IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 14, 2019, accepted March 3, 2019, date of publication March 7, 2019, date of current version March 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903654

Exploiting the Massive MIMO Channel Structural Properties for Minimization of Channel Estimation Error and Training Overhead

SAMER BAZZI[1], STELIOS STEFANATOS[2], LUC LE MAGOAROU[3], SALAH EDDINE HAJRI[4], MOHAMAD ASSAAD[4], (Senior Member, IEEE), STÉPHANE PAQUELET[3], GERHARD WUNDER[2], (Senior Member, IEEE), AND WEN XU[1], (Senior Member, IEEE)

[1]European Research Center, Huawei Technologies Duesseldorf GmbH, 80992 Munich, Germany
[2]Heisenberg Communication and Information Theory Group, Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany
[3]b<>com, 35510 Rennes, France
[4]Laboratoire des Signaux et Systemes, CNRS, CentraleSupelec, 91190 Gif-sur-Yvette, France

Corresponding author: Samer Bazzi (samer.bazzi@huawei.com)

This work was supported in part by the framework of the Horizon 2020 Project ONE5G receiving funds from the European Union under Grant ICT-760809.

ABSTRACT Exploiting massive multiple-input-multiple-output (MIMO) gains come at the expense of obtaining accurate channel estimates at the base station. However, conventional channel estimation techniques do not scale well with increasing number of antennas and incur an unacceptably large training overhead in many applications. This calls for training designs and channel estimation techniques that efficiently exploit the physical properties of the massive MIMO channel as captured by sophisticated system/channel models. In this paper, we present designs that exploit the sparsity of the angle and delay domain representation of the massive MIMO channel as well as the low-rank property of the channel covariance, while also providing the connection between the sparse angle-delay representation and low-rank covariance property. Numerous multiuser scenarios are investigated including uplink, downlink, and single- and multi-cell communications, with the designs aiming at minimizing the channel estimation error or maximizing achievable rates with reduced training overhead. Theoretical analysis and numerical performance results indicate significant reduction of training overhead over conventional techniques while achieving similar performance. The presented methods demonstrate the importance of exploiting fundamental channel properties and reveal important insights on the interplay/tradeoff between training overhead and performance that can serve as guidelines for the design of future massive MIMO communication systems.

INDEX TERMS Channel sparsity, correlated fading, channel estimation, training design, compressive sensing, pilot contamination, performance bounds, massive MIMO.

I. INTRODUCTION

A. BACKGROUND

Massive multiple-input-multiple-output (MIMO) technology is a cornerstone of future communication systems and is crucial for meeting 5G requirements [1]. By coherent processing of the signals over a large number of cheap antenna elements at the base station (BS), massive MIMO systems focus the radiated energy on intended targets and discriminate received signals using transmit precoding and receive combining, respectively. High spatial multiplexing capabilities can be obtained with simple and cost efficient transceiver design, which make massive MIMO systems very appealing. Nevertheless, massive MIMO gains depend heavily on accurate channel estimation, which is not a straightforward task. At first sight, the number of parameters to estimate scales with the number of transmit and receive antennas, which may be very large in massive MIMO systems. For this reason, classical estimation methods such as least squares (LS) estimation may not be appropriate, especially when only a limited number of observations can be obtained due to channel aging. To overcome this limitation, some information/structure about the channel has to be used to regularize the problem.

One way to regularize the channel estimation problem is to use a parametric channel model, which exploits the fact that a signal arrives at the receiver via a limited number of

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

distinct (resolvable) paths [2]. Therefore, posing the channel estimation problem as that of identifying the channel paths properties (gain, delay, angle) immediately implies improvement of the channel estimation procedure over conventional approaches as the number of unknowns to be estimated (significantly) decreases. Another way is to exploit the prior distribution about the channel, yielding Bayesian estimation [3], [4]. If the channel covariance matrix is rank deficient, the number of parameters that need to be estimated effectively decreases, as shown in [3].[1]

In both cases, exploiting the underlying channel structural properties results in a reduced number of parameters to estimate. This directly translates to a reduced overhead of any training based scheme. This is clearly beneficial for any communication system and any envisioned use case or scenario (e.g. enhanced mobile broadband [5], ultra-reliable low-latency communications [6], wireless sensor networks [7], [8], etc.) as it results in larger effective throughputs and lower latencies required to decode the desired signals, and is inherently robust to channel aging. Unfortunately, analytical insights on the necessary or even sufficient overheads for reliable channel estimation remain missing for a number of important multiuser scenarios. This holds for compressive sensing (CS) approaches exploiting the parametric channel model as well as Bayesian estimation approaches.

Exploiting the channel structural properties can be further leveraged to enable massive connectivity, which is an important 5G requirement. This can be understood in the context of training overhead reduction as well. By developing effective clustering techniques which separate users based on prior information, training sequences can be reused among clusters with minimum uplink (UL) pilot contamination. Thus, high connection densities can be achieved with low training overheads.

B. CONTRIBUTIONS AND PAPER ORGANIZATION

This paper presents several solutions that exploit the channel structural properties for training overhead reduction in numerous massive MIMO scenarios [9]–[14]. These works are part of the outcomes of the Horizon 2020 ONE5G project [15]. The solutions address different use cases and apply different sets of tools that exploit different levels of prior knowledge at the BS. The results are illustrated numerically for various channel models and BS configurations. Besides presenting detailed discussions of the aforementioned works, the paper complements these works as well. For instance, a new problem formulation is provided in Sec. III, while new comparisons are provided in Section V. Note that previous papers (see, e.g., [16]) have presented a high level overview of low-overhead channel estimation schemes without discussing training overhead scaling nor training sequence reuse aspects. The latter are the focus of this work.

Section II introduces the general channel estimation problem formulation, the parametric channel model used in CS approaches, and the covariance matrix structure for channels obeying the parametric model. The potential for training overhead reduction by exploiting the parametric channel model, possibly combined with covariance matrix information, is highlighted for the simple single-link case. This motivates the investigation of sophisticated channel estimation and training designs for the multiuser cases considered in this paper.

Section III presents a study aimed at finding the optimal number of virtual, or dominant, paths that are sufficient to accurately represent the channel. The study and numerical investigations show that realistic channels can indeed be accurately presented with a small number of virtual paths (at least in high frequency bands), even when simulators consider path numbers in the order of hundreds. It thus serves as a motivation for the subsequent sections, which address different aspects of training design exploiting the found result.

For the parametric channel model, we analytically study the scaling of training overhead (pilot subcarriers) for reliable channel estimation in the UL wideband scenario in Section IV. The study is critically facilitated by the concept of hierarchical sparsity, introduced therein. By partitioning users such that users within the same group utilize the same pilot subcarriers, the sufficient scaling rule highlights that the training overhead scales logarithmically with the number of subcarriers but it is independent of the number of active users per group. This is an effect that appears only in massive MIMO scenarios (i.e., the overhead would scale proportionally to these parameters in the SISO case). Essentially, with massive MIMO, the bulk of the training overhead is shifted to the spatial domain (by considering a large number of observed antennas).

The paper then considers multiuser training designs exploiting low-rank covariance information for narrowband channels in the following sections. We present a sufficient scaling rule of the downlink (DL) training overhead for reliable DL channel estimation in Section V. For practical channels with correlated entries, the found sufficient overhead depends on the ranks of the users' covariance matrices and the overlap between their range spaces, and may be much smaller than the number of BS antennas. We then go beyond the channel estimation sum mean-square-error (MSE) metric and consider training designs operating on the achievable sum rate metric, which is a more important metric in many wireless systems' applications.

Finally, to mitigate (intra-cell) pilot contamination effects in massive connectivity setups, we present a novel spatial domain grouping scheme that allows for a high training sequence reuse in Sec. VI. We base our approach on the spatial basis provided by the unitary discrete Fourier transform (DFT) matrix, and perform user grouping based on the users' covariance information. We then tackle the multi-cell case and present an inter-cell pilot decontamination solution using graph theory. Despite using different tools and addressing

[1] Sec. V-C addresses the full rank case and shows how covariance information may still be used towards training overhead reduction.

different use cases, Secs. IV, V, and VI share the conclusion that the training overheads can be made much smaller than the number of transmit antennas (sum of users' antennas in the UL, number of BS antennas in the DL) if channel structural information is properly exploited.

We wrap up and point out to interesting research directions in Section VII. We finally point out that since the paper addresses different scenarios with different levels of prior knowledge, the relevant works related to each scenario are included in the corresponding section.

C. NOTATION

Vectors (matrices) will be denoted by small (upper) case bold letters. Any vector $\boldsymbol{x}$ will always be treated as a single-column matrix. $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$ and $(\cdot)^{*}$ denote transpose, Hermitian, and complex conjugate, respectively. The matrix consisting of the first K columns of $\boldsymbol{X}$ will be denoted as $\boldsymbol{X}^{1:K}$ and $[\boldsymbol{X}]_{l,m}$ is the (l, m)th entry of $\boldsymbol{X}$. We also define for convenience the vector

$$\boldsymbol{f}_K(x) \triangleq [1, e^{j2\pi x/K}, \ldots, e^{j2\pi(K-1)x/K}]^{\mathrm{T}}, \tag{1}$$

and the $K \times K$ DFT matrix

$$\boldsymbol{F}_K \triangleq [\boldsymbol{f}_K^*(0), \boldsymbol{f}_K^*(1), \ldots, \boldsymbol{f}_K^*(K-1)].$$

The set $\{1, 2, \ldots, K\}$ is denoted as $[K]$ and $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. $\mathbf{I}_K$ denotes the $K \times K$ identity matrix. $\mathcal{D}(\boldsymbol{x})$ is the diagonal matrix with main diagonal $\boldsymbol{x}$. The vector resulting by stacking the columns of $\boldsymbol{X}$ is denoted by $\mathrm{vec}(\boldsymbol{X})$. $\mathbb{C}^{N_1 \cdot N_2 \cdots N_\ell}$ denotes the space of complex-valued, multilevel block vectors consisting of N_1 blocks, each containing N_2 blocks, . . ., each containing $N_{\ell-1}$ blocks of N_ℓ elements (for a total of $N_1 N_2 \cdots N_\ell$ elements). A vector $\boldsymbol{x}$ is called s-sparse if $|\mathrm{supp}(\boldsymbol{x})| = s$. $\|\boldsymbol{x}\|$ and $\|\boldsymbol{X}\|_F$ are the Euclidean and Frobenius norms of vector $\boldsymbol{x}$ and matrix $\boldsymbol{X}$, respectively. The eigenvalue decomposition of a positive semi-definite matrix $\boldsymbol{X} \in \mathbb{C}^{N \times N}$ will be denoted as $\boldsymbol{X} = \boldsymbol{U}_{\boldsymbol{X}} \mathcal{D}(\lambda_{\boldsymbol{X},1}, \lambda_{\boldsymbol{X},2}, \ldots, \lambda_{\boldsymbol{X},N}) \boldsymbol{U}_{\boldsymbol{X}}^{\mathrm{H}}$, with the eigenvalues ordered in decreasing order, i.e., $\lambda_{\boldsymbol{X},1} \geq \lambda_{\boldsymbol{X},2} \geq \ldots \geq \lambda_{\boldsymbol{X},N} \geq 0$, and the eigenvector matrix being unitary ($\boldsymbol{U}_{\boldsymbol{X}}^{\mathrm{H}} \boldsymbol{U}_{\boldsymbol{X}} = \boldsymbol{U}_{\boldsymbol{X}} \boldsymbol{U}_{\boldsymbol{X}}^{\mathrm{H}} = \mathbf{I}_N$). The rank and range space of $\boldsymbol{X}$ are denoted as $\mathrm{rank}(\boldsymbol{X})$ and $\mathrm{ran}(\boldsymbol{X})$, respectively.

II. GENERAL CHANNEL ESTIMATION MODELS

We consider in this paper the standard massive MIMO setting where BSs in the network are equipped with M antennas each and users have single antennas [5]. Transmissions are performed via orthogonal-frequency-division-multiplexing (OFDM) with N subcarriers and a cyclic prefix length of N_{cp} samples that is greater than the maximum delay spread experienced by any user.

In all the scenarios that will be considered in the following, we focus on training-based channel estimation schemes. The observed signal at the receiver side for an arbitrary BS-user link (either DL or UL) can always be written as the standard linear model

$$\boldsymbol{y} = \boldsymbol{A}\,\mathrm{vec}(\boldsymbol{H}) + \boldsymbol{n} \tag{2}$$

where $\boldsymbol{y} \in \mathbb{C}^T$ with T a number proportional to the training overhead, $\boldsymbol{H} \in \mathbb{C}^{M \times N}$ is the spatial-frequency response of the unknown channel that is assumed time-invariant within the transmission interval, $\boldsymbol{n} \in \mathbb{C}^T$ represents additive noise, and $\boldsymbol{A} \in \mathbb{C}^{T \times MN}$ is the sensing/training matrix that is known at the receiver and whose structure depends on the specific transmission scenario (UL or DL) and transmitted training symbols.

Without any prior information about the channel and noise, i.e., with $\boldsymbol{H}$ and $\boldsymbol{n}$ treated as arbitrary, an estimate $\hat{\boldsymbol{H}}$ of the channel can be obtained by an LS approach [17], i.e.,

$$\mathrm{vec}(\hat{\boldsymbol{H}}) = \arg \min_{\boldsymbol{x} \in \mathbb{C}^{MN}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2. \tag{3}$$

The LS estimate is unbiased, and under the common assumption of the noise vector consisting of independent and identically distributed (i.i.d.) elements, it is additionally the best linear estimate [17]. However, an LS estimate can only be obtained when $\boldsymbol{A}$ has a left inverse, which requires $T \geq MN$ [17]. It follows that LS estimation requires a training overhead that scales proportional to the channel ambient dimension MN. With M in the order of 100 or more in a massive MIMO setting, this overhead quickly becomes unacceptable even under narrowband signaling ($N = 1$), especially for scenarios requiring frequent training (high mobility) and/or a massive number of users.

