This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

Dual-Polarized FDD Massive MIMO: A Comprehensive Framework

Mahdi Barzegar Khalilsarai[⋆], *Member, IEEE,* Tianyu Yang[⋆], Saeid Haghighatshoar[†], *Member, IEEE,* Xinping Yi[‡], *Member, IEEE,* and Giuseppe Caire[⋆], *Fellow, IEEE*

***Abstract*—We propose a comprehensive scheme for realizing a massive multiple-input multiple-output (MIMO) system with dual-polarized antennas in frequency division duplexing (FDD) mode. Dual-polarized arrays are commonly employed due to the favorable property that, in principle, they can double the number of channel spatial degrees of freedom with a less-than-proportional increase in array size. However, processing a dual-polarized massive MIMO channel is demanding due to the high channel dimension and the lack of Uplink-Downlink (UL-DL) channel reciprocity in FDD mode. In particular, the difficulty arises in common channel training and DL precoding in a multi-user setup . To address this, we develop a unified framework consisting of three steps: (1) covariance estimation to efficiently estimate the UL covariance from noisy UL pilots; (2) a UL-DL covariance transformation method that obtains the DL covariance from the estimated UL covariance, eliminating the need for DL channel covariance training via pilot transmission; (3) a joint multi-user DL channel training method, which enables the BS to estimate *effective DL channels* given any protocol-specific pilot dimension and to use them for interference-free DL beamforming and data transmission. Through extensive simulations, we show that our scheme is applicable to a variety of communication scenarios in terms of the number of antennas, UL and DL pilot dimensions, and angular scattering properties. Unlike the common trend in the literature, we do *not* make strong structural assumptions about the wireless channel (such as angular sparsity), ensuring a general treatment of the problem.**

***Index Terms*—Active channel sparsification, Channel covariance estimation, Dual-polarized FDD massive MIMO, Multi-user channel training, Uplink-Downlink covariance transformation.**

I. Introduction

Massive multiple-input multiple-output (MIMO) antenna systems promise high data rates as well as link reliability in prospective generations of wireless communication systems [1, 2]. The characteristic property of these systems is the deployment of a large number ($M \gg 1$) of antennas at the base station (BS), resulting in substantial improvements in terms of beamforming and multiplexing gains, while also increasing the array size. Since most wireless networks are currently based on *frequency division duplexing* (FDD), implementing a massive MIMO system in FDD mode is an appealing proposition. Besides, many network developers consider using dual-polarized (DP) antenna elements in the array, since this offers a doubling of the number of inputs with a less-than-proportional increase in array size [3, 4]. Using DP antennas results in UL and DL channels of dimension $2M$. In order to perform DL beamforming, the BS needs to obtain *fresh* channel estimates at each time-frequency coherence block and for all the K users. In a *time division duplexing* (TDD) system, this can be obtained from K mutually orthogonal UL pilots transmitted by the users and exploiting UL-DL channel reciprocity, which readily yields a DL channel estimate. This is unfortunately not the case in FDD systems, since UL and DL transmissions occur over disjoint frequency bands, violating instantaneous channel reciprocity. Therefore, the BS spends a fraction T_{dl} out of the total resource dimension T to broadcast pilots to the users. Upon receiving the pilots, the users send their T_{dl} "measurements" to the BS via closed-loop feedback, using which the BS estimates the DL channels.

Channel estimation in this way is generally challenging, since the channel dimension is large ($2M \gg 1$) and the pilot dimension is limited to the size of the coherence block ($T_{\text{dl}} \leq T$), which is used not only for pilots but also for data transmission. For example, in a standard LTE setup the users are scheduled over resource blocks containing 14 OFDM symbols and 12 subcarriers, making a total of $T = 14 \times 12 = 168$ dimensions [5]. With a DP array of, say, $M = 100$ antennas, the number of coefficients to be estimated amounts to $2M = 200$ which is larger than the block size $T = 168$ and (much) larger than the pilot dimension. As a result, since the number of linear channel measurements is less than the estimand vector, the BS can not obtain the channel state via conventional methods such as the Least Squares (LS) estimation. Hence the question is how to efficiently precode data to the users given a fixed DL pilot dimension that is small relative to the channel dimension. Our answer to these questions involves several steps that are outlined as follows.

A. Channel Covariance Estimation

Channel covariance knowledge at the BS either for UL or DL is crucial not only for designing efficient DL precoders, but also for a variety of tasks including minimum mean squared error (MMSE) channel estimation and user grouping . During UL, each user transmits a number of orthogonal pilots to the BS, which in turn uses the set of observed channel samples to estimate the UL channel covariance. The simplest and most common estimator is the sample covariance. It is well-known that in scenarios (such as the one in hand with a massive array $M \gg 1$), in which the number of samples (N) is comparable

[1] Communications and Information Theory Group (CommIT), Technische Universität Berlin ({m.barzegarkhalilsarai, tianyu.yang, caire}@tu-berlin.de).

[2] Saeid Haghighatshoar is currently with the Swiss Center for Electronics and Microtechnology (CSEM), however his contribution to this work was made while he was with the CommIT group (saeid.haghighatshoar@csem.ch).

[3] Department of Electrical Engineering and Electronics, University of Liverpool (xinping.yi@liverpool.ac.uk).

to the signal dimension ($2M$), the sample covariance can be substantially improved by taking into account the covariance structure. There are several ways to exploit the structure in estimation, including recent methods that consider low-rank and sparse covariance models. For example, methods based on rank minimization or nuclear norm minimization (for low-rank covariances), and ℓ_0-pseudo-norm or ℓ_1-norm minimization (for sparse covariances) or combinations thereof are proposed [6–8].

In this paper we do not assume the DP channel covariance to be low-rank or sparse as is postulated in many works in the literature in similar problems [9–12], but as we will show it follows a Kronecker-type form and is given by an integral transform involving a positive semidefinite matrix-valued function of the angle of arrival (AoA). This function, coined as the *dual-polarized angular spread function* (DP-ASF), represents the channel angular power density in H and V polarizations as well as the cross-correlation between the two. Our approach to covariance estimation is based on a parametric representation of the DP-ASF in terms of a linear combination of elementary, limited-support density functions, whose coefficients are estimated given independent DP channel samples $\{\mathbf{h}_{\text{ul}}(i)\}_{i=1}^{N}$. This parametric model is general, in that, it incorporates specular as well as diffuse angular scattering and does not assume unverified polarization properties. The estimation is carried out via a convex program, which enforces the positive semidefinite property on the solution. After estimating the DP-ASF, an estimate of the UL covariance is readily given by a simple integral transform.

B. Uplink-Downlink Channel Covariance Transformation

In addition to the UL covariance, the BS needs to obtain an estimate of the DL covariance for all users both to obtain a reliable estimate of user DL channels and to design a DL precoder for multi-user data transmission. In an FDD system, UL and DL covariances are different and therefore the DL covariance has to be estimated via DL training and UL closed-loop feedback. This process is not efficient since the overhead of transmitting DL pilots, receiving feedback from the users and then estimating the DL covariance is too large. In order to estimate the DL covariance, we propose a UL-DL covariance transformation method, which hinges upon a phenomenon known as *angular channel reciprocity*: the angular power density as seen from the array is the same for UL and DL, resulting in the DP-ASF to be identical during UL and DL. The concept of angular channel reciprocity is well-established in the literature (e.g., [13, 14]) considered for the single-polarized array, and what we propose here is its natural extension to the DP array. Having an estimate of the DP-ASF from the previous step, we use angular reciprocity and a change of the array response from UL to DL to obtain an estimate of the DL covariance.

C. Downlink Channel Training and Precoding via Active Sparsification

In order to achieve the gains of massive MIMO, it is necessary for the BS to estimate (train) instantaneous user DL channels and perform interference-free DL beamforming. While channel training is an easy task with small MIMO arrays, it becomes increasingly challenging when the number of antennas grows large. This is especially an issue in FDD mode, where instantaneous channel reciprocity does not hold, and UL and DL channels corresponding to different frequency bands are virtually uncorrelated random vectors. The DL channels can not be obtained simply from their UL counterparts and therefore, the BS has to probe the channel in the DL by broadcasting pilot symbols, receive feedback from the users and finally estimate the DL channel. In order to estimate a $2M$-dimensional DP channel with any conventional method and without assumptions such as channel sparsity, the BS needs to transmit at least $2M$ pilot symbols and receive their feedback in the UL to have "stable" channel estimates. On the other hand, as explained earlier the time-frequency resources of a single coherence block are used for both channel training and data transmission. Dedicating a number T_{dl} of a total of T coherence block dimensions to DL training introduces a *pre-log factor* of $\max\{0, 1 - T_{\text{dl}}/T\}$ in the sum-rate. When the pilot dimension is comparable to the large number of antennas, this factor will be small or in the worst case, equal to zero.

This dimensionality problem is not solved even by resorting to the channel sparsity assumption and various compressed sensing (CS) techniques (see e.g. [9] and [10]). First, the channel sparsity postulate may not be always verified, especially in rich scattering environments and in the presence of diffuse scattering (in fact sources such as [15] call this assumption the *sparsity hypothesis*). Therefore, CS techniques are always at the mercy of environmental properties, as to whether the channel is indeed sparse or not. Second, even if the sparsity assumption holds, the number of measurements necessary for accurate sparse recovery might be still high, exceeding the available DL pilot dimension.

To resolve this issue, we adopt and extend the *active channel sparsification* (ACS) approach first proposed by some of the authors in [13] for single-polarized arrays. Given user DL covariances and for a given pilot dimension T_{dl}, the idea of ACS is to design a sparsifying precoder that jointly reduces the number of significant angular components of all the user channels to less than T_{dl}, while at the same time maximizing the rank of the sparsified effective channel matrix. This enables, as it will be shown, stable recovery of the effective user channels and simultaneously maximizing the system multiplexing gain, which is proportional to the channel matrix rank. Using the ACS method, we are not at the mercy of channel's sparsity features and we do not make any assumptions thereof. ACS is deployed via first identifying a set of common virtual beams among all the users for channel representation and forming a user-virtual beam bipartite graph. Then we prove a result, relating the channel matrix rank to the maximal matching size in the graph. Finally, the sparsifying precoder is given by selecting a subset of users and virtual beams as the solution to a *mixed integer linear program* (MILP). The block diagram of Fig. 1 summarizes the steps described above.

Perhaps the most relevant work that addresses the FDD massive MIMO problem is the method proposed in [11]. This work assumes channels with discrete and sparse multipath

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

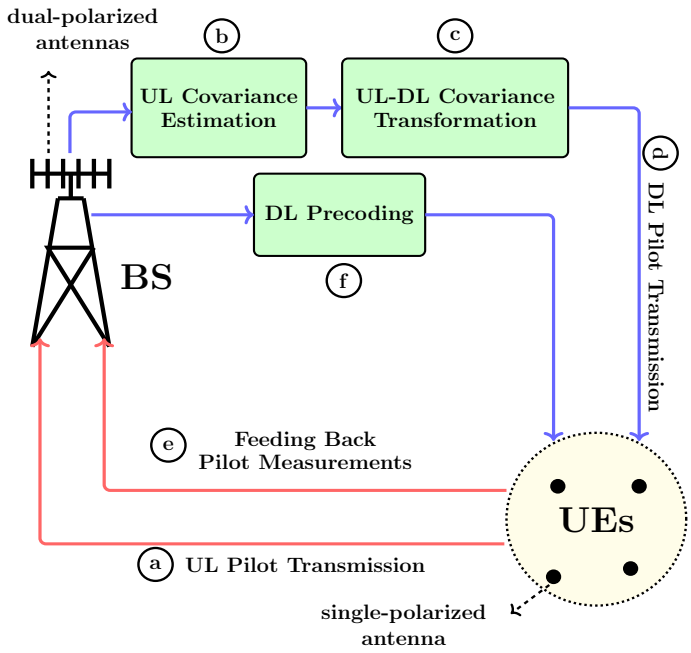


Fig. 1: Overall block diagram of our scheme.

components and proposes to estimate the channel parameters, namely the AoAs and complex coefficients of the signal paths at the user side via a tensor factorization method. This method suffers from the same shortcoming as CS-based methods in that its success heavily depends on the sparsity assumption. In fact the multipath identifiablity bound given in this work gives guarantees for channels that are extremely sparse even when the pilot dimensions is large (see Theorem 6 in [11]). Since parameter estimation is done at the user side, this bound is yet more restrictive when the number of antennas at the user side is small, as is the case in today's smartphones and other user equipment. In contrast, our method does not depend on channel sparsity and applies to cases with diffuse as well as discrete scattering. It is also computationally more efficient, since it performs computation at the BS side, which has access to much more computational resources than what is available in a user's device.

D. Organization

The paper is organized as follows. In Section II we introduce the dual-polarized channel model. In Section III we develop our channel covariance estimator. Section IV discusses UL-DL covariance transformation. In Section V we elaborate on the ACS method. Various empirical results in Section VI conclude the paper.

