_{d,g}$ reception vector, and $\mathbf{n}_g$ is the corresponding noise. Based on the ZF precoding theory, we need to assume that $|\mathcal{S}_g| \geq N_{d,g}$,[11] then the ZF precoder is given by

$$\mathbf{P}_{g,ZF} = \mathbf{H}_{g,v} \left( \mathbf{H}_{g,v}^H \mathbf{H}_{g,v} \right)^{-1}, \tag{8}$$

or by applying the Singular Value Decomposition (SVD)[12] of $\mathbf{H}_{g,v} = \mathbf{U}_{g,v} \mathbf{\Sigma}_{g,v} \mathbf{V}_{g,v}^H$, where $\mathbf{U}_{g,v}$, $\mathbf{\Sigma}_{g,v}$, and $\mathbf{V}_{g,v}$ are the matrices of left singular vectors, the singular values, and the right singular vectors of $\mathbf{H}_{g,v}$, respectively, we get

$$\mathbf{P}_{g,ZF} = \mathbf{U}_{g,v} \begin{bmatrix} \tilde{\mathbf{\Sigma}}_{g,v}^{-1} \\ \mathbf{0} \end{bmatrix} \mathbf{V}_{g,v} = \tilde{\mathbf{U}}_{g,v} \tilde{\mathbf{\Sigma}}_{g,v}^{-1} \mathbf{V}_{g,v}^H, \tag{9}$$

where $\tilde{\mathbf{\Sigma}}_{g,v}$, $\tilde{\mathbf{U}}_{g,v}$ represent the part of non-zero singular values of $\mathbf{\Sigma}_{g,v}$, and its corresponding columns of $\mathbf{U}_{g,v}$, respectively. However, this type of ZF precoder will be employing more power than the original system without precoding. In other words, the transmitted power required in ZF precoding is $\mathrm{tr}(\mathbf{P}_{g,ZF} \mathbf{P}_{g,ZF}^H) = \sum_{i=1}^{N_{d,g}} \frac{1}{s_{g,v,i}^2} \neq N_{d,g}$, where $s_{g,v,i}$ is the $i$th non-zero singular value of $\mathbf{H}_{g,v}$. We thus need to normalize the precoder by $\gamma = \sqrt{\frac{N_{d,g}}{\mathrm{tr}((\mathbf{H}_{g,v}^H \mathbf{H}_{g,v})^{-1})}} = \sqrt{\frac{N_{d,g}}{\sum_{i=1}^{N_{d,g}} \frac{1}{s_{g,v,i}^2}}}$, resulting in the following receiving vector

$$\tilde{\mathbf{y}}_{d,g} = \gamma \mathbf{x}_{d,g} + \mathbf{n}_{d,g}, \tag{10}$$

which shows an SNR loss, due to the effect of smaller singular values. However, one can see an advantage of ZF precoding due to the creation of $N_{d,g}$ independent streams in the above equation. In this paper, besides comparing the new ZF-PGP precoder to the ZF one, we will also employ the following additional benchmark precoders: a) The RZF precoder (RZFP) [25], b) The Serial Sub-group Combining precoder (SSCP) which normalizes the overall power of the ZF precoder (ZFP) every two consecutive singular values, and c) Parallel Sub-group Combining precoder (PSCP) which normalizes the ZFP power by grouping pairs of extreme singular values. In the following development of the ZF-PGP precoder, we assume that the BS knows the channels perfectly. However, this perfect CSI at the BS assumption is made to allow for the highest ZF-type precoding performance, because when this is not true, the ZF-type precoders considered herein will suffer from cross-interference that will result in significant loss in performance. For the ZF-PGP though, since we have the downlink estimation issue explained below, we can get a sense of its overall performance even with estimated, erroneous CSI, by increasing the level of errors on the downlink alone. This is done in the next section.

*Theorem 1: For the outer ZF-PGP precoder and under the JSDM-FA sparsity-unleashing methodology, the optimal PGP-WG precoder is determined by re-weighting the optimal PGP powers of the VCM channel, due to the presence of the ZF part, while the right PGP-WG singular vectors remain the same as the original PGP-WG solution, i.e., the ones determined without considering a ZF precoder part.*

*Proof:* The downlink received vector equation, after projecting to the VCM, for group $g$ ($1 \leq g \leq G$) is

$$\mathbf{y}_{d,g} = \mathbf{H}_{g,v}^H \mathbf{P}_{g,ZF} \mathbf{P}_{g,PGP} \mathbf{x}_g + \mathbf{n}_g, \tag{11}$$

where $\mathbf{P}_{g,ZF}$, $\mathbf{P}_{g,PGP}$ represent the ZF, and PGP-WG[13] precoder, respectively. Here, the ZF part is employed in order to equalize the MIMO channel. It is well-known that a mutual information maximizing linear precoder only depends on the singular values of the channel [20]. We use the fact that the

[11] In the opposite case the ZF precoder does not exist, however one can apply CFSDM in conjunction with ZF precoding to smaller size sub-groups, in order to be able to offer service to all users in the group.

[12] As a convention, we always employ the SVD with singular values ranked from maximum to minimum going down the main diagonal, and thus the singular vectors follow the same ranking. We call this the natural SVD in the sequel.

[13] A short review of PGP-WG [8] is included in Appendix E.

ZFP part satisfies $\mathbf{H}_{d,g}^H \mathbf{P}_{g,ZF} = \mathbf{I}_{N_{d,g}}$. Then, the receiving equation for all antennas after introducing the PGP part can be written as

$$\mathbf{y}_{d,g} = \mathbf{P}_{g,PGP}\mathbf{x}_{d,g} + \mathbf{n}_{d,g}. \tag{12}$$

Using SVD the expressions for $\mathbf{P}_{g,ZF}$, $\mathbf{P}_{g,PGP}$ can be written as $\mathbf{P}_{g,ZF} = \mathbf{U}_{g,v}\begin{bmatrix}\tilde{\mathbf{\Sigma}}_{g,v}^{-1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}_{g,v}^H$, and $\mathbf{P}_{g,PGP} = \mathbf{\Sigma}_{g,PGP}\mathbf{V}_{g,PGP}^H$, respectively.[14] Thus,

$$\begin{aligned}\mathbf{P}_{g,ZF}\mathbf{P}_{g,PGP} &= \mathbf{U}_{g,v}\begin{bmatrix}\tilde{\mathbf{\Sigma}}_{g,v}^{-1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}_{g,v}^H\mathbf{\Sigma}_{g,PGP}\mathbf{V}_{g,PGP}^H\\ &= \mathbf{U}_{g,v,N_{d,g}}\tilde{\mathbf{\Sigma}}_{g,v}^{-1}\mathbf{V}_{g,v}^H\mathbf{\Sigma}_{g,PGP}\mathbf{V}_{g,PGP}^H, \end{aligned}\tag{13}$$

where in order to simplify notation, we defined $\mathbf{U}_{g,v,N_{d,g}}$ to represent the first $N_{d,g}$ highest singular vectors of $\mathbf{U}_{g,v}$. Note that due to employing two precoders in tandem, the power constraint for the combined $\mathbf{P}_{g,ZF}\mathbf{P}_{g,PGP}$ precoder that comprises both the ZF and PGP parts becomes $\mathrm{tr}(\mathbf{P}_{g,ZF}\mathbf{P}_{g,PGP}\mathbf{P}_{g,PGP}^H\mathbf{P}_{g,ZF}^H) = N_{d,g}$, or equivalently, after using properties of the matrix trace, we can write for the overall precoder power constraint

$$\mathrm{tr}\left(\mathbf{V}_{g,v}\tilde{\mathbf{\Sigma}}_{g,v}^{-2}\mathbf{V}_{g,v}^H\mathbf{\Sigma}_{g,PGP}^2\right) = N_{d,g}. \tag{14}$$