A. CHANNEL ESTIMATION BASED ON THE PARAMETRIC MODEL

The key to training overhead reduction is the observation that $\boldsymbol{H}$ is not an arbitrary matrix, but that its elements exhibit correlation as has been experimentally confirmed [18]. Inspired by the far-field waveform propagation physics, the most common as well as simplest model that (approximately) captures these correlation properties expresses $\boldsymbol{H}$ as a sum of rank one matrices with a clearly defined structure, namely [19]–[21]

$$\boldsymbol{H} = \sum_{p=1}^{P} c_p \boldsymbol{e}_M(\theta_p, \phi_p) \boldsymbol{f}_N^{\mathrm{H}}(\tau_p). \tag{4}$$

In this model, P, $c_p \in \mathbb{C}$, $\theta_p \in [-\pi, \pi]$, $\phi_p \in [-\pi, \pi]$, and $\tau_p \in [0, N_{cp} - 1]$ are parameters representing the number of propagation paths and the gain, azimuth angle, elevation angle, and normalized delay (with respect to the system sampling period) of the p-th propagation path, respectively.[2] Each path angle tuple represents the direction of departure or direction of arrival of the path, for the DL and UL transmission cases, respectively. The elements of the steering vector $\boldsymbol{e}_M(\theta, \phi) \in \mathbb{C}^M$ capture the known antenna array geometry and are given by [22]

$$[\boldsymbol{e}_M(\theta, \phi)]_m = \mathrm{e}^{-\mathrm{j}\frac{2\pi}{\lambda}\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{u}(\theta,\phi)}, \quad m = 1, \ldots, M, \tag{5}$$

where $\boldsymbol{u}(\theta, \phi) \in \mathbb{R}^3$ denotes the unit vector with azimuth and elevation angles θ, ϕ, respectively, $\boldsymbol{a}_m \in \mathbb{R}^3$ denotes the

[2] It is assumed that $\theta_p \neq \theta_{p'}$ and/or $\phi_p \neq \phi_{p'}$, for all $p \neq p'$, otherwise there would exist paths that could be equivalently merged into a single path.

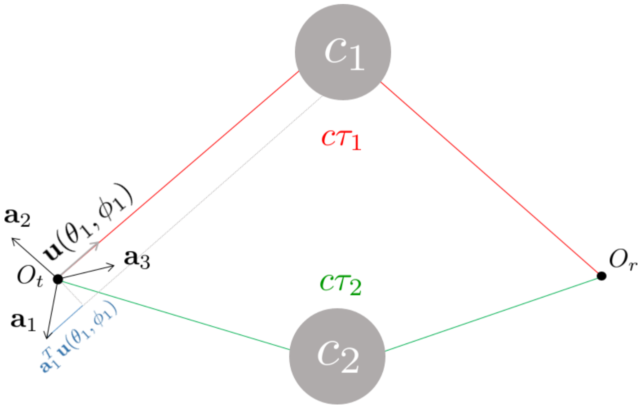

FIGURE 1. System representation considering two paths of complex gains c_1 and c_2 and three transmit antennas. For the first path, the direction of departure $u(\theta_1, \phi_1)$ is shown, as well as the path length difference that determines the phase shift in the steering vector (5) for the first transmit antenna located at a_1 (in blue). Path lengths causing delays τ_1 and τ_2 are also shown (in red and green).

position of the m-th BS antenna element with respect to the centroid of the array, and $\lambda > 0$ is the carrier wavelength. Fig. 1 illustrates this model for the case $P = 2$.

Note that the channel model of (4) is the obvious extension to the MIMO setting of the single antenna model considered in standard textbooks, describing the channel as a sum of propagation paths [23]. Although the model applies to any carrier frequency used today by practical communication systems, the number of significant paths decreases, in principle, with increasing frequency (e.g. the millimeter wave band frequencies [24]). Of course, propagation with few significant paths is also possible with smaller carrier frequencies (e.g., microwave band) under certain use cases, as verified by a multitude of legacy works on channel estimation exploiting this fact (see, e.g., [25]–[28]).

The model of (4) implies a well defined structure for the channel realization that should be exploited by the channel estimator, even if no other a priori (e.g., statistical) information is available. Consideration of this model effectively transforms the channel matrix estimation problem to the problem of estimating the $4P$ path parameters $\{(c_p, \theta_p, \phi_p, \tau_p)\}_{p=1}^{P}$. As will be shown in Sec. III, a (very) small value of P can be considered by the channel estimator at least for sufficiently high carrier frequencies, i.e., the channel is sparse, rendering the number of unknowns to be estimated much smaller than the channel ambient dimension.

Of particular importance due to its implementation simplicity as well as analytical tractability is the case of a uniform linear array (ULA) where the antenna elements are distributed along a line in $\mathbb{R}^3$ with equal spacing. Assuming, w.l.o.g., that this line belongs to the horizontal plane, the ULA cannot resolve elevation angles, i.e., is independent of ϕ, and can only discriminate among azimuth angles in $\theta \in [-\pi/2, \pi/2]$. With an element spacing of $\lambda/2$, its steering vector is given by [22]

$$\boldsymbol{e}_M(\theta) = \boldsymbol{f}_M(M \sin\theta/2) \tag{6}$$

for $\theta \in [-\pi/2, \pi/2]$, up to a unit-modulus constant, which is of no importance for our purposes as it can be subsumed by the complex channel path gains.

B. BAYESIAN CHANNEL ESTIMATION

When the probability density function that describes $\boldsymbol{H}$ is known, a Bayesian channel estimation framework can be employed, typically with the objective of minimizing the Bayesian estimation MSE

$$\text{MSE} \triangleq \mathbb{E}\left(\|\boldsymbol{H} - \hat{\boldsymbol{H}}\|_F^2\right) \tag{7}$$

where the expectation is taken over the joint distribution of $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}$. It is well known that the optimal estimate is the conditional channel mean given the received signal [17]. In case a complete statistical description of the channel is not available or the calculation of the optimal estimate is computationally complex, the linear minimum mean square error (LMMSE) estimator is commonly employed due to its efficient implementation and utilization of statistical information that is limited to the mean and covariance of the channel.

Clearly, LMMSE estimators can be used for any training matrix choice. When the channel covariance matrix is low-rank, the channel exhibits important structural properties that could be exploited within the context of training matrix optimization and overhead reduction. Considering for simplicity and w.l.o.g. narrowband transmission ($N = 1$) and writing the channel as a column vector $\boldsymbol{h} \in \mathbb{C}^M$, the channel covariance matrix equals

$$\boldsymbol{C} \triangleq \mathbb{E}\left((\boldsymbol{h} - \mathbb{E}(\boldsymbol{h}))(\boldsymbol{h} - \mathbb{E}(\boldsymbol{h}))^{\mathrm{H}}\right) \tag{8}$$

$$= \boldsymbol{U}_{\boldsymbol{C}}\, \mathcal{D}(\lambda_{\boldsymbol{C},1}, \lambda_{\boldsymbol{C},2}, \ldots, \lambda_{\boldsymbol{C},M})\, \boldsymbol{U}_{\boldsymbol{C}}^{\mathrm{H}}. \tag{9}$$

The gain offered by the knowledge of the covariance matrix can be seen by application of the Karhunen-Loève (KL) expansion which expresses the channel as [29]

$$\boldsymbol{h} = \mathbb{E}(\boldsymbol{h}) + \boldsymbol{U}_{\boldsymbol{C}}^{1:\mathrm{rank}(\boldsymbol{C})}\, \boldsymbol{h}_{\mathrm{KL}} \tag{10}$$

where $\boldsymbol{h}_{\mathrm{KL}} \in \mathbb{C}^{\mathrm{rank}(\boldsymbol{C})}$ is the vector of $\mathrm{rank}(\boldsymbol{C})$ KL coefficients that are uncorrelated zero-mean random variables of variance equal to the non-zero eigenvalues of $\boldsymbol{C}$. This representation effectively transforms the channel estimation problem to that of estimating $\mathrm{rank}(\boldsymbol{C})$ variables, which, when $\mathrm{rank}(\boldsymbol{C}) \ll M$ (highly correlated channel entries), provides significant training overhead reductions when $\boldsymbol{C}$ is known at the BS. Since the columns of $\boldsymbol{U}_{\boldsymbol{C}}^{1:\mathrm{rank}(\boldsymbol{C})}$ provide a known basis for the channel representation, sending training symbols along the columns of $\boldsymbol{U}_{\boldsymbol{C}}^{1:\mathrm{rank}(\boldsymbol{C})}$—i.e., a training overhead of $\mathrm{rank}(\boldsymbol{C})$ slots—is sufficient to estimate $\boldsymbol{h}_{\mathrm{KL}}$, and thus $\boldsymbol{h}$.[3]

C. COVARIANCE MATRIX OF CHANNELS OBEYING THE PARAMETRIC MODEL

The above discussion was so far general in the sense that it holds for arbitrary channel models. Let us now consider a

[3] The same training solution is obtained in [30], by explicitly minimizing the channel estimation MSE when LMMSE estimators are used.

channel based on the parametric model (4) and look at its covariance matrix. Under the common assumption that the path gains are uncorrelated zero mean variables, the covariance matrix equals

$$\boldsymbol{C} = \sum_{p=1}^{P} \mathbb{E}(|c_p|^2)\, \mathbb{E}\left(\boldsymbol{e}_M(\theta_p, \phi_p)\boldsymbol{e}_M^{\mathrm{H}}(\theta_p, \phi_p)\right). \tag{11}$$

In time-invariant scenarios (resp. limited mobility conditions), the path angles become (resp. can be approximated as) fixed and the covariance matrix reduces to

$$\boldsymbol{C} = \sum_{p=1}^{P} \mathbb{E}(|c_p|^2)\, \boldsymbol{e}_M(\theta_p, \phi_p)\boldsymbol{e}_M^{\mathrm{H}}(\theta_p, \phi_p). \tag{12}$$

In the sparse channel case ($P < M$), the knowledge of $\boldsymbol{C}$ implies the knowledge of 1) the path angles, and 2) the path gains' variances.[4] Clearly, the knowledge of the angles and gains' variances implies the knowledge of $\boldsymbol{C}$ in all cases (sparse or non-sparse channels).

From (12), we observe that $\boldsymbol{C}$ is of rank P. Assuming a sparse channel [small P in (4)], the covariance matrix captures this fundamental sparsity, manifested as a low-rank property. Here, Bayesian estimation reduces to estimating the P complex path gains $\{c_p\}_{p=1}^{P}$ [cf. (4)]. The latter exhibit a one-to-one mapping to the P complex entries of $\boldsymbol{h}_{\mathrm{KL}}$ [cf. (10)].

Whether the parametric model or Bayesian model is used for estimation, we observe that the number of parameters to estimate and resulting overhead is $\mathcal{O}(P)$, which is independent of the channel ambient dimension. Since Bayesian estimation incorporates prior information (knowledge of angle values and paths' statistics) into the model, it results in a lower number of parameters to be estimated than that of non-Bayesian estimation (e.g., one based solely on the parametric model). Note that a reduced number of parameters can be exploited in two ways:

1) A reduced training overhead, compared to the overheads of other (non-Bayesian) methods, or
2) An estimation performance that is superior to other non-Bayesian methods, for a fixed training overhead.

In this paper, we mainly consider the Bayesian estimation potential for training overhead reduction.

III. SPARSE REPRESENTATION OF THE MASSIVE MIMO CHANNEL

Many CS works are based on the assumption that the wireless channel is sparse (small P). In the previous section, we have presented conditions under which this translates into a covariance matrix of low-rank, also given by P. The latter assumption is the starting point of many other works as well. Nonetheless, it is worth questioning whether the sparse assumption really holds in practice.

We aim at answering this question in this section, by considering the estimation of the channel of one arbitrary user in the DL. In order to simplify the discussion, narrowband transmissions are considered, resulting in the channel being represented by a column vector $\boldsymbol{h} \in \mathbb{C}^M$ obtained as the right hand side of (4) with $N = 1$.

A. SYSTEM MODEL

For channel estimation purposes, the BS reserves T slots for transmission of training (pilot) symbols. The corresponding training matrix is denoted $\boldsymbol{S} \triangleq [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T] \in \mathbb{C}^{M\times T}$, where $\boldsymbol{s}_t^* \in \mathbb{C}^M$ is the training sequence transmitted in training slot t. The training sequences are considered to be mutually orthogonal and of equal norm, i.e., $\boldsymbol{s}_i^{\mathrm{H}}\boldsymbol{s}_j = P_{\mathrm{t}}\delta_{ij}, i, j \in \{1, 2, \ldots, T\}$, where $P_t > 0$ is the transmit power. Note that the orthogonality requirement restricts the number of training slots as $T \leq M$, which is a reasonable constraint towards low training overhead designs (small T).

The received training signal at the user equals

$$\boldsymbol{y} = \boldsymbol{S}^{\mathrm{H}}\boldsymbol{h} + \boldsymbol{n} \tag{13}$$

where $\boldsymbol{n}$ is a noise vector with i.i.d. complex Gaussian elements of zero mean and variance σ^2. For convenience we will define the training signal-to-noise (SNR) as

$$\mathrm{SNR} \triangleq \frac{P_{\mathrm{t}}}{\sigma^2}. \tag{14}$$

Based on $\boldsymbol{y}$, the task of the user is to obtain an estimate of $\boldsymbol{h}$. When there is no prior/structural information about $\boldsymbol{h}$, the channel is treated as deterministic and an LS channel estimation approach appears natural. However, we do have structural information about the channel, as described in Sec. II-A. It is therefore reasonable to consider an estimation procedure that provides channel estimates with the same structure as that suggested by the fundamental physical model of (4) [32]–[34].

As the actual number of propagation paths P is not known, the estimator a priori assumes a number of $\tilde{P}$ paths for the channel model, effectively treating the observation in (13) as

$$\boldsymbol{y} = \boldsymbol{S}^{\mathrm{H}}\tilde{\boldsymbol{h}} + \tilde{\boldsymbol{n}}, \tag{15}$$

where

$$\tilde{\boldsymbol{h}} \triangleq \sum_{p=1}^{\tilde{P}} \tilde{c}_p \boldsymbol{e}_M\left(\tilde{\theta}_p, \tilde{\phi}_p\right) \tag{16}$$

is the vector to be estimated, with $\{(\tilde{c}_p, \tilde{\theta}_p, \tilde{\phi}_p)\}_{p=1}^{\tilde{P}}$ representing the path parameters of the model and $\tilde{\boldsymbol{n}} \triangleq \boldsymbol{n} + \boldsymbol{S}^{\mathrm{H}}(\boldsymbol{h} - \tilde{\boldsymbol{h}})$ represents the effective noise. Denoting

$$\hat{\tilde{\boldsymbol{h}}} \triangleq \sum_{p=1}^{\tilde{P}} \hat{\tilde{c}}_p \boldsymbol{e}_M\left(\hat{\tilde{\theta}}_p, \hat{\tilde{\phi}}_p\right) \tag{17}$$

the channel estimate where $\{(\hat{\tilde{c}}_p, \hat{\tilde{\theta}}_p, \hat{\tilde{\phi}}_p)\}_{p=1}^{\tilde{P}}$ are the estimated path parameters of the assumed model, note that the channel estimation error $\hat{\tilde{\boldsymbol{h}}} - \boldsymbol{h}$ is the result of

1) the error in estimating the $3\tilde{P}$ model parameters, and
2) the error due to model mismatch (i.e., $\tilde{\boldsymbol{h}} \neq \boldsymbol{h}$).

[4] These can be obtained via a Vandermonde decomposition of $\boldsymbol{C}$ [31].

In the following, these two error types are characterized towards obtaining insights on the channel estimation performance as well as the selection of $\tilde{P}$.

