This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

Low Complexity Hybrid Precoding Strategies for Millimeter Wave Communication Systems

Cristian Rusu, Roi Méndez-Rial, Nuria González-Prelcic and Robert W. Heath Jr.

***Abstract*—Millimeter communication systems use large antenna arrays to provide good average received power and to take advantage of multi-stream MIMO communication. Unfortunately, due to power consumption in the analog front-end, it is impractical to perform beamforming and fully digital precoding at baseband. Hybrid precoding/combining architectures have been proposed to overcome this limitation. The hybrid structure splits the MIMO processing between the digital and analog domains, while keeping the performance close to that of the fully digital solution. In this paper we introduce and analyze several algorithms that efficiently design hybrid precoders and combiners starting from the known optimum digital precoder/combiner, which can be computed when perfect channel state information is available. We propose several low complexity solutions which provide different trade-offs between performance and complexity. We show that the proposed iterative solutions perform better in terms of spectral efficiency and/or are faster than previous methods in the literature. All of them provide designs which perform close to the known optimal digital solution. Finally, we study the effects of quantizing the analog component of the hybrid design and show that even with coarse quantization, the average rate performance is good.**

I. Introduction

The millimeter wave (mmWave) band is the spectral frontier for the next generation commercial wireless local area and cellulars networks [1]. MmWave solutions promise to deliver high data rates by exploiting the large bandwidth available at carrier frequencies from 30 GHz to 300 GHz for both indoor [2] and outdoor [3] wireless systems. As an example, WirelessHD [4] technology and the new Wireless Local Area Network (WLAN) standard IEEE 802.11ad [5], already operate in the unlicensed 60 GHz band using channels with a bandwidth of approximately 2 GHz. Coupled with advanced modulation and multiple-input multiple-output (MIMO) techniques, these systems are capable of delivering data rates up to 28 Gbps in case of WirelessHD and 6.75 Gbps for IEEE 802.11ad. Future applications of mmWave also include 5G cellular systems [6]–[9] where gigabit-per-second data rates are expected to be achieved.

At mmWave frequencies, large antenna arrays are used both at the transmitter and the receiver to provide enough antenna aperture [10]. In a classical MIMO communication system, fully digital processing is made possible by connecting to every antenna dedicated baseband and radio-frequency (RF) hardware. In a mmWave MIMO communication system making use of large arrays, this approach consumes a lot of power due to components such as analog-to-digital converters (ADCs) and RF devices [11]. Hybrid MIMO architectures [12]–[16] overcome this limitation by splitting the MIMO processing between the analog and digital domains. The analog precoding stage processes the received signals at the different antenna elements feeding a number of radio-frequency (RF) chains much lower than the array size. A mmWave prototype with a hybrid beamforming architecture has already been designed and tested in [17].

The first work on single user hybrid precoding specially designed for mmWave communications was [18]. The hybrid precoder/combiner design was formulated as a sparsity constrained factorization in the style of simultaneous matching pursuit [19] or dictionary learning [20], accounting for the sparse structure of mmWave channels. The proposed solution in [18] has two distinct phases. First, the fully digital precoder is computed assuming channel state information, a procedure that is well understood and for which efficient solutions are available. This solution is known to be the optimum unconstrained precoder from a data-rate maximization perspective. Second, the fully digital solution is factorized into analog and digital parts each with the constraints that reflect the hardware design. The resulting optimization problem is hard to solve due to the constant magnitude constraints that reflect the analog hardware components. This difficulty was overcome by searching for the analog components in a predefined dictionary that reflects the geometry of the antenna array used at transmission. In this way, the sparse scattering structure of mmWave channels is exploited. The main limitation of the algorithm is the high computational complexity.

The same algorithmic approach was used in [21] to design a hybrid solution approximating the minimum mean squared error transmit/receive processing in MIMO interference channels. Unfortunately, the dictionary constraint leads to hybrid designs that perform well below the fully digital solution. Furthermore, the computational complexity is high. In [22], a codebook based hybrid precoding optimization problem is considered in the context of an array-of-subarrays architecture. The sparse nature of the mmWave channel is exploited to propose a computationally efficient way of finding the near-optimal precoder in a given codebook. In [23], the single data stream case analyzed in [15] was extended to multiple streams and the conditions under which the performance of the optimal digital precoder can be matched are shown. The work in [22], [23], and [15], though, does not exploit the sparse nature of mmWave channels.

While computing the digital component is relatively straight

C. Rusu, R. Méndez-Rial and N. González-Prelcic are with the University of Vigo, Spain, Email: {crusu,roimr,nuria}@gts.uvigo.es. R. Heath is with The University of Texas at Austin, TX 78712, Email: rheath@utexas.edu.

This work was supported by the Spanish Government and the European Regional Development Fund (ERDF) under project TACTICA and COMPASS; the National Science Foundation under Grant No. NSF-CCF-1319556; and by a gift from Huawei Technologies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

forward, one of the major difficulties when constructing the hybrid precoder is imposing the unit magnitude constraint on the entries of the analog component. There are several solutions proposed to overcome this difficulty. For example, singular value decompositions used in a greedy fashion were proposed in [24] to update the analog component. The method performs well experimentally but no theoretical guarantees are provided to justify its performance. In [25], a heuristic was proposed to maximize directly the spectral efficiency. They outline a numerical optimization procedure that converges to local minimum points, but unfortunately works only when the number of data streams equals the number of RF chains. Another solution presented in [26] performs a simplex 1-D iterative local search for every element of the analog precoder. This solution approaches the performance of the digital precoder, but some concerns are raised regarding its efficiency due to the large number of entries that are updated separately, using for each one a series of iterations.

In this paper, *assuming perfect channel state information at both the transmitter and receiver and a total power constraint for the antenna arrays, we propose methods for the design of the hybrid precoder and combiner that are fast, while still obtaining near-optimal performance in terms of the achieved data rates*. We develop four methods that provide different trade-offs between computational complexity and performance. They differ in the proposed optimization techniques, which are dependent on assumptions on dimensions (number of streams and RF chains), and on the structure of the analog and digital components (e.g., scaled unitary/semi-unitary or no internal structure).

The first method, called HD–AM (hybrid design by alternating minimization), performs well in terms of approximating the performance of the full digital solution and enjoys low computational complexity. Unfortunately, it can be applied only in the special case when the number of data streams equals the number of RF chains and a scaled unitary assumption is made for the digital component.

In this case, the analog component of the precoding matrix can be decoupled in the mathematical formulation and both the analog and the digital precoders can be optimized separately. We guarantee the convergence of this method to a local optimum point. When applicable, this method is preferred, since it enjoys both low complexity and high performance. The following two proposed solutions will not exhibit any constraint on the relationship between the number of RF chains and that of the streams.

The second procedure, called HD–FUM (hybrid design by fast unitary matching), improves upon the previously introduced idea of searching for the analog component in an overcomplete dictionary [18], by avoiding computationally expensive matching pursuit iterations. This method is the fastest among the designs proposed in this paper and in previous work. Unfortunately, due to the dictionary constraints, it performs the worst in approximating the fully digital solution. The algorithm also depends on the array geometry through the dictionary, which is not required in the other algorithms. With this approach we are able to quickly construct analog and digital components that are both semi-unitary. We highlight their potential use for the channel estimation problem.

The third proposed approach for the hybrid precoder design, called HD–CVXR (hybrid design by convex relaxation), is optimized for high performance (best approximation of the fully digital solution). It is the most general in scope, making no dimensionality assumptions or other internal constraints. From a computational perspective, this approach is relative slower than the previously two proposed techniques. But we expect it to perform the best, since no internal assumptions are made in the optimization problem. The main idea of this approach is to relax the hard, non-convex, optimization problem at hand with its convex relaxation. By solving a series of these problems (which include an iterative update of each column of the analog precoder by convex optimization, and an update of the baseband precoder by least squares), the solution is improved at each step. A mathematical general analysis of the convergence of this algorithm though is challenging.

Finally, we also develop a variant of HD–CVXR called HD-LSR (hybrid design by least squares relaxation). It is based on the idea of using a heuristic relaxation like in HD-CVXR, but the RF precoder is now updated entry by entry. The algorithm is also general, no assumption is made on the problem dimensions or the structure of the analog and baseband precoders. In this special case we are able to prove the convergence to a local optimum point.

We compare all the proposed methods in the simulations section. The algorithm based on alternating minimization, HD-AM, is preferred if it is applicable, because of its low complexity and the high spectral efficiency. The dictionary based solution, HD-FUM, has the lowest computational complexity, but provides high errors and lower spectral efficiency when trying to approximate the optimum precoder. All the proposed methods provide a higher spectral efficiency than previous solutions described in [18] and [24].

Bold lower-case letters like $\mathbf{a}$ denote vectors and its i^{th} entry is a_i, bold upper-case letters like $\mathbf{A}$ denote complex matrices, variables $|\mathbf{A}|$, $\mathbf{A}^T$ and $\mathbf{A}^*$ denote the elementwise absolute value, transpose and conjugate transpose respectively, $(\mathbf{A})_{i,j}$ denotes the (i,j) entry of $\mathbf{A}$, $(\mathbf{A})_j$ denotes the j^{th} column of $\mathbf{A}$. The set $\mathcal{U}_{n\times m}$ denotes the set of all complex matrices of size $n \times m$ that have unit magnitude entries, and any other calligraphic upper-case letters denote sets. The Frobenius norm of $\mathbf{A}$ is given by $\|\mathbf{A}\|_F^2 = \sum_i \sum_j |(\mathbf{A})_{i,j}|^2$, the nuclear norm of $\mathbf{A}$ is given by $\|\mathbf{A}\|_* = \sum_i \sigma_i$, where σ_i is the i^{th} singular value of $\mathbf{A}$. We use the notation $\mathcal{CN}(\mathbf{a}, \mathbf{A})$ to denote a complex circularly symmetric Gaussian random vector with mean $\mathbf{a}$ and covariance $\mathbf{A}$, while $\mathbb{C}^{n\times m}$ denotes the set of complex matrices of size n by m.

II. Formulation of the hybrid precoder/combiner design problems

Consider the mmWave hybrid MIMO system in Figure 1. The transmitter sends N_{s} data streams using N_{t} antennas to the receiver, which has N_{r} antennas. Mutistream transmission using L_{t} RF transmit chains with $N_{\text{s}} \leq L_{\text{t}} \leq N_{\text{t}}$ is assumed, and the hardware is mirrored at the receiver, where L_{r} RF chains are available such that $N_{\text{s}} \leq L_{\text{r}} \leq N_{\text{r}}$. The digital

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

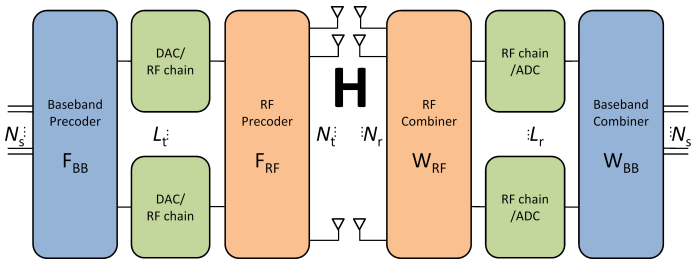

Figure 1. Block diagram of the transmitter-receiver single user mmWave system architecture that uses hybrid precoder and combiner.

precoder $\mathbf{F}_{\text{BB}}$ is designed assuming infinite precision and no constraints, while the analog precoder $\mathbf{F}_{\text{RF}}$ is assumed to have elements of equal magnitude, i.e., only phase shifting is performed in the analog domain. The total power constraint is enforced by normalizing $\mathbf{F}_{\text{BB}}$ such that $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$. While in this paper we consider only a total power limit, recent work on hybrd precoding considers also a per antenna power constraint [27]. As the analog precoder $\mathbf{F}_{\text{RF}}$ may be applied at some intermediate frequency or at the RF frequency, we represent it using its complex baseband equivalent.

