Uplink Pilots for Multiuser MIMO with Mixed Grant Free and Grant Based Transmissions

Nassar Ksairi, Merouane Debbah

Uplink Pilots for Multiuser MIMO with Mixed Grant Free and Grant Based Transmissions

Nassar Ksairi and Mérouane Debbah
Mathematical and Algorithmic Sciences Lab, Paris Research Center,
Huawei Technologies France SASU, 92100 Boulogne-Billancourt, France.
Emails: {nassar.ksairi, merouane.debbah}@huawei.com

***Abstract*—Pilot based acquisition of channel state information (CSI) is a challenging problem in multiuser multiple-input multiple-output (MU-MIMO) systems that allow unscheduled intermittent transmissions from user terminals (UTs). These challenges stem mainly from the need to account for i) possible intra-cell pilot collisions or interference which are characteristic of unscheduled uplink transmissions and ii) the fact that these same pilots are typically also needed for user activity detection at the base station (BS). The problem gets even more challenging in situations with both scheduled (grant based) and unscheduled (grant free) concurrent uplink transmissions because of the fact that grant free and grant based transmissions are typically different in terms of key performance indicators (KPIs). In this article, we present an assignment and multiplexing scheme for uplink pilot sequences from grant free and grant based transmissions that take place on the same time-frequency resources and we numerically assess its performance in terms of both CSI quality and user activity detection probability. The scheme is based on a novel method that can extend any set of state-of-the-art orthogonal pilots to generate a hybrid (orthogonal/non-orthogonal) pilot set with a built-in trade-off between the maximum achievable sum rate for grant based transmissions and the maximum tolerable level of pilot interference for grant free transmissions.**

I. Introduction

In current cellular systems, uplink data is transmitted in a scheduled (grant based) manner i.e., uplink data transmission can only take place after the time-frequency resources needed for it have already been reserved by the network and signaled to the emitting device following the reception of *scheduling request* from that device[1]. With the emergence of wireless applications requiring the deployment of massive machine type communications (MTC) and ultra-reliable low-latency communications (URLLC) it becomes necessary for cellular systems to provide support for unscheduled (grant free) transmissions. Indeed, grant free transmission of data packets with small payloads is known to be more efficient in terms of throughput than its grant based counterpart [1]. Moreover, grant free transmission could be crucial for meeting the latency requirements of URLLC applications in order to avoid the extra delay needed for resource reservation [2].

Whether the transmission is grant based or grant free, uplink pilots need to be sent by the active UTs along with their uplink data. The CSI obtained from these pilots is needed for coherent detection of the transmitted symbols. Typically, active UTs that are scheduled to transmit on the same time-frequency resources are assigned exclusive mutually orthogonal pilot sequences so that the channels of their respective links to the BS can be estimated without interference from other co-scheduled UTs. Since pilot assignment can be done in a persistent or semi-persistent manner that is independent of data grants, this exclusive orthogonal pilot assignment can still be used for unscheduled grant free transmissions. However, due to the sporadic nature of uplink data traffic in most scenarios that require grant free transmissions, exclusive assignment of orthogonal pilots to all grant free active UTs turns out to be very inefficient with respect to (w.r.t.) resource utilization. To remedy this inefficiency, it was proposed in [3], [4] to get rid of exclusivity when assigning pilots to grant free UTs and replace it with pilot hopping patterns that are user specific and which define predetermined sequences of orthogonal pilots to send during the different transmission slots of a multi-slot frame. This way, if a pilot collision occurs in one or several slots, user identification is still possible at the end of the frame reception. Moreover, the high channel estimation error that happens as a result of these collisions is averaged out if the number of transmission slots per frame is large enough. While this solution is compatible with MTC, it cannot be applied in most delay sensitive applications which are typically characterized by a transmission that occupies a single slot. A pilot random access solution that is more fit for one-slot grant free transmissions was proposed in [5]. This solution consists in assigning exclusive but non-orthogonal pilot sequences to the different UTs. Indeed, non-orthogonality makes it possible to assign distinct pilots to a large number of active UTs while keeping the number of symbols per sequence much smaller than the number of pilots. Of course, the use of non-orthogonal pilots will entail intra-cell pilot contamination and hence multiuser interference that will affect user activity detection and channel estimation at the BS. The potential performance degradation due to this interference can be alleviated by resorting to compressive sensing methods [6] and joint channel estimation [7].

In all the above-mentioned works, the cases where grant free and grant based transmissions can take place on the same time-frequency resources are not addressed. This co-existence is crucial from a resource utilization efficiency perspective in scenarios where the total volume of grant free traffic is not large enough to justify being reserved an exclusive portion

[1]Scheduling requests can only be sent by *active* devices that have already been registered as such by the network via a *random access* procedure.

of the available time-frequency resources. Note that in such cases the way the (limited) set of available uplink pilots is divided between the two *grant classes* i.e., the grant free and the grant based groups of UTs, will have a huge effect on the throughput performance of grant based transmissions and the reliability performance of grant free transmissions. For instance, reserving more pilots for grant free transmissions to reduce pilot collision probability will reduce the maximum number of MIMO transmission layers available for grant based transmissions and hence their achievable throughput.

In this paper, we propose a hybrid multiplexing scheme that mixes orthogonality (between the pilots of two different grant classes) and non-orthogonality (within the same class) to achieve the sought trade-off between grant based throughput performance and grant free reliability performance more efficiently than a simple orthogonal or non-orthogonal multiplexing scheme. Indeed, the proposed scheme allows different degrees of non-orthogonality i.e., of pilot resource overloading, within the two grant classes, while at the same time protecting the pilot signals of each class from the interference generated by pilot sequence non-orthogonality of the other class. For instance, this property can be exploited to achieve the same value of grant free user activity detection probability and the same level of grant free transmission reliability as those achieved with an orthogonal pilot assignment scheme while reserving a much larger portion of the available pilot resources for grant based transmissions, thus granting them more spatial transmission layers than the orthogonal scheme.