II. Channel Model

We consider a *uniform linear array* (ULA) of M dual-polarized antenna elements that communicates with a user that has one single-polarized antenna. The Uplink (UL) channel can be represented as

$$\mathbf{h}_{\text{ul}} = \begin{bmatrix} \mathbf{h}_{\text{ul,H}} \\ \mathbf{h}_{\text{ul,V}} \end{bmatrix} \in \mathbb{C}^{2M}, \tag{1}$$

where $\mathbf{h}_{\text{ul,H}} \in \mathbb{C}^M$ is the channel vector corresponding to the M H-polarized antenna ports and $\mathbf{h}_{\text{ul,V}} \in \mathbb{C}^M$ is the one corresponding to the M V-polarized ports. The channel for either polarization is a superposition of random gains along a continuum of AoAs, weighted by the *antenna element response* which for antenna m is given by $a_m = e^{j\pi m \frac{2d\sin(\theta)}{\lambda_{\text{ul}}}}$, where d is the antenna spacing, $\theta \in [-\theta_{\max}, \theta_{\max}]$ is the AoA, $\theta_{\max}$ is the maximum array angular aperture and λ_{ul} is the wave-length of the electromagnetic wave at the UL carrier frequency. Taking the antenna spacing to be the standard $d = \frac{\lambda_{\text{ul}}}{2\sin\theta_{\max}}$ and with the change of variables $\xi = \frac{\sin\theta}{\sin\theta_{\max}} \in [-1, 1]$, the antenna element response admits the simpler form

$$a_m = e^{jm\pi\xi}, \ m = 0, \dots, M-1, \tag{2}$$

where ξ is understood as the *normalized* AoA parameter. Then, we can express H and V channel vectors as

$$\mathbf{h}_{\text{ul,H}} = \int_{-1}^{1} W_{\text{H}}(\xi)\mathbf{a}_{\text{ul}}(\xi)d\xi, \tag{3a}$$

$$\mathbf{h}_{\text{ul,V}} = \int_{-1}^{1} W_{\text{V}}(\xi)\mathbf{a}_{\text{ul}}(\xi)d\xi \tag{3b}$$

where $\mathbf{a}(\xi) = [1, e^{j\pi\xi}, \dots, e^{j\pi(M-1)\xi}]^{\mathsf{T}}$ is the *array response vector*, and W_{H} and W_{V} are random processes representing the angular gains along each AoA for H and V polarizations, respectively. We assume W_{H} and W_{V} to be zero-mean, circularly symmetric, complex Gaussian processes with the following autocorrelations:

$$\mathbb{E}\left[W_{\text{H}}(\xi)W_{\text{H}}^*(\xi')\right] = \gamma_{\text{H}}(\xi)\delta(\xi - \xi'), \tag{4a}$$

$$\mathbb{E}\left[W_{\text{V}}(\xi)W_{\text{V}}^*(\xi')\right] = \gamma_{\text{V}}(\xi)\delta(\xi - \xi'), \tag{4b}$$

where we have adopted the *wide-sense stationary uncorrelated scattering* (WSSUS) model, which assumes stationary second-order channel statistics (over reasonably short time intervals)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

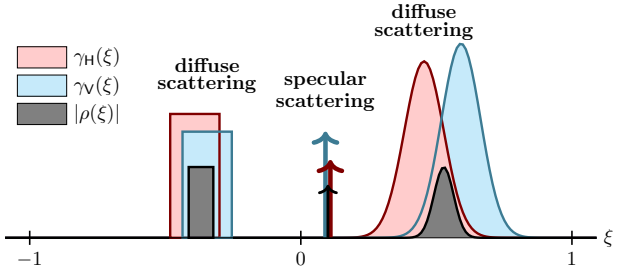


Fig. 2: An example of H and V ASFs as well as the H-V cross-correlation modulus. The blue shaded function highlights $\gamma_{\mathsf{V}}(\xi)$, the red one highlights $\gamma_{\mathsf{H}}(\xi)$ and the black one highlights $|\rho(\xi)|$.

and uncorrelated angular scattering gains [16]. The functions[1] γ_{H} and γ_{V} are both real and non-negative, representing the channel power density received along each AoA for H and V polarizations, respectively. We call these horizontal and vertical angular spread functions (ASFs) (see Fig. 2). In practice, the H and V links can *not* be entirely isolated from each other and therefore, there exists a leakage of channel power between the two. This implies that, for each AoA, the random gains $W_{\mathsf{H}}(\xi)$ and $W_{\mathsf{V}}(\xi)$ are correlated such that we have

$$\mathbb{E}\left[W_{\mathsf{H}}(\xi)W_{\mathsf{V}}^{*}(\xi')\right] = \rho(\xi)\delta(\xi-\xi'), \tag{5}$$

where ρ is a generally complex-valued function.

Remark *1:* It is conventional in the literature to model the channel vector explicitly in terms of its line-of-sight (LoS) and non-line-of-sight (NLoS) components as

$$\mathbf{h}_{\mathrm{ul,(H/V)}} = \sqrt{\alpha}\,\mathbf{h}^{\mathrm{LoS}}_{\mathrm{ul,(H/V)}} + \sqrt{1-\alpha}\,\mathbf{h}^{\mathrm{NLoS}}_{\mathrm{ul,(H/V)}},$$

for an energy normalization scalar $\alpha \in [0,1]$ [11, 18]. Furthermore, each component is modeled as a superposition of the array responses to waves impinging from discrete, separable paths, represented as $\mathbf{h}^{\mathrm{LoS}}_{\mathrm{ul,(H/V)}} = \beta^{\mathrm{LoS}}_{\mathrm{ul,(H/V)}}\mathbf{a}_{\mathrm{ul}}(\xi_{\mathrm{LoS}})$ and $\mathbf{h}^{\mathrm{NLoS}}_{\mathrm{ul,(H/V)}} = \sum_{i=1}^{p-1}\beta^{\mathrm{NLoS}}_{\mathrm{ul,(H/V)},i}\mathbf{a}_{\mathrm{ul}}(\xi_{\mathrm{NLoS},i})$, where p is the total number of paths, $\xi_{\mathrm{LoS}}, \{\xi_{\mathrm{NLoS},i}\}_i$ are the AoAs and $\beta^{\mathrm{LoS}}_{\mathrm{ul,(H/V)}}$, $\{\beta^{\mathrm{NLoS}}_{\mathrm{ul,(H/V)},i}\}_i$ are complex-valued gains. The model that we use in (3) includes this decomposition as a special case, in which the angular gain process has non-zero variance over a finite set of discrete AoAs corresponding to LoS and NLoS signal paths. The H and V ASF's in this case would be delta trains, where the location of each delta corresponds to an LoS or NLoS AoA and the positive gains are the corresponding signal powers. △

From (1) and (3) the dual-polarized UL channel can be more conveniently expressed as

$$\begin{aligned}\mathbf{h}_{\mathrm{ul}} &= \int_{-1}^{1}\begin{bmatrix}\mathbf{a}_{\mathrm{ul}}(\xi) & \mathbf{0}\\ \mathbf{0} & \mathbf{a}_{\mathrm{ul}}(\xi)\end{bmatrix}\begin{bmatrix}W_{\mathsf{H}}(\xi)\\ W_{\mathsf{V}}(\xi)\end{bmatrix}d\xi\\ &= \int_{-1}^{1}\left(\mathbf{I}_2\otimes\mathbf{a}_{\mathrm{ul}}(\xi)\right)\mathbf{w}(\xi)d\xi,\end{aligned} \tag{6}$$

[1]We use the term "function" with some abuse of terminology. An accurate term would be "distribution" in the sense of generalized functions [17], since we also consider Dirac's delta which is not a function in the conventional sense.

where $\otimes$ denotes Kronecker product, and $\mathbf{w}(\xi) := [W_{\mathsf{H}}(\xi), W_{\mathsf{V}}(\xi)]^{\mathsf{T}}$. The channel covariance can be computed according to (6) as

$$\mathbf{\Sigma}^{\mathrm{ul}}_{\mathbf{h}} = \mathbb{E}\left[\mathbf{h}_{\mathrm{ul}}\mathbf{h}_{\mathrm{ul}}^{\mathsf{H}}\right] = \int_{-1}^{1}\mathbf{\Gamma}(\xi)\otimes\mathbf{A}_{\mathrm{ul}}(\xi)d\xi, \tag{7}$$

where we have defined the rank-1 matrix $\mathbf{A}_{\mathrm{ul}}(\xi) = \mathbf{a}_{\mathrm{ul}}(\xi)\mathbf{a}_{\mathrm{ul}}(\xi)^{\mathsf{H}}$, and the matrix-valued function

$$\mathbf{\Gamma}(\xi) = \mathbb{E}\left[\mathbf{w}(\xi)\mathbf{w}(\xi')^{\mathsf{H}}\right] = \begin{bmatrix}\gamma_{\mathsf{H}}(\xi) & \rho(\xi)\\ \rho(\xi)^{*} & \gamma_{\mathsf{V}}(\xi)\end{bmatrix} \in \mathbb{C}^{2\times 2}, \tag{8}$$

which is positive semidefinite (PSD) for all $\xi \in [-1,1]$. We call $\mathbf{\Gamma}(\xi)$ the *dual-polarized angular spread function* (DP-ASF) and we note that, similar to the role played by the ASF in a single-polarized array, the DP-ASF captures the angular spectral properties of the channel, i.e. the power density along H and V links and the power leakage density between the two.

III. Uplink Channel Covariance Estimation

Suppose that the BS receives N noisy pilots in the UL as

$$\mathbf{y}_{\mathrm{ul}}(i) = \mathbf{h}_{\mathrm{ul}}(i)\,x_n + \mathbf{z}(i),\ i = 1,\ldots,N, \tag{9}$$

where $x_i = \sqrt{P}$ is the pilot symbol, $\mathbf{z}(i) \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_{2M})$ is the additive white Gaussian noise (AWGN) vector at the i-th transmission with N_0 being the noise variance per element, and $\mathbf{h}_{\mathrm{ul}}(i)$ is the i-th random channel realization. With orthogonal pilot transmission over distinct time-frequency coherence blocks, we can safely assume that the channel realizations $\mathbf{h}(i)$, $n = 1,\ldots,N$ are independent. A simple estimator of the UL channel covariance $\mathbf{\Sigma}^{\mathrm{ul}}_{\mathbf{h}}$ is given by the sample covariance matrix

$$\widehat{\mathbf{\Sigma}}^{\mathrm{ul}}_{\mathbf{h}} = \widehat{\mathbf{\Sigma}}^{\mathrm{ul}}_{\mathbf{y}} - N_0\mathbf{I}_{2M} := \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_{\mathrm{ul}}(i)\mathbf{y}_{\mathrm{ul}}(i)^{\mathsf{H}} - N_0\mathbf{I}_{2M}, \tag{10}$$

The sample covariance is a consistent estimator of the true covariance and converges to it for relatively large number of samples ($N \gg 2M$), obtaining which is affordable in the case of small MIMO channels. However, for a dual-polarized massive MIMO channel with $2M \gg 1$, this condition is hardly met and instead, the number of samples is in the order of the channel dimension ($N = \mathcal{O}(2M)$). In these regimes of dimensionality, it is well-known that one can considerably improve the sample covariance estimator, for example by exploiting the covariance structure. In particular, here we are interested in covariance matrices that belong to the set of feasible DP MIMO covariances of a ULA defined as

$$\mathcal{C}_{\mathrm{ul}} := \left\{\int_{-1}^{1}\mathbf{\Phi}(\xi)\otimes\mathbf{A}_{\mathrm{ul}}(\xi)d\xi,\ \mathbf{\Phi}:[-1,1]\to\mathbb{S}^2_{+}\right\}, \tag{11}$$

where $\mathbf{\Phi}$ is a generic DP-ASF and $\mathbb{S}^2_{+}$ denotes the set of 2×2 PSD matrices. This structure will be considered in estimation of the covariance as follows.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

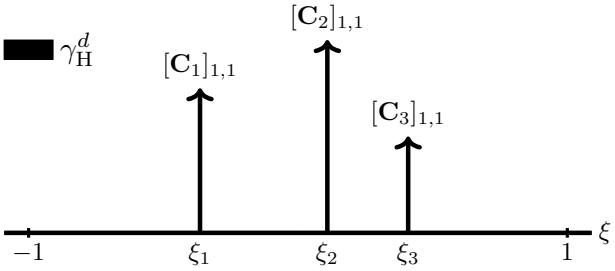


Fig. 3: An example of the discrete component of the ASF corresponding to the H polarization

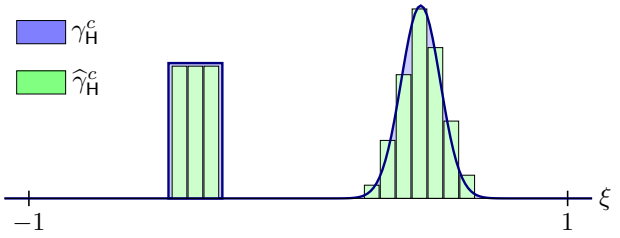


Fig. 4: An example of the diffuse component of the ASF corresponding to the H polarization and its approximation with rectangular densities.

A. Parametric Representation of the DP-ASF

The DP-ASF of a channel models the received power density over each AoA. This power density in turn depends on the scattering properties of the environment: partly it comes from line of sight (LoS) propagation, specular reflection and wedge diffraction in the environment that occupy narrow angular intervals, while the rest of the power comes from diffuse scattering, occupying wide angular intervals [3] (see Fig. 2). In order to distinguish between these two types of multipath effects, we decompose the DP-ASF into "discrete" and "continuous" (diffuse) components:

$$\mathbf{\Gamma} = \mathbf{\Gamma}_d + \mathbf{\Gamma}_c, \tag{12}$$

where $\mathbf{\Gamma}_c$ is the continuous component and $\mathbf{\Gamma}_d$ is the discrete component. For the discrete part, the parametric form is simply given by a train of weighted delta functions:

$$\mathbf{\Gamma}_d(\xi) = \begin{bmatrix} \gamma_{\mathrm{H}}^d(\xi) & \rho_d(\xi) \\ \rho_d^*(\xi) & \gamma_{\mathrm{V}}^d(\xi) \end{bmatrix} = \sum_{i=1}^{r} \mathbf{C}_i \delta(\xi - \xi_i),\ \xi \in [-1,1], \tag{13}$$

where $\mathbf{C}_i \succeq \mathbf{0},\ i = 1,\dots,r$ are 2×2 PSD matrices, $\xi_i,\ i = 1,\dots,r$ are discrete AoAs, and γ_{H}^d, γ_{V}^d, and ρ_d are discrete components of the H and V ASF's and the cross-correlation term, respectively. Fig. 3 illustrates an example of γ_{H}^d with $r = 3$ discrete AoAs.