Assuming a narrowband block fading channel operation as in [14], [18], [28], [29], the received signal is

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{s}$ represents the symbol vector of size N_{s}, ρ is the average received power and $\mathbf{H}$ is the $N_{\text{r}} \times N_{\text{t}}$ channel matrix, $\mathbf{n}$ is the noise vector i.i.d. as $\mathcal{CN}(0, \sigma^2)$. In this paper, the input symbol vector is normalized such that $\mathbb{E}[\mathbf{s}\mathbf{s}^*] = N_{\text{s}}^{-1}\mathbf{I}_{N_{\text{s}}}$. As in previous work on hybrid precoding [18], [24]–[26], [30]–[32], perfect knowledge of $\mathbf{H}$ will be assumed in this paper.

Although mmWave channels can be described using multipath models employed at lower frequencies [33], the beamspace (virtual) representation [29] is a natural choice due to the highly directional nature of propagation at mmWave. In addition, mmWave channels are expected to have limited scattering, which leads to a sparse channel structure [34], [35]. To incorporate this fact, we use a clustered channel model with N_{cl} scattering clusters, each of which contribute N_{ray} propagation paths. Therefore, in this paper we adopt the narrowband clustered channel which has been shown to provide an accurate representation of narrowband mmWave channels [36]–[38]

$$\mathbf{H} = \sqrt{\frac{N_{\text{t}}N_{\text{r}}}{N_{\text{cl}}N_{\text{ray}}}} \sum_{i=1}^{N_{\text{cl}}} \sum_{\ell=1}^{N_{\text{ray}}} \beta_{i,\ell}\mathbf{a}_{\text{r}}(\alpha_{i\ell}^{\text{r}})\mathbf{a}_{\text{t}}^*(\alpha_{i\ell}^{\text{t}}), \tag{2}$$

where $\beta_{i,\ell}$ is the complex gain, whereas $\mathbf{a}_{\text{t}}(\alpha_{i\ell}^{\text{t}})$ and $\mathbf{a}_{\text{r}}(\alpha_{i\ell}^{\text{r}})$ are the antenna array steering and response vectors at the transmitter and receiver evaluated at the corresponding azimuth angles of departure or arrival. $\mathbf{H}$ can be approximated using the beamspace channel representation [34], [35] as

$$\mathbf{H} \approx \mathbf{A}_{\text{r}}\mathbf{H}_{\text{v}}\mathbf{A}_{\text{t}}^*, \tag{3}$$

where $\mathbf{H}_{\text{v}} \in \mathbb{C}^{N_{\text{r}} \times N_{\text{t}}}$ contains the path gains of the quantized spatial frequencies. The dictionary matrices $\mathbf{A}_{\text{t}} = \left[\mathbf{a}_{\text{t}}(\alpha_1^{\text{t}}) \ \ldots \ \mathbf{a}_{\text{t}}(\alpha_{N_{\text{t}}}^{\text{t}})\right] \in \mathbb{C}^{N_{\text{t}} \times N_{\text{t}}}$ and $\mathbf{A}_{\text{r}} = \left[\mathbf{a}_{\text{r}}(\alpha_1^{\text{r}}) \ \ldots \ \mathbf{a}_{\text{r}}(\alpha_{N_{\text{r}}}^{\text{r}})\right] \in \mathbb{C}^{N_{\text{r}} \times N_{\text{r}}}$ contain the transmitter and receiver array response vectors evaluated on a grid for the angles of departure and arrival of N_{t} and N_{r} points respectively.

The array geometry will influence the choice of dictionary matrices. For example, when uniform linear arrays (ULAs) are assumed, $\mathbf{A}_{\text{r}}$ ($\mathbf{A}_{\text{t}}$) contains Vandermonde vectors of complex exponentials on a fine grid, i.e., $\mathbf{A}_{\text{r}}$ is an Inverse Discrete Fourier Transform (IDFT) matrix. We exploit the ULA geometry explicitly in one of our algorithms; the others do not use this information.

The receiver applies the $N_{\text{r}} \times L_{\text{r}}$ analog combining matrix $\mathbf{W}_{\text{RF}}$ and the $L_{\text{r}} \times N_{\text{s}}$ baseband combining matrix $\mathbf{W}_{\text{BB}}$ to the received signal. The dimensions satisfy $N_{\text{s}} \leq L_{\text{r}} \leq N_{\text{r}}$ to use a limited number of RF chains and a low dimensional digital combiner following the analog-to-digital converters (ADCs). The post-processed received signal after hybrid combining is

$$\tilde{\mathbf{y}} = \sqrt{\rho}\mathbf{W}_{\text{BB}}^*\mathbf{W}_{\text{RF}}^*\mathbf{H}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{s} + \mathbf{W}_{\text{BB}}^*\mathbf{W}_{\text{RF}}^*\mathbf{n}. \tag{4}$$

The goal is to design $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}, \mathbf{W}_{\text{RF}}$ and $\mathbf{W}_{\text{BB}}$ to maximize the spectral efficiency R achieved with Gaussian signaling over the mmWave channel [39]:

$$R = \log_2 \left|\mathbf{I}_{N_{\text{s}}} + \frac{\rho}{N_{\text{s}}}\mathbf{R}_n^{-1}\mathbf{W}_{\text{BB}}^*\mathbf{W}_{\text{RF}}^*\mathbf{H}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}} \times \mathbf{F}_{\text{BB}}^*\mathbf{F}_{\text{RF}}^*\mathbf{H}^*\mathbf{W}_{\text{RF}}\mathbf{W}_{\text{BB}}\right|, \tag{5}$$

where $\mathbf{R}_n = \sigma_n^2\mathbf{W}_{\text{BB}}^*\mathbf{W}_{\text{RF}}^*\mathbf{W}_{\text{RF}}\mathbf{W}_{\text{BB}}$ is the noise covariance matrix after combining. To avoid maximizing this expression over all unknowns $\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}, \mathbf{W}_{\text{RF}}$ and $\mathbf{W}_{\text{BB}}$ we consider the approach in [18], which proposes to decouple the problem by considering the transmitter and receiver sides separately. In [18] it is shown experimentally that the performance of the decoupled solution is very close to that of the fully digital precoder/combiner. Even considering this decoupling strategy, the individual optimization problems remain non-convex and hard. Therefore, attacking directly the joint problem does not seem a reasonable approach.

Consider the singular value decomposition of the channel matrix $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. In the absence of hardware limitations and considering equal power allocation across streams, the optimum unconstrained precoder $\mathbf{F}_{\text{opt}}$, that maximizes the data rate is given by the first N_{s} columns in $\mathbf{V}$ associated with the highest singular values in $\mathbf{\Sigma}$. The discussion and the methods presented in this paper can be extended to scenarios with unequal power allocation (for the optimum unconstrained precoder $\mathbf{V}\mathbf{\Gamma}$ where $\mathbf{\Gamma}$ is a diagonal obtained from waterfilling power allocation) and dynamic ways of choosing the number of streams N_{s}.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

The hybrid precoding problem at the transmitter is the design of $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$, given that the optimum unconstrained precoder is readily computed from perfect knowledge of $\mathbf{H}$. After some manipulation, [18] shows that this can be done by solving

$$\begin{aligned} \underset{\mathbf{F}_{\text{RF}} \in \mathcal{U}_{N_{\text{t}} \times L_{\text{t}}},\, \mathbf{F}_{\text{BB}}}{\text{minimize}} \quad & \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}. \end{aligned} \tag{6}$$

In this paper, we propose several new optimization strategies to solve (6). We identify assumptions (on the dimensions and on the internal structures of the analog and digital components) such that fast numerical optimization procedures can be deployed to solve (6), while achieving good spectral efficiency. The main difficulty is handling the hardware constraint $\mathbf{F}_{\text{RF}} \in \mathcal{U}_{N_{\text{t}} \times L_{\text{t}}}$ in a computationally efficient manner and such that the solutions reached are nearly optimal, i.e., performance close to that of $\mathbf{F}_{\text{opt}}$. For the analog component $\mathbf{F}_{\text{RF}}$ we also deal with the loss in performance due to the quantization of the phase shifters.

While we do focus on the hybrid design of the precoder, the methods proposed in this paper extend to the hybrid design of the combiner. With the precoder fixed, the optimum unconstrained combiners may be signed to either to maximize the data rate ($\mathbf{W}_{\text{opt}}$) or to minimize the mean squared error between transmitted and received signals ($\mathbf{W}_{\text{MMSE}}$). $\mathbf{W}_{\text{opt}}$ consists of left singular vectors from the decomposition of the channel matrix while from [40]

$$\mathbf{W}_{\text{MMSE}}^* = \frac{1}{\sqrt{\rho}} \left(\mathbf{F}_{\text{BB}}^* \mathbf{F}_{\text{RF}}^* \mathbf{H}^* \mathbf{H} \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}} + \frac{\sigma_n^2 N_{\text{s}}}{\rho} \mathbf{I}_{N_{\text{s}}} \right)^{-1} \times \mathbf{F}_{\text{BB}}^* \mathbf{F}_{\text{RF}}^* \mathbf{H}^*. \tag{7}$$

For brevity we only describe the hybrid design of the precoder using $\mathbf{F}_{\text{opt}}$. The same methods can be applied for the hybrid design of the combiner given either $\mathbf{W}_{\text{opt}}$ or $\mathbf{W}_{\text{MMSE}}$.

III. Properties of the optimal unconstrained precoder

Before proposing methods to approach the problem in (6), we establish several properties of the optimum unconstrained precoder $\mathbf{F}_{\text{opt}}$. In this way, we uncover additional structure to use in the fast design of the hybrid precoders $\mathbf{F}_{\text{BB}}$ and $\mathbf{F}_{\text{RF}}$.

First, irrespective of the channel model used, the optimum $\mathbf{F}_{\text{opt}}$ is a semi-unitary matrix, as it corresponds to the first N_{s} columns of the right singular vectors of $\mathbf{H}$. Therefore all $(\mathbf{F}_{\text{opt}})_j, j = 1, \ldots, N_{\text{s}}$ have unit ℓ_2 norm, which gives $\|\mathbf{F}_{\text{opt}}\|_F^2 = N_{\text{s}}$, $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$ in (6). Further since $\mathbf{F}_{\text{opt}}^* \mathbf{F}_{\text{opt}} = \mathbf{I}$, its factorization should also have a semi-unitary structure, i.e., $(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}})^* (\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}) \approx \mathbf{I}$. This observation justifies the additional orthogonality constraints that we will consider when solving (6).