Let us define $\mathbf{M} \doteq \mathbf{V}_{g,v}\tilde{\mathbf{\Sigma}}_{g,v}^{-2}\mathbf{V}_{g,v}^H$, then equivalently the power constraint becomes

$$\mathrm{tr}\left(\mathbf{M}\mathbf{\Sigma}_{g,PGP}^2\right) = N_{d,g}. \tag{15}$$

We thus see that due to the ZF precoder part, the overall precoding power constraint becomes more involved, resulting in a much more difficult optimization problem. In contrast, the original PGP-WG method would entail power constraints similar to the ones in (37) which are much easier to accommodate. In addition, due to the presence of small singular values in the channel, especially in correlated channels, there will be a significant power loss in the power constraint described by (14). Now, since $\mathbf{\Sigma}_{g,PGP}$ is a diagonal matrix, we can write (15) as

$$\sum_{m=1}^{N_{d,g}} \mathbf{M}_{m,m}s_{PGP,m}^2 = N_{d,g}, \tag{16}$$

where $s_{PGP,m}$ is the $m$th singular value of $\mathbf{P}_{g,PGP}$. On the other hand, the performance of a mutual information maximizing precoder is determined by $\mathbf{\Sigma}_{g,PGP}$ and $\mathbf{V}_{g,PGP}$ according to (12). Incorporating the new re-weighted constraints of (16) into the PGP-WG framework presented in Appendix E is feasible, but it would result in additional complexity. However, using $w_m \doteq \sqrt{|\mathbf{M}_{m,m}|}$, for $m = 1, 2, \cdots N_{d,g}$, which are calculated in closed form in Appendix C, a simplification is possible by casting the power constraint on the PGP part as an ordinary one. By defining $\mathbf{\Sigma}_{g,v,eff} = \mathrm{diag}[s_{g,v,eff,1}, \cdots, s_{g,v,eff,N_{d,g}}]$, $\mathbf{\Sigma}_{g,PGP,eff} = \mathrm{diag}[s_{g,PGP,eff,1}, \cdots, s_{g,PGP,eff,N_{d,g}}]$, where $s_{g,v,eff,m} \doteq \frac{1}{w_m}$, $s_{g,PGP,eff,m} \doteq s_{PGP,m}w_m$, respectively, for $m = 1, 2, \cdots, N_{d,g}$, we have the equivalent model fitting within the original PGP-WG type power constraint as follows

$$\begin{aligned}\tilde{\mathbf{y}}_{d,g} &= \mathbf{\Sigma}_{g,v,eff}\mathbf{P}_{g,PGP,eff}\mathbf{x}_{d,g} + \mathbf{n}_{d,g}\\ &= \mathbf{\Sigma}_{g,v,eff}\mathbf{\Sigma}_{g,PGP,eff}\mathbf{V}_{g,PGP}^H\mathbf{x}_{d,g} + \mathbf{n}_{d,g}, \end{aligned}\tag{17}$$

where the optimal precoder needs to solve

$$\begin{aligned}&\underset{\mathbf{\Sigma}_{g,PGP,eff},\mathbf{V}_{g,PGP}}{\text{maximize}} \quad I(\mathbf{x}_{d,g};\tilde{\mathbf{y}}_{d,g})\\ &\text{subject to} \quad \mathrm{tr}(\mathbf{\Sigma}_{g,PGP,eff}^2) = N_{d,g}. \end{aligned}\tag{18}$$

Thus, the optimization is within the framework of PGP-WG type of problems. Then, one can readily apply the original PGP-WG algorithm to solve (18). This way, a major simplification to the solution of the original ZF-PGP problem has been achieved. □

Based on the previous theorem proof, it is straightforward to prove the following corollary.

*Corollary 1: The outer ZF-PGP precoder is equivalently determined by using an effective channel singular value matrix given by $\mathbf{\Sigma}_{g,v,eff}$, followed by an ordinary PGP-WG precoder.*

This corollary demonstrates that the actual ZF-PGP precoder is determined in a straightforward manner by using an effective channel $\mathbf{\Sigma}_{g,v,eff}$ and then by determining its corresponding WG-PGP. The diagonal effective channel singular value matrix is determined by the inverses of the $w_m$, $m = 1, 2, \cdots, N_{d,g}$, which are given by (see Appendix A) $w_m = \sqrt{|\mathbf{M}_{m,m}|} = \sqrt{\sum_{m'=1}^{N_{d,g}} \frac{|(\mathbf{V}_{g,v})_{m,m'}|^2}{s_{g,v,m'}^2}}$. The overall pseudo code for the ZF-PGP precoder is shown in Algorithm 1 above.

**Algorithm 1** Outer ZF-PGP Precoder Algorithm

**procedure** OUTER ZF-PGP PRECODER($\mathbf{H}_{g,v}^H$)
  Determine the SVD of $\mathbf{H}_{g,v} = \mathbf{U}_{g,v}\mathbf{\Sigma}_{g,v}\mathbf{V}_{g,v}^H$
  Determine $\mathbf{U}_{PGP} = \mathbf{I}$ using (34)
  Determine $w_m \doteq \sqrt{|\mathbf{M}_{m,m}|}$, for $m = 1, 2, \cdots N_{d,g}$, using (29)
  Determine $\mathbf{\Sigma}_{g,v,eff} = \mathrm{diag}[s_{g,v,eff,1}, \cdots, s_{g,v,eff,N_{d,g}}]$ and $\mathbf{\Sigma}_{g,PGP,eff} = \mathrm{diag}[s_{g,PGP,eff,1}, \cdots, s_{g,PGP,eff,N_{d,g}}]$, where $s_{g,v,eff,m} \doteq \frac{1}{w_m}$, $s_{g,PGP,eff,m} \doteq s_{PGP,m}w_m$, respectively, for $m = 1, 2, \cdots N_{d,g}$
  Solve the $N_{d,g}$ PGP-WG sub-problems emanating from (18) and determine the PGP-WG precoder by applying (35) and (36)
**end procedure**

Regarding the performance advantages of ZF-PGP, we need to stress the following facts. For uncorrelated MIMO channels, i.e., when full independence exists between the $\mathbf{H}_{g,v}$ entries and with the condition number of the channel being relatively low [26], then a ZF-PGP precoder does not perform well, while a ZF one performs extremely well. The original pioneering works of [27], [28], and others have addressed independent Gaussian channels and capitalized on the very desirable property of the ZF precoder to offer separate data streams to each user. When a MIMO channel presents correlations, for example by using a channel model for ULA herein and with

[14]The left singular vectors of a mutual information maximizing precoder with power constraint is $\mathbf{I}_{N_{d,g}}$ for the channel in (12).

a high condition number, due to massive MIMO, ZF and RZF gains evaporate rapidly and a PGP-WG precoder performs much better as demonstrated in the next section. In addition, when the number of UEs becomes close to the number of available VCMBs in the group $\mathcal{S}_g$, although a PGP-WG precoder achieves excellent performance [7], it requires a joint decoding at the receiving UEs, and thus it becomes impractical. Under these conditions, a ZF-PGP outer MU-MIMO precoder can achieve excellent performance with simplified receiver operation due to the independent data streams sent to each UE.