Notations: The $N \times N$ identity matrix is denoted by $\mathbf{I}_N$ while notation $\mathbf{1}_{N\times M}$ (respectively $\mathbf{0}_{N\times M}$) stands for the $N \times M$ matrix with all its entries set to one (respectively to zero). Notation $\mathrm{diag}(\mathbf{x})$, for some $N \times 1$ vector $\mathbf{x}$, is used to designate the $N\times N$ diagonal matrix whose diagonal elements are the components of $\mathbf{x}$. Reciprocally, $\mathrm{diag}\,(\mathbf{M})$, for some $N \times N$ matrix $\mathbf{M}$, designates the $N \times 1$ vector composed of the diagonal elements of $\mathbf{M}$. Finally, $\mathrm{diag}\,(\mathbf{M}_1, \ldots, \mathbf{M}_L)$, for some $N \times N$ matrices $\{\mathbf{M}_l\}_{l=1\cdots L}$, stands for the $LN \times LN$ block-diagonal matrix with $\{\mathbf{M}_l\}_{l=1\cdots L}$ as its diagonal blocks.

II. System Model

Consider *uplink* transmission in a wireless system consisting of one BS equipped with M antennas (indexed using $m \in \{1, \ldots, M\}$) serving single-antenna user terminals (UTs) that have data to transmit. Signal transmission is done using orthogonal frequency division multiplexing (OFDM) with N_{FFT} subcarriers, N_{CP}-long cyclic prefix and a total bandwidth of W Hz. Assume that the OFDM resource grid is structured into resource blocks (RBs) each containing T OFDM symbols. In the sequel, we focus on transmissions that are taking place on one of the N_{RB} RBs on which co-existence between grant free and grant based transmissions is allowed (see Figure 1). Denote by $\mathcal{K}_1$ and $\mathcal{K}_2$ the sets of UTs transmitting on the considered RB respectively with and without grant. Define $K_1 \overset{\text{def.}}{=} |\mathcal{K}_1|$ and $K_2 \overset{\text{def.}}{=} |\mathcal{K}_2|$ and note that while K_1 is deterministic, K_2 is random since the number of grant free transmissions is not a priori known at the BS. Nonetheless,

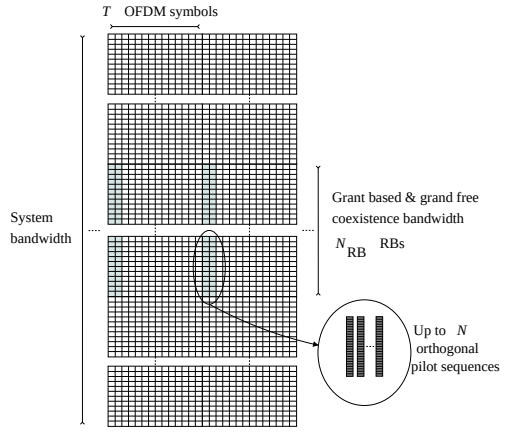

Fig. 1. Grant free and grant based coexistence region in the resource grid

it is reasonable to assume that K_2 is upper-bounded by a deterministic value $K_2^{\max}$ that is known to both the BS and the UTs. This knowledge can be the outcome of the initial access and connection setup procedures during which the BS can signal the index of one or several RBs on which each UT is allowed to transmit[2]. Denote by $\mathcal{K}_2^{\max}$ the set of indexes of all the grant free UTs allowed to transmit on the RB under investigation and note that $\mathcal{K}_2 \subset \mathcal{K}_2^{\max}$ and $|\mathcal{K}_2^{\max}| = K_2^{\max}$. Typically, the relevant performance measure for the UTs in $\mathcal{K}_1$ is their achievable sum throughput while the UTs in $\mathcal{K}_2$ have a minimum acceptable reliability constraint e.g., block error rate (BLER) and miss-detection/false-alarm rate values that should not be exceeded.

The wireless link from UT k to BS antenna m is a multipath channel with $L \geq 1$ dominant paths that remain constant within the duration of an OFDM symbol. Let $h_{k,m,t}(.)$ designate the continuous-time impulse response of this channel during the t-th OFDM symbol ($t \in \{1, \ldots, T\}$) and write

$$h_{k,m,t}(\tau) = \sum_{l=0}^{L-1} \alpha_{k,m,l,t}\delta(\tau - \tau_l),\ \forall \tau \in \mathbb{R}, \tag{1}$$

where τ_l is the delay of the l-th path and $\alpha_{k,m,l,t}$ is its complex-valued gain. Path gains $\{\alpha_{k,m,l,t}\}_{l\in\{0\cdots L-1\}}$ are modeled, as in [8], as mutually independent zero-mean random variables with variances $\sigma_l^2 \overset{\text{def.}}{=} \mathbb{E}\left[|\alpha_{k,m,l,t}|^2\right]$ that satisfy $\sum_{l=0}^{L-1}\sigma_l^2 = 1$, $\mathbb{E}\left[\alpha_{k,m,l,t}\alpha_{j,m,l,s}^*\right] = 0$ for all $j \neq k$ and

$$\begin{aligned}\mathbb{E}\left[\alpha_{k,m,l,t}\alpha_{k,m,l,s}^*\right] &= \sigma_l^2 J_0\left(2\pi f_k^D\left(N_{\mathrm{FFT}} + N_{\mathrm{CP}}\right)T_s(t-s)\right)\\ &\overset{\text{def.}}{=} \sigma_l^2 r_k^{\alpha}(t-s),\end{aligned} \tag{2}$$

where f_k^D is the maximum Doppler frequency shift and where $T_s \overset{\text{def.}}{=} 1/W$. Now, define $\mathbf{h}_{k,l,t} \overset{\text{def.}}{=} [h_{k,1,l,t} \cdots h_{k,M,l,t}]^{\mathrm{T}}$, $\boldsymbol{\alpha}_{k,l,t} \overset{\text{def.}}{=} [\alpha_{k,1,l,t} \cdots \alpha_{k,M,l,t}]^{\mathrm{T}}$ and assume $\boldsymbol{\alpha}_{k,l,t} \sim$

[2]Note that the BS needs to adapt the value of $K_2^{\max}$ to the current load of traffic from grant free UTs. For instance, $K_2^{\max}$ should be increased in order to avoid denials of service in scenarios with relatively high densities of UTs that need to transmit in a grant free manner.