In contrast to the discrete component, we can not assume a parametric description of $\mathbf{\Gamma}_c$ in terms of delta functions. Instead, we define a dictionary of n density functions with small support[2]

$$\Psi_c := \{\psi_i(\xi) \geq 0\, \forall \xi,\ |\mathrm{supp}(\psi_i)| \ll 1 : i = 1,\dots,n\}, \tag{14}$$

using which we approximate $\mathbf{\Gamma}_c$ as

$$\mathbf{\Gamma}_c(\xi) = \begin{bmatrix} \gamma_{\mathrm{H}}^d(\xi) & \rho_d(\xi) \\ \rho_d^*(\xi) & \gamma_{\mathrm{V}}^d(\xi) \end{bmatrix} \approx \sum_{i=1}^{n} \mathbf{C}_i' \psi_i(\xi), \tag{15}$$

for $\xi \in [-1,1]$ where similar to (13) $\mathbf{C}_i',\ i = 1,\dots,n$ are 2×2 PSD matrices. If Ψ_c is suitably chosen and is large enough ($n \gg 1$), then one can find the coefficients $\mathbf{C}_i'$ such that the approximation error in (15) is small. A simple choice for Ψ_c is to consider it as a collection of a large number of densities with non-overlapping support $\mathrm{supp}(\psi_i) \cap \mathrm{supp}(\psi_j) = \varnothing, i \neq j$. As we will shortly see, this results in a simple constraint in the estimation of the coefficient matrices. Also there are various options to choose the densities. Two examples are rectangular densities[3] $\psi_i^{\mathtt{rect}}(\xi) = \mathtt{rect}_{[-1+\frac{2(i-1)}{n}, -1+\frac{2i}{n}]}$, $i = 1,\dots,n$, and truncated Gaussian densities $\psi_i^{\mathtt{gauss}}(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\xi - \frac{2i-1}{n})^2}{2\sigma^2}) \mathtt{rect}_{[-1+\frac{2(i-1)}{n}, -1+\frac{2i}{n}]}, i = 1,\dots,n$, for $\xi \in [-1,1]$. Fig. 4 illustrates an example of approximation with rectangular densities.

Using (7), (13) and (15), we can derive a similar discrete-continuous decomposition for the UL channel covariance as

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{h}}^{\mathrm{ul}} &= \mathbf{\Sigma}_{\mathbf{h}}^{\mathrm{ul,d}} + \mathbf{\Sigma}_{\mathbf{h}}^{\mathrm{ul,c}} \\ &= \int_{-1}^{1} \mathbf{\Gamma}_d(\xi) \otimes \mathbf{A}_{\mathrm{ul}}(\xi) d\xi + \int_{-1}^{1} \mathbf{\Gamma}_c(\xi) \otimes \mathbf{A}_{\mathrm{ul}}(\xi) d\xi \\ &\approx \sum_{i=1}^{r} \mathbf{C}_i \otimes \mathbf{A}_{\mathrm{ul}}(\xi_i) + \sum_{i=1}^{n} \mathbf{C}_i' \otimes \mathbf{A}_{\mathrm{ul},i}', \end{aligned} \tag{16}$$

where we have defined $\mathbf{A}_{\mathrm{ul},i}' = \int_{-1}^{1} \psi_i(\xi) \mathbf{A}_{\mathrm{ul}}(\xi) d\xi \in \mathbb{C}^{M\times M}$. If the discrete AoAs $\{\xi_i\}_{i=1}^r$ were known, we could claim via Eq. (16) that estimating $\mathbf{\Sigma}_{\mathbf{h}}^{\mathrm{ul}}$ is equivalent to estimating the coefficient matrices $\{\mathbf{C}_i\}_{i=1}^r$ and $\{\mathbf{C}_i'\}_{i=1}^n$. In order to make this strategy plausible, one needs to first estimate the discrete AoAs $\{\xi_i\}_{i=1}^r$.

B. Estimating Discrete AoAs

We propose a heuristic eigenspace method for estimating discrete AoAs, inspired by the Multiple Signal Classification (MUSIC) algorithm, which is a well-known spectral estimation method [19]. In a standard case (e.g. the single-polarized channel) the pilot measurements covariance is given by a matrix $\mathbf{\Sigma}_{\mathbf{y}}^{\mathrm{ul}} = \sum_{i=1}^{r} c_i \mathbf{a}_{\mathrm{ul}}(\xi_i) \mathbf{a}_{\mathrm{ul}}(\xi_i)^{\mathsf{H}} + \mathbf{\Sigma}_c$ where $\{c_i\}_i$ are positive scalars, $\{\xi_i\}_i$ are discrete AoAs, and $\mathbf{\Sigma}_c$ is the covariance of the part of channel resulting from diffuse scattering plus the additive noise covariance. MUSIC first computes the sample covariance $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{\mathrm{ul}}$ from a set of noisy pilots. Then it computes the eigen-decomposition of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{\mathrm{ul}}$ as $\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{U}}^{\mathsf{H}}$, where $\mathbf{U} \in \mathbb{C}^{M\times M}$ is a unitary matrix and $\widehat{\mathbf{D}} \in \mathbb{R}_+^{M\times M}$ is diagonal with real, non-negative elements that are sorted descendingly. Define the "noise subspace" as the space spanned by the last $M - r$ columns of $\widehat{\mathbf{U}}$. It is shown that under certain separability conditions on the discrete AoAs, the ratio of power between the discrete and diffuse parts of the ASF, and the noise power, MUSIC can recover the discrete AoAs by

[2]The support of a function f on a domain Ω is defined as $\mathrm{supp}(f) = \{x \in \Omega : f(x) \neq 0\}$.

[3]For interval $\mathcal{A}$, $\mathtt{rect}_{\mathcal{A}}$ is the indicator function over $\mathcal{A}$, i.e. $\mathtt{rect}_{\mathcal{A}} = 1_{\mathcal{A}}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

minimizing the "pseudo-spectrum" $\eta(\xi) = \left\|\mathbf{U}_{\text{noi}}^{\mathsf{H}}\,\mathbf{a}(\xi)\right\|^2$ over the candidate AoAs $\xi \in [-1, 1]$ [20, 21]. We can extend this idea to the present case, where the dual-polarized channel covariance consists of Kronecker-product components of the form $\mathbf{C}_i \otimes \mathbf{a}_{\text{ul}}(\xi_i)\mathbf{a}_{\text{ul}}(\xi_i)^{\mathsf{H}}$ by defining the pseudo-spectrum as

$$\eta(\xi) = \left\|\mathbf{U}_{\text{noi}}^{\mathsf{H}}\,(\mathbf{1} \otimes \mathbf{a}(\xi))\right\|^2, \tag{17}$$

where $\mathbf{1} = [1, 1]^{\mathsf{T}}$ and $\mathbf{U}_{\text{noi}}^{\mathsf{H}}$ represents the noise subspace of the DP channel sample covariance, containing the $2M - 2r$ eigenvectors of the sample covariance associated with its $2M - 2r$ smallest eigenvalues. We estimate the discrete AoAs as the r minimizers of η in (17) over $\xi \in [-1, 1]$ that have the smallest pseudo-spectrum value.

Note that in practice the number of discrete AoAs (r) is not given, but it can be learned over time considering the fact that in large dimensions, the largest eigenvalues of the channel sample covariance correspond to discrete signal paths and counting them gives an estimate ($\widehat{r}$) of the number of such paths. This is not always easy to implement due to the ambiguity in determining the largest eigenvalues, as this always depends on a relative measure of the sum-power of the largest eigenvalues compared to the rest. But we note that, overestimating the number of spikes is better than underestimating it and should be preferred. If "fake" spikes (i.e., false positives) appear in the set of estimated discrete AoAs, they will be eventually associated with small coefficients in the next coefficient estimation step. However, if a true spike is not detected, then we may not get an accurate covariance estimate as no term in the parametric expansion (16) will compensate for the contribution of the missing spike.

Recalling (16), now we can say that estimating $\mathbf{\Sigma}_{\mathbf{h}}^{ul}$ is equivalent to estimating the $n + \widehat{r}$ coefficient parameters, namely $\{\mathbf{C}_i\}_{i=1}^{\widehat{r}}$ and $\{\mathbf{C}'_i\}_{i=1}^{n}$.

C. Estimating DP-ASF Coefficients

Let us first reformulate the channel covariance parametric description in a simpler form. Define the known $M \times M$ matrices $\mathbf{S}_i^{\text{ul}} = \mathbf{A}_{\text{ul}}(\widehat{\xi}_i)$ for $i = 1, \ldots, \widehat{r}$ and $\mathbf{S}_i^{\text{ul}} = \mathbf{A}'_{\text{ul},i-\widehat{r}}$ for $i = \widehat{r}+1, \ldots, \widehat{r}+n$. Also define their associated unknown coefficients as $\mathbf{W}_i = \mathbf{C}_i$ for $i = 1, \ldots, \widehat{r}$ and $\mathbf{W}_i = \mathbf{C}'_{i-\widehat{r}}$ for $i = \widehat{r}+1, \ldots, \widehat{r}+n$. Then according to the right-hand-side of (16) we can formulate the parametric expression of the covariance as

$$\widetilde{\mathbf{\Sigma}}_{\mathbf{h}}^{\text{ul}}(\{\mathbf{W}_i\}_{i=1}^{\widehat{r}+n}) \approx \sum_{i=1}^{\widehat{r}+n} \mathbf{W}_i \otimes \mathbf{S}_i^{\text{ul}}. \tag{18}$$

Now, the problem is to estimate the coefficient matrices $\{\mathbf{W}_i \in \mathbb{S}_+^2\}_{i=1}^{n+\widehat{r}}$, given noisy pilot measurements $\{\mathbf{y}_{\text{ul}}(j)\}_{j=1}^{N}$ in (9). Our proposition for performing this task is based on minimizing the the difference between the channel sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{h}}^{\text{ul}}$ and $\widetilde{\mathbf{\Sigma}}_{\mathbf{h}}^{\text{ul}}$ as a function of the coefficients. We perform the minimization by constraining the coefficients to be PSD. Formally, we have the following optimization problem:

$$\begin{aligned}\{\widehat{\mathbf{W}}_i\}_{i=1}^{\widehat{r}+n} = \underset{\{\mathbf{W}_i\}_{i=1}^{\widehat{r}+n}}{\arg\min} \quad & \|\widehat{\mathbf{\Sigma}}_{\mathbf{h}}^{\text{ul}} - \sum_{i=1}^{\widehat{r}+n} \mathbf{W}_i \otimes \mathbf{S}_i^{\text{ul}}\|_{\mathsf{F}}^2 \\ \text{subject to} \quad & \mathbf{W}_i \succeq \mathbf{0},\ i = 1, \ldots, \widehat{r}+n.\end{aligned} \tag{19}$$

We call this problem a *positive semi-definite least-squares* (PSD-LS) program. The PSD-LS is convex and can be solved using standard software. Then we obtain the covariance estimate simply by using (18) and replacing $\mathbf{W}_i$ with $\widehat{\mathbf{W}}_i$, which yields

$$\mathbf{\Sigma}_{\mathbf{h}}^{\text{ul}\,\star} = \sum_{i=1}^{\widehat{r}+n} \widehat{\mathbf{W}}_i \otimes \mathbf{S}_i^{\text{ul}}. \tag{20}$$

Note that solving (19) also provides an estimate of the DP-ASF using (13) and (15) as

$$\widehat{\mathbf{\Gamma}}(\xi) = \sum_{i=1}^{\widehat{r}} \widehat{\mathbf{W}}_i\,\delta(\xi - \widehat{\xi}_i) + \sum_{i=1}^{n} \widehat{\mathbf{W}}_{\widehat{r}+i}\,\psi_i(\xi). \tag{21}$$

IV. UL-DL Covariance Transformation

Estimating DL channel covariance is necessary for MMSE channel estimation and multi-user common channel training. Similar to (3), the DL channel for H and V polarizations can be expressed respectively as

$$\mathbf{h}_{\text{dl,H}} = \int_{-1}^{1} W_{\mathsf{H}}^{\text{dl}}(\xi)\mathbf{a}_{\text{dl}}(\xi)d\xi, \tag{22a}$$

$$\mathbf{h}_{\text{dl,V}} = \int_{-1}^{1} W_{\mathsf{V}}^{\text{dl}}(\xi)\mathbf{a}_{\text{dl}}(\xi)d\xi, \tag{22b}$$

where

$$\mathbf{a}_{\text{dl}}(\xi) = [1, e^{j\pi\nu\xi}, \ldots, e^{j\pi(M-1)\nu\xi}]^{\mathsf{T}} \in \mathbb{C}^M \tag{23}$$

is the DL array response and $\nu = \frac{\lambda_{\text{ul}}}{\lambda_{\text{dl}}} = \frac{f_{\text{dl}}}{f_{\text{ul}}}$ is the DL to UL carrier frequency ratio. The factor ν appears in the exponents because of the change in the wavelength from UL to DL: the response of antenna element m in DL in terms of the AoA variable θ is equal to $e^{j\pi m\frac{2d\sin(\theta)}{\lambda_{\text{dl}}}}$. Setting the antenna spacing to $d = \frac{\lambda_{\text{ul}}}{2\sin\theta_{\max}}$ as before and considering the change of variables $\xi = \frac{\sin\theta}{\sin\theta_{\max}}$ we have $e^{j\pi m\frac{2d\sin(\theta)}{\lambda_{\text{dl}}}} = e^{j\pi m\frac{\lambda_{\text{ul}}}{\lambda_{\text{dl}}}\xi}$ which results in formula (23) for the DL array response vector. We assume the H and V random angular gain processes $W_{\mathsf{H}}^{\text{dl}}$ and $W_{\mathsf{V}}^{\text{dl}}$ in (22) to be zero-mean complex Gaussian. These process depend only on the angular location of the scatterers and their response to the electromagnetic wave. While a random realization of such a response can be different from UL to DL, it seems reasonable to assume that their spectral properties remain the same for UL and DL bands that are not significantly far from each other. Fortunately, UL and DL bands in an FDD system are located close to each other in comparison to their carrier frequencies. For example, in a certain FDD UL-DL band we have $f_{\text{ul}} = 1920$ MHz and $f_{\text{dl}} = 2110$ MHz [5], which means that we have

$$\text{normalized UL-DL carrier spacing} = \frac{|f_{\text{dl}} - f_{\text{ul}}|}{(f_{\text{dl}} + f_{\text{ul}})/2} = 0.0942,$$

which is small. Therefore, it is reasonable to assume that the autocorrelations and cross-correlation of the angular gain processes remain the same in UL and DL, i.e.

$$\mathbb{E}\left[W_{\mathsf{H}}^{\text{dl}}(\xi)W_{\mathsf{H}}^{\text{dl}\,*}(\xi')\right] = \gamma_{\mathsf{H}}(\xi)\delta(\xi-\xi'), \tag{24a}$$

$$\mathbb{E}\left[W_{\mathsf{V}}^{\text{dl}}(\xi)W_{\mathsf{V}}^{\text{dl}\,*}(\xi')\right] = \gamma_{\mathsf{H}}(\xi)\delta(\xi-\xi'), \tag{24b}$$

and $\mathbb{E}\left[W_{\mathsf{H}}^{\text{dl}}(\xi)W_{\mathsf{V}}^{\text{dl}\,*}(\xi')\right] = \rho(\xi)\delta(\xi-\xi')$. As explained before, this means that the DP-ASF remains the same in UL and DL, a property known as angular channel reciprocity [11, 14, 22]. While the DP-ASF is the same for UL and DL, the covariances are still different. Similar to the calculation in (7), the DL channel covariance can be expressed as

$$\mathbf{\Sigma}_{\mathbf{h}}^{\text{dl}} = \mathbb{E}\left[\mathbf{h}_{\text{dl}}\mathbf{h}_{\text{dl}}^{\mathsf{H}}\right] = \int_{-1}^{1} \mathbf{\Gamma}(\xi)\otimes\mathbf{A}_{\text{dl}}(\xi)d\xi, \tag{25}$$

where $\mathbf{A}_{\text{dl}}(\xi) = \mathbf{a}_{\text{dl}}(\xi)\mathbf{a}_{\text{dl}}(\xi)^{\mathsf{H}}$.

