Spatially Sparse Precoding in Millimeter Wave MIMO Systems

Omar El Ayach, *Member, IEEE,* Sridhar Rajagopal, *Senior Member, IEEE,* Shadi Abu-Surra, *Member, IEEE,* Zhouyue Pi, *Senior Member, IEEE,* and Robert W. Heath, Jr., *Fellow, IEEE*

***Abstract*—Millimeter wave (mmWave) signals experience orders-of-magnitude more pathloss than the microwave signals currently used in most wireless applications and all cellular systems. MmWave systems must therefore leverage large antenna arrays, made possible by the decrease in wavelength, to combat pathloss with beamforming gain. Beamforming with multiple data streams, known as precoding, can be used to further improve mmWave spectral efficiency. Both beamforming and precoding are done digitally at baseband in traditional multi-antenna systems. The high cost and power consumption of mixed-signal devices in mmWave systems, however, make analog processing in the RF domain more attractive. This hardware limitation restricts the feasible set of precoders and combiners that can be applied by practical mmWave transceivers. In this paper, we consider transmit precoding and receiver combining in mmWave systems with large antenna arrays. We exploit the spatial structure of mmWave channels to formulate the precoding/combining problem as a sparse reconstruction problem. Using the principle of basis pursuit, we develop algorithms that accurately approximate optimal unconstrained precoders and combiners such that they can be implemented in low-cost RF hardware. We present numerical results on the performance of the proposed algorithms and show that they allow mmWave systems to approach their unconstrained performance limits, even when transceiver hardware constraints are considered.**

***Index Terms*—Millimeter wave, multiple-input multiple-output (MIMO), antenna arrays, beamforming, precoding, cellular communication, sparsity, sparse reconstruction, basis pursuit, limited feedback.**

I. Introduction

THE capacity of wireless networks has thus far scaled with the increasing data traffic, primarily due to improved area spectral efficiency (bits/s/Hz/m^2) [1]. A number of physical layer enhancements such as multiple antennas, channel coding, and interference coordination, as well as the general trend toward network densification have all been instrumental in achieving this efficiency [1], [2]. Since there seems to be little scope for further gains at the physical layer, and since the widespread deployment of heterogeneous networks is not without challenges [3], these techniques alone may not be sufficient to meet future traffic demands. As a result, increasing the spectrum available for commercial wireless systems, potentially by exploring new less-congested spectrum bands, is a promising solution to increase network capacity.

Millimeter wave (mmWave) communication, for example, has enabled gigabit-per-second data rates in indoor wireless systems [4], [5] and fixed outdoor systems [6]. More recently, advances in mmWave hardware [7] and the potential availability of spectrum has encouraged the wireless industry to consider mmWave for the access link in outdoor cellular systems [8], [9]. A main differentiating factor in mmWave communication is that the ten-fold increase in carrier frequency, compared to the current majority of wireless systems, implies that mmWave signals experience an orders-of-magnitude increase in free-space pathloss. An interesting redeeming feature in mmWave systems, however, is that the decrease in wavelength enables packing a large number of antenna elements into small form factors. Large arrays can provide the beamforming gain needed to overcome pathloss and establish links with reasonable signal-to-noise ratio (SNR). Further, large arrays may enable precoding multiple data streams which could improve spectral efficiency and allow systems to approach capacity [10], [11].

While the fundamentals of precoding are the same regardless of carrier frequency, signal processing in mmWave systems is subject to a set of non-trivial practical constraints. For example, traditional multiple-input multiple-output (MIMO) processing is often performed digitally at baseband, which enables controlling both the signal's phase and amplitude. Digital processing, however, requires dedicated baseband and RF hardware for each antenna element. Unfortunately, the high cost and power consumption of mmWave mixed-signal hardware precludes such a transceiver architecture at present, and forces mmWave systems to rely heavily on analog or RF processing [7], [8]. Analog precoding is often implemented using phase shifters [7], [8], [12] which places constant modulus constraints on the elements of the RF precoder. Several approaches have been considered for precoding in such low-complexity transceivers [13]–[27]. The work in [13]–[15] considers antenna (or antenna subset) selection which has the advantage of replacing phase shifters with even simpler analog switches. Selection, however, provides limited array

Manuscript received May 10, 2013; revised September 20, 2013; accepted November 25, 2013. The associate editor coordinating the review of this paper and approving it for publication was X. Dong.

O. El Ayach is with Qualcomm Technologies Inc., San Diego, CA, 92122 (e-mail: oelayach@qti.qualcomm.com).

S. Rajagopal, S. Abu-Surra, and Z. Pi are with Samsung Research America - Dallas, Richardson, TX, 75082 USA (e-mail: {sasurra, srajagop, zpi}@sta.samsung.com).

R. Heath is with The University of Texas at Austin, Austin, TX 78712 USA (e-mail: rheath@utexas.edu).

This work was done while the first author was with The University of Texas at Austin and with Samsung Research America - Dallas. The authors at The University of Texas at Austin were supported in part by the Army Research Laboratory contract W911NF-10-1-0420 and National Science Foundation grant 1218338 and 1319556. This work has appeared in part at the 2012 IEEE International Communications Conference (ICC).

Digital Object Identifier 10.1109/TWC.2014.011714.130846

1536-1276/14$31.00 © 2014 IEEE

gain and performs poorly in correlated channels such as those experienced in mmWave [16]. To improve performance over correlated channels, the work [17]–[19] considers beam steering solutions in which phase shifters are used to optimally orient an array's response in space, potentially based on statistical channel knowledge. The strategies in [17]–[19], however, are in general suboptimal since beam steering alone cannot perfectly capture the channels dominant eigenmodes. The work in [20]–[25] develops iterative precoding algorithms for systems that leverage analog processing, and [26] further proposes simple analytical solutions. Further hardware limitations have also been considered in [27], for example, which focuses on analog receiver processing with only quantized phase control and finite-precision analog-to-digital converters. The work in [20]–[27], however, is not specialized to mmWave MIMO systems with large antenna arrays. Namely, the work in [20]–[27] does not leverage the structure present in mmWave MIMO channels and adopts models that do not fully capture the effect of limited mmWave scattering and large tightly-packed arrays [28]–[31].

In this paper, we focus on the precoding insight and solutions that can be derived from jointly considering the following three factors: (i) precoding with RF hardware constraints, (ii) the use of large antenna arrays, and (iii) the limited scattering nature of mmWave channels. We consider single-user precoding for a practical transceiver architecture in which a large antenna array is driven by a limited number of transmit/receive chains [8], [10], [11], [32]. In such a system, transmitters have the ability to apply high-dimensional (tall) RF precoders, implemented via analog phase shifters, followed by low-dimensional (small) digital precoders that can be implemented at baseband. We adopt a realistic clustered channel model that captures both the limited scattering at high frequency and the antenna correlation present in mmWave antenna arrays [28]–[31]. We show that the joint treatment of practical hardware architectures and realistic channel models can yield simple precoding solutions with near-optimal spectral efficiency. We note that a similar observation is made by the authors of [33]–[37] in which low-complexity hybrid analog-digital transceivers are constructed by leveraging the concept of beamspace MIMO in which a number of dominant orthogonal propagation paths are selected and are further digitally combined at baseband.