Second, it is possible to exploit the structure of the channel itself to uncover additional properties of the unconstrained precoder. Assuming the channel structure of (3), and assuming the ULA geometry, then any right singular vector of the channel matrix (2), denoted here for simplicity of exposition by $\mathbf{v}$, obeys

$$\begin{aligned} & \mathbf{H}^* \mathbf{H} \mathbf{v} = \sigma^2 \mathbf{v} \quad \Rightarrow \quad \mathbf{A}_{\text{t}} \mathbf{H}_{\text{v}}^* \mathbf{H}_{\text{v}} \mathbf{A}_{\text{t}}^* \mathbf{v} = \sigma^2 \mathbf{v} \\ & \Rightarrow \; \mathbf{H}_{\text{v}}^* \mathbf{H}_{\text{v}} \mathbf{A}_{\text{t}}^* \mathbf{v} = \sigma^2 \mathbf{A}_{\text{t}}^* \mathbf{v} \quad \Rightarrow \quad \mathbf{H}_{\text{v}}^* \mathbf{H}_{\text{v}} \tilde{\mathbf{v}} = \sigma^2 \tilde{\mathbf{v}}, \end{aligned} \tag{8}$$

where we have denoted $\tilde{\mathbf{v}}$ as the Fourier transform of $\mathbf{v}$. Given that $\mathbf{H}_{\text{v}}$ has, in general, a sparse structure for mmWave communication channels, it follows that the product $\mathbf{H}_{\text{v}}^* \mathbf{H}_{\text{v}}$ is also sparse. This has two implications: i) either the singular value σ is zero, meaning that the channel matrix does not have full rank and therefore a limited number of spatial degrees of freedom; ii) or, if the singular value is non-zero then $\tilde{\mathbf{v}}$ must have less than N_{nzr} non-zero entries where N_{nzr} is the number of rows of $\mathbf{H}_{\text{v}}^* \mathbf{H}_{\text{v}}$ that have at least one non-zero element (we call them here non-zero rows), i.e., $\|\tilde{\mathbf{v}}\|_0 \leq N_{\text{nzr}}$. In essence, the singular vectors associated with the non-zero singular values have sparse Fourier transforms.

Moreover, given that we have the induced norm of the Fourier matrix is $\|\mathbf{A}_{\text{t}}^*\|_{1,\infty} = N_{\text{t}}^{-1/2}$ we can conclude that

$$\|\mathbf{A}_{\text{t}}^* \tilde{\mathbf{v}}\|_\infty \leq \frac{\|\tilde{\mathbf{v}}\|_1}{\sqrt{N_{\text{t}}}} \leq \sqrt{\frac{\|\tilde{\mathbf{v}}\|_0}{N_{\text{t}}}} \leq \sqrt{\frac{N_{\text{nzr}}}{N_{\text{t}}}}, \tag{9}$$

where we have used that $\|\tilde{\mathbf{v}}\|_1 \leq \sqrt{\|\tilde{\mathbf{v}}\|_0} \|\tilde{\mathbf{v}}\|_2$ (an inequality stemming from the numerical sparsity concept [41]) and that $\|\tilde{\mathbf{v}}\|_2 = 1$, since the unitary $\mathbf{A}_{\text{t}}^*$ preserves the ℓ_2 norm. Since the components of the optimum procoder are the Fourier transforms of sparse vectors, we know that they themselves cannot be sparse and are actually well-spread. In other words, they do not exhibit a localized support. Result (9) puts an upper bound on the largest spikes that we can encounter. By considering that $\tilde{\mathbf{v}}$ is N_{nzr}-sparse vector consisting of standard IID Gaussian random variables, a similar result can be obtained $\frac{\mathbb{E}\|\tilde{\mathbf{v}}\|_1}{\sqrt{N_{\text{t}}}} \sim \sqrt{\frac{2\|\tilde{\mathbf{v}}\|_0}{\pi N_{\text{t}}}} \leq \sqrt{\frac{2N_{\text{nzr}}}{\pi N_{\text{t}}}} \leq 0.798\sqrt{\frac{N_{\text{nzr}}}{N_{\text{t}}}}$, that describes the expected ℓ_1 behavior of the singular vectors of the channel matrix $\mathbf{H}$. Here we have used that for an ℓ_2 normalized i.i.d. random Gaussian generic vector $\mathbf{x}$, we have that $\mathbb{E}[\|\mathbf{x}\|_1 \|\mathbf{x}\|_2^{-1}] \sim \sqrt{2\pi^{-1}\|\mathbf{x}\|_0}$.

Therefore, for each component i of the optimum precoder, by (9), it follows that

$$\|(\mathbf{F}_{\text{opt}})_i\|_\infty \leq \sqrt{\frac{N_{\text{nzr}}}{N_{\text{t}}}}. \tag{10}$$

Coupled with the fact that the singular vectors have unit ℓ_2 norm, we conclude that the right singular vector are approximately flat, i.e., very unlikely to observe large magnitude peaks. In the worse case scenario, if the sparsity assumption does not hold and $N_{\text{nzr}} \approx N_{\text{t}}$, then we approach the upper limit of $\|(\mathbf{F}_{\text{opt}})_i\|_\infty \approx 1$. In the best case scenario, if $N_{\text{nzr}} = 1$, then we know that the Fourier transform of a spike results in a flat magnitude Fourier transform vector, and therefore, we reach the lower limit of $\|(\mathbf{F}_{\text{opt}})_i\|_\infty = N_{\text{t}}^{-1/2}$.

The properties of the optimum unconstrained precoder highlighted in this section will help better understand the performance of the optimization methods proposed next, and justify the orthogonality assumptions that we make. Notice that, unlike for the optimum unitary $\mathbf{W}_{\text{opt}}$, in the case of the combiner $\mathbf{W}_{\text{MMSE}}$ in (7), some of the properties described in this section do not hold. In this paper, all the proposed methods

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

work for both $\mathbf{F}_{\text{opt}}$ and $\mathbf{W}_{\text{MMSE}}$ but their performance depends on the structure of these unconstrained precoder and combiner.

IV. Proposed hybrid precoding algorithms

In this section, we develop new solutions to the hybrid precoder design problem. Several assumptions on the dimensions of the problem and other internal constraints lead to different optimization techniques that are appropriate. We explore here these techniques.

A. HD–AM: Hybrid Design by Alternating Minimization when $L_t = N_s$

We deal now with solving (6) assuming two additional constraints: that dimensions obey $L_{\text{t}} = N_{\text{s}}$, i.e., the number of streams equals the number of RF chains, and the baseband $\mathbf{F}_{\text{BB}}$ is scaled unitary (that is, a matrix with orthogonal columns of the same magnitude, after the energy normalization step).

Therefore, we would like to solve the following optimization problem related to (6):

$$\begin{aligned} \underset{\mathbf{F}_{\text{RF}}\in\mathcal{U}_{N_{\text{t}}\times L_{\text{t}}},\ \mathbf{F}_{\text{BB}}}{\text{minimize}} \quad & \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 \\ \text{subject to} \quad & \mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^* = \mathbf{F}_{\text{BB}}^*\mathbf{F}_{\text{BB}} = c\mathbf{I} \\ & \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}. \end{aligned} \tag{11}$$

where c is a constant.

Following ideas from the frame design and dictionary learning literature [20], [42], we propose to solve (11) by an alternating optimization procedure. The idea is to tackle iteratively two optimization problems that have a direct, exact, solution:

- Keep $\mathbf{F}_{\text{RF}}$ fixed and solve for unitary $\mathbf{F}_{\text{BB}}$:

$$\underset{\mathbf{F}_{\text{BB}};\ \mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^*=\mathbf{F}_{\text{BB}}^*\mathbf{F}_{\text{BB}}=\mathbf{I}}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2. \tag{12}$$

  The solution to this problem, which is called the orthonormal Procrustes problem [43], is given by $\mathbf{F}_{\text{BB}} = \mathbf{V}\mathbf{U}^*$ resulting from the singular value decomposition of $\mathbf{F}_{\text{opt}}^*\mathbf{F}_{\text{RF}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$.
- Keep unitary $\mathbf{F}_{\text{BB}}$ fixed and solve:

$$\underset{\mathbf{F}_{\text{RF}}\in\mathcal{U}_{N_{\text{t}}\times L_{\text{t}}}}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2. \tag{13}$$

  The solution here entails a two step procedure. First, notice that the objective function can be rewritten as

$$\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = \|\mathbf{F}_{\text{opt}}\mathbf{F}_{\text{BB}}^* - \mathbf{F}_{\text{RF}}\|_F^2, \tag{14}$$

  due to the unitary structure of $\mathbf{F}_{\text{BB}}$. Now the $\mathbf{F}_{\text{RF}}$ is isolated. This decoupling allows for the optimization of each entry in $\mathbf{F}_{\text{RF}}$ separately. The minimizer in $\mathcal{U}_{N_{\text{t}}\times L_{\text{t}}}$ of problem (13) is given by normalizing each entry of the variable separately, i.e., $\mathbf{F}_{\text{RF}} = (\mathbf{F}_{\text{opt}}\mathbf{F}_{\text{BB}}^*) \oslash |\mathbf{F}_{\text{opt}}\mathbf{F}_{\text{BB}}^*|$.

The proposed method, which is called Hybrid Design by Alternating Minimization (HD–AM), is described in detail in Algorithm 1. Notice that the result of this algorithm is a unitary baseband precoder and a general analog precoder constrained only by the hardware architecture. The unitary structure of $\mathbf{F}_{\text{BB}}$ is what allows the fast solutions to problems (13) and (12). As previously explained, this is a natural additional constraint for the digital component. The final energy normalization in this situation reduces to an adjustment by $N_{\text{t}}^{-1/2}$ since $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = \|\mathbf{F}_{\text{RF}}\|_F^2 = N_{\text{t}}N_{\text{s}}$ again due to the scaled unitary $\mathbf{F}_{\text{BB}}$. After the energy normalization step, $\mathbf{F}_{\text{BB}}$ is scaled unitary, not unitary.

Notice from (5) that the unitary assumption on $\mathbf{F}_{\text{BB}}$ makes it irrelevant to the maximization of the mutual information, i.e., any unitary baseband structure just cancels out. Fortunately, even with a unitary $\mathbf{F}_{\text{BB}}$, the minimization of $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2$ is closely related to the maximization of the mutual information in (5), since the local Euclidean property of the Grassmann manifold still holds [18], [44], and therefore chordal distances can be approximated well, locally, with Frobenius distances.

Also, after the last iteration of the algorithm we have the opportunity to update $\mathbf{F}_{\text{BB}}$ via the least squares. This will produce a non-unitary solution $\mathbf{F}_{\text{BB}}$, and the Frobenius norm objective function will be further reduced. Furthermore, a non-unitary $\mathbf{F}_{\text{BB}}$ will also have an impact in (5).

Remark 1. (The convergence of HD–AM to local minimum points): Notice that each step of the alternating optimization process in Algorithm 1 solves the steps exactly. Therefore, no increase in the objective function can be achieved and thus the error term is guaranteed to decrease monotonically with each iteration. Coupled with the fact that the objective function is bounded below, this shows that the proposed method converges to local optimum points. ■

Remark 2. (The computational complexity of HD–AM): Algorithm 1 involves operations that can be efficiently executed. The construction of $\mathbf{F}_{\text{RF}}$ only implies complex matrix multiplication and elementwise division operations, while the construction of $\mathbf{F}_{\text{BB}}$ necessitates a singular value decomposition on a small square matrix of size N_{s}. All these procedures are well understood and can be efficiently implemented – for example when $N_{\text{s}} = 2$ even a closed form solution exists for the singular value decomposition. Therefore, HD–AM is faster than previous methods based on matching pursuit [18] and greedy methods [24] or other iterative procedures [26]. ■

The structures we consider do not just reflect in the low computational complexity of HD–AM, but also, in its performance analysis. It has been shown in [15] (see proof of Theorem 1) that when $N_{\text{s}} = 1$ and $L_{\text{t}} = 2$ it is possible to perfectly reconstruct $\mathbf{F}_{\text{opt}}$ using a hybrid architecture. A remarks was made in [24] and [25] that in the general case of N_{s} streams, it is possible to obtain perfect reconstruction with $L_{\text{t}} = 2N_{\text{s}}$ RF chains, simply by extending the result in [15] for each component of $\mathbf{F}_{\text{opt}}$. This is a worst case scenario. If the components of $\mathbf{F}_{\text{opt}}$ themselves have already unit magnitude, perfect reconstruction is possible with just $L_{\text{t}} = N_{\text{s}}$ RF chains. In the other cases some reconstruction error might exist.

