IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 3, 2020, accepted March 12, 2020, date of publication March 16, 2020, date of current version March 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981280

User Grouping for the Uplink of Multiuser Hybrid mmWave MIMO

DARIAN PÉREZ-ADÁN, ÓSCAR FRESNEDO, (Member, IEEE), JOSÉ P. GONZÁLEZ-COMA, (Member, IEEE), AND LUIS CASTEDO, (Senior Member, IEEE)

Department of Computer Engineering, University of A Coruña, 15001 A Coruña, Spain
CITIC Research Center, University of A Coruña, 15001 A Coruña, Spain

Corresponding author: Darian Pérez-Adán (d.adan@udc.es)

This work has been funded by the Xunta de Galicia (ED431G2019/01), the Agencia Estatal de Investigación of Spain (TEC2016-75067-C4-1-R, RED2018-102668-T, PID2019-104958RB-C42) and ERDF funds of the EU (AEI/FEDER, UE), and the predoctoral grant BES-2017-081955.

ABSTRACT Hybrid analog/digital schemes for precoding/combining have proved to be a low-complexity and/or low-power strategy to obtain reasonable beamforming gains in multiuser millimeter-wave (mmWave) multiple-input multiple-output (MIMO) systems. Hybrid precoding/combining performs jointly baseband processing and analog processing in the radio frequency (RF) domain. In these systems, the number of RF chains limits the maximum number of streams simultaneously handled by the transceivers. In the uplink of a multiuser mmWave MIMO system, the hardware reduction based on hybrid transceivers is limited by the number of data streams that must be simultaneously served by the centralized node. Most works approach hybrid transceiver design by considering more RF chains than data streams, an unrealistic assumption when the number of nodes is large. On the other hand, statistically independent information is conventionally assumed in multiuser mmWave systems. This assumption does not hold in scenarios like wireless sensor networks (WSNs), where the sources produce correlated information. In this paper, by enabling inter-user correlation exploitation, we propose a grouping approach to handle a high number of individual sources with a limited number of RF chains through distributed quantizer linear coding (DQLC) mappings. The allocation of the users per group and the hybrid design of the combiner at the common central node to serve the grouped users is also analyzed. We also propose a hybrid minimum mean square error (MMSE) combining design in order to exploit the spatial correlation between the sources in a conventional uncoded mmWave uplink. Simulation results show the performance advantages of the proposed approaches in various hardware-constrained system settings.

INDEX TERMS Millimeter-wave communications, multiuser channels, joint source-channel coding, hybrid combining, source correlation.

I. INTRODUCTION

Millimeter-wave (mmWave) MIMO systems are being considered for future wireless communications systems [1], [2]. The unfavorable free-space omnidirectional path loss in mmWave band due to the small wavelengths can be compensated with large gains obtained with antenna arrays having a huge number of elements, i.e., with massive MIMO technologies [3]. A major issue in massive MIMO is that dedicating one RF chain per antenna leads to inefficient and unaffordable solutions in terms of RF cost and power consumption [1], [4]. This problem is commonly handled by decoupling the fully digital precoder/combiner into a baseband part and an analog processing RF part utilized specifically to change the antenna signals phase via variable phase shifters [5]. This strategy is extensively considered in the state-of-the-art under the name of hybrid analog-digital architecture for mmWave and allows to significantly reduce the number of RF chains [4], [6]–[10].

According to their connectivity level, hybrid mmWave MIMO architectures can be distinguished in two classes [11]: fully-connected structure (FCS) [12]–[14] where each antenna is connected to each deployed RF chain,

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

and partially-connected structure (PCS) [15]–[18] which consists on connecting each RF chain only to some antenna groups in order to reduce even more the power consumption to the detriment of the beamforming gain. However, precoding/combining strategies for both schemes assume at least the same number of RF chains as individual data streams simultaneously handled, i.e., $N_{\mathrm{RF}} \geq K_{\mathrm{s}}$. Specifically, FCS leads to the same performance as that of totally digital beamforming under the condition $N_{\mathrm{RF}} \geq 2K_{\mathrm{s}}$ [19], where N_{RF} is the number of RF chains of the transceiver and K_{s} is the number of data streams to be handled.

Multiuser hybrid precoding/combining has been investigated in [12], [13], [20], [21]. Although these approaches were developed for the downlink, they can be implemented in the uplink by invoking the mean square error (MSE) duality to transform the combiners in the downlink into the precoders in the uplink [22]. In [12], an iterative projected gradient (PG) algorithm has been developed to approach the hybrid precoding for multiuser wideband channels. In [21], a first step is designed to maximize the desired signal power of each user by using the analog precoder, while the inter-user interference is then cancelled in a second step following a zero-forcing (ZF) strategy to design the baseband precoder. The algorithms proposed in [13] and [20] have been developed for multi-stream transmissions in narrowband and wideband scenarios, respectively. These algorithms are based on a greedy approach where the user data streams are allocated iteratively to the available RF chains by evaluating the impact of allocating the new stream on the overall system performance according to a specific metric (e.g., sum-rate). In these algorithms, the first step is developed to maximize the desired signal power per user and partially cancel the inter-user interference. Then, the residual interference is cancelled in a second step by adopting a ZF strategy for the baseband precoder. Nevertheless, the hybrid factorization in these works is performed under the assumption $K_{\mathrm{s}} = N^{\mathrm{r}}_{\mathrm{RF}}$.

Nowadays, applications like WSNs or the Internet of Things (IoT) demand the capability of serving simultaneously a huge amount of nodes. Considering that the number of RF chains must usually be at least equal to the number of transmitted streams, this situation would lead to the need of a tremendous amount of RF chains at the common receive node, $N^{\mathrm{r}}_{\mathrm{RF}}$. These requirements become even more severe when working with critical data which need to be sent with a minimum latency. For this reason, in this work, we explore a new strategy for the design of practical mmWave massive MIMO systems which are able to deal with these requirements in terms of hardware complexity and delay. Specifically, we propose a novel approach which enables the transmission of a number of streams significantly larger than the number of RF chains available at the receiver with minimum delay. This approach is based on the idea of using some appropriate non-orthogonal multiple access (NOMA) technique to ensure that the streams corresponding to several users are effectively superimposed during the transmission so that the information can then be decoded with an acceptable level of distortion by means of a single RF chain at reception.

Since we are mainly interested in applications having minimum latency, an appealing candidate for the encoding operation is the use of DQLC [23], [24]. This mapping function was proposed for the distributed encoding of Gaussian symbols in a multiple access channel (MAC) scenario where all the users transmit their encoded symbols simultaneously. Another suitable properties of this mapping function are that the encoding and decoding of the user information can be performed with negligible delay and low computational cost, as well as it is able to exploit the source correlation between different users (spatial correlation). The latter feature is also important since there exists a large number of scenarios where this premise occurs. Therefore, the proposed DQLC-based scheme provides a suitable solution to reduce the number of required RF chains ($N^{\mathrm{r}}_{\mathrm{RF}}$) in the context of mmWave massive MIMO applications with a large number of potential transmitters and delay constraints.

The design of the DQLC-based scheme poses interesting challenges in the different parts of the communication link. First, it is required to define some coherent strategy to gather the users to be served by the same RF chain at reception. In addition, digital combiners should be designed to cancel inter-group interferences since the DQLC decoding is quite sensitive to this type of interferences. Finally, we focus on the common base station (BS) combiner design and derive a hybrid solution to exploit the intra-group correlation via a MMSE combiner instead of the conventional ZF strategy applied in [13], [20], [21].

A. CONTRIBUTIONS

In this paper, we propose a low-complexity system which enables the correlation exploitation and allows the reduction of the number of RF chains for multiuser mmWave massive MIMO communications by considering a FCS-based hybrid combiner at the common BS. For completeness, we also address the design of practical solutions to exploit the source correlation in a less restrictive scenario where the number of RF chains is allowed to be equal to the number of available streams to be transmitted. In such a scenario, a MMSE-based hybrid design of the combiner inspired by [13] and [21] is proposed to exploit the inter-user correlation for the uplink of conventional ungrouped mmWave MIMO systems.

The main contributions of this work are summarized as follows:

- Proposing a hybrid MMSE combiner to exploit the spatial correlation in the uplink of ungrouped correlated sources in mmWave.
- Proposing a user grouping approach to reduce hardware complexity through the use of DQLC mapping to superimpose user symbols in hybrid multiuser mmWave MIMO systems.
- Proposing a scheduling algorithm to define the grouped served users and allocate the users per group.

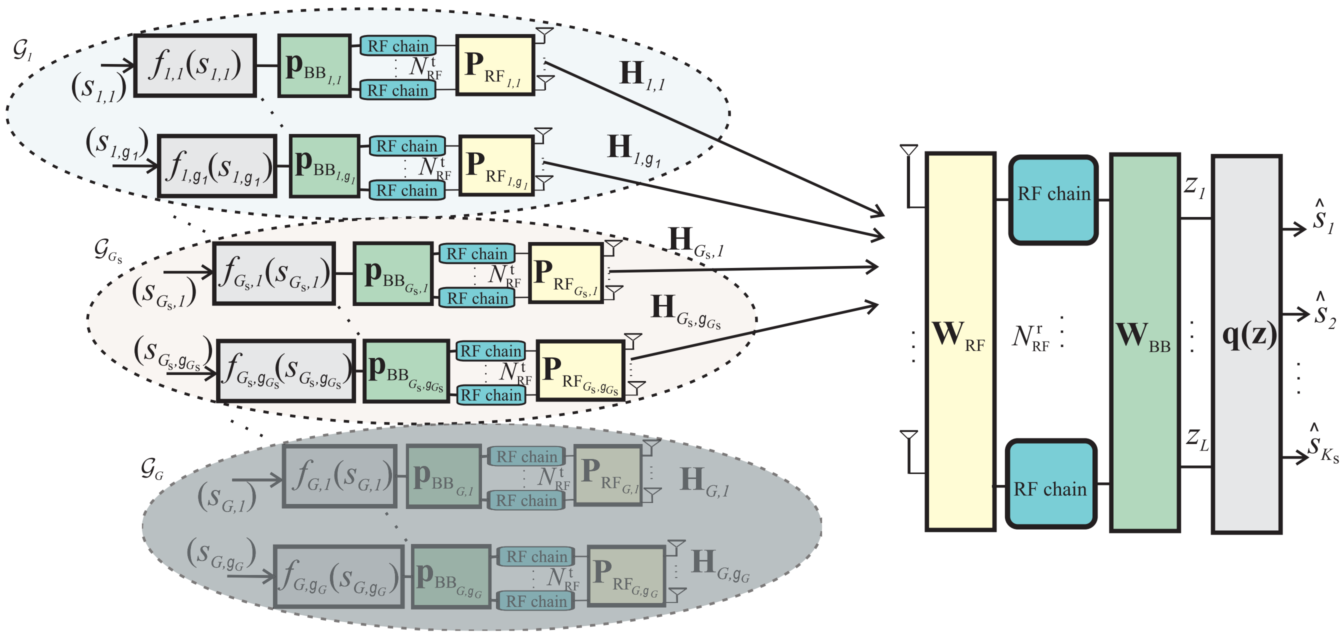

FIGURE 1. Block diagram of the multiuser mmWave MIMO system with G groups that contain g_i users each.

- Proposing a novel approach to face the hybrid combiner design for user grouping in mmWave MIMO systems.

B. ORGANIZATION

The rest of the paper is structured as follows. The system model is detailed in Section II. The grouping and the allocation algorithms, as well as the scheduling policy to serve the users, are analyzed in Section III. A novel hybrid combining approach to serve the grouped users is proposed in Section IV. Precoding and combining design for conventional systems with ungrouped correlated sources are discussed in Section V. The computational complexity of the proposed algorithms is analyzed in Section VI. Simulation results are presented in Section VII, and Section VIII is devoted to the conclusions.

C. NOTATION

The following notation is employed in this paper: a is a scalar, $\mathbf{a}$ is a vector, and $\mathbf{A}$ is a matrix. $[\mathbf{A}]_{i,j}$ is the entry on the i-th row and the j-th column of $\mathbf{A}$, whereas $[\mathbf{A}]_{i,:}$ represents the i-th row of $\mathbf{A}$. Transpose, conjugate transpose, and pseudoinverse of $\mathbf{A}$ are represented by $\mathbf{A}^T$, $\mathbf{A}^*$, and $\mathbf{A}^\dagger$, respectively, whereas $\|\mathbf{A}\|_F$ denotes the Frobenius norm of $\mathbf{A}$. span $(\cdot)$ represents the subspace spanned by the columns of the input matrix. blkdiag $(\cdot)$ is the operator that constructs a block diagonal matrix from input matrices. Calligraphic letters are employed to denote sets and sequences, $|\mathcal{A}|$ represents the cardinality of $\mathcal{A}$ and $\mathcal{A} \setminus b$ represents the exclusion of the element b from $\mathcal{A}$. Finally, $\lfloor \cdot \rceil$ stands for the rounding operation and expectation is denoted by $\mathbb{E}[\cdot]$.

II. SYSTEM MODEL

Figure 1 shows the uplink of a multiuser mmWave system where K users send data to a common BS. The set $\mathcal{K} = \{1, \ldots, K\}$ contains the subset of active users, $\mathcal{K}_s$, and the subset of idle users, $\mathcal{K}_I$, such that $\mathcal{K} = \mathcal{K}_s \cup \mathcal{K}_I$. We assume that these users are divided into G groups by using an appropriate scheduling algorithm and according to a given performance criterion. Accordingly, $\mathcal{G}_i$, $\forall\ i = 1, \ldots, G$, represents the i-th group of users. As observed in Figure 1, $G_s \leq G$ groups are simultaneously served, i.e., $\mathcal{K}_s = \cup_{i=1}^{G_s} \mathcal{G}_i$. We assume that $G_I = G - G_s$ groups of users are idle, leading to $\mathcal{K}_I = \cup_{i=1}^{G_I} \mathcal{G}_i$. We denote the number of active and inactive users by $K_s = |\mathcal{K}_s|$ and $K_I = |\mathcal{K}_I|$, respectively. Finally, the vector which contains the number of users per group is denoted by $\mathbf{g} = [g_1, \ldots, g_{G_s}, \ldots, g_G]$, where $g_i = |\mathcal{G}_i|$ is the number of users allocated in the i-th group. Note that the following equality always holds $\sum_{i=1}^{G_s} g_i = K_s$.