$\mathcal{CN}\left(\mathbf{0}, \mathbf{R}_{k,l}^{\alpha}\right)$. Finally, assume as in [9] and [10] that the correlation between the two vectors $\boldsymbol{\alpha}_{k,l,t_1}$ and $\boldsymbol{\alpha}_{k,l,t_2}$ whenever $t_1 \neq t_2$ is separable into time domain and space domain multiplicative terms. More precisely, $\mathbb{E}\left[\boldsymbol{\alpha}_{k,l,t_1}\boldsymbol{\alpha}_{k,l,t_2}^{\mathrm{H}}\right] = r_k^{\alpha}(t_1 - t_2)\mathbf{R}_{k,l}^{\alpha}$. From (1), the frequency domain channel coefficient between UT k and the m-th BS antenna at subcarrier n during OFDM symbol t, denoted as $H_{k,m,t}$, is given by

$$H_{k,m,t,n} = \sum_{l=0}^{L-1} \alpha_{k,m,l,t} e^{-2\pi\imath \frac{\tau_l n}{N_{\mathrm{FFT}} T_s}}. \tag{3}$$

Now, define $\mathbf{H}_{k,t,n} \overset{\text{def.}}{=} [H_{k,1,t,n} \cdots H_{k,M,t,n}]^{\mathrm{T}}$ and denote by $x_{k,t,n}$ the sample transmitted on the n-th subcarrier during the t-th OFDM symbol which satisfies

$$\mathbb{E}\left[|x_{k,t,n}|^2\right] = P_k. \tag{4}$$

The corresponding $M \times 1$ vector of samples received at the BS array then writes as

$$\mathbf{y}_{t,n} = \sum_{k\in\mathcal{K}_1} \sqrt{g_k}\mathbf{H}_{k,t,n}x_{k,t,n} + \sum_{k\in\mathcal{K}_2} \sqrt{g_k}\mathbf{H}_{k,t,n}x_{k,t,n} + \mathbf{z}_{t,n}, \tag{5}$$

where $\mathbf{z}_{t,n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_M)$ is a vector of noise samples and g_k is the large-scale fading factor. Assume that coherent detection e.g., maximum-ratio combining (MRC), is performed at the BS using pilot based channel estimates of vectors $\mathbf{H}_{k,t,n}$. For that sake, $N_p \geq 1$ resource elements (REs) within each RB are reserved for pilot transmission. Denote by $\{(t_i, n_i)\}_{i=1\cdots N_p}$ the set of positions of these REs and let $\mathbf{p}_k \overset{\text{def.}}{=} \begin{bmatrix} x_{k,t_1,n_1} & \cdots & x_{k,t_{N_p},n_{N_p}} \end{bmatrix}^{\mathrm{T}}$ designate the pilot sequence transmitted by UT k. These sequences are also used at the BS to determine which subset $\mathcal{K}_2 \subset \mathcal{K}_2^{\max}$ is currently active on the considered RB. The interference terms in (5) hint at a possible trade-off between the achievable performances in each grant class. For instance, getting rid of inter class interference during channel estimation requires reserving exclusive portions of the pilot time, frequency and code resources for each grant class i.e., sequences assigned to UTs from different classes should be orthogonal. However, since pilot resources are scarce, reserving a larger portion of them to one grant class for the sake of better performance results in a smaller portion for the other. In the next section, we propose a hybrid scheme to generate $\mathbf{p}_k$ that can achieve any required level of trade-off between the two classes more efficiently than both orthogonal and non-orthogonal pilot assignment methods.

III. Grant Class Aware Pilot Design

Consider a set $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ of N orthogonal pilot sequences each of length N_p and normalized such that

$$\frac{1}{N_p}\mathbb{E}\left[\|\mathbf{u}_u\|^2\right] = P_k, \quad \forall u \in \{1, \ldots, N\}. \tag{6}$$

Such a set can be composed using N orthogonal sequences obtained with a combination of time division multiplexing (TDM), frequency division multiplexing (FDM) and code division multiplexing (CDM) or using N phase-shifted versions of the same Zadoff-Chu (ZC) sequence (as in LTE [11]). Next, split this set into two disjoint subsets, namely $\mathcal{U}_1$ and $\mathcal{U}_2$, the first containing $N_1 < N$ sequences while the second is composed of $N_2 = N - N_1$ sequences.