Returning to the $N_{\text{s}} = L_{\text{t}}$ case and with the choice of the solution described from (12), the objective function value is

$$\begin{aligned} \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 &= \|\mathbf{F}_{\text{opt}}\|_F^2 + N_{\text{t}}^{-1}\|\mathbf{F}_{\text{RF}}\|_F^2 \\ &\quad - 2N_{\text{t}}^{-1/2}\text{tr}(\mathbf{F}_{\text{opt}}^*\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}) \\ &= N_{\text{s}} + L_{\text{t}} - 2N_{\text{t}}^{-1/2}\text{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{V}\mathbf{U}^*) \\ &= 2N_{\text{s}} - 2N_{\text{t}}^{-1/2}\text{tr}(\mathbf{\Sigma}) = 2N_{\text{s}} - 2N_{\text{t}}^{-1/2}\|\mathbf{F}_{\text{opt}}^*\mathbf{F}_{\text{RF}}\|_*. \end{aligned} \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

Therefore, the minimization of the objective function is equivalent to maximizing the nuclear norm of $\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}$. Well-known results from the matrix analysis literature bound the nuclear norm using the Frobenius norm as

$$\|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_F \leq \|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_* \leq \sqrt{N_{\text{s}}}\|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_F, \quad (16)$$

while from the non-negativity of (15) we know that

$$\|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_* \leq N_{\text{s}}\sqrt{N_{\text{t}}}. \quad (17)$$

Let us consider a simple example where $N_{\text{s}} = L_{\text{t}}$ and we consider the $\mathbf{F}_{\text{RF}}$ from the initialization of HD–AM, i.e., $\mathbf{F}_{\text{RF}} = \mathbf{F}_{\text{opt}} \oslash |\mathbf{F}_{\text{opt}}|$. In this situation, each entry (i,j) of $\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}$ has the form $(\mathbf{F}_{\text{opt}})_i^{*}(\mathbf{F}_{\text{RF}})_j$. From Holder's inequality we have that

$$|(\mathbf{F}_{\text{opt}})_i^{*}(\mathbf{F}_{\text{RF}})_j| \leq \|(\mathbf{F}_{\text{opt}})_i\|_1\|(\mathbf{F}_{\text{RF}})_j\|_\infty = \|(\mathbf{F}_{\text{opt}})_i\|_1. \quad (18)$$

Therefore we have that

$$\|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_F \leq \sqrt{N_{\text{s}}\sum_{i=1}^{N_{\text{s}}}\|(\mathbf{F}_{\text{opt}})_i\|_1^2} \leq N_{\text{s}}\sqrt{N_{\text{t}}}, \quad (19)$$

which matches the upper bound on the nuclear norm from (17). The last inequality is due to the fact that the ℓ_1 norm of a ℓ_2 constrained vector takes the maximum value of $\sqrt{N_{\text{t}}}$ – this is achieved when the entries of the vector all have the same magnitude of $N_{\text{t}}^{-1/2}$. This extreme is intuitive: if the optimum unconstrained precoder has unit magnitude entries then $\mathbf{F}_{\text{BB}} = N_{\text{t}}^{-1/2}\mathbf{I}$ (since $\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}$ is a multiple of the identity matrix) and in fact we reach the upper limit in (16) which is $\sqrt{N_{\text{s}}}\|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_F = N_{\text{s}}\sqrt{N_{\text{t}}}$. Therefore, exact representation is achieved. At the other extreme consider the case when $\|(\mathbf{F}_{\text{opt}})_i\|_1 = 1$, the lowest possible value for any ℓ_2 constrained vector which is achieved when $(\mathbf{F}_{\text{opt}})_i \subset \{\mathbf{e}_k\}_{k=1}^{N_{\text{t}}}$. Therefore, the Frobenius norm in (19) is bounded as $\|\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}}\|_F \leq N_{\text{s}}$. In this case, using the best upper bound of (16), the objective function of (15) is bounded below by $2N_{\text{s}}(1-\sqrt{N_{\text{s}}N_{\text{t}}^{-1}})$. As the number of transmit antennas usually greatly exceeds the number of streams in a mmWave communication system, asymptotically the bound becomes $2N_{\text{s}}$ which is quite loose, considering that the target optimum precoder is constrained like $\|\mathbf{F}_{\text{opt}}\|_F^2 = N_{\text{s}}$.

As a general note, the performance of the hybrid factorization considered here is bounded by the ℓ_1 and ℓ_∞ norms of the components of the optimum constrained precoder $\mathbf{F}_{\text{opt}}$. The bounds previously described (9) on the ℓ_∞ norm of the components of $\mathbf{F}_{\text{opt}}$ show that these are approximately flat in the case of sparse scattering and therefore suitable for hybrid factorization – in the extreme case of $N_{\text{nzr}} = 1$ we know that Fourier transform of impulses have perfectly flat magnitude and therefore exact factorization is possible while in all other cases some reconstruction error is bounded as the square root of N_{nzr}.

The solution presented in this section is fast and, as we will show, performs very well. Unfortunately, this solution only holds for the special case of $L_{\text{t}} = N_{\text{s}}$. If $N_{\text{s}} < L_{\text{t}}$ then the development in (14) does not hold in general, and the computation of $\mathbf{F}_{\text{RF}}$ cannot be done simply. It is possible to extend HD–AM to cases where $N_{\text{s}} < L_{\text{t}}$. One approach is to deal with the square case and then, by using a greedy approach similar to [24] for example, extend the digital precoder with $(L_{\text{t}} - N_{\text{s}})$ components.

We move now to explore other solutions to the hybrid design problem that deal directly with any choice of the number of data streams and RF chains.

Algorithm 1 – Hybrid Design by Alternating Minimization (HD–AM)

Input: The optimum unconstrained solution $\mathbf{F}_{\text{opt}} \in \mathbb{C}^{N_{\text{t}}\times N_{\text{s}}}$ and the maximum number of iterations K.

Output: Analog $\mathbf{F}_{\text{RF}} \in \mathcal{U}_{N_{\text{t}}\times L_{\text{t}}}$ and baseband (orthogonal) $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{N_{\text{s}}\times N_{\text{s}}}$ such that $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2$ is reduced and $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$.

Initialization: Take initial precoder $\mathbf{F}_{\text{RF}} = \mathbf{F}_{\text{opt}} \oslash |\mathbf{F}_{\text{opt}}|$.

Iterations $1, \ldots, K$:

- Update $\mathbf{F}_{\text{BB}}$ by solving (12) whose solution is given by $\mathbf{F}_{\text{BB}} = \mathbf{V}\mathbf{U}^{*}$ from $\mathbf{F}_{\text{opt}}^{*}\mathbf{F}_{\text{RF}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{*}$.
- Update $\mathbf{F}_{\text{RF}}$ by solving (13) whose solution is given by $\mathbf{F}_{\text{RF}} = (\mathbf{F}_{\text{opt}}\mathbf{F}_{\text{BB}}^{*}) \oslash |\mathbf{F}_{\text{opt}}\mathbf{F}_{\text{BB}}^{*}|$.

For better performance, give up the unitary structure and update by least squares $\mathbf{F}_{\text{BB}} = (\mathbf{F}_{\text{RF}}^{*}\mathbf{F}_{\text{RF}})^{-1}\mathbf{F}_{\text{RF}}^{*}\mathbf{F}_{\text{opt}}$.
Normalize $\mathbf{F}_{\text{BB}} = \sqrt{N_{\text{s}}}\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^{-1}\mathbf{F}_{\text{BB}}$.

B. HD–FUM: Hybrid Design by Fast Unitary Matching

We have seen in the previous section that, when dimensions allow it, a mathematical simplification leads to a fast solution to the hybrid precoder design problem.

Unfortunately, the HD–AM algorithm cannot be applied in the general case, i.e. when the number of RF chains does not match the number of streams, and the difficulty of designing a hybrid precoder with the hardware constraints for $\mathbf{F}_{\text{RF}}$ has to be addressed in a different manner.

It has been shown in [18], that the problem of finding the precoder that maximizes the mutual information with the hardware constraints associated with mmWave MIMO architectures (6) can be approximated by solving

$$\begin{aligned} \underset{\mathbf{S},\ \mathbf{F}_{\text{BB}}}{\text{minimize}} \quad & \|\mathbf{F}_{\text{opt}} - (\mathbf{D}\mathbf{S})\mathbf{F}_{\text{BB}}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{S}\|_0 = L_{\text{t}},\ \|(\mathbf{D}\mathbf{S})\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}, \end{aligned} \quad (20)$$

where $\|\bullet\|_0$ is the ℓ_0 pseudo-norm accounting for the number of non-zero elements, $\mathbf{D}$ of size $N_{\text{t}} \times N_{\text{cl}}N_{\text{ray}}$ is the dictionary of array response vectors that is used to define the analog precoder as a selection of columns from the dictionary, i.e., $\mathbf{F}_{\text{RF}} = \mathbf{D}\mathbf{S}$, where $\mathbf{S}$ of size $N_{\text{cl}}N_{\text{ray}} \times L_{\text{t}}$ selects L_{t} columns from the dictionary $\mathbf{D}$, which depends on the array geometry. Note that HD–FUM is the only algorithm that requires knowledge of the array geometry; the others schemes proposed in this paper are geometry-independent.

In this section, we explore simplifications of this strategy that reduce the computational complexity of the solution to (20) proposed in [18]. We modify the process of searching columns of the overcomplete dictionary $\mathbf{D}$ to searching columns of unitary matrices, subsets of this overcomplete

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

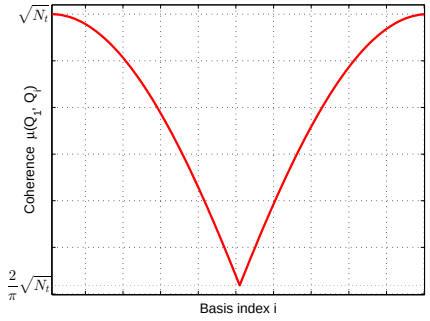

Figure 2. The coherences (22) between $\mathbf{Q}_1$ and all the other orthonormal bases $\mathbf{Q}_i$ defined within the dictionary $\mathbf{D}$ as in (21).

dictionary, such that simple correlations replace the matching pursuit iterations. This approach was first suggested in [31].