We assume that each user sends a single stream of discrete-time continuous-amplitude symbols to a common receiver with N_r antennas. We also assume that each user is equipped with N_t transmit antennas. The source symbols of the K users are represented by the vector $\mathbf{s} = [s_1, s_2, \ldots, s_K]^T$ which is assumed to follow a zero-mean spatially correlated multivariate complex-valued Gaussian distribution with covariance matrix $\mathbf{C}_s = \mathbb{E}[\mathbf{s}\mathbf{s}^*]$, such that $[\mathbf{C}_s]_{k,k} = 1, \ \forall\ k$, and $[\mathbf{C}_s]_{i,j} = \rho_{i,j}$, $0 \leq \rho_{i,j} \leq 1$, $\forall\ i,j$ with $i \neq j$.

At each channel use, each active user sends one complex-valued encoded symbol as $f_{i,j}(s_{i,j})$, $\forall\, i = 1, \ldots, G_s$, $\forall j = 1, \ldots, g_i$, where $f_{i,j}(\cdot)$ represents the mapping function that encodes $s_{i,j}$, the source symbol of the j-th user in the i-th group. Note that sub-index i is employed to index the considered group, whereas sub-index j identifies the j-th user accommodated in the i-th group, i.e., this pair of indices actually represents the user $\mathcal{G}_i(j)$.

Therefore, the vector corresponding to all encoded served user symbols per channel use is represented by

$$\mathbf{f}(\mathbf{s}) = \left[f_{1,1}(s_{1,1}), \ldots, f_{1,g_1}(s_{1,g_1}), \ldots, f_{G_\mathrm{s},g_{G_\mathrm{s}}}(s_{G_\mathrm{s},g_{G_\mathrm{s}}}) \right]^T,$$

where we also assume that the encoded symbols satisfy the condition $\mathbb{E}\left[|f_{i,j}\left(s_{i,j}\right)|^2\right] \leq 1$.

After encoding the source information, the resulting encoded symbols are precoded prior to be transmitted over the channel. Hybrid precoding is considered at the users due to the hardware constraints. The hybrid precoder of the j-th user in the i-th group is denoted as $\mathbf{p}_{\mathrm{H}i,j} = \mathbf{P}_{\mathrm{RF}i,j}\mathbf{p}_{\mathrm{BB}i,j}$, and it is implemented by using $N_{\mathrm{RF}}^{\mathrm{t}} = 2$ transmit chains. We assume this amount of RF chains per user, which is enough to lead essentially to the performance obtained by the unconstrained precoder implementation, since we focus on the single-stream scenario (see [19, Appendix A]). The baseband precoder is $\mathbf{p}_{\mathrm{BB}i,j} \in \mathbb{C}^{N_{\mathrm{RF}}^{\mathrm{t}} \times 1}$ and the analog precoder is $\mathbf{P}_{\mathrm{RF}i,j} \in \mathcal{P}_{\mathrm{RF}}$ such that $\mathcal{P}_{\mathrm{RF}} \subset \mathbb{C}^{N_\mathrm{t} \times N_{\mathrm{RF}}^{\mathrm{t}}}$ is the set of feasible RF precoder matrices with unit modulus entries. An individual power constraint is imposed at each user, such that $\|\mathbf{P}_{\mathrm{RF}i,j}\mathbf{p}_{\mathrm{BB}i,j}\|_F^2 \leq T_{i,j},\ \forall\, i,j$.

At the receiver, the K_s served data streams are collected by deploying N_r receive antennas. Therefore, the received signal reads as

$$\mathbf{y} = \sum_{i=1}^{G_\mathrm{s}} \sum_{j=1}^{g_i} \mathbf{H}_{i,j}\mathbf{P}_{\mathrm{RF}i,j}\mathbf{p}_{\mathrm{BB}i,j} f_{i,j}(s_{i,j}) + \mathbf{n}, \tag{1}$$

where $\mathbf{H}_{i,j} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ describes the mmWave channel response of the j-th user in the i-th group, and the vector $\mathbf{n} = \left[n_1, n_2, \ldots, n_{N_\mathrm{r}}\right]^T$ represents the complex-valued additive white Gaussian noise (AWGN) such that $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_\mathrm{n}^2\mathbf{I})$. The received signal in (1) can also be rewritten in a more compact way as

$$\mathbf{y} = \mathbf{H}\mathbf{P}_\mathrm{H}\mathbf{f}(\mathbf{s}) + \mathbf{n}, \tag{2}$$

by considering the stacked channel matrix $\mathbf{H} = [\mathbf{H}_{1,1}, \ldots, \mathbf{H}_{G_\mathrm{s},g_{G_\mathrm{s}}}]$, whereas the matrix containing all the hybrid precoders of the served users is described by $\mathbf{P}_\mathrm{H} = \mathrm{blkdiag}\left(\mathbf{p}_{\mathrm{H}1,1}, \ldots, \mathbf{p}_{\mathrm{H}G_\mathrm{s},g_{G_\mathrm{s}}}\right)$. For convenience, we define the equivalent channel response for the i-th group as

$$\tilde{\mathbf{H}}_i = \left[\mathbf{H}_{i,1}\mathbf{p}_{\mathrm{H}i,1}, \ldots, \mathbf{H}_{i,g_i}\mathbf{p}_{\mathrm{H}i,g_i}\right], \quad \forall\, i = 1, \ldots, G_\mathrm{s}, \tag{3}$$

such that the overall equivalent matrix $\tilde{\mathbf{H}} \in \mathbb{C}^{N_\mathrm{r} \times K_\mathrm{s}}$ can be constructed as

$$\tilde{\mathbf{H}} = \left[\tilde{\mathbf{H}}_1, \ldots, \tilde{\mathbf{H}}_{G_\mathrm{s}}\right]. \tag{4}$$

Hence, (2) can alternatively be written as

$$\mathbf{y} = \tilde{\mathbf{H}}\mathbf{f}(\mathbf{s}) + \mathbf{n}. \tag{5}$$

We assume that the receiver has $N_{\mathrm{RF}}^{\mathrm{r}} \leq K$ available RF chains. Because of this hardware limitation, a hybrid combiner $\mathbf{W}_\mathrm{H} = \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}$ is implemented at the BS to decouple the superimposed MAC signal before applying the demapping functions to estimate the source symbols. The signal after this combining processing can be described as

$$\mathbf{z} = \mathbf{W}_\mathrm{H}^*\mathbf{y} = \mathbf{W}_{\mathrm{BB}}^*\mathbf{W}_{\mathrm{RF}}^*\mathbf{y}, \tag{6}$$

where $\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}}^{\mathrm{r}} \times G_\mathrm{s}}$ represents the baseband combiner and $\mathbf{W}_{\mathrm{RF}} \in \mathcal{W}_{\mathrm{RF}}$ denotes the RF combiner with constant value entry constraints. $\mathcal{W}_{\mathrm{RF}} \subset \mathbb{C}^{N_\mathrm{r} \times N_{\mathrm{RF}}^{\mathrm{r}}}$ represents the set of feasible RF combiners with that property.

It is important to highlight that the hybrid combiner actually produces G_s channel symbols, which are then processed by the demapping functions

$$\mathbf{q}(\mathbf{z}) = [q_1(z_1), \ldots, q_{G_\mathrm{s}}(z_{G_\mathrm{s}})]^T$$

to obtain an estimation of the served user symbols $\hat{\mathbf{s}} = [\hat{s}_{1,1}, \ldots, \hat{s}_{G_\mathrm{s},g_{G_\mathrm{s}}}]^T$ by employing just $N_{\mathrm{RF}}^{\mathrm{r}} = G_\mathrm{s}$ RF chains. Thus, the demapping function $q_i(z_i) : \mathbb{C} \rightarrow \mathbb{C}^{g_i}$ will provide an estimation for the source symbols transmitted by the g_i users at the i-th group as

$$\hat{\mathbf{s}}_i = q_i(z_i) = [\hat{s}_{i,1}, \ldots, \hat{s}_{i,g_i}]^T, \quad \forall\, i = 1, \ldots, G_\mathrm{s}. \tag{7}$$

Since we are considering complex-valued continuous-amplitude source symbols, the information will recover with a certain level of distortion. In this work, the observed distortion is measured as the MSE between the source symbols and the estimated ones, i.e.,

$$\xi = \frac{1}{K_\mathrm{s}} \sum_{i=1}^{G_\mathrm{s}} \sum_{j=1}^{g_i} |s_{i,j} - \hat{s}_{i,j}|^2. \tag{8}$$

An interesting discussion arises when considering scenarios where the number of total users is significantly larger than the number of available RF chains at reception, i.e., $N_{\mathrm{RF}}^{\mathrm{r}} \ll K$. In this situation, we can opt for two opposite allocation policies. One is to exactly select the same number of active users as the number of available RF chains (i.e., $g_i = 1,\ \forall\, i$ and $G_\mathrm{s} = N_{\mathrm{RF}}^{\mathrm{r}}$). In this case, the objective will be to recover the source symbols with the minimum possible distortion although this may imply the number of inactive users to be very large. The other is to gather the users into large groups in order to serve most of the users simultaneously with the available RF chains. In this case, we neglect the impact on the communication reliability. An intermediate solution would be to find a trade-off between the number of served users and the level of distortion obtained after decoding the user symbols. In Section III, we will address the problem of designing the different components of the communication system and the scheduling algorithm to obtain a suitable balance between these two performance indices: the number of served users and signal distortion. In addition, the first scenario is considered in Section V for completeness.

A. CHANNEL MODEL

We consider the following geometric channel model suitable for mmWave propagation [1], [25], [26]

$$\mathbf{H}_{i,j} = \gamma \sum_{n=1}^{N_{\text{cl}}} \sum_{m=1}^{N_{\text{ray}}} \beta_{n,m} \mathbf{a}_{\text{BS}}(\phi_{n,m}^{\text{BS}}, \theta_{n,m}^{\text{BS}}) \mathbf{a}_{\text{t}}^{*}(\phi_{n,m}^{\text{t}}, \theta_{n,m}^{\text{t}}). \quad (9)$$

This narrowband block-fading channel model comprises N_{cl} scattering clusters, constituted by N_{ray} rays each [27] with $\gamma = \sqrt{N_{\text{t}} N_{\text{r}} / N_{\text{cl}} N_{\text{ray}}}$. The parameter $\beta_{n,m}$ represents the complex path gain of the m-th ray in the n-th cluster. $\phi^{\text{t}}(\theta^{\text{t}})$ are the azimuth (elevation) angles of departure (AoD) at each transmitter and $\phi^{\text{BS}}(\theta^{\text{BS}})$ represent the azimuth (elevation) angles of arrival (AoA) at the BS. The transmit and receive array response vectors are denoted by $\mathbf{a}_{\text{t}}(\phi^{\text{t}}, \theta^{\text{t}})$ and $\mathbf{a}_{\text{BS}}(\phi^{\text{BS}}, \theta^{\text{BS}})$. We assume a uniform square planar array (USPA) at both ends, hence, the array response vectors are defined as [28]

$$\mathbf{a}_{\text{USPA}}(\phi, \theta) = \frac{1}{\sqrt{N}} \Big[1, \ldots, e^{j\frac{2\pi}{\lambda} d(p \sin(\phi) \sin(\theta) + q \cos(\theta))}, \ldots, e^{j\frac{2\pi}{\lambda} d\left(\left(\sqrt{N}-1\right)\sin(\phi)\sin(\theta) + \left(\sqrt{N}-1\right)\cos(\theta)\right)}\Big]^{T}, \quad (10)$$

where $\sqrt{N} \times \sqrt{N}$ represents the antenna array size, λ is the wavelength and d stands for the inter antenna spacing, which is fixed to $\lambda/2$. The indices $0 \leq p < \sqrt{N}$ and $0 \leq q < \sqrt{N}$ are employed to determine the position of the antenna elements in the array.

III. USER GROUPING IN mmWave HYBRID SYSTEMS

Existing works on hybrid analog/digital MIMO transceivers have mainly focused on the case $K_{\text{s}} \leq N_{\text{RF}}^{\text{r}} < 2K_{\text{s}}$. In order to circumvent this constraint and reduce N_{RF}^{r}, we consider a design strategy where K users are gathered into G groups so that the users at each group use a DQLC mapping to encode and superimpose their source symbols. Considering this encoding strategy, the information corresponding to all the users in the same group could be recovered from a single RF chain at the receiver by designing the mapping functions and the different filters in an appropriate way. Note that this approach allows reducing N_{RF}^{r} down to the number of served user groups G_{s} ($G_{\text{s}} \leq K_{\text{s}}$). In addition, DQLC encoding presents some additional advantages such as zero delay and efficient exploitation of the spatial correlation at each user group.

A. DISTRIBUTED QUANTIZER LINEAR CODING (DQLC)

DQLC is a joint source-channel coding (JSCC) technique proposed for the non-orthogonal transmission of multivariate Gaussian sources over the MAC [23]. This mapping constitutes a specific case of vector quantizer linear coding (VQLC) by imposing a zero-delay encoding constraint [29]. The conventional implementation of DQLC establishes that $g_i - 1$ users of the i-th group transmit a quantized version of its symbol whereas the symbol of the remaining user is just scaled by a power factor so that it can be placed between two quantization steps of the encoded user symbols. Therefore, the DQLC mapping function for the i-th group is mathematically defined as

$$f_{i,j}(s_{i,j}) = \begin{cases} \alpha_{i,j} \left\lfloor \dfrac{s_{i,j}}{\Delta_{i,j}} - \dfrac{1}{2} \right\rceil + \frac{1}{2}, & j < g_i \\ \alpha_{i,j} s_{i,j}, & j = g_i \end{cases}, \quad (11)$$

where $\alpha_{i,j}$ represents a gain factor and $\Delta_{i,j}$ represents the quantization step of the quantizer employed for the j-th user in the i-th group. As mentioned, a specific DQLC mapping is individually applied into the $G_{\text{s}} \leq G$ groups containing the users served, and therefore the parameters $\alpha_{i,j}$ and $\Delta_{i,j}$ must conveniently be optimized for the g_i users at each group. A detailed description about the DQLC implementation and parameter optimization can be found in [23], [24], [30].