One way to estimate the DL covariance is to send pilots from the BS to the user, receive measurements via closed-loop feedback from the user and estimate the instantaneous channels. Then the BS can accumulate these estimates and compute the sample covariance. Due to the high channel dimension, this strategy incurs a large overhead both because estimating a single channel requires sending many pilots and that the sample covariance requires many samples to converge. We solve this issue by estimating the DL covariance from the UL pilots in (9) and leveraging angular channel reciprocity. With this approach DL covariance estimation does not require additional transmission of DL pilots, and in this sense is estimated for free from the naturally received UL pilots. Recall that in the previous section we estimated the DP-ASF from UL pilots, denoted as $\widehat{\mathbf{\Gamma}}$ in (21). Plugging $\widehat{\mathbf{\Gamma}}$ instead of $\mathbf{\Gamma}$ in (25) we estimate the DL covariance as

$$\mathbf{\Sigma}_{\mathbf{h}}^{\text{dl}\,\star} = \int_{-1}^{1} \widehat{\mathbf{\Gamma}}(\xi)\otimes\mathbf{A}_{\text{dl}}(\xi)d\xi = \sum_{i=1}^{n+\widehat{r}} \widehat{\mathbf{W}}_i\otimes\mathbf{S}_i^{\text{dl}}, \tag{26}$$

where $\mathbf{S}_i^{\text{dl}} = \mathbf{A}_{\text{dl}}(\widehat{\xi}_i)$ for $i = 1,\ldots,\widehat{r}$ and $\mathbf{S}_i^{\text{dl}} = \int_{-1}^{1}\psi_i(\xi)\mathbf{A}_{\text{dl}}(\xi)d\xi$ for $i = \widehat{r}+1,\ldots,\widehat{r}+n$.

V. Downlink Channel Training and Multi-User Precoding

Besides the problem of covariance estimation, the BS is required to transmit multiplexed data to several users in the DL. An interference-free transmission is possible only if the BS has access to the instantaneous DL channel state information (CSI) for all users to construct a precoder. Since channel reciprocity does not hold in FDD mode, the instantaneous DL CSI is obtained via common DL training (pilot transmission) of the user channels and feeding back the measurements to the BS during UL. The challenge is that, for a dual-polarized massive MIMO system with a channel dimension of $2M \gg 1$, the number of pilots used for DL training must be large to enable channel estimation. This results in a substantial reduction of DL sum-rate, since a large portion of the resources is spent on channel training and not data transmission. Also feeding back the measurements to the BS consumes a considerable part of resources in UL and may result in large delays.

To address this issue, we propose a common training scheme that enables DL data multiplexing for any pilot dimension. The point is that the BS does *not* need to estimate the full-dimensional channel in order to be able to multiplex data to the users. It is sufficient that it learns the *effective* user channels, namely a low-dimensional projection of the channels. The effective channels are obtained as a concatenation of the true channel with a *sparsifying precoder* that depends only on the user channel covariances. Here we propose a design of this precoder for dual-polarized channels, based on the idea of *active channel sparsification* (ACS) that was formerly developed for single-polarized channels in [13].

We can formalize the idea behind ACS as follows. To jointly train the DL channels, the BS transmits a pilot matrix $\mathbf{\Psi}$ of dimension $T_{\text{dl}}\times M'$, where T_{dl} is a fixed pilot dimension such that each row $\mathbf{\Psi}_{i,:}$ represents a pilot signal that is transmitted from the $M' \leq 2M$ inputs of a precoding matrix $\mathbf{B}$ of dimension $M'\times 2M$. The integer M' is a suitable intermediate dimension that, as we will see later, is determined within the design. The observed training symbols at user k can be expressed via the T_{dl}-dimensional vector

$$\mathbf{y}_{\text{dl},k} = \mathbf{\Psi}\mathbf{B}\mathbf{h}_{\text{dl},k} + \mathbf{z}_k = \mathbf{\Psi}\widetilde{\mathbf{h}}_{\text{dl},k} + \mathbf{z}_k, \tag{27}$$

where $\mathbf{h}_{\text{dl},k}$ is the DL channel vector of user k for $k = 1,\ldots,K$, $\mathbf{z}_k \sim \mathcal{CN}(\mathbf{0}, N_0\mathbf{I}_{T_{\text{dl}}})$ is the AWGN with element-wise variance N_0, and the pilot and precoding matrices are normalized such that $\text{tr}(\mathbf{\Psi}\mathbf{B}\mathbf{B}^{\mathsf{H}}\mathbf{\Psi}^{\mathsf{H}}) = T_{\text{dl}}P_{\text{dl}}$, where P_{dl} is the BS transmit power resulting in the DL signal-to-noise ratio (SNR) to be equal to $\text{SNR} = \frac{P_{\text{dl}}}{N_0}$. In (27) we have also defined the effective channel vector $\widetilde{\mathbf{h}}_{\text{dl},k} := \mathbf{B}\mathbf{h}_{\text{dl},k}$ as the concatenation of the precoder with the true channel. In the ACS method, our intention is to design $\mathbf{B}$ as a sparsifying precoder, such that each vector $\widetilde{\mathbf{h}}_{\text{dl},k}$, $k = 1,\ldots,K$ is sufficiently sparse and yet $\widetilde{\mathbf{H}} = [\mathbf{h}_{\text{dl},1},\ldots,\mathbf{h}_{\text{dl},K}]$ forms an effective channel matrix with a rank that is as large as possible. In this way, each effective channel can be estimated using the fixed (and possibly small) pilot overhead T_{dl}, but the BS is still able to transmit multiple data streams in the DL.

A. Necessity of Channel Sparsification

The channel vector of user k admits the Karhunen-Loève (KL) expansion $\mathbf{h}_{\text{dl},k} = \sum_{m=1}^{2M} g_{k,m}\sqrt{\lambda_{k,m}}\,\mathbf{u}_m^{(k)}$, where $g_{k,m} \sim \mathcal{CN}(0,1)$ are i.i.d. complex Gaussian variables, $\mathbf{u}_m^{(k)}$ is the m-th eigenvector of user k DL channel covariance and $\lambda_{k,m}$ is its associated eigenvalue. Define the vector of eigenvalues of user k as $\boldsymbol{\lambda}_k \in \mathbb{R}_+^{2M}$ and define the support of $\boldsymbol{\lambda}_k$ as $\mathcal{S}_k = \{m : \lambda_{k,m} \neq 0\}$ with a size $s_k = |\mathcal{S}_k|$, which specifies the covariance rank. The following lemma yields necessary and sufficient conditions for the stable estimation of $\mathbf{h}_{\text{dl},k}$, where by estimation stability we mean that the estimation error vanishes as the noise variance tends to zero. The proof can be found in [13].

Lemma *1:* Consider the channel vector $\mathbf{h}_{\text{dl},k}$ with support set $\mathcal{S}_k$. Let $\widehat{\mathbf{h}}_{\text{dl},k}$ denote any estimator for $\mathbf{h}_{\text{dl},k}$ based on the observation $\mathbf{y}_{\text{dl},k} = \mathbf{\Psi}\mathbf{h}_{\text{dl},k} + \mathbf{z}_k$ (note that this coincides with (27) by replacing $\mathbf{B} = \mathbf{I}_{2M}$, i.e., without the sparsifying precoder). Let $\mathbf{R}_e = \mathbb{E}[(\mathbf{h}_{\text{dl},k} - \widehat{\mathbf{h}}_{\text{dl},k})(\mathbf{h}_{\text{dl},k} - \widehat{\mathbf{h}}_{\text{dl},k})^{\mathsf{H}}]$

denote the corresponding estimation error covariance matrix. If $T_{\text{dl}} \geq s_k$ there exist pilot matrices $\mathbf{\Psi} \in \mathbb{C}^{T_{\text{dl}} \times 2M}$ for which $\lim_{N_0 \downarrow 0} \text{tr}(\mathbf{R}_e) = 0$ for all support sets $\mathcal{S}_k : |\mathcal{S}_k| = s_k$. Conversely, for any support set $\mathcal{S}_k : |\mathcal{S}_k| = s_k$ any pilot matrix $\mathbf{\Psi} \in \mathbb{C}^{T_{\text{dl}} \times 2M}$ with $T_{\text{dl}} < s_k$ yields $\lim_{N_0 \downarrow 0} \text{tr}(\mathbf{R}_e) > 0$. □

Stable channel estimation is necessary in order to achieve high spectral efficiency in the high-SNR regime. In fact, if the estimation mean-squared error (MSE) of the user channels does not vanish as $N_0 \downarrow 0$, the system self-interference due to imperfect channel knowledge grows proportionally to the signal power and we have an interference-limited multi-user system, which is undesirable. An implication of Lemma 1 is that, if $T_{\text{dl}} < s_k$ for some user k, then any scheme that relies on channel sparsity will fail to yield a stable channel estimate. This includes, for example, the sophisticated compressed sensing (CS) methods, which simply can not stably estimate a s_k-sparse channel from $T_{\text{dl}} < s_k$ measurements. Therefore, one constraint for designing the sparsifying precoder $\mathbf{B}$ is that, once it is applied to the channel vector, the sparsity of the resulting effective channel is less than or equal to the available pilot dimension T_{dl}.

B. Virtual Beam Representation

From the discussion above, we conclude that the user channels should be sparsified, i.e. they should be transformed such that their KL expansion contains fewer non-zero eigenvalues. In a multi-user setup this should be done for every user channel. On the other hand, the KL expansion of the channels are generally different from each other and there is no common eigenbasis shared among them. In this case it is extremely difficult to design a precoder that jointly sparsifies the user channels. Fortunately, in the case of dual-polarized ULA channels there exists an approximate common eigenbasis that enables joint sparsification.

To see this, note that we can express a dual-polarized ULA covariance as the block matrix

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{h}_{\text{dl}} \mathbf{h}_{\text{dl}}^{\mathsf{H}}] = \begin{bmatrix} \mathbf{\Sigma}_{\mathsf{HH}} & \mathbf{\Sigma}_{\mathsf{HV}} \\ \mathbf{\Sigma}_{\mathsf{VH}} & \mathbf{\Sigma}_{\mathsf{VV}} \end{bmatrix}, \tag{28}$$

where $\mathbf{\Sigma}_{\mathsf{HH}} = \mathbb{E}[\mathbf{h}_{\text{dl,H}} \mathbf{h}_{\text{dl,H}}^{\mathsf{H}}]$, $\mathbf{\Sigma}_{\mathsf{VV}} = \mathbb{E}[\mathbf{h}_{\text{dl,V}} \mathbf{h}_{\text{dl,V}}^{\mathsf{H}}]$, and $\mathbf{\Sigma}_{\mathsf{HV}} = \mathbf{\Sigma}_{\mathsf{VH}}^{\mathsf{H}} = \mathbb{E}[\mathbf{h}_{\text{dl,H}} \mathbf{h}_{\text{dl,V}}^{\mathsf{H}}]$, where $\mathbf{h}_{\text{dl,H}}$ and $\mathbf{h}_{\text{dl,V}}$ are generic H and V DL channel vectors, respectively. The diagonal blocks $\mathbf{\Sigma}_{\mathsf{HH}}$ and $\mathbf{\Sigma}_{\mathsf{VV}}$ are Hermitian Toeplitz matrices of dimension M. The well-known Szegö theorem states that for a Hermitian Toeplitz matrix of dimension $M \gg 1$, there exists a circulant matrix that approximately has the same eigenvalue distribution as the Toeplitz matrix [23, 24]. On the other hand the eigenvectors of a circulant matrix are given by the columns of the DFT matrix of the same size. It follows that, for large dimensions, the DFT matrix approximately diagonalizes the Toeplitz covariance. More concretely, we can compute the circulant approximation of $\mathbf{\Sigma}_{\mathsf{HH}}$ and $\mathbf{\Sigma}_{\mathsf{VV}}$ by first defining the vectors $\mathring{\boldsymbol{\lambda}}_{\mathsf{H}}, \mathring{\boldsymbol{\lambda}}_{\mathsf{V}} \in \mathbb{R}^M$ as

$$[\mathring{\boldsymbol{\lambda}}_{\mathsf{H}}]_m = [\mathbf{F}^{\mathsf{H}} \mathbf{\Sigma}_{\mathsf{HH}} \mathbf{F}]_{m,m}, \quad [\mathring{\boldsymbol{\lambda}}_{\mathsf{V}}]_m = [\mathbf{F}^{\mathsf{H}} \mathbf{\Sigma}_{\mathsf{VV}} \mathbf{F}]_{m,m} \tag{29}$$

where $\mathbf{F} \in \mathbb{C}^{M \times M}$ is the DFT matrix with elements $[\mathbf{F}]_{m,n} = \frac{1}{\sqrt{M}} e^{j2\pi \frac{(m-1)(n-1)}{M}}$, $m, n = 1, 2, \ldots, M$. Then we can express the circulant approximations by $\mathring{\mathbf{\Sigma}}_{\mathsf{HH}} = \mathbf{F}\, \text{diag}\left(\mathring{\boldsymbol{\lambda}}_{\mathsf{H}}\right) \mathbf{F}^{\mathsf{H}}$ and $\mathring{\mathbf{\Sigma}}_{\mathsf{VV}} = \mathbf{F}\, \text{diag}\left(\mathring{\boldsymbol{\lambda}}_{\mathsf{V}}\right) \mathbf{F}^{\mathsf{H}}$. From the Szegö theorem we have $\mathbf{\Sigma}_{\mathsf{HH}} \approx \mathbf{F}\, \text{diag}\left(\mathring{\boldsymbol{\lambda}}_{\mathsf{H}}\right) \mathbf{F}^{\mathsf{H}}$, and $\mathbf{\Sigma}_{\mathsf{VV}} \approx \mathbf{F}\, \text{diag}\left(\mathring{\boldsymbol{\lambda}}_{\mathsf{V}}\right) \mathbf{F}^{\mathsf{H}}$. It follows that the H and V channel vectors admit a (approximate) representation over the columns of $\mathbf{F} = [\mathbf{f}_0, \ldots, \mathbf{f}_{M-1}]$ as $\mathbf{h}_{\mathsf{H}} \approx \mathbf{F}\mathbf{g}_{\mathsf{H}}$, $\mathbf{h}_{\mathsf{V}} \approx \mathbf{F}\mathbf{g}_{\mathsf{V}}$, where $\mathbf{g}_{\mathsf{H}} \sim \mathcal{CN}\left(\mathbf{0}, \text{diag}\left(\mathring{\boldsymbol{\lambda}}_{\mathsf{H}}\right)\right)$ and $\mathbf{g}_{\mathsf{V}} \sim \mathcal{CN}\left(\mathbf{0}, \text{diag}\left(\mathring{\boldsymbol{\lambda}}_{\mathsf{V}}\right)\right)$ are i.i.d complex Gaussian vectors.