Recall that the N_t–element steering vectors of an ULA with inter-element spacing $d = \lambda/2$, that make up the dictionary $\mathbf{D}$, is $\mathbf{a}(\phi) = \frac{1}{\sqrt{N_\text{t}}}\begin{bmatrix}1 & e^{jkd\sin\phi} & \dots & e^{jkd(N_\text{t}-1)\sin\phi}\end{bmatrix}^T$, where $k = 2\pi/\lambda$ is the wavenumber and λ is the wavelength. The inner products, in absolute value, between any two such distinct vectors, assuming the sine terms are uniformly quantized in N points in the interval $[-1, 1)$ (i.e., $\sin\phi_\ell = -1 + 2\ell/N$, for $\ell = 0, \dots, N-1$), are given by: $|\mathbf{a}(\phi_\ell)^*\mathbf{a}(\phi_i)| = \frac{1}{N_\text{t}}\left|\frac{\sin(\pi(\ell-i)N_\text{t}/N)}{\sin(\pi(\ell-i)/N)}\right|$. The parameter N is the angular resolution. Higher N leads to better performance, since the discretization on the grid is finer. Unfortunately, it comes with higher computational cost and the lost of recovery performance guarantees based on the mutual coherence [45], since $|\mathbf{a}(\phi_\ell)^*\mathbf{a}(\phi_i)| \to 1$ as $N \to \infty$. In general, the objective is to solve the problem on a grid as fine as possible, i.e. N as large as possible. Because the inner products only depend on the distance between the two steering vectors $(\ell - i)$ – this actually leads to a circulant Hermitian Gram matrix $\mathbf{G} = \mathbf{D}^*\mathbf{D}$. Because of this, when $\alpha = N/N_\text{t}$ is an integer, it is possible to construct α unitary matrices $\mathbf{Q}_\ell$ of size $N_\text{t} \times N_\text{t}$ by choosing columns of $\mathbf{D}$ equally spaced with distance N_t. In this way, $\mathbf{D}$ is viewed, after column permutation, as a concatenation of α unitary matrices

$$\mathbf{D} = \begin{bmatrix}\mathbf{Q}_1 & \mathbf{Q}_2 & \dots & \mathbf{Q}_\alpha\end{bmatrix}. \tag{21}$$

We are interested in these unitary matrices since $\mathbf{F}_\text{opt}$ is semi-unitary (since this matrix has orthonormal columns but is not square, i.e., $\mathbf{F}_\text{opt}^*\mathbf{F}_\text{opt} = \mathbf{I}$ but $\mathbf{F}_\text{opt}\mathbf{F}_\text{opt}^* \neq \mathbf{I}$). By this argument we search for $\mathbf{F}_\text{RF}$ among the α unitary matrices from $\mathbf{D}$. We move to solve (20) with an additional constraint that the solution $\mathbf{F}_\text{RF}$ needs to be a unitary matrix (but still composed of steering vectors). This means that we are searching for $\mathbf{F}_\text{RF}$ among the matrices $\mathbf{Q}_\ell$. The method proposed in [31] computed all α correlations of each $\mathbf{Q}_\ell$ with the target signals and keeps the basis that concentrates most of the energy in a sparse support of fixed size N_s. We now further develop this approach to take full advantage of the structures in $\mathbf{D}$.

Remark 3. (Efficient sparse recovery in the dictionary D): Our objective is to represent sparsely the signals in one of the basis $\mathbf{Q}_\ell$ of $\mathbf{D}$. Now we consider properties that provide some insight into how to search for this basis.

Consider the coherence between two orthobases given by

$$\mu(\mathbf{Q}_i, \mathbf{Q}_\ell) = \sqrt{N_\text{t}} \max_{\mathbf{q}_i \in \mathbf{Q}_i, \mathbf{q}_\ell \in \mathbf{Q}_\ell} |\mathbf{q}_i^* \mathbf{q}_\ell|, \tag{22}$$

for all columns $\mathbf{q}_i$ and $\mathbf{q}_\ell$ of the bases $\mathbf{Q}_i$ and $\mathbf{Q}_\ell$ respectively. The coherence μ, which is bounded by $1 \leq \mu \leq \sqrt{N_\text{t}}$, is used to provide guarantees about the capacity of each basis to construct a sparse representation of a given signal. The classical results can be summarized as: given that a signal is represented in two orthobases $\mathbf{Q}_i$ with support in the set Γ_i and $\mathbf{Q}_\ell$ with support in Γ_ℓ, the uncertainty principles state

$$|\Gamma_i| + |\Gamma_\ell| \leq \frac{2\sqrt{N_\text{t}}}{\mu(\mathbf{Q}_i, \mathbf{Q}_\ell)}, \;\; |\Gamma_i| + |\Gamma_\ell| \sim \frac{N_\text{t}}{\mu(\mathbf{Q}_i, \mathbf{Q}_\ell)^2}, \tag{23}$$

where the first result holds for *all* sets Γ_i, Γ_j and the second result, called the robust uncertainty principle, holds for *most* sets Γ_i, Γ_j [46]. Given the coherence, these results provide insights into the possibility of two bases being able to provide a sparse representation of the same signal. High coherence leads to similar sparse representations while low coherence denotes the fact that there are signals that cannot be sparse in both bases.

Since the full dictionary $\mathbf{D}$ is constructed on a fine grid, adjacent bases display high coherence, while distant bases display lower coherence. This is shown in Figure 2 where the upper limit is the upper limit of coherence, i.e., $\sqrt{N_\text{t}}$. The lower limit of coherence in Figure 2 occurs for $\mathbf{Q}_1$ at $\mathbf{Q}_\ell$, where $\ell = \alpha/2 + 1$ assuming an even α. For references different than $\mathbf{Q}_1$, the coherences follow the same shape in Figure 2 but circularly shifted. Thus the lower bound of coherence is given by computing the Dirichlet kernel value when $N = \alpha N_\text{t}$ and letting $N_\text{t} \to \infty$

$$\frac{\mu(\mathbf{Q}_1, \mathbf{Q}_{\alpha/2+1})}{\sqrt{N_\text{t}}} = \lim_{N_\text{t}\to\infty} \frac{1}{N_\text{t}} \sin^{-1}\left(\frac{\pi}{2N_\text{t}}\right) = \frac{2}{\pi}. \tag{24}$$

As such, we may approximate the lower limit in Figure 2 as $2\pi^{-1}\sqrt{N_\text{t}} \approx 0.6366\sqrt{N_\text{t}}$. This is large when compared to the lower coherence limit of 1, that is well known to be achievable, e.g., $\mu(\mathbf{F}, \mathbf{I}) = 1$, being $\mathbf{F}$ the Fourier matrix. Still, this is to be expected, since all basis belong to the same class of complex exponentials. Ultimately, this means for the sets in (23) that

$$|\Gamma_i| + |\Gamma_\ell| \leq \pi \text{ and } \; |\Gamma_i| + |\Gamma_\ell| \sim 0.25\pi^2. \tag{25}$$

Because all the bases $\mathbf{Q}_\ell$ are highly correlated, any signal representations in any two of them will be very similar (in terms of the support and of the representation error). Unfortunately, this means that some class of signals will be poorly represented in all bases $\mathbf{Q}_\ell$.

Furthermore, we have that $\mu(\mathbf{F}, \mathbf{Q}_1) = \sqrt{N_\text{t}}$ (therefore both basis display similar sparse representations capabilities) and that $\mu(\mathbf{I}, \mathbf{Q}_\ell) = 1$ for all ℓ (all basis of the dictionary $\mathbf{D}$ are maximally different from the identity matrix). Since for $\mathbf{F}_\text{RF}$ we keep L_t columns from the selected basis, the representation performance depends on the relationship between L_t and N_nzr. Still, as depicted in Figure 2, we expect that the energy in the compact representations of size L_t will vary smoothly with the basis and that they exhibit overall good diversity to cover

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

Algorithm 2 – Hybrid Design by Fast Unitary Matching (HD–FUM)

Input: The optimum unconstrained solution $\mathbf{F}_{\text{opt}} \in \mathbb{C}^{N_{\text{t}} \times N_{\text{s}}}$, the number of RF chains L_{t}, the maximum number of iterations K and the resolution α.

Output: Analog $\mathbf{F}_{\text{RF}} \subset \mathbf{D}$ of size $N_{\text{t}} \times L_{\text{t}}$ and baseband $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{L_{\text{t}} \times N_{\text{s}}}$ such that $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2$ is reduced and $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$.

Initialization: Compute correlations $\mathbf{Q}_1^*\mathbf{F}_{\text{opt}}$ or $\mathbf{Q}_{\alpha/2+1}^*\mathbf{F}_{\text{opt}}$ and keep the indices of the largest L_{t} correlations in magnitude in the set $\mathcal{I}$.

- **Compute semi-unitary analog $\mathbf{F}_{\text{RF}}$**:
  - For each ℓ evaluate the energy compaction capability of $\mathbf{Q}_\ell$ in the set $\mathcal{I}$: Compute correlations $\mathbf{r}_\ell = \text{norms}(\mathbf{Q}_\ell^*\mathbf{F}_{\text{opt}})$ only on the set $\mathcal{I}$, where the function computes the 2–norm for each row of the resulting product – $\mathbf{r}_\ell$ is of size L_{t}. Compute the overall energy contribution of $\mathbf{Q}_\ell$ on the set $\mathcal{I}$ to: $c_\ell = \|\mathbf{r}_\ell\|_2$. Select $\ell_{\max} = \arg\max_{\ell} c_\ell$.
  - Set $\mathbf{F}_{\text{RF}}$ as L_{t} columns of $\mathbf{Q}_{\ell_{\max}}$ indexed in $\mathcal{I}$.
- **Compute baseband $\mathbf{F}_{\text{BB}}$**: With the previously computed semi-unitary $\mathbf{F}_{\text{RF}}$ compute the baseband either by solving a least squares problem $\mathbf{F}_{\text{BB}} = \mathbf{F}_{\text{RF}}^*\mathbf{F}_{\text{opt}}$ or compute a semi-unitary $\mathbf{F}_{\text{BB}}$ by the comments made in Remark 5.

Normalize baseband $\mathbf{F}_{\text{BB}} = \sqrt{N_{\text{s}}}\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^{-1}\mathbf{F}_{\text{BB}}$.

well a significant proportion of sparse signals. The comments just made justify the choice made for the dictionary $\mathbf{D}$, and therefore we should expect good performance when using these structures.

For a computationally efficient implementation of the proposed method, we now make several observations:

1) for even N_{t}: $\mathbf{Q}_1^*\mathbf{F}_{\text{opt}} = \sqrt{N_{\text{t}}}\begin{bmatrix}\mathbf{J}_{N_{\text{t}}/2+1}\\ \mathbf{J}_{N_{\text{t}}/2-1}\end{bmatrix}\text{IFFT}(\mathbf{F}_{\text{opt}})$.
2) for odd N_{t}: $\mathbf{Q}_{\alpha/2+1}^*\mathbf{F}_{\text{opt}} = \sqrt{N_{\text{t}}}\begin{bmatrix}\mathbf{J}_{(N_{\text{t}}+1)/2}\\ \mathbf{J}_{(N_{\text{t}}-1)/2}\end{bmatrix}\text{IFFT}(\mathbf{F}_{\text{opt}})$.

where $\mathbf{J}_n$ denotes the anti-diagonal square matrix of size $n \times n$. Therefore, an opportunity to efficiently compute the correlations presents itself for $\ell \in \{1, \alpha/2+1\}$ – by efficient we mean computational complexity of $O(N_{\text{t}} \log N_{\text{t}})$. Now, since by the arguments previously made, all $\mathbf{Q}_\ell$ are highly coherent, after computing $\mathbf{Q}_1^*\mathbf{F}_{\text{opt}}$ or $\mathbf{Q}_{\alpha/2+1}^*\mathbf{F}_{\text{opt}}$ we can consider the support of size L_{t} known, and the same in each of the other basis. Therefore, we compute only the correlations with these positions, not with the full basis – this reduces the computational complexity of computing the correlations in any other of the basis from $O(N_{\text{t}}^2)$ to $O(L_{\text{t}}N_{\text{t}})$. ■

Therefore, we can construct an algorithm to search efficiently among the $\mathbf{Q}_\ell$ for the analog component $\mathbf{F}_{\text{RF}}$. Also, due to the extra structure from Figure 2, it is possible to avoid having to do a full sweep for all ℓ. The baseband $\mathbf{F}_{\text{BB}}$ can be computed from a least squares problem or with an additional semi-unitary structure.