In [30], in-depth treatment about this optimization is provided, where only one group of users is considered. In this work, we consider several groups and therefore the potential inter-group interference has to be taken into account in the parameter design. We follow an alternative approach to [30] by considering the signal-to-interference-plus-noise ratio (SINR)—affected by the inter-group interference due to the hybrid implementation of the combiner at the BS—as the metric to optimize the DQLC parameters. In the proposed setup, the symbols at the input of the demapping operation are given by

$$\mathbf{z} = \mathbf{W}_{\text{H}}^{*} \mathbf{H} \mathbf{P}_{\text{H}} \mathbf{f}(\mathbf{s}) + \mathbf{W}_{\text{H}}^{*} \mathbf{n} = \mathbf{R} \mathbf{f}(\mathbf{s}) + \tilde{\mathbf{n}}, \quad (12)$$

where

$$\mathbf{R} = \mathbf{W}_{\text{H}}^{*} \mathbf{H} \mathbf{P}_{\text{H}} = \mathbf{W}_{\text{H}}^{*} \tilde{\mathbf{H}} \quad (13)$$

represents the equivalent channel response for the encoded symbols after filtering, and $\tilde{\mathbf{n}} = \mathbf{W}_{\text{H}}^{*} \mathbf{n}$ is the equivalent AWGN with the same statistics as the original noise. Therefore, the SINR values can directly be computed from the equivalent channel matrix $\mathbf{R}$ and from the noise variance. From the above equation, we can decompose the input signal at the demapping operation into its individual components as

$$z_i = \sum_{j=1}^{K_{\text{s}}} [\mathbf{R}]_{i,j} f_j(s_j) + \tilde{\mathbf{n}}_i, \quad \forall\, i = 1, \ldots, G_{\text{s}}, \quad (14)$$

which can be rearranged as

$$z_i = \sum_{j=1}^{g_i} [\mathbf{R}]_{i,[l(i)+j]} f_{i,j}(s_{i,j}) + \sum_{r \neq i}^{G_{\text{s}}} \sum_{j}^{g_r} [\mathbf{R}]_{i,[l(r)+j]} f_{r,j}(s_{r,j}) + \tilde{\mathbf{n}}_i, \quad (15)$$

where the auxiliary indices $l(\cdot)$ determine the first component of the equivalent channel matrix corresponding to the group given by the argument, and they can be computed in a simple way as $l(i) = \sum_{t=1}^{i-1} g_t + 1$. Note that the first term in (15) corresponds to the desired signal for the i-th group, the second

term represents the interference caused by the signal transmitted by other groups, and the third term is the i-th component of the equivalent noise after filtering.

At the receiver side, different methods have been adopted in order to perform DQLC demapping. One is sequential decoding where an estimation of the quantized symbols is first computed, and the obtained symbols are then used to estimate the scaled symbol (cf. [29, Section III]). Another is the approximated MMSE estimation with sphere decoding [24, Section III], which in general exhibits a better performance for more than two users per group. Therefore, we will use this latter decoding algorithm since it can be applied to an arbitrary number of users per group (cf. [24] or [30] for details).

B. USER GROUPING

As introduced in the previous sections, the key point to reduce the hardware requirements in the considered mmWave-based massive MIMO system is the grouping of users into disjoint groups and the application of an optimized DQLC scheme to encode the source symbols at each group. Therefore, the design of an adequate allocation policy to define the grouping configuration is essential to achieve a good trade-off between the system performance (i.e., observed distortion) and the number of served users. In this subsection, the proposed user grouping allocation by considering the correlation and the channel state information (CSI) is presented. The scheduling policy involves both the grouping operation, that determines the number of users per group, and the allocation of the specific users that will be accommodated in each group.

The design of the proposed scheduling policy is clearly influenced by some important issues related to the experimental behaviour of DQLC mappings over wireless channels. In particular, the following considerations should be taken into account:

- From [30], [31] it is clear that the DQLC mapping is able to improve the system performance by exploiting spatial correlation. Hence, the scheduling approach will be designed to guarantee the exploitation of the inter-user correlation per group. Thus, it is always preferable to allocate users with high levels of correlation in the same group, if possible.
- Another interesting issue related to DQLC performance can be observed from (15). As will be seen below, the receive combiner should be designed to cancel the interferences from other groups, in which case (15) boils down to

$$z_i = \sum_{j=1}^{g_i} [\mathbf{R}]_{i,[l(i)+j]} f_{i,j}(s_{i,j}) + \tilde{\mathbf{n}}_i. \tag{16}$$

  when the inter-group interferences are perfectly cancelled. According to the particular way of encoding the source symbols with DQLC to deal with the superimposition caused by the non-orthogonal transmission over the MAC, the distortion of the recovered symbols will be lower when the components of the equivalent channel matrix involved in the above equation point to the same direction.

  As observed in (16), each received symbol comprises two different parts: the sum of all encoded symbols in a group weighted by the equivalent channel responses, and the noise component. As shown in [23], [30], the key for an appropriate DQLC decoding is to ensure that the sum of the encoded symbols corresponding to the next users does not cause that the quantized symbol of the considered user crosses to a different quantization interval. Note that DQLC decoding could be interpreted as particular form of successive interference cancellation, where the sum of the symbols for the next users would be the remaining interference, and the decoding would only be possible if such interference did not move the considered quantized symbol to a different interval. In this way, the probability of a crossing event will be lower when the channel coefficients point to the same direction. For this reason, a reasonable scheduling strategy would be to gather in the same group those users whose channel subspaces intersect. Hence, it will be possible to find a common precoding direction providing reasonable equivalent channel gains for some of them. In this sense, a particular metric is employed to provide a joint measure of the degree of similarity among the user channels and the joint gain of such channels.
- As shown in [24], [30], the performance of DQLC-based systems degrades as the number of users increases in the MAC. This fact inevitably leads to a trade-off between the number of users per group and the system performance. The impact of adding a new user to a group can be determined from the approximation to the sum-MSE for DQLC transmissions considering a single group, which was derived in [30]. From the analysis of this error expression and from the results obtained in [30], it is possible to sense that the adding of a new user approximately leads to a duplication of the error (for high signal-to-noise ratio (SNR) values). Indeed, the specific increase on the distortion with each new user actually depends on the SNR value, but the relevant fact is that, for a given SNR, this increase is similar regardless of the number of users in the group. Thus, the performance loss of adding a new user to a group with 2 users is similar to adding it to a group with 6 users irrespective of the considered SNR value. Moreover, the gain with respect to an uncoded transmission of the source symbols becomes negligible beyond a certain number of users per group (about 7). Considering these two arguments, it is important to prevent oversized groups which can lead to a performance degradation. In this sense, the maximum number of users per group can explicitly be limited by introducing the parameter $g_{\max} \leq 7$.
- We have experimentally observed that a balanced configuration of the number of users per group

provides better performance than unbalanced ones. This behaviour can also be explained from the analysis of the approximation for the sum-MSE expression in [30], since the fact of increasing the group size necessarily implies to increase the size of the quantization intervals employed at the encoding operation, which significantly penalizes the system performance (higher distortions). Hence, considering the sum-MSE as the performance metric and the assumption that each new user implies to double the overall error, it is easy to see that the resulting distortion will be lower when the individual contributions are balanced. To illustrate this point, let us consider an example scenario where the distortion observed for each user is 0.1 when using a DQLC scheme with 2 users per group. In addition, we aim at serving 6 users with 2 groups. In this case, we could consider two grouping options: a) 3 users per group or b) 2 users in a group and 4 users in the other. For the first possibility, the overall error will be 6×0.2, whereas in the latter it would be $2 \times 0.1 + 4 \times 0.4$. The same analysis could be done under the more general assumption on the impact of the number of users per group explained in the previous point.

- Finally and related to the importance of preventing huge distortions caused by the fact that a quantized symbol is erroneously decoded (i.e., if it is detected as a point on an erroneous interval), it is important to ensure that the quantized symbols correspond to those users in the group whose channel matrices have larger singular values [30].

Taking into account all previous issues, we have developed a scheduling procedure which comprises two intertwined parts: the grouping strategy and the allocation policy. The grouping algorithm must determine the group configuration, i.e., the number of groups and the number of users per group, whereas the allocation algorithm is responsible of selecting which users are grouped into each group. The different steps of these two proposed algorithms are described in Algorithm 1 and Algorithm 2.

1) GROUPING ALGORITHM

As observed, Algorithm 1 determines the vector $\mathbf{g}$, which contains the number of users per group, given the set of all available users, $\mathcal{K}$, and the number of available RF chains at reception, $N_{\mathrm{RF}}^{\mathrm{r}}$. Basically, the aim of the algorithm is to find a balanced group configuration under the premise that the number of groups, G, should be as close as the number of available RF chains with $G \geq N_{\mathrm{RF}}^{\mathrm{r}}$. This condition is imposed to limit the number of users which cannot be served simultaneously.

In the first iteration, the size of the first group is determined considering the total number of users and the number of available RF chains (step 4). Note that this size is limited to $g_{\max}$ users. Next, g_1 users will be allocated to the current group by invoking the allocation algorithm (step 5). Note that the actual number of users allocated to the group does not necessarily have to be equal to the size determined a priori for such a group. As we will see later, this situation occurs because it was not possible to find users with enough correlation or whose channels are sufficiently similar. After this step, the sequence and the number of unallocated users are updated (steps 7 and 8), and the counter ℓ for the number of available groups is updated (steps 9 to 12). This counter is initially set to $N_{\mathrm{RF}}^{\mathrm{r}}$ and it is decreased by one at each iteration. However, it may be necessary to occasionally increase it if the available groups are not enough to gather all the remaining users due to the constraint on the maximum group size. This sequence of steps is repeated at each algorithm iteration until all the users are allocated to any group.

Algorithm 1 Grouping

Input: $\mathcal{K}, N_{\mathrm{RF}}^{\mathrm{r}}, g_{\max}$

1: **Initialize:** $i = 0$, $g_i = 0 \; \forall \, i$, $\ell = N_{\mathrm{RF}}^{\mathrm{r}}$, $K = |\mathcal{K}|$
2: **repeat**
3: $\quad i \leftarrow i + 1$
4: $\quad \hat{g}_i = \min\left(g_{\max}, \left\lceil \frac{K}{\ell} \right\rceil\right)$
5: $\quad \mathcal{G}_i \leftarrow$ **Algorithm 2** with $\hat{g}_i$ and $\mathcal{K}$
6: $\quad g_i \leftarrow |\mathcal{G}_i|$
7: $\quad \mathcal{K} \leftarrow \mathcal{K} \setminus \mathcal{G}_i$
8: $\quad K \leftarrow K - g_i$
9: $\quad$ **if** $g_i \neq \hat{g}_i$ **then**
10: $\quad\quad \ell \leftarrow \lceil K / g_{\max} \rceil$
11: $\quad$ **else**
12: $\quad\quad \ell \leftarrow \ell - 1$
13: **until** $K = 0$

Output: $\mathbf{g} = [g_1, \ldots, g_{G\mathrm{s}}, \ldots, g_G]$, $\{\mathcal{G}_i\}_{i=1}^{G}$

2) ALLOCATION ALGORITHM

Algorithm 2 allocates the users to their corresponding groups. As observed, this algorithm provides the sequence of users $\mathcal{G}_i$ for the i-th group given the set of users which have not been allocated yet, the source correlation matrix, the channel responses, the desirable size for the group and a set of design parameters.

As mentioned in the beginning of this section, two different factors decisively impact into the performance of the DQLC-based systems: the level of correlation between the source symbols corresponding to the users which are gathered into the same group, and the similarity of their channel matrices (i.e., the ability for aligning their channels). The impact of a high correlation level is in general more positive than that of the similarity, although it depends on the overall correlation in the system. For example, the impact of the first factor is obviously lacking for scenarios with uncorrelated sources. In order to model this behaviour, we introduce the factors δ_ρ and δ_{s} which define the weight of the correlation criterion and of the similarity one, respectively, in the proposed metric for selecting users for a same group. In addition, a correlation threshold γ_ρ and a similarity threshold γ_{s} are introduced to discard some users that, even when the constraint of the

maximum users per group is fulfilled, could severely degrade the system performance when including them in the group.

In general, the optimization of these design parameters would lead to an exhaustive search in a four-dimension parameter space or to highly non-convex optimization problems involving non-linearities and discontinuities due to the DQLC mapping function. For this reason, we have considered a heuristic approach – based on the insight provided by the issues related to the DQLC behaviour – to lower the computational cost for the selection of the parameters. Note also that the thresholds γ_ρ and γ_s drive the general behaviour of the scheduling procedure: either incorporating as many users as possible by choosing low values for these thresholds, or introducing a smaller number of users but guaranteeing a high level of the received signal quality. Thus, the tuning of these parameters is conducted depending on the overall source correlation and the system requirements.

Algorithm 2 summarizes the steps of the proposed allocation strategy. As observed, this algorithm firstly includes the user with the best channel by evaluating the largest singular value of the user channels as follows

$$\mathcal{G}(1) = k \;\text{ s.t. }\; \|\mathbf{H}_{\mathcal{K}(k)}\|_2 \geq \|\mathbf{H}_{\mathcal{K}(j)}\|_2, \quad \forall\, j \neq k. \quad (17)$$

As observed, the sequence $\mathcal{G}$ will include the users which are already allocated to the considered group at each iteration. It is worth remarking that we introduce a change in the notation for the allocation algorithm to highlight that the group configuration is not definitively established. Therefore, instead of using index i for the groups and j for the users in the group, we will use $\mathcal{K}(k)$ and $\mathcal{G}(k)$ to refer to the k-th user in the set of unallocated users and the k-th user already allocated to the i-th group, respectively.