In the proposed scheme, subset $\mathcal{U}_1$ is reserved for grant based UTs while $\mathcal{U}_2$ is assigned to the grant free class. However, the sequences of the latter subset are not directly used by the grant free UTs as their pilot sequences. Instead, the actual pilot sequence transmitted by each one of these UTs is a linear combinations of *all* the sequences of $\mathcal{U}_2$. More precisely, let $j_k \in \{1, \ldots, N\}$ be the index of the pilot sequence from $\mathcal{U}_1$ assigned to UT $k \in \mathcal{K}_1$ and $u_1, \ldots, u_{N_2}$ be the indexes of the sequences making up the subset $\mathcal{U}_2$. Then we can write

$$\mathbf{p}_k = \mathbf{u}_{j_k} \in \mathcal{U}_1, \qquad \forall k \in \mathcal{K}_1, \tag{7}$$

$$\mathbf{p}_k = w_{k,1}\mathbf{u}_{u_1} + \ldots + w_{k,N_2}\mathbf{u}_{u_{N_2}}, \qquad \forall k \in \mathcal{K}_2. \tag{8}$$

Here, $\mathbf{w}_k \overset{\text{def.}}{=} \begin{bmatrix} w_{k,1} & \cdots & w_{k,N_2} \end{bmatrix}^{\mathrm{T}}$ is a UT specific signature that can be signaled to the UT during the connection setup procedure. Of course, the resulting pilot sequences $\mathbf{p}_k$ and $\mathbf{p}_j$ for any $k, j \in \mathcal{K}_2^{\max}$ are mutually non-orthogonal. However, thanks to the orthogonality of the original baseline sequences, this non-orthogonality only results in intra-class pilot contamination. Moreover, the level of this intra-class pilot contamination as well as the proportions of pilot resources assigned to each grant class are all configurable parameters that can in principle be optimized to yield the best trade-off between the relevant performance metrics of the UTs of each grant class. Note that this novel scheme was originally proposed in previous works of ours e.g. [12], as a solution for multiplexing *data* transmissions originating from multiple service classes.

Remark 1. *In cellular systems, $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ are not arbitrary sequences with the only requirement to be mutually orthogonal. They should be both orthogonal within one cell and with low cross-correlation to the pilot sequences in neighboring cells. One important byproduct of the proposed pilot design is that $\mathbf{p}_k$ inherit the multicell interference mitigation properties of the baseline sequences since only these sequences appear, from a network perspective, to be transmitted.*

One way to guarantee that the transmit power constraint in (4) is respected is by setting

$$\|\mathbf{w}_k\|^2 = 1, \ \forall k \in \{1, \ldots, K_2^{\max}\}. \tag{9}$$

Indeed, if (6) and (9) are respected, then the power constraint in (4) is satisfied by all the components of the pilot sequence $\mathbf{p}_k$. Now, let $\left\{k_1, \ldots, k_{K_2^{\max}}\right\} = \mathcal{K}_2^{\max}$ and define

$$\mathbf{W} \overset{\text{def.}}{=} \begin{bmatrix} \mathbf{w}_{k_1} & \cdots & \mathbf{w}_{k_{K_2^{\max}}} \end{bmatrix}. \tag{10}$$

One practical way to obtain matrices $\mathbf{W}$ that satisfy (9) is to use *pseudo-noise* (PN) generators [11] with different initializations to generate the different columns of $\mathbf{W}$. However, since PN sequences are binary, the number of different N_2-long sequences we can get will be upper bounded by 2^{N_2}. A more systematic method is to select $\{\mathbf{w}_k\}_{k=1\cdots K_2^{\max}}$ as a

collection of $K_2^{\max}$ points on the surface of a unit-radius N_2-dimensional complex-valued sphere with a minimal pairwise angle that is as large as possible. This problem, which is shown to be relevant from a performance enhancing perspective in Section IV, can be solved in advance for different configurations of $(N_2, K_2^{\max})$ using for instance the methodology of [13] and the outcome of this offline optimization can be stored in look-up tables.

Remark 2. *The larger the value of N_2, the smaller the largest cross-correlation among pairs of sequences from the set $\{\mathbf{w}_k\}_{k\in\mathcal{K}_2^{\max}}$ (and by extension among pairs of sequences from the set $\{\mathbf{p}_k\}_{k\in\mathcal{K}_2^{\max}}$). There is thus a trade-off between the level of residual pilot contamination within the grant free class and the number $N_1 = N - N_2$ of spatial layers available for transmissions from the grant based class. The quantification of this trade-off is left for future works.*

IV. User Activity Detection and Channel Estimation Using Grant Class Aware Pilots

Thanks to the mutual orthogonality of sequences $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$, channel estimation for any UT $k \in \mathcal{K}_1$ can be done using any state-of-the-art method independently of grant free transmissions. As for UTs $k \in \mathcal{K}_2$, they first need to be identified. For that sake, let $\mathbb{1}_{\mathcal{A}}(.)$ denote the indicator function of set $\mathcal{A}$ and define the vector of concatenated channel vectors from all the grant free UTs that are allowed to transmit on the considered RB as

$$\mathbf{H} \overset{\text{def.}}{=} \Big[\mathbb{1}_{\mathcal{K}_2}(k_1)\sqrt{g_{k_1}P_{k_1}}\left(\mathbf{H}_{k_1}\right)^{\mathrm{T}} \quad \cdots \quad \mathbb{1}_{\mathcal{K}_2}(k_{K_2^{\max}})\sqrt{g_{k_{K_2^{\max}}}P_{k_{K_2^{\max}}}}\left(\mathbf{H}_{k_{K_2^{\max}}}\right)^{\mathrm{T}}\Big]^{\mathrm{T}} \tag{11}$$

where for any $k \in \mathcal{K}_2^{\max}$

$$\mathbf{H}_k \overset{\text{def.}}{=} \begin{bmatrix}\mathbf{H}_{k,t_1,n_1}^{\mathrm{T}} & \cdots & \mathbf{H}_{k,t_{N_p},n_{N_p}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}. \tag{12}$$