We exploit the sparse-scattering structure of mmWave channels to formulate the design of hybrid RF/baseband precoders as a sparsity constrained matrix reconstruction problem [38]–[43]. Initial results on this precoding approach were presented in [44]. In this paper, we formalize the mmWave precoding problem and show that, instead of directly maximizing mutual information, near-optimal hybrid precoders can be found via an optimization that resembles the problem of sparse signal recovery with multiple measurement vectors, also known as the simultaneously sparse approximation problem [45]–[48]. We thus provide an algorithmic precoding solution based on the concept of orthogonal matching pursuit [39], [41], [49]. The algorithm takes an optimal unconstrained precoder as input and approximates it as linear combination of beam steering vectors that can be applied at RF (and combined digitally at baseband). Further, we extend this sparse precoding approach to receiver-side processing and show that designing hybrid minimum mean-square error (MMSE) combiners can again be cast as a simultaneously sparse approximation problem and solved via basis pursuit [50], [51]. We argue that, in addition to providing practical near-optimal precoders, the proposed framework is particularly amenable for limited feedback operation and is thus not limited to genie-aided systems with perfect transmitter channel knowledge [52]. The generated precoders can be efficiently compressed using simple scalar quantizers (for the arguments of the beam steering vectors) and low-dimensional Grassmannian subspace quantizers (used to quantize the baseband precoder) [52]–[54]. We briefly describe the construction of the limited feedback codebooks required, but defer the analysis of limited feedback performance to future work. Finally, we present simulation results on the performance of the proposed strategy and show that it allows mmWave systems to approach their unconstrained performance limits even when practical transceiver constraints are considered.

We use the following notation throughout this paper: $\mathbf{A}$ is a matrix; $\mathbf{a}$ is a vector; a is a scalar; $\mathbf{A}^{(i)}$ is the i^{th} column of $\mathbf{A}$; $(\cdot)^T$ and $(\cdot)^*$ denote transpose and conjugate transpose respectively; $\|\mathbf{A}\|_F$ is the Frobenius norm of $\mathbf{A}$, $\operatorname{tr}(\mathbf{A})$ is its trace and $|\mathbf{A}|$ is its determinant; $\|\mathbf{a}\|_p$ is the p-norm of $\mathbf{a}$; $[\mathbf{A} \mid \mathbf{B}]$ denotes horizontal concatenation; $\operatorname{diag}(\mathbf{A})$ is a vector formed by the diagonal elements of $\mathbf{A}$; $\mathbf{I}_N$ is the $N \times N$ identity matrix; $\mathbf{0}_{M\times N}$ is the $M \times N$ all-zeros matrix; $\mathcal{CN}(\mathbf{a};\mathbf{A})$ is a complex Gaussian vector with mean $\mathbf{a}$ and covariance matrix $\mathbf{A}$. Expectation is denoted by $\mathbb{E}[\cdot]$ and the real part of a variable is denoted by $\Re\{\cdot\}$.

II. System Model

In this section, we present the mmWave signal and channel model considered in this paper.

A. System Model

Consider the single-user mmWave system shown in Fig. 1 in which a transmitter with N_{t} antennas communicates N_{s} data streams to a receiver with N_{r} antennas [32]. To enable multi-stream communication, the transmitter is equipped with $N_{\mathrm{t}}^{\mathrm{RF}}$ transmit chains such that $N_{\mathrm{s}} \leq N_{\mathrm{t}}^{\mathrm{RF}} \leq N_{\mathrm{t}}$. This hardware architecture enables the transmitter to apply an $N_{\mathrm{t}}^{\mathrm{RF}} \times N_{\mathrm{s}}$ baseband precoder $\mathbf{F}_{\mathrm{BB}}$ using its $N_{\mathrm{t}}^{\mathrm{RF}}$ transmit chains, followed by an $N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{RF}}$ RF precoder $\mathbf{F}_{\mathrm{RF}}$ using analog circuitry. The discrete-time transmitted signal is therefore given by $\mathbf{x} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s}$ where $\mathbf{s}$ is the $N_{\mathrm{s}} \times 1$ symbol vector such that $\mathbb{E}\left[\mathbf{s}\mathbf{s}^*\right] = \frac{1}{N_{\mathrm{s}}}\mathbf{I}_{N_{\mathrm{s}}}$. Since $\mathbf{F}_{\mathrm{RF}}$ is implemented using analog phase shifters, its elements are constrained to satisfy $(\mathbf{F}_{\mathrm{RF}}^{(i)}\mathbf{F}_{\mathrm{RF}}^{(i)*})_{\ell,\ell} = {N_{\mathrm{t}}}^{-1}$, where $(\cdot)_{\ell,\ell}$ denotes the ℓ^{th} diagonal element of a matrix, i.e., all elements of $\mathbf{F}_{\mathrm{RF}}$ have equal norm. The transmitter's total power constraint is enforced by normalizing $\mathbf{F}_{\mathrm{BB}}$ such that $\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2 = N_{\mathrm{s}}$; no other hardware-related constraints are placed on the baseband precoder.

For simplicity, we consider a narrowband block-fading propagation channel as in [10], [25], [32], [34], which yields a received signal

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

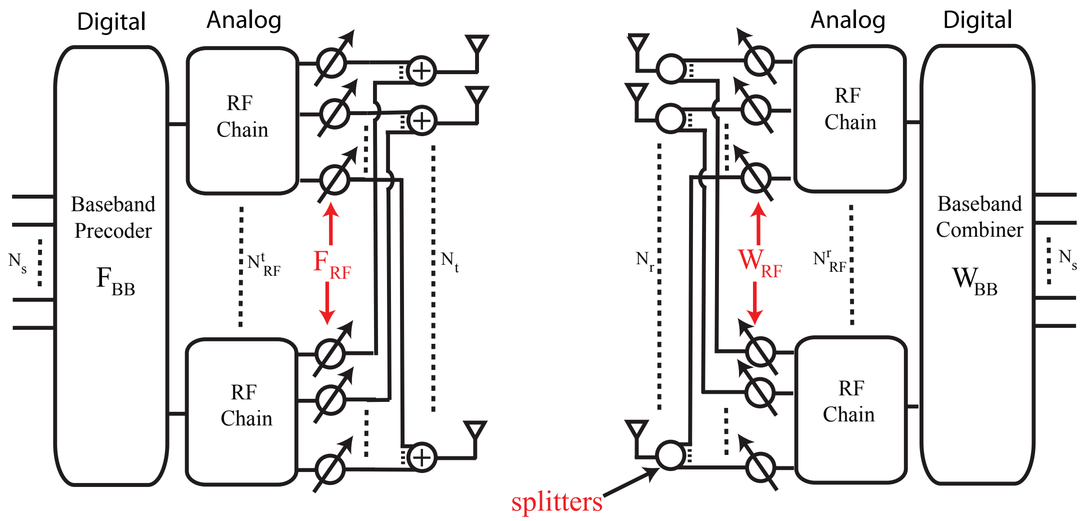


Fig. 1. Simplified hardware block diagram of mmWave single user system with digital baseband precoding followed by constrained radio frequency precoding implemented using RF phase shifters.

where $\mathbf{y}$ is the $N_\mathrm{r} \times 1$ received vector, $\mathbf{H}$ is the $N_\mathrm{r} \times N_\mathrm{t}$ channel matrix such that $\mathbb{E}\left[\|\mathbf{H}\|_F^2\right] = N_\mathrm{t} N_\mathrm{r}$, ρ represents the average received power, and $\mathbf{n}$ is the vector of i.i.d $\mathcal{CN}(0, \sigma_\mathrm{n}^2)$ noise. In writing (1), we implicitly assume perfect timing and frequency recovery. Moreover, to enable precoding, we assume that the channel $\mathbf{H}$ is known perfectly and instantaneously to both the transmitter and receiver. In practical systems, channel state information (CSI) at the receiver can be obtained via training [17], [55]–[57] and subsequently shared with the transmitter via limited feedback [52]; an efficient limited feedback strategy is presented in Section V. Techniques for efficient mmWave channel estimation that potentially leverage the geometric nature of mmWave channels [55]–[57], as well as the rigorous treatment of frequency selectivity, are still an ongoing topic of research.