B. PARAMETER ESTIMATION ERROR

In order to isolate the parameter estimation error, we assume that the model mismatch error is negligible compared to the noise, i.e., $\tilde{\boldsymbol{n}} \approx \boldsymbol{n}$, effectively implying that $\tilde{\boldsymbol{h}} \approx \boldsymbol{h}$. Under this assumption, it can be shown that the MSE of any unbiased estimate of $\tilde{\boldsymbol{h}}$ is lower bounded as [9]

$$\mathbb{E}\|\tilde{\boldsymbol{h}} - \hat{\tilde{\boldsymbol{h}}}\|^2 \geq 2\tilde{P}/\mathrm{SNR}. \tag{18}$$

This bound is achievable by a training sequence design satisfying the condition [9]

$$\mathrm{ran}(\boldsymbol{S}^*) \supseteq \mathrm{span}\left(\bigcup_{p=1}^{\tilde{P}} \left\{\boldsymbol{e}_{M,p}, \frac{\partial \boldsymbol{e}_{M,p}}{\partial \theta_p}, \frac{\partial \boldsymbol{e}_{M,p}}{\partial \phi_p}\right\}\right) \tag{19}$$

where $\boldsymbol{e}_{M,p}$ is used here to denote $\boldsymbol{e}_M(\theta_p, \phi_p)$. This condition can be interpreted as specifying a subspace of beamforming directions in $\mathbb{C}^M$ that the training sequences should span. Note that this condition can be satisfied without any knowledge about this space by using training sequences that span the whole $\mathbb{C}^M$, i.e., $\boldsymbol{S} = \sqrt{P_t}\boldsymbol{Q}$, where $\boldsymbol{Q} \in \mathbb{C}^{M\times M}$ is an arbitrary unitary matrix. This approach requires $T = M$ training slots. When channel information is available such that the space is known, the optimality condition can be satisfied with $T \leq 3\tilde{P}$ slots, potentially achieving great overhead reduction with no performance degradation.[5]

An important feature of the MSE bound is that it is proportional to $\tilde{P}$, which captures the well-known fact that increasing the number of parameters to estimate results in worst performance [17]. This observation motivates the consideration of a small $\tilde{P}$ by the estimator. However, an arbitrarily small $\tilde{P}$ will cause a large model mismatch error. Thus, an essential issue that is investigated next is how small can $\tilde{P}$ be selected while the model selection mismatch is still negligible.

C. MODEL MISMATCH ERROR

As a measure of the model mismatch error we consider the norm $||\boldsymbol{h} - \tilde{\boldsymbol{h}}||$.[6] Unfortunately, exact characterization of this metric is not available, since identification of the value of $\tilde{\boldsymbol{h}}$ that minimizes the model mismatch error for an arbitrary $\boldsymbol{h}$ is an NP-hard problem [35]. However, a closed-form upper bound can be obtained by considering a, possibly suboptimal, $\tilde{\boldsymbol{h}}$ obtained under the principle of path merging as follows:

1) Partition the set of P true path parameter tuples $\{(c_p, \theta_p, \phi_p)\}_{p=1}^{P}$ into $\tilde{P} \leq P$ subsets $\{\mathcal{R}_p\}_{p=1}^{\tilde{P}}$, where each subset $\mathcal{R}_p$ contains at least one tuple.
2) For each subset $\mathcal{R}_p, p = 1, 2, \ldots, \tilde{P}$, assign a virtual path with an angle tuple $(\tilde{\theta}_p, \tilde{\phi}_p)$ and set the corresponding path gain equal to

$$\tilde{c}_p = \arg\min_{\gamma\in\mathbb{C}} \|\gamma\, \boldsymbol{e}_M(\tilde{\theta}_p, \tilde{\phi}_p) - \sum_{(c,\theta,\phi)\in\mathcal{R}_p} c\, \boldsymbol{e}_M(\theta, \phi)\|^2$$

The mismatch error of this approach has been characterized in [10], resulting in the bound

$$\left\|\boldsymbol{h} - \tilde{\boldsymbol{h}}\right\| \leq \kappa \sum_{p=1}^{\tilde{P}} \sum_{(c,\theta,\phi)\in\mathcal{R}_p} |c| \left\|\boldsymbol{u}(\tilde{\theta}_p, \tilde{\phi}_p) - \boldsymbol{u}(\theta, \phi)\right\| \tag{20}$$

where $\kappa \triangleq \frac{2\pi}{\lambda}\sqrt{\sum_{m=1}^{M} \left\|\boldsymbol{a}_m\right\|^2}$.

The form of the bound reveals that the model mismatch error depends critically on how the path angles of the true channel are distributed as well as how these are merged for the construction of $\tilde{\boldsymbol{h}}$. In particular, the error can be made zero when the paths of each subset $\mathcal{R}_p$ are collinear, i.e., have identical angles, and the corresponding virtual path angles are set to the same values. This is trivially the case when $\tilde{P} = P$ (with each subset consisting of only one path). With $\tilde{P} < P$, a mismatch error is unavoidable, however, this error can be small when the paths can be partitioned into groups containing approximately collinear paths.

Clearly, increasing $\tilde{P}$ towards P reduces the bound towards zero as there is increased flexibility in identifying and merging approximately collinear paths. In particular, it was shown in [10] that for a channel realization with angle tuples distributed uniformly over the sphere $[-\pi, \pi]^2$, which represents the worst case for the path-merging scheme, the mismatch error scales as $\mathcal{O}(1/\tilde{P})$ as $\tilde{P}$ increases. Recalling that the parameter estimation error is increasing with $\tilde{P}$, it follows that there exists a value of $\tilde{P}$ that provides the optimal tradeoff between the model mismatch and parameter estimation errors. This value is evaluated numerically in the following section.

D. NUMERICAL RESULTS

In this section, we numerically identify the optimal $\tilde{P}$ for a simulated scenario where the channel $\boldsymbol{h}$ is generated using the NYUSIM realistic channel simulator [19] assuming a millimeter-wave carrier frequency (28 GHz) and a distance of 30 meters between the BS and the user. The BS is equipped with a square uniform planar array (UPA) of $M = 64$ $\lambda/2$ separated elements. For each channel realization, the number of propagation paths P is between fifty and a hundred and the channel norm is normalized to 1. For channel estimation, the BS employs the training matrix $\boldsymbol{S} = P_t \boldsymbol{I}_M$ and the user applies the standard orthogonal matching pursuit (OMP) algorithm [36], which provides estimates of the form (17) for a given number $\tilde{P}$ of virtual paths.

Figure 2 shows the MSE $\mathbb{E}(\|\boldsymbol{h} - \hat{\boldsymbol{h}}\|^2)$, averaged over many realizations of $\boldsymbol{h}$, as a function of the number of virtual paths $\tilde{P}$ considered by the estimator. As a (non-achievable) lower performance bound, we also plot the MSE of an OMP-generated

[5] Training overhead reduction will be demonstrated in Secs. V and VI under a more general framework exploiting channel covariance matrix information.

[6] The model mismatch error is characterized under the assumption of perfect knowledge of $\boldsymbol{h}$ and is treated independently of the parameter estimation problem.

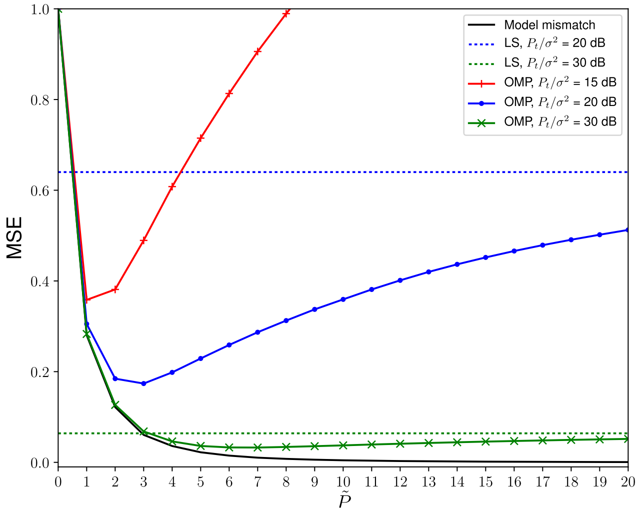

FIGURE 2. Normalized MSE performance of parametric channel estimation via OMP as a function of the considered number of virtual paths $\tilde{P}$.

estimate when $\boldsymbol{h}$ is observed directly and without noise. Note that in this artificial case the estimation error is only due to the model mismatch. It can be seen that performance is strongly dependent on $\tilde{P}$, verifying the analysis. For small values of $\tilde{P}$, the MSE is dominated by the model mismatch and almost coincides with the lower bound. For large values of $\tilde{P}$, it is the variance term that dominates, with the MSE increasing proportionally to $\tilde{P}$ and $1/\text{SNR}$ as suggested by (18). A particularly interesting observation is that the optimal (or close to optimal) value of $\tilde{P}$ is no larger than 7 for all cases of SNR, even though P can reach up to 100. In addition, the MSE achieved with the optimal $\tilde{P}$ is lower than the MSE achieved by the standard LS estimator, which equals M/SNR. This clearly suggests that the wireless channel can be safely treated (and modeled) in algorithm development as sparse, i.e., consisting of only a few propagation paths. On the one hand, this can be used towards improving the MSE compared to the LS approach as shown here. On the other hand, it is shown in the next sections how this sparse representation can be exploited towards reducing the training overhead in several multiuser scenarios.

IV. TRAINING SEQUENCE DESIGN AND SCALING LAWS FOR UPLINK WIDEBAND CHANNEL ESTIMATION

In this section, we consider the problem of multiuser channel estimation in the massive MIMO UL wideband scenario ($N \gg 1$). As shown in the previous section, the wireless channel can be treated as sparse for estimation purposes, which suggests incorporation of approaches and techniques from the mature field of CS [37]. Indeed, by reformulating the channel estimation problem in a format compatible to the one considered in CS, a few recent publications have proposed CS-inspired algorithms for UL massive MIMO wideband channel estimation [34], [38], [39]. In these works, it is numerically demonstrated that this approach provides excellent performance with low training overhead. Nonetheless, an analytical characterization of the latter remains missing due to the Kronecker structure of the resulting sensing matrix [cf. (23)]. Building on the sparse channel model from the previous section and introducing the concept of hierarchical sparsity, we present a sufficient asymptotic scaling of the training overhead for reliable channel estimation in the following.

A. SYSTEM MODEL

A BS equipped with a ULA is considered, that arbitrarily partitions its served users to groups of K users, with K an integer parameter to be designed in the following. For UL channel estimation purposes, each user group is assigned an exclusive set of training subcarriers and all users within a group transmit their training symbols on these subcarriers and on the same OFDM symbol. Considering an arbitrary user group in the following, let $\mathcal{T} \subseteq [N]$ denote the set of its $T \triangleq |\mathcal{T}|$ dedicated training subcarriers. We also assume that only $K_a \leq K$ users are active.

Let $\boldsymbol{P}_{\mathcal{T}} \in \{0, 1\}^{T \times N}$ denote the frequency-domain *sampling* matrix obtained by extracting the T rows of $\mathbf{I}_N$ corresponding to $\mathcal{T}$. The $M \times T$ space-frequency observation matrix based on which the BS will estimate the user channels is

$$\boldsymbol{Y} = \sqrt{P_t} \sum_{k=1}^{K} \boldsymbol{H}_k \boldsymbol{P}_{\mathcal{T}}^{\mathrm{T}} \mathcal{D}(\boldsymbol{s}_k) + \boldsymbol{N}, \tag{21}$$

where $\boldsymbol{s}_k \in \mathbb{C}^T$ is the training sequence of user k consisting of unit modulus elements, $\boldsymbol{H}_k \in \mathbb{C}^{M \times N}$ is its channel transfer matrix whose structure follows the model of (4) for the active users, whereas it is a zero matrix for the inactive users, and $\boldsymbol{N} \in \mathbb{C}^{M \times T}$ is the noise matrix of i.i.d. complex Gaussian elements of zero mean and variance σ^2. It is also assumed that each user channel has the same number P of propagation paths that is independent of M and N. The BS knows P and that only K_a users are active, but not exactly which of them (the identity of the active users will be implicitly obtained from the channel estimates). The channels are considered sparse, i.e., P is significantly smaller than N or M. (As discussed in the previous section, in scenarios where P is large, sparse virtual channels can be considered in place of the actual channels with a small degradation of performance.)

B. CHANNEL ESTIMATION FORMULATION AS A COMPRESSIVE SENSING PROBLEM

Operation with an asymptotically large number of antennas M and number of subcarriers N, keeping the subcarrier spacing constant, is considered. Note that this implies a proportional increase of the cyclic prefix length N_{cp}, where it is assumed for simplicity that N/N_{cp} is always an integer. In this regime, the following assumptions are employed.

1) The (normalized) delay τ of an arbitrary channel path of an arbitrary (active) user takes N_{cp} discrete values from the set $\{0, \ldots, N_{cp} - 1\}$ whereas its (azimuth)

angle θ is such that $\sin(\theta)$ takes M discrete values from the set $\{-1, -1+\frac{1}{M}, -1+\frac{2}{M} \ldots, -\frac{1}{M}\}$.

2) No two channel paths belonging to either the same user or different users have the same azimuth angle value.

Although the first assumption is not exact for finite M, N, it can be treated as an approximation that is accurate in the asymptotic regime as the considered discrete sets provide a finely grained sampled version of the continuous delay and angle domains. The second assumption simply reflects the intuition that the angle of arrivals of any two paths, even of the same user, are not expected to be exactly the same.

The benefit of these assumptions is that they allow to express the channel matrix $\boldsymbol{H}_k$ of an arbitrary active user k as

$$\boldsymbol{H}_k = \boldsymbol{F}_M^{\mathrm{H}} \boldsymbol{X}_k \left(\boldsymbol{F}_N^{1:N_{cp}}\right)^{\mathrm{T}}$$

which follows directly from (4) and (5). $\boldsymbol{X}_k \in \mathbb{C}^{M \times N_{cp}}$ is the delay-angular representation of the channel of user k whose (m,n)-th element is non-zero and equal $[\boldsymbol{X}_k]_{m,n} = jc_{p,k}$, where $c_{p,k}$ is the gain of the p-th path, if and only if this path has a delay equal to $(n-1) \in \{0, \ldots, N_{cp}-1\}$ and an angle whose sine is equal to $-1+(m-1)/M$, $m \in [M]$.