Remark 4. (The computation of a semi-unitary $\mathbf{F}_{\text{BB}}$): The computation of a square ($L_{\text{t}} = N_{\text{s}}$) unitary $\mathbf{F}_{\text{BB}}$ was described in (12). Its solution is given by $\mathbf{F}_{\text{BB}} = \mathbf{V}\mathbf{U}^*$ resulting from $\mathbf{F}_{\text{opt}}^*\mathbf{F}_{\text{RF}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. From the calculation in (15), if we allow for a semi-unitary solution $\mathbf{F}_{\text{BB}}$ of size $L_{\text{t}} \times N_{\text{s}}$ with $N_{\text{s}} < L_{\text{t}}$, one choice is $\mathbf{F}_{\text{BB}} = \mathbf{V}\bar{\mathbf{U}}^*$ where $\bar{\mathbf{U}}$ contains the first N_{s} columns of $\mathbf{U}$. Choosing the singular vectors associated with the largest singular values ensures a maximal decrease in the objective function, i.e., a maximal $\text{tr}(\mathbf{\Sigma})$. This semi-unitary $\mathbf{F}_{\text{BB}}$ reaches the lowest objective function. ■

The detailed proposed method is presented in Algorithm 2.

Remark 5. (Refining the semi-unitary solutions $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$): The constraints that both the analog and baseband components of the hybrid solution have to be semi-unitary (or unitary when the choice of dimensions allows it) seems restrictive, especially when considering that the goal is to maximize spectral efficiency (5). Taking into account the channel model (2), the channel estimation problem can be written as $\mathbf{Y} = \mathbf{A}\text{vec}(\mathbf{H}_{\text{v}})$ [47], [48] where the measurement matrix is

$$\mathbf{A} = (\mathbf{F}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{S})^T \otimes (\mathbf{W}_{\text{BB}}^*\mathbf{W}_{\text{RF}}^*\mathbf{F}^*). \tag{26}$$

Notice that $\mathbf{A}\mathbf{A}^* = \mathbf{I} \otimes \mathbf{I} = \mathbf{I}$ when the analog and digital components and the training set $\mathbf{S}$ are semi-unitary and thus $\mathbf{A}$ is a tight frame. Tight frames have been recently shown to reduce the mean squared error [49] when performing CS reconstruction with either matching or basis pursuit. ■

Once the analog $\mathbf{F}_{\text{RF}}$ has been computed by HD–FUM, a computational cheap way to improve the performance is to allow for a local search around the components indexed in $\mathcal{I}$. Details of such a procedure are outlined in [31]. We call this approach HD–FUM + Local Search (LocS). This step could be necessary since assuming both analog and digital components are semi-unitary might be too restrictive and lead to large errors in the approximation (6).

Notice that each iteration of HD–FUM only computes correlations (matrix-matrix multiplication) and ℓ_2 norms avoiding therefore expensive matrix decompositions. Furthermore, due to the structure of the 1D search space, a sequential search is avoided and thus the solution can be found quickly. As a consequence, HD–FUM is in practice faster, or presents comparable running speed, with the proposed HD–AM algorithm. Unfortunately, HD–FUM will perform worse in terms of rate maximization than the, much slower but more general, matching pursuit solution presented in [18].

The operation of HD–FUM depends on the geometry of the arrays deployed. In this section, we derived the dictionary $\mathbf{D}$ from a ULA and exploited its interpretation as a concatenation of unitary structures (as in (21)). Extending the algorithm to work with other array geometries may be challenging, unless they have special structure. For example, it could be modified to work with an uniform planar array (UPA), which can be written in terms of unitary structures as its array manifold is a Kronecker product of ULA manifolds. Alternatively, the array can be written in terms of a generalized ULA [50], but we leave the details for future work.

C. HD–CVXR: Hybrid Design by Convex Relaxation

One of the main difficulties, from a mathematical optimization perspective, in solving (6) is the unitary constraint

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

Algorithm 3 – Hybrid Design by Convex Optimization Relaxation (HD–CVXR)

Input: The optimum unconstrained solution $\mathbf{F}_{\text{opt}} \in \mathbb{C}^{N_{\text{t}} \times N_{\text{s}}}$, the number of RF chains L_{t} and the maximum number of iterations K.

Output: Analog $\mathbf{F}_{\text{RF}} \in \mathcal{U}_{N_{\text{t}} \times L_{\text{t}}}$ and baseband $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{L_{\text{t}} \times N_{\text{s}}}$ such that $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2$ is reduced and $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$.

Initialization: Construct $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$ randomly or by using HD–AM (when dimensions allow it) or HD–FUM.

Iterations $1, \ldots, K$:

- Update of $\mathbf{F}_{\text{RF}}$: for each column $(\mathbf{F}_{\text{RF}})_j$
  - Solve for variable $\mathbf{x} \in \mathbb{C}^{N_{\text{t}}}$:

$$\underset{\mathbf{x};\ |\mathbf{x}| - \mathbf{1} \leq \beta \mathbf{1}}{\text{minimize}} \quad \|\mathbf{R}_j - \mathbf{x}(\mathbf{F}_{\text{BB}})_j\|_F^2, \tag{27}$$

  where we define $\mathbf{R}_j = \mathbf{F}_{\text{opt}} - (\mathbf{F}_{\text{RF}})_{-j}(\mathbf{F}_{\text{BB}})_{-j}$.
  - Normalize and update $(\mathbf{F}_{\text{RF}})_j = \mathbf{x} \oslash |\mathbf{x}|$.
- Update by least squares $\mathbf{F}_{\text{BB}} = (\mathbf{F}_{\text{RF}}^* \mathbf{F}_{\text{RF}})^{-1} \mathbf{F}_{\text{RF}}^* \mathbf{F}_{\text{opt}}$.

Normalize baseband $\mathbf{F}_{\text{BB}} = \sqrt{N_{\text{s}}} \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^{-1} \mathbf{F}_{\text{BB}}$.

of the analog component $\mathbf{F}_{\text{RF}} \in \mathcal{U}_{N_{\text{t}} \times L_{\text{t}}}$. The dictionary based solution HD–FUM solves this problem by restricting the columns of $\mathbf{F}_{\text{RF}}$ to belong to a finite set, the unit magnitude columns of the dictionary $\mathbf{D}$. The first solution proposed in this paper (HD–AM) deals with the unitary constraint in a general; sadly, it can be applied only with a particular choice of the dimensions (the number of RF chains L_{t} equals the number of streams N_{s}).

In this section we describe a general approach for the optimization of $\mathbf{F}_{\text{RF}}$ based on a convex optimization relaxation techniques. Ideally, with a fixed $\mathbf{F}_{\text{BB}}$, we would like to solve exactly the non-convex optimization problem in (11), which omitting the total energy constraint, can be written as:

$$\underset{\mathbf{F}_{\text{RF}};\ |(\mathbf{F}_{\text{RF}})_{i,j}| = 1,\ \forall\ (i,j)}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2. \tag{28}$$

Since all the constraints of keeping the entries on the complex unit circle are non-convex, the first idea is to create an allowed trust region where to search for the solution. This way, the problem becomes:

$$\underset{\mathbf{F}_{\text{RF}};\ |(|(\mathbf{F}_{\text{RF}})_{i,j}| - 1)| \leq \beta,\ \forall\ (i,j)}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2, \tag{29}$$

and is followed by the normalization $\mathbf{F}_{\text{RF}} = \mathbf{F}_{\text{RF}} \oslash |\mathbf{F}_{\text{RF}}|$ to return the solution to the set $\mathcal{U}_{N_{\text{t}} \times L_{\text{t}}}$. The parameter $0 < \beta \ll 1, \beta \in \mathbb{R}$ defines the size of the trust region. Unfortunately, the relaxation in (29) is again non-convex since the trust region around the unit circle is non-convex. To solve this last issue we finally reach the convex optimization problem:

$$\underset{\mathbf{F}_{\text{RF}};\ |(\mathbf{F}_{\text{RF}})_{i,j}| - 1 \leq \beta,\ \forall\ (i,j)}{\text{minimize}} \quad \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2, \tag{30}$$

followed necessarily by the same normalization as for (29) for all $N_{\text{t}}L_{\text{t}}$ variables. A difficulty is the choice of the parameter β: a small value means that the relaxation made is more precise but we restrict the search space and therefore a larger number of iterations may be needed. In general, we would like to perform the optimization in (30) with a value of β as large as possible while guaranteeing other properties, like a monotonically decreasing objective function for example. Unfortunately, no obvious way of selecting the parameter β presents itself. The heuristic employed in this paper is to choose a value for β small enough as to ensure, empirically, convergence of the method with the possible cost of running the algorithm for a larger number of iterations. The value we consider in this paper is $\beta = 0.05$.

The choice presents itself here: optimize over all indices (i,j) or restrict the optimization per column or even per entry and cover the whole $\mathbf{F}_{\text{RF}}$ through an iterative process. For clarity of exposition we define the following structures:

- $(\mathbf{F}_{\text{RF}})_j$ is the j^{th} column of $\mathbf{F}_{\text{RF}}$ and $(\mathbf{F}_{\text{RF}})_{-j}$ is the matrix composed of the columns of $\mathbf{F}_{\text{RF}}$ except for the j^{th} column.
- $(\mathbf{F}_{\text{BB}})_j$ is the j^{th} line of $\mathbf{F}_{\text{BB}}$ and $(\mathbf{F}_{\text{BB}})_{-j}$ is the matrix composed of the lines of $\mathbf{F}_{\text{BB}}^*$ expect for the j^{th} line.

The detailed proposed algorithm, called HD–CVXR, which alternates between updating each column of $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$ is described in Algorithm 3.

Remark 6. (An initialization for HD–CVXR): A computational inexpensive initialization is to take for $\mathbf{F}_{\text{RF}}$ the first N_{s} columns of $\mathbf{F}_{\text{opt}}$ and normalize them such that each entry has unit magnitude. The other columns of $\mathbf{F}_{\text{RF}}$, up to L_{t}, can be taken randomly or normalized random linear combinations of the first N_{s} columns of $\mathbf{F}_{\text{opt}}$. In the general setting, the baseband $\mathbf{F}_{\text{BB}}$ is computed simply by the least squares. ■

Algorithm 4 – Hybrid Design by Least Squares Relaxation (HD–LSR)

Input: The optimum unconstrained solution $\mathbf{F}_{\text{opt}} \in \mathbb{C}^{N_{\text{t}} \times N_{\text{s}}}$, the number of RF chains L_{t} and the maximum number of iterations K.

Output: Analog $\mathbf{F}_{\text{RF}} \in \mathcal{U}_{N_{\text{t}} \times L_{\text{t}}}$ and baseband $\mathbf{F}_{\text{BB}} \in \mathbb{C}^{L_{\text{t}} \times N_{\text{s}}}$ such that $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2$ is reduced and $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$.

Initialization: Construct $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$ randomly or by using HD–AM (when dimensions allow it) or HD–FUM.

Iterations $1, \ldots, K$: For each entry of $\mathbf{F}_{\text{RF}}$:

- Update $(\mathbf{F}_{\text{RF}})_{i,j}$ and normalize by (34).
- With the new $\mathbf{F}_{\text{RF}}$, update the baseband by least squares $\mathbf{F}_{\text{BB}} = (\mathbf{F}_{\text{RF}}^* \mathbf{F}_{\text{RF}})^{-1} \mathbf{F}_{\text{RF}}^* \mathbf{F}_{\text{opt}}$.

Normalize baseband $\mathbf{F}_{\text{BB}} = \sqrt{N_{\text{s}}} \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^{-1} \mathbf{F}_{\text{BB}}$.

D. HD–LSR: Hybrid Design by Least Squares Relaxation

From the previous discussion it should be clear that some heuristic relaxation needs to be applied to deal with the hard unit magnitude constraints that plague the entries of $\mathbf{F}_{\text{RF}}$. The approach presented in HD–CVXR is to update iteratively each column of $\mathbf{F}_{\text{RF}}$. Alternatively, the whole $\mathbf{F}_{\text{RF}}$ could be updated simultaneously by the same process, although this might lead to large optimization problems that require substantial computational power and memory storage. Let us explore the other extreme possibility: update $\mathbf{F}_{\text{RF}}$ entry by entry.