Also, let $\{j_1, \ldots, j_{K_1}\} = \mathcal{K}_1$ and denote by $\mathbf{G}$ the vector of concatenated channel vectors from all the UTs in $\mathcal{K}_1$. The N_pM samples received at the BS on the positions of the pilot symbols can be written by referring to (5) and (11) as

$$\mathbf{Y} = \mathbf{P}\mathbf{H} + \mathbf{Q}\mathbf{G} + \mathbf{Z}, \tag{13}$$

where $\mathbf{Z} \sim \mathcal{CN}\left(\mathbf{0}_{1\times N_pM}, \sigma^2\mathbf{I}_{N_pM}\right)$ and

$$\begin{aligned}\mathbf{P} &\overset{\text{def.}}{=} \left[\mathrm{diag}\left(\mathbf{p}_{k_1}\otimes\mathbf{1}_{1\times M}\right)\cdots\mathrm{diag}\left(\mathbf{p}_{k_{K_2^{\max}}}\otimes\mathbf{1}_{1\times M}\right)\right],\\ \mathbf{Q} &\overset{\text{def.}}{=} \left[\mathrm{diag}\left(\mathbf{p}_{j_1}\otimes\mathbf{1}_{1\times M}\right)\cdots\mathrm{diag}\left(\mathbf{p}_{j_{K_1}}\otimes\mathbf{1}_{1\times M}\right)\right].\end{aligned} \tag{14}$$

We propose the following heuristic for the identification of $\mathcal{K}_2$ out of $\mathcal{K}_2^{\max}$ and for the estimation of the corresponding channel vectors. Start by assuming that all the UTs in $\mathcal{K}_2^{\max}$ are active. In this case, a linear minimum mean-square error (LMMSE) channel estimation which assumes that $\mathcal{K}_2 = \mathcal{K}_2^{\max}$ can be used to get $\hat{\mathbf{H}}$, the estimate of the concatenated channels vector $\mathbf{H}$, based on $\mathbf{Y}$ as

$$\begin{aligned}\hat{\mathbf{H}} &\overset{\text{def.}}{=} \mathbb{E}\left[\mathbf{H}\mathbf{Y}^{\mathrm{H}}|\mathcal{K}_2 = \mathcal{K}_2^{\max}\right]\left(\mathbb{E}\left[\mathbf{Y}\mathbf{Y}^{\mathrm{H}}|\mathcal{K}_2 = \mathcal{K}_2^{\max}\right]\right)^{-1}\mathbf{Y}\\ &= \mathbf{R}_{\mathbf{H}}^{\max}\mathbf{P}^{\mathrm{H}}\left(\mathbf{P}\mathbf{R}_{\mathbf{H}}^{\max}\mathbf{P}^{\mathrm{H}} + \mathbf{Q}\mathbf{R}_{\mathbf{G}}\mathbf{Q}^{\mathrm{H}} + \sigma^2\mathbf{I}_{NpM}\right)^{-1}\mathbf{Y},\end{aligned} \tag{15}$$

where the expectation in (15) is w.r.t. the joint distribution of $\mathbf{H}$ and $\mathbf{Y}$ conditioned on $\mathcal{K}_2 = \mathcal{K}_2^{\max}$, where $\mathbf{R}_{\mathbf{G}} \overset{\text{def.}}{=} \mathbb{E}\left[\mathbf{G}\mathbf{G}^{\mathrm{H}}\right]$, and where the second line follows because $\mathbb{E}\left[\mathbf{H}\mathbf{G}^{\mathrm{H}}\right] = \mathbb{E}\left[\mathbf{G}\mathbf{H}^{\mathrm{H}}\right] = \mathbf{0}$ due to the assumptions made in Section II. Here, we defined $\mathbf{R}_{\mathbf{H}}^{\max} \overset{\text{def.}}{=} \mathbb{E}\left[\mathbf{H}\mathbf{H}^{\mathrm{H}}|\mathcal{K}_2 = \mathcal{K}_2^{\max}\right]$ so that

$$\mathbf{R}_{\mathbf{H}}^{\max} = \mathrm{diag}\left(\mathbf{R}_{k_1}, \ldots, \mathbf{R}_{k_{K_2^{\max}}}\right), \tag{16}$$

with

$$\mathbf{R}_k \overset{\text{def.}}{=} g_kP_k\left[\sum_{l=0}^{L-1} r_k^{\alpha}(t_i - t_j)\mathbf{R}_{k,l}^{\alpha}e^{-2\pi\imath\frac{\tau_l(n_i-n_j)}{N_{\mathrm{FFT}}T_s}}\right]_{1\le i,j\le N_p}. \tag{17}$$

Of course, not all the UTs are necessarily simultaneously active on the same RB i.e., in general $\mathcal{K}_2 \neq \mathcal{K}_2^{\max}$. Moreover, the components of $\hat{\mathbf{H}}$ corresponding to inactive UTs will not have zero values. This is due both to additive noise and to pilot signal interference from active UTs caused by the (by design) non-orthogonality of the pilot sequences assigned to $\mathcal{K}_2^{\max}$. However, proper setting of the relevant system and pilot scheme parameters, namely $K_2^{\max}, N, N_2$, and M, and of the matrix $\mathbf{W}$ can guarantee, roughly speaking, that the components of $\hat{\mathbf{H}}$ corresponding to inactive UTs are much smaller in absolute value than their active UTs counterparts and that the components of $\hat{\mathbf{H}}$ corresponding to active UTs are close enough to their actual value. More precisely, define the error vector $\tilde{\mathbf{H}} \overset{\text{def.}}{=} \mathbf{H} - \hat{\mathbf{H}}$,. Then it is straightforward to show by referring to (15) that $\mathbf{R}_{\tilde{\mathbf{H}}} \overset{\text{def.}}{=} \mathbb{E}\left[\tilde{\mathbf{H}}\tilde{\mathbf{H}}^{\mathrm{T}}|\mathcal{K}_2\right]$ writes as