As $\boldsymbol{X}_k$ has either $P \ll MN_{cp}$ non-zero (active user) or only zero (inactive user) elements, it is a sparse matrix. This naturally suggests application of compressive sensing methods for algorithm design and performance analysis. In particular, by considering a vectorized version of the observed signal, the system model equation (21) can be equivalently and more conveniently written as

$$\begin{aligned} \boldsymbol{y} &\triangleq (1/\sqrt{TMP_t})\mathrm{vec}(\boldsymbol{Y}^{\mathrm{T}}) && (22)\\ &= (1/\sqrt{TM})(\boldsymbol{F}_M^{\mathrm{H}} \otimes \boldsymbol{A}_\tau)\boldsymbol{x} + \boldsymbol{n}, && (23) \end{aligned}$$

where $\boldsymbol{x} \triangleq \mathrm{vec}\left([\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K]^{\mathrm{T}}\right)$ contains the unknown delay-angle representations of all user channels and is a sparse vector with K_aP non-zero elements, $\boldsymbol{n} \triangleq (1/\sqrt{TMP_t})\mathrm{vec}(\boldsymbol{N}^{\mathrm{T}})$, and

$$\boldsymbol{A}_\tau \triangleq \left[\mathcal{D}(\boldsymbol{s}_1)\boldsymbol{P}_{\mathcal{T}}\boldsymbol{F}_N^{1:N_{cp}}, \ldots, \mathcal{D}(\boldsymbol{s}_K)\boldsymbol{P}_{\mathcal{T}}\boldsymbol{F}_N^{1:N_{cp}}\right]. \quad (24)$$

The motivation for this problem formulation is a well-known result in CS theory [37], which, when applied to this setting, states that, a necessary requirement for perfect recovery of $\boldsymbol{x}$ from $\boldsymbol{y}$ in the absence of noise is [37, Th. 11.6]

$$TM = \mathcal{O}\left(K_aP\log(KN_{cp}M)\right), \text{ for } KN_{cp}M \to \infty. \quad (25)$$

Equation (25) suggests that, in the massive MIMO setting, the necessary training overhead T remarkably scales as $\mathcal{O}(1)$ as N, M increase, i.e., it is independent of the number of users and channel paths. Intuitively, the burden of channel estimation is shifted to the spatial dimension.

However, achievability of the theoretical bound of (25) depends crucially on the *sensing matrix* $(1/\sqrt{TM})(\boldsymbol{F}_M^{\mathrm{H}} \otimes \boldsymbol{A}_\tau)$ that appears in (23). One commonly employed sufficient condition is that the sensing matrix should satisfy the, so called, restricted isometry property (RIP) [37], which would indeed be the case (with high probability) if its elements were, e.g., independent and Gaussian distributed. Unfortunately, there is limited flexibility in designing such a sensing matrix since, by default, it has a Kronecker product structure and the design of the user signatures can only affect the constituent matrix $\boldsymbol{A}_\tau$ under the specific block structure of (24). Due to exactly this issue, a rigorous characterization of the sufficient overhead requirements for massive MIMO is missing in [40].

$$[\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3] = \left[\underbrace{\begin{matrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \times & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \times & \cdot & \cdot & \cdot \end{matrix}}_{\boldsymbol{X}_1} \vdots \underbrace{\begin{matrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{matrix}}_{\boldsymbol{X}_2} \vdots \underbrace{\begin{matrix} \times & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \times & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{matrix}}_{\boldsymbol{X}_3} \right]$$

FIGURE 3. A toy example of a realization of X with $M = 5$, $N_{cp} = 4$, for a case with $K = 3$ users out of which only $K_a = 2$ (first and third) users are active. The channel of each active user has $P = 2$ paths. Zero elements in X are represented by dots (·), whereas the K_aP, not necessarily equal, non-zero elements are represented by the symbol ×. By assumption (which holds in practice for asymptotically large M), no two paths have the same angle, therefore X has exactly K_aP non-zero rows for all channel realizations.

C. EXPLOITING THE CHANNEL HIERARCHICAL SPARSITY

The key to a rigorous characterization of the overhead requirements as well as an algorithm design that achieves them (in a noisy setting) is the observation that for an active user k, under assumption 2, (a) the P non-zero elements of $\boldsymbol{X}_k$ belong strictly to P different rows of $\boldsymbol{X}_k$ and (b) for any other active user $k' \neq k$, the set of non-zero rows of $\boldsymbol{X}_k$ and $\boldsymbol{X}_{k'}$ do not overlap. This results in matrix $[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k]$ having, for every realization of the user channels, exactly K_aP non-zero rows, each with only a single non-zero element. An example is shown in Fig. 3. This, in turn, implies that the unknown vector $\boldsymbol{x}$ appearing in (23) is not simply sparse but its sparsity pattern has a well defined structure. In particular, first note that $\boldsymbol{x} \in \mathbb{C}^{M \cdot K \cdot N_{cp}}$,[7] i.e., $\boldsymbol{x}$ is a 3-level block (compound) vector: it consists of M blocks (level 1) corresponding to the range of channel angle values, each of which consists of K blocks (level 2) corresponding to the K users, each of which consists of N_{cp} elements (level 3) corresponding to the range of delay values. It is easy to see that only K_aP level-1 blocks are non-zero, each non-zero level-1 block contains only 1 level-2 non-zero block, and each non-zero level-2 block has only 1 non-zero level-3 element.

We say that $\boldsymbol{x}$ is $(K_aP, 1, 1)$-hierarchically-sparse in $\mathbb{C}^{M \cdot K \cdot N_{cp}}$ (or simply $(K_aP, 1, 1)$-hi-sparse). Note that the notion of hierarchical sparsity is a restriction of the standard notion of sparsity: A hierarchically sparse vector is sparse but the reverse does not necessarily hold. An example of an $(2, 1, 1)$-hi-sparse vector in $\mathbb{C}^{5 \cdot 2 \cdot 3}$ is shown in Fig. 4 (bottom vector).

Clearly, the hierarchical sparsity of $\boldsymbol{x}$ is a property that *should be exploited* in algorithm design and analysis as it provides significant restrictions on its support, compared to the standard notion of sparsity. Towards this end, the

[7] Recall the notation of multi-level block vectors described in Sec. I-C.

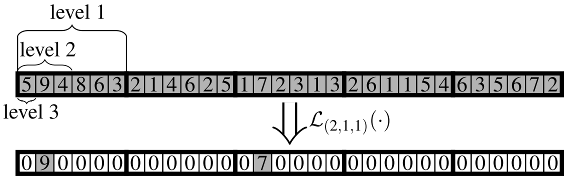

FIGURE 4. A 3-level block vector $\tilde{x} \in \mathbb{C}^{5\cdot 2\cdot 3}$ and its best (2, 1, 1)-hi-sparse approximation $\mathcal{L}_{(2,1,1)}(\tilde{x})$. Note that $\mathcal{L}_{(2,1,1)}(\tilde{x})$ is different from the best least-squares approximation of $\tilde{x}$ by a vector with 2 non-zero elements.

low-complexity, iterative hard thresholding (IHT) algorithm [37] is modified as shown in Algorithm 1, referred to as hierarchical IHT (HiIHT).

Algorithm 1 HiIHT Channel Estimation

```
Require: $\boldsymbol{y}, \boldsymbol{A} \triangleq (1/\sqrt{TM})(\boldsymbol{F}_M^{\mathrm{H}} \otimes \boldsymbol{A}_\tau), K, P$.
$i = 0, \hat{\boldsymbol{x}}^{(0)} = \boldsymbol{0} \in \mathbb{C}^{M\cdot K\cdot N_{cp}}$
repeat
    $i = i + 1$,
    $\hat{\boldsymbol{x}}^{(i)} = \mathcal{L}_{(K_a P,1,1)}\left(\hat{\boldsymbol{x}}^{(i-1)} + \boldsymbol{A}^{\mathrm{H}}\left(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}^{(i-1)}\right)\right)$
until stopping criterion is met at $i = i^*$
return $(K_a P, 1, 1)$-hi-sparse $\hat{\boldsymbol{x}}^{(i^*)}$
```

Within iteration i of HiIHT, two steps are performed. First, a standard gradient descent step towards decreasing the quadratic error $\|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}^{(i-1)}\|^2$ of the previous iteration estimate is computed. However, since the result will in general be non-sparse, operator $\mathcal{L}_{(K_a P,1,1)}(\cdot)$ is subsequently applied, which computes the least-squares projection of an arbitrary vector onto the space of $(K_a P, 1, 1)$-hi-sparse vectors. A toy example of the action of $\mathcal{L}_{(K_a P,1,1)}(\cdot)$ for the case $M = 5, K = K_a = 2, N_{cp} = 3, P = 1$, is presented in Fig. 4. As discussed in [11], the projection operation can be performed very efficiently with negligible computational cost compared to the operations required by gradient descent computation.

D. TRAINING DESIGN AND OVERHEAD SCALING

The HiIHT algorithm was presented for an arbitrary selection of $\mathcal{T}$ and $\{\boldsymbol{s}_k\}_{k\in[K]}$. Towards (optimal) system design, a rigorous characterization of the HiIHT performance was obtained in [11], identifying the, so-called, Hierarchical RIP (HiRIP) constant of the sensing matrix as a key parameter for achieving reliable channel estimation. The HiRIP constant can be considered as another one of the many constants characterizing a matrix (such as, e.g., the condition number). In particular, it provides an indication of the effect a matrix has when applied to hierarchically sparse multilevel vectors and is a specialization of the well-known RIP constant in CS theory [37].

With δ denoting the HiRIP constant of A and with $M, N, N_{cp} \gg P$, the sequence of estimates $\{\hat{\boldsymbol{x}}^{(i)}\}$ generated by the HiIHT algorithm satisfies [11]

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}^{(i)}\| \leq \left(\sqrt{3}\delta\right)^i \|\boldsymbol{x}\| + \frac{2.18}{1 - \sqrt{3}\delta}\|\boldsymbol{n}\|, \quad i \geq 0, \tag{26}$$

as long as

$$\delta < 1/\sqrt{3}. \tag{27}$$

It follows that any training design resulting in a sensing matrix whose HiRIP constant satisfies (27) is sufficient for HiIHT to achieve reliable *channel estimation*, in the sense of perfect recovery in the noiseless ($\|\boldsymbol{n}\| = 0$) case and bounded error in the noisy ($\|\boldsymbol{n}\| > 0$) case. One such design was shown in [11] to be that with

- $K = N/N_{cp}$ users per group,
- the T training subcarriers of a group selected randomly and uniformly out of the N total subcarriers, and
- *frequency shifted* user training sequences of the form

$$\boldsymbol{s}_k = \boldsymbol{P}_{\mathcal{T}}\mathcal{D}\left(\boldsymbol{f}_N^{\mathrm{H}}((k-1)N_{cp})\right)\boldsymbol{s}, k \in [K] \tag{28}$$

  where $\boldsymbol{s} \in \mathbb{C}^N$ is an arbitrary vector of unit-modulus elements.

Even though the design is not obtained as the solution of an optimization problem and, therefore, cannot be claimed to be optimal, it has the benefit of a very efficient HiIHT implementation. It is easy to see that $\boldsymbol{A}_\tau = \boldsymbol{F}_N$ under this design, which allows for computation of the gradient step in HiIHT by means of a two-dimensional FFT with a complexity order $\mathcal{O}(MN\log(MN))$ per iteration. This allows for application with (very) large N and/or M and one such case will be demonstrated in the next section. In addition, this design allows to obtain rigorous sufficient conditions for the training overhead required for reliable massive MIMO channel estimation. In particular, it can be shown that for asymptotically large M, N, a training overhead

$$T \geq \min\left\{C\log^4(N), N\right\}, \tag{29}$$

where C is a constant, is sufficient [11].

It is noted that the overhead scaling of (29) is only sufficient and, therefore, may be larger that the necessary and sufficient one (which is unknown). Even so, this result indicates that reliable channel estimation is possible with a training overhead that is independent of both K_a and P, similar to what was suggested by the theoretical bound of (25), which is particularly appealing as it implies a robust training design without the need for training reconfiguration with changing K_a and/or P. Essentially, this result verifies the intuition that the massive number of antennas and corresponding observations can compensate for a small training overhead that would be inadequate in a conventional (e.g., single antenna) setting. This indicates the significant advantage of employing massive MIMO in terms of reducing the training overhead for channel estimation.

E. NUMERICAL RESULTS

In order to demonstrate the analytical insights presented in the previous subsection, the mean square error (MSE) performance of HiIHT, defined as $\sum_{k=1}^{K}\mathbb{E}(\|\boldsymbol{H}_k - \hat{\boldsymbol{H}}_k\|^2)/MN$ is numerically evaluated. The channel transfer matrix estimate of user k is obtained as $\hat{\boldsymbol{H}}_k \triangleq \boldsymbol{F}_M^{\mathrm{H}} \hat{\boldsymbol{X}}_k (\boldsymbol{F}_N^{1:N_{cp}})^{\mathrm{T}}$ based on the HiIHT estimate $\hat{\boldsymbol{X}}_k$ of the corresponding delay-angular channel representation. A setting with $N = 1024$ subcarriers, $M = 256$ antennas, $N_{cp} = 256$ cyclic prefix samples and $P = 3$ propagation paths per user channel was considered (results are similar for any other parameters such that $M, N, N_{cp} \gg P$). The delay/angle values of each active user channel were randomly and uniformly generated satisfying the assumptions described in subsection IV-B, whereas the gains were generated as independent complex Gaussian random variables of zero mean and variance $1/P$ (all active users experience the same SNR). The SNR is set equal to 10 dB.

Figure 5 shows the MSE performance with varying T for the case of $K = N/N_{cp} = 4$ and variable number of randomly selected active users K_a. It can be seen, that, in all cases, performance is excellent with a training overhead in the order of only about 1% the number of subcarriers. This value should be compared with the overhead of about $D/N\% = 25\%$ required in conventional single antenna OFDM channel estimation without exploiting channel sparsity [41]. Note that estimation performance degrades as K_a increases, which is expected as there are more parameters to be estimated [17]. However, this degradation is graceful and, most importantly, the minimum training overhead required for reliable estimation is the same for all K_a, in line with the analytical insights of the previous subsection.

In addition, the performance with the same system parameters as before but with $K = K_a = 1$ is also shown for both HiIHT as wel as IHT (that latter does not take into account the hierarchically sparse property of the channel). It can be seen that HiIHT performance improves in the sense that the minimum number of training sequences to achieve good performance is reduced. This is because a reduced number of parameters are estimated and there is no ambiguity in which user is active. IHT is seen to require significant larger minimum training overhead than HiIHT, signifying the importance of exploiting the channel structure for its estimation. Note that for sufficiently large T, IHT and HiIHT perform the same as the number of observations is so large that the prior structural information employed by HiIHT becomes insignificant.

V. TRAINING SEQUENCE DESIGN AND SCALING LAWS FOR DOWNLINK NARROWBAND CHANNEL ESTIMATION

In the following, we begin our treatment of Bayesian methods, i.e. methods exploiting user covariance information at the BS (and possibly at the user). We note that the analysis, although applicable to the channel model of (4) and corresponding covariance matrix structure of (12), is actually more general and applies to arbitrary channel models.