From the discussion above we conclude that the dual-polarized DL channel vector of the k-th user $\mathbf{h}_{\text{dl},k}$ is related to its corresponding channel coefficients as

$$\mathbf{h}_{\text{dl},k} \approx \widetilde{\mathbf{F}} \mathbf{g}_k, \ k = 1, \ldots, K \tag{30}$$

where we have defined $\widetilde{\mathbf{F}} := \mathbf{I}_2 \otimes \mathbf{F} \in \mathbb{C}^{2M}$. The columns of $\widetilde{\mathbf{F}}$ represent the set of common "virtual beams" for the dual-polarized channel among all users. For every m, the elements $[\mathbf{g}_{\mathsf{H},k}]_m$ and $[\mathbf{g}_{\mathsf{V},k}]_m$ can be in general correlated, due to the correlation between horizontal and vertical channels.

C. User-Virtual Beam Bipartite Graph

We now introduce a graphical model that encodes the *power profile* of each user along the common virtual beams. Define $\mathcal{G} = (\mathcal{V}, \mathcal{K}, \mathcal{E})$ as a bipartite graph with two color classes $\mathcal{V}$ and $\mathcal{K}$, where $\mathcal{V}$ is a node set of dimension $2M$, representing the set of virtual beams and $\mathcal{K}$ is a node set of dimension K, representing the users. Also $(k, v) \in \mathcal{E}$ if and only if $[\boldsymbol{\lambda}_k]_v > 0$, where $\boldsymbol{\lambda}_k := [\boldsymbol{\lambda}_{\mathsf{H},k}^{\mathsf{T}}, \boldsymbol{\lambda}_{\mathsf{V},k}^{\mathsf{T}}]^{\mathsf{T}}$. Therefore, the *biadjacency* matrix of this graph is given by a $2M \times K$ binary matrix $\mathbf{A}$ for which $[\mathbf{A}]_{v,k} = 1$ if and only if $(v, k) \in \mathcal{E}$.

Recall that the DL channel matrix $\mathbf{H} = [\mathbf{h}_{\text{dl},1}, \ldots, \mathbf{h}_{\text{dl},K}] \in \mathbb{C}^{2M \times K}$ is related to the matrix of "angular" channel gains $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_K] \in \mathbb{C}^{2M \times K}$ as $\mathbf{H} = \widetilde{\mathbf{F}}\mathbf{G}$, where each column of $\widetilde{\mathbf{F}}$ is a virtual beam vector. Particularly interesting is the relation between $\mathbf{H}$ and the bipartite graph $\mathcal{G}$, summarized in Theorem 1. The following definition and lemma are required for proving this theorem.

Definition *1:* [matching] A matching is a set of edges in a graph that do not share any endpoints. A perfect matching is a matching that connects all nodes of the graph [25].

Lemma *2:* [Rank and perfect matchings] Let $\mathbf{Q}$ denote an $r \times r$ matrix with some elements identically zero, and the non-identically zero elements drawn from a continuous distribution, such that an element $[\mathbf{Q}]_{i,j}$ is independent from all elements that are not in the same row or column with it (it may or may not be dependent on elements in the same row or same column). Consider a bipartite graph $\mathcal{Q}$ with biadjacency matrix $\widetilde{\mathbf{A}}$ such that $[\widetilde{\mathbf{A}}]_{i,j} = 1$ if $[\mathbf{Q}]_{i,j}$ is not identically zero, and $[\widetilde{\mathbf{A}}]_{i,j} = 0$ otherwise. Then, $\mathbf{Q}$ has rank r with probability 1 if and only if $\mathcal{Q}$ contains a perfect matching. □

Proof: The determinant of $\mathbf{Q}$ is given by the expansion $\det(\mathbf{Q}) = \sum_{\iota \in \boldsymbol{\pi}_r} \text{sgn}(\iota) \prod_i [\mathbf{Q}]_{i,\iota(i)}$, where ι is a permutation of the set $\{1, 2, \ldots, r\}$, where $\boldsymbol{\pi}_r$ is the set of all such permutations and where $\text{sgn}(\iota)$ is either 1 or -1. From the construction of $\mathcal{Q}$, it is clear that the product $\prod_i [\mathbf{Q}]_{i,\iota(i)}$ is non-zero only if the edge subset $\{(i, \iota(i)), i = 1, \ldots, r)\}$ is a perfect matching. Hence, if $\mathcal{Q}$ contains a perfect matching,

then $\det(\mathbf{Q}) \neq 0$ with probability 1 (and $\text{rank}(\mathbf{Q}) = r$), since the non-identically zero entries of $\mathbf{Q}$ are drawn from a continuous distribution, such that all elements involved in the product $\prod_i [\mathbf{Q}]_{i,\iota(i)}$ are independent (no two elements from either the same row or the same column are involved in this product). If it does not contain a perfect matching, then $\det(\mathbf{Q}) = 0$ and therefore $\text{rank}(\mathbf{Q}) < r$. ■

Theorem *1:* The rank of $\mathbf{H}$ is given, with probability 1, by the side-length of the largest square intersection sub-matrix whose associated sub-graph in $\mathcal{G}$ contains a perfect matching. □

Proof: Note that since $\mathbf{H} = \widetilde{\mathbf{F}}\mathbf{G}$ and $\widetilde{\mathbf{F}}$ is unitary, the rank of $\mathbf{H}$ is equal to that of $\mathbf{G}$. In addition, the rank of $\mathbf{G}$ is equivalent to the largest order of any non-zero minor in $\mathbf{G}$,[4] i.e. the side-length of the largest non-singular square sub-matrix of $\mathbf{G}$. The elements of $\mathbf{G}$ are either identically zero or drawn from a Gaussian distribution with zero mean and a variance $[\boldsymbol{\lambda}_k]_m$ for some $(m, k) \in [2M] \times [K]$. We also know that an element in $\mathbf{Q}$ can be correlated with (at most) one element in the same column. Now, according to Lemma 2 any such sub-matrix $\mathbf{Q}$ is non-singular (has rank equal to its side-length) if and only if its associated sub-graph $\mathcal{Q} \subseteq \mathcal{G}$ contains a perfect matching. This concludes the proof. ■

Theorem 1 implies that the rank of the channel matrix is given, with probability 1, by the size of a certain matching in the user-virtual beam bipartite graph $\mathcal{G}$. This matching is contained in a sub-graph of $\mathcal{G}$ that specifies the selected users and virtual beams. In particular, we want to maximize the size of this matching, since it corresponds to the rank of the channel matrix and therefore the system multiplexing gain. On the other hand, the estimation stability criterion induces a restriction on all candidate sub-graphs: for each user in the selected sub-graph, the number of virtual beams connected to it through an edge must be less than the pilot dimension T_{dl} so that stable channel estimation is possible according to Lemma 1 (*stability constraint*). Also the *effective* channel of a selected user should be sufficiently "strong" so that beamforming data to that user is reasonable and does not compromise the total sum-rate (*power constraint*). The trade-off between the rank objective and these constraints results in an optimization problem, solving which yields the desired sparsifying precoder.

D. Active Channel Sparsification

Let $\mathcal{G} = (\mathcal{V}, \mathcal{K}, \mathcal{E})$ denote the user-virtual beam bipartite graph as previously defined and let $\mathcal{M}(\mathcal{V}', \mathcal{K}')$ denote a matching in $\mathcal{G}$ involving node subsets $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{K}' \subseteq \mathcal{K}$. Take T_{dl} to be the available DL pilot dimension. Then, maximizing the effective channel matrix rank subject to stability and power constraints can be cast as the following optimization problem:

$$\underset{\mathcal{V}' \subseteq \mathcal{V}, \mathcal{K}' \subseteq \mathcal{K}}{\text{maximize}} \quad |\mathcal{M}(\mathcal{V}', \mathcal{K}')| \tag{31a}$$

$$\text{subject to} \quad \deg_{\mathcal{G}'}(k) \leq T_{\text{dl}} \qquad \forall\ k \in \mathcal{K}', \tag{31b}$$

$$\sum_{m \in \mathcal{N}_{\mathcal{G}'}(k)} [\mathbf{W}]_{m,k} \geq P_0, \qquad \forall\ k \in \mathcal{K}', \tag{31c}$$

where $\mathcal{G}'$ is an induced sub-graph of $\mathcal{G}$ corresponding to the node subsets $\mathcal{V}'$ and $\mathcal{K}'$, $\deg_{\mathcal{G}'}(k)$ denotes the degree of node k in $\mathcal{G}'$, $\mathcal{N}_{\mathcal{G}'}(k)$ denotes the set of nodes that are neighbor to k in $\mathcal{G}'$, and $P_0 \geq 0$ is a predefined power threshold. Constraint (31b) ensures that the number of virtual beams contributing to the effective channel of user k is less than or equal the pilot dimension T_{dl}. Recall that this number is equal to the effective channel dimension and therefore the constraint satisfies the estimation stability condition (see Lemma 1). Constraint (31c) is a power constraint, which ensures that if a user is chosen to be served (i.e., is in the solution sub-graph), then it should have sufficient total power (at least P_0) along the selected virtual beams.

The optimization in (31) can be seen as a specific selection of the induced su-graph $\mathcal{G}'$ that contains a "large" perfect matching. We can model this selection process by assigning binary variables to the nodes of the user-virtual beam bipartite graph and solving a mixed integer linear program (MILP). This program is given as follows:

$$\underset{x_m, y_k, z_{m,k}}{\text{maximize}} \quad \sum_{m \in \mathcal{V}} \sum_{k \in \mathcal{K}} z_{m,k} + \delta \sum_{m \in \mathcal{V}} x_m \tag{32a}$$

$$\text{subject to} \quad z_{m,k} \leq [\mathbf{A}]_{m,k} \quad \forall m \in \mathcal{V}, k \in \mathcal{K}, \tag{32b}$$

$$\sum_{k \in \mathcal{K}} z_{m,k} \leq x_m \quad \forall m \in \mathcal{V}, \tag{32c}$$

$$\sum_{m \in \mathcal{V}} z_{m,k} \leq y_k \quad \forall k \in \mathcal{K}, \tag{32d}$$

$$\sum_{m \in \mathcal{V}} [\mathbf{A}]_{m,k} x_m \leq T_{\text{dl}} y_k + 2M(1 - y_k) \ \forall k \in \mathcal{K} \tag{32e}$$

$$P_0\, y_k \leq \sum_{m \in \mathcal{V}} [\mathbf{W}]_{m,k} x_m \quad \forall k \in \mathcal{K}, \tag{32f}$$

$$x_m \leq \sum_{k \in \mathcal{K}} [\mathbf{A}]_{m,k} y_k \quad \forall m \in \mathcal{V}, \tag{32g}$$

$$x_m, y_k \in \{0, 1\} \quad \forall m \in \mathcal{V}, k \in \mathcal{K}, \tag{32h}$$

$$z_{m,k} \in [0, 1] \quad \forall m \in \mathcal{V}, k \in \mathcal{K}, \tag{32i}$$

where $0 < \delta < \frac{1}{2M}$ is a small positive scalar. The binary variables $\{x_m\}_{m=1}^{2M}$ represent the virtual beams and the binary variables $\{y_k\}_{k=1}^{K}$ represent the users. The solution is given by the set of nodes $\mathcal{V}^\star = \{m : x_m^\star = 1\}$ and $\mathcal{K}^\star = \{k : y_k^\star = 1\}$, with $\{x_m^\star\}_{m=1}^{2M}$ and $\{y_k^\star\}_{k=1}^{K}$ being a solution of (32).

The MILP introduced in (32) can be solved for most practical array dimensions (for example, up to $M = 128$) using standard solvers. We have used the built-in "intlinprog" routine in MATLAB to perform our simulations, provided in Section VI. The solution of (32) determines the set of users as well as virtual beams that are to be probed and served: a user k is probed and served if and only if $y_k^\star = 1$; similarly, a virtual beam m is probed and served if and only if $x_m^\star = 1$. Fig. 5 provides a miniature example, in which we have $K = 2$ users, $2M = 6$ virtual beams and $T_{\text{dl}} = 2$. Here the maximum matching size is equal to two, and by omitting beams number 2 and 5 (red crosses), the MILP satisfies the constraint (31b),

[4] a minor of $\mathbf{G}$ is the determinant of some square sub-matrix of $\mathbf{G}$.

