Low Complexity Hybrid Sparse Precoding and Combining in Millimeter Wave MIMO Systems

Cristian Rusu†, Roi Méndez-Rial†, Nuria González-Prelcic†, and Robert W. Heath Jr.‡
† Universidade de Vigo, Email: {crusu,roimr,nuria}@gts.uvigo.es
‡ The University of Texas at Austin, Email: rheath@utexas.edu

***Abstract*—Millimeter wave (mmWave) multiple-input multiple-output (MIMO) communication with large antenna arrays has been proposed to enable gigabit per second communication for next generation cellular systems and local area networks. A key difference relative to lower frequency solutions is that in mmWave systems, precoding/combining can not be performed entirely at digital baseband, due to the high cost and power consumption of some components of the radio frequency (RF) chain. In this paper we develop a low complexity algorithm for finding hybrid precoders that split the precoding/combining process between the analog and digital domains. Our approach exploits sparsity in the received signal to formulate the design of the precoder/combiners as a compressed sensing optimization problem. We use the properties of the matrix containing the array response vectors to find first an orthonormal analog precoder, since sparse approximation algorithms applied to orthonormal sensing matrices are based on simple computations of correlations. Then, we propose to perform a local search to refine the analog precoder and compute the baseband precoder. We present numerical results demonstrate substantial improvements in complexity while maintaining good spectral efficiency.**

I. Introduction

Millimeter wave (mmWave) is the new spectral frontier for next generation cellular networks and wireless local area networks [1], [2], [3], [4]. An important requirement in mmWave systems is the use of large arrays at the transmitter and receiver to provide a reasonable link budget. The antennas form a multiple-input multiple-output (MIMO) communication link that can be configured for different objectives. The de facto approach is spatial directivity, which provides beamforming gain needed to achieve a reasonable signal-to-noise ratio (SNR) at the receiver. MmWave channels though also have the ability to support spatial multiplexing of multiple data streams due to scattering and polarization [5], [6], [7], [8]. Unfortunately, power and cost requirements in the mmWave analog front-end make it challenging to implement the typical MIMO precoding transceiver found in lower frequency systems, which is implemented in entirely in baseband. A solution is the hybrid precoding framework, where the precoding/combining process is divided between analog and digital domains [9], [10], [11].

This work was partially funded by the Spanish Government and the European Regional Development Fund (ERDF) under projects TACTICA and COMPASS (TEC2013-47020-C2-1-R) by the Galician Regional Government and AtlantTIC.

This material is based upon work supported in part by the National Science Foundation under Grant No. NSF-CCF-1319556.

A popular design of hybrid precoders for mmWave channels based on variable phase shifters was proposed in [9] for a particular mmWave system model incorporating: i) the constraints on the analog precoder/combiner, ii) presence of large antenna arrays, and iii) the limited scattering nature of the mmWave channel. The design of the precoders and combiners is formulated as a sparsity seeking optimization problem with hardware constraints. It resembles the problem of sparse signal recovery via multiple measurement vectors (MMV), also known as the simultaneous sparse recovery problem (S–OMP) [12]. The approach in [9] is elegant yet solving for the precoders still results in high complexity. A limitation of the work in [9], is that perfect channel state information is assumed at the receiver. This has been overcome in work on adaptive channel estimation [10], where the mmWave channel estimation problem is formulated as a compressed sensing problem, so that the channel parameters are estimated using standard CS tools. Training beamforming and combining vectors during the channel estimation phase are designed using a multi-resolution codebook. The main limitation of this work is that it assumes known array geometries for both the transmitter and receiver. Further investigation is also needed to obtain lower complexity solutions to both the channel estimation and the hybrid analog/digital precoding design problems. Hybrid precoding structures based on the use of variable phase shifters have been proposed earlier for general MIMO architectures in [13], but do not take into account the characteristics of millimeter wave propagation or leverage sparsity of the received signal. A related concept called beamspace MIMO communication has been proposed in [14], which uses a high-resolution discrete lens array for analog spatial beamforming. This avoids the need for phase shifters but does not have uniform performance across a broad range of angles.

In this paper we propose a low-complexity solution to the hybrid precoding optimization problem posed in [9]. We take into account the full structure of the optimization problem by exploiting the semi-unitary optimum precoder (optimum in the absence of hardware constraints). This structure reduces significantly the search space in the array manifold and thus leads to a lower complexity procedure versus that found in [9]. The reduction in complexity is due to an orthogonal matching step that fits the optimum precoder with the closest semi-unitary structure in the array manifold that emulates its behavior. The orthogonal matching step eliminates the need for the, slow, greedy matching pursuit steps deployed in the previous

approach [9]. This step is then followed by a local search that further improves the solution by using either a fast one-by-one selection procedure or a full matching pursuit search but both only on a reduced section of the array manifold, around the semi-unitary solution previously found. Numerical results show that the computational advantage comes with no significant performance degradation in the proposed method as compared to previous results.

II. Problem Formulation

A. System model

Consider the mmWave system shown in Figure 1. The transmitter sends N_{s} data streams using N_{t} antennas to the receiver, which has N_{r} antennas. The transmitter has $N_{\mathrm{t}}^{\mathrm{RF}}$ RF transmit chains. Due to the high cost and power consumption associated with providing each antenna with an RF chain and a digital-to-analog converter (DAC) capable of handling the high frequencies and bandwidths of mmWave systems, $N_{\mathrm{s}} \leq N_{\mathrm{t}}^{\mathrm{RF}} \leq N_{\mathrm{t}}$. In the hybrid precoding approach, the transmitter applies two precoders: the $N_{\mathrm{t}}^{\mathrm{RF}} \times N_{\mathrm{s}}$ digital baseband precoder $\boldsymbol{F}_{\mathrm{BB}}$ and the $N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{RF}}$ analog precoder $\boldsymbol{F}_{\mathrm{RF}}$. The digital precoder $\boldsymbol{F}_{\mathrm{BB}}$ is designed assuming infinite precision while the analog precoder $\boldsymbol{F}_{\mathrm{RF}}$ is assumed to have elements of equal norm assuming only phase shifting is performed in the analog domain. The total power constraint is enforced by normalizing $\boldsymbol{F}_{\mathrm{BB}}$ such that $\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\|_F^2 = N_{\mathrm{s}}$. While the analog precoder $\boldsymbol{F}_{\mathrm{RF}}$ may be applied at some intermediate frequency or at the RF frequency, we represent it using its complex baseband equivalent.