Then, the algorithm iteratively includes the next users by considering the metric

$$m_k = \delta_\rho \tilde{\rho}_k + \delta_\mathrm{s} C_{\mathrm{sim}k}, \quad (18)$$

where $\tilde{\rho}_k$ is the mean of the cross-correlation of the users in $\mathcal{G}$ and the candidate user $\mathcal{K}(k)$, whereas the parameter $C_{\mathrm{sim}k}$ measures the convenience of including user $\mathcal{K}(k)$ in $\mathcal{G}$ according to its channel response. As mentioned, this parameter should represent the joint gain of such channels. Following this premise, we define the composite channel for the L users already in the sequence $\mathcal{G}$ as $\mathbf{H}_\mathrm{c} = \left[\mathbf{H}^T_{\mathcal{G}(1)}, \mathbf{H}^T_{\mathcal{G}(2)}, \ldots, \mathbf{H}^T_{\mathcal{G}(L)}\right]^T$. Then, the composite channel including the candidate user k reads as

$$\mathbf{H}'_\mathrm{c} = \left[\mathbf{H}^T_\mathrm{c} \;\; \mathbf{H}^T_{\mathcal{K}(k)}\right]^T. \quad (19)$$

We now employ the auxiliary vector $\mathbf{p}_i$ defined as the right singular vector associated with the larger singular value of the decomposition of $\mathbf{H}_\mathrm{c}$ and, similarly, $\mathbf{p}'_i$ for $\mathbf{H}'_\mathrm{c}$ to compute the parameter $C_{\mathrm{sim}k}$ as

$$C_{\mathrm{sim}k} = \frac{\sum_{z=1}^{L} \|\mathbf{H}_{\mathcal{G}(z)}\mathbf{p}_i\|^2}{\sum_{z=1}^{L} \|\mathbf{H}_{\mathcal{G}(z)}\mathbf{p}'_i\|^2 + \|\mathbf{H}_{\mathcal{K}(k)}\mathbf{p}'_i\|^2}. \quad (20)$$

Algorithm 2 Allocation

Input: $\mathcal{K}$, g, $\mathbf{C}_\mathrm{s}$, $\{\mathbf{H}_k\}_{k=1}^K$, γ_ρ, γ_s, δ_ρ, δ_s

$k = \arg\max_{k\in\mathcal{K}} \|\mathbf{H}_{\mathcal{K}(k)}\|_2$
$\mathcal{G}(1) \leftarrow k$
$\mathcal{K} \leftarrow \mathcal{K} \setminus \mathcal{G}(1)$
$\mathbf{H}_\mathrm{c} = \mathbf{H}_k$
$[\mathbf{S}, \Sigma, \mathbf{D}] \leftarrow \mathrm{svd}(\mathbf{H}_\mathrm{c})$
$\mathbf{p}_1 = [\mathbf{D}]_{:,1}$
for $i = 2 : g$ **do**
  for k $= 1{:}|\mathcal{K}|$ **do**
    $\mathcal{G}' = \mathcal{G} \cup \mathcal{K}(k)$
    $\mathbf{H}'_\mathrm{c} = \left[\mathbf{H}^T_\mathrm{c} \;\; \mathbf{H}^T_{\mathcal{K}(k)}\right]^T$
    $[\mathbf{S}, \Sigma, \mathbf{D}] \leftarrow \mathrm{svd}(\mathbf{H}'_\mathrm{c})$
    $\mathbf{p}' = [\mathbf{D}]_{:,1}$
    $C_{\mathrm{sim}k} \leftarrow$ Compute with (20)
    $\tilde{\rho}_k \leftarrow \frac{2!(|\mathcal{G}'|-2)!}{|\mathcal{G}'|!} \sum_{i\in\mathcal{G}'} \sum_{j\in\mathcal{G}', i>j} \rho_{i,j}$
  $\tilde{\mathcal{G}}(i) = \arg\max_{k\in\mathcal{K}} \; |\delta_\rho \tilde{\rho}_k + \delta_\mathrm{s} C_{\mathrm{sim}k}|$
  if $C_{\mathrm{sim}\tilde{\mathcal{G}}(i)} > \gamma_\mathrm{s}$ and $\tilde{\rho}_{\tilde{\mathcal{G}}(i)} > \gamma_\rho$ **then**
    $\mathcal{G}(i) = \tilde{\mathcal{G}}(i)$
    $\mathcal{K} \leftarrow \mathcal{K} \setminus \tilde{\mathcal{G}}(i)$
    $\mathbf{H}_\mathrm{c} = \left[\mathbf{H}^T_\mathrm{c} \;\; \mathbf{H}^T_{\mathcal{G}(i)}\right]^T$
    $[\mathbf{S}, \Sigma, \mathbf{D}] \leftarrow \mathrm{svd}(\mathbf{H}_\mathrm{c})$
    $\mathbf{p}_i = [\mathbf{D}]_{:,1}$
  else return

Output: $\mathcal{G}$

Note that this metric prioritizes users with channels lying in similar subspaces, and takes into account the channel norm to avoid users with low SNRs.

After evaluating the metric in (18) for the set of all the unallocated users, the algorithm selects those users with the highest values for this metric as candidates for being included into the considered group. Then, if the value for the measure obtained for $C_{\mathrm{sim}k}$ and the average cross-correlation $\tilde{\rho}_k$ exceed the given thresholds, that user will actually be included in the sequence $\mathcal{G}$.

3) REMARKS

As a result of applying the previous grouping and allocation algorithms, we will obtain the sequences $\mathcal{G}_i$, $\forall\, i = 1, \ldots, G$ with the user indices per group, and the vector $\mathbf{g}$ which stacks the number of users per group. Thus, the proposed scheduling procedure provides a suitable grouping configuration with G groups and g_i users for the i-th group. The users in each group are also sorted to ensure that the quantized users correspond to those whose equivalent channel matrices have larger singular values.

Next, $G_\mathrm{s} = N^\mathrm{r}_\mathrm{RF}$ groups of users will be accommodated by considering the hardware constraint at the BS (N^r_RF). It is remarkable that Algorithm 1 provides the groups sorted according to the metric m_k. Hence, only the users

corresponding to the first G_s groups will be selected to transmit their encoded symbols after the scheduling procedure.

In summary, the proposed scheduling is characterized by:

- Its objective is to serve the maximum number of users but they must be allocated to the groups in a consistent way to minimize the impact on the system performance.
- The grouping configuration must be as balanced as possible according to a consistent user allocation.
- The number of groups must be as small as possible (with $G \geq N_{RF}^{r}$).
- The trade-off between the number of served users and the resulting distortion in the decoded symbols can be managed through the thresholds γ_ρ and γ_s. By choosing small values for such thresholds, we allow large groups incorporating some users which could negatively increase the impact on the observed distortion. Conversely, by choosing large threshold values, only users with minimum impact on the observed distortion are allowed to be incorporated into the served groups. This leads to groups with few users such that the actual number of served users will be small.
- Although the behaviour of the scheduling algorithm is fundamentally determined by the above thresholds, an additional constraint is introduced by limiting the maximum size of the groups to g_{max}. The value of this parameter should be chosen only to prevent oversized groups in scenarios where the number of users is much larger than the number of RF chains, but not to explicitly define the trade-off between served users and system performance.

C. UNCONSTRAINED PRECODING AND COMBINING

In this section, the design of the user precoders and the receive combiner is addressed for the DQLC-based grouped system considered in the previous subsections. Unconstrained fully digital precoding for MIMO systems has been studied for correlated sources [32], [33]. In particular, we choose the projected Gradient precoding strategy [32, Section III] in order to exploit the inter-user correlation at each group. On the other hand, the digital combiner should be designed to cancel the inter-group interferences in order to avoid penalizing the performance of the DQLC scheme. This strategy hence ensures to split the communication system into non-interfering groups where the users of a group transmit their source symbols using an optimized DQLC scheme, whereas a single RF chain is employed to recover such symbols at the receiver.

The unconstrained digital combiner at the receiver, $\mathbf{W}^* \in \mathbb{C}^{G_s \times N_r}$, is designed to satisfy the following condition

$$[\mathbf{W}^*]_{l,:}\mathbf{H}_{i,j}\mathbf{p}_{\mathrm{H}i,j} = 0, \quad \forall\, i \neq l, \text{ with } l, i = 1, \ldots, G_s, \text{ and } j = 1, \ldots, g_i. \tag{21}$$

Thus, the l-th row of the combiner is onto the nullspace of the equivalent channels corresponding to the users of the remaining $G_s - 1$ groups. Note that the inter-user interference is just cancelled between those users contained in different groups.

In order to exploit the spatial correlation in the groups, while the inter-group interference is cancelled, Algorithm 3 has been implemented. As observed, at the i-th iteration, the precoders for the users of the i-th group are computed by means of the gradient-based precoding strategy developed in [32] considering the part of the correlation matrix corresponding to the users of such a group $\mathbf{C}_{cl}$.

In a similar way, the i-th row of the combiner is also calculated under the premise of satisfying the condition in (21). We follow an approach similar to [13] where the filter is computed in two-separate stages, but adapting it to the requirements of the proposed grouped system. First, a projector $\mathbf{T} \in \mathbb{C}^{N_r \times N_r}$ is used to project the user channels of the i-th group to the orthogonal subspace of the equivalent channels of the users in the $i-1$ previous groups. This projector is initially defined as $\mathbf{T} = \mathbf{I}_{N_r}$ and is then updated at each iteration as

$$\mathbf{T} = \mathbf{T} - \mathbf{B}_i\mathbf{B}_i^*, \tag{22}$$

where $\mathbf{B}_i$ is the basis for the span of the projected equivalent channels at the i-th group, i.e. $\mathbf{T}\tilde{\mathbf{H}}_i$, with

$$\tilde{\mathbf{H}}_i = \left[\mathbf{H}_{i,1}\mathbf{p}_{\mathrm{H}_{i,1}}, \ldots, \mathbf{H}_{i,g_i}\mathbf{p}_{\mathrm{H}_{i,g_i}}\right]. \tag{23}$$

Then, each column of $\tilde{\mathbf{W}}$ can be calculated as the conventional MMSE combiner taking into account only the users contained in the group, i.e.,

$$[\tilde{\mathbf{W}}]_{:,i} = \left(\frac{1}{\sigma_n^2}\mathbf{I}_{N_r} + \mathbf{h}_{\mathcal{G}_i}\mathbf{h}_{\mathcal{G}_i}^*\right)^{-1}\mathbf{h}_{\mathcal{G}_i}, \tag{24}$$

with $\mathbf{h}_{\mathcal{G}_i} = \mathbf{T}\sum_{j=1}^{g_i}\mathbf{H}_{i,j}\mathbf{p}_{\mathrm{H}i,j}$.

Next, in a second step, the residual interferences can be cancelled by projecting the candidate combiner columns onto the nullspace of the equivalent channels of the users scheduled in other groups. Defining $\mathbf{N}_i$ as the basis for the subspace spanned by the columns of the matrix

$$\mathbf{I}_{\mathcal{G}_i} = \left[\tilde{\mathbf{H}}_1, \ldots, \tilde{\mathbf{H}}_{i-1}, \tilde{\mathbf{H}}_{i+1}, \ldots, \tilde{\mathbf{H}}_{G_S}\right],$$

we eventually obtain the columns of the filter as

$$[\mathbf{W}]_{:,i} = \left(\mathbf{I} - \mathbf{N}_i\mathbf{N}_i^*\right)[\tilde{\mathbf{W}}]_{:,i}. \tag{25}$$

Hence, the digital combiner $\mathbf{W}^*$, yields a MAC signal decoupled and stacked in the vector $\mathbf{z} \in \mathbb{C}^{G_s \times 1}$ such that the i-th entry will be

$$z_i = [\mathbf{W}^*]_{i,:}\left(\sum_{j=1}^{g_i}\mathbf{H}_{i,j}\mathbf{p}_{\mathrm{H}_{i,j}}f_{i,j}(s_{i,j}) + \mathbf{n}_i\right) \quad \forall\, i = 1, \ldots, G_s, \tag{26}$$

and each element z_i contains a weighted sum of the g_i encoded symbols from the i-th group. The individual symbols are then estimated by using the demapping function $q_i(z_i)$.

Note that Algorithm 3 employs a matrix factorization algorithm to determine, the hybrid precoders, and the digital and hybrid combiners, considering the limited number of RF chains.

Algorithm 3 GSA

Input: $\{\mathbf{H}_k\}_{k=1}^{K_s}, \mathbf{C}_s, N_{RF}^r, \mathbf{g}$

Initialize: $\mathbf{P}_H = [\,], \mathbf{W} = [\,], \mathbf{T} = \mathbf{I}_{N_r}$, $\mathbf{H}_{comp} = [\,], \mathbf{H}_{inter} = [\,], l = 0$
repeat
  $l \leftarrow l + 1$
  for $u = 1, \ldots, g_l$ **do**
    $\mathbf{H}_{comp} = [\mathbf{H}_{comp}\ \ \mathbf{T}\mathbf{H}_{l,u}]$
  $\mathbf{C}_{cl} \leftarrow$ Inter-user correlation per group
  $\mathbf{P}_{Hcl} = \text{blkdiag}\left(\mathbf{p}_{Hl,1}, \ldots, \mathbf{p}_{Hl,g_l}\right) \leftarrow$ Gradient precoding $\left(\mathbf{H}_{comp}, \mathbf{C}_{cl}\right)$ [32]
  $\mathbf{H}_{inter} = [\mathbf{H}_{inter}\ \ \mathbf{T}\tilde{\mathbf{H}}_i]$
  $\mathbf{B}_i \leftarrow$basis for span($\mathbf{H}_{inter}$)
  $\mathbf{T} = \mathbf{T} - \mathbf{B}_i\mathbf{B}_i^*$
  $\mathbf{h}_{\mathcal{G}_l} = \mathbf{T}\sum_{j=1}^{g_l}\mathbf{H}_{l,j}\mathbf{p}_{Hl,j}$
  $\tilde{\mathbf{W}} = [\tilde{\mathbf{W}}, \left(\frac{1}{\sigma_n^2}\mathbf{I}_{N_r} + \mathbf{h}_{\mathcal{G}_l}\mathbf{h}_{\mathcal{G}_l}^*\right)^{-1}\mathbf{h}_{\mathcal{G}_l}]$
  $\mathbf{P}_H = \text{blkdiag}\,(\mathbf{P}_H, \mathbf{P}_{Hcl})$
until $l = N_{RF}^r$
for $l = 1, 2, \ldots, G_s - 1$ **do**
  $\mathbf{I}_{\mathcal{G}_l} = \left[\tilde{\mathbf{H}}_1, \ldots, \tilde{\mathbf{H}}_{l-1}, \tilde{\mathbf{H}}_{l+1}, \ldots, \tilde{\mathbf{H}}_{G_S}\right]$
  $\mathbf{N}_l \leftarrow$basis for span $\left(\mathbf{I}_{\mathcal{G}_l}\right)$
  $[\mathbf{W}]_{:,l} = \left(\mathbf{I} - \mathbf{N}_l\mathbf{N}_l^*\right)[\tilde{\mathbf{W}}]_{:,l}$
$\mathbf{R} = \mathbf{W}^*\mathbf{H}\mathbf{P}_H$
$\mathbf{W}_H = \mathbf{W}_{RF}\mathbf{W}_{BB} \leftarrow$ **Algorithm 4** $(\mathbf{H}, \mathbf{P}_H, \mathbf{W}, \mathbf{R})$

Output: $\mathbf{P}_H, \mathbf{W}_H$

D. COMBINING WITH A LIMITED NUMBER OF RF CHAINS

The functionality of the combiner at the BS—which is critical to face the demapping process—can be performed by a hybrid combiner. By deploying $N_{RF}^r \geq 2\,G_s$ RF chains, we can use the closed-form expression in [19] therefore canceling completely the inter-group interference with the resulting hybrid combiner. When $G_s \leq N_{RF}^r < 2G_s$, a factorization algorithm from [12], [14] or [16] can be employed to decouple the overall digital combiner into the baseband and RF components. In this case, a gap in the system performance will be observed because of the inter-group interference cannot be totally cancelled. In Section IV we show that the performance offered by the algorithms in [12], [14], [16] for the grouped system can be considerably exceeded by following a different approach to address the optimization problem when $N_{RF}^r = G_s$.