### *B. Mismatched Downlink Decoding*

We focus on errors on the downlink of the system with each UE being sent pilot data using the optimized ZF-PGP precoder the BS has already determined, so that the UE can estimate the employed precoder, then use the precoder estimate in order to decode the received information data. This is due to the fact that the UE needs to know the optimal precoder determined by the BS in order to decode the data. This estimation process contains errors. In order to assess the system's robustness to this kind of errors, we employ a more general method used extensively in the literature [14], [29] to model the estimation errors. The applied technique helps decouple the employed estimation technique from its corresponding performance, as explained in Section IV. Based on the imperfect channel estimate, the subgroup receiver determines the precoder and employs it at the decoding process. Due to the estimation errors, this results in a mismatch in the UE decoding process. We calculate the corresponding mismatched mutual information with the downlink receiver operating under mismatch, due to the erroneous estimate and assess the corresponding performance loss. We summarize the JSDM-FA sparsity-unleashing representation methodology from [7] below to facilitate the readability of the paper.

Let us consider the downlink mismatched precoded information decoding under ZF-PGP. Denote by $\mathbf{P}$, $\hat{\mathbf{P}}$, $p_{\mathbf{P},\mathbf{y}}(\mathbf{y}|\mathbf{x})$, $p_{\hat{\mathbf{P}},\mathbf{y}}(\mathbf{y}|\mathbf{x})$, the optimal precoder employed on the downlink, the estimation-based precoder, the pdf of the actual downlink PGP-WG subgroup channel, and the pdf of the estimated downlink PGP-WG subgroup channel, respectively. The mismatched mutual information, $I_{msm}(\mathbf{x};\mathbf{y})$, on the subgroup downlink under the presented mismatch can be found employing the results in [30] and using

$$\begin{aligned} I_{msm}(\mathbf{x};\mathbf{y}) = &-\frac{1}{M^{N_t}}\sum_{\mathbf{x}_m}\mathbb{E}_{\mathbf{P},\mathbf{y}|\mathbf{x}_m}\left\{\log_2\left(\sum_{\mathbf{x}_k}\frac{1}{M^{N_t}}\right.\right. \\ &\left.\left.\times\frac{1}{\pi^{N_r}\sigma^{2N_r}}\exp(-\frac{1}{\sigma^2}||\mathbf{y}-\hat{\mathbf{P}}\mathbf{x}_k||^2)\right)\right\} \\ &+\frac{1}{M^{N_t}}\sum_{\mathbf{x}_m}\mathbb{E}_{\mathbf{P},\mathbf{y}|\mathbf{x}_m}\left\{\log_2\left(\frac{1}{\pi^{N_r}\sigma^{2N_r}}\right.\right. \\ &\left.\left.\times\exp(-\frac{1}{\sigma^2}||\mathbf{y}-\hat{\mathbf{P}}\mathbf{x}_k||^2)\right)\right\}, \end{aligned} \tag{19}$$

where $\mathbf{P} = \tilde{\mathbf{\Sigma}}_{g,v,eff}\mathbf{\Sigma}_{g,PGP,eff}\mathbf{V}_{g,PGP}$, $\hat{\mathbf{P}} = \hat{\tilde{\mathbf{\Sigma}}}_{g,v,eff}\hat{\mathbf{\Sigma}}_{g,PGP,eff}\hat{\mathbf{V}}_{g,PGP}$, respectively, and where the metrics within the argument of the $\log_2$ function reflect the fact that estimated channels are employed and in our case $N_r = 2$ receiving antennas are employed by each subgroup. Realizing that under $\mathbf{P},\mathbf{y}|\mathbf{x}_m$, $\mathbf{y} = \mathbf{P}\mathbf{x}_m + \mathbf{n}$, and after some straightforward simplifications, we can easily find that

$$\begin{aligned} I_{msm}(\mathbf{x};\mathbf{y}) = N_t\log_2(M) - N_r - \frac{1}{M^{N_t}}\sum_{\mathbf{x}_m}\mathbb{E}_{\mathbf{n}}\left\{\log_2\left(\sum_{\mathbf{x}_k}\right.\right. \\ \left.\left.\times\exp(-\frac{1}{\sigma^2}||\mathbf{n}+\mathbf{P}\mathbf{x}_m-\hat{\mathbf{P}}\mathbf{x}_k||^2)\right)\right\} \\ -\frac{1}{\sigma^2}||\mathbf{P}-\hat{\mathbf{P}}||_F^2, \end{aligned} \tag{20}$$

where $||\mathbf{P}-\hat{\mathbf{P}}||_F^2 = \mathrm{tr}\left((\mathbf{P}-\hat{\mathbf{P}})(\mathbf{P}-\hat{\mathbf{P}})^H\right)$ is the Frobenius norm square of $\mathbf{P}-\hat{\mathbf{P}}$. The third term in the above equation can be very accurately approximated by employing the Gauss-Hermite approximation in a fashion similar to [8], with a simple modification to the expression presented there, i.e., substituting $\mathbf{P}\mathbf{x}_m - \hat{\mathbf{P}}\mathbf{x}_k$ for $\mathbf{P}(\mathbf{x}_m - \mathbf{x}_k)$. The details are omitted here due to lack of space. The final expression upon employing the Gauss-Hermite approximation is presented in Appendix D.

We observe that due to the channel estimation induced mismatch, the mutual information expression experiences two impacts: First, it has a modified term in the summation of the expectations over $\mathbf{n}$ that reflects the mismatch between the actual channel and its estimation, and second, it contains a Frobenius norm square of the MSE of the estimation, normalized by $\mathrm{SNR}_{s,d}$. This second term shows that there is a limit of how well an increase of $\mathrm{SNR}_{s,d}$ will compensate for estimation errors.

For ZF-PGP to be efficient, very accurate CSI is assumed available at the BS for the reasons explained above. However, in 5G due to the concept of Cloud-RAN (C-RAN) there are multiple reasons that open the possibility of additional error sources to be present. For example, although an accurate channel estimate needs to be present at the BS, the ZF-PGP optimal precoder determination might take place at the C-RAN and employing cheap, inaccurate processors, or the additional propagation delay between the C-RAN and the BS might result in a suboptimal PGP-WG precoder. In order to assess the system's robustness to this kind of error, we employ a more general method used extensively in the literature to model the estimation errors [14], [29]. The applied technique helps decouple the employed estimation technique from its corresponding performance, as explained in Section IV. Based on the imperfect channel estimate, the subgroup receiver determines the precoder and employs it at the decoding process. Due to the estimation errors, this results in a mismatch and then we calculate the corresponding mismatched mutual information with the downlink receiver operating under mismatch, due to the erroneous estimate and assess the corresponding performance loss.

## IV. Numerical Results

In this section, we present numerical results regarding the precoders presented above and their comparison with

other widely used precoders which we use as benchmarks. More explicitly, we use the ZFP, SSCP, PSCP, and RZFP as benchmarks. In addition, we compare the performance of the ZF-PGP with PGP-WG [7]. In all cases, there are group channels created based on the JSDM-FA sparsity-unleashing methodology [7] which we employ to get the desired results using the same approach presented in [8]. We only show results for QAM with $M = 16$ and Quadrature Phase Shift Keying (QPSK) (in one scenario), however the same approach we used in [7], [8] for higher modulation size of $M = 64$ can be used here as well. More general cases than using two antennas per UE are easily accommodated using methods presented in [22], by splitting appropriately the number of UE antennas in groups. The model used for CSI errors on the downlink is as follows. In all cases we use the VAAC that we have used in previous work [7], [8]. When results with mismatched decoding at the receivers are used, the following model applies. The BS determines the near-optimal precoder for this subgroup and sends this to subgroup $s_g$ $(1 \leq s_g \leq K_g)$, $1 \leq g_s \leq \frac{N_{d,g}}{2}$ of the JSDM-FA sparsity-unleashing methodology using the ZF-PGP approach as

$$\mathbf{Y}_{s_g} = \mathbf{\Sigma}_{s_g} \mathbf{P}_{s_g,o} \mathbf{X}_{p_{s_g}} + \mathbf{N}_{s_g}, \tag{21}$$