The receiver uses its $N_\mathrm{s} \le N_\mathrm{r}^\mathrm{RF} \le N_\mathrm{r}$ RF chains and analog phase shifters to obtain the processed received signal

$$\tilde{\mathbf{y}} = \sqrt{\rho}\mathbf{W}_\mathrm{BB}^*\mathbf{W}_\mathrm{RF}^*\mathbf{H}\mathbf{F}_\mathrm{RF}\mathbf{F}_\mathrm{BB}\mathbf{s} + \mathbf{W}_\mathrm{BB}^*\mathbf{W}_\mathrm{RF}^*\mathbf{n}, \qquad (2)$$

where $\mathbf{W}_\mathrm{RF}$ is the $N_\mathrm{r} \times N_\mathrm{r}^\mathrm{RF}$ RF combining matrix and $\mathbf{W}_\mathrm{BB}$ is the $N_\mathrm{r}^\mathrm{RF} \times N_\mathrm{s}$ baseband combining matrix. Similarly to the RF precoder, $\mathbf{W}_\mathrm{RF}$ is implemented using phase shifters and therefore is such that $(\mathbf{W}_\mathrm{RF}^{(i)}\mathbf{W}_\mathrm{RF}^{(i)*})_{\ell,\ell} = N_\mathrm{r}{}^{-1}$. When Gaussian symbols are transmitted over the mmWave channel, the spectral efficiency achieved is given by [58]

$$R = \log_2\left(\left|\mathbf{I}_{N_\mathrm{s}} + \frac{\rho}{N_\mathrm{s}}\mathbf{R}_\mathrm{n}^{-1}\mathbf{W}_\mathrm{BB}^*\mathbf{W}_\mathrm{RF}^*\mathbf{H}\mathbf{F}_\mathrm{RF}\mathbf{F}_\mathrm{BB} \right.\right.$$
$$\left.\left. \times \mathbf{F}_\mathrm{BB}^*\mathbf{F}_\mathrm{RF}^*\mathbf{H}^*\mathbf{W}_\mathrm{RF}\mathbf{W}_\mathrm{BB}\right|\right), \qquad (3)$$

where $\mathbf{R}_\mathrm{n} = \sigma_\mathrm{n}^2\mathbf{W}_\mathrm{BB}^*\mathbf{W}_\mathrm{RF}^*\mathbf{W}_\mathrm{RF}\mathbf{W}_\mathrm{BB}$ is the noise covariance matrix after combining.

B. Channel Model

The high free-space pathloss that is a characteristic of mmWave propagation leads to limited spatial selectivity or scattering. Similarly, the large tightly-packed antenna arrays that are characteristic of mmWave transceivers lead to high levels of antenna correlation. This combination of tightly packed arrays in sparse scattering environments makes many of the statistical fading distributions used in traditional MIMO analysis inaccurate for mmWave channel modeling. For this reason, we adopt a narrowband clustered channel representation, based on the extended Saleh-Valenzuela model, which allows us to accurately capture the mathematical structure present in mmWave channels [28], [29], [31], [59], [60].

Using the clustered channel model, the matrix channel $\mathbf{H}$ is assumed to be a sum of the contributions of N_cl scattering clusters, each of which contribute N_ray propagation paths to the channel matrix $\mathbf{H}$. Therefore, the discrete-time narrowband channel $\mathbf{H}$ can be written as

$$\mathbf{H} = \gamma\sum_{i,\ell}\alpha_{i\ell}\Lambda_\mathrm{r}(\phi_{i\ell}^\mathrm{r}, \theta_{i\ell}^\mathrm{r})\Lambda_\mathrm{t}(\phi_{i\ell}^\mathrm{t}, \theta_{i\ell}^\mathrm{t})\mathbf{a}_\mathrm{r}(\phi_{i\ell}^\mathrm{r}, \theta_{i\ell}^\mathrm{r})\mathbf{a}_\mathrm{t}(\phi_{i\ell}^\mathrm{t}, \theta_{i\ell}^\mathrm{t})^*, \qquad (4)$$

where γ is a normalization factor such that $\gamma = \sqrt{N_\mathrm{t}N_\mathrm{r}/N_\mathrm{cl}N_\mathrm{ray}}$ and $\alpha_{i\ell}$ is the complex gain of the ℓ^th ray in the i^th scattering cluster, whereas $\phi_{i\ell}^\mathrm{r}$ ($\theta_{i\ell}^\mathrm{r}$) and $\phi_{i\ell}^\mathrm{t}$ ($\theta_{i\ell}^\mathrm{t}$) are its azimuth (elevation) angles of arrival and departure respectively. The functions $\Lambda_\mathrm{t}(\phi_{i\ell}^\mathrm{t}, \theta_{i\ell}^\mathrm{t})$ and $\Lambda_\mathrm{r}(\phi_{i\ell}^\mathrm{r}, \theta_{i\ell}^\mathrm{r})$ represent the transmit and receive *antenna element gain* at the corresponding angles of departure and arrival. Finally, the vectors $\mathbf{a}_\mathrm{r}(\phi_\ell^\mathrm{r}, \theta_\ell^\mathrm{r})$ and $\mathbf{a}_\mathrm{t}(\phi_{i\ell}^\mathrm{t}, \theta_{i\ell}^\mathrm{t})$ represent the normalized receive and transmit array response vectors at an azimuth (elevation) angle of $\phi_{i\ell}^\mathrm{r}$ ($\theta_{i\ell}^\mathrm{r}$) and $\phi_{i\ell}^\mathrm{t}$ ($\theta_{i\ell}^\mathrm{t}$) respectively.

In Section VI, we assume that $\alpha_{i\ell}$ are i.i.d. $\mathcal{CN}(0, \sigma_{\alpha,i}^2)$ where $\sigma_{\alpha,i}^2$ represents the average power of the i^th cluster. The average cluster powers are such that $\sum_{i=1}^{N_\mathrm{cl}}\sigma_{\alpha,i}^2 = \gamma$ where γ is a normalization constant that satisfies $\mathbb{E}\left[\|\mathbf{H}\|_F^2\right] = N_\mathrm{t}N_\mathrm{r}$ [29]. The N_ray azimuth and elevation angles of departure, $\phi_{i\ell}^\mathrm{t}$ and $\theta_{i\ell}^\mathrm{t}$, within the cluster i are assumed to be randomly distributed with a uniformly-random mean cluster angle of ϕ_i^t and θ_i^t respectively, and a constant angular spread (standard deviation) of $\sigma_{\phi^{\mathrm{t}}}$ and $\sigma_{\theta^{\mathrm{t}}}$ respectively. The azimuth and elevation angles of arrival, $\phi^{\mathrm{r}}_{i\ell}$ and $\theta^{\mathrm{r}}_{i\ell}$, are again randomly distributed with mean cluster angles of $(\phi^{\mathrm{r}}_{i},\ \theta^{\mathrm{r}}_{i})$ and angular spreads $(\sigma_{\phi^{\mathrm{t}}},\ \sigma_{\theta^{\mathrm{r}}})$. While a variety of distributions have been proposed for the angles of arrival and departure in clustered channel models, the Laplacian distribution has been found to be a good fit for a variety of propagation scenarios [61], and will thus be adopted in the numerical results of Section VI. Similarly, a number of parametrized mathematical models have been proposed for the functions $\Lambda_{\mathrm{t}}(\phi^{\mathrm{t}}_{i\ell},\theta^{\mathrm{t}}_{i\ell})$ and $\Lambda_{\mathrm{r}}(\phi^{\mathrm{r}}_{i\ell},\theta^{\mathrm{r}}_{i\ell})$. For example, if the transmitter's antenna elements are modeled as being ideal sectored elements [62], $\Lambda_{\mathrm{t}}(\phi^{\mathrm{t}}_{i\ell},\theta^{\mathrm{t}}_{i\ell})$ would be given by

$$\Lambda_{\mathrm{t}}(\phi^{\mathrm{t}}_{i\ell},\theta^{\mathrm{t}}_{i\ell}) = \begin{cases} 1 & \forall \phi^{\mathrm{t}}_{i\ell} \in [\phi^{\mathrm{t}}_{\mathrm{min}},\phi^{\mathrm{t}}_{\mathrm{max}}], \forall \theta^{\mathrm{t}}_{i\ell} \in [\theta^{\mathrm{t}}_{\mathrm{min}},\theta^{\mathrm{t}}_{\mathrm{max}}],\\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