Assuming narrowband operation as in [9], the transmitted signal is $\boldsymbol{x} = \boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\boldsymbol{s}$, where $\boldsymbol{s}$ represents the symbol vector. In this paper, the input symbol vector is normalized such that $\mathbb{E}[\boldsymbol{s}\boldsymbol{s}^H] = \frac{1}{N_{\mathrm{s}}}\boldsymbol{I}_{N_{\mathrm{s}}}$. Using ρ to denote the average received power, $\boldsymbol{H}$ the $N_{\mathrm{r}} \times N_{\mathrm{t}}$ channel matrix, and $\boldsymbol{n}$ a vector with IID $\mathcal{CN}(0, \sigma^2)$ entries, the received complex baseband signal of dimension $N_{\mathrm{r}} \times 1$ is

$$\boldsymbol{y} = \sqrt{\rho}\boldsymbol{H}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\boldsymbol{s} + \boldsymbol{n}. \tag{1}$$

MmWave channels are expected to have limited scattering with only a few scattering clusters. To incorporate this fact, we use a clustered channel model with N_{cl} scattering clusters, each of which contribute N_{ray} propagation paths. The clustered model is widely used as a MIMO channel model [15], [16], [17] and is also used in mmWave [18], [2]. With a clustered model, the channel matrix is

$$\boldsymbol{H} = \sqrt{\frac{N_{\mathrm{t}}N_{\mathrm{r}}}{N_{\mathrm{cl}}N_{\mathrm{ray}}}} \sum_{i=1}^{N_{\mathrm{cl}}} \sum_{\ell=1}^{N_{\mathrm{ray}}} \alpha_{i,\ell}\boldsymbol{a}_{\mathrm{r}}(\phi_{i,\ell}^{\mathrm{r}})\boldsymbol{a}_{\mathrm{t}}(\phi_{i,\ell}^{\mathrm{t}})^H. \tag{2}$$

where $\alpha_{i,\ell}$ is the complex gain of the l^{th} ray in the i^{th} cluster, whereas $\boldsymbol{a}_{\mathrm{t}}(\phi_{i,\ell}^{\mathrm{t}})$ and $\boldsymbol{a}_{\mathrm{r}}(\phi_{i,\ell}^{\mathrm{r}})$ are the antenna array response vectors at the transmitter and receiver evaluated at the l^{th} path i^{th} cluster azimuth angles of departure or arrival (we assume uniform linear arrays, whose responses do not depend on the elevation angle).

The receiver applies the $N_{\mathrm{r}} \times N_{\mathrm{r}}^{\mathrm{RF}}$ analog combining matrix $\boldsymbol{W}_{\mathrm{RF}}$ and the $N_{\mathrm{r}}^{\mathrm{RF}} \times N_{\mathrm{s}}$ baseband combining matrix $\boldsymbol{W}_{\mathrm{BB}}$.

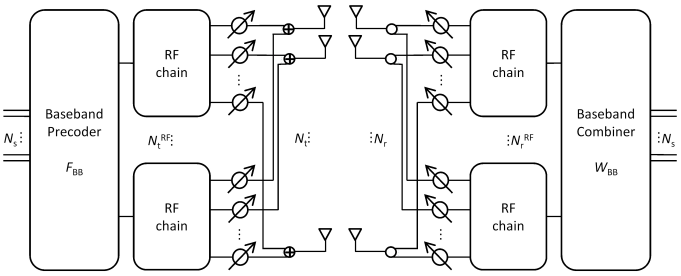

Fig. 1: Block diagram of a mmWave single user system with hybrid precoding: baseband precoding and radio frequency precoding with RF phase shifters. Dimensions follow $N_{\mathrm{s}} \leq N_{\mathrm{t}}^{\mathrm{RF}} \leq N_{\mathrm{t}}$ and $N_{\mathrm{s}} \leq N_{\mathrm{r}}^{\mathrm{RF}} \leq N_{\mathrm{r}}$.

The dimensions satisfy $N_{\mathrm{s}} \leq N_{\mathrm{r}}^{\mathrm{RF}} \leq N_{\mathrm{r}}$ to use a limited number of RF chains and a low dimensional digital combiner following the analog-to-digital converter (ADC). The post-processed received signal after the hybrid combining structure is

$$\tilde{\boldsymbol{y}} = \sqrt{\rho}\boldsymbol{W}_{\mathrm{BB}}^H\boldsymbol{W}_{\mathrm{RF}}^H\boldsymbol{H}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\boldsymbol{s} + \boldsymbol{W}_{\mathrm{BB}}^H\boldsymbol{W}_{\mathrm{RF}}^H\boldsymbol{n}. \tag{3}$$

B. The precoder and combiner design problem

There are many potential metrics to be used in the design of the hybrid precoding and combing matrices. In this paper, we are interested in designs that maximize the achievable sum rate with low computational complexity.