where $\mathbf{Y}_{g_s}$, $\mathbf{\Sigma}_{g_{g,v_s}}$, $\mathbf{P}_{g_s,o}$, $\mathbf{X}_{p_{g_s}}$, and $\mathbf{N}_{g_s}$ represent the received data, the channel singular values for the PGP-WG subgroup, the ZF-PGP determined precoder for the PGP-WG subgroup, the pilots, and the noise, respectively, all for subgroup $g_s$. In order to model the errors of the estimation process we use for the estimate of $\mathbf{P}_{g_s,o}$, $\hat{\mathbf{P}}_{g_s,o}$, the following model [14], [29]:

$$\hat{\mathbf{P}}_{s_g,o} = \sqrt{1-\tau^2}\mathbf{P}_{g_s,o} + \tau \mathbf{N}_{est}, \tag{22}$$

with $\mathbf{N}_{est}$ representing the random errors being a $2 \times 2$ matrix of complex, independent, Gaussian random variables of mean 0 and variance 1, and with $\tau$ being an estimation quality parameter in $[0, 1]$ with 0 representing ideal estimation and 1 representing fully erroneous estimation. We use $\tau = 0.1,\ 0.2,$ and $0.3$ in the results in this paper. For cases where $\tau \geq 0.2$ is employed, we can make the argument that due to the higher error level, one can model possible channel estimation errors as well and use the corresponding results as showing robustness to both type of errors (at the BS and UE). The groups presented are developed from a massive MIMO system with $N_u = 100$ antennas using the JSDM-FA sparsity-unleashing methodology presented in [7], according to the relevant equations presented in Section II. When no comparison is made with the actual channel GCMI in the results, in general, we need to employ the symbol received $\mathrm{SNR}$, $\mathrm{SNR}_{s,d,g}$ as the resource variable, however in order to be able to compare our results fairly with pre-existing precoding works, we set $\mathrm{SNR}_{s,d,g} = \mathrm{SNR}_{s,u}$, i.e., path loss and shadowing are neglected. When comparisons are made to the actual channel GCMI in the results, we employ as resource the information bit based $\mathrm{SNR}$, denoted as $\mathrm{SNR}_b$.

In Fig. 3 we present results using group $G1$ of [7] for the achievable mutual information with PGP-WG, ZFP, SSCP, PSCP, and RZF, all with ideal CSI. The resulting downlink channel is a $4 \times 8$ MIMO channel, i.e., $N_{d,g} = 4,\ |\mathcal{S}_g| = 8$,

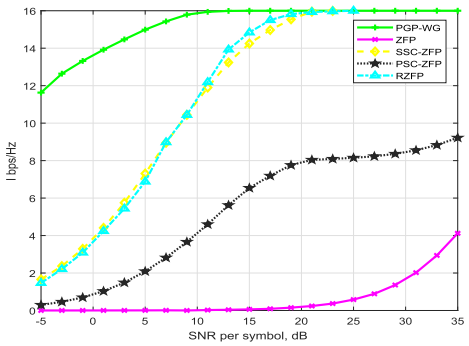



Fig. 3. $I(\mathbf{x};\mathbf{y})$ results for PGP-WG, ZFP, SSCP, PSCP, and RZFP cases for the channel in $G_1$ of [7] in conjunction with QAM $M = 16$ modulation.

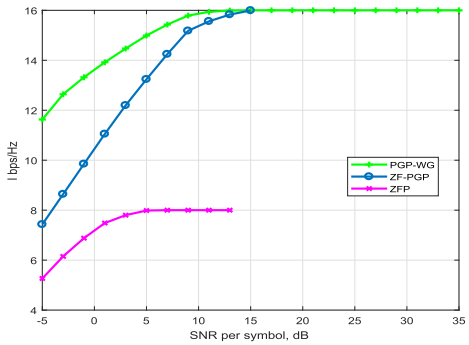



Fig. 4. $I(\mathbf{x};\mathbf{y})$ results for ZF-PGP, PGP-WG, and ZFP cases for the channel in $G_1$ of [7] in conjunction with QAM $M = 16$ modulation and CFSDM.

using an angular spread of $\pm 4°$. We see what is typical behavior in correlated channels, i.e., PGP-WG offers much higher mutual information than all the others, albeit at the cost of requiring demodulation cooperation among the UEs. It thus becomes appealing to consider precoding techniques that aim to bridge the gap between ZFP and PGP-WG type of precoding techniques. This channel presents very high correlation and ZF-PGP cannot offer an efficient solution. Instead, we resort to CFSDM in order to reduce the channel correlation and present our first ZF-PGP results in Fig. 4. Here we present results for the same channel as in Fig. 3, but employing two different subcarriers, in a CFSDM fashion, resulting in two CFSDM groups of 1 UE each. We use perfect CSI and employ VAAC for the ZF-PGP (explained in Appendix F) with $N_{TV} = 2$ additional transmitting antennas added per CFSDM group. We divide the sum of the two CFSDM groups' mutual information by two (the number of subcarriers employed in CFSDM) in order to calculate the resulting spectral efficiency. We see that ZF-PGP significantly outperforms ZFP, but due to the exploitation of two subcarriers, the performance of ZF has been improved from the non-CFSDM case dramatically. In addition, the performance of ZF-PGP approaches the performance of PGP-WG at high $\mathrm{SNR}$.

In Fig. 5 we present results for another small group. Here the group is created based on the JSDM-FA framework [7] and comprises $N_{d,g} = 4,\ |\mathcal{S}_g| = 6$, the group was developed with an angular spread of $\pm 10°$. We employ both $M = 16$ QAM and Quadrature Phase Shift Keying (QPSK) modulation

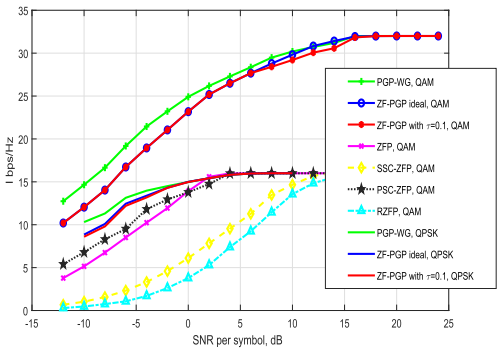



Fig. 5. $I(\mathbf{x};\mathbf{y})$ results for ZF-PGP (ideal and with channel estimation errors ($\tau = 0.1$)), PGP-WG, ZFP, SSCP, PSCP, and RZFP cases for the $4 \times 6$ group channel in conjunction with QAM $M = 16$ and QPSK modulation.

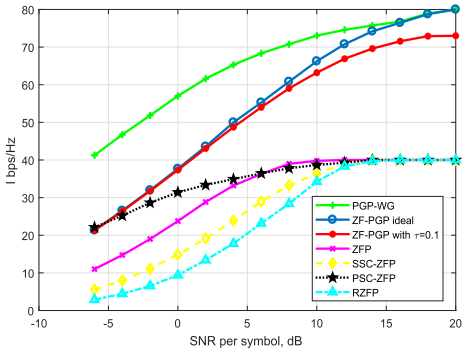



Fig. 6. $I(\mathbf{x};\mathbf{y})$ results for ZF-PGP (ideal and with channel estimation errors ($\tau = 0.1$)), ZFP, SSCP, PSCP, and RZFP cases for the $30 \times 14$ CFSDM subgroup 3 channel in conjunction with QAM $M = 16$ modulation and CFSDM.