where we have assumed unit gain over the sector defined by $\phi^{\mathrm{t}}_{\ell} \in [\phi^{\mathrm{t}}_{\mathrm{min}},\phi^{\mathrm{t}}_{\mathrm{max}}]$ and $\theta^{\mathrm{t}}_{\ell} \in [\theta^{\mathrm{t}}_{\mathrm{min}},\theta^{\mathrm{t}}_{\mathrm{max}}]$ without loss of generality. The receive antenna element gain $\Lambda_{\mathrm{r}}(\phi^{\mathrm{r}}_{i\ell},\theta^{\mathrm{r}}_{i\ell})$ is defined similarly over the azimuth sector $\phi^{\mathrm{r}}_{i\ell} \in [\phi^{\mathrm{r}}_{\mathrm{min}},\phi^{\mathrm{r}}_{\mathrm{max}}]$ and elevation sector $\theta^{\mathrm{r}}_{i\ell} \in [\theta^{\mathrm{r}}_{\mathrm{min}},\theta^{\mathrm{r}}_{\mathrm{max}}]$. Alternatively, instead of considering the simplified model in (5), the functions $\Lambda_{\mathrm{t}}(\phi^{\mathrm{t}}_{i\ell},\theta^{\mathrm{t}}_{i\ell})$ and $\Lambda_{\mathrm{r}}(\phi^{\mathrm{r}}_{i\ell},\theta^{\mathrm{r}}_{i\ell})$ can be replaced by the well-known far field radiation patterns for commonly used antennas such as patch or half-wave dipole antennas [63].

The array response vectors $\mathbf{a}_{\mathrm{t}}(\phi^{\mathrm{t}}_{i\ell},\theta^{\mathrm{t}}_{i\ell})$ and $\mathbf{a}_{\mathrm{r}}(\phi^{\mathrm{r}}_{\ell},\theta^{\mathrm{r}}_{\ell})$ are a function of the transmit and receiver antenna array structure only, and are thus independent of the antenna element properties. While the algorithms and results derived in the remainder of this paper can be applied to arbitrary antenna arrays, we give the following two illustrative examples of commonly-used antenna arrays for completeness. For an N-element uniform linear array (ULA) on the y-axis[1], the array response vector can be written as [63]

$$\mathbf{a}_{\mathrm{ULAy}}(\phi) = \frac{1}{\sqrt{N}}\left[1,\ e^{jkd\sin(\phi)},\ \ldots,\ e^{j(N-1)kd\sin(\phi)}\right]^{T}, \tag{6}$$

where $k = \frac{2\pi}{\lambda}$ and d is the inter-element spacing. Note that we do not include θ in the arguments of $\mathbf{a}_{\mathrm{ULAy}}$ as the array's response is invariant in the elevation domain. In the case of a uniform planar array (UPA) in the yz-plane[2] with W and H elements on the y and z axes respectively, the array response vector is given by [63]

$$\begin{aligned}\mathbf{a}_{\mathrm{UPA}}(\phi,\theta) = \frac{1}{\sqrt{N}}\,[\,1,\ \ldots,\ & e^{jkd(m\sin(\phi)\sin(\theta)+n\cos(\theta))},\\ & \ldots,\ e^{jkd((W-1)\sin(\phi)\sin(\theta)+(H-1)\cos(\theta))}\,]^{T},\end{aligned} \tag{7}$$

where $0 \leq m < W$ and $0 \leq n < H$ are the y and z indices of an antenna element respectively and the antenna array size is $N = WH$. Considering uniform planar arrays is of interest in mmWave beamforming since they (i) yield smaller antenna array dimensions, (ii) facilitate packing more antenna elements in a reasonably-sized array, and (iii) enable beamforming in the elevation domain (also known as 3D beamforming).

III. Spatially Sparse Precoding for the Single User mmWave Channel

We seek to design hybrid mmWave precoders $(\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}})$ that maximize the spectral efficiency expression in (3). Directly maximizing (3), however, requires a joint optimization over the four matrix variables $(\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}},\mathbf{W}_{\mathrm{RF}},\mathbf{W}_{\mathrm{BB}})$. Unfortunately, finding global optima for similar constrained joint optimization problems is often found to be intractable [64], [65]. In the case of mmWave precoding, the non-convex constraints on $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{RF}}$ makes finding an exact solution unlikely. To simplify transceiver design, we temporarily decouple the joint transmitter-receiver optimization problem and focus on the design of the hybrid precoders $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$. Therefore, in lieu of maximizing spectral efficiency, we design $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ to maximize the mutual information achieved by Gaussian signaling over the mmWave channel

$$\mathcal{I}(\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}) = \log_2\left(\left|\mathbf{I}+\frac{\rho}{N_{\mathrm{s}}\sigma^2_{\mathrm{n}}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}^{*}_{\mathrm{BB}}\mathbf{F}^{*}_{\mathrm{RF}}\mathbf{H}^{*}\right|\right). \tag{8}$$

We note here that abstracting receiver operation, and focusing on mutual information instead of the spectral efficiency expression in (3), effectively amounts to assuming that the receiver can perform optimal nearest-neighbor decoding based on the N_{r}-dimensional received signal $\mathbf{y}$. Unfortunately, such a decoder is impossible to realize with practical mmWave systems in which decoders do not have access to the N_{r}-dimensional signal. In practical mmWave systems, received signals must be combined in the analog domain, and possibly in the digital domain, before any detection or decoding is performed. For this reason, we revisit the problem of designing practical mmWave receivers in Section IV.

Proceeding with the design of $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$, the precoder optimization problem can be stated as

$$\begin{aligned}(\mathbf{F}^{\mathrm{opt}}_{\mathrm{RF}},\mathbf{F}^{\mathrm{opt}}_{\mathrm{BB}}) = \underset{\mathbf{F}_{\mathrm{RF}},\ \mathbf{F}_{\mathrm{BB}}}{\arg\max}\ & \mathcal{I}(\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}),\\ \text{s.t.}\quad & \mathbf{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}},\\ & \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|^2_F = N_{\mathrm{s}},\end{aligned} \tag{9}$$

where $\mathcal{F}_{\mathrm{RF}}$ is the set of feasible RF precoders, i.e., the set of $N_{\mathrm{t}} \times N^{\mathrm{RF}}_{\mathrm{t}}$ matrices with constant-magnitude entries. To the extent of the authors' knowledge, no general solutions to (9) are known in the presence of the non-convex feasibility constraint $\mathbf{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}}$. Therefore, we propose to solve an approximation of (9) in order to find practical near-optimal precoders that can be implemented in the system of Fig. 1.

We start by examining the mutual information achieved by the hybrid precoders $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ and rewriting (8) in terms of the "distance" between $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ and the channel's optimal unconstrained precoder $\mathbf{F}_{\mathrm{opt}}$. To do so, define the channel's ordered singular value decomposition (SVD) to be $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{*}$ where $\mathbf{U}$ is an $N_{\mathrm{r}} \times \mathrm{rank}(\mathbf{H})$ unitary matrix, $\boldsymbol{\Sigma}$ is a $\mathrm{rank}(\mathbf{H}) \times \mathrm{rank}(\mathbf{H})$ diagonal matrix of singular values arranged in decreasing order, and $\mathbf{V}$ is a $N_{\mathrm{t}} \times \mathrm{rank}(\mathbf{H})$ unitary

[1] Following the standard notation used in [63], we use the terms x, y, and z-axes to refer to the axes of the standard Cartesian coordinate system defined at the antenna array itself. Similarly, we adopt the standard notation and conventions for the polar coordinate system when dealing with the angles ϕ^{r}, ϕ^{t}, θ^{r}, and θ^{t}).