The convex optimization in (27) updates the full column $(\mathbf{F}_{\text{RF}})_j$ in one stroke. An alternative option is to update each entry $(\mathbf{F}_{\text{RF}})_{i,j}$ of this column separately. Consider the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

objective function from (27)

$$\mathbf{F}_{\text{opt}}-(\mathbf{F}_{\text{RF}})_{-j}(\mathbf{F}_{\text{BB}})_{-j}-(\mathbf{F}_{\text{RF}})_{j}(\mathbf{F}_{\text{BB}})_{j}=\mathbf{R}-(\mathbf{F}_{\text{RF}})_{j}(\mathbf{F}_{\text{BB}})_{j}.$$

If we now treat each entry $(\mathbf{F}_{\text{RF}})_{i,j}$ of $(\mathbf{F}_{\text{RF}})_j$ separately then we reach a series of N_{t} decoupled optimization problems of the form:

$$\underset{\gamma\in\mathbb{C}}{\text{minimize}} \quad \|(\mathbf{R})_i-\gamma\left((\mathbf{F}_{\text{RF}})_{i,j}(\mathbf{F}_{\text{BB}})_j\right)\|_F^2, \tag{31}$$

where $(\mathbf{R})_i$ is the i^{th} row of $\mathbf{R}$. For simplicity and clarity we can just denote the optimization problem (31) as

$$\begin{aligned}&\underset{\gamma\in\mathbb{C}}{\text{minimize}} \quad \|\mathbf{y}-\gamma\mathbf{x}\|_F^2\\ &\text{with fixed } \mathbf{x}=(\mathbf{F}_{\text{RF}})_{i,j}(\mathbf{F}_{\text{BB}})_j \text{ and } \mathbf{y}=(\mathbf{R})_i.\end{aligned} \tag{32}$$

The least squares solution to this problem is given by

$$\gamma_{\text{LS}}=(\mathbf{x}^*\mathbf{y})\|\mathbf{x}\|_2^{-2}. \tag{33}$$

In general this solution does not have unit magnitude. Therefore we propose to replace it by $\gamma=(\mathbf{x}^*\mathbf{y})|\mathbf{x}^*\mathbf{y}|^{-1}$. This is a natural modification since γ is the closest solution to the complex unit circle in the ℓ_2 sense to γ_{LS}. The final updated entry is

$$(\mathbf{F}_{\text{RF}})_{i,j}=(\mathbf{x}^*\mathbf{y})|\mathbf{x}^*\mathbf{y}|^{-1}(\mathbf{F}_{\text{RF}})_{i,j}. \tag{34}$$

Based on these observation we propose an algorithm called Hybrid Design by Least Squares Relaxation (HD–LSR), a variant of HD–CVXR for the single entry optimization case. The full procedure is described in Algorithm 4.

Remark 8. (The convergence of HD–LSR to local minimum points): We know that γ_{LS} is the best unconstrained solution in terms of reducing the objective function of (31). The question now is if after normalization we can still guarantee a decrease in this objective function. Mathematically this question is equivalent to asking if

$$\|\mathbf{y}-\mathbf{x}\|_F^2-\|\mathbf{y}-\gamma\mathbf{x}\|_F^2\geq 0. \tag{35}$$

Developing for the trace definition of the Frobenius norm and using (34) we reach that

$$|\mathbf{x}^*\mathbf{y}|-\Re(\mathbf{x}^*\mathbf{y})\geq 0, \tag{36}$$

which is always true and therefore the proposed solution (34) necessarily leads to a smaller objective function value with each iteration. Therefore, with each iteration the objective function value is monotonically decreasing and since it is bounded below we have that HD–LSR monotonically converges to a local optimum point. ■

Since the elements $(\mathbf{F}_{\text{RF}})_{i,j}$ are updated sequentially and each step is followed by the update of the baseband component $\mathbf{F}_{\text{BB}}$ some attention has to be dedicated to the computational complexity of the proposed solution. Considering a general least squares solution for $\mathbf{F}_{\text{BB}}$ this does not have to be computed entirely at each iteration. We can use knowledge from previous iterations to update the pseudoinverse by using well known results from the matrix computation literature [51] based on the Sherman-Morrison-Woodbury formulas.

Since the update of $\mathbf{F}_{\text{RF}}$ is done sequentially the subsequent calculation of $\mathbf{F}_{\text{BB}}$ can be done efficiently, using information from the previous iteration. After the initialization of $\mathbf{F}_{\text{RF}}$ the first computation of $\mathbf{F}_{\text{BB}}$ is done by solving the full problem

$$(\mathbf{F}_{\text{RF}}^*\mathbf{F}_{\text{RF}})\mathbf{F}_{\text{BB}}=\mathbf{F}_{\text{RF}}^*\mathbf{F}_{\text{opt}}. \tag{37}$$

Let us consider now that $\mathbf{F}_{\text{RF}}$ has been updated on column j, and the goal is to update $\mathbf{F}_{\text{BB}}$. Having the solution from the previous iteration, there are two ways in which the update of $\mathbf{F}_{\text{BB}}$ is simplified:

1) The right hand term $\mathbf{F}_{\text{RF}}^*\mathbf{F}_{\text{opt}}$ needs to be recomputed only on the j^{th} line, reducing the computation complexity from $O(N_{\text{s}}L_{\text{t}}N_{\text{t}})$ to $O(N_{\text{s}}N_{\text{t}})$.
2) Solving the remaining linear system for $\mathbf{F}_{\text{BB}}$ is done efficiently assuming $\mathbf{F}_{\text{RF}}$ is full rank and that some factorization of the previous positive definite $\mathbf{F}_{\text{RF}}^*\mathbf{F}_{\text{RF}}$ is available and now we need to compute a factorization of $(\mathbf{F}_{\text{RF}}+\delta\mathbf{e}_j^T)^*(\mathbf{F}_{\text{RF}}+\delta\mathbf{e}_j^T)$ where $\delta\in\mathbb{C}^{N_{\text{t}}\times 1}$ is the update on the j^{th} column of $\mathbf{F}_{\text{RF}}$ and $\mathbf{e}_j$ is the j^{th} component of the standard basis of $\mathbb{C}^{N_{\text{t}}}$. The linear algebra literature has studied the problem of rank–1 updates to QR and LU available factorizations (see Sections 3.2 and 3.3 of [52]). These methods lower the complexity of computing the factorization from scratch $O(L_{\text{t}}^3)$ to $O(L_{\text{t}}^2)$.

Since in mmWave systems make use of large antenna arrays then dimensions usually obey $L_{\text{t}}\ll N_{\text{t}}$ the reduction of the computational complexity is not significant since the complexity of the overall problem is dictated mainly by the number of antennas N_{t}. As the computational complexity of producing the least squares solution depends only linearly on N_{t} we can afford a seldom update of $\mathbf{F}_{\text{BB}}$.

Out of the proposed solutions, HD–CVXR and HD–LSR are the most general. No assumption is made on the problem dimensions or the internal structure of the analog and baseband components. Therefore, we expect this approach to perform best in terms of data rates, i.e., from the proposed solutions HD–CVXR and HD–LSR will produce the closest factorizations $\mathbf{F}_{\text{opt}}\approx\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}$ with the unit magnitude constraint on the analog component.

V. Results

In this section we provide numerical evidence of the performance of the proposed methods and comparisons with previous methods from the literature. A major objective is to establish trade-offs spectral efficiency-complexity for the different proposed algorithms.

Throughout this section we assume the clustered narrowband channel model (2) with N_{cl} clusters and N_{ray} propagation paths per cluster assuming all clusters are of equal power satisfying the normalization constraint $\mathbb{E}[\|\mathbf{H}\|_F^2]=N_{\text{t}}N_{\text{r}}$. The angles of departure and arrival are normal randomly distributed with mean cluster angle uniformly randomly distributed in $[0,2\pi]$, while the angle spread is set to 7.5. The total power constraint is fixed for all precoders with equal power allocation per stream and SNR$=\rho\sigma_n^{-2}$. We show results for low and medium SNR levels, since these are the settings where mmWave systems usually operate. We assume ULAs at the transmitter and receiver. We present now numerical simulations that show the performance of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2016.2614495, IEEE Transactions on Wireless Communications

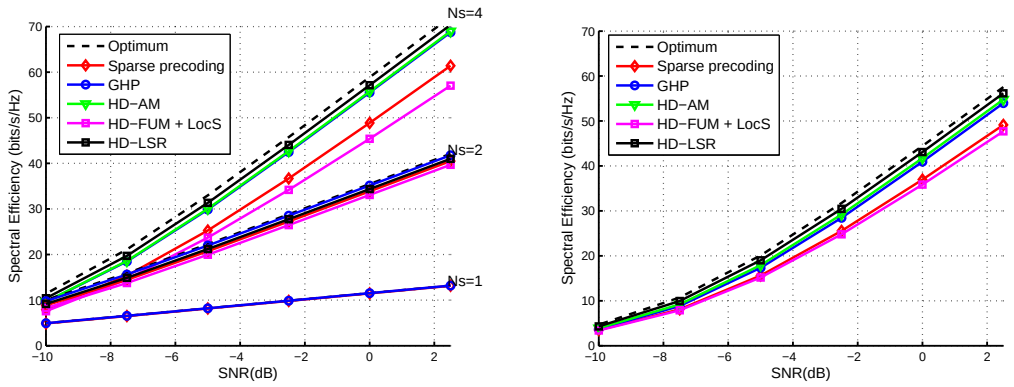

Figure 3. Average spectral efficiency (5) achieved by the proposed methods as compared to the optimum and previous methods from the literature GHP [24] and sparse precoding [18]. The results are averaged over 100 random realizations of the channel. (a) The experimental setup is: $N_{\mathrm{t}} = N_{\mathrm{r}} = 128$ antennas, $L_{\mathrm{t}} = L_{\mathrm{r}} = 4$ RF chains and $N_{\mathrm{s}} \in \{1, 2, 4\}$ streams while the channel model has $N_{\mathrm{cl}} = 3$ and $N_{\mathrm{ray}} = 18$. (b) Experimental setup analogous to (a) for $N_{\mathrm{t}} = N_{\mathrm{r}} = 64$ antennas with $N_{\mathrm{s}} = 8$ data streams and $L_{\mathrm{t}} = L_{\mathrm{r}} = 8$ RF chains.

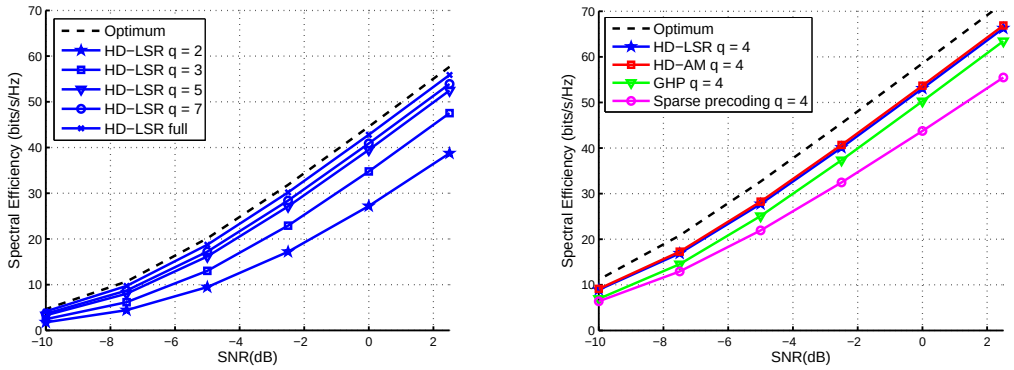

Figure 4. Average spectral efficiency when the analog component of the hybrid precoder and combiner is quantized. We also show for perspective the performance of the optimum fully digital design and the hybrid design with full precision. (a) The precoder is designed via HD–LSR and quantized with q levels. The experimental setup is the same as in Figure 3(b). (b) The hybrid precoder and combiner are designed via HD–LSR, HD–AM, GHP and Sparse precoding that are quantized with $q = 4$ levels. The experimental setup is the same as in Figure 3(a).

proposed methods when maximizing the spectral efficiency as defined in (5). Given that we assume the channel realization is known, we perform the singular value decomposition to retrieve the optimum unconstrained fully digital precoder $\mathbf{F}_{\mathrm{opt}}$ and combiner $\mathbf{W}_{\mathrm{opt}}$. Then we derive their hybrid factorizations using the proposed methods and we compare with the recently previously proposed methods from the literature.