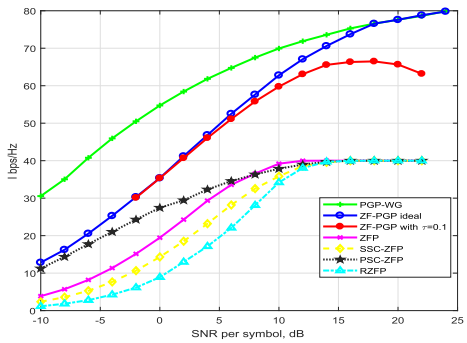



Fig. 7. $I(\mathbf{x};\mathbf{y})$ results for ZF-PGP (ideal and with channel estimation errors ($\tau = 0.1$)), PGP-WG, ZFP, SSCP, PSCP, and RZFP cases for the $40 \times 12$ group CFSDM subgroup 1 channel and in conjunction with QAM $M = 16$ modulation CFSDM.

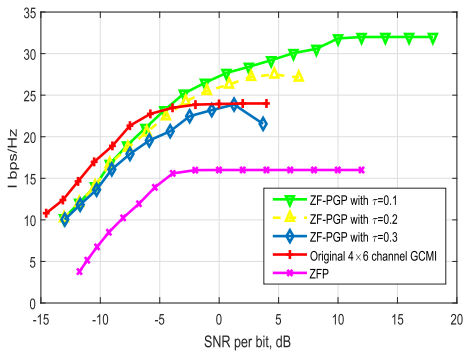



Fig. 8. $I(\mathbf{x};\mathbf{y})$ results with respect to $\mathrm{SNR}_b$ for ZF-PGP (with channel estimation errors ($\tau = 0.1,\ 0.2,$ and $0.3$)), ZFP, and the original channel GCMI for the $4 \times 6$ group channel and in conjunction with QAM $M = 16$ modulation.

for the PGP-WG and ZF-PGP cases. The results presented for ZF-PGP take advantage of the VAAC that can be very efficiently employed by ZF-PGP and PGP-WG, but it is not possible for the other techniques presented. The example shown uses $N_{TV} = N_{d,g} = 4$ added at the transmitter based on VAAC. We observe the excellent performance of ZF-PGP, even with estimation errors. In addition, we see that in high $\mathrm{SNR}$, ZF-PGP follows PGP-WG very closely. The gain of ZF-PGP over ZFP with $\tau = 0.1$ level of errors are 42.8%, and 75%, at $\mathrm{SNR}_s = 5\ dB$, and $15\ dB$, respectively. Finally, QPSK exhibits the same type of behavior as $M = 16$ QAM, but its overall throughput is lower, due to its smaller constellation size, however we observe that the error behavior of QPSK is similar to the one with QAM.

In Fig. 6 we show results for a large group, i.e., $N_{d,g} = 30,\ |\mathcal{S}_g| = 14$, created with an angular spread of $\pm 4^\circ$ with the JSDM-FA framework. Because $N_{d,g} > |\mathcal{S}_g|$ in this case, CFSDM is required in order to implement the ZFP type of precoders. We divide the UE in three CFSDM subgroups each comprising 10 antennas. We present results for the third subgroup. The PGP-WG and ZF-PGP systems use $N_{TV} = 10$. We observe the excellent performance of ZF-PGP, even with estimation errors. In addition, we see that in high $\mathrm{SNR}$, ZF-PGP follows PGP-WG very closely. The gain of ZF-PGP with $\tau = 0.1$ level of errors over ZFP are 42.8%, and 75%, at $\mathrm{SNR}_s = 5\ dB$, and $15\ dB$, respectively. In Fig. 7 we show results for a system with $N_{d,g} = 40,\ |\mathcal{S}_g| = 12$, created with an angular spread of $\pm 4^\circ$ with the JSDM-FA framework. Here, again CFSDM is needed. Thus, we divide the groups into four equal size subgroups and simulate the first subgroup. The VAAC employs $N_{TV} = 10$. The resulting gain of ZF-PGP with $\tau = 0.1$ level of errors over ZFP at $\mathrm{SNR}_s = 10\ dB$ is greater than $50\%$. We observe that the estimation errors have a more profound effect in high $\mathrm{SNR}$ in this case, resulting in a reversing effect as the achievable spectral efficiency peaks in value, then it starts decreasing. This is a typical behavior under the model used, as the estimation errors are independent than the $\mathrm{SNR}$ in the model. This can be the case, for example, if the downlink channel estimation employs a constant estimation $\mathrm{SNR}$, independent of the data transmission $\mathrm{SNR}$.

In all cases considered so far, the gains offered by ZF-PGP in low $\mathrm{SNR}$ over ZFP are even higher than the ones in high $\mathrm{SNR}$.

In the rest of the results, we compare the performance of ZF-PGP with the original channel GCMI and ZFP, albeit based on $\mathrm{SNR}_b$. This approach gives a better idea about the power efficiency of different precoding techniques. In Fig. 6, 7, and 8 we present the corresponding results for the groups used in Fig. 3, 4, and 5, respectively.

In Fig. 8 we observe that for all $\mathrm{SNR}_b > 4\ dB$, ZF-PGP with error level equal to $0.1,\ 0.2,$ and $0.3$ offers higher

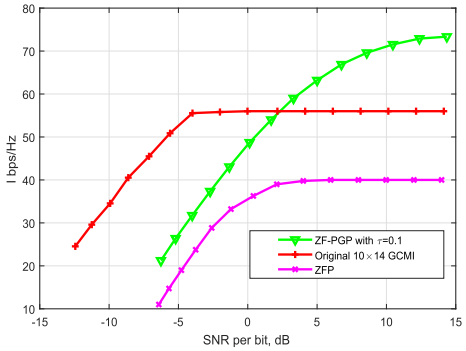



Fig. 9. $I(\mathbf{x};\mathbf{y})$ results with respect to $\mathrm{SNR}_b$ for ZF-PGP (with channel estimation errors ($\tau = 0.1$)), ZFP, and the original channel GCMI for the $30 \times 14$ group CFSDM subgroup 3 channel and in conjunction with QAM $M = 16$ modulation and CFSDM.

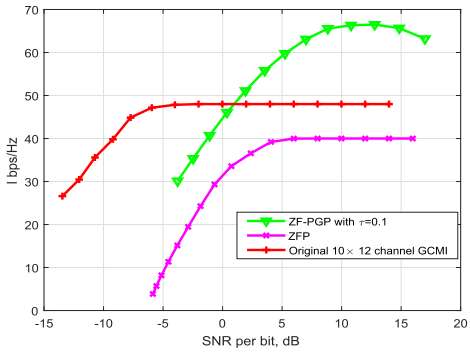



Fig. 10. $I(\mathbf{x};\mathbf{y})$ results with respect to $\mathrm{SNR}_b$ for ZF-PGP (with channel estimation errors ($\tau = 0.1$)), ZFP, and the original channel GCMI for the $40 \times 12$ group CFSDM subgroup 1 channel and in conjunction with QAM $M = 16$ modulation and CFSDM.

spectral efficiency than the original channel GCMI with ideal CSI at the receiving UEs. Due to the increased error levels, we can safely contend that if estimation errors are present on the uplink channel estimation, the ZF-PGP with estimation errors at both the BS and at the UEs will still perform significantly better than the ZFP with ideal CSI, a major conclusion in this paper. In Fig. 9 the same type of behavior is observed, i.e., for $\mathrm{SNR} > 3\ dB$ a ZF-PGP system with error level $\tau = 0.1$ offers higher spectral efficiency than the original channel GCMI with ideal CSI at the UEs. Finally, in Fig. 10 we see that ZF-PGP with error level $\tau = 0.1$ offers higher spectral efficiency than the original channel GCMI with ideal CSI at the UEs. In all cases considered with regards to $\mathrm{SNR}_b$, ZF-PGP with errors is much more power efficient than ZFP as shown by the plots. It is worthwhile stressing that PGP-WG employed in [7] is always equal or better in spectral efficiency over the original GCMI one with respect to power efficiency ($\mathrm{SNR}_b$).