In the conventional precoding paradigm, a single optimal RF precoder $\boldsymbol{F}_{\mathrm{opt}}$ and combiner $\boldsymbol{W}_{\mathrm{opt}}$ would be applied leading to

$$\tilde{\boldsymbol{y}} = \sqrt{\rho}\boldsymbol{W}_{\mathrm{opt}}^H\boldsymbol{H}\boldsymbol{F}_{\mathrm{opt}}\boldsymbol{s} + \boldsymbol{W}_{\mathrm{opt}}^H\boldsymbol{n}. \tag{4}$$

The mutual information maximizing solution (assuming Gaussian signaling) is given by the usual water filling strategy, where $\boldsymbol{F}_{\mathrm{opt}}$ consists of weighted columns that correspond to the right singular values of $\boldsymbol{H}$. The design of the combining matrix in this framework is flexible: it takes the form of the left singular vectors of $\boldsymbol{H}$ multiplied by any nonsingular matrix on the left [13].

Maximizing the spectral efficiency in the hybrid precoding case requires maximizing

$$R = \log_2 \left| \boldsymbol{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_s}\boldsymbol{R}_n^{-1}\boldsymbol{W}_{\mathrm{BB}}^H\boldsymbol{W}_{\mathrm{RF}}^H\boldsymbol{H}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}} \right. \\ \left. \boldsymbol{F}_{\mathrm{BB}}^H\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{H}^H\boldsymbol{W}_{\mathrm{RF}}\boldsymbol{W}_{\mathrm{BB}} \right|,$$

over all precoding and combining matrices, being $\boldsymbol{R}_n = \sigma_n^2\boldsymbol{W}_{\mathrm{BB}}^H\boldsymbol{W}_{\mathrm{RF}}^H\boldsymbol{W}_{\mathrm{RF}}\boldsymbol{W}_{\mathrm{BB}}$ the noise covariance matrix after combining. Since maximizing the mutual information involves a joint optimization of four matrix variables with non-convex constraints for $\boldsymbol{F}_{\mathrm{RF}}$ and $\boldsymbol{W}_{\mathrm{RF}}$, finding global optima is very difficult.

The solution for the design of the analog and digital precoders and combiners in [9] simplifies the joint transmitter-receiver optimization problem by decoupling the design into the transmitter and the receiver, solving similar optimization problems. The objective is to obtain an equally good solution in terms of spectral efficiency but with lower computational complexity. Because we propose a low complexity solution

to the algorithm in [9], we summarize the key steps of the algorithm here.

a) Transmitter: The goal is the design of $\boldsymbol{F}_{\text{RF}}\boldsymbol{F}_{\text{BB}}$ to maximize the mutual information achieved with Gaussian signaling over the mmWave channel:

$$\log_2\left|\boldsymbol{I}_{N_{\text{s}}}+\frac{\rho}{N_{\text{s}}\sigma_n^2}\boldsymbol{H}\boldsymbol{F}_{\text{RF}}\boldsymbol{F}_{\text{BB}}\boldsymbol{F}_{\text{BB}}^H\boldsymbol{F}_{\text{RF}}^H\boldsymbol{H}^H\right|. \tag{5}$$

In the absence of hardware limitations and considering equal power allocation across streams, the optimum precoder that maximizes (5) is given by: $\boldsymbol{F}_{\text{opt}}=\boldsymbol{V}(:,1:N_{\text{s}})$ where $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ is the singular value decomposition. $\boldsymbol{F}_{\text{opt}}$ is the set of the N_{s} columns in $\boldsymbol{V}$ associated with the highest singular values in $\boldsymbol{\Sigma}$.

It has been shown in [9] that the problem of finding the precoder that maximizes the mutual information with the hardware constraints associated to mmWave architectures can be well approximated by solving

$$\begin{aligned}\underset{\boldsymbol{A}_{\text{t}},\tilde{\boldsymbol{F}}_{\text{BB}}}{\arg\min}\quad & \|\boldsymbol{F}_{\text{opt}}-\boldsymbol{A}_{\text{t}}\tilde{\boldsymbol{F}}_{\text{BB}}\|_F\\ \text{subject to}\quad & \|\text{diag}(\tilde{\boldsymbol{F}}_{\text{BB}}\tilde{\boldsymbol{F}}_{\text{BB}}^H)\|_0=N_{\text{t}}^{\text{RF}}\\ & \|\boldsymbol{A}_{\text{t}}\tilde{\boldsymbol{F}}_{\text{BB}}\|_F^2=N_{\text{s}},\end{aligned} \tag{6}$$

where $\|\bullet\|_0$ is the ℓ_0 pseudo-norm accounting for the number of non-zero elements, $\boldsymbol{A}_{\text{t}}$ of size $N_{\text{t}}\times N_{\text{cl}}N_{\text{ray}}$ is the matrix of array response vectors and $\tilde{\boldsymbol{F}}_{\text{BB}}$ of size $N_{\text{cl}}N_{\text{ray}}\times N_{\text{s}}$ has only N_{t}^{RF} non-zero rows (the sparsity constraint) and their product has an energy constraint. We denote by $\boldsymbol{F}_{\text{BB}}$ the $N_{\text{t}}^{\text{RF}}\times N_{\text{s}}$ matrix $\tilde{\boldsymbol{F}}_{\text{BB}}$ restricted to the rows that are non-zero and with $\boldsymbol{F}_{\text{RF}}$ the subset of columns from $\boldsymbol{A}_{\text{t}}$ that participate in the solution. This problem consists in finding N_{t}^{RF} array response vectors and their optimal baseband combination. The $\boldsymbol{A}_{\text{t}}$ matrix contains the set of feasible RF precoders, i.e., the steering vectors.