[2] Note that placing the UPA on the yz-plane (and the earlier UPA along the y-axis) is a completely arbitrary choice that was made to simplify notation. All results hold regardless of array orientation

matrix. Using the SVD of $\mathbf{H}$ and standard mathematical manipulation, (8) can be rewritten as

$$\mathcal{I}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) = \log_2\left(\left|\mathbf{I} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}^2\mathbf{V}^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}\right|\right). \tag{10}$$

Further, defining the following two partitions of the matrices $\mathbf{\Sigma}$ and $\mathbf{V}$ as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{bmatrix}, \qquad \mathbf{V} = [\mathbf{V}_1 \quad \mathbf{V}_2], \tag{11}$$

where $\mathbf{\Sigma}_1$ is of dimension $N_{\mathrm{s}} \times N_{\mathrm{s}}$ and $\mathbf{V}_1$ is of dimension $N_{\mathrm{t}} \times N_{\mathrm{s}}$, we note that the optimal unconstrained unitary precoder for $\mathbf{H}$ is simply given by $\mathbf{F}_{\mathrm{opt}} = \mathbf{V}_1$. Further note that the precoder $\mathbf{V}_1$ cannot in general be expressed as $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ with $\mathbf{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}}$, and thus cannot be realized in the mmWave architecture of interest. If the hybrid precoder $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ can be made sufficiently "close" to the optimal precoder $\mathbf{V}_1$, however, the mutual information resulting from $\mathbf{F}_{\mathrm{opt}}$ and $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ can be made comparable. In fact, to simplify the forthcoming treatment of $\mathcal{I}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$, we make the following system assumption.

Approximation 1: We assume that the mmWave system parameters $(N_{\mathrm{t}}, N_{\mathrm{r}}, N_{\mathrm{t}}^{\mathrm{RF}}, N_{\mathrm{r}}^{\mathrm{RF}})$, as well as the parameters of the mmWave propagation channel $(N_{\mathrm{cl}}, N_{\mathrm{ray}}, \ldots)$, are such that the hybrid precoders $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ can be made sufficiently "close" to the optimal unitary precoder $\mathbf{F}_{\mathrm{opt}} = \mathbf{V}_1$. Mathematically, this "closeness" is defined by the following two equivalent approximations:

1) The eigenvalues of the matrix $\mathbf{I}_{N_{\mathrm{s}}} - \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{BB}}^*\mathbf{V}_1$ are small. In the case of mmWave precoding, this can be equivalently stated as $\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \approx \mathbf{I}_{N_{\mathrm{s}}}$.[3]
2) The singular values of the matrix $\mathbf{V}_2^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ are small; alternatively $\mathbf{V}_2^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \approx \mathbf{0}$.

This approximation is similar to the high-resolution approximation used to simplify the analysis of limited feedback MIMO systems by assuming that codebooks are large enough such that they contain codewords that are sufficiently close to the optimal unquantized precoder [54]. In the case of mmWave precoding, this approximation is expected to be tight in systems of interest which include: (i) a reasonably large number of antennas N_{t}, (ii) a number of transmit chains $N_{\mathrm{s}} < N_{\mathrm{t}}^{\mathrm{RF}} \leq N_{\mathrm{t}}$, and (iii) correlated channel matrices $\mathbf{H}$.

Functionally, Approximation 1 allows us to further simplify the mutual information $\mathcal{I}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}})$. To do so, we use the partitions defined in (11) and further define the following partition of the matrix $\mathbf{V}^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}$ as

$$\begin{aligned}
&\mathbf{V}^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V} \\
&\quad = \begin{bmatrix} \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1, & \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_2 \\ \mathbf{V}_2^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1, & \mathbf{V}_2^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_2 \end{bmatrix} \\
&\quad = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix},
\end{aligned}$$

which allows us to approximate the mutual information achieved by $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ as

$$\begin{aligned}
&\mathcal{I}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) \\
&\quad = \log_2\left(\left|\mathbf{I} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}^2\mathbf{V}^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}\right|\right) \\
&\quad = \log_2\left(\left|\mathbf{I} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\begin{bmatrix} \mathbf{\Sigma}_1^2, & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2^2 \end{bmatrix}\begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix}\right|\right) \\
&\quad \overset{(a)}{=} \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\mathbf{Q}_{11}\right|\right) \\
&\qquad + \log_2\left(\left|\mathbf{I} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_2^2\mathbf{Q}_{22}\right.\right. \\
&\qquad\quad \left.\left. - \frac{\rho^2}{{N_{\mathrm{s}}}^2\sigma_{\mathrm{n}}^4}\mathbf{\Sigma}_2^2\mathbf{Q}_{21}\left(\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\mathbf{Q}_{11}\right)^{-1}\mathbf{\Sigma}_1^2\mathbf{Q}_{12}\right|\right) \\
&\quad \overset{(b)}{\approx} \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1\right|\right),
\end{aligned} \tag{12}$$

where (a) is a result of using the Schur complement identity for matrix determinants and (b) follows from invoking Approximation 1 which implies that $\mathbf{Q}_{12}$, $\mathbf{Q}_{21}$ and $\mathbf{Q}_{22}$ are approximately zero. Using (12), mutual information can be further simplified by writing

$$\begin{aligned}
&\mathcal{I}(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}) \\
&\overset{(a)}{\approx} \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\right|\right) \\
&\quad + \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} - \left(\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\right)^{-1}\right.\right. \\
&\qquad\qquad \left.\left. \times \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\left(\mathbf{I}_{N_{\mathrm{s}}} - \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1\right)\right|\right) \\
&\overset{(b)}{\approx} \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\right|\right) \\
&\quad - \mathrm{tr}\left(\left(\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\right)^{-1}\right. \\
&\qquad\qquad \left. \times \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\left(\mathbf{I}_{N_{\mathrm{s}}} - \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1\right)\right) \\
&\overset{(c)}{\approx} \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\right|\right) \qquad (13) \\
&\quad - \mathrm{tr}\left(\mathbf{I}_{N_{\mathrm{s}}} - \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1\right) \\
&= \log_2\left(\left|\mathbf{I}_{N_{\mathrm{s}}} + \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2\right|\right) - \left(N_{\mathrm{s}} - \|\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2\right),
\end{aligned} \tag{14}$$

where (a) is exact given (12) and can be obtained by defining the matrices $\mathbf{B} = \frac{\rho}{N_{\mathrm{s}}\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_1^2$ and $\mathbf{A} = \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{RF}}^*\mathbf{V}_1$

[3] For the eigenvalues of $\mathbf{I}_{N_{\mathrm{s}}} - \mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{BB}}^*\mathbf{V}_1$ to be small, we need $\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \approx \mathbf{\Psi}$ where $\mathbf{\Psi}$ is *any* $N_{\mathrm{s}} \times N_{\mathrm{s}}$ *unitary matrix* (not necessarily $\mathbf{I}_{N_{\mathrm{s}}}$). However, if $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ is a valid precoder with $\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \approx \mathbf{\Psi}$, then so is the rotated precoder $\mathbf{F}_{\mathrm{RF}}\widetilde{\mathbf{F}}_{\mathrm{BB}} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{\Psi}^*$ for which we have $\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\widetilde{\mathbf{F}}_{\mathrm{BB}} \approx \mathbf{I}_{N_{\mathrm{s}}}$. Since $\mathbf{F}_{\mathrm{BB}}$ can be arbitrarily rotated, the conditions $\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{BB}}^*\mathbf{V}_1 \approx \mathbf{I}_{N_{\mathrm{s}}}$ and $\mathbf{V}_1^*\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \approx \mathbf{I}_{N_{\mathrm{s}}}$ can be considered equivalent in this case without loss of generality.