## V. Conclusions

In this paper, we propose a new, robust combined Zero-Forcing Per-Group Precoding (ZF-PGP) method that achieves very high spectral efficiency gains in comparison with other massive MU-MIMO precoding techniques, while it simultaneously offers individually separate streams to reach individual UE, i.e., it obliterates the need for coordinated, joint decoding by the group's UEs. We show that for correlated channels and/or when there are more antennas in a group than the number of VCMBs in the group, the near-optimal ZF-PGP is determined from an effective channel singular value matrix which is calculated easily from the singular values and the right singular vectors of the group's uplink virtual channel in the JSDM-FA decomposition. We show that at high SNR ZF-PGP approaches the original PGP-WG near-optimal performance. In addition, ZF-PGP is shown to be robust to estimation errors. The gains of ZF-PGP over ZF type precoders and GCMI are documented by many examples and shown to be significant, i.e., higher than 50% in high SNR even with errors. Finally, ZF-PGP offers higher performance than the original channel GCMI one, although the latter is assumed with perfect CSI at the receiving UEs, thus demonstrating an overall excellent performance. Our future work will look at incorporating additional quality-of-service (QoS) constraints for each UE in a group in the outer precoding design as well as studying the impact of millimiter wave (mmWave) propagation effect to the ZF-PGP performance in a cell with random UE deployment.

## Appendix A
## JSDM-FA Low Dimension Inner Precoder

Assume without loss in generality that a ULA deployed at the BS along the $z$ direction and for flat fading,[15] i.e., $B < B_{COH}$, where $B$ and $B_{COH}$ are the RF signal bandwidth and the coherence bandwidth of the channel, respectively. The channel employed in JSDM-FA is similar to the Wide-Sense Stationary Uncorrelated Scattering (WSSUS) channel in [31] and the channel in [10]. Each user group on the uplink transmits from the same "cluster" of elevation angles $\theta_g \in [\bar{\theta}_g - \Delta\theta, \bar{\theta}_g + \Delta\theta]$, with $\bar{\theta}_g$ being the mean of $\theta_g$, distributed uniformly in the support interval, thus each user's $k^{(g)}$ of group $g$, ($1 \le k^{(g)} \le K_g$ and $1 \le g \le G$) transmitting antenna $n$ channel, $\mathbf{h}_{u,g,k,n} = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} \beta_{lgkn} \mathbf{a}(\theta_{lgkn})$, where $\mathbf{a}(\theta_{lgkn}) = [1, \exp(-j2\pi D \cos(\theta_{lgkn})), \cdots, \exp(-j2\pi D(N_u - 1) \times \cos(\theta_{lgkn}))]^T$ is the array response vector, where each $\theta_{lgkn}$ is independently selected and uniformly distributed in the group's angular support $[\bar{\theta}_g - \Delta\theta, \bar{\theta}_g + \Delta\theta]$ of its group, with $D = d/\lambda$ representing the normalized distance of successive array elements, $\lambda$ is the wavelength, $\theta_{lgkn}$ is the elevation (arrival) angle of the $l$ path of group $g$ $k$ user's $n$ receiving antenna, and the path gains $\beta_{lgkn}$ are independent complex Gaussian random variables with zero mean and variance 1. This channel model is similar to the one in [32], the Wide-Sense Stationary Uncorrelated Scattering (WSSUS) model in [31], the model employed in [33], and other works, and its gist is the standard antenna array model [10]. Note that the channel model adopted here is not Gaussian if $L$ is relatively small [32] and that it promotes group channel correlation due to clustering the group's elevation angles in the interval $[\bar{\theta}_g - \Delta\theta, \bar{\theta}_g + \Delta\theta]$. The VCM representation, presented in [34], is formed by projecting the original channel $\mathbf{H}_u$ to the

[15] The case of frequency selective fading can be treated as in [7] without changing the results presented herein for ZF-PGP.

$N_u$ dimensional space formed by the $N_u \times N_u$ DFT matrix $\mathbf{F}_{N_u}$. For massive MIMO systems, i.e., when $N_u \gg 1$, by projecting each group channel $\mathbf{H}_g$ on the DFT virtual channel space [34], we get $\tilde{\mathbf{H}}_{g,v} = \mathbf{F}_{N_u}^H \mathbf{H}_g$, where $\mathbf{F}_{N_u}$ is the DFT matrix of order $N_u$. Since each group attains the same angular behavior, over all users and antennas in the group, only a few, consecutive elements of $\tilde{\mathbf{H}}_g$ are significant [7]. This comes as a result of the fact that significant angular components need to be in the main lobe of the response vector (see [7] for details). By employing a size $|\mathcal{S}_g| \times N_u$ selection matrix $\mathbf{S}_g$[16]

$$\mathbf{H}_{g,v} = \mathbf{S}_g^T \tilde{\mathbf{H}}_g = \mathbf{S}_g^T \mathbf{F}_{N_u}^H \mathbf{H}_g, \tag{23}$$

where the group $g$ virtual channel matrix $\mathbf{H}_{g,v}$ is a reduced size, $r_g \times N_{d,g}$, matrix, with $r_g = |\mathcal{S}_g|$ being the number of significant angular components in group $g$, due to the sparsity available in the angular domain. We can then write for the uplink group $g$ channel matrix $\mathbf{H}_g$,

$$\mathbf{H}_g = \mathbf{F}_{N_u} \mathbf{S}_g \mathbf{S}_g^T \mathbf{F}_{N_u}^H \mathbf{H}_g = \mathbf{F}_{N_u, \mathcal{S}_g} \mathbf{H}_{g,v}, \tag{24}$$

where $\mathbf{F}_{N_u,\mathcal{S}_g}$ represents the selected columns of $\mathbf{F}_{N_u}$ due to its sparse representation in the angular domain. Finally, due to non-overlapping of the support sets, i.e., $\mathcal{S}_n \cap_{m \neq n} \mathcal{S}_m = \emptyset$, we see that the system becomes approximately orthogonal inter-group wise, i.e., $\sum_{m \neq g} \mathbf{H}_{d,g} \mathbf{H}_{d,m}^H \approx 0$. Then,

$$\mathbf{y}_d = \begin{bmatrix} \mathbf{H}_{1,v}^H \mathbf{F}_{N_u,\mathcal{S}_1}^H \\ \mathbf{H}_{2,v}^H \mathbf{F}_{N_u,\mathcal{S}_2}^H \\ \vdots \\ \mathbf{H}_{G,v}^H \mathbf{F}_{N_u,\mathcal{S}_G}^H \end{bmatrix} \begin{bmatrix} \mathbf{F}_{N_u,\mathcal{S}_1} & \mathbf{F}_{N_u,\mathcal{S}_2} & \cdots & \mathbf{F}_{N_u,\mathcal{S}_G} \end{bmatrix} \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_3 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{G-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}_G \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_G \end{bmatrix} + \mathbf{n}, \tag{25}$$

where for $1 \leq g \leq G$, $\mathbf{H}_{g,v}^H$ is a size $N_{d,g} \times |\mathcal{S}_g|$ matrix, $\mathbf{F}_{N_u,\mathcal{S}_G}$ is a size $|\mathcal{S}_g| \times N_u$ matrix, $\mathbf{P}_g$ is a size $|\mathcal{S}_g| \times |\mathcal{S}_g|$ matrix, and $\mathbf{x}_g$ is the group $g$ downlink symbol vector of size $|\mathcal{S}_g| \times 1$. The corresponding per-group $g$ ($1 \leq g \leq G$) receiving vector equation becomes

$$\mathbf{y}_{d,g} = \mathbf{H}_{g,v}^H \mathbf{P}_g \mathbf{x}_{d,g} + \mathbf{n}_{d,g}, \tag{26}$$

where, after normalization, the noise variance per component is equal to $\tilde{\sigma}_d^2 = \frac{1}{\mathrm{SNR}_u}$, as discussed in Section II.