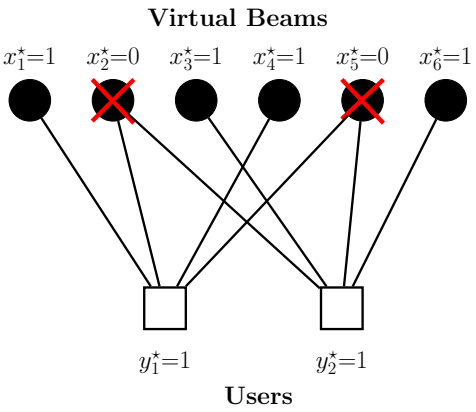


Fig. 5: an example of a user-virtual beam bipartite graph with $K = 2$ users and $2M = 6$ virtual beams. The red crosses denote inactive (i.e., eliminated) beams after solving the MILP with $T_{\text{dl}} = 2$.

since now each user is connected to 2 ($\leq T_{\text{dl}} = 2$) active beams.

E. Common DL Channel Training and Multi-User Precoding

Using the MILP solution, let us define $\mathcal{V}^\star = \{m : x_m^* = 1\} := \{m_1, m_2, \ldots, m_{M'}\}$ as the set of M' "active" virtual beams (with cardinality $|\mathcal{V}^\star| = M'$) and $\mathcal{K}^\star = \{k : y_k^* = 1\} := \{k_1, k_2, \ldots, k_{K'}\}$ as the set of K' active users. We design the sparsifying precoding matrix in (27) as

$$\mathbf{B} = \widetilde{\mathbf{F}}_{\mathcal{V}^\star}^{\mathsf{H}}, \tag{33}$$

where $\widetilde{\mathbf{F}}_{\mathcal{V}^\star}$ is the $2M \times M'$ matrix consisting of the columns of $\widetilde{\mathbf{F}}$ whose indices are in $\mathcal{V}^\star$. The effective DL channel vector of user k is given by the concatenation of this precoder with the full-dimensional channel, so that we have $\widetilde{\mathbf{h}}_{\text{dl},k} = \mathbf{B}\mathbf{h}_{\text{dl},k} \approx \widetilde{\mathbf{F}}_{\mathcal{V}^\star}^{\mathsf{H}} \widetilde{\mathbf{F}} \mathbf{g}_k$, where the approximation is only due to the approximate virtual beam representation in (30). It is easy to show that, the vector $\widetilde{\mathbf{h}}_{\text{dl},k}$ is of dimension M', and has significantly large components only over a subset of $\{1, 2, \ldots, M'\}$ determined by the intersection of $\mathcal{V}^\star$ and the support of $\mathbf{g}_k$, i.e. by $\mathcal{V}^\star \cap \mathcal{J}_k$. Recall that satisfying constraint (31b) ensures that $|\mathcal{V}^\star \cap \mathcal{J}_k| \leq T_{\text{dl}}$, so that one can stably recover the effective channel vector by taking T_{dl} linearly independent pilot measurements via the matrix $\mathbf{\Psi}$ (see Lemma 1). A convenient choice is to let the DL pilot matrix $\mathbf{\Psi}$ to be proportional to a random unitary matrix of dimension $T_{\text{dl}} \times M'$, such that $\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{H}} = P_{\text{dl}} \mathbf{I}_{T_{\text{dl}}}$. Once user k collects its pilot signal measurements in the form of the T_{dl}-dimensional vector $\mathbf{y}_{dl,k}$, it feeds them back to the BS in T_{dl} UL channel uses via analog unquantized feedback (this type of feedback is analyzed in e.g. [26, 27]). Upon receiving the noisy pilot measurements $\mathbf{y}_{\text{dl},k} = \mathbf{\Psi}\mathbf{B}\mathbf{h}_{\text{dl},k} + \mathbf{z}_k$ for any user $k \in \{1, \ldots, K\}$, the BS can obtain the minimum mean squared error (MMSE) estimate of the $2M$-dimensional DP channel $\mathbf{h}_{\text{dl},k}$ as

$$\widehat{\mathbf{h}}_{\text{dl},k} = \mathbf{\Sigma}_{\mathbf{h}\mathbf{y},k} \mathbf{\Sigma}_{\mathbf{y}\mathbf{y},k}^{-1} \mathbf{y}_{\text{dl},k}, \tag{34}$$

where $\mathbf{\Sigma}_{\mathbf{h}\mathbf{y},k} = \mathbb{E}\left[\mathbf{h}_{\text{dl},k}\mathbf{y}_{\text{dl},k}^{\mathsf{H}}\right] = \mathbf{\Sigma}_k^{\text{dl}} \mathbf{B}^{\mathsf{H}} \mathbf{\Psi}^{\mathsf{H}}$ and $\mathbf{\Sigma}_{\mathbf{y}\mathbf{y},k} = \mathbb{E}\left[\mathbf{y}_{\text{dl},k}\mathbf{y}_{\text{dl},k}^{\mathsf{H}}\right] = \mathbf{\Psi}\mathbf{B}\mathbf{\Sigma}_k^{\text{dl}}\mathbf{B}^{\mathsf{H}}\mathbf{\Psi}^{\mathsf{H}} + N_0 \mathbf{I}_{T_{\text{dl}}}$.

F. Beamforming and Data Transmission

Without loss of generality, let us assume that the BS wants to serve the first K' users, using a beamforming scheme that is ideally interference-free. We consider zero-forcing beamforming (ZFBF) for this purpose, where the ZFBF matrix $\mathbf{V}_{\mathsf{ZF}}$ is given by the column-normalized version of the Moore-Penrose pseudoinverse of the estimated effective channel matrix defined as $\widehat{\mathbf{H}}_{\mathsf{eff}} = \mathbf{B}\widehat{\mathbf{H}} = \mathbf{B}\left[\widehat{\mathbf{h}}_{\text{dl},1}, \widehat{\mathbf{h}}_{\text{dl},2}, \ldots, \widehat{\mathbf{h}}_{\text{dl},K'}\right] \in \mathbb{C}^{M' \times K'}$, so that we have $\mathbf{V}_{\mathsf{ZF}} = \widehat{\mathbf{H}}_{\mathsf{eff}}^{\dagger} \mathbf{J}^{1/2}$, where $\widehat{\mathbf{H}}_{\mathsf{eff}}^{\dagger} = \widehat{\mathbf{H}}_{\mathsf{eff}} \left(\widehat{\mathbf{H}}_{\mathsf{eff}}^{\mathsf{H}} \widehat{\mathbf{H}}_{\mathsf{eff}}\right)^{-1}$ and $\mathbf{J}$ is a diagonal matrix, normalizing the columns of $\mathbf{V}_{\mathsf{ZF}}$. A channel use of the DL precoded data transmission phase at the k-th user receiver takes on the form

$$r_k = \mathbf{h}_{\text{dl},k}^{\mathsf{H}} \mathbf{B}^{\mathsf{H}} \mathbf{V}_{\mathsf{ZF}} \mathbf{P}^{1/2} \mathbf{s} + n_k, \tag{35}$$

where $\mathbf{s} \in \mathbb{C}^{K' \times 1}$ is a vector of unit-energy user data symbols, $\mathbf{P}$ is a diagonal matrix defining the power allocation to the DL data streams and $n_k \sim \mathcal{CN}(\mathbf{0}, N_0)$ is the AWGN. The transmit power constraint is given by $\text{tr}(\mathbf{B}^{\mathsf{H}} \mathbf{V}_{\mathsf{ZF}} \mathbf{P} \mathbf{V}_{\mathsf{ZF}}^{\mathsf{H}} \mathbf{B}) = \text{tr}(\mathbf{V}_{\mathsf{ZF}}^{\mathsf{H}} \mathbf{V}_{\mathsf{ZF}} \mathbf{P}) = \text{tr}(\mathbf{P}) = P_{\text{dl}}$, where we used $\mathbf{B}\mathbf{B}^{\mathsf{H}} = \mathbf{I}_{M'}$ and the fact that $\mathbf{V}_{\mathsf{ZF}}^{\mathsf{H}} \mathbf{V}_{\mathsf{ZF}}$ has unit diagonal elements by construction. We use the simple uniform power allocation $[\mathbf{P}]_{k,k} = \frac{P_{\text{dl}}}{K'}$ to each k-th user data stream. The received symbol at user k receiver is given by $r_k = b_{k,k}\mathbf{s}_k + \sum_{\ell \neq k} b_{k,\ell}\mathbf{s}_\ell + n_k$, where the coefficients $b_{k,1}, \ldots, b_{k,K'}$ are given by the elements of the $1 \times K'$ row vector $\mathbf{h}_{\text{dl},k}^{\mathsf{H}} \mathbf{B}\mathbf{V}\mathbf{P}^{1/2}$ in (35). In the presence of an accurate channel estimation we expect that $b_{k,k} \approx \sqrt{[\mathbf{J}]_{k,k}[\mathbf{P}]_{k,k}}$ and $b_{k,\ell} \approx 0$ for $\ell \neq k$. However, this is not a given, since in general there typically exists a non-negligible channel estimation error. For simplicity, in order to calculate the ergodic sum-rate, here we assume that the coefficients $b_{k,1}, \ldots, b_{k,K'}$ are known to the corresponding receiver k. Including the DL training overhead, this yields the rate expression (see [28]):

$$R_{\text{sum}} = \left(1 - \frac{T_{\text{dl}}}{T}\right) \sum_{k=1}^{K'} \mathbb{E}\left[\log\left(1 + \frac{|b_{k,k}|^2}{N_0 + \sum_{\ell \neq k} |b_{k,\ell}|^2}\right)\right]. \tag{36}$$

VI. Simulation Results

In this section, we empirically examine the performance of our scheme in different aspects of dual-polarized UL channel covariance estimation, UL-DL covariance transformation and common multi-user DL channel training and precoding. We compare the covariance estimation performance of our method with the sample covariance estimator in terms of the mean normalized Frobenius norm error, defined as

$$E_{\text{NF}} = \mathbb{E}\left\{\frac{||\mathbf{\Sigma}_{\mathbf{h}} - \widehat{\mathbf{\Sigma}}_{\mathbf{h}}||_{\mathsf{F}}}{||\mathbf{\Sigma}_{\mathbf{h}}||_{\mathsf{F}}}\right\}, \tag{37}$$

where $\mathbf{\Sigma}_{\mathbf{h}}$ is the true channel covariance and $\widehat{\mathbf{\Sigma}}_{\mathbf{h}}$ is its estimate and where the expectation is taken over several sources of randomness in the channel, namely, random ASFs, random channel realizations in the sample set and random additive noise.

We consider a BS equipped with a ULA of M antennas with $\lambda_{\text{ul}}/2$ spacing. To examine the covariance estimation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

performance, we suppose $N = 2\kappa M$ independent samples of the $2M$-dimensional dual-polarized channel are available, where $\kappa = \frac{N}{2M}$ denotes the ratio between the sample set size and the channel dimension. The number of density functions used to approximate the continuous DP-ASF in (15) is set to $n = 3M$. In order to produce semi-random Horizontal and Vertical ASFs we consider the following generative model:

$$\begin{aligned}\gamma_{\mathsf{H}}(\xi) = &\frac{\alpha}{|\mathcal{I}_1| + |\mathcal{I}_2|}\left(\mathtt{rect}_{\mathcal{I}_1}(\xi) + \mathtt{rect}_{\mathcal{I}_2}(\xi)\right) \\ &+ \frac{1-\alpha}{2}\left(\delta(\xi - \xi_1) + \delta(\xi - \xi_2)\right),\end{aligned} \tag{38}$$

where for an interval $\mathcal{I} \subset [-1, 1]$, we have defined the rectangular function as $\mathtt{rect}_{\mathcal{I}}(\xi) = 1$ for $\xi \in \mathcal{I}$ and $\mathtt{rect}_{\mathcal{I}}(\xi) = 0$ for $\xi \notin \mathcal{I}$. The intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ are subsets of $[-1, 1]$, each of length $|\mathcal{I}_1|$ and $|\mathcal{I}_2|$, respectively, where the lengths are chosen uniformly at random between 0.1 and 0.4, i.e. $|\mathcal{I}_j| \sim \mathbb{U}([0.1, 0.4])$, independently for $j = 1$ and $j = 2$. Besides, $\xi_1, \xi_2 \in [-1, 1]$ denote discrete AoAs, generated independently and uniformly at random over $[-1, 1]$. The scalar $\alpha \in [0, 1]$ denotes what we call the continuous-to-discrete ASF ratio. Basically, since $\int_{-1}^{1} \frac{1}{|\mathcal{I}_1|+|\mathcal{I}_2|}\left(\mathtt{rect}_{\mathcal{I}_1}(\xi) + \mathtt{rect}_{\mathcal{I}_2}(\xi)\right) d\xi = 1$ and $\int_{-1}^{1} \frac{1}{2}\left(\delta(\xi - \xi_1) + \delta(\xi - \xi_2)\right) d\xi = 1$, α controls the contribution of the continuous part versus the discrete part to the overall ASF: for $\alpha = 0$ we have a purely discrete ASF, for $\alpha = 1$ we have a purely continuous one and for $\alpha \in (0, 1)$ we have a mixture of the two. Similarly, we generate the vertical ASF as:

$$\begin{aligned}\gamma_{\mathsf{V}}(\xi) = &\frac{\alpha}{|\mathcal{I}'_1| + |\mathcal{I}'_2|}\left(\mathtt{rect}_{\mathcal{I}'_1}(\xi) + \mathtt{rect}_{\mathcal{I}'_2}(\xi)\right) \\ &+ \frac{1-\alpha}{2}\left(\delta(\xi - \xi'_1) + \delta(\xi - \xi'_2)\right),\end{aligned} \tag{39}$$

Since it is natural for the horizontal and vertical ASFs to overlap in their support, we assume the discrete AoAs to be the same, i.e. $\xi'_1 = \xi_1$ and $\xi'_2 = \xi_2$, and we assume $\mathcal{I}'_1$ and $\mathcal{I}'_2$ to be slightly shifted versions of $\mathcal{I}_1$ and $\mathcal{I}_2$ as $\mathcal{I}'_1 = \mathcal{I}_1 + 0.1$ and $\mathcal{I}'_2 = \mathcal{I}_2 + 0.1$. Finally, we assume the cross-correlation function $\rho(\xi)$ to take on the form $\rho(\xi) = \beta\sqrt{\gamma_{\mathsf{H}}(\xi)\,\gamma_{\mathsf{V}}(\xi)}$, where $\beta \in [0, 1]$ is a scalar that controls the cross-correlation level between H and V channels. This is a simplifying assumption on the form of $\rho(\xi)$, which does not undermine the generality of the DP-ASF, and satisfies the necessary condition $|\rho(\xi)|^2 \leq \gamma_{\mathsf{H}}(\xi)\gamma_{\mathsf{V}}(\xi)$ for the DP-ASF $\mathbf{\Gamma}(\xi)$ to be a PSD matrix-valued function for all $\xi \in [-1, 1]$. In addition, we can change the cross-correlation between H and V channels simply by changing β. The larger β is, the more correlated the polar channels are.