IV. HYBRID COMBINING FOR USER GROUPING IN mmWave MIMO SYSTEMS

The problem of hybrid transceiver design is typically formulated as the solution to the following optimization problem [12], [14], [16]

$$\min_{\mathbf{W}_{BB}, \mathbf{W}_{RF}} \|\mathbf{W} - \mathbf{W}_{RF}\mathbf{W}_{BB}\|_F^2, \quad \text{s.t. } \mathbf{W}_{RF} \in \mathcal{W}_{RF} \tag{27}$$

which represents a non-convex optimization problem due to the constraint on the RF combiner. Since the aim of the hybrid combiner is also to minimize the inter-group interferences even satisfying the condition in (21), we rather focus on preserving the structure of the matrix which integrates the equivalent channel responses of the served users and the unconstrained digital combiner, i.e., $\mathbf{R} = \mathbf{W}^*\mathbf{H}\mathbf{P}_H$. Note that the approach in (27) only guarantees that the hybrid combiner is close to the digital one in the Euclidean space, but it does not impose any constraint on the structure of the joint equivalent response.

The desirable structure for the matrix $\mathbf{R} \in \mathbb{C}^{G_s \times K_s}$ is

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & \ldots & r_{1,g_1} & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & r_{2,1} & \ldots & r_{2,g_2} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & 0 & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & r_{G_s,1} & \ldots & r_{G_s,g_{G_s}} \end{bmatrix}$$

where the zero entries represents inter-group interference, while the non-zero elements are the desirable post-combining equivalent channel gains. Following this premise, we state an optimization problem where the cost function aims at minimizing the difference between the joint equivalent response considering the hybrid combiner and the above matrix $\mathbf{R}$. Thus, we first define the distortion between the response obtained with the digital combiner in (25) and the response obtained with the hybrid combiner, i.e.,

$$d\left(\mathbf{W}_{RF}, \mathbf{W}_{BB}\right) = \|\mathbf{R} - \mathbf{W}_{BB}^*\mathbf{W}_{RF}^*\mathbf{H}\mathbf{P}_H\|_F^2. \tag{28}$$

Then, the optimization problem can be stated as

$$\min_{\mathbf{W}_{BB}, \mathbf{W}_{RF}} \|\mathbf{R} - \mathbf{W}_{BB}^*\mathbf{W}_{RF}^*\mathbf{H}\mathbf{P}_H\|_F^2. \quad \text{s.t. } \mathbf{W}_{RF} \in \mathcal{W}_{RF} \tag{29}$$

We next determine a PG algorithm to solve the non-convex optimization problem in (29) under the assumption $N_{RF}^r = G_s$. The gradient of the cost function (28) is

$$\frac{\partial d}{\partial \mathbf{W}_{RF}^*} = \mathbf{W}_{BB}^*\mathbf{R}\mathbf{P}_H^*\mathbf{H}^* + \mathbf{W}_{BB}^*\mathbf{W}_{BB}\mathbf{W}_{RF}\mathbf{H}\mathbf{P}_H\mathbf{P}_H^*\mathbf{H}^*. \tag{30}$$

Then, at each iteration of the algorithm, the unconstrained solution is given by

$$\tilde{\mathbf{W}}_{RF}^* = \mathbf{W}_{RF}^* - \mu\frac{\partial d}{\partial \mathbf{W}_{RF}^*}, \tag{31}$$

which is then projected onto the set of feasible solutions $\mathcal{W}_{RF}$ according to the aforementioned RF hardware constraints. The initial matrix is given by the projection of the digital combiner $\mathbf{W}$ onto the set $\mathcal{W}_{RF}$, and the step size μ is diminished in order to reach a local optimum. The well-known least squares (LS) solution is employed to update the baseband combiner by using the closed-form expression

$$\mathbf{W}_{BB} = \mathbf{R}(\mathbf{P}_H^*\mathbf{H}^*\mathbf{W}_{RF})^\dagger. \tag{32}$$

The iterative algorithm is stopped when the distortion d falls below a certain threshold value δ or when the maximum number of iterations ϵ is reached. Algorithm 4 summarizes the

Algorithm 4 PG

Input: $\mathbf{H} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t} K_\mathrm{s}}$, $\mathbf{P}_\mathrm{H} \in \mathbb{C}^{N_\mathrm{t} K_\mathrm{s} \times K_\mathrm{s}}$, $\mathbf{W} \in \mathbb{C}^{N_\mathrm{r} \times G_\mathrm{s}}$, $\mathbf{R} \in \mathbb{C}^{G_\mathrm{s} \times K_\mathrm{s}}$, μ_0, δ, ϵ

1: **Initialize:** $\ell \leftarrow 0$
2: $[\mathbf{W}_\mathrm{RF}]_{i,j}^{(0)} = \frac{1}{\sqrt{N_\mathrm{r}}} \exp\left(j \arg\left([\mathbf{W}]_{i,j}\right)\right), \ \forall\, i,j$
3: $\mu \leftarrow \mu_0$
4: **repeat**
5: $\quad \ell \leftarrow \ell + 1$
6: $\quad \tilde{\mathbf{W}}_\mathrm{RF}^{*} \leftarrow \mathbf{W}_\mathrm{RF}^{*\,(\ell-1)} - \mu \frac{\partial d}{\partial \mathbf{W}_\mathrm{RF}^{*}}$
7: $\quad [\mathbf{W}_\mathrm{RF}]_{i,j}^{(\ell)} = \exp\left(j \arg\left([\tilde{\mathbf{W}}_\mathrm{RF}]_{i,j}\right)\right), \ \forall\, i,j$
8: $\quad \mathbf{W}_\mathrm{BB}^{(\ell)} = \mathbf{R}\left(\mathbf{P}_\mathrm{H}^{*}\mathbf{H}^{*}\mathbf{W}_\mathrm{RF}^{(\ell)}\right)^{\dagger}$
9: $\quad$ **if** $d\left(\mathbf{W}_\mathrm{RF}^{(\ell-1)}, \mathbf{W}_\mathrm{BB}^{(\ell-1)}\right) \leq d\left(\mathbf{W}_\mathrm{RF}^{(\ell)}, \mathbf{W}_\mathrm{BB}^{(\ell)}\right)$ **then**
10: $\quad\quad \mu \leftarrow \mu/2$
11: **until** $d\left(\mathbf{W}_\mathrm{RF}^{(\ell)}, \mathbf{W}_\mathrm{BB}^{(\ell)}\right) < \delta$ or $\ell \geq \epsilon$
12: $\mathbf{W}_\mathrm{H} = \mathbf{W}_\mathrm{RF}\mathbf{W}_\mathrm{BB}$

Output: $\mathbf{W}_\mathrm{H}$

proposed strategy to solve the hybrid factorization problem for the combiner at the BS.

V. HYBRID COMBINER FOR CORRELATED SOURCES

In this section, we move to a scenario where the number of RF chains is assumed equal to the number of users, i.e., $N_\mathrm{RF}^\mathrm{r} = K$, and where the user information is spatially correlated. Therefore, we can consider the conventional ungrouped solution where all the users simultaneously transmit one source symbol per channel use employing an uncoded scheme, that is, the mapping function applied to each user is just a scale factor to satisfy the power constraints. Note that the scenario approached in this section can be seen as a one-user-per-group system where the equality $N_\mathrm{RF}^\mathrm{r} = K = G_\mathrm{s}$ also holds. In this scenario, we focus on the common hybrid receiver design at the BS, but exploiting the source correlation. Hybrid transceivers for multiuser has been addressed in various scenarios [13], [21], [34] using ZF strategies. In these schemes, unlike the group-based approach, the inter-user interference must be cancelled for uncorrelated sources, but this strategy would no longer be adequate in the presence of correlation.

A. PROPOSED HYBRID MMSE COMBINER

We propose a combiner design where the conventional cancellation of the inter-user interference using a ZF strategy—which completely ignores the spatial correlation—is replaced with a MMSE combiner which exploits the source correlation by incorporating the correlation matrix $\mathbf{C}_\mathrm{s}$. The user precoders stacked in $\mathbf{P}_\mathrm{H}$ are calculated by considering the gradient-based precoding strategy proposed in [32] in order to also exploit the spatial correlation.

In the computation of the hybrid combiner, the RF component $\mathbf{W}_\mathrm{RF}$ is first obtained by projecting the digital

Algorithm 5 Hybrid MMSE

Input: $\{\mathbf{H}_k\}_{k=1}^{K_\mathrm{s}}$, $\mathbf{C}_\mathrm{s}$
1: $\mathbf{P}_\mathrm{H} = \mathrm{blkdiag}\left(\mathbf{p}_{\mathrm{H}1}, \ldots, \mathbf{p}_{\mathrm{H}K_\mathrm{s}}\right) \leftarrow$ Gradient precoding [32]
2: $\mathbf{W}_\mathrm{MMSE} = \left(\mathbf{H}\mathbf{P}_\mathrm{H}\mathbf{C}_\mathrm{s}\mathbf{P}_\mathrm{H}^{*}\mathbf{H}^{*} + \sigma_\mathrm{n}^2\mathbf{I}_{N_\mathrm{r}}\right)^{-1}\mathbf{H}\mathbf{P}_\mathrm{H}\mathbf{C}_\mathrm{s}$
3: $[\mathbf{W}_\mathrm{RF}]_{i,j} = \frac{1}{\sqrt{N_\mathrm{r}}} \exp\left(j \arg\left([\mathbf{W}_\mathrm{MMSE}]_{i,j}\right)\right), \ \forall\, i,j$
4: $\mathbf{R}_\mathrm{RF} = \mathbf{W}_\mathrm{RF}^{*}\tilde{\mathbf{H}}$
5: $\mathbf{W}_\mathrm{BB} = \left(\mathbf{R}_\mathrm{RF}\mathbf{C}_\mathrm{s}\mathbf{R}_\mathrm{RF}^{*} + \sigma_\mathrm{n}^2\mathbf{W}_\mathrm{RF}^{*}\mathbf{W}_\mathrm{RF}\right)^{-1}\mathbf{R}_\mathrm{RF}\mathbf{C}_\mathrm{s}$
6: $\mathbf{W}_\mathrm{H} = \mathbf{W}_\mathrm{RF}\mathbf{W}_\mathrm{BB}$

Output: $\mathbf{W}_\mathrm{H}$

MMSE combiner $\mathbf{W}_\mathrm{MMSE} \in \mathbb{C}^{N_\mathrm{r} \times K_\mathrm{s}}$ onto the set of feasible combiners—which consists on keeping only the phases of its entries—. The baseband component $\mathbf{W}_\mathrm{BB} \in \mathbb{C}^{N_\mathrm{RF}^\mathrm{r} \times K_\mathrm{s}}$ is then derived as

$$\mathbf{W}_\mathrm{BB} = \left(\mathbf{R}_\mathrm{RF}\mathbf{C}_\mathrm{s}\mathbf{R}_\mathrm{RF}^{*} + \sigma_\mathrm{n}^2\mathbf{W}_\mathrm{RF}^{*}\mathbf{W}_\mathrm{RF}\right)^{-1}\mathbf{R}_\mathrm{RF}\mathbf{C}_\mathrm{s}, \quad (33)$$

by using a MMSE-based approach, incorporating the correlation matrix $\mathbf{C}_\mathrm{s}$ and employing the equivalent channels comprising the RF combiner, i.e.,

$$\mathbf{R}_\mathrm{RF} = \mathbf{W}_\mathrm{RF}^{*}\tilde{\mathbf{H}}, \quad (34)$$

where $\tilde{\mathbf{H}} = \left[\mathbf{H}_1\mathbf{p}_{\mathrm{H}1}, \ldots, \mathbf{H}_{K_\mathrm{s}}\mathbf{p}_{\mathrm{H}K_\mathrm{s}}\right]$ stacks the equivalent channel vectors for the K_s served users in the ungrouped system.

The Algorithm 5 summarizes the procedure to compute the hybrid combiner which exploits the inter-user correlation for an uncoded system with $K_\mathrm{s} = K$ served users and $N_\mathrm{RF}^\mathrm{r} = K_\mathrm{s}$ receive chains at the BS.

VI. COMPUTATIONAL COMPLEXITY OF THE PROPOSED ALGORITHMS

In this section, the computational complexity of algorithms 2, 3, 4 and 5 is analyzed. The Algorithm 1 has been excluded from this analysis since the computational cost of the operations required in this algorithm is negligible. Table 1 summarizes the steps with highest computational cost for each algorithm and the overall complexity order.

The steps 11, 13 and 20 have been included to analyze the computational complexity of Algorithm 2. As observed, steps 11 and 20 require to perform an SVD, and therefore their complexity order is $\mathcal{O}\left(N_\mathrm{r}N_\mathrm{t}^2 g_\mathrm{max}\right)$, whereas the complexity of step 13 is $\mathcal{O}\left(N_\mathrm{r}N_\mathrm{t} g_\mathrm{max}\right)$. Since Algorithm 1 is an iterative procedure where each operation is repeated g_max iterations at most, we conclude that the complexity order for this algorithm is $\mathcal{O}\left(N_\mathrm{r}N_\mathrm{t}^2 g_\mathrm{max}^2\right)$.

The step 5 (channel projections), the step 7 (gradient precoding), the step 9 (basis for the subspace span($\mathbf{H}_\mathrm{inter}$)), the step 12 (MMSE combiner), the step 17 (basis for the subspace span($\mathbf{I}_{\mathcal{G}_l}$)) and the step 19 (composite channel matrix) are considered to evaluate the computational complexity of Algorithm 3. In this algorithm, the steps 7 (computation of the user precoding by using the gradient

TABLE 1. Computational complexity of the proposed algorithms.