b) Receiver: Assuming $\boldsymbol{F}_{\text{RF}}\boldsymbol{F}_{\text{BB}}$ fixed, design $\boldsymbol{W}_{\text{RF}}\boldsymbol{W}_{\text{BB}}$ to minimize the mean squared error (MMSE) between transmitted and processed received signals. The optimum MMSE combiner is well known [19] as

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

The solution proposed in [9] consists of finding hybrid combiners that minimize

$$\begin{aligned}\underset{\boldsymbol{W}_{\text{RF}},\boldsymbol{W}_{\text{BB}}}{\arg\min}\quad & \|\mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^H\right]^{1/2}(\boldsymbol{W}_{\text{MMSE}}-\boldsymbol{W}_{\text{RF}}\boldsymbol{W}_{\text{BB}})\|_F\\ \text{subject to}\quad & \boldsymbol{W}_{\text{RF}}\in\mathcal{W}_{\text{RF}},\end{aligned} \tag{8}$$

where $\mathcal{W}_{\text{RF}}$ is the set of $N_{\text{r}}\times N_{\text{r}}^{\text{RF}}$ matrices with constant-gain phase-only entries.

Note that the optimization problems in (6) and (8) to be solved at the transmitter and receiver, respectively, are similar. The solution proposed in [9] involves a greedy strategy based on the Orthogonal Matching Pursuit (OMP) approach, which can be ultimately seen as a variant of Simultaneous OMP (SOMP) [12], and connected to the problem of sparse representations of multiple-measurement vectors (MMV) [20].

In this paper, we find a method that produces good results when solving these optimization problems (6) and avoids the use of the, slow, greedy steps. High complexity reduction is achieved without performance degradation, as shown in Section IV.

III. Low Complexity Hybrid Precoding/Combining Solutions

A. The proposed method

In this section we explain the proposed low complexity precoding algorithm in detail; the derivation of the combiner is similar and is omitted for brevity. The key idea to reduce computational complexity when solving the optimization problem in (6) is to reduce the process of searching columns of the overcomplete matrix $\boldsymbol{A}_{\text{t}}$ to searching columns of orthonormal matrices, subsets of this overcomplete matrix, such that simple correlations replace the matching pursuit iterations.

Recall that the N_{t}–element steering vectors of an uniform linear array (ULA) take the form

$$\boldsymbol{a}_{\text{ULA}}(\phi)=\frac{1}{\sqrt{N_{\text{t}}}}\begin{bmatrix}1 & e^{jkd\sin\phi} & \dots & e^{jkd(N_{\text{t}}-1)\sin\phi}\end{bmatrix}^T. \tag{9}$$

The dot products, in absolute value, between any two such distinct vectors, assuming the sine terms are uniformly distributed 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:

$$\begin{aligned}|\boldsymbol{a}_{\text{ULA}}(\phi_\ell)^H\boldsymbol{a}_{\text{ULA}}(\phi_i)|&=\frac{1}{N_{\text{t}}}\left|\frac{1-e^{2j\pi\frac{(\ell-i)N_{\text{t}}}{N}}}{1-e^{2j\pi\frac{(\ell-i)}{N}}}\right|\\ &=\frac{1}{N_{\text{t}}}\left|\frac{\sin(\pi(\ell-i)N_{\text{t}}/N)}{\sin(\pi(\ell-i)/N)}\right|.\end{aligned} \tag{10}$$

The parameter N is the angular resolution. In general, the objective is to solve the problem on a grid as fine as possible, i.e. N as large as possible. It is clear that the dot products only depend on the distance between the two steering vectors $(\ell-i)$ –this actually leads to a circulant Hermitian Gram matrix $\boldsymbol{G}=\boldsymbol{A}_{\text{t}}^H\boldsymbol{A}_{\text{t}}$. Because of this, when $\alpha=N/N_{\text{t}}$ is an integer it is possible to construct α orthonormal matrices of size $N_{\text{t}}\times N_{\text{t}}$ by choosing columns of $\boldsymbol{A}_{\text{t}}$ equally spaced with distance N_{t} (modulo wrapping). In this way $\boldsymbol{A}_{\text{t}}$ is viewed as a concatenation of α orthonormal matrices. We are interested in these orthonormal matrices since $\boldsymbol{F}_{\text{opt}}$ is semi-unitary (since this matrix has orthonormal columns but is not square, i.e., $\boldsymbol{F}_{\text{opt}}^H\boldsymbol{F}_{\text{opt}}=\boldsymbol{I}_{N_{\text{s}}}$ but $\boldsymbol{F}_{\text{opt}}\boldsymbol{F}_{\text{opt}}^H\neq\boldsymbol{I}_{N_{\text{t}}}$). Therefore, its approximation needs to obey $(\boldsymbol{F}_{\text{RF}}\boldsymbol{F}_{\text{BB}})^H(\boldsymbol{F}_{\text{RF}}\boldsymbol{F}_{\text{BB}})\approx\boldsymbol{I}_{N_{\text{s}}}$ and thus $\boldsymbol{F}_{\text{RF}}^H\boldsymbol{F}_{\text{RF}}$ and $\boldsymbol{F}_{\text{BB}}^H\boldsymbol{F}_{\text{BB}}$ need to closely approximate identity matrices. By this argument we search for $\boldsymbol{F}_{\text{RF}}$ among the α orthonormal matrices from $\boldsymbol{A}_{\text{t}}$. We move to solve (6) with an additional constraint that the solution $\boldsymbol{F}_{\text{RF}}$ needs to be an orthonormal matrix (but still composed of steering vectors). The proposed two-step process is described next.