A. UL Covariance Estimation Error

The first experiment compares the UL covariance estimators. We consider a ULA of size $M = 32$. To perform a Monte-Carlo simulation, we generate 100 random DP-ASFs according to the model explained earlier. For each random DP-ASF, we generate N independent samples of the channel as $\mathbf{h}_{\text{ul}}(1), \ldots, \mathbf{h}_{\text{ul}}(N)$ and AWGN vectors $\mathbf{z}(1), \ldots, \mathbf{z}(N)$ to generate the noisy pilot signals $\mathbf{y}_{\text{ul}}(i) = \mathbf{h}_{\text{ul}}(i) + \mathbf{z}(i)$, $i = 1, \ldots, N$. We repeat this for 50 different realizations of channel and noise, each time estimating the covariance given pilot signals and computing the estimation error. Therefore, the UL covariance estimation error is eventually averaged over $100 \times 50 = 5000$ random instances to empirically compute the error metric in (37). Fig. 6 compares the normalized Frobenius norm error as a function of the sampling ratio (left figure) as well as the SNR (right figure). The error figures show that the method based on PSD-LS considerably improves estimation accuracy in comparison to the sample covariance estimator. The main reason is that, PSD-LS captures the structure of the dual-polarized covariance (see (19)): it enforces the Kronecker structure by adopting the parametric covariance form $\sum_{i=1}^{n+\widehat{r}} \mathbf{W}_i \otimes \mathbf{S}_i$ and it constraints the coefficients $\mathbf{W}_i$, $i = 1, \ldots, n + \widehat{r}$ to be PSD in accordance with the DP-ASF being a PSD matrix-valued function.

B. UL-DL Covariance Transformation Error

The second part of our proposed scheme involves UL to DL covariance transformation as explained in Section IV. Using the same simulation setup as introduced earlier, we study the DL covariance estimation error. In order to separately study the error of covariance transformation and that of UL covariance estimation from random channel samples, we consider two cases: in the first case we assume that the true UL covariance is given, perform the transformation and compute the error. In the second case, we assume that only the noisy pilot signals $\mathbf{y}_{\text{ul}}(1), \ldots, \mathbf{y}_{\text{ul}}(N)$ are given. Obviously, the estimation error is expected to be larger in the second case. Mathematically, in the first case we replace $\widehat{\mathbf{\Sigma}}^{\text{ul}}_{\mathbf{h}}$ with $\mathbf{\Sigma}^{\text{ul}}_{\mathbf{h}}$ in (19) and estimate the ASF parametric form, whereas in the second case we compute $\widehat{\mathbf{\Sigma}}^{\text{ul}}_{\mathbf{h}}$ as $\widehat{\mathbf{\Sigma}}^{\text{ul}}_{\mathbf{h}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_{\text{ul}}(i)\mathbf{y}_{\text{ul}}(i)^{\mathsf{H}} - N_0\mathbf{I}$. Finally, we also plot the error measures for UL covariance estimation from the noisy pilots to compare it to the other two other cases. Fig. 7 illustrates the error vs sampling ratio (left figure) and error vs SNR curves (right figure). The figures show that, given a precise estimate of the UL covariance, the DL covariance can be estimated with a low error. In other words, the dominant source of error lies not in the UL-DL covariance transformation module, but in estimating the UL covariance from noisy pilots. This shows how effective the covariance transformation algorithm is. It also points to the more reasonable way of estimating the DL covariance. Collecting DL channel samples and using them to estimate the DL covariance is inefficient since it consumes too many resources to gather enough channel samples for a precise estimate of the covariance, especially since DL pilot measurements must be sent to the BS via closed-loop feedback. Instead, the BS can take in a sufficiently high number of UL channel samples, accurately estimate the UL covariance and perform UL-DL covariance estimation to obtain the DL covariance with much less error.

C. Sum-Rate Assessment of ACS

The third part of the implementation developed in this work was dedicated to an efficient common DL channel training

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications

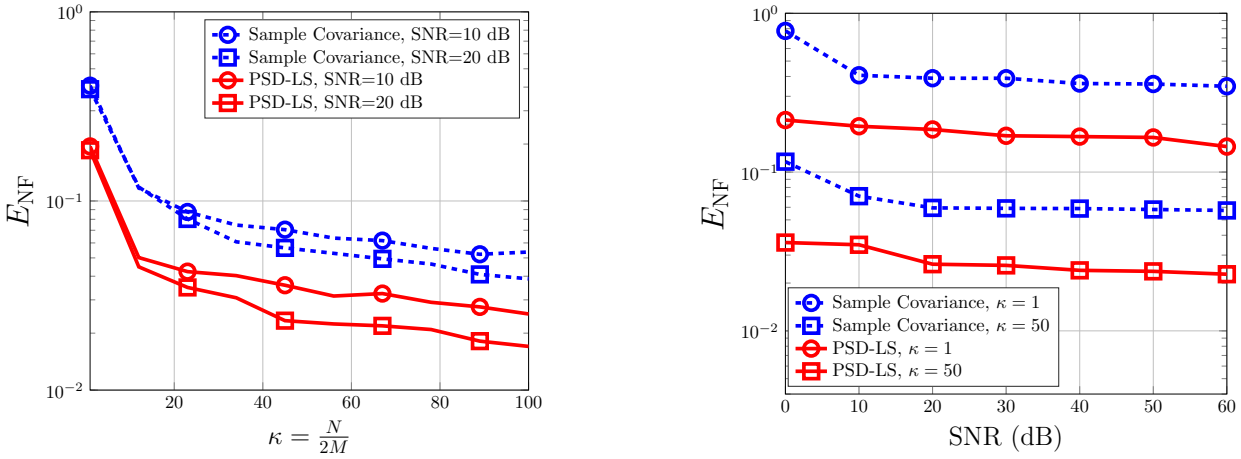


Fig. 6: Channel covariance estimation error vs the sampling ratio (left) and SNR (right) for $M = 32$.

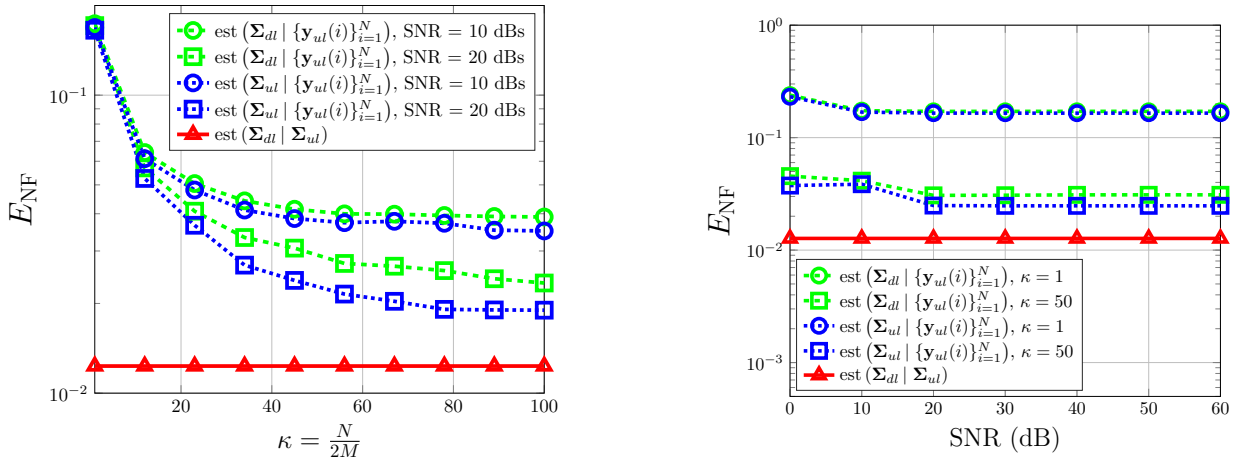


Fig. 7: Channel covariance transformation error vs sampling ratio (left) and SNR (right) with $M = 32$. est$(\mathbf{X}|\mathbf{Y})$ denotes the estimate of $\mathbf{X}$ given $\mathbf{Y}$.

and multi-user precoding. For any DL pilot dimension, the ACS approach enables the BS to stably estimate effective user channel vectors while maximizing the effective channel matrix rank. In this section we present results to study the performance of ACS in terms of sum-rate, for various DL pilot dimensions and SNR values. As a multi-user scenario, we consider K users, with covariances that are generated as follows. Define the four rectangular functions: $\mathcal{I}_1(\xi) = \mathtt{rect}_{[-0.8,-0.6]}$, $\mathcal{I}_2(\xi) = \mathtt{rect}_{[-0.45,-0.25]}$, $\mathcal{I}_3(\xi) = \mathtt{rect}_{[0.1,0.3]}$, $\mathcal{I}_4(\xi) = \mathtt{rect}_{[0.5,0.7]}$. Each of these functions represents angular power density of a single scatterer in the environment. We assume that the DP-ASF components of a single generic user are (semi-)randomly generated as

$$\gamma_{\mathsf{H}}(\xi) = \frac{\alpha}{Z}\left(\mathtt{rect}_{\mathcal{I}_i}(\xi) + \mathtt{rect}_{\mathcal{I}_j}(\xi)+\right) + \frac{1-\alpha}{2}(\delta(\xi-\xi_1)+\delta(\xi-\xi_2)), \tag{40}$$

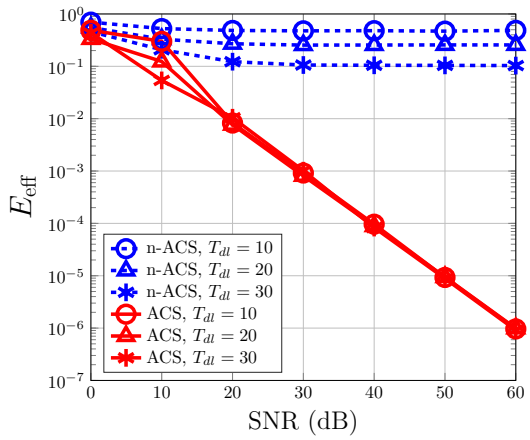
where $i, j \in \{1,2,3,4\}$ are uniformly generated random indices, $\alpha = 0.5$ is the continuous-to-discrete ASF ratio, Z is a normalizing scalar such that $\int_{-1}^{1} \frac{1}{Z}\left(\mathtt{rect}_{\mathcal{I}_i}(\xi) + \mathtt{rect}_{\mathcal{I}_i}(\xi)\right) d\xi = 1$, and ξ_1, ξ_2 are discrete AoAs, generated independently and uniformly at random over $[-1, 1]$. In order to generate the vertical ASF, similar to the previous section, we assume that the support of the continuous part of γ_{V} is a slightly shifted version of the support of the continuous part of γ_{H}. We also assume that they share the same support for their discrete part. Then we have

$$\gamma_{\mathsf{V}}(\xi) = \frac{\alpha}{Z}\left(\mathtt{rect}_{\mathcal{I}'_i}(\xi) + \mathtt{rect}_{\mathcal{I}'_j}(\xi)+\right) + \frac{1-\alpha}{2}(\delta(\xi-\xi'_1)+\delta(\xi-\xi'_2)), \tag{41}$$

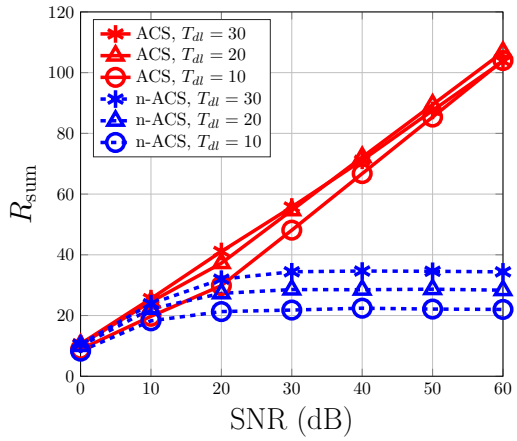
where $\mathcal{I}'_i = \mathcal{I}_i + 0.1$, $\mathcal{I}'_j = \mathcal{I}_j + 0.1$ and $\xi'_1 = \xi_1$, $\xi'_2 = \xi_2$. Besides, we suppose the H-V cross-correlation function to take on the form $\rho(\xi) = \beta\sqrt{\gamma_{\mathsf{H}}(\xi)\,\gamma_{\mathsf{V}}(\xi)}$.