and noting that $\mathbf{I}+\mathbf{BA} = (\mathbf{I}+\mathbf{B})(\mathbf{I}-(\mathbf{I}+\mathbf{B})^{-1}\mathbf{B}(\mathbf{I}-\mathbf{A}))$. The simplification in (b) follows from Approximation 1 which implies that the eigenvalues of the matrix $\mathbf{X} = (\mathbf{I}_{N_{\rm s}} + \frac{\rho}{N_{\rm s}\sigma_{\rm n}^2}\mathbf{\Sigma}_1^2)^{-1}\frac{\rho}{N_{\rm s}\sigma_{\rm n}^2}\mathbf{\Sigma}_1^2\left(\mathbf{I}_{N_{\rm s}} - \mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\mathbf{F}_{\rm BB}^*\mathbf{F}_{\rm RF}^*\mathbf{V}_1\right)$ are small and thus allows us to use the following approximation $\log_2|\mathbf{I}_{N_{\rm s}} - \mathbf{X}| \approx \log_2(1-{\rm tr}(\mathbf{X})) \approx -{\rm tr}(\mathbf{X})$. Finally (c) follows from adopting a high *effective-SNR* approximation which implies that $(\mathbf{I}+\frac{\rho}{N_{\rm s}\sigma_{\rm n}^2}\mathbf{\Sigma}_1^2)^{-1}\frac{\rho}{N_{\rm s}\sigma_{\rm n}^2}\mathbf{\Sigma}_1^2 \approx \mathbf{I}_{N_{\rm s}}$ and yields the final result in (14).[4] We notice that the first term in (14) is the mutual information achieved by the optimal precoder $\mathbf{F}_{\rm opt} = \mathbf{V}_1$ and that the dependence of $\mathcal{I}(\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB})$ on the hybrid precoder $\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$ is now captured in the second and final term of (13) and (14).

In cases where $\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$ is made exactly unitary, we note that the second term in (13) and (14) is nothing but the squared chordal distance between the two points $\mathbf{F}_{\rm opt} = \mathbf{V}_1$ and $\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$ on the Grassmann manifold. Since Approximation 1 states the these two points are "close", we can exploit the manifold's locally Euclidean property to replace the chordal distance by the Euclidean distance $\|\mathbf{F}_{\rm opt} - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F$ [66]. Therefore, near-optimal hybrid precoders that approximately maximize $\mathcal{I}(\mathbf{F}_{\rm RF}, \mathbf{F}_{\rm BB})$ can be found by instead minimizing $\|\mathbf{F}_{\rm opt} - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F$. In fact, even without treating $\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$ as a point on the Grassmann manifold, Approximation 1 implies that $\|\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2$, and consequently (14), can be approximately maximized by instead maximizing ${\rm tr}\left(\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\right)$.[5] Since maximizing ${\rm tr}\left(\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\right)$ is again equivalent to minimizing $\|\mathbf{F}_{\rm opt} - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F$, the precoder design problem can be rewritten as

$$\begin{aligned}(\mathbf{F}_{\rm RF}^{\rm opt}, \mathbf{F}_{\rm BB}^{\rm opt}) = \mathop{\arg\min}_{\mathbf{F}_{\rm BB}, \mathbf{F}_{\rm RF}} & \|\mathbf{F}_{\rm opt} - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F, \\ \text{s.t.} \quad & \mathbf{F}_{\rm RF} \in \mathcal{F}_{\rm RF}, \\ & \|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2 = N_{\rm s},\end{aligned} \tag{15}$$

which can now be summarized as finding the projection of $\mathbf{F}_{\rm opt}$ onto the set of hybrid precoders of the form $\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$ with $\mathbf{F}_{\rm RF} \in \mathcal{F}_{\rm RF}$. Further, this projection is defined with respect to the standard Frobenius norm $\|\cdot\|_F^2$. Unfortunately, the complex non-convex nature of the feasible set $\mathcal{F}_{\rm RF}$ makes finding such a projection both analytically (in closed form) and algorithmically intractable [69]–[72].

To provide near-optimal solutions to the problem in (15), we propose to exploit the structure of the mmWave MIMO channels generated by the clustered channel model in Section II-B. Namely, we leverage the following observations on mmWave precoding:

1) *Structure of optimal precoder*: Recall that the optimal unitary precoder is $\mathbf{F}_{\rm opt} = \mathbf{V}_1$, and that the columns of the unitary matrix $\mathbf{V}$ form an orthonormal basis for the channel's row space.

2) *Structure of clustered mmWave channels*: Examining the channel model in (4), we note that the array response vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t}), \forall i, \ell,\ \theta_{i\ell}^{\rm t})$ also form a finite spanning set for the channel's row space. In fact, when $N_{\rm cl}N_{\rm ray} \leq N_{\rm t}$, we note that the array response vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t})$ will be linearly independent with probability one and will thus form another *minimal basis* for the channel's row space when $N_{\rm cl}N_{\rm ray} \leq \min(N_{\rm t}, N_{\rm r})$.
Note: To establish the linear independence of the vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t})$, consider the case of uniform linear arrays. When ULAs are considered, the $N_{\rm t} \times N_{\rm cl}N_{\rm ray}$ matrix formed by the collection of vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t})\ \forall i, \ell$ will be a Vandermonde matrix which has full rank whenever the angles $\phi_{i\ell}^{\rm t}$ are distinct. This event occurs with probability one when $\phi_{i\ell}^{\rm t}$ are generated from a continuous distribution. Linear independence can be established in the case of UPAs by writing their response vectors as a Kronecker product of two ULA response vectors [19].

3) *Connection between* $\mathbf{F}_{\rm opt}$ *and* $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t}\ \ \theta_{i\ell}^{\rm t})$: Regardless of whether $N_{\rm cl}N_{\rm ray} \leq N_{\rm t}$ or not, observation 1 implies that the columns of the optimal precoder $\mathbf{F}_{\rm opt} = \mathbf{V}_1$ are related to the vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t})$ through a linear transformation. As a result, the columns of $\mathbf{F}_{\rm opt}$ can be written as linear combinations of $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t}), \forall i, \ell$.

4) *Vectors* $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t}\ \ \theta_{i\ell}^{\rm t})$ *as columns of* $\mathbf{F}_{\rm RF}$: Recall that the vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t})$ are constant-magnitude phase-only vectors which can be applied at RF using analog phase shifters. Therefore, the mmWave transmitter can apply $N_{\rm t}^{\rm RF}$ of the vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t},\ \theta_{i\ell}^{\rm t})$ at RF (via the RF precoder $\mathbf{F}_{\rm RF}$), and form arbitrary linear combinations of them using its digital precoder $\mathbf{F}_{\rm BB}$. Namely, it can construct the linear combination that minimizes $\|\mathbf{F}_{\rm opt} - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F$.

Therefore, by exploiting the structure of $\mathbf{H}$, we notice that near-optimal hybrid precoders can be found by further restricting $\mathcal{F}_{\rm RF}$ to be the set of vectors of the form $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t}, \theta_{i\ell}^{\rm t})$ and solving

$$\begin{aligned}(\mathbf{F}_{\rm RF}^{\rm opt}, \mathbf{F}_{\rm BB}^{\rm opt}) = \arg\min & \|\mathbf{F}_{\rm opt} - \mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F, \\ \text{s.t.} \quad & \mathbf{F}_{\rm RF}^{(i)} \in \left\{\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t}, \theta_{i\ell}^{\rm t}),\ \forall i, \ell\right\}, \\ & \|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2 = N_{\rm s},\end{aligned} \tag{16}$$

which amounts to finding the best low dimensional representation of $\mathbf{F}_{\rm opt}$ using the basis vectors $\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t}, \theta_{i,\ell}^{\rm t})$. We note here that the set of basis vectors can be extended to include array response vectors $\mathbf{a}_{\rm t}(\cdot,\cdot)$ in directions other than $\{(\phi_{i\ell}^{\rm t}, \theta_{i\ell}^{\rm t})|\ 1 \leq i \leq N_{\rm cl},\ 1 \leq \ell \leq N_{\rm ray}\}$, though the effect of this basis extension is typically negligible. Similarly, in cases where $N_{\rm cl}N_{\rm ray} > N_{\rm t}$, and $\{\mathbf{a}_{\rm t}(\phi_{i\ell}^{\rm t}, \theta_{i\ell}^{\rm t})|\ 1 \leq i \leq N_{\rm cl},\ 1 \leq \ell \leq N_{\rm ray}\}$ forms and overcomplete representation of the channel's right singular space, it is possible to reduce the redundancy in the spanning set for example by using the orthogonal steering vectors leveraged in [33]–[37] for which observations 2-4 also hold. In any case, the precoding problem consists of selecting the "best" $N_{\rm t}^{\rm RF}$ array response vectors and finding their optimal baseband combination. Finally, we note that the constraint of $\mathbf{F}_{\rm RF}^{(i)}$ can be embedded directly into the optimization objective to obtain

[4] Note here that it is not the nominal SNR $\frac{\rho}{N_{\rm s}\sigma_{\rm n}^2}$ that is assumed to be high. This would be a problematic assumption in mmWave systems. It is, however, only the *effective-SNRs* in the channel's dominant $N_{\rm s}$ subspaces that are assumed to be sufficiently high. This is a reasonable assumption since these effective SNRs include the large array gain from mmWave beamforming.