$$\begin{aligned}\mathbf{R}_{\tilde{\mathbf{H}}} = \mathbf{R}_{\mathbf{H}} - \mathbf{R}_{\mathbf{H}}^{\max}\mathbf{P}^{\mathrm{H}}&\left(\mathbf{P}\mathbf{R}_{\mathbf{H}}^{\max}\mathbf{P}^{\mathrm{H}} + \mathbf{Q}\mathbf{R}_{\mathbf{G}}\mathbf{Q}^{\mathrm{H}} + \sigma^2\mathbf{I}_{NpM}\right)^{-1}\\ &\times\left(\mathbf{P}\mathbf{R}_{\mathbf{H}}\mathbf{P}^{\mathrm{H}} + \mathbf{Q}\mathbf{R}_{\mathbf{G}}\mathbf{Q}^{\mathrm{H}} + \sigma^2\mathbf{I}_{NpM}\right)\\ &\times\left(\mathbf{P}\mathbf{R}_{\mathbf{H}}^{\max}\mathbf{P}^{\mathrm{H}} + \mathbf{Q}\mathbf{R}_{\mathbf{G}}\mathbf{Q}^{\mathrm{H}} + \sigma^2\mathbf{I}_{NpM}\right)^{-1}\mathbf{P}\mathbf{R}_{\mathbf{H}}^{\max}.\end{aligned} \tag{18}$$

Here, we defined $\mathbf{R}_{\mathbf{H}} \overset{\text{def.}}{=} \mathbb{E}\left[\mathbf{H}\mathbf{H}^{\mathrm{H}}|\mathcal{K}_2\right]$ so that

$$\mathbf{R}_{\mathbf{H}} = \mathrm{diag}\left(\mathbb{1}_{\mathcal{K}_2}(k_1)\mathbf{R}_{k_1}, \ldots, \mathbb{1}_{\mathcal{K}_2}(k_{K_2^{\max}})\mathbf{R}_{k_{K_2^{\max}}}\right). \tag{19}$$

One way to guarantee good detection and channel estimation performances is to choose $K_2^{\max}, N, N_2, M$ and $\mathbf{W}$ such that[3]

$$\mathbb{E}\left[\mathbf{R}_{\tilde{\mathbf{H}}}\right] \preceq \epsilon^{\text{thr}}\mathrm{diag}\left(g_{k_1}P_{k_1}\mathbf{I}_{N_pM}, \ldots, g_{k_{K_2^{\max}}}P_{k_{K_2^{\max}}}\mathbf{I}_{N_pM}\right) \tag{20}$$

for some predefined performance threshold ϵ^{thr}. The expectation in (20) is w.r.t. the distribution of the random variables $\{\mathbb{1}_{\mathcal{K}_2}(k_j)\}_{j=1\cdots K_2^{\max}}$. If the condition in (20) holds, it allows us to perform activity detection of a UT k_j ($j \in$

[3]Notation $\mathbf{A} \preceq \mathbf{B}$, for two positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$, stands for the property that $\mathbf{B} - \mathbf{A}$ is positive semi-definite.

$\{1,\ldots,K_2^{\max}\}$) by means of comparing the norm of its estimated channel vector i.e., $\left\|\left[\hat{\mathbf{H}}\right]_j\right\|^2$, to $N_p M g_{k_j} P_{k_j} \epsilon^{\text{thr}}$. In practice, the value of ϵ^{thr} should be chosen based on the tolerated level of *false alarm* i.e., the probability of identifying a UT $k \in \mathcal{K}_2^{\max}$ as belonging to $\mathcal{K}_2$ while k is in reality currently inactive. One possible criterion is to set ϵ^{thr} such that correct user detection probability is maximized while the false-alarm probability is smaller than a given value. Finding an exact solution to this problem remains a difficult task that is left for future work. In the absence of an exact solution, the following conservative[4] condition can replace (20)

$$\max_{k\in\mathcal{K}_2^{\max}} \sum_{j\in\mathcal{K}_2^{\max}\setminus\{k\}} \mathbb{E}\left[\mathbb{1}_{\mathcal{K}_2}(j)\right] \frac{g_j P_j}{g_k P_k} \left|\mathbf{w}_k^{\mathrm{H}}\mathbf{w}_j\right|^2 + \frac{\sigma^2}{g_k P_k N_p} < \epsilon^{\text{thr}}. \quad (21)$$

Indeed, when $\mathbf{H}_{k,t_i,n_i}$ is constant w.r.t. (t_i, n_i) for all $i \in \{1,\ldots,N_p\}$, then the argument of the maximum in the left-hand side of (21) is the variance of the estimation error associated with any component of $\mathbf{H}_{k,t_i,n_i}$ when *single-user* least-squares channel estimation i.e., de-correlating the received pilot signal on each antenna element with $\mathbf{p}_k$, is used instead of joint LMMSE. Here, we used the fact that $\mathbf{p}_k^{\mathrm{H}}\mathbf{p}_j = 0$ for all $(k,j) \in \mathcal{K}_2^{\max} \times \mathcal{K}_1$ and that $\mathbf{p}_k^{\mathrm{H}}\mathbf{p}_j = \mathbf{w}_k^{\mathrm{H}}\mathbf{w}_j$ for all $(k,j) \in \left(\mathcal{K}_2^{\max}\right)^2$ due to (8) and to the orthogonality of sequences $\{\mathbf{u}_1,\ldots,\mathbf{u}_N\}$. Note that the condition in (21) justifies the sphere surface packing method presented in Section III. The proposed detection and channel estimation method is summarized by Algorithm 1. Once $\hat{\mathbf{H}}_k$ is obtained,

Algorithm 1 User activity detection and channel estimation for grant free UTs based on grant class aware pilots

Initialization: $\mathcal{K}_2 \leftarrow \emptyset$
Compute $\hat{\mathbf{H}}$ using (15)
for $1 \le j \le K_2^{\max}$ **do**
  $\left[\hat{\mathbf{H}}\right]_j \leftarrow \left[\left[\hat{\mathbf{H}}\right]_{(j-1)N_pM+1} \cdots \left[\hat{\mathbf{H}}\right]_{jN_pM}\right]^{\mathrm{T}}$
  if $\frac{1}{g_{k_j}P_{k_j}}\left\|\left[\hat{\mathbf{H}}\right]_j\right\|^2 > \epsilon^{\text{thr}}$ **then**
    $\mathcal{K}_2 \leftarrow \mathcal{K}_2 \cup \{k_j\}$
    $\hat{\mathbf{H}}_{k_j} \leftarrow \frac{1}{g_{k_j}P_{k_j}}\left[\hat{\mathbf{H}}\right]_j$
  end if
end for

channel estimates on data carrying resource elements can be deduced using, for instance, LMMSE interpolation.