The proposed algorithm is divided into two distinct steps. In step A we propose to design an orthonormal RF precoder $\boldsymbol{F}_{\text{RF}}$, without any concern to the baseband coder. This simplifies

Algorithm 1 Hybrid Analog-Digital Design by Orthogonal Matching and Local Search (OM+LS).
Input: the optimal unconstrained precoder $\boldsymbol{F}_{\text{opt}}$, the resolution N, the sparsity target N_{t}^{RF} and the local search length (L)
Output: $\boldsymbol{F}_{\text{RF}}$ and $\boldsymbol{F}_{\text{BB}}$ such that the objective function of (6) is as low as possible under the given constraints.

- **Step A.** Compute an orthonormal $\boldsymbol{F}_{\text{RF}}$:
  - Compute all correlations $\boldsymbol{R} = \boldsymbol{A}_{\text{t}}^H \boldsymbol{F}_{\text{opt}}$.
  - With $\alpha = N/N_{\text{t}}$, for $\ell = 1, \ldots, \alpha$:
    - Build orthogonal matrix $\boldsymbol{Q}_\ell$ by selecting N_{t} columns α–apart (mod N) of $\boldsymbol{A}_{\text{t}}$ starting from index ℓ.
    - Compute correlations $\boldsymbol{r}_\ell = \text{norms}(\boldsymbol{Q}_\ell^H \boldsymbol{F}_{\text{opt}})$, where the function computes the 2–norm for each row of the resulting product. $\boldsymbol{r}_\ell$ is of size N_{t}.
    - Compute the overall energy contribution of $\boldsymbol{Q}_\ell$ to: $c_\ell = \sum_{k=1}^{N_{\text{s}}} |\boldsymbol{r}_\ell|_{[k]}^2$ where $|\boldsymbol{r}_\ell|_{[k]}$ represents the k^{th} largest entry of $\boldsymbol{r}_\ell$, in absolute value.
  - Select $\ell_{\max} = \arg\max_{\ell} c_\ell$.
  - Set $\boldsymbol{F}_{\text{RF}}$ as the N_{t}^{RF} columns of $\boldsymbol{Q}_{\ell_{\max}}$ that have the maximum entries in $\boldsymbol{r}_{\ell_{\max}}$, sorted according to these values in descending order.
- **Step B.** Based on the orthonormal $\boldsymbol{F}_{\text{RF}}$ with support S on the columns of $\boldsymbol{A}_{\text{t}}$ (i.e., $\boldsymbol{F}_{\text{RF}} = (\boldsymbol{A}_{\text{t}})_S$) construct the new pair $(\boldsymbol{F}_{\text{RF}}, \boldsymbol{F}_{\text{BB}})$ based on a local search procedure:
  - For $\ell = 1, \ldots, N_{\text{t}}^{\text{RF}}$:
    - Remove the ℓ^{th} column from the support S, to produce the support S'.
    - Construct the solution $(\boldsymbol{F}_{\text{RF}}, \boldsymbol{F}_{\text{BB}})$ on the support S' : $\boldsymbol{F}_{\text{RF}} = (\boldsymbol{A}_{\text{t}})_{S'}, \boldsymbol{F}_{\text{BB}} = (\boldsymbol{F}_{\text{RF}}^H \boldsymbol{F}_{\text{RF}}) \backslash \boldsymbol{F}_{\text{RF}}^H \boldsymbol{F}_{\text{opt}}$.
    - Compute the normalized residual $\boldsymbol{F}_{\text{res}} = \boldsymbol{F}_{\text{opt}} - \boldsymbol{F}_{\text{RF}} \boldsymbol{F}_{\text{BB}}, \boldsymbol{F}_{\text{res}} = \boldsymbol{F}_{\text{res}} \| \boldsymbol{F}_{\text{res}} \|_F^{-1}$.
    - Add to the support S a new index from the set $S_L = \{S(\ell) - L, \ldots, S(\ell) + L\}$ (modulo N) of columns from $\boldsymbol{A}_{\text{t}}$ that maximally correlates with the current residual $\boldsymbol{F}_{\text{res}}$ – i.e., maximum entry in $\text{norms}((\boldsymbol{A}_{\text{t}})_{S_L}^H \boldsymbol{F}_{\text{res}})$.
  - Compute final solution $(\boldsymbol{F}_{\text{RF}}, \boldsymbol{F}_{\text{BB}})$ on the support S.

greatly the computational complexity needed since sparse approximation algorithms (like OMP) applied to orthonormal sensing matrices are reduced to the computation of the correlations and the selection of N_{t}^{RF} columns of $\boldsymbol{A}_{\text{t}}$ that produce the highest correlations. Since we are not interested in imposing explicitly the orthogonal constraint on $\boldsymbol{F}_{\text{RF}}$, in step B we further reduce the objective function by starting a local search, of length L, around the selected steering vectors from $\boldsymbol{A}_{\text{t}}$ (the support set S) to find a better support for the RF precoder. The baseband precoder is computed this time at every step, similarly to the OMP solution. The whole approach is similar to a block learning mechanism presented in [21].

Notice another immediate approach also follows here. Instead of step B of the proposed method, suppose that we apply the SOMP approach using the frequency dictionary $\boldsymbol{A}_{\text{t}}$ restricted to the full index set $S_J = \bigcup_{j=1}^{N_{\text{t}}^{\text{RF}}} S_j$ of size $(2L+1)N_{\text{t}}^{\text{RF}}$, and not on the full resolution N, as described in step B. The full correlations still need to be computed, but the subsequent steps involve a working dimension $(2L+1)N_{\text{t}}^{\text{RF}} \ll N$. Generally N is large since we want to produce results under a good resolution. In this approach, step A acts like a grid reduction step. We call this approach OM+SOMP, since it differs from the initial approach by replacing step B with an SOMP approach and not a sequential atom update. In terms of complexity OM+LS is simpler than OM+SOMP but in terms of performance we expect the latter to do better.