Figure 3 shows the spectral efficiency results for the proposed methods and previous methods from the literature for various number of data streams $N_{\mathrm{s}} \in \{1, 2, 4\}$. For a low number of streams N_{s} the performance of all the methods are very close to the optimum fully digital solution. The best performing methods are the proposed ones, followed closely by GHP.

Finally we examine the effects of quantization when designing the analog component of the hybrid precoder and combiner. Figure 4(a) shows the spectral efficiency of HD–LSR, that now also uses q levels of quantization for the entries of the analog component. The entries are quantized immediately after they are computed. The results show that moderate levels of quantization already provide very good results, comparable to the full, un-quantized, HD–LSR. The quantization is done after the update of each entry $(\mathbf{F}_{\mathrm{RF}})_{i,j}$ by choosing the closest element, in the ℓ_2 norm sense, from a dictionary of q equally spaced elements placed on the complex unit circle. Figure 4(b) compares the proposed methods with previously introduced algorithms when quantization is done with $q = 4$ levels. For each method, the quantization is performed immediately after any update to $\mathbf{F}_{\mathrm{RF}}$, while for the dictionary based method quantization is done element-wise after the construction of the transmit and receive dictionaries. Observe that HD–FUM cannot be applied since the quantization destroys the internal dictionary structure exploited by this method. The dictionary based method performs the worst followed by the GHP, and then, with very similar performance, the two proposed methods HD–AM and HD–LSR. In the last steps of HD–AM and HD–FUM we give up the orthogonal structure of $\mathbf{F}_{\mathrm{BB}}$ which we compute via the least squares for better performance.

In terms of the running times of the proposed algorithms we report the following: HD–FUM takes approximately 0.0001 seconds, HD–AM takes 0.0009 seconds, HD–LSR takes approximately 0.6 seconds while the slowest is HD–CVXR takes

Table I
A BRIEF SUMMARY OF THE PROPOSED AND COMPETING METHODS. WE SHOW THE SPECTRAL EFFICIENCY AND ALSO THE APPROXIMATE ASYMPTOTIC COMPUTATIONAL COMPLEXITY, KEEPING ONLY THE DOMINANT COMPONENTS OF THE CALCULATIONS. THE ASTERISK DENOTES THAT THE COMPLEXITY IS PER ITERATION.

Method	Constraints	Complexity	Spectral efficiency
HD–AM*	$L_{\mathrm{t}} = N_{\mathrm{s}}$	medium $O(N_{\mathrm{s}}^2 N_{\mathrm{t}})$	high
HD–FUM	$\mathbf{F}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{RF}}$ are dictionary based	low $O(N_{\mathrm{t}}(\log N_{\mathrm{t}}) + N_{\mathrm{s}} L_{\mathrm{t}})$	medium
HD–CVXR*	none	very high $O(N_{\mathrm{t}}^3 N_{\mathrm{s}})$	high
HD–LSR*	none	high $O(N_{\mathrm{s}} N_{\mathrm{t}}(L_{\mathrm{t}} + N_{\mathrm{s}}))$	high
GHP [24]	none	medium $O(N_{\mathrm{s}} L_{\mathrm{t}} N_{\mathrm{t}})$	high
Sparse precoding [18]	$\mathbf{F}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{RF}}$ are dictionary based	high $O(N_{\mathrm{s}} L_{\mathrm{t}} N_{\mathrm{t}}^2)$	medium

40 seconds using the CVX interior point solver[1] which has polynomial complexity. These running times are averaged over the 500 realizations, implemented in Mathworks' Matlab©. In Table I we summarize the characteristics of the proposed methods as compared to previous methods from the literature. The estimate of the complexity included in Table I refers to the complexity associated to compute the precoder or the combiner for a given number of transmit or receive antennas and RF chains. The complexity of computing both precoder and combiner doubles the number of operations, but the order of complexity is the same. Regarding performance, as previously explained, HD–FUM performs best in terms of computational complexity, while HD–LSR provides the highest spectral efficiency in all the simulation settings. When applicable, HD–AM is always preferred, since it performs very close to HD–LSR with a lower complexity.

VI. CONCLUSIONS

In this paper we proposed four different algorithms for single user hybrid precoding and combining, using full channel state information. Our general approach was to iteratively solve an optimization problem that acts as an approximate solution to the hard problem of hybrid design. In each case we are able to exploit mathematical structures to provide low complexity solutions that do not require, in general, knowledge of the transmit/receive array architectures and that outperform previously proposed methods from the literature. HD–AM is the best solution when the dimensions of the system allow to use it (the number of RF chains equals the number of streams), since it provides a high spectral efficiency at a medium computational complexity. HD–FUM exhibits the lowest complexity, but the spectral efficiency that provides is lower than that of the other proposed algorithms; it is recommended when reducing complexity is the key specification in the mmWave system. HD-CVXR and HD-LSR provide the highest spectral efficiency because no assumption is made on the structure of the precoder, but their complexity is very high and high respectively. The final choice of a preferred algorithm will depend on the targeted implementation complexity, power consumption, and other system features.

[1] www.cvxr.com

REFERENCES

[1] T. S. Rappaport, R. W. Heath Jr., R. C. Daniels, and J. Murdock, *Millimeter Wave Wireless Communications*. Prentice-Hall, September 2014.

[2] S. Yong and C. Chong, "An overview of multigigabit wireless through millimeter wave technology: Potentials and technical challenges," *EURASIP Journal on Wireless Communications and Networking*, 2007.

[3] P. B. Papazian, G. A. Hufford, R. J. Achatz, and R. Hoffman, "Study of the local multipoint distribution service radio channel," *IEEE Trans. Broadcast.*, vol. 43, no. 2, pp. 175–184, 1997.

[4] "WirelessHD Specification Overview," Tech. Rep., 2009. [Online]. Available: http://www.wirelesshd.org

[5] "ISO/IEC/IEEE International Standard for Information technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements-part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band (adoption of IEEE Std 802.11ad-2012)," *ISO/IEC/IEEE 8802-11:2012/Amd.3:2014(E)*, pp. 1–634, March 2014.

[6] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, 2011.

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

[8] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[9] T. Bai and R. W. Heath Jr., "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb 2015.

[10] R. W. Heath, N. Gonzlez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems," *IEEE J. Sel.Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, April 2016.

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

[12] J. Nsenga, A. Bourdoux, and F. Horlin, "Mixed analog/digital beamforming for 60 GHz MIMO frequency selective channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2010, pp. 1–6.

[13] F. Gholam, J. Via, and I. Santamaria, "Beamforming design for simplified analog antenna combining architectures," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2373–2378, 2011.

[14] Z. Pi, "Optimal transmitter beamforming with per-antenna power constraints," in *IEEE International Conference on Communications*, 2012, pp. 3779–3784.

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

[16] V. Venkateswaran and A. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4131–4143, 2010.

[17] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, and K. Cheun, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, 2014.

[18] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, March 2014.

[19] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

[20] K. Engan, S. O. Aase, and J. H. Husøy, "Method of optimal directions for frame design," in *Proc. IEEE ICASSP*, 1999, pp. 2443–2446.

[21] M. Kim and Y. H. Lee, "MSE-based hybrid RF/baseband processing for millimeter-wave communication systems in MIMO interference channels," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, 2015.

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

[23] E. Zhang and C. Huang, "On achieving optimal rate of digital precoder by RF-baseband codesign for MIMO systems," in *Vehicular Technology Conference*, 2014, pp. 1–5.

[24] R. Méndez-Rial, C. Rusu, N. González-Prelcic, and R. W. Heath, Jr., "Dictionary-free hybrid precoders and combiners for mmWave MIMO systems," in *IEEE Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2015.

[25] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale MIMO systems," in *Proc. of the IEEE International Conf. on Acoustics, Speech and Signal Processing*, April 2015.

[26] C.-E. Chen, "An iterative hybrid transceiver design algorithm for millimeter wave MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 2162–2337, June 2015.

[27] R. Lopez-Valcarce, N. Gonzlez-Prelcic, C. Rusu, and R. Heath Jr., "Hybrid precoders and combiners for mmWave MIMO systems with per-antenna power constraints," in *Proc. GLOBECOM*, 2016.

[28] E. Torkildson, U. Madhow, and M. Rodwell, "Indoor millimeter wave MIMO: Feasibility and performance," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4150–4160, December 2011.

[29] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, July 2013.

[30] A. Alkhateeb, O. E. 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. 4553, pp. 831–846, 2014.

[31] C. Rusu, R. Méndez-Rial, N. González-Prelcic, and R. W. Heath Jr, "Low complexity hybrid sparse precoding and combining in millimeter wave MIMO systems," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Jun. 2015.

[32] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.

[33] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2007.

[34] 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.

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

[36] P. Smulders and L. Correia, "Characterisation of propagation in 60 GHz radio channels," *Electronics Communication Engineering Journal*, vol. 9, no. 2, pp. 73–80, 1997.

[37] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60 GHz indoor channel," *IEEE J. Sel. Areas. Communications*, vol. 20, no. 3, pp. 620 – 630, April 2002.

[38] V. Raghavan and A. M. Sayeed, "Sublinear capacity scaling laws for sparse MIMO channels," *IEEE Trans. Inform. Th.*, pp. 345 – 364, Jan. 2011.

[39] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.

[40] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation, vol. 1*. Prentice Hall, 2000.

[41] M. E. Lopes, "Estimating unknown sparsity in compressed sensing," *Journal Machine Learning Research*, vol. 28, pp. 217–225, 2013.

[42] J. A. Tropp, I. S. Dhillon, R. W. Heath Jr., and T. Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 188–209, 2005.

[43] P. Schonemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[44] O. El Ayach, R. W. Heath Jr., S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec 2013, pp. 3476–3480.

[45] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231–2242, 2004.

[46] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[47] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath Jr., "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *Proc. Asilomar*, 2014.

[48] C. Rusu, N. G. Prelcic, and R. W. Heath Jr., "Low resolution adaptive compressed sensing for mmWave MIMO receivers," in *Proc. Asilomar*, 2015.

[49] C. Wei, M. R. D. Rodrigues, and I. J. Wassell, "On the use of unit-norm tight frames to improve the average MSE performance in compressive sensing applications"," *IEEE Signal Processing Letters*, vol. 19, pp. 8–11, 2012.

[50] M. Costa, A. Richter, and V. Koivunen, "Unified Array Manifold Decomposition Based on Spherical Harmonics and 2-D Fourier Basis," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4634–4645, Sept 2010.

[51] W. W. Hager, "Updating the inverse of a matrix," *SIAM Review*, vol. 31, no. 2, pp. 221–239, 1989.

[52] A. Bjorck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.