V. Numerical Results

Numerical results were obtained assuming a LTE-like system with RBs composed each of 12 sub-carriers over $T = 14$ OFDM symbols. Users' channels follow the Extended Vehicular A (EVA) model [14] with $f_k^D = 70$ Hz, $\max_l \tau_l = 25$ μs and $\mathbf{R}_{k,l}^{\alpha} = \sigma_l^2 \mathbf{I}_M$. The number N_p of pilot symbols per RB is equal to 24 and they span two OFDM symbols of the RB. As in LTE, the baseline pilots $\{\mathbf{u}_1,\ldots,\mathbf{u}_N\}$ are ZC sequences with distinct phase shifts. In the sequel, $N = 8$. All the following results have been obtained by averaging over 2×10^4 transmit time intervals (TTIs). During each TTI, four grant based UTs and a number K_2 of grant free UTs are simultaneously transmitting 32 bits of information on one RB using quadrature phase shift keying (QPSK) modulation and Turbo channel coding. In what follows, when K_2 is assumed fixed then the active K_2 grant free UTs are randomly chosen among a total of $K_2^{\max} = 8$ UTs. In Figure 2, the conditional detection probability (conditioned on a fixed value of K_2, namely $K_2 = 3$ and $K_2 = 6$) performance of Algorithm 1 is plotted as function of the signal-to-noise ratio (SNR) (defined as $\text{SNR}_k \overset{\text{def.}}{=} \frac{g_k P_k}{\sigma^2}$ and assumed here to be the same for all the active UTs as could be the case when uplink power control is applied) and is compared to a lower bound obtained assuming that only a smaller pool of possible grant free transmitters, namely of cardinality $K_2^{\max} = N_2 = 4$, needs to be supported. For such a small value, there is no need for the proposed scheme since baseline orthogonal pilot sequences can instead be used. The detection threshold ϵ^{thr} was set to 0.04 and the signature sequences were generated using an approximate solution to the sphere surface packing problem characterized by an average $\left|\mathbf{w}_k^{\mathrm{H}}\mathbf{w}_j\right|^2$ (averaged on all possible pairs of distinct signatures) that is approximately equal to 0.12. While these values violate the conservative condition in (21), they yielded a false alarm rate smaller than 5×10^{-3} on the whole range of considered SNR values hinting that (20) might be satisfied. Figure 2 shows that detection probability stays within an acceptable margin from the lower bound even while the proposed scheme supports collision free detection from a pool of $K_2^{\max} = 8$ grant free emitters i.e., twice the load supported by the scheme achieving the lower-bound.

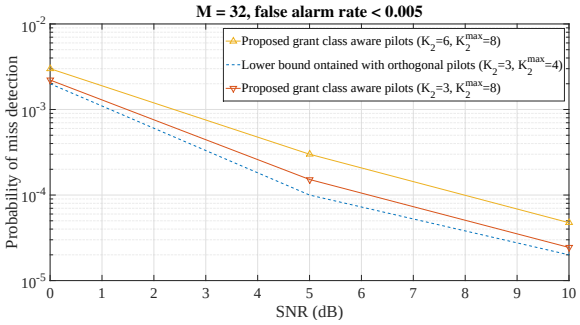

Fig. 2. User activity detection probability of the proposed pilot design conditioned on $K_2 = 3$ and $K_2 = 6$ respectively

To investigate the effect of the grant class aware pilot design on channel estimation quality, we simulate the BLER of a reference grant free UT (denoted in the sequel as $k_0 \in \mathcal{K}_2$) when MRC is done at the BS using the channel estimates obtained with the proposed pilot design and we pilot it as function of SNR_{k_0}. As we did in Figure 2, the simulated

[4]since based on single-user instead of joint channel estimation

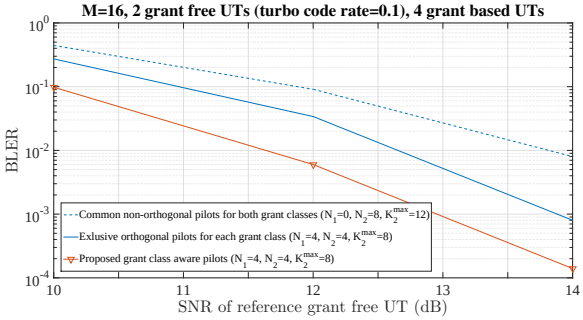

Fig. 3. BLER of a grant free transmission conditioned on $K_2 = 2$ and assuming four grant based transmissions