Algorithm	Operation	Complexity
Algorithm 2 Allocation	Compute svd $(\mathbf{H}'_{\mathrm{c}})$ (step 11)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 g_{\max}\right)$
	Compute $C_{\mathrm{sim}\,k}$ (20) (step 13)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}} g_{\max}\right)$
	Compute svd $(\mathbf{H}_{\mathrm{c}})$ (step 20)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 g_{\max}\right)$
	Overall	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 g_{\max}^2\right)$
Algorithm 3 GSA	Compute projections (step 5)	$\mathcal{O}\left(N_{\mathrm{r}}^2 N_{\mathrm{t}} K_{\mathrm{s}}\right)$
	Compute gradient precoding [32] (step 7)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 K_{\mathrm{s}}^2 \epsilon\right)$
	Compute basis for the subspace span $(\mathbf{H}_{\mathrm{inter}})$ (step 9)	$\mathcal{O}\left(N_{\mathrm{r}} g_{\max}^2\right)$
	Compute MMSE combiner (step 12)	$\mathcal{O}\left(N_{\mathrm{r}}^3\right)$
	Compute basis for the subspace span $\left(\mathbf{I}_{\mathcal{G}_l}\right)$ (step 17)	$\mathcal{O}\left(N_{\mathrm{r}}(K_{\mathrm{s}} - g_l)^2\right)$
	Compute composite channel matrix $(\mathbf{R})$ (13) (step 19)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}K_{\mathrm{s}}^2\right)$
	Overall	$\mathcal{O}\left(N_{\mathrm{r}}^3 N_{\mathrm{RF}}^{\mathrm{r}}\right) + \mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 N_{\mathrm{RF}}^{\mathrm{r}} K_{\mathrm{s}}^2 \epsilon\right)$
Algorithm 4 PG	Compute RF combiner (step 6)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}N_{\mathrm{RF}}^{\mathrm{r}} K_{\mathrm{s}}\right)$
	Compute baseband combiner (step 8)	$\mathcal{O}\left({N_{\mathrm{RF}}^{\mathrm{r}}}^3\right)$
	Overall	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}N_{\mathrm{RF}}^{\mathrm{r}} K_{\mathrm{s}}\epsilon\right)$
Algorithm 5 Hybrid MMSE	Compute gradient precoding [32] (step 1)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 K_{\mathrm{s}}^2 \epsilon\right)$
	Compute MMSE combiner (step 2)	$\mathcal{O}\left(N_{\mathrm{r}}^3\right)$
	Compute composite channel matrix $(\mathbf{R}_{\mathrm{RF}})$ (34) (step 4)	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{RF}}^{\mathrm{r}} K_{\mathrm{s}}\right)$
	Compute baseband combiner (step 5)	$\mathcal{O}\left(K_{\mathrm{s}}^3\right)$
	Overall	$\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}^2 K_{\mathrm{s}}^2 \epsilon\right) + \mathcal{O}\left(N_{\mathrm{r}}^3\right)$

algorithm [32], which requires ϵ iterations at most) and 12 (computation of each MMSE combiner row) are the more complex operations, leading to an overall computational complexity order $\mathcal{O}(N_{\mathrm{r}}^3 N_{\mathrm{RF}}^{\mathrm{r}}) + \mathcal{O}(N_{\mathrm{r}}N_{\mathrm{t}}^2 N_{\mathrm{RF}}^{\mathrm{r}} K_{\mathrm{s}}^2 \epsilon)$ by considering that these steps are repeated $N_{\mathrm{RF}}^{\mathrm{r}}$ times. The steps 5, 9, 17 and 19 present a complexity order of $\mathcal{O}\left(N_{\mathrm{r}}^2 N_{\mathrm{t}} K_{\mathrm{s}}\right)$, $\mathcal{O}\left(N_{\mathrm{r}} g_{\max}^2\right)$, $\mathcal{O}\left(N_{\mathrm{r}}(K_{\mathrm{s}} - g_l)^2\right)$ and $\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}K_{\mathrm{s}}^2\right)$, respectively. Although the complexity of these operations is increased by the number of iteration needed in the Algorithm 3 to accommodated the user groups per RF chains, they still involve a lower computational complexity than the one reached in steps 7 and 12.

The complexity order of the steps 6 and 8 in Algorithm 4 is $\mathcal{O}\left(N_{\mathrm{r}}N_{\mathrm{t}}N_{\mathrm{RF}}^{\mathrm{r}} K_{\mathrm{s}}\right)$ and $\mathcal{O}({N_{\mathrm{RF}}^{\mathrm{r}}}^3)$, respectively. Therefore, the main contribution to the computational complexity corresponds to the operation computed in the step 6 to obtain the RF combiner as the number of RF chains is assumed to be a small value in hybrid systems. It is remarkable that the computational complexity of the whole algorithm is increased by the number of iterations ϵ required to reach

the convergence, which yields to an overall complexity order $\mathcal{O}(N_{\mathrm{r}}N_{\mathrm{t}}N_{\mathrm{RF}}^{\mathrm{r}}K_{\mathrm{s}}\epsilon)$.

Finally, the computational complexity of Algorithm 5 is also analyzed. The complexity of the steps 1, 2, 4 and 5 is summarized in Table 1. As observed, the steps 1 and 2 require the highest computational complexity, since they involve the computation of the gradient precoding [32] and the MMSE combiner. The complexity order for these operations is $\mathcal{O}(N_{\mathrm{r}}N_{\mathrm{t}}^2K_{\mathrm{s}}^2\epsilon)$ and $\mathcal{O}(N_{\mathrm{r}}^3)$, respectively. Since the complexity of the steps 4 ($\mathcal{O}(N_{\mathrm{r}}N_{\mathrm{RF}}^{\mathrm{r}}K_{\mathrm{s}})$) and 5 ($\mathcal{O}(K_{\mathrm{s}}^3)$) is lower than those of the steps 1 and 2, the overall complexity order for this algorithm will be $\mathcal{O}(N_{\mathrm{r}}N_{\mathrm{t}}^2K_{\mathrm{s}}^2\epsilon) + \mathcal{O}(N_{\mathrm{r}}^3)$.

VII. SIMULATION RESULTS

In this section, we evaluate the performance and impact of the proposed solutions by providing numerical results and comparisons to other suitable schemes.

Let us consider an exponential correlation model where the coefficients of the covariance matrix are given by $[\mathbf{C}_{\mathrm{s}}]_{i,j} = \rho^{|i-j|}$, $\forall\ i,j$, with ρ the correlation factor. At each time instant, the user symbols are assumed to be generated from a multivariate circularly-symmetric Gaussian distribution with zero mean and covariance matrix according to the exponential correlation model. In the computer experiments, we assume that the maximum size of the user groups is limited to $g_{\max} = 4$, since this value provides enough flexibility to the scheduling algorithm while mitigating the potential loss caused by larger groups. A MIMO configuration with $N_{\mathrm{t}} = 16$ antennas per user and $N_{\mathrm{r}} = 49$ antennas at the BS is considered. In order to randomly simulate the user channels, the maximum and the minimum number of clusters and rays per user is set to $N_{\mathrm{cl}_{\max}} = 4$, $N_{\mathrm{ray}_{\max}} = 3$ and $N_{\mathrm{cl}_{\min}} = 1$, $N_{\mathrm{ray}_{\min}} = 1$, respectively. Then, the channel response of the j-th user in the i-th group is randomly modeled by assigning $N_{\mathrm{cl}\,i,j} \in \{N_{\mathrm{cl}_{\min}}, \ldots, N_{\mathrm{cl}_{\max}}\}$ clusters compounded by $N_{\mathrm{ray}\,i,j} \in \{N_{\mathrm{ray}_{\min}}, \ldots, N_{\mathrm{ray}_{\max}}\}$ rays. The path gain parameter $\beta_{n,m}$ is modelled as a random variable following a standard complex-valued Gaussian distribution, that is, $\beta_{n,m} \sim \mathcal{CN}(0, 1)$. The angles of departure and arrival (AoA/AoD) follow the Laplacian distribution where the mean angles are uniformly randomly distributed in $[0, 2\pi)$, whereas the angular spread is fixed to 10 degrees like in [16].

The performance of the considered communication systems is assessed in terms of the average signal-to-distortion ratio (SDR) which is computed as

$$\mathrm{SDR\ (dB)} = 10\log_{10}\left(\frac{1}{\hat{\xi}_{\mathrm{sum}}}\right),$$

where

$$\hat{\xi}_{\mathrm{sum}} = \frac{1}{NK_{\mathrm{s}}}\sum_{n=1}^{N}\sum_{i=1}^{G_{\mathrm{s}}}\sum_{j=1}^{g_i}|s_{n,i,j} - \hat{s}_{n,i,j}|^2 \qquad (35)$$

represents the average MSE between the source symbols and the estimated ones obtained after the demapping operation. The results reported in this section were computed by averaging the SDR over $N = 1000$ channel realizations.

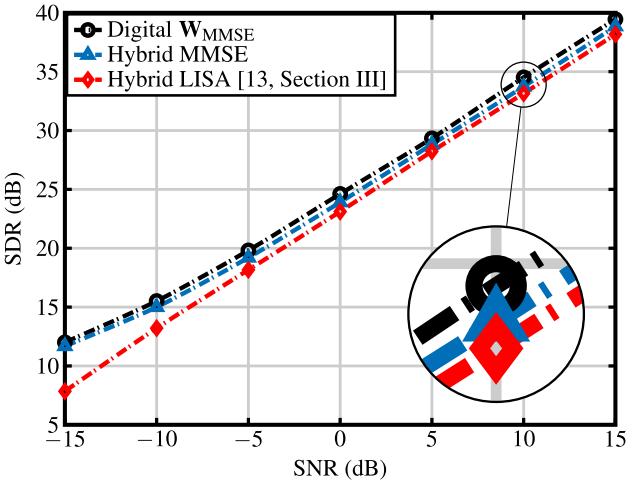

FIGURE 2. Performance of the different hybrid strategies for the uncoded scenario with $K = 10$ users, $N_{\mathrm{RF}}^{\mathrm{r}} = 10$ and $\rho = 0.80$.

Without loss of generality, we assume $\sigma_{\mathrm{n}}^2 = 1$, such that the SNR per user can be defined from its corresponding power constraint, i.e., $\mathrm{SNR\ (dB)}_{i,j} = 10\log_{10}(T_{i,j})$, $\forall\ i = 1, \ldots, G_{\mathrm{s}}$ and $j = 1, \ldots, g_i$. For simplicity, the same power constraint is imposed to all the users such that a single SNR value can be considered for all the users in the simulations. Finally, the parameter ϵ (maximum number of iterations) in Algorithm 4 is set to 1000 and the same number of iterations is fixed for the other algorithms of matrix factorization in [12], [14] and [16] as well as for the gradient precoding algorithm in [32].

Several experiments have been carried out to evaluate the performance of the different proposed methods. First, we evaluate the hybrid MMSE-based strategy for uncoded systems developed in Section V. The performance achieved by the proposed PG algorithm in Section IV to solve the hybrid combining for user grouping is compared to the conventional strategies for matrix factorizations derived in [12, Algorithm 1], [14, Algorithm 4] and [16, MO-AltMin Algorithm]. Next, we evaluate the performance gain obtained with the scheduling and the allocation policies presented in Section III. In this context, we also analyze the impact of the thresholds γ_ρ and γ_{s} as well as the weight factors δ_ρ and δ_{s}. Finally, the performance of DQLC techniques is compared to a conventional NOMA approach – based on the uncoded transmission of the source symbols with power allocation – in the grouped system.

A. UNCODED SCENARIO

In this section, we evaluate the behaviour of the proposed hybrid MMSE approach for uncoded correlated sources. Figure 2 shows the SDR obtained by considering different strategies for the design of the hybrid combiner in the uncoded scenario with $K = 10$ users, $N_{\mathrm{RF}}^{\mathrm{r}} = 10$ RF chains and correlation factor $\rho = 0.8$. In this setup, all the users are assumed to transmit their source information simultaneously,

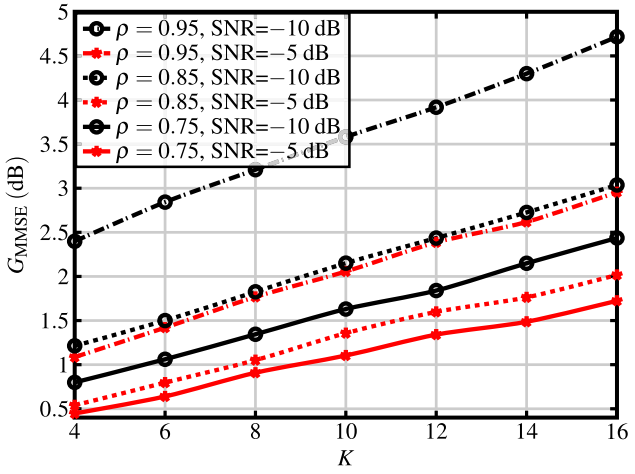

FIGURE 3. Performance gain G_{MMSE} (dB) of the proposed hybrid MMSE algorithm over H-LISA vs number of users for different correlation factors $\rho \in \{0.75, 0.85, 0.95\}$, and SNR (dB) $\in \{-10$ dB, -5 dB$\}$.

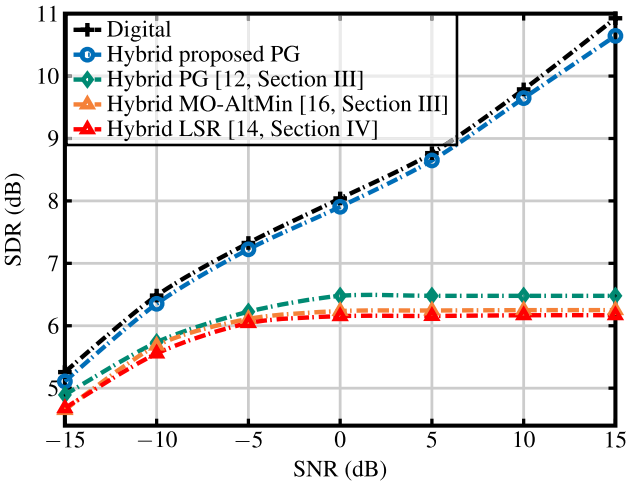

FIGURE 4. Performance obtained for different strategies of combining in a scenario with $K = 6$ users, $N_{\text{RF}}^{\text{r}} = 2$, $\rho = 0.80$, and scheduling parameters $\gamma_\rho = 0$, $\gamma_{\text{s}} = 0$, $\delta_\rho = 0.2$ and $\delta_{\text{s}} = 0.8$.

i.e., $K_{\text{s}} = K$. The conventional fully digital MMSE combiner, the Hybrid-Linear Successive Allocation (H-LISA) algorithm and the strategy proposed in Algorithm 5 for hybrid implementation are compared. In addition, the gradient precoding strategy [32] has jointly been used with the digital MMSE combiner and the proposed Hybrid MMSE combiner. Recall that the approach H-LISA, inspired on digital Linear Successive Allocation (LISA) for traditional digital precoding, is based on two stages to finally cancel the inter-user interference by applying ZF in the second stage [13, Section III]. The algorithms from [12], [14], [16] present significantly worse performance than H-LISA for the case $N_{\text{RF}}^{\text{r}} = K_{\text{s}}$, and hence they have not been included in the figure. As expected, the hybrid implementation based on MMSE provides higher performance than that of the hybrid implementation based on the H-LISA algorithm, especially for low SNR values. On the other hand, the loss of the proposed MMSE-based hybrid design with respect to the fully digital MMSE combiner is almost negligible for all the range of SNR values. Thus, the impact of the proposed factorization algorithm on the system performance is minimum for the restrictive scenario given by the constraint $N_{\text{RF}}^{\text{r}} = K$.