[5] This is since the magnitude of $\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$'s off-diagonal entries is negligible and all $\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$'s diagonals must be made close to one. Thus $\|\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\|_F^2$, i.e., the $\ell 2$ norm of $\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}$'s diagonals, can be maximized by optimizing ${\rm tr}\left(\mathbf{V}_1^*\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}\right)$, i.e., the $\ell 1$ norm of the diagonals [43], [67], [68].

Algorithm 1 Spatially Sparse Precoding via Orthogonal Matching Pursuit

```
Require: $\mathbf{F}_{\mathrm{opt}}$
1:  $\mathbf{F}_{\mathrm{RF}} = \text{Empty Matrix}$
2:  $\mathbf{F}_{\mathrm{res}} = \mathbf{F}_{\mathrm{opt}}$
3:  for $i \le N_{\mathrm{t}}^{\mathrm{RF}}$ do
4:      $\mathbf{\Psi} = \mathbf{A}_t^* \mathbf{F}_{\mathrm{res}}$
5:      $k = \arg\max_{\ell=1,\ \ldots,\ N_{\mathrm{cl}}N_{\mathrm{ray}}} (\mathbf{\Psi}\mathbf{\Psi}^*)_{\ell,\ell}$
6:      $\mathbf{F}_{\mathrm{RF}} = \left[\mathbf{F}_{\mathrm{RF}} | \mathbf{A}_t^{(k)}\right]$
7:      $\mathbf{F}_{\mathrm{BB}} = \left(\mathbf{F}_{\mathrm{RF}}^* \mathbf{F}_{\mathrm{RF}}\right)^{-1} \mathbf{F}_{\mathrm{RF}}^* \mathbf{F}_{\mathrm{opt}}$
8:      $\mathbf{F}_{\mathrm{res}} = \frac{\mathbf{F}_{\mathrm{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}}{\|\mathbf{F}_{\mathrm{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F}$
9:  end for
10: $\mathbf{F}_{\mathrm{BB}} = \sqrt{N_{\mathrm{s}}} \frac{\mathbf{F}_{\mathrm{BB}}}{\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F}$
11: return $\mathbf{F}_{\mathrm{RF}}$, $\mathbf{F}_{\mathrm{BB}}$
```

the following equivalent problem

$$\begin{aligned} \widetilde{\mathbf{F}}_{BB}^{\mathrm{opt}} &= \arg\min_{\widetilde{\mathbf{F}}_{\mathrm{BB}}} \|\mathbf{F}_{\mathrm{opt}} - \mathbf{A}_{\mathrm{t}}\widetilde{\mathbf{F}}_{\mathrm{BB}}\|_F, \\ &\quad \text{s.t.} \quad \|\mathrm{diag}(\widetilde{\mathbf{F}}_{\mathrm{BB}}\widetilde{\mathbf{F}}_{\mathrm{BB}}^*)\|_0 = N_{\mathrm{t}}^{\mathrm{RF}}, \\ &\quad\quad\quad \|\mathbf{A}_{\mathrm{t}}\widetilde{\mathbf{F}}_{\mathrm{BB}}\|_F^2 = N_{\mathrm{s}}, \end{aligned} \tag{17}$$

where $\mathbf{A}_{\mathrm{t}} = \left[\mathbf{a}_{\mathrm{t}}(\phi_{1,1}^{\mathrm{t}}, \theta_{1,1}^{\mathrm{t}}),\ \ldots,\ \mathbf{a}_{\mathrm{t}}(\phi_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{t}}, \theta_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{t}})\right]$ is an $N_{\mathrm{t}} \times N_{\mathrm{cl}}N_{\mathrm{ray}}$ matrix of array response vectors and $\widetilde{\mathbf{F}}_{\mathrm{BB}}$ is an $N_{\mathrm{cl}}N_{\mathrm{ray}} \times N_{\mathrm{s}}$ matrix. The matrices $\mathbf{A}_{\mathrm{t}}$ and $\widetilde{\mathbf{F}}_{\mathrm{BB}}$ act as auxiliary variables from which we obtain $\mathbf{F}_{\mathrm{RF}}^{\mathrm{opt}}$ and $\mathbf{F}_{\mathrm{BB}}^{\mathrm{opt}}$ respectively. Namely, the sparsity constraint $\|\mathrm{diag}(\widetilde{\mathbf{F}}_{\mathrm{BB}}\widetilde{\mathbf{F}}_{\mathrm{BB}}^*)\|_0 = N_{\mathrm{t}}^{\mathrm{RF}}$ states that $\widetilde{\mathbf{F}}_{\mathrm{BB}}$ cannot have more than $N_{\mathrm{t}}^{\mathrm{RF}}$ non-zero rows. When only $N_{\mathrm{t}}^{\mathrm{RF}}$ rows of $\widetilde{\mathbf{F}}_{\mathrm{BB}}$ are non zero, only $N_{\mathrm{t}}^{\mathrm{RF}}$ columns of the matrix $\mathbf{A}_{\mathrm{t}}$ are effectively "selected". As a result, the baseband precoder $\mathbf{F}_{\mathrm{BB}}^{\mathrm{opt}}$ will be given by the $N_{\mathrm{t}}^{\mathrm{RF}}$ non-zero rows of $\widetilde{\mathbf{F}}_{\mathrm{BB}}^{\mathrm{opt}}$ and the RF precoder $\mathbf{F}_{\mathrm{RF}}^{\mathrm{opt}}$ will be given by the corresponding $N_{\mathrm{t}}^{\mathrm{RF}}$ columns of $\mathbf{A}_{\mathrm{t}}$.

Essentially, we have reformulated the problem of jointly designing $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$ into a sparsity constrained matrix reconstruction problem with one variable. Although the underlying motivation differs, and so does the interpretation of the different variables involved in (17), the resulting problem formulation is identical to the optimization problem encountered in the literature on sparse signal recovery. Thus, the extensive literature on sparse reconstruction can now be used for hybrid precoder design [39], [41]. To see this more clearly, note that in the simplest case of single stream beamforming, (17) simplifies to

$$\begin{aligned} \widetilde{\mathbf{f}}_{BB}^{\mathrm{opt}} &= \arg\min_{\widetilde{\mathbf{f}}_{\mathrm{BB}}} \|\mathbf{f}_{\mathrm{opt}} - \mathbf{A}_t\widetilde{\mathbf{f}}_{\mathrm{BB}}\|_F, \\ &\quad \text{s.t.} \quad \|\widetilde{\mathbf{f}}_{\mathrm{BB}}\|_0 = N_{\mathrm{t}}^{\mathrm{RF}}, \quad \|\mathbf{A}_t\widetilde{\mathbf{f}}_{\mathrm{BB}}\|_F^2 = N_{\mathrm{s}}, \end{aligned} \tag{18}$$

in which the sparsity constraint is now on the vector $\widetilde{\mathbf{f}}_{\mathrm{BB}}$. This beamforming problem can be solved, for example, by relaxing the sparsity constraint and using convex optimization to solve its $\ell2 - \ell1$ relaxation. Alternatively, (18) can be solved using tools from [39]–[42], [49].