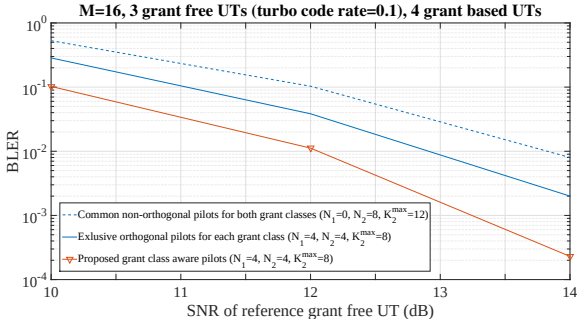

Fig. 4. BLER of a grant free transmission conditioned on $K_2 = 3$ and assuming four grant based transmissions

error probabilities are conditional probabilities, conditioned on a fixed value of K_2 ($K_2 = 2$ in Figure 3 and $K_2 = 4$ in Figure 4) but not on a fixed set $\mathcal{K}_2$. The thus obtained BLER curve is then compared to two conventional pilot schemes, namely *orthogonal pilot multiplexing* and *shared pilot multiplexing*. In *orthogonal pilot multiplexing*, half the baseline sequences are assigned to grant based UTs while the remaining sequences are assigned to grant free UTs. Note that whenever $K_2^{\max} > N/2$, the probability of a pilot collision between two grant free transmissions with this scheme is non-zero. In *non-orthogonal pilot multiplexing*, 12 non-orthogonal PN pilot sequences are generated to accommodate the $K_1 + K_2^{\max}$ UTs that are authorized to transmit on the considered RB. Figures 3 and 4 were generated assuming that the BS has successfully identified the active K_2 UTs from within the pool of $K_2^{\max}$ possible grant free transmitters. As expected, the gain in BLER performance from the proposed grant class aware pilot design w.r.t. the orthogonal pilot multiplexing scheme is higher for larger values of K_2 (at target BLER=10^{-3}, the gain is around 1.8 dB for $K_2 = 2$ and around 2.5 dB for $K_2 = 3$). This is because, in the orthogonal construction, the probability of two or more grant free transmissions with the same pilot sequence taking place simultaneously increases as K_2 increases. The figures also show that both the orthogonal and the proposed pilot designs outperform the non-orthogonal scheme in terms of BLER performance of the grant free UTs. This can be explained by the fact that the lower cross-correlation among the non-orthogonal sequences resulting from the higher value of N_2 in the non-orthogonal scheme ($N_2 = 8$ in the latter scheme as opposed to $N_2 = 4$ in the orthogonal and the proposed schemes) is not sufficient to counter the effect of the additional interference originating from the pilot signals of the grant based UTs (an interference that is nonexistent in the orthogonal and the proposed designs).

VI. Conclusions

In this article, we presented a new scheme to generate pilot sequences for uplink transmissions in MU-MIMO systems in which grant free and grant based transmissions are allowed to co-exist. The scheme is based on a novel method that can extend any set of orthogonal pilots by generating new pilots with a built-in trade-off between the relevant performance metrics of the two grant classes. We finally showed, using realistic simulations, that this grant class aware pilot design outperforms conventional pilot multiplexing schemes in terms of both CSI quality and user activity detection probability of the grant free transmissions, while guaranteeing that grant based transmissions get the same number of spatial layers as offered by those schemes.

References

[1] H. S. Dhillon, L. Lampe, H. Huang, and H. Visawanathan, "Wide-Area Wireless Communication Challenges for the Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 168-174, Feb. 2017.

[2] P. Popovski, J. J. Nielsen, Č. Stefanović, E. de Carvalho, E. Strömy, K. F. Trillingsgaard, A.-S. Bana, D. Min Kim, R. Kotaba, J. Park, René B. Sørensen, *Ultra-Reliable Low-Latency Communication (URLLC): Principles and Building Blocks*. Avilable: http://arXiv:1708.07862, Aug. 2017.

[3] E. de Carvalho, E. Björnson, J. H. Sørensen, E. G. Larsson, and P. Popovski, *Random Pilot and Data Access in Massive MIMO for Machine-type Communications*, Avilable: http://arXiv:1606.02080, Sept. 2017.

[4] E. de Carvalho, E. Björnson, E. G. Larsson, and P. Popovski, *Random Access for Massive MIMO Systems with Intra-Cell Pilot Contamination*, in *ICASSP*, Shanghai, Mar. 2016.

[5] L. Liu and W. Yu, *Massive Device Connectivity with Massive MIMO*, in *ISIT*, Aachen, June 2017.

[6] G. Wunder, P. Jung, and M. Ramadan, *Compressive Random Access Using A Common Overloaded Control Channel*, Avilable: http://arXiv:1504.05318, Sept. 2016.

[7] H. Wang, W. Zhang, Y. Liu, Q. Xu, and P. Pan "On Design of Non-Orthogonal Pilot Signals for a Multi-Cell Massive MIMO System," *IEEE Wireless Commun. Letters*, vol. 4, no. 2, Apr. 2015.

[8] A. K. Sadek, W. Su, and K. J. R. Liu, *Maximum Achievable Diversity for MIMO-OFDM Systems with Arbitrary Spatial Correlation*, in *GLOBECOM*, Dallas, Dec. 2004.

[9] A. Hedayat, H. Shah, and A. Nosratinia, "Analysis of Space-Time Coding in Correlated Fading Channels," *IEEE Trans. on Wireless Commun.*, vol. 4, no. 6, pp. 2882-2891, Nov. 2005.

[10] C.-X. Wang, X. Hong, H. Wu, and W. Xu, "Spatial-Temporal Correlation Properties of the 3GPP Spatial Channel Model and the Kronecker MIMO Channel Model," *EURASIP Journal on Wireless Communications and Networking*, vol. 2007, no. 1, Jan. 2007.

[11] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*. Cambridge University Press, 2009.

[12] N. Ksairi and M. Debbah, *Performance Analysis of Multi-Service Oriented Multiple Access Under General Channel Correlation*, in *GLOBECOM*, Singapore, Dec. 2017.

[13] A. Decurninge and M. Guillaud, *Cube-Split: Structured Quantizers on the Grassmannian of Lines*, in *WCNC*, San Francisco, Mar. 2017.

[14] The 3rd Generation Partnership Project (3GPP), *Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception*, Sep. 2015. Available: http://www.3gpp.org/.