Recall that the potential of reduced training overhead in the Bayesian setup was observed in the single-user case in Sec. II-B when the low-rank channel covariance matrix is known at the BS. Even though training overhead reduction, compared to a naive LS estimation, is expected in a multiuser scenario by an approach where the BS sequentially transmits per-user optimized training sequences, one expects that optimization of the training sequences jointly over the user covariance matrices (e.g., exploiting any possible overlaps between user covariance range spaces) would provide greater gains.

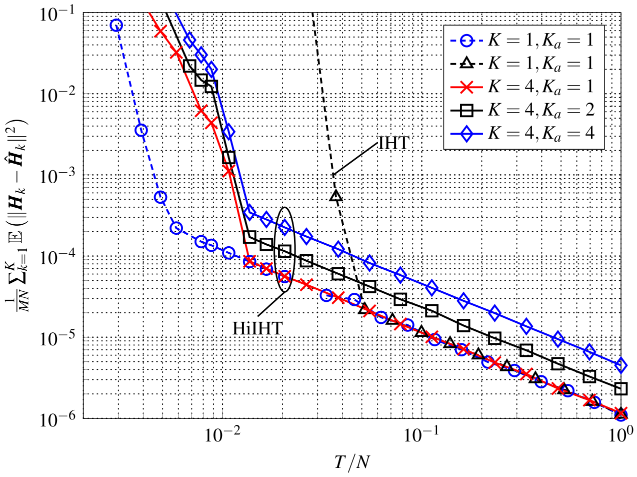

FIGURE 5. **Multiuser channel MSE of HiHTP estimator as a function of training overhead ($N = 1024$ subcarriers, $M = 256$ antennas, $N_{cp} = 256$ cyclic prefix samples, $P = 3$ propagation paths per user channel,** SNR **= 10dB). K_a ($\leq K$) refers to the number of active users.**

This principle is investigated and formalized in this section under a Bayesian channel estimation framework for a narrowband DL system (cf. Section III) serving $K \leq M$ users and all the channel covariance matrices $\boldsymbol{C}_k$, $k = 1, \ldots, K$, available at the BS (e.g., by means of a feedback mechanism or exploiting statistical reciprocity of UL and DL channels), whereas each user knows the covariance matrix of its own channel alone. Two different approaches for the training design are considered: MSE-aware and rate-aware. The treatment is this section is especially relevant for frequency-division-duplex (FDD) systems. In case there are errors in estimating the covariance matrices, then the presented analysis [e.g. (33)] does not exactly hold. Nonetheless, as covariance matrices are in general constant over multiple coherence intervals, they can be estimated very accurately even under the presence of inter-cell interference (see, e.g., [42], [43]). Any small residual errors would not affect the validity of the performed analysis.

A. MSE-AWARE TRAINING DESIGN

With covariance matrix information at the users, LMMSE channel estimation is possible. For the k-th user, its estimated channel equals

$$\hat{\boldsymbol{h}}_k = \boldsymbol{W}_k \boldsymbol{y}_k \tag{30}$$

where $\boldsymbol{y}_k \in \mathbb{C}^T$ is the observed signal at user k as given in (13) and

$$\boldsymbol{W}_k \triangleq \boldsymbol{C}_k \boldsymbol{S} \left(\boldsymbol{S}^{\mathrm{H}} \boldsymbol{C}_k \boldsymbol{S} + \sigma^2 \mathbf{I}_T \right)^{-1} \tag{31}$$

is the LMMSE estimator of user k. The corresponding MSE equals

$$\mathrm{MSE}_k = \mathrm{tr}(\boldsymbol{C}_{\epsilon,k}), \tag{32}$$

where $\boldsymbol{C}_{\epsilon,k} \triangleq \mathbb{E}\left((\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k)(\boldsymbol{h}_k - \hat{\boldsymbol{h}}_k)^{\mathrm{H}} \right)$ is the covariance matrix of the channel estimation error given by [17]

$$\boldsymbol{C}_{\epsilon,k} = \boldsymbol{C}_k - \boldsymbol{C}_k \boldsymbol{S} \left(\boldsymbol{S}^{\mathrm{H}} \boldsymbol{C}_k \boldsymbol{S} + \sigma^2 \mathbf{I}_M \right)^{-1} \boldsymbol{S}^{\mathrm{H}} \boldsymbol{C}_k. \tag{33}$$

Towards jointly optimizing the training matrix $\boldsymbol{S}$ over all users, a reasonable objective function is the sum MSE $\sum_{k=1}^{K} \mathrm{MSE}_k$. As this is a well-defined function of $\boldsymbol{S}$, the training matrix can be optimized for any pre-selected training overhead T. Unfortunately, the optimal $\boldsymbol{S}$ cannot be found in closed form, rendering a numerical approach necessary. Although sophisticated iterative algorithms that show good performance have been proposed for that purpose [44], [45], they provide insights on neither the structure of the optimal $\boldsymbol{S}$ and its explicit dependency on $\{\boldsymbol{C}_k\}_{k=1}^{K}$, nor on the minimum training duration that achieves a vanishing channel estimation sum MSE as the training SNR $P_{\mathrm{t}}/\sigma^2 \rightarrow \infty$.[8]

Towards obtaining such analytical insights, it is easy to see from (32) and (33) that the training matrix that minimizes MSE_k of an arbitrary user k is of the form

$$\boldsymbol{S} = \sqrt{P_t} (\boldsymbol{U}_{\boldsymbol{C}_k}^{1:T}) \boldsymbol{Q} \tag{34}$$

for any given T, where $\boldsymbol{Q} \in \mathbb{C}^{T \times T}$ is an arbitrary unitary matrix (see also [30]). Substituting (34) into (31) results in

$$\boldsymbol{W}_k = \boldsymbol{S}\,\mathcal{D}\left(\frac{\lambda_{\boldsymbol{C}_k,1}}{\lambda_{\boldsymbol{C}_k,1} + \frac{\sigma^2}{P_t}}, \ldots, \frac{\lambda_{\boldsymbol{C}_k,T}}{\lambda_{\boldsymbol{C}_k,T} + \frac{\sigma^2}{P_t}} \right) \rightarrow \boldsymbol{S} \ \text{ as } \ P_t/\sigma^2 \rightarrow \infty. \tag{35}$$

It follows that when $\boldsymbol{S}$ is optimized to minimize MSE_k in the single-user case, the LMMSE estimator of user k reduces to $\boldsymbol{S}$ in the large SNR regime. This motivates a heuristic design for the multiuser case where all users, instead of the LMMSE estimator (31), simply use $\boldsymbol{W}_k = \boldsymbol{S}, k = 1, 2, \ldots, K$, and $\boldsymbol{S}$ is to be optimized by the BS. Even though this approach is suboptimal, it simplifies the design of $\boldsymbol{S}$ and provides insights on the minimum sought training overhead as shown next.

Defining $\boldsymbol{C}_{\mathrm{sum}} \triangleq \sum_{k=1}^{K} \boldsymbol{C}_k$, the sum MSE with $\boldsymbol{W}_k = \boldsymbol{S}$ reads [12]

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}\left(\|\boldsymbol{h}_k - \boldsymbol{S}\boldsymbol{y}_k\|^2 \right) &= \mathrm{tr}(\boldsymbol{C}_{\mathrm{sum}}) - \mathrm{tr}\left(\boldsymbol{S}^{\mathrm{H}} \boldsymbol{C}_{\mathrm{sum}} \boldsymbol{S} \right) + KT \frac{\sigma^2}{P_t} \\ &\geq \sum_{i=T+1}^{M} \lambda_{\boldsymbol{C}_{\mathrm{sum}},i} + KT \frac{\sigma^2}{P_t},\end{aligned} \tag{36}$$

where the lower bound (36) is achieved with

$$\boldsymbol{S} = \sqrt{P_t} (\boldsymbol{U}_{\boldsymbol{C}_{\mathrm{sum}}}^{1:T}) \boldsymbol{Q} \tag{37}$$

where $\boldsymbol{Q} \in \mathbb{C}^{T \times T}$ is an arbitrary unitary matrix.[9]

The achieved lower bound (36) gives interesting insights on the sought T in the asymptotically high SNR operation. As $P_t/\sigma^2 \rightarrow \infty$, the second term vanishes and it is readily seen that setting $T = \mathrm{rank}(\boldsymbol{C}_{\mathrm{sum}})$ drives the first term and thus (36) to 0. Denoting $\mathrm{rank}(\boldsymbol{C}_k) = r_k$, it holds that

$$\max(r_1, \ldots, r_K) \leq \mathrm{rank}(\boldsymbol{C}_{\mathrm{sum}}) \leq \min\left\{ \sum_k r_k, M \right\}. \tag{38}$$

Note that the lower bound corresponds to the case where the covariance range spaces of users with the smaller covariance ranks completely lie in the covariance range space of the user with the largest covariance rank. In practical systems, this may happen when all users are located in a relatively small area and experience similar large-scale fading conditions. The upper bound holds by the rank subadditivity property and corresponds to the case where the covariance range spaces are orthogonal, or do overlap but cover the complete $\mathbb{C}^M$ space. This suggests potentially much smaller training overhead than M for smaller K, smaller covariance ranks, and larger overlap of covariance range spaces.

Since the LMMSE estimator is the optimal linear estimator for any training matrix, it follows that using (37) with $T = \mathrm{rank}(\boldsymbol{C}_{\mathrm{sum}})$ in combination with LMMSE estimators also drives the channel estimation sum MSE to zero asymptotically. In fact, $\mathrm{rank}(\boldsymbol{C}_{\mathrm{sum}})$ is an upper bound of the minimum training duration when LMMSE filters are used. In addition, numerical evaluation of the actual sum MSE with LMMSE estimators and the (suboptimal) training design of (37) shows practically the same performance with the sum MSE achieved by the design obtained by direct numerical minimization of the sum LMMSE w.r.t $\boldsymbol{S}$ [12].

B. RATE-AWARE TRAINING DESIGN

The (sum) MSE metric considered in the previous subsection for the design of $\boldsymbol{S}$ is a robust and reasonable measure to evaluate the channel estimation accuracy. However, in many wireless communications' applications, the achievable data rate is the main metric of interest. Therefore, it is natural to consider training designs that directly operate on the data rate. Towards this end, consider the received signal at an arbitrary user k during the data transmission phase

$$y_{k,\mathrm{data}} = \boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{z}_k d_k + \sum_{l=1, l \neq k}^{K} \boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{z}_l d_l + n_k \tag{39}$$

where $d_k \in \mathbb{C}$ and $\boldsymbol{z}_k \in \mathbb{C}^M$ are the data symbol and precoder for transmission to user k and n_k a zero mean com-

[8] This problem is not of interest in the uncorrelated channel entries case, as the corresponding minimum training duration is given by M.

[9] The resulting sum MSE depends linearly on T in the finite SNR regime, which is uncommon. This is due to the suboptimal design and is discussed in detail in [12]. Since the main goal here is to characterize the minimum training duration in the asymptotic SNR regime, this issue plays no role in subsequent derivations.

plex Gaussian noise of variance σ^2. The precoders are power normalized such that $\sum_{k=1}^{K} \|z_k\|^2 = P_t$, i.e. same transmit power in both data transmission and training phases. We also make the assumption that the channels follow a Gaussian distribution in this subsection.

The first term in (39) is the intended signal towards user k while the second term represents inter-user interference. As expected, its properties depend on the employed precoding vectors, which are selected based on the channel state information (CSI) available at the BS. For an FDD system, the CSI is obtained by means a of training phase where the users estimate the channels, followed by a feedback phase where the users inform the BS of their channel estimates. Since both steps are prone to errors, they result in imperfect CSI at the BS and consequently reduced transmitted rates due to the precoder mismatch with the actual channels.

In order to focus on the effect of channel estimation errors, we assume perfect feedback, i.e., the BS knows $\{\hat{\boldsymbol{h}}_k\}_{k=1}^{K}$. It can then be shown [13] that the the rate loss ΔR_k experienced by user k compared to the ideal CSI case can be expressed as the sum of two terms: a dominant term due to inter-user interference ($\Delta_{\text{int}} R_k$), and a minor term due to the beamforming loss ($\Delta_{\text{beamf}} R_k$). When zero-forcing (ZF) precoders are employed, the average rate loss (over all possible channel realizations) due to inter-user interference is upper bounded by [13]

$$\mathbb{E}\left(\Delta_{\text{int}} R_k\right) < \log_2\left(1 + P_t\, \lambda_{\boldsymbol{C}_{\epsilon,k},1}/\sigma^2\right) \text{bits/sec/Hz} \tag{40}$$

where $\lambda_{\boldsymbol{C}_{\epsilon,k},1}$ is the largest eigenvalue of the channel estimation error covariance matrix [cf. (33)]. In order to avoid a rate saturation effect for user k due to inter-user interference at high SNR, the bound of (40) reveals that it is sufficient to scale $\lambda_{\boldsymbol{C}_{\epsilon,k},1}$ inversely proportional to the SNR P_t/σ^2, i.e., it is sufficient for the condition

$$\lambda_{\boldsymbol{C}_{\epsilon,k},1} \leq c\,\sigma^2/P_t \tag{41}$$

to hold, where $c = \mathcal{O}(1)$ is a constant. Under this condition, it is further shown in [13] that the average beamforming loss $\mathbb{E}(\Delta_{\text{beamf}} R_k)$ converges to zero as $P_t/\sigma^2 \to \infty$. This results in the average rate loss for user k upper bounded as

$$\mathbb{E}(\Delta R_k) < \log_2(1 + c)\text{bits/sec/Hz} \tag{42}$$

for asymptotically large SNR, given that (41) holds. Nonetheless, it is numerically observed that the bound holds for low SNR values as well.

By joint consideration of all K user rates, the training design problem can be posed as the identification of the (possibly non-unique) $\boldsymbol{S} \in \mathbb{C}^{M\times T}$ with the minimum T that results in (41) holding for all k. Though this problem is challenging and has no closed-form solutions in general, an intuitive suboptimal solution can be obtained. Similar to the previous section, the starting point is again the consideration of an arbitrary user k and the observation that the optimal training matrix that achieves (41) is equal to $\boldsymbol{S} = \sqrt{P_t}\boldsymbol{U}_{\boldsymbol{C}_k}^{1:T_k}$, for some sufficiently large T_k.[10] Then, noting that (41) is also satisfied by an augmented matrix $[\boldsymbol{U}_{\boldsymbol{C}_k}^{1:T_k}, \boldsymbol{S}']$ for any arbitrary $\boldsymbol{S}'$ with M rows,[11] a training design that satisfies (41) for all k is

$$\boldsymbol{S} = \sqrt{P_t}\left[\boldsymbol{U}_{\boldsymbol{C}_1}^{1:T_1}, \boldsymbol{U}_{\boldsymbol{C}_2}^{1:T_2}, \ldots, \boldsymbol{U}_{\boldsymbol{C}_K}^{1:T_K}\right]. \tag{43}$$

This simple design procedure, essentially exploiting the per-user channel correlation structure, requires a training overhead $T = \sum_{k=1}^{K} T_k$, which can be much smaller than M for small K and/or $\{T_k\}_{k=1}^{K}$ and/or low SNR values.