## Appendix B
## Received Group Data Equation With Large-Scale Channel Effects

When large-scale channel effects are considered, one needs to invoke a common path loss factor $a_{p,g}$, and a common shadowing factor $a_{s,g}$ to all the UEs in a group. This results in a modified virtual channel representation for group $g$ ($1 \leq g \leq G$) given by $\tilde{\mathbf{H}}_{g,v}^H = \sqrt{a_{p,g} a_{s,g}} \mathbf{H}_{g,v}^H$. In other words, the received vector equation becomes

$$\mathbf{y}_{d,g} = \sqrt{a_{p,g} a_{s,g}} \mathbf{H}_{g,v}^H \mathbf{P}_g \mathbf{x}_{d,g} + \mathbf{n}_{d,g}, \tag{27}$$

from which after normalizing by $\sqrt{a_{p,g} a_{s,g}}$, we get

$$\tilde{\mathbf{y}}_{d,g} = \mathbf{H}_{g,v}^H \mathbf{P}_g \mathbf{x}_{d,g} + \mathbf{n}_{d,g}, \tag{28}$$

where the noise, due to the normalization, has a variance per component equal to $\frac{1}{\mathrm{SNR}_{s,d,g}}$, with $\mathrm{SNR}_{s,d,g} = \frac{a_{p,g} a_{s,g} E_{s,u}}{N_0} = a_{p,g} a_{s,g} \mathrm{SNR}_{s,u}$ being the average received SNR of the group.

## Appendix C
## Expressions for $w_m$

From Section III the expression for $\mathbf{M}_{m,m}$ with $m = 1, 2, \cdots, N_{d,g}$ is

$$\mathbf{M}_{m,m} = (\mathbf{V}_{g,v})_{m,\cdot} \tilde{\mathbf{\Sigma}}_{g,v}^{-2} (\mathbf{V}_{g,v})_{m,\cdot}^H = \sum_{m'=1}^{N_{d,g}} \frac{|(\mathbf{V}_{g,v})_{m,m'}|^2}{s_{g,v,m'}^2}, \tag{29}$$

where $s_{g,v,m}$ ($1 \leq m \leq N_{d,g}$) is the $m$th diagonal element of $\tilde{\mathbf{\Sigma}}_{g,v}$. Upon taking the square root of (29), the desired expression is obtained.

## Appendix D
## Gauss-Hermite Approximation to $I_{msm}$

In expression (20), we can apply a similar Gauss-Hermite approximation to the one applied in [7]. Using this, we can approximate (20) as

$$I(\mathbf{x}; \mathbf{y}) \approx N_t \log_2(M) - \frac{N_r}{\log(2)} - \frac{1}{M^{N_t}} \sum_{k=1}^{M^{N_t}} \hat{f}_k, \tag{30}$$

where

$$\hat{f}_k = \left(\frac{1}{\pi}\right)^{N_r} \sum_{k_{r1}=1}^{L} \sum_{k_{i1}=1}^{L} \cdots \sum_{k_{rN_r}=1}^{L} \sum_{k_{iN_r}=1}^{L} c(k_{r1}) c(k_{i1}) \cdots c(k_{rN_r}) c(k_{iN_r}) g_k(\sigma n_{k_{r1}}, \sigma n_{k_{i1}}, \cdots, \sigma n_{k_{rN_r}}, \sigma n_{k_{iN_r}}), \tag{31}$$

with

$$g_k(\sigma v_{k_{r1}}, \sigma v_{k_{i1}}, \cdots, \sigma v_{k_{rN_r}}, \sigma v_{k_{iN_r}}) \tag{32}$$

being the value of the function

$$\log_2 \left( \sum_m \exp\left(-\frac{1}{\sigma^2} ||\mathbf{n} - \hat{\mathbf{H}}_{g,v}^H \hat{\mathbf{P}} \mathbf{x}_k + \mathbf{H}_{g,v}^H \mathbf{P} \mathbf{x}_m||^2\right) \right) \tag{33}$$

evaluated at $\mathbf{n}_e = \sigma \mathbf{v}(\{k_{rv},\ k_{iv}\}_{v=1}^{N_r})$, where for the model in this paper, $N_t = N_r = N_{d,g}$.

[16] A selection matrix $\mathbf{S}^T$ 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, with the specific $\mathbf{e}_i$ vectors employed defined by the desired selection. Such a matrix has the property that $\mathbf{S}^T \mathbf{S} = \mathbf{I}$.

## APPENDIX E
## GENERIC PGP-WG PRECODER DETERMINATION

The PGP-WG precoder can be determined as follows [22]. In group $g$ there are $|\mathcal{S}_g|$ VCMBs. We assume an even number of VCMBs without loss of generality. For ZF precoding, we need to assume that $N_{d,g} = 2K_g \leq |\mathcal{S}_g|$ (in the opposite case we need to apply CFSDM to accommodate additional downlink users). PGP-WG develops an $N_{d,g} \times N_{d,g}$ near-optimal precoder as follows. It determines the downlink precoder by first employing SVD $\mathbf{P} = \mathbf{U}_{PGP}\mathbf{\Sigma}_{PGP}\mathbf{V}_{PGP}^H$. The left singular vector matrix, $\mathbf{U}_{PGP}$, is determined from the $N_{d,g}$ first (largest) right eigenvectors of the group's downlink channel right singular vectors [20], i.e.,

$$\mathbf{U}_{PGP} = \mathbf{U}_{g,v,N_{d,g}}, \tag{34}$$

where the subscript $N_{d,g}$ emphasizes the selection of the singular vectors corresponding to the largest (non-zero) singular vectors. In PGP-WG, the precoder right singular vector matrix has a block diagonal structure, where each block has size $2 \times 2$. Thus, the input and output antennas are partitioned in $K_g$ subgroups, each subgroup comprising two output antennas and two input symbols. Since different users need to receive different data symbols, we assign same user antennas to one user subgroup. The input symbols are grouped based on the non-zero group's channel singular values, with the optimal way being forming the singular value groups in the PGP by combining the most distant (in value) singular values, then the best performance is achieved by the PGP [22]. Finally, in PGP-WG,

$$\begin{aligned}\mathbf{V}_{PGP} &= \mathrm{diag}[\mathbf{V}_1, \cdots, \mathbf{V}_{N_{d,g}}] \\ &= \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{V}_{N_{d,g}} \end{bmatrix}\end{aligned} \tag{35}$$

is a unitary matrix, and the diagonal matrix

$$\mathbf{\Sigma}_{PGP} = \mathrm{diag}[\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_{N_{d,g}}] \tag{36}$$

satisfies $\sum_{i=1}^{N_{d,g}} \mathrm{tr}(\mathbf{\Sigma}_i^2) = N_{d,g}$, due to SVD. Then, the PGP-WG approach performs a sequence of $K_g$, size $2 \times 2$ globally optimal precoding determinations, i.e., it solves the following $N_{d,g}$ optimization sub-problems, one for each partition $i$ subgroup ($i = 1, 2, \cdots, N_{d,g}$),

$$\begin{aligned} \underset{\mathbf{P}_i}{\text{maximize}} \quad & I(\mathbf{x}_{s_i}; \mathbf{y}_{s_i}) \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{\Sigma}_i^2) = 2, \end{aligned} \tag{37}$$

with $\mathbf{P}_i = \mathbf{\Sigma}_i \mathbf{V}_{P_i}$. The power constraint has been modified from its original form, because $\mathrm{tr}(\mathbf{P}_i \mathbf{P}_i^H) = \mathrm{tr}(\mathbf{\Sigma}_i^2)$. Thus, PGP-WG needs to run $N_{d,g}$ globally optimal precoders of size $2 \times 2$, which was done very efficiently for QAM type modulations in the past [8].