Assuming a dual-polarized ULA with antennas, we generate semi-random DP-ASFs for K users as explained above. Then we compute their covariances using (7). In order to isolate the effect of sparsification from the other parts of the implementation (UL covariance estimation, UL-DL covariance transformation), we assume that the true DL covariance for each user is available at the BS. For a given DL pilot dimension, we implement ACS by designing the DL precoder via the MILP in (32) for common training and estimation of the effective channels. Next the users are served through a ZFBF

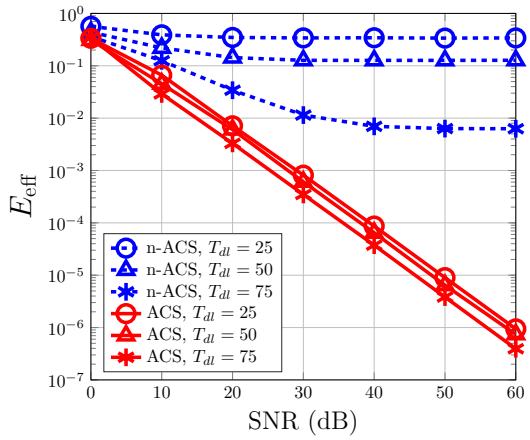
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications



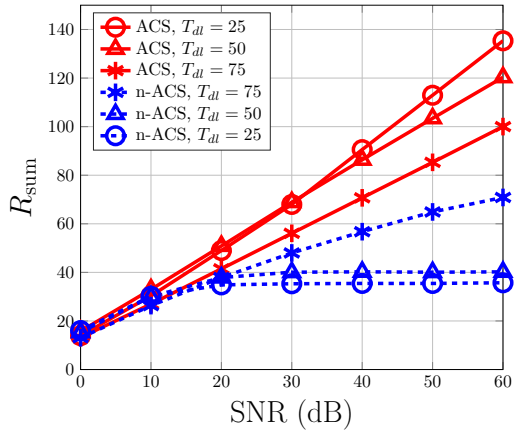
(a) $M = 64,\ K = 8$



(b) $M = 64,\ K = 8$



(c) $M = 32,\ K = 6$



(d) $M = 32,\ K = 6$

Fig. 8: Effective channel estimation error vs SNR (left figures), and sum-rate vs SNR (right figures) comparison, with $M = 64$ and $K = 8$ for the upper row and $M = 32$ and $K = 6$ for the lower row. "n-ACS" refers to the non-ACS method, implemented using random pilot vectors and no sparsification.

scheme and the sum-rate is computed via (36). We compare the performance of ACS with that of non-ACS training. The latter case is equivalent to setting the precoder in (27) to $\mathbf{B} = \mathbf{I}_{2M}$, i.e. *not* sparsifying the channels. Apart from the sum-rate metric, we also compute the mean squared error (MSE) of estimating the effective channels via the following formula:

$$E_{\text{eff}} = \frac{1}{|\mathcal{K}^\star|} \sum_{k \in \mathcal{K}^\star} \mathbb{E}\left\{ \frac{\left\| \mathbf{B}\left(\mathbf{h}_{\text{dl},k} - \widehat{\mathbf{h}}_{\text{dl},k}\right) \right\|^2}{\|\mathbf{B}\mathbf{h}_{\text{dl},k}\|^2} \right\} \tag{42}$$

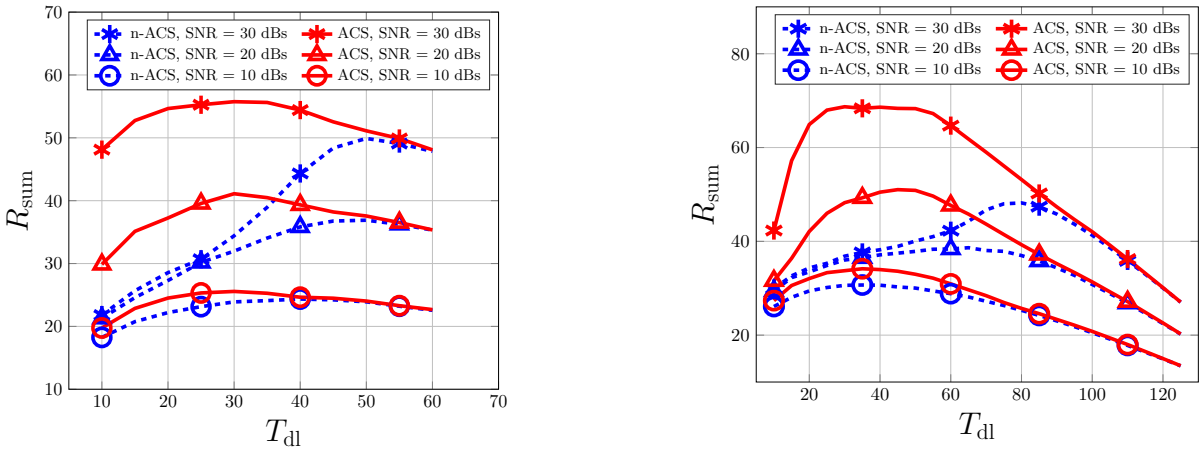
where we recall that $\mathcal{K}^\star$ is the set of users selected to be served by the MILP.

See the results of Fig. 8, in which we have plotted the effective error and sum-rate curves as a function of SNR for two different system setups, where in one the array size is $M = 32$ and serves $K = 6$ users, and in the other the array size is $M = 64$ and serves $K = 8$ users. In each case, we illustrate the results for different values of DL pilot dimension T_{dl}. First, note that with the ACS method, the effective channel estimation error decreases linearly with the the increase of $\log(\text{SNR})$ (or the SNR in dBs). In contrast, with n-ACS this error is saturated to a fixed value and does not decrease by increasing the SNR. This behavior is a direct outcome of Lemma 1, which states that if the pilot dimension is less than the sparsity order of the effective channel $\mathbf{Bh}$, then the estimation error does not tend to zero with increasing the SNR. Conversely, a stable estimation is possible if the pilot dimension is larger than the sparsity order of the effective channel, which is enabled by ACS through the MILP.

Stable estimation of the effective channel is also important in achieving an interference-free DL transmission. This can be seen by comparing the sum-rate curves of the ACS and non-ACS methods (Figs. 8d and 8b) in the high-SNR regime. With the non-ACS method, the sum-rate saturates to a fixed value as SNR increases, demonstrating an interference-limited behavior. However, with ACS the sum-rate increases linearly with $\log(\text{SNR})$, achieving much higher sum-rates in the medium-to-high-SNR regime.

Fig. 9 illustrates the sum-rate vs pilot dimension curves for the two setups as before and for various SNR values. The point of this figure is to show the relationship between the pilot dimension and the sum-rate. From (36) we note that the pilot dimension controls a trade-off in consuming time-frequency resources: increasing the pilot dimension results

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3099727, IEEE Transactions on Wireless Communications



(a) $M = 32,\ K = 6$ (b) $M = 64,\ K = 8$

Fig. 9: sum-rate vs pilot dimension curves.

in better channel estimation and therefore less interference, which increases the argument inside the logarithm in (36), but it decreases the pre-log factor $1 - T_{\text{dl}}/T$ and leaves fewer resources for data transmission. Therefore, we expect that there exists an optimal pilot dimension which maximizes the sum-rate for any given setup. This can be seen from the curves of Fig. 9. Note that in all setups the ACS method achieves higher sum-rates compared to the non-ACS method for the same pilot dimension, in some cases achieving almost twice the sum-rate of the non-ACS method.

VII. Conclusion

We proposed a thorough implementation of a multi-user FDD massive MIMO system with dual-polarized antenna elements. We addressed the dimensionality challenge of such systems through a three-step process: (1) UL covariance estimation from limited, noisy UL channel samples, (2) UL-DL covariance transformation, and (3) active channel sparsification and multi-user precoding for DL channel training and interference-free beamforming. Using error and sum-rate metrics we showed that our approach is successful for implementing dual-polarized FDD massive MIMO systems, overcoming the curse of prohibitively large dimensions and limited time-frequency resources.

References

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

[2] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE journal of selected topics in signal processing*, vol. 8, no. 5, pp. 742–758, 2014.

[3] V. Degli-Esposti, V.-M. Kolmonen, E. M. Vitucci, and P. Vainikainen, "Analysis and modeling on co-and cross-polarized urban radio propagation for dual-polarized MIMO wireless systems," *IEEE transactions on antennas and propagation*, vol. 59, no. 11, pp. 4247–4256, 2011.

[4] W. Xu, X. Wu, X. Dong, H. Zhang, and X. You, "Dual-polarized massive MIMO systems under multi-cell pilot contamination," *IEEE Access*, vol. 4, pp. 5998–6013, 2016.

[5] S. Sesia, I. Toufik, and M. Baker, *LTE-the UMTS long term evolution: from theory to practice*. John Wiley & Sons, 2011.

[6] B. Kang, V. Monga, and M. Rangaswamy, "Rank-constrained maximum likelihood estimation of structured covariance matrices," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 1, pp. 501–515, 2014.

[7] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[8] X. Luo, "Recovering model structures from large low rank and sparse covariance matrix estimation," *arXiv preprint arXiv:1111.1133*, 2011.

[9] X. Rao and V. K. Lau, "Distributed compressive csit estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3261–3271, 2014.

[10] Y. Ding and B. D. Rao, "Dictionary learning-based sparse channel representation and estimation for FDD massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5437–5451, 2018.

[11] C. Qian, X. Fu, N. D. Sidiropoulos, and Y. Yang, "Tensor-based channel estimation for dual-polarized massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6390–6403, 2018.

[12] J. Dai, A. Liu, and V. K. Lau, "FDD massive MIMO channel estimation with arbitrary 2D-array geometry," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2584–2599, 2018.

[13] M. Barzegar Khalilsarai, S. Haghighatshoar, X. Yi, and G. Caire, "FDD massive MIMO via UL/DL channel covariance extrapolation and active channel sparsification," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 121–135, 2018.

[14] L. Miretti, R. L. G. Cavalcante, and S. Stanczak, "FDD massive MIMO channel spatial covariance conversion using projection methods," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3609–3613.

[15] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 114–123, 2016.

[16] J. G. Proakis and M. Salehi, *Digital communications*. McGraw-hill New York, 2001, vol. 4.

[17] I. Gelfand and G. Shilov, "Generalized functions, volume 1, properties and operators," 1964.

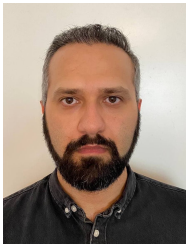
[18] D. Zhu, J. Choi, and R. W. Heath, "Two-dimensional AoD and AoA acquisition for wideband millimeter-wave systems with dual-polarized MIMO," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7890–7905, 2017.

[19] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Transactions on Acoustics, speech, and signal processing*, vol. 37, no. 5, pp. 720–741, 1989.

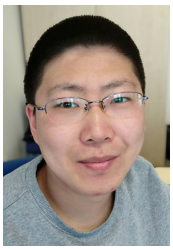
[20] O. Najim, P. Vallet, G. Ferré, and X. Mestre, "On the statistical performance of MUSIC for distributed sources," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2016, pp. 1–5.

[21] M. Barzegar Khalilsarai, T. Yang, S. Haghighatshoar, and G. Caire, "Structured channel covariance estimation for dual-polarized massive MIMO arrays," in *WSA 2020; 24th International ITG Workshop on*

Smart Antennas. VDE, 2020, pp. 1–6.
[22] H. Xie, F. Gao, S. Jin, J. Fang, and Y.-C. Liang, "Channel estimation for TDD/FDD massive MIMO systems with channel covariance computing," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 4206–4218, 2018.
[23] R. M. Gray, "Toeplitz and circulant matrices: A review," 2006.
[24] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing–the large-scale array regime," *IEEE transactions on information theory*, vol. 59, no. 10, pp. 6441–6463, 2013.
[25] R. Diestel, "Graph theory," *Grad. Texts in Math*, vol. 101, 2005.
[26] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2845–2866, 2010.
[27] M. Kobayashi, N. Jindal, and G. Caire, "Training and feedback optimization for multiuser MIMO downlink," *IEEE Transactions on Communications*, vol. 59, no. 8, pp. 2228–2240, 2011.
[28] G. Caire, "On the ergodic rate lower bounds with applications to massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3258–3268, 2018.



Mahdi Barzegar Khalilsarai (S'17) received his B.Sc. degree in Electrical Engineering in 2013 from the University of Tehran, Tehran, Iran, his M.Sc. degree in Communication Systems in 2015 from Sharif University of Technology, Tehran, Iran, and his Ph.D. degree in 2021 in Telecommunications from Technische Universität Berlin, Berlin, Germany. His research interests include signal processing, convex optimization, compressed sensing, machine learning, and their applications in wireless systems.



Tianyu Yang (S'17) received his M.Sc. degree in Electrical Engineering, Information Technology and Computer Engineering from RWTH Aachen University, Germany, in 2017. From 2017 to 2019, he was a research assistant at the Institute for Theoretical Information Technology, RWTH Aachen University. He is currently pursuing the Ph.D. degree at the Communication and Information Theory Chair, Technical University of Berlin. His research interests include massive MIMO systems, unmanned aerial vehicle communications, mobile edge computing systems, and machine learning applications in wireless networks.

Saeid Haghighatshoar (S'12–M'15) received the B.Sc. degree in Electrical Engineering (Electronics) in 2007 and the M.Sc. degree in Electrical Engineering (Communication Systems) in 2009, both from Sharif University of Technology, Tehran, Iran, and the Ph.D. degree in Computer and Communication Sciences from École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2014. From 2015 to 2020, he was a Post-Doctoral Researcher with the Communications and Information Theory (CommIT) group at Technische Universität Berlin, Berlin, Germany. His research interests lie in Information Theory, Communication Systems, Wireless Communication, Optimization Theory, and Compressed Sensing.

Xinping Yi (Member, IEEE) received the Ph.D. degree in electronics and communications from Télécom ParisTech, Paris, France, in 2015. He is currently a Lecturer (Assistant Professor) with the Department of Electrical Engineering and Electronics, University of Liverpool, U.K. Prior to Liverpool, he was a Research Associate with Technische Universität Berlin, Berlin, Germany, from 2014 to 2017, a Research Assistant with EURECOM, Sophia Antipolis, France, from 2011 to 2014, and a Research Engineer with Huawei Technologies, Shenzhen, China, from 2009 to 2011. His main research interests include information theory, graph theory, and machine learning, and their applications in wireless communications and artificial intelligence.

Giuseppe Caire (S '92 – M '94 – SM '03 – F '05) was born in Torino in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992, and the Ph.D. from Politecnico di Torino in 1994. He has been a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994-1995, Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin, Germany.

He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society and Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, the Vodafone Innovation Prize in 2015, an ERC Advanced Grant in 2018, the Leonard G. Abraham Prize for best IEEE JSAC paper in 2019, the IEEE Communications Society Edwin Howard Armstrong Achievement Award in 2020, and he is a recipient of the 2021 Leibinz Prize of the German National Science Foundation (DFG). Giuseppe Caire is a Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.