Going one step further it was observed in [13] that further training overhead gains can be achieved by also exploiting the cross user correlation structure. In particular, it was shown that if some $\boldsymbol{S}$ satisfies condition (41), then so does any other $\boldsymbol{S}'$ satisfying $\text{ran}(\boldsymbol{S}) \subseteq \text{ran}(\boldsymbol{S}')$. It follows that if the range spaces of the constituent matrices of the design in (43) partially overlap, further reduction of the training overhead can be achieved by considering a training matrix that satisfies

$$\text{ran}(\boldsymbol{S}) = \text{ran}\left(\left[\boldsymbol{U}_{\boldsymbol{C}_1}^{1:T_1}, \ldots, \boldsymbol{U}_{\boldsymbol{C}_K}^{1:T_K}\right]\right), \tag{44}$$

ultimately resulting in a training overhead equal to the dimension of this range space, and that can be potentially be much smaller than $\sum_k T_k$. Note that T_k increases with increasing SNR; therefore, T has to be an increasing function of SNR as well, if (42) were to hold for all SNR values.

Since $T_k \leq \text{rank}(\boldsymbol{C}_k)$, the main take away is that it is possible to further reduce the training overhead (compared to the design of Sec. V-A) if the inter-user interference power resulting from not training some carefully chosen covariance eigenvectors (namely, eigenvectors $T_k + 1, \ldots, \text{rank}(\boldsymbol{C}_k)$) is comparable to or smaller than the noise power. This is illustrated numerically in Sec. V-D.

C. REMARKS ON THE FULL RANK COVARIANCE MATRIX CASE

For completeness, we address the case of full rank covariance matrices. Here, the feasibility of the proposed schemes highly depends on the eigenvalue spread. If the latter is large and many eigenvalues are (very) small, then a numerical rank r_{num} of $\boldsymbol{C}_{\text{sum}}$ can be defined as shown in the next section. The value of r_{num} can be (much) smaller than M. Training along the r_{num} eigenvectors of $\boldsymbol{C}_{\text{sum}}$ provides a satisfactory performance in the finite SNR regime. Initial measurement campaigns corroborate the large eigenvalue spread property [46], [47]. In case r_{num} is close to M and/or the eigenvalue spread is not large, then the scheme of Sec. V-A may not provide large overhead reductions. Nonetheless, the rate-aware scheme of Sec. V-B can still be used to reduce the training overhead for practical SNR values. One such example is discussed in detail in [13, Sec. VI-F].

[10] Up to a unitary matrix ambiguity which can w.l.o.g. be set to the identity matrix.

[11] This follows intuitively as introducing more training slots can only improve channel estimation performance with standard LMMSE estimators. A formal proof is provided in [13].

D. NUMERICAL RESULTS

We consider a BS with a UPA of $M = 128$ elements distributed in 8 rows and 16 columns and mounted at an elevation of $h = 50$ m. The antenna element spacing is set to $\lambda/2$ in both horizontal and vertical directions. The BS serves $K = 8$ users present in a 120° sector with mean angles of arrivals $\{-52.5^\circ, -37.5^\circ, \ldots, 52.5^\circ\}$ in the azimuth direction. User k is $150 + 25 \bmod(k-1, 4)$ meters away from the BS and a path loss exponent of 2 is considered. We use the Laplacian angular spectrum (LAS) correlation model [48] with horizontal and vertical standard deviations of 12 and 6 degrees, respectively. The resulting covariance matrices have many small eigenvalues but no eigenvalues which exactly equal 0; thus, a numerical rank has to be defined similarly to [12]. Here, we define the numerical rank as the smallest number of eigenvalues that contribute at least 99.99% of the channel power (covariance trace). The resulting numerical covariance ranks are given by 32, 35, 37, 39, 40, 42, and 45.

In Fig. 6, we plot the sum rate of all users obtained by the MSE-based training solution of Sec. V-A for the derived fixed training overhead $T = \text{rank}(\boldsymbol{C}_{\text{sum}}) = 51$ slots for various SNR values, where the rank is obtained numerically as described above. Further, we plot the sum rate of the rate-based training solution of Sec. V-B with $c = 1$, i.e. a training solution resulting in an average rate loss smaller than 1 bit/sec/Hz for each user. LMMSE estimators are employed by users in both cases. In addition, the perfect CSI sum rate and the sum rate lower bound (obtained as the perfect CSI sum rate minus $K \log_2(1 + c)$) are plotted. Fig. 7 plots the corresponding training overheads of both proposed designs. When looking at the results of the MSE aware design of Sec. V-A, one can see that exploiting the low-rank covariance structure and the intersection between the covariance range spaces allows obtaining accurate channel estimates that closely follow the perfect CSI sum rates with a much smaller training overhead (51) than the number of BS antennas (128). Note that a per-user optimized sequence solution does not provide any training overhead reductions here, since $\sum_k \text{rank}(\boldsymbol{C}_k) > M$. This stresses the importance of exploiting intersections of covariance range spaces.

However, even greater overhead reductions can be achieved by the design of Sec. V-B as depicted in Fig. 7 and with only negligible performance degradation as the design is explicitly matched to the achievable rate and operating SNR. Note that this design has an adaptive training overhead as previously discussed. Further, even though operation in finite SNR is considered, the sum rate loss compared to the ideal CSI case is contained within the asymptotic upper bound of (42). As a final remark, we note that Fig. 6 highlights the effect of CSI quality on precoding, and does not explicitly take the training overhead needed to obtain the CSI into account when effective spectral efficiencies are calculated. Therefore, the design of Sec. V-B would be superior to the one of Sec. V-A in any finite coherence block regime when this is considered. The next section will show results in the finite coherence block regime to further highlight this aspect.

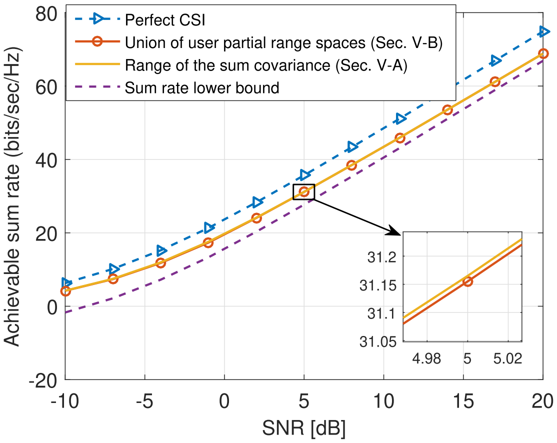

FIGURE 6. Comparison of different proposed training designs exploiting covariance information for a scenario with $K = 8$ users, $M = 128$ BS antennas. The corresponding training overheads are shown in Fig. 7.

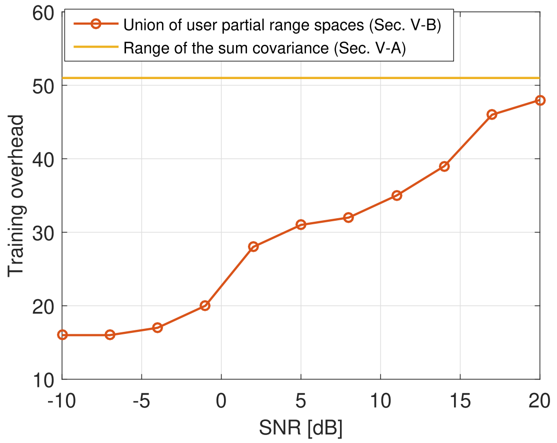

FIGURE 7. Training overheads of the proposed designs.

VI. A SPATIAL BASIS COVERAGE APPROACH FOR UL TRAINING AND SCHEDULING

The previous section illustrated how the partial intersection between covariance range spaces can be exploited to reduce the training overhead in a DL multiuser scenario. In this section, we continue our treatment of Bayesian methods and present a scenario where another property of covariance matrices can be exploited for training overhead reduction, namely the (approximate) orthogonality between the dominant eigenvectors of carefully selected users.

A. SYSTEM MODEL

Consider a narrowband UL multiuser, multi-cell setting of $N_c > 1$ cells. Recall that in the single-cell case, a single channel use ($T = 1$) is sufficient for an accurate channel estimate at the BS, when a single user is considered. However, when $K > 1$ users are served by the cell, a naive LS estimation approach leads to a minimum overhead $T = K$ and, ultimately, $T = N_c K$ when inter-cell interference is significant. This overhead may be unacceptable in massive connectivity scenarios and ultra dense networks. Towards reducing the overhead in these scenarios, Xie *et al.* [49] have proposed a greedy scheduling approach that allocates the same training sequences to spatially separated users, while imposing guard intervals between their spatial signatures. The analysis in [50] confirms the soundness of exploiting the channel spatial structure to reduce the impact of pilot contamination. Therein, the authors showed that, using multi-cell MMSE precoding/combining, massive MIMO capacity increases without bound as the number of antennas increases when the same training sequence is allocated to users with linearly independent covariance matrices. A user assisted clustering was proposed in [51], where a DL probing phase is used to obtain the cluster information from the users based on the instantaneous channels. This scheme thus requires control information to be sent in the DL and UL directions. A location-aware pilot allocation algorithm that exploits the behavior of line-of-sight (LOS) interference among users was proposed in [52]. Therein, a low-complexity algorithm was developed to allocate the same pilot sequence to the users with small LOS interference.

Similar to the treatment in the previous section, we exploit the channel covariance information in the following towards improving the channel estimation performance while keeping the control information at a minimum. In addition, we propose a novel grouping scheme to achieve higher training sequence reuse, while discarding the imposed separation in [49] for the sake of full exploitation of the angular range.

We do not optimize the training sequences as was done in the previous sections. Instead, we assume that the training sequence of each user is selected from an arbitrary set of $T \leq K$ orthogonal training sequences $\{\boldsymbol{s}_t\}_{t=1}^{T}$, each of length T and norm P_t. Considering the single-cell scenario first (no inter-cell interference), the received signal $\boldsymbol{Y} \in \mathbb{C}^{M\times T}$ at the BS during the training phase equals

$$\boldsymbol{Y} = \textstyle\sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{s}_{\chi(k)}^{\mathrm{H}} + \boldsymbol{N} \tag{45}$$

where $\chi : \{1,\ldots,K\} \rightarrow \{1,\ldots,T\}$ is a mapping function that assigns users to training sequences and $\boldsymbol{N} \in \mathbb{C}^{M\times T}$ is the noise matrix with i.i.d. complex Gaussian entries as in the previous sections. A straightforward LS estimation of $\boldsymbol{h}_k$ equals

$$\begin{aligned}\hat{\boldsymbol{h}}_k &= \frac{1}{\sqrt{P_t}}\boldsymbol{Y}\boldsymbol{s}_{\chi(k)} \\ &= \boldsymbol{h}_k + \sum_{\substack{l\neq k,\\ \chi(l)=\chi(k)}} \boldsymbol{h}_l + \frac{1}{\sqrt{P_t}}\boldsymbol{N}\boldsymbol{s}_{\chi(k)}.\end{aligned} \tag{46}$$

It can be seen that the channel estimate is not only affected by noise but also by pilot contamination due to multiple users using the same training sequence when $K > T$. Clearly, pilot contamination can be completely eliminated when $T = K$ and with a simple user-to-sequence assignment $\chi(k) = k$. However, when a value of $T < K$ can only be afforded, an assignment rule that minimizes pilot contamination is necessary. The problem of finding this rule is considered next.

B. EXPLOITING CHANNEL STRUCTURE FOR OPTIMAL TRAINING ASSIGNMENT

As in the previous section, we exploit the structural properties of the user channels as captured by their covariance matrices. In particular, it follows from (10) that the BS, instead of the estimation of $\boldsymbol{h}_k$, may equivalently consider the estimation of the KL coefficients $\boldsymbol{h}_{k,\mathrm{KL}} \in \mathbb{C}^{\mathrm{rank}(\boldsymbol{C}_k)}$ that can be obtained from (46) and (10) as

$$\begin{aligned}\hat{\boldsymbol{h}}_{k,\mathrm{KL}} &= \boldsymbol{U}_{\boldsymbol{C}_k}^{\mathrm{H}}\hat{\boldsymbol{h}}_k \\ &= \boldsymbol{h}_{k,\mathrm{KL}} + \boldsymbol{U}_{\boldsymbol{C}_k}^{\mathrm{H}}\left(\sum_{\substack{l\neq k,\\ \chi(l)=\chi(k)}} \boldsymbol{U}_{\boldsymbol{C}_l}\boldsymbol{h}_{l,\mathrm{KL}} + \frac{1}{\sqrt{P_t}}\boldsymbol{N}\boldsymbol{s}_{\chi(k)}\right)\end{aligned} \tag{47}$$

where we have assumed zero-mean channels w.l.o.g. Denoting $\mathcal{G}_t \subseteq \{1,2,\ldots,K\}$ the group (set) of users that are assigned training sequence t, it directly follows from (47) that the training interference affecting the channel estimates for the users in this group is (approximately) zero if the range space of the matrices $\{\boldsymbol{U}_{\boldsymbol{C}_k}\}_{k\in\mathcal{G}_t}$ are (approximately) mutually orthogonal. This observation strongly motivates assigning users with approximately orthogonal covariance eigenspaces the same training sequence.

We apply this approach in the context of a BS equipped with a ULA. In this context, the eigenvectors of each channel covariance converge to columns of the DFT matrix $\boldsymbol{F}_M$ as $M \rightarrow \infty$ [53]. For finite and large M this provides a good approximation; thus, it is convenient to consider this DFT eigenspace structure and approximate the dominant eigenvectors of $\boldsymbol{C}_k$ with the following set of vectors, referred to as spatial signature:

$$\hat{\boldsymbol{U}}_{\boldsymbol{C}_k} \triangleq \left\{\boldsymbol{f}_M(s) : \frac{Q_{k,s}}{\mathbb{E}[||\boldsymbol{h}_k||_2^2]} \geq \alpha, s = 0,1,\ldots,M-1\right\} \tag{48}$$

where $Q_{k,s} \triangleq \boldsymbol{f}_M^{\mathrm{H}}(s)\boldsymbol{C}_k\boldsymbol{f}_M(s)$ quantifies the average energy of $\boldsymbol{h}_k$ that is aligned to $\boldsymbol{f}_M(s)$ and $\alpha \in (0,1)$ is a tunable threshold parameter. The user grouping performed by the BS will be based on their spatial signatures. The value of α quantifies the number of DFT beams that will be included in each user spatial signature. Defining a value for α amounts to addressing an important trade-off between accuracy and complexity. On one hand, a small value of α will result in high dimension spatial signatures that can convey more information about the channel covariance matrix. However, any subsequent optimization will be impacted by an increase in complexity. On the other hand, a high value for α simplifies the representation of the channel spatial structure by including only a few strongest beams. This results in a relatively poor representation of the channel spatial domain structure which may reduce the gains from any subsequent optimization.