The same discussion applies at the receiver for the design of the pair $(\boldsymbol{W}_{\text{RF}}, \boldsymbol{W}_{\text{BB}})$ by solving

$$\begin{aligned} &\arg\min_{\boldsymbol{A}_{\text{r}}, \tilde{\boldsymbol{W}}_{\text{BB}}} \|\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^H]^{-1/2} (\boldsymbol{W}_{\text{MMSE}} - \boldsymbol{A}_{\text{r}} \tilde{\boldsymbol{W}}_{\text{BB}})\|_F \\ &\text{subject to } \|\text{diag}(\tilde{\boldsymbol{W}}_{\text{BB}} \tilde{\boldsymbol{W}}_{\text{BB}}^H)\|_0 = N_{\text{r}}^{\text{RF}} \end{aligned} \tag{11}$$

using the same two step approach of orthogonal matching followed by a local search strategy depicted in Algorithm 1 for the new objective function.

B. Computational complexity

Equipped with the two proposed strategies, orthogonal matching to reduce the array manifold space followed by either a fast local search (OM+LS) or a slower full matching pursuit search (OMP+SOMP) in the reduced manifold, we now establish how they improve on the previous solution to problem (6) described in [9]. The computational complexity of the proposed method is dominated by the computation of the correlations $\boldsymbol{R} = \boldsymbol{A}_{\text{t}}^H \boldsymbol{F}_{\text{opt}}$. The other computations of step A only represent fast manipulations of the elements of this matrix –note that all products $\boldsymbol{Q}_\ell^H \boldsymbol{F}_{\text{opt}}$ are contained in $\boldsymbol{R}$. Step B is dominated by correlation computations of subsets of $2L+1$ columns from $\boldsymbol{A}_{\text{t}}$ with the current residual $\boldsymbol{F}_{\text{res}}$. The correlations in step B are computed only for $(2L+1)$ steering vectors (instead of the full N), again N_{t}^{RF} times. The local search parameter should obey $1 \leq L \leq \lceil \alpha/2 \rceil$ and generally takes low values. All correlations of the type $\boldsymbol{F}_{\text{RF}}^H \boldsymbol{F}_{\text{opt}}$ are computed for both algorithms only once when the full correlations $\boldsymbol{A}_{\text{t}}^H \boldsymbol{F}_{\text{opt}}$ are found. Considering that $\alpha < N_{\text{t}}^{\text{RF}} \ll N_{\text{t}} \ll N$, with $N = \alpha N_{\text{t}}$ the two methods (we omit OM+LS for brevity) take approximately

$$\begin{aligned} C_{\text{SOMP}} &\approx 2\alpha N_{\text{t}}^{\text{RF}} N_{\text{t}}^2 N_{\text{s}} (N_{\text{t}} + 1) \\ C_{\text{OM+SOMP}} &\approx 2\alpha N_{\text{s}} N_{\text{t}}^2 + 2\alpha N_{\text{s}} N_{\text{t}} + 2\alpha N_{\text{t}} + \\ &\qquad 2N_{\text{s}} N_{\text{t}}^{\text{RF}} (N_{\text{t}}^{\text{RF}} - 1)(\alpha + 1)(N_{\text{t}} + 1) \end{aligned} \tag{12}$$

number of operations. The speed up comes from the fact that the full correlation matrix in the proposed method is computed only once (in step A) while for the SOMP approach its computation needs to take place N_{t}^{RF} times. Notice that both computationally expensive procedures of step A are included in the first step of the regular SOMP, thus the (N_{t}^{RF}-1) factor for the run of the matching pursuit steps of OM+SOMP.

We now leverage the speed-up provided by OM+SOMP to increase the resolution and produce better solutions. Consider that SOMP runs for a resolution $N_1 = \alpha_1 N_{\text{t}}$ while OM+SOMP runs for a different resolution $N_2 = \alpha_2 N_{\text{t}}$, we would like to know how large can α_2 be such that complexity

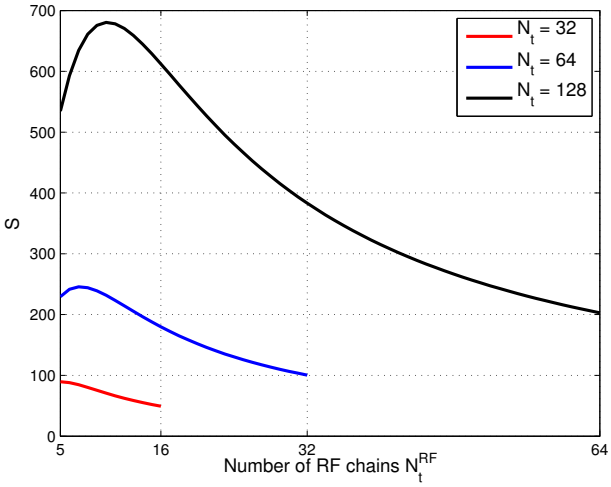

Fig. 2: Evolution of speedup S when OM+SOMP and SOMP are used with the same angular resolution $N = 4N_{\mathrm{t}}$ for fixed $N_s = 4$.

of the proposed method does not surpass that of SOMP. Considering the local search length $L = \alpha_1/d, d \geq 2$, in the case of OM+SOMP we analyze and compare the complexities in (12) to reach:

$$\alpha_2 \leq \left\lfloor \frac{dN_{\mathrm{t}}^{\mathrm{RF}} N_{\mathrm{s}}(N_{\mathrm{t}}+1)(\alpha_1 N_{\mathrm{t}} - N_{\mathrm{t}}^{\mathrm{RF}})}{N_{\mathrm{t}}^{\mathrm{RF}}(N_{\mathrm{t}}^{\mathrm{RF}}-1)N_{\mathrm{s}}(N_{\mathrm{t}}+1) + dN_{\mathrm{t}}(N_{\mathrm{s}}(N_{\mathrm{t}}+1)+1)} \right\rfloor. \quad (13)$$

Keeping the value of α_2 under this bound guarantees that the computational complexity of the proposed will be lower than that of SOMP.