We now evaluate the performance gain of the proposed hybrid MMSE combiner over the H-LISA algorithm. The gain is defined as G_{MMSE} (dB) $=$ $\text{SDR}_{\text{H-MMSE}}$ (dB) $-$ SDR_{HL} (dB), where SDR_{HL} represents the SDR (dB) obtained by using the H-LISA algorithm and $\text{SDR}_{\text{H-MMSE}}$ the one provided by the proposed hybrid MMSE combining approach. In Figure 3, this gain is plotted versus the number of served users, K_{s}, considering several correlation factors and two particular SNR values, SNR (dB) $\in \{-10$ dB, -5 dB$\}$. As expected, the performance gain obtained by using the proposed MMSE solution is higher for the lower SNR levels, since the MMSE exploits the spatial correlation. Figure 3 also illustrates that the performance gain increases with the correlation factor and the number of users.

B. HYBRID COMBINING DESIGN

The performance achieved by the proposed PG algorithm in Section IV to solve the hybrid combining for user grouping is compared to that of the conventional strategies for matrix factorizations derived in [12, Algorithm 1], [14, Algorithm 4] and [16, MO-AltMin Algorithm].

Figure 4 shows the SDR achieved by considering $K = 6$ users, $N_{\text{RF}}^{\text{r}} = 2$, $\rho = 0.8$ and different strategies for the combining process. Specifically, the fully digital combiner, the strategies of factorization employed in [12], [14], [16], and the proposed hybrid combining are compared. In this simulation we have considered $\gamma_\rho = 0$ and $\gamma_{\text{s}} = 0$, i.e., $K_{\text{s}} = K$ users are served at each channel realization. We have also considered $\delta_\rho = 0.8$ and $\delta_{\text{s}} = 0.2$ because those values provide good performance for the grouped system in the considered correlation level. In any way, the same grouping and allocation policies are considered for all the cases. It is shown that the proposed hybrid strategy does not lead to a significant loss in terms of SDR regarding the digital combiner (less than 1 dB). Figure 4 also illustrates that the proposed hybrid algorithm significantly outperforms the combiners based on the factorization of the digital solution in [12], [14] and [16], which saturate in the medium and high SNR regime. This saturation of the system performance when employing the algorithms in [12], [14] and [16] is motivated because the parameter optimization of DQLC is performed according to the SINR of the users per group, which saturates when using these matrix factorization algorithms. Note that these algorithms disregard the desired structure before the demapping functions, and thus the interference between users corresponding to different groups will not be properly cancelled.

C. SCHEDULING

We now evaluate the performance gain provided by the scheduling procedure designed in Section III. We also

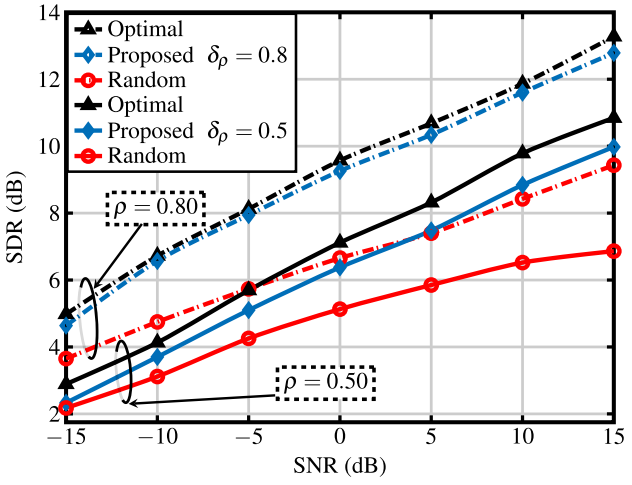

FIGURE 5. Performance for $K = 5$ users, $N_{RF}^{r} = 2$ and $\rho = 0.80$ and different policies of scheduling and allocation.

analyze the impact of the thresholds γ_ρ and γ_s as well as the weight factors δ_ρ and δ_s. Recall that the optimization of these design parameters would lead to an exhaustive search in a four-dimension parameter space or to highly non-convex optimization problems involving non-linearities and discontinuities in the DQLC mapping function. We have considered a heuristic approach to choose these parameters, which is based on the insight provided by the issues related to the DQLC behaviour presented in Section III. In this section, some numerical results are presented to confirm the validity of the considered approach.

In the first experiment, the proposed scheduling to determine the groups and accommodate the users is evaluated by comparison with two benchmark scenarios. System performance for three scheduling policies are plotted in Figure 5: 1) random grouping and allocation, 2) the proposed scheduling following jointly Algorithm 1 and Algorithm 2, and 3) the optimal combinatorial strategy. We assume a small number of users ($K = 5$) to be served in order to compute the performance of all the grouping and allocation possibilities, and select the optimal solution with an affordable computational complexity. The proposed PG algorithm is employed to perform the hybrid combining process.

The performance comparison is carried out for two different correlation factors $\rho \in \{0.50, 0.80\}$, so that the corresponding design parameters of the proposed scheduling should be determined for each case. In particular, such parameters are $\gamma_\rho = 0.5$, $\gamma_s = 0.35$, $\delta_\rho = 0.8$ and $\delta_s = 0.2$ for $\rho = 0.80$, whereas we consider $\gamma_\rho = 0.2$, $\gamma_s = 0.35$, $\delta_\rho = 0.5$ and $\delta_s = 0.5$ for $\rho = 0.50$. As observed, the similarity threshold is the same because the same channel model is used, while the rest of parameters are conveniently adjusted depending on the correlation level. For the sake of fairness, the same number of served users K_s is imposed on each strategy at each channel realization.

As observed in Figure 5, the proposed scheduling provides a remarkable performance gain with respect to the random

TABLE 2. Gap (dB) between the performance obtained by the optimal scheduling and the proposed policies for $\rho = 0.80$.

SNR (dB)	$\delta_\rho = 1$ $\delta_s = 0$	$\delta_\rho = \mathbf{0.8}$ $\delta_s = \mathbf{0.2}$	$\delta_\rho = 0.5$ $\delta_s = 0.5$	$\delta_\rho = 0.3$ $\delta_s = 0.7$
-15	0.68	**0.34**	0.41	0.61
-10	0.69	**0.16**	0.50	1.11
-5	0.69	**0.19**	0.57	1.01
0	0.55	**0.32**	0.69	1.10
5	0.68	**0.34**	0.61	0.90
10	0.95	**0.27**	0.56	0.89
15	0.90	**0.49**	0.74	1.05

strategy for both correlation factors and in the whole range of SNRs. However, the more interesting conclusion is that the proposed policies closely approach the performance of the optimal combinatorial solution with much lower computational cost. These results are especially important since they confirm the suitability of the proposed scheduling procedure for the DQLC-based grouped system, at least for a small number of users. It is not possible to ensure that this behaviour holds for scenarios with a larger number of users because of the impossibility of implementing the combinatorial strategy with an affordable computational complexity. For the considered scenario, the number of grouping possibilities boils down in general to two possible configurations, but the number of allocation options is already 30 in the worst case, when $K_s = 3$, since in this case $g_1 = 1$ and $g_2 = 2$, and the number of allocation possibilities is

$$K \times \frac{(K-1)!}{g_2!\,(K-1-g_2)!}, \tag{36}$$

i.e., for the first group, we can select one of the $K = 5$ users, and then we have to select $g_2 = 2$ users among the four available ones (combinations without repetition). Note that a combinatorial explosion is inevitable as the number of available users to be allocated increases. For a specific group configuration, the number of allocation possibilities is

$$\prod_{i=1}^{G_s} \frac{(K - \sum_{j=1}^{i-1} g_j)!}{g_i!\,(K - \sum_{j=1}^{i} g_j)!}. \tag{37}$$

On the other hand, Table 2 and Table 3 show the performance loss in dB of the proposed scheduling with respect to the optimal combinatorial solution for $\rho = 0.80$ and $\rho = 0.50$, respectively. As observed, different parameter configurations are included in these tables showing that the weight factors δ_ρ and δ_s depend on the overall correlation of the users. As expected, it is preferable to lower the weight of the correlation-based metric, δ_ρ, as the overall correlation in the system decreases. Another interesting observation is that the loss of the proposed scheduling with different weight factors is quite stable, and therefore the use of the optimal values for these parameters is not critical.

The extreme situation occurs when the source symbols are uncorrelated, i.e., $\rho = 0$. Table 4 shows the performance

TABLE 3. Gap (dB) between the performance obtained by the optimal scheduling and the proposed policies for $\rho = 0.50$.

SNR (dB)	$\delta_\rho = 1$ $\delta_s = 0$	$\delta_\rho = 0.8$ $\delta_s = 0.2$	$\delta_\rho = \mathbf{0.5}$ $\delta_s = \mathbf{0.5}$	$\delta_\rho = 0.3$ $\delta_s = 0.7$
-15	0.60	0.71	**0.56**	0.61
-10	0.67	0.58	**0.43**	0.58
-5	0.64	0.46	**0.41**	0.65
0	0.84	**0.64**	0.74	0.74
5	1.28	0.93	**0.91**	1.02
10	1.01	1.05	**0.95**	0.99
15	1.15	0.97	**0.86**	0.99

TABLE 4. Gap (dB) between the performance obtained by the optimal scheduling and the proposed policies for $\rho = 0$.

SNR (dB)	$\delta_\rho = 1$ $\delta_s = 0$	$\delta_\rho = 0$ $\delta_s = 1$	Random
-15	0.88	**0.52**	1.06
-10	0.87	**0.58**	1.32
-5	1.05	**0.77**	1.71
0	1.24	**0.96**	2.25
5	1.89	**0.99**	2.72
10	2.21	**1.14**	3.27
15	2.45	**1.12**	3.66

loss regarding the optimal scheduling as in the previous case but for $\rho = 0$ and considering three different strategies: 1) the proposed scheduling with $\delta_\rho = 1$ and $\delta_s = 0$; 2) the proposed scheduling with $\delta_\rho = 0$ and $\delta_s = 1$; and 3) the random scheduling. Note that the first approach would correspond to the case where the grouping configuration is obtained according to Algorithm 1. The user allocation, however, is actually random since the users are not correlated. The second approach corresponds to the proposed strategy with grouping and allocation depending on the channel similarity metric. Hence, the difference between both approaches is an interesting measure of the impact of allocating the users with the proposed Algorithm 2, even when no correlation is present. As observed, this gain goes from 0.3 dB to 1.3 dB for high SNR values. On the other hand, the difference between the second and fourth columns is a measure of the gain provided by the proposed grouping procedure.

In the next experiment, we measure the gain of the proposed scheduling with respect to the random strategy as the number of available users K increases. This gain is defined as

$$G_{\text{sc}}\ (\text{dB}) = \text{SDR}_{\text{sc}}(\text{dB}) - \text{SDR}_{\text{rm}}(\text{dB}), \tag{38}$$

where $\text{SDR}_{\text{sc}}(\text{dB})$ and $\text{SDR}_{\text{rm}}(\text{dB})$ represent the SDR (dB) obtained by the proposed policies and the random policies, respectively. Figure 6 plots the gains obtained for $N_{\text{RF}}^{\text{r}} = 4$, $\rho \in \{0.50, 0.80, 0.95\}$ and SNR (dB) = 10 dB, against the number of users K. The thresholds and the weight factors employed in the scheduling procedure are shown in

TABLE 5. Values of the scheduling parameters for the different correlation factors.

Parameter	$\rho = 0.95$	$\rho = 0.8$	$\rho = 0.5$
δ_ρ	0.8	0.8	0.5
δ_s	0.2	0.2	0.5
γ_ρ	0.8	0.6	0.3
γ_s	0.35	0.35	0.35

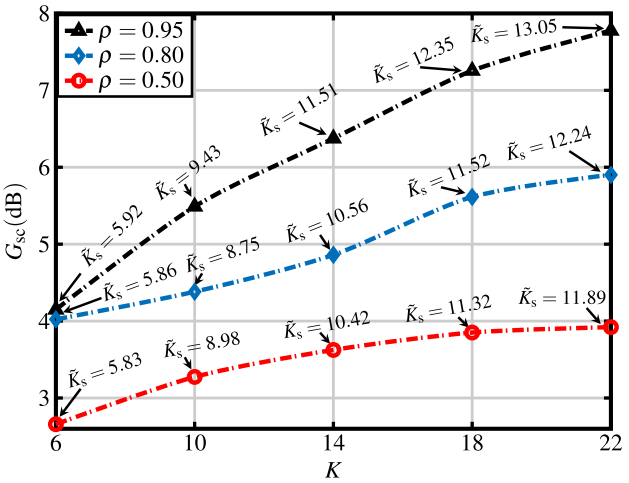

FIGURE 6. G_{sc} (dB) versus K with $N_{\text{RF}}^{\text{r}} = 4$, SNR (dB) = 10 dB and $\rho \in \{0.95, 0.80, 0.50\}$.

Algorithm 5 for the considered correlation factors. The mean number of served users, $\tilde{K}_s$, is also represented at each SNR level. As intuitively expected, the performance gain offered by the proposed algorithms increases with the number of users. In that case, the performance of the random scheduling will be increasingly degraded since the probability of grouping low-correlated users with different and low-capacity channels is higher. Figure 6 also shows that the gain G_{sc} (dB) increases when the correlation becomes higher. This is because the spatial correlation exploitation is also incorporated in the proposed grouping and allocation policies.