Besides providing good approximations of covariance eigenvectors for finite and large M, the spatial signature proves useful for other reasons. For instance, even though the BS is able to form a channel covariance estimate on its own, the spatial signature defines a unified eigenbasis for all users that tremendously simplifies the problem formulation as will be seen by examining (53). Additionally, this reduces complexity since the eigenvalue decomposition required to obtain the (exact) eigenvectors is avoided.

Treating users with strictly non-overlapping spatial signatures as candidates for being assigned the same training sequence, a natural criterion for performing the sequence assignment is to maximize the number of users that can be supported given the training overhead T. Targeting sum rate maximization, it is also natural to select users with the largest received energy at the BS. Introducing the KT group assignment variables

$$x_{k,t} = \begin{cases} 1, & \text{if user } k \text{ is assigned to group } \mathcal{G}_t, \\ 0, & \text{otherwise,} \end{cases} \tag{49}$$

for $k \in \{1, \ldots, K\}$, $t \in \{1, \ldots, T\}$, the resulting problem can be formulated as the maximization of the total average received energy by the users subject to following constraints:

$$\underset{\{x_{k,t}\}}{\text{maximize}} \quad \sum_{t=1}^{T}\sum_{k=1}^{K} x_{k,t} Q_k \tag{50}$$

$$\text{subject to} \quad \sum_{k=1}^{K} x_{k,t} \leq U_t, t = 1, 2 \ldots, T, \tag{51}$$

$$\sum_{t=1}^{T} x_{k,t} \leq 1, k = 1, 2 \ldots, K, \tag{52}$$

$$\sum_{k=1}^{K} x_{k,t}\mathbb{I}_{k,s} \leq 1, t = 1, \ldots, T, s = 0, \ldots, M-1 \tag{53}$$

where $\mathbb{I}_{k,s} \triangleq \mathbb{I}(\boldsymbol{f}_M(s) \in \hat{\boldsymbol{U}}_{\mathcal{C}_k})$ and $Q_k \triangleq \sum_{s=1}^{M} \mathbb{I}_{k,s} Q_{k,s}$ is the total channel energy of user k over its spatial signature. Here, $\mathbb{I}$ is the standard indicator function defined by $\mathbb{I}(1) = 1$ and $\mathbb{I}(0) = 0$. Constraint (51) restricts the total number of the users per (sequence) group, effectively imposing a reuse factor U_t for sequence t, while constraint (52) guarantees that each user can be assigned to at most one group. The (approximate) mutual orthogonality between users within each group is guaranteed by the constraints in (53) which restrict each vector of the DFT basis to be considered at most once in each sequence group. In other words, users with overlapping spatial signatures cannot be assigned the same training sequence. Here, it can be seen how the defined spatial signatures simplify the problem formulation and solution. By (53), identification of (approximately) orthogonal covariance channels is simply achieved by checking if a given DFT column belongs to the spatial signatures of at least two users. A much more complex constraint would need to be introduced if the exact users' covariance eigenvectors are used.

It can be shown that the above problem is NP-hard, i.e., its global optimal solution cannot be found by means of a polynomial time solvable algorithm. An approximate, yet efficient algorithm was proposed in [14] based on an extension of the generalized maximum coverage algorithm in [54]. The proposed approach uses two nested greedy phases that are combined with several instances of a knapsack problem. The algorithm is guaranteed to provide, at least, a fraction $(1 - (\frac{T-1}{T})^T)\frac{\frac{3}{2} - \frac{e^{-2}}{2}}{1-e^{-2}}$ of the optimal solution.

Finally, we mention that the constraint (53) might leave users from being served, i.e., it might happen that $x_{k,t} = 0 \; \forall t$ for a given user k. If fairness is of concern, one could optimize the set of selected K users before running the proposed algorithm.

C. A GRAPHICAL APPROACH FOR INTER-CELL INTERFERENCE MANAGEMENT

The previous discussion considered a single-cell system, i.e., it ignored interference generated from other cells in the system. With $\mathcal{G}_{t,c}$ denoting the users served by cell $c \in \{1, \ldots, N_c\}$ that are assigned training sequence t, this approach implicitly performs a system-wide user grouping with sequence t assigned to users in the set $\mathcal{G}_t \triangleq \cup_{c=1}^{N_c} \mathcal{G}_{t,c}$.

This independent-per-cell sequence assignment can be sufficient in sparse networks with large cells, where the inter-cell interference power is expected to be much smaller than the intra-cell interference power. However, inter-cell interference becomes a significant issue in dense networks. In such a scenario and with channel covariance information available for every user-BS link in the system, one expects improved performance with a joint sequence assignment over the cells.

Towards an improved system-wide sequence assignment, we leverage the per-cell user grouping described previously in a cross-cell sequence allocation procedure that operates in two steps:

1) Each of the N_c cells groups its served users according to the procedure described above, i.e., ignoring inter-cell interference. However, no sequences are assigned to the groups at this stage.
2) The TN_c user groups $\{\mathcal{G}_{t,c}\}$ are partitioned into T super groups $\{\mathcal{G}_t\}$ of user groups with minimum (ideally, zero) mutual cross-cell interference.

Towards an efficient user grouping in the second stage of the algorithm, we employ a graph-based approach. An undirected weighted interference graph with TN_c vertices corresponding to the user groups $\{\mathcal{G}_{t,c}\}$ obtained in the first step is constructed. The vertices are connected via edges with weights that quantify the level of mutual cross-cell interference between the user groups. A simple example is shown in Fig. 8.

For two arbitrary vertices $\mathcal{G}_{t,c}, \mathcal{G}_{t',c'}$, we define the weight as

$$w_{\mathcal{G}_{t,c},\mathcal{G}_{t',c'}} \triangleq \min_{k \in \mathcal{G}_{t,c}, k' \in \mathcal{G}_{t',c'}} \left\{ I_{(k',c')\to(k,c)} + I_{(k,c)\to(k',c')} \right\} \tag{54}$$

when $c \neq c'$, where

$$I_{(k',c')\to(k,c)} \triangleq \frac{\mathrm{tr}(\hat{\boldsymbol{U}}^H_{\mathcal{C}^c_{k,c}} \hat{\boldsymbol{U}}_{\mathcal{C}^c_{k',c'}})}{\|\hat{\boldsymbol{U}}_{\mathcal{C}^c_{k,c}}\|_F \|\hat{\boldsymbol{U}}_{\mathcal{C}^c_{k',c'}}\|_F}, \tag{55}$$

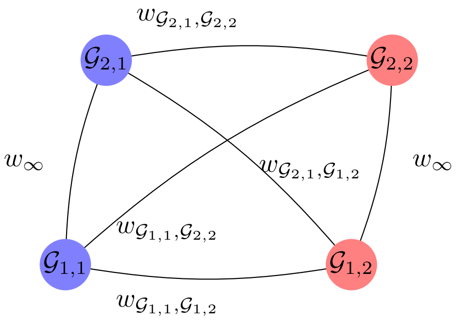
$w_{\mathcal{G}_{2,1},\mathcal{G}_{2,2}}$
$\mathcal{G}_{2,1}$
$\mathcal{G}_{2,2}$
w_∞
$w_{\mathcal{G}_{2,1},\mathcal{G}_{1,2}}$
w_∞
$w_{\mathcal{G}_{1,1},\mathcal{G}_{2,2}}$
$\mathcal{G}_{1,1}$
$\mathcal{G}_{1,2}$
$w_{\mathcal{G}_{1,1},\mathcal{G}_{1,2}}$

FIGURE 8. Example of an interference graph with $T = 2$ and $N_c = 2$.

with $\boldsymbol{C}_{k,c}^{c'}$ denoting the covariance matrix of the channel between user k of cell c and the BS of cell c', and $\hat{\boldsymbol{U}}_{\boldsymbol{C}_{k,c}^{c'}}$ representing the spatial signature of user k of cell c, as viewed by the BS of cell c', obtained similarly to (48), with $\boldsymbol{C}_{k,c}^{c'}$ in place of $\boldsymbol{C}_k$. For the weight corresponding to user groups of the same cell, we assign a very large weight w_∞ so as to guarantee that they do not receive the same training sequences by the algorithm to be described in the following.

The weight expression of (54) is inspired by the single and weighted average linkage, commonly used in hierarchical clustering [55] and the similarity measure used in [56] and [57]. Note that $I_{(k',c')\rightarrow(k,c)}$ can be regarded as a measure of the interference user k' of cell c' would introduce to the channel estimate of user k of cell c if both users had the same training sequence. Intuitively, a small weight $w_{\mathcal{G}_{t,c},\mathcal{G}_{t',c'}}$ indicates that users in $\mathcal{G}_{t,c}$ and $\mathcal{G}_{t',c'}$ have channels resulting in low levels of cross-cell pilot contamination, in case they are allocated the same training sequence. This suggests that the same training sequence should be allocated to user groups linked with small interference weights. A natural approach to construct super groups with these property is to consider an MAX-T-CUT problem formulation, with the T user super groups $\mathcal{G}_t$ identified as the ones maximizing

$$\sum_{1\le r<s<T}\ \sum_{\substack{\mathcal{G}_{t,c}\in\mathcal{G}_r,\\ \mathcal{G}_{t',c'}\in\mathcal{G}_s}} w_{\mathcal{G}_{t,c},\mathcal{G}_{t',c'}} \tag{56}$$

As this problem is NP-hard, we use the low complexity greedy approach in [58]. This algorithm provides an efficient partitioning of the interference graph with at least a fraction $(1 - \frac{1}{T})$-approximation of the optimal allocation solution. A note on the complexity of the proposed scheme is now in order. Indeed, optimizing cross-cell pilot allocation needs centralized knowledge of the channel spatial information for all links in the system. The latter may require a non-negligible signaling overhead between the BSs. In addition, one should take into consideration the needed computation time to perform the optimization. Nevertheless, the proposed approach is based on the second order statistics of the channel which are characterized by an extended stationarity time. This means that the network will be able to exchange the needed information and to perform cross-cell pilot allocation without reducing its efficiency. In addition, the DFT-based structure of the spatial signatures results in simplifying the task of spatial information centralization since DFT codebooks can be used.

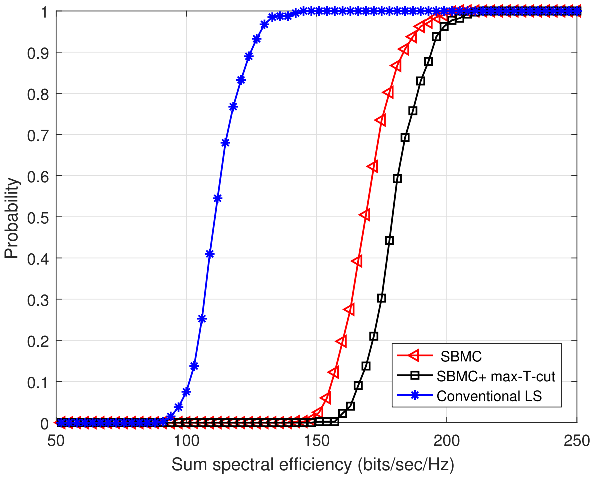

FIGURE 9. Comparison of CDFs of achievable sum spectral efficiency for $T = 6$, $U_t = 3$, and $\text{SNR}_t = 10$ dB.

D. NUMERICAL RESULTS AND DISCUSSION

In this section, we provide numerical results demonstrating the performance of the proposed sequence assignment scheme in a multi-cell context. We consider a network that consists of $N_c = 4$ hexagonal cells, each containing at its center a BS that is equipped with ULA of $M = 128$ antennas elements with $\lambda/2$ spacing. Each cell has a radius 0.5 Km, from center to vertex, and serves $K = 25$ randomly located users. The channel between each user and each serving BS is independently generated according to the model of (4) with $P = 100$ paths with angles uniformly distributed in the interval $\left[\bar{\theta} - 10°, \bar{\theta} + 10°\right]$ where $\bar{\theta}$ denotes the azimuth mean angle of arrival. The path-loss coefficient is set to 3.5 and the parameter α in (48) is set to 0.05. The latter indicates that users spatial signature should contain any DFT beam that spans at least 5% of the channel energy. This value was selected in order to strike a good compromise between performance and complexity.

For each channel coherence interval, set here to 128 channel uses, a training phase is implemented first in order to obtain a channel estimate for each user by its serving BS. A data transmission stage then follows, where all users simultaneously send their data symbols and each BS detects the symbols by means of maximum ratio combining, treating the estimated channel as the actual one. The performance of the schemes in Secs. VI. B and VI. C, respectively, is compared to that of a conventional TDD massive MIMO system. In this baseline system, each BS randomly assigns a set of orthogonal training sequences to its scheduled users to completely eliminate intra-cell pilot contamination, but the effects of inter-cell pilot contamination are neglected.

Figure 9 shows the impact of the proposed approach on the CDF of the achievable sum spectral efficiency of scheduled users. To provide a fair comparison, all schemes schedule the same users, set to $T \times U_t$ per cell. The only difference lies in the adopted UL training approach. While the number of training sequences for the proposed algorithm was set to $T = 6$ and the sequence reuse factor in (51) was set to $U_t = 3$, the users in the baseline scheme will be using randomly assigned orthogonal training sequences of length $T \times U_t$. We see that the proposed spatial signature based user grouping without inter-cell interference consideration enables a considerable improvement in spectral efficiency, achieving 5% outage rate around 154 bits/sec/Hz instead of 98 bits/sec/Hz provided by the conventional training scheme. This gain is due to the reduced training overhead T (6 instead of $T \times U_t$), which results in the utilization of a larger portion of the coherence interval for data transmission while simultaneously enabling accurate channel estimation with the proposed user grouping. When inter-cell interference is also taken into account, the performance can be further improved, achieving a 5% outage rate around 164 bits/sec/Hz. However, this performance is obtained at the cost of global channel covariance information for all links in the system, which may be costly.

VII. CONCLUSION

This work shows how channel structural information can be utilized towards optimizing the channel estimation MSE or achievable rates with minimum training overheads, which is of critical importance in a multiuser massive MIMO setting. Starting from the parametric channel model defined in Sec. II, we showed in Sec. III how the channel can be effectively considered as sparse in many scenarios of interest (e.g. high carrier frequencies). Building on this finding and further introducing the concept of hierarchical sparsity, we studied sufficient training scaling laws in single-cell UL wideband systems in Sec. IV. We found that a training overhead that scales logarithmically with the number of subcarriers but independently of the number of channel paths and users is sufficient for reliable channel estimation. This is a huge benefit offered by the use of multiple antennas in a massive MIMO setting, which is not available in conventional MIMO. Using the developed connection between sparse channels and low-rank channel covariances in Sec. II-C, we have further exploited the finding of Sec. III towards minimizing the training overhead in a Bayesian framework with channel covariance knowledge at the BS in Secs. V and VI. Sec. V studied the training overhead scaling laws in a single-cell DL scenario, and showed that a sufficient training overhead for reliable channel estimation depends on the ranks of the users' covariance matrices and the overlap between their range spaces. For covariance matrices with small ranks and/or significant overlap, the resulting overhead can be significantly smaller than the number of BS antennas, which would be the overhead required under a naive LS-based channel estimation approach. We additionally showed how the training overhead can be further reduced when the data rate metric is directly optimized. Finally, Sec. VI considered a multi-cell UL scenario and designed a novel spatial user grouping scheme that mitigates intra-cell as well as inter-cell pilot contamination while supporting high training sequence reuse factors. The proposed algorithm exploits the (approximate) orthogonality between dominant covariance eigenvectors of carefully selected users and is facilitated by the concept of spatial signatures, introduced therein.