Alternatively, consider that OM+SOMP and SOMP runs with the same angular resolution $N = \alpha N_{\mathrm{t}}$. In this case we show the achieved speed-up ($S = C_{\mathrm{SOMP}}/C_{\mathrm{OM+SOMP}}$). Figure 2 show the evolution of S for various dimensions.

IV. Numerical Results

In this section we provide several Monte Carlo simulation results to illustrate the performance of the hybrid precoder/combiner solution presented in the previous section.

We consider the narrowband clustered channel model in (2) with $N_{\mathrm{cl}} = 6$ clusters and $N_{\mathrm{ray}} = 8$ propagation paths per cluster. For purposes of simulations, the entries of $\boldsymbol{H}$ denoted by $\alpha_{i,\ell}$ are assumed to be IID $\mathcal{CN}(0, \sigma_{\alpha,i}^2)$ where $\sigma_{\alpha,i}^2$ is the average power of the i^{th} cluster. We set $\sigma_{\alpha_{i,\ell}}^2 = \sigma_{\alpha}^2$, all clusters with equal power satisfying the normalization constraint $\mathbb{E}[\|\boldsymbol{H}\|_F^2] = N_{\mathrm{t}}N_{\mathrm{r}}$. The scaling factor in front of the summation in (2) is used to ensure that $\mathbb{E}[\|\boldsymbol{H}\|_F^2] = N_{\mathrm{t}}N_{\mathrm{r}}$. The angles of departure and arrival $\phi_{i,\ell}^{\mathrm{t}}, \phi_{i,\ell}^{\mathrm{r}}$ are normal random distributed with mean cluster angle $\phi_i^{\mathrm{t}}, \phi_i^{\mathrm{r}}$ uniformly random distributed in $[0, 2\pi]$. The angle spread is set to $\sigma_\phi^{\mathrm{t}} = \sigma_\phi^{\mathrm{r}} = 7.5°$. Two ULAs with $N_{\mathrm{t}} = 32/64$ and $N_{\mathrm{r}} = 32/64$ are considered for transmission and reception. The same total power constraint is fixed for all precoders with equal power allocation per stream and the signal to noise ratio is given by $\mathrm{SNR} = \frac{\rho}{\sigma_n^2}$.

Fig. 3 shows the spectral efficiency achieved by the hybrid analog digital precoders SOMP and OM+SOMP, together with

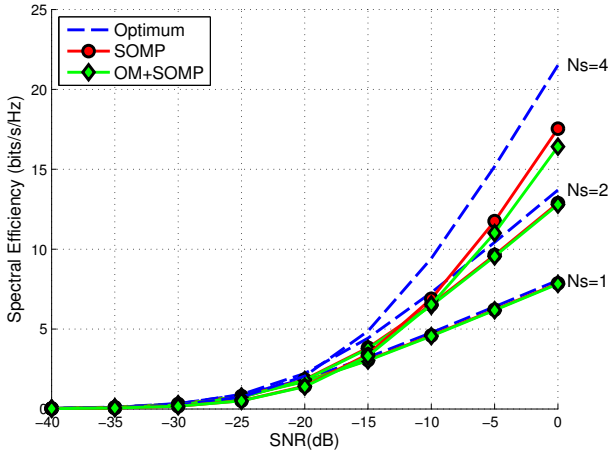

Fig. 3: Spectral efficiency achieved by different precoders for ULA system with 32 transmit/receive antennas. The cluster channel has $N_{\mathrm{cl}} = 6$, $N_{\mathrm{ray}} = 8$ with an angular spread of $7.5°$. $N_{\mathrm{t}}^{\mathrm{RF}} = 4, N_{\mathrm{r}}^{\mathrm{RF}} = 4$ RF chains are considered and $N_{\mathrm{s}} \in \{1, 2, 4\}$ data streams.

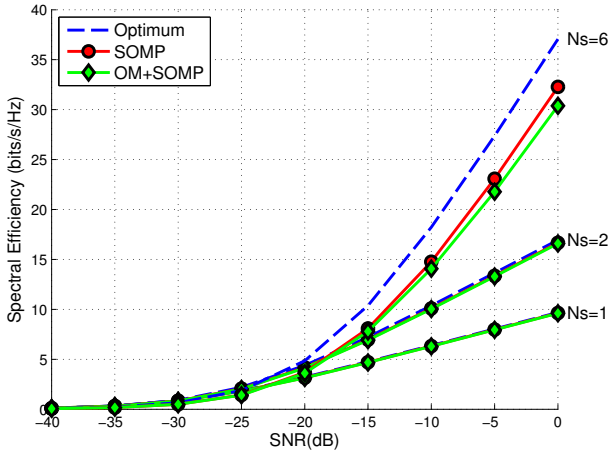

Fig. 4: Spectral efficiency achieved by different precoders for ULA system with 64 transmit/receive antennas. The cluster channel has $N_{\mathrm{cl}} = 6$, $N_{\mathrm{ray}} = 8$ with an angular spread of $7.5°$. $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 6$ RF chains are considered and $N_{\mathrm{s}} \in \{1, 2, 6\}$ data streams.