Finally, we evaluate the impact of varying the threshold parameters γ_ρ and γ_s on the number of served users for the grouped system. As commented, we can balance the trade-off between the number of served users and the system performance by adjusting these thresholds conveniently. Figure 7 shows the impact of increasing the correlation threshold, γ_ρ, for $K = 16$, $N_{\text{RF}}^{\text{r}} = 4$ and $\rho = 0.9$. The results confirm that a low correlation threshold prioritizes the maximization of the number of served users, whereas a high correlation threshold would preserve the level of signal quality by grouping just the most correlated sources in the same group. The choice of the weight parameters δ_ρ and δ_s also plays an important role in order to deciding the grouping and the allocation of the users. Figure 7 also shows that increasing the similarity weight factor implies to reduce the number of served users,

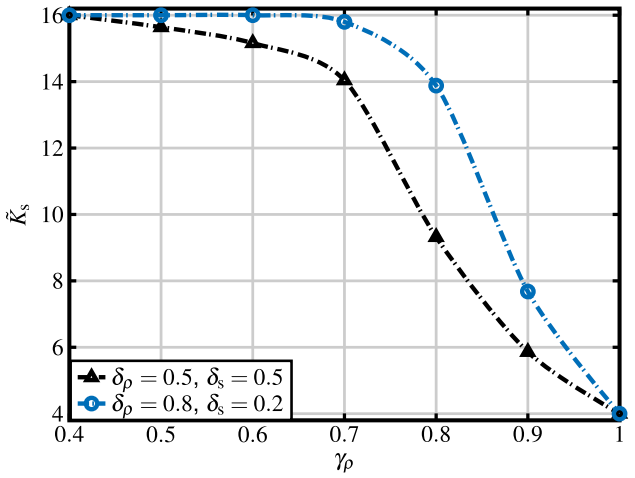

FIGURE 7. Mean served users $\tilde{K}_s$ versus the correlation threshold γ_ρ, $K = 16$, $N_{RF}^r = 4$ and $\rho = 0.9$.

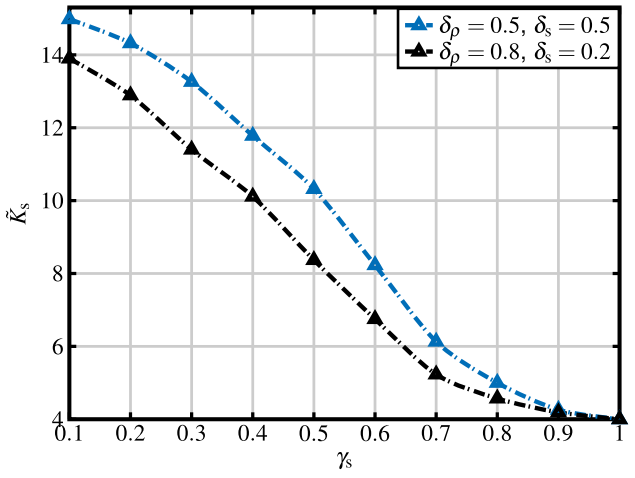

FIGURE 8. Mean served users $\tilde{K}_s$ versus the similarity threshold γ_s for $K = 16$, $N_{RF}^r = 4$ and $\rho = 0.9$.

as observed for the setting $\delta_\rho = 0.5$, $\delta_s = 0.5$. This is because the ordering of the sets of candidate users to be allocated gives more weight to the channel similarity and, therefore, users with lower correlation level can be prioritized, although they will then be disregarded when applying the correlation threshold.

A similar effect is shown in Figure 8 where the mean value of served users, $\tilde{K}_s$, is plotted versus the similarity threshold, γ_s. It can be observed that increasing γ_s reduces K_s. Conversely, by giving less importance to the channel similarity through the weight factor δ_s, the probability of discarding users increases, because users with low channel similarity can be prioritized in the allocation ordering.

D. NOMA CODING SCHEME

We now evaluate the performance of the DQLC-based encoding scheme employed to implement the grouping approach. We specifically compare the performance of the

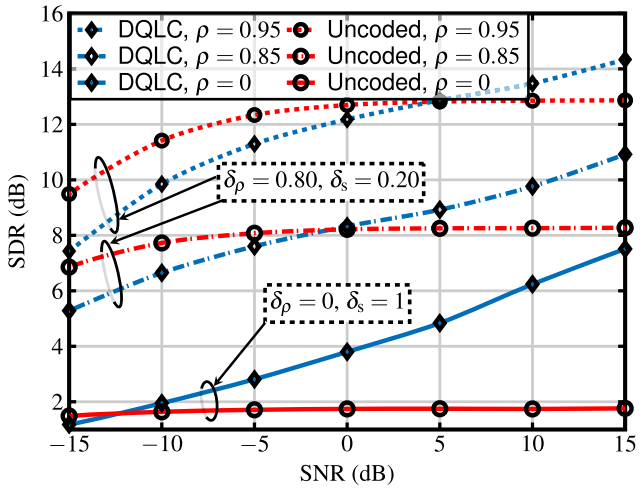

FIGURE 9. SDR (dB) performance for $K = 24$ users, $N_{RF}^r = 8$ and $\rho \in \{0, 0.85, 0.95\}$, $\gamma_\rho = 0$, $\gamma_s = 0$.

grouped system with DQLC to that of a grouped scheme with a conventional NOMA approach based on the uncoded transmission of the source symbols with an appropriate power allocation [35], [36]. The same scheduling procedure is applied in both cases to determine the number of groups and users per group. In order to ensure the same number of served users, K_s, for both approaches, we set $\gamma_\rho = 0$ and $\gamma_s = 0$, i.e., a conservative strategy is assumed where no users are disregarded in the allocation stage. In addition, the parameters δ_ρ and δ_s are equally tuned depending on the correlation factor: $\delta_\rho = 0$ and $\delta_s = 1$ for $\rho = 0$, and $\delta_\rho = 0.8$ and $\delta_s = 0.2$ for $\rho \in \{0.85, 0.95\}$. Finally, the PG algorithm is considered for the hybrid combiner at the BS.

Figure 9 shows the SDR (dB) achieved for the two NOMA strategies considering $K = 16$, $N_{RF}^r = 8$ and three different correlation factors $\rho \in \{0, 0.85, 0.95\}$. As observed, the performance of the uncoded scheme is in general higher for the low SNR regime, but saturates above some SNR value depending on the correlation. In other words, the uncoded scheme is better than the DQLC scheme just for low SNR values in highly correlated scenarios. These results match to the theoretical statements derived in [35] for the non-orthogonal transmission over the MAC and the experimental results obtained in [23], [30] for the same scenario. The key point is that the uncoded transmission provides the best performance below a certain SNR threshold, but is no longer optimal above it. Also, this threshold moves to a lower value as the source correlation decreases. This behaviour is clearly observed in the results shown in Figure 9.

VIII. CONCLUSION

This work addressed the problem of user grouping for the uplink of multiuser hybrid mmWave MIMO. A hybrid analog-digital MMSE precoding/combining approach that exploits the spatial correlation in the mmWave uplink when

$N_{RF}^{r} = K_s$ has been derived. User grouping and allocation have been proposed to support the case $N_{RF}^{r} < K_s$.

The resulting scheduling procedure can be configured in a simple way to perform at different points in the trade-off between the number of served users and symbol distortion. Finally, a new approach for the computation of the hybrid combiner from the fully digital one has been presented.

Results show that the proposed scheduling procedure provides reasonable gains compared to a random allocation policy and, in addition, it closely approaches the optimal combinatorial solution. The hybrid design of the BS combiner offers large gains over the conventional algorithms of matrix factorization. Finally, the results show the feasibility of the grouped system to serve simultaneously a large number of them with a lower number of RF chains.

REFERENCES

[1] T. S. Rappaport, R. W. Heath, Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Upper Saddle River, NJ, USA: Prentice-Hall, Sep. 2014.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[4] R. W. Heath, Jr., N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[5] P. Sudarshan, N. Mehta, A. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, Dec. 2006.

[6] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybridly connected structure for hybrid beamforming in mmWave massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 662–674, Feb. 2018.

[7] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog–digital beamforming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3311–3326, May 2016.

[8] E. Zhang and C. Huang, "On achieving optimal rate of digital precoder by RF-baseband codesign for MIMO systems," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2014, pp. 1–5.

[9] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, Jr., "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[10] S. Sun, T. Rappaport, R. W. Heath, Jr., A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 110–121, Dec. 2014.

[11] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[12] J. P. Gonzalez-Coma, J. Rodriguez-Fernandez, N. Gonzalez-Prelcic, L. Castedo, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for frequency selective multiuser mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 353–367, May 2018.

[13] W. Utschick, C. Stockle, M. Joham, and J. Luo, "Hybrid LISA precoding for multiuser millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 752–765, Feb. 2018.

[14] C. Rusu, R. Mendez-Rial, N. Gonzalez-Prelcic, and R. W. Heath, Jr., "Low complexity hybrid precoding strategies for millimeter wave communication systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8380–8393, Dec. 2016.

[15] N. Li, Z. Wei, H. Yang, X. Zhang, and D. Yang, "Hybrid precoding for mmWave massive MIMO systems with partially connected structure," *IEEE Access*, vol. 5, pp. 15142–15151, 2017.

[16] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[17] J. Singh and S. Ramakrishna, "On the feasibility of codebook-based beamforming in millimeter wave systems with multiple antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2670–2683, May 2015.

[18] J. Zhang, X. Huang, V. Dyadyuk, and Y. Guo, "Massive hybrid antenna array for millimeter-wave cellular communications," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 79–87, Feb. 2015.

[19] X. Zhang, A. F. Molisch, and S.-Y. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4091–4103, Nov. 2005.

[20] J. P. Gonzalez-Coma, W. Utschick, and L. Castedo, "Hybrid LISA for wideband multiuser millimeter-wave communication systems under beam squint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1277–1288, Feb. 2019.

[21] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[22] T. E. Bogale and L. Vandendorpe, "Robust sum MSE optimization for downlink multiuser MIMO systems with arbitrary power constraint: Generalized duality approach," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1862–1875, Apr. 2012.

[23] P. A. Floor, A. N. Kim, T. A. Ramstad, I. Balasingham, N. Wernersson, and M. Skoglund, "On joint source-channel coding for a multivariate Gaussian on a Gaussian MAC," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1824–1836, May 2015.

[24] P. Suarez-Casal, O. Fresnedo, and L. Castedo, "DQLC optimization for joint source channel coding of correlated sources over fading MAC," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1292–1296.

[25] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[26] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[27] L. M. Correia and P. F. M. Smulders, "Characterisation of propagation in 60 GHz radio channels," *Electron. Commun. Eng. J.*, vol. 9, no. 2, pp. 73–80, Apr. 1997.

[28] C. A. Balanis, *Antenna Theory: Analysis and Design*, 3rd ed. Hoboken, NJ, USA: Wiley, 2005.

[29] P. A. Floor, A. N. Kim, T. A. Ramstad, and I. Balasingham, "On transmission of multiple Gaussian sources over a Gaussian MAC using a VQLC mapping," in *Proc. IEEE Inf. Theory Workshop*, Sep. 2012, pp. 50–54.

[30] O. Fresnedo, P. Suarez-Casal, and L. Castedo, "Transmission of spatio-temporal correlated sources over fading multiple access channels with DQLC mappings," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5604–5617, Aug. 2019.

[31] P. A. Floor, A. N. Kim, N. Wernersson, T. A. Ramstad, M. Skoglund, and I. Balasingham, "Zero-delay joint source-channel coding for a bivariate Gaussian on a Gaussian MAC," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 3091–3102, Oct. 2012.

[32] P. Suarez-Casal, J. P. Gonzalez-Coma, O. Fresnedo, and L. Castedo, "Design of linear precoders for correlated sources in MIMO multiple access channels," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6110–6122, Dec. 2018.

[33] K.-H. Lee and D. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. 24, no. 12, pp. 1283–1290, Dec. 1976.

[34] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, Jr., "Hybrid MMSE precoding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, 2017.

[35] A. Lapidoth and S. Tinguely, "Sending a bivariate Gaussian source over a Gaussian MAC with feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1852–1864, Apr. 2010.

[36] M. Gastpar, "Uncoded transmission is exactly optimal for a simple Gaussian 'sensor' network," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5247–5251, Nov. 2008.

DARIAN PÉREZ-ADÁN received the B.S. degree in telecommunications and electronics engineering from the Technological University of Havana José Antonio Echeverría (CUJAE), Cuba, in 2017. He is currently pursuing the Ph.D. degree with the University of A Coruña (UDC), Spain, where he has been with the Group of Electronic Technology and Communications, since 2018. His research interests include signal processing for millimeter-wave and multiuser communications.

ÓSCAR FRESNEDO (Member, IEEE) received the degree in computer engineering and the Ph.D. degree in computer engineering from the University of A Coruña, Spain, in 2007 and 2014, respectively. Since 2007, he has been with the Group of Electronic Technology and Communications (GTEC), Department of Electronics and Systems, University of A Coruña, where he had the benefit of a FPI scholarship granted by the Spanish Government, from 2008 to 2012. He has published 13 articles in international technical journals as well as more than 30 articles in relevant international conferences and workshops in the area of communications and signal processing. His main research interests are in the design of coding schemes, analog joint source-channel coding, multiuser communications, and image processing. He has participated as a Research member in more than ten research projects and contracts granted by regional, national and European administrations. He has also received the Best Student Paper Award at the 14th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Darmstadt, in 2013.

JOSÉ P. GONZÁLEZ-COMA (Member, IEEE) was born in Marín, Spain. He received the degree in computer engineering and the Ph.D. degree from the University of A Coruña, Spain, in 2009 and 2015, respectively. He was appointed as a Visiting Researcher at the Associate Institute for Signal Processing, Technische Universität München, Germany, in 2012, and at the Signal Processing in Communications Group (UVIGO), Spain, in 2017. Since 2017, he has been with the Research Center on Information and Communication Technologies, University of A Coruña, where he is currently a Postdoctoral Researcher. He received the FPI Grant from the Ministerio de Ciencia e Innovación. His main research interests are in channel estimation and precoding in massive multi-in multi-out systems, and millimeter-wave communications.

LUIS CASTEDO (Senior Member, IEEE) received the Ph.D. degree in telecommunications engineering from the Technical University of Madrid, Spain, in 1993. Since 1994, he has been a Faculty Member with the Department of Computer Engineering, University of A Coruña (UDC), Spain, where he became a Professor, in 2001, and acted as the Chairman, from 2003 to 2009. He had previously held several research appointments at the University of Southern California (USC) and École supérieure d'électricité (SUPELEC). From 2014 to 2018, he was a Manager of the Communications and Electronic Technologies (TEC) program in the State Research Agency, Spain. He has also been the Principal Researcher of more than 50 research projects funded by public organisms and private companies. He has coauthored more than 300 articles in peer-reviewed international journals and conferences. His research interests are signal processing for wireless communications and prototyping of digital communication equipment. His articles have received three Best Student Paper Awards at the IEEE/ITG Workshop on Smart Antennas, in 2007, at the IEEE International Workshop on Signal Processing Advances in Wireless Communications, in 2013, and at the IEEE International Conference on Internet of Things (iThings), in 2017. He has been the General Co-Chair of the 8th IEEE Sensor Array and Multichannel Signal Processing Workshop, in 2014, and the 27th European Signal Processing Conference, in 2019.