Future possible lines of work include, for methods exploiting the sparse angle-delay representation, designing algorithms that work efficiently for different number of paths per user and without the a priori number of paths knowledge at the BS. For methods exploiting the low-rank covariance structure, interesting avenues for future work include considering non-orthogonal sequences with optimized power levels in the aim of further reducing the training overhead.

ACKNOWLEDGEMNT

The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

REFERENCES

[1] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

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

[3] J. H. Kotecha and A. M. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 546–557, Feb. 2004.

[4] Y. Liu, T. F. Wong, and W. W. Hager, "Training signal design for estimation of correlated MIMO channels with colored interference," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1486–1497, Apr. 2007.

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

[6] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[7] D. Ciuonzo, P. S. Rossi, and S. Dey, "Massive MIMO channel-aware decision fusion," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 604–619, Feb. 2015.

[8] F. Jiang, J. Chen, A. L. Swindlehurst, and J. A. López-Salcedo, "Massive MIMO for wireless sensing with a coherent multiple access channel," *IEEE Trans. Signal Process.*, vol. 63, no. 12, pp. 3005–3017, Jun. 2015.

[9] L. Le Magoarou and S. Paquelet, "Parametric channel estimation for massive MIMO,"in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, Jun. 2018, pp. 30–34.

[10] L. Le Magoarou and S. Paquelet. (2018). "Bias-variance tradeoff in MIMO channel estimation." [Online]. Available: https://arxiv.org/abs/1804.07529

[11] G. Wunder, S. Stefanatos, A. Flinth, I. Roth, and G. Caire, "Low-overhead hierarchically-sparse channel estimation for multiuser wideband massive MIMO," *IEEE Trans. Wireless Commun.*, to be published. [Online]. Available: https://arxiv.org/abs/1806.00815

[12] S. Bazzi and W. Xu, "Low-complexity channel estimation in correlated massive MIMO channels," in *Proc. 22nd Int. ITG Workshop Smart Antennas (WSA)*, Bochum, Germany, Mar. 2018, pp. 1–6.

[13] S. Bazzi and W. Xu, "On the amount of downlink training in correlated massive MIMO channels," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2286–2299, May 2018.

[14] S. E. Hajri and M. Assaad, "A spatial basis coverage approach for uplink training and scheduling in massive MIMO systems," *IEEE Trans. Wireless Commun.*, to be published. [Online]. Available: https://arxiv.org/abs/1804.10934

[15] F. Schaich *et al.*, "The ONE5G approach towards the challenges of multi-service operation in 5G systems," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Porto, Portugal, Jun. 2018, pp. 1–6.

[16] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313–7321, Nov. 2016.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing Estimation Theory*, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[18] L. Liu *et al.*, "The COST 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.

[19] S. Sun, G. R. MacCartney. Jr., and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5g wireless communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[20] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.

[21] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901 v14.1.0, 3GPP, 2017.

[22] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part IV: Optimum Array Processing*. Hoboken, NJ, USA: Wiley, 2002.

[23] D. Tse and P. Viswanath, *Fundamental of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[24] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 7, pp. 2207–2225, Jul. 2016.

[25] J. L. Paredes, G. R. Arce, and Z. Wang, "Ultra-wideband compressed sensing: Channel estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 3, pp. 383–395, Oct. 2007.

[26] B. Yang, K. B. Letaief, R. S. Cheng, and Z. Cao, "Channel estimation for OFDM transmission in multipath fading channels based on parametric channel modeling," *IEEE Trans. Commun.*, vol. 49, no. 3, pp. 467–479, Mar. 2001.

[27] W. Dongming, H. Bing, Z. Junhui, G. Xiqi, and Y. Xiaohu, "Channel estimation algorithms for broadband MIMO-OFDM sparse channel," in *Proc. IEEE Pers., Indoor Mobile Radio Commun. (PIMRC)*, Beijing, China, Sep. 2003, pp. 1929–1933.

[28] M. R. Raghavendra and K. Giridhar, "Improving channel estimation in OFDM systems for sparse multipath channels," *IEEE Signal Process. Lett.*, vol. 12, no. 1, pp. 52–55, Jan. 2005.

[29] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, 2nd ed. Norwell, MA, USA: Kluwer, 1993.

[30] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 802–814, Oct. 2014.

[31] Z. Yang, L. Xie, and P. Stoica, "Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 2701–3685, Apr. 2016.

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

[33] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[34] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath, Jr., "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1996–2009, Sep. 2017.

[35] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, no. 6, pp. 948–958, 2010.

[36] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[37] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York, NY, USA: Springer, 2013.

[38] S. Haghighatshoar and G. Caire, "Massive MIMO pilot decontamination and channel interpolation via wideband sparse channel estimation," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8316–8332, Dec. 2017.

[39] L. You, X. Gao, A. L. Swindlehurst, and W. Zhong, "Channel acquisition for massive MIMO-OFDM with adjustable phase shift pilots," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1461–1476, Mar. 2016.

[40] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2909–2913.

[41] O. Edfors, M. Sandell, J.-J. van de Beek, S. K. Wilson, and P. O. Borjesson, "OFDM channel estimation by singular value decomposition," *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 931–939, Jul. 1998.

[42] K. Upadhya and S. A. Vorobyov, "Covariance matrix estimation for massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 546–550, Apr. 2018.

[43] D. Neumann, M. Joham, and W. Utschick, "Covariance matrix estimation in massive MIMO," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 863–867, Jun. 2018.

[44] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2868–2881, May 2015.

[45] S. Bazzi and W. Xu, "Downlink training sequence design for FDD multiuser massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4732–4744, Sep. 2017.

[46] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in real propagation environments: Do all antennas contribute equally?" *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 3917–3928, Nov. 2015.

[47] X. Gao, F. Tufvesson, and O. Edfors, "Massive MIMO channels—Measurements and models," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 280–284.

[48] A. F. Molisch, *Wireless Communications*, 2nd ed. Piscataway, NJ, USA: IEEE Press, 2011.

[49] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3170–3184, Apr. 2017.

[50] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

[51] S. E. Hajri, M. Assaad, and G. Caire, "Scheduling in massive MIMO: User clustering and pilot assignment," in *Proc. Allerton Conf. Commun., Control, Comput.*, Sep. 2016, pp. 107–114.

[52] N. Akbar, S. Yan, N. Yang, and J. Yuan, "Location-aware pilot allocation in multicell multiuser massive MIMO networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7774–7778, Aug. 2018.

[53] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[54] R. Cohen and L. Katzir, "The generalized maximum coverage problem," *Inf. Process. Lett.*, vol. 108, no. 1, pp. 15–22, 2008.

[55] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access J.*, vol. 2, no. 1, pp. 947–959, Sep. 2014.

[56] A. Maatouk, S. E. Hajri, M. Assaad, H. Sari, and S. Sezginer, "Graph theory based approach to users grouping and downlink scheduling in FDD massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[57] A. Maatouk, S. E. Hajri, M. Assaad, and H. Sari, "On optimal scheduling for joint spatial division and multiplexing approach in FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1006–1021, Feb. 2019.

[58] S. Sahni and T. Gonzalez, "P-complete approximation problems," *J. Assoc. Comput. Machinery*, vol. 23, no. 3, pp. 555–565, Jul. 1976.

SAMER BAZZI received the B.E. degree in computer and communications engineering from the American University of Beirut, in 2008, and the M.Sc. and Dr.-Ing. degrees in communications engineering from the Technische Universität München (TUM), in 2010 and 2016, respectively. From 2010 to 2014, he was a member of the DOCOMO Euro-Labs, Munich, Germany, and an External Dr.-Ing. Candidate with the Chair of Signal Processing Methods, TUM, where he worked on coordinated multi-point techniques and precoding for interference channels. In 2015, he joined the European Research Center, Huawei Technologies Duesseldorf GmbH, where he is currently a Senior Engineer. His research interests include multiple-input-multiple-output systems, parameter estimation, interference management, and general signal processing techniques for wireless communications.

STELIOS STEFANATOS received the Diploma degree in physics and the M.S. degree in communications engineering from the National Kapodistrian University of Athens (NKUA), Greece, and the Ph.D. degree in communications engineering from the University of Piraeus, Greece. He was a Research Associate with NKUA, the University of Piraeus, and the Athena Research and Innovation Center, Greece. He is currently a Research Associate with the Freie Universität Berlin, Germany. His research interests include communication theory, stochastic geometry analysis of wireless networks, statistical signal processing, and resource allocation for wireless communications.

LUC LE MAGOAROU received the Ph.D. degree in signal processing and the M.Sc. degree in electrical engineering from the National Institute of Applied Sciences (INSA), Rennes, France, in 2016 and 2013, respectively. He was with the PANAMA Research Group, Inria, Rennes. He is currently a Postdoctoral Researcher with b<>com, Rennes. His main research interests include signal processing, machine learning, and multiple-input-multiple-output communication systems.

SALAH EDDINE HAJRI completed a double degree in September 2014 where he obtained his Engineering diploma from Sup'com, Tunis, Tunisia, and his M.Sc. degree from Supélec, Gif-sur-Yvette, France, both in Wireless Digital Communication Systems. He received the Ph.D. degree in networks, information and communications, from CentraleSupélec in 2018. He is currently with the TCL Chair on 5G Systems, and also with the LSS Lab, CentraleSupélec, focuses on the enhancement of 5G new radio. His research interests include resource optimization and cross-layer design in wireless networks, massive multiple-input-multiple-output systems, and machine learning enablers of 5G new radio.

MOHAMAD ASSAAD (SM'15) received the M.Sc. and Ph.D. degrees in telecommunications from Telecom ParisTech, Paris, France, in 2002 and 2006, respectively. Since 2006, he has been with the Telecommunications Department, CentraleSupelec, where he is currently a Professor and holds the TCL Chair on 5G. He has co-authored one book and more than 90 publications in journals and conference proceedings. His research interests include mathematical modeling of communication networks, multiple-input-multiple-output systems, resource management, and cross-layer design in wireless networks. He has served as a TPC Member or a TPC Co-Chair for top international conferences. He is an Editor of IEEE Wireless Communications Letters and the *Journal of Communications and Information Networks*. He has given in the past successful tutorials on 5G systems at various conferences including IEEE ISWCS'15 and IEEE WCNC'16 conferences.

STÉPHANE PAQUELET received the B.Sc. degree from the Ecole Polytechnique, Paris, France, in 1996, and the M.Sc. degree from Telecom Paris, Paris, France, in 1998. He was involved in the fields of cryptology and signal processing for electronic warfare with Thales and led UWB Research and Development with Mitsubishi Electric, from 2002 to 2007, where he proposed two pioneering transceivers for short-range/high data rates and large-range/low data rates, including telemetry. He was with Renesas Design France, until 2014, where he developed multi-standard reconfigurable RF-IC. Since 2015, he has been leading wireless activities for IRT with b<>com.

GERHARD WUNDER (M'04–SM'11) received the Dipl.-Ing. degree (Hons.) in electrical engineering and the Ph.D. degree (Dr.-Ing.) *(summa cum laude)* from the Technical University of Berlin, in 1999 and 2003, respectively, and the Habilitation degree (venia legendi), in 2007. He became a Research Group Leader with the Fraunhofer Heinrich-Hertz-Institut, Berlin. In 2007, he became a Privatdozent (Associate Professor). He was a Visiting Professor with the Georgia Institute of Technology (Prof. Jayant), Atlanta, GA, USA, and with Stanford University (Prof. Paulraj), Palo Alto, CA, USA. In 2009, he was a Consultant with Alcatel-Lucent Bell Labs (Prof. Stolyar), Murray Hill, NJ, USA, and with Alcatel-Lucent Bell Labs (Dr. Valenzuela), Crawford Hill, NJ, USA. In 2015, he became a Heisenberg Fellow, granted for the first time to a Communication Engineer. He is currently the Head of the Heisenberg Communications and Information Theory Group, Free University of Berlin. He is a Coordinator and a Principal Investigator with the FP7 Project 5GNOW on 5G new waveforms (received Outstanding Excellence from EC) and PROPHYLAXE on the Internet of Things Physical Layer Security funded by German BMBF. He has been a member of the project management teams of H2020 projects FANTASTIC-5G (also received Outstanding Excellence) and ONE5G, both regarded as flagship projects within the European 5GPPP framework. He is a member of the IEEE TWC's Executive Editorial Committee. In 2011, he received the 2011 Award for Outstanding Scientific Publication in the field of communication engineering at the German Communication Engineering Society (ITG Award 2011). He has co-chaired numerous international renowned workshops, conference tracks, and special issues, particularly in the context of 5G. He was a Co-Chair of the IEEE GLOBECOM 2017 Signal Processing for Communications Symposium. He has been nominated together with Dr. Müller (BOSCH Stuttgart) and Prof. Paar (Ruhr University Bochum) for the Deutscher Zukunftspreis 2017 on behalf of the PROPHYLAXE Project.

WEN XU (SM'03) received the B.Sc. and M.Sc. degrees in electrical engineering from the Dalian University of Technology, China, in 1982 and 1985, respectively, and the Dr.-Ing. degree in electrical engineering from the Technische Universität München (TUM), Germany, in 1996. From 1995 to 2006, he was with Siemens Mobile (later BenQ Mobile), Munich, where he was the Head of the Algorithms and Standardization Lab. As a competence center, the lab was responsible for physical layer and multimedia signal processing, and partly protocol stack aspects, of 2G, 3G, and 4G mobile terminals, and actively involved in standardization activities of ETSI, 3GPP, DVB, and ITU. From 2007 to 2014, he was with Infineon Technologies AG (later Intel Mobile Communications GmbH), Neubiberg, where he focused on wireline and wireless system concepts/architectures and software/hardware implementations. In 2014, he joined the European Research Center, Huawei Technologies Duesseldorf GmbH, Munich, where he is currently the Head of Radio Access Technologies Department. His research interests include signal processing, source/channel coding, and wireless communications systems in general. He is a member of the Verband der Elektrotechnik, Elektronik, Informationstechnik, Germany.