the the optimum unconstrained solution for different SNR values. We assume a system with $N_{\mathrm{t}} = N_{\mathrm{r}} = 32$ ULAs for the transmitter and receiver, $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 6$ RF chains and $N_{\mathrm{s}} \in \{1, 2, 4\}$ data streams. Fig. 4 plots again the spectral efficiency for a different setup: $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 6$ and $N_{\mathrm{s}} \in \{1, 2, 6\}$. We see that the proposed precoder OM+SOMP achieves spectral efficiencies that are very closed to those achieved by SOMP for all the set ups. For low number of streams OM+SOMP almost perfectly matches SOMP, while the gap between both methods increases for high SNR and high number of data streams. At the same time, the efficiencies obtained by both hybrid precoders are similar to the optimum unconstrained solution for low number of data streams. The differences between the optimum unconstrained solution and the unconstrained ones, however, becomes non-negligible when the number of streams

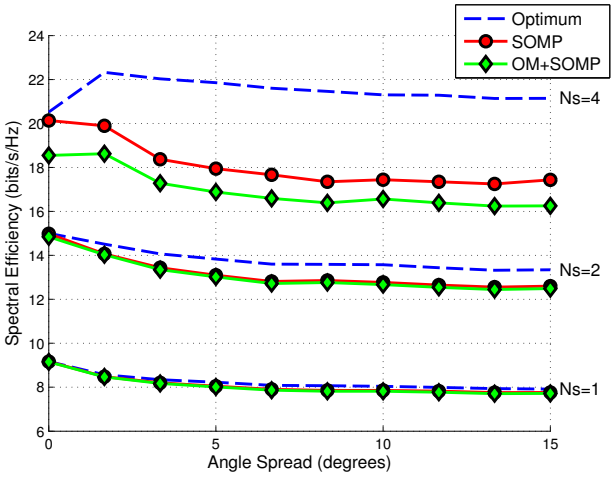

Fig. 5: Spectral efficiency as a function of the Angle Spread. We assume an ULA system with 32 transmit/receive antennas, $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 4$ RF chains and $N_{\mathrm{s}} \in \{1, 2, 4\}$ data streams. The cluster channel has $N_{\mathrm{cl}} = 6$, $N_{\mathrm{ray}} = 8$ and varying angular spread. We have the SNR$= 0$ dB.

equals the number of available RF chains.

Fig. 5 shows the spectral efficiency as a function of the angle spread of the channel scatterers. We assume a system with $N_{\mathrm{t}} = N_{\mathrm{r}} = 32$ antennas, $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 6$ RF chains and $N_{\mathrm{s}} \in \{1, 2, 4\}$ data streams. The signal to noise ratio is fixed to SNR$= 0$ dB. We see that the performance of the hybrid precoders decreases with an increase of the angle spread, while the gap between OM+SOMP and SOMP remains constant.

In all cases, OM+SOMP and SOMP run with the same angular resolution N and thus OM+SOMP achieves a considerable speedup without sacrificing the performance in terms of the spectral efficiency.

V. Conclusions

In this paper we developed a new optimization algorithm for the design of hybrid precoders and combiners for mmWave MIMO systems. Our two solutions incorporate constraints that account for the practical hardware limitations at these frequencies: analog beamforming based on quantized variable phase shifters and the use of a limited number of RF chains. The main innovation in our work is to exploit the array geometry in a way that allows us to reduce the search complexity and thus the overall complexity of the algorithm. Simulation results show that the spectral efficiency achieved by using the new algorithms is comparable to the unconstrained solution, yet with substantially lower overall complexity.

References

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

[2] T. Rappaport, R. W. Heath Jr., R. Daniels, and J. Murdock, *Millimeter wave wireless communications*. Prentice Hall, 2014.

[3] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

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

[5] E. Torkildson, C. Sheldon, U. Madhow, and M. Rodwell, "Millimeter-wave spatial multiplexing in an indoor environment," in *IEEE GLOBECOM Workshops*, December 2009, pp. 1–6.

[6] C. Sheldon, M. Seo, E. Torkildson, M. Rodwell, and U. Madhow, "Four-channel spatial multiplexing over a millimeter-wave line-of-sight link," in *IEEE GLOBECOM Workshops*, June 2009, pp. 389–392.

[7] T. Rappaport, F. Gutierrez, E. Ben-Dor, J. Murdock, Y. Qiao, and J. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 1850–1859, 2013.

[8] S. Sun, T. Rappaport, R. W. Heath Jr., A. Nix, and S. Rangan, "MIMO for millimeter wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Communications Magazine*, December 2014.

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

[10] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath Jr, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, October 2014.

[11] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5g cellular communications: theoretical feasibility and prototype results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, February 2014.

[12] A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing (Special Issue on Sparse Approximations in Signal and Image Processing)*, pp. 572–588, 2006.

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

[14] A. Sayeed and J. Brady, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," in *Proc. of 2013 IEEE Global Telecommunications Conference (GLOBECOM)*, Atlanta, GA, 2013.

[15] J. V. Wallace and M. A. Jensen, "Modeling the indoor MIMO wireless channel," *IEEE Transactions on Antennas and Propagation*, vol. 50, no. 5, pp. 592–599, 2002.

[16] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kysti, D. Laurenson, G. Matz, A. F. Molisch, C. Oestges, and H. Ozcelik, "Survey of channel and radio propagation models for wireless MIMO systems," *EURASIP Journal on Wireless Communications and Networking*, 2007.

[17] A. Forenza, D. J. Love, and R. W. Heath, "Simplified spatial correlation models for clustered MIMO channels with different array configurations," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 4, pp. 1924–1934, 2007.

[18] C. Gustafson, K. Haneda, S. Wyne, and F. Tufvesson, "On mm-Wave multipath clustering and channel modeling," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 3, pp. 1445–1455, 2014.

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

[20] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[21] C. Rusu and B. Dumitrescu, "Block orthonormal overcomplete dictionary learning," in *21st European Signal Processing Conference*, 2013, pp. 1–5.