## APPENDIX F
## VAAC RATIONALE

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

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{38}$$

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

$$\mathbf{y} = \mathbf{U}_H \mathbf{\Sigma}_H \mathbf{V}_H^H \mathbf{x} + \mathbf{n}, \tag{39}$$

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

$$\begin{aligned}\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_a \end{bmatrix} &= \begin{bmatrix} \mathbf{U}_H & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_H & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_H^H & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N \end{bmatrix} \\ &\quad \times \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_a \end{bmatrix} + \begin{bmatrix} \mathbf{n} \\ \mathbf{n}_a \end{bmatrix},\end{aligned} \tag{40}$$

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 (40) 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 (40) with PGP-WG. The PGP-WG algorithm will optimize and assign an amplitude diagonal matrix as per PGP-WG $\mathbf{\Sigma}_{P_i} = \mathrm{diag}[\sqrt{2},\ 0]$, $i = 1, 2, \cdots, K_g$ for each sub-group of the PGP-WG, i.e., no power sent to the noise-only antenna. On the other hand, PGP-WG will also determine the optimal unitary precoder matrix to each sub-group in the PGP-WG, thus it will be multiplexing optimally two QAM symbols to each actual transmitting antenna of the original MIMO system in (39).

## REFERENCES


[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[3] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

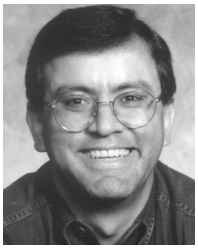
[4] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[5] 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, no. 10, pp. 6441–6463, Oct. 2013.

[6] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct. 2014.

[7] T. Ketseoglou and E. Ayanoglu, "Downlink precoding for massive MIMO systems exploiting virtual channel model sparsity," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 1925–1939, May 2018.
[8] T. Ketseoglou and E. Ayanoglu, "Linear precoding gain for large MIMO configurations with QAM and reduced complexity," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4196–4208, Oct. 2016.
[9] M. A. Girnyk, M. Vehkapera, and L. K. Rasmussen, "Large-system analysis of correlated MIMO multiple access channels with arbitrary signaling in the presence of interference," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2060–2073, Apr. 2014.
[10] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 502–513, Mar. 2000.
[11] W. Weichselberger, M. Herdin, H. Ozcelik, and E. Bonek, "A stochastic MIMO channel model with joint correlation of both link ends," *IEEE Trans. Wireless Commun.*, vol. 5, no. 1, pp. 90–100, Jan. 2006.
[12] 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 Trans. Wireless Commun.*, vol. 14, no. 2, pp. 983–997, Feb. 2015.
[13] Y. Wu, D. W. K. Ng, C. W. Wen, R. Schober, and A. Lozano, "Low-complexity MIMO precoding for finite-alphabet signals," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4571–4584, Jul. 2017.
[14] S. Zarei, W. Gerstacker, and R. Schober, "Low-complexity hybrid linear/tomlinson-harashima precoding for downlink large-scale MU-MIMO systems," in *Proc. IEEE Globecom Workshops*, Dec. 2016, pp. 1–7.
[15] 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 Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7309–7324, Nov. 2018.
[16] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.
[17] 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 Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7523–7537, Nov. 2017.
[18] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "Secrecy rate beamforming for multicell networks with information and energy harvesting," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 677–689, Feb. 2017.
[19] 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," *Proc. IEEE*, vol. 106, no. 10, pp. 1779–1833, Oct. 2018.
[20] C. Xiao, Y. R. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3301–3314, Jul. 2011.
[21] A. Molisch, *Wireless Communications*. New York, NY, USA: Wiley, 2011.
[22] T. Ketseoglou and E. Ayanoglu, "Linear precoding for MIMO with LDPC coding and reduced complexity," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2192–2204, Apr. 2015.
[23] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 303–318, Jan. 2017.
[24] S. Haghighatshoar and G. Caire, "Low-complexity massive MIMO subspace estimation and tracking from low-dimensional projections," *IEEE Trans. Signal Process.*, vol. COMM-66, no. 7, pp. 1832–1844, Apr. 2018.
[25] K.-H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. 24, no. 12, pp. 1283–1290, Dec. 1976.
[26] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cabridge Univ. Press, 2005.
[27] C. B. Peel, B. M. Hockwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.
[28] A. Wiesel, Y. C. Eldar, and S. Shamai (Shitz), "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.
[29] M. N. Boroujerdi, S. Haghighatshoar, and G. Caire, "Low-complexity statistically robust precoder/detector computation for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6516–6530, Oct. 2018.
[30] T. R. M. Fischer, "Some remarks on the role of inaccuracy in Shannon's theory of information transmission," in *Proc. 8th Trans. Prague Conf.*, 1978, pp. 211–226.
[31] R. B. Ertel *et al.*, "Overview of spatial channel models for antenna array communication systems," *IEEE Pers. Commun.*, vol. 5, no. 1, pp. 10–22, Feb. 1998.
[32] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, Mar. 2013.
[33] H. Xie, F. Gao, S. Zhang, and S. Jin, "Spatial-temporal BEM and channel estimation strategy for massive MIMO time-varying systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
[34] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

**Thomas Ketseoglou** (S'85–M'91–SM'96) received the B.S. degree from the University of Patras, Patras, Greece, in 1982, the M.S. degree from the University of Maryland, College Park, MD, USA, in 1986, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, USA, in 1990, all in electrical engineering. He was in the wireless communications industry, including senior level positions with Siemens, Ericsson, Rockwell, and Omnipoint. From 1996 to 1998, he has participated in TIA TR45.5 (now 3GPP2) 3G standardization, making significant contributions to the cdma2000 standard. He has been an inventor and a co-inventor of several essential patents in wireless communications. Since 2003, he has been with the Electrical and Computer Engineering Department, California State Polytechnic University, Pomona, CA, USA, where he is currently a Professor. In 2011, he spent his sabbatical leave at the Digital Technology Center, University of Minnesota, Minneapolis, MN, USA, where he taught digital communications and performed research on network data and machine learning techniques. He is also a part-time Lecturer with the University of California at Irvine. His teaching and research interests are in wireless communications, signal processing, and machine learning, with current emphasis on multiple-input multiple-output (MIMO), optimization, localization, and link prediction.



**Ender Ayanoglu** (S'82–M'85–SM'90–F'98) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1980, and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, USA, in 1982 and 1986, respectively, all in electrical engineering. He was with the Communications Systems Research Laboratory, part of AT&T Bell Laboratories, Holmdel, NJ, USA, until 1996, and Bell Labs, Lucent Technologies, until 1999. From 1999 to 2002, he was a Systems Architect with Cisco Systems, Inc. Since 2002, he has been a Professor with the Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA, USA, where he served as the Director of the Center for Pervasive Communications and Computing and held the Conexant-Broadcom Endowed Chair from 2002 to 2010. He was a recipient of the IEEE Communications Society Stephen O. Rice Prize Paper Award in 1995 and the IEEE Communications Society Best Tutorial Paper Award in 1997. From 1990 to 2002, he served on the Executive Committee of the IEEE Communications Society Communication Theory Committee, where he was the Chair from 1999 to 2001. From 1993 to 2014, he was an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, and served as the Editor-in-Chief from 2004 to 2008. He is currently a Senior Editor of IEEE TRANSACTIONS ON COMMUNICATIONS.