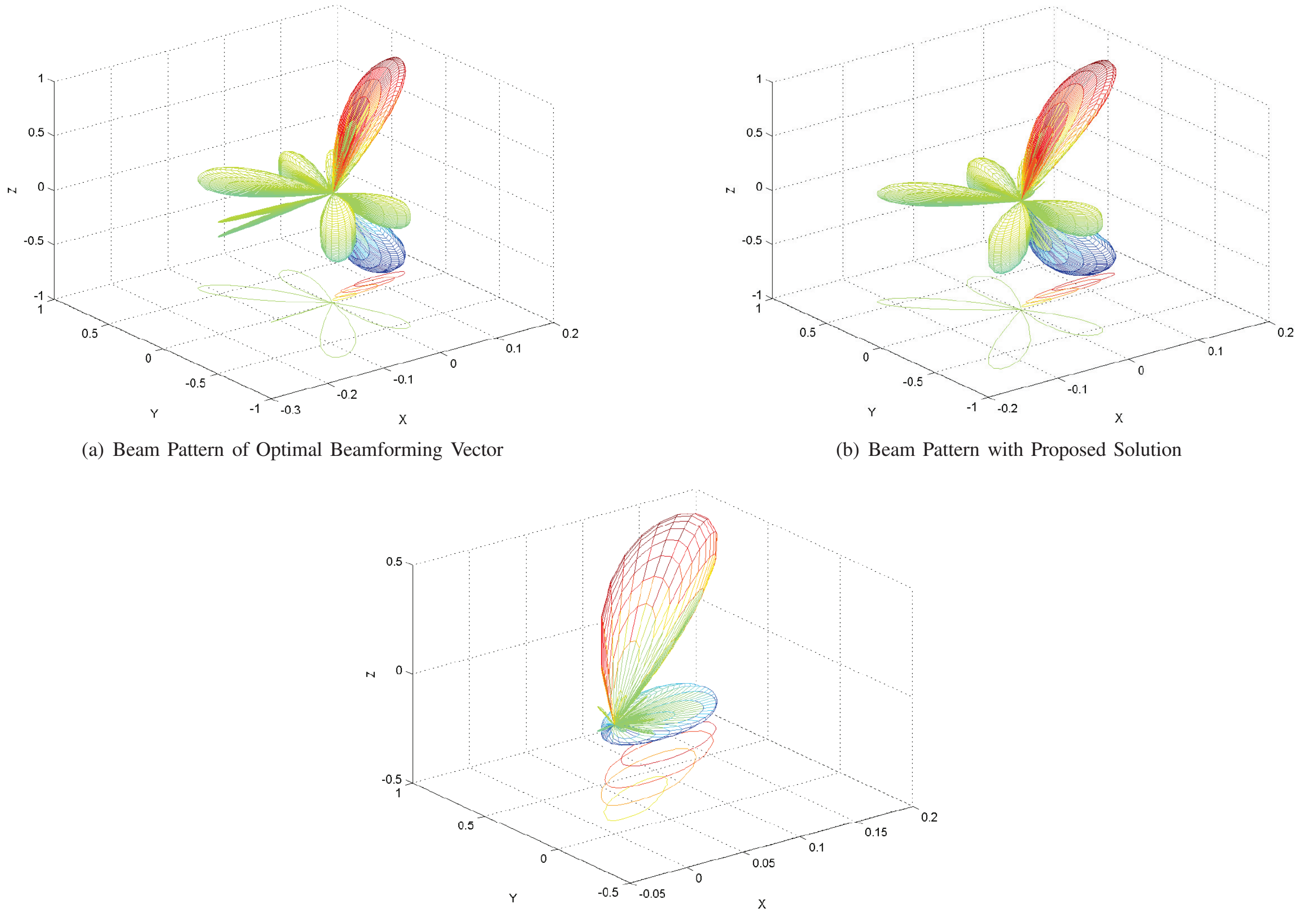
In the more general case of $N_{\mathrm{s}} > 1$, the problem in (17) is equivalent to the problem of sparse signal recovery with multiple measurement vectors, also known as the simultaneously sparse approximation problem [45]–[48]. So, for the general case of $N_{\mathrm{s}} \geq 1$, we present an algorithmic solution based on the well-known concept of orthogonal matching pursuit [39], [41], [49]. The pseudo-code for the precoder solution is given in Algorithm 1. In summary, the precoding algorithm starts by finding the vector $\mathbf{a}_{\mathrm{t}}(\phi_{i\ell}^{\mathrm{t}}, \theta_{i\ell}^{\mathrm{t}})$ along which the optimal precoder has the maximum projection. It then appends the selected column vector $\mathbf{a}_{\mathrm{t}}(\phi_{i\ell}^{\mathrm{t}}, \theta_{i\ell}^{\mathrm{t}})$ to the RF precoder $\mathbf{F}_{\mathrm{RF}}$. After the dominant vector is found, and the least squares solution to $\mathbf{F}_{\mathrm{BB}}$ is calculated in step 7, the contribution of the selected vector is removed in step 8 and the algorithm proceeds to find the column along which the "residual precoding matrix" $\mathbf{F}_{\mathrm{res}}$ has the largest projection. The process continues until all $N_{\mathrm{t}}^{\mathrm{RF}}$ beamforming vectors have been selected. At the end of the $N_{\mathrm{t}}^{\mathrm{RF}}$ iterations, the algorithm would have (i) constructed an $N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{RF}}$ RF precoding matrix $\mathbf{F}_{\mathrm{RF}}$, and (ii) found the optimal $N_{\mathrm{t}}^{\mathrm{RF}} \times N_{\mathrm{s}}$ baseband precoder $\mathbf{F}_{\mathrm{BB}}$ which minimizes $\|\mathbf{F}_{\mathrm{opt}} - \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2$. Step 10 ensures that the transmit power constraint is exactly satisfied.

To gain more intuition about the proposed precoding framework, Fig. 2 plots the beam patterns generated by a transmitter with a 256-element planar array for an example channel realization with 6 rays (or equivalently 6 clusters with an angular spread of 0) using (i) the channel's optimal unconstrained precoder, (ii) the proposed precoding strategy with $N_{\mathrm{t}}^{\mathrm{RF}} = 4$, and (iii) the beam steering vector in the channel's dominant physical direction. We observe that in practical mmWave channels, optimal precoders do in fact generate spatially sparse beam patterns and thus may be accurately approximated by a finite combination of array response vectors. Further, Fig. 2 indicates that Algorithm 1 succeeds in generating beam patterns which closely resemble those generated by $\mathbf{F}_{\mathrm{opt}}$. Therefore, Algorithm 1 succeeds in selecting the best $N_{\mathrm{t}}^{\mathrm{RF}}$ steering directions and forming appropriate linear combinations of the selected response vectors. This beam pattern similarity will ultimately result in favorable spectral efficiency performance as shown in Section VI.

Having presented the proposed precoding framework, we conclude this section with the following design remarks.

Remark 2: We note that the mmWave terminals need not know the exact angles $(\phi_{i\ell}^{\mathrm{t}}, \theta_{i\ell}^{\mathrm{t}})$ that make up the channel matrix $\mathbf{H}$, and need not use the matrix $\mathbf{A}_{\mathrm{t}}$ as defined earlier. We have only used this finite basis for simplicity of exposition. In general, the mmWave terminals can instead select basis vectors of the form $\mathbf{a}_{\mathrm{t}}(\phi, \theta)$ using any finite set of representative azimuth and elevation directions (such as a set of equally spaced angles for example). This approach avoids having to decompose $\mathbf{H}$ into its geometric representation and is naturally suited for limited feedback operation. This approach will be discussed further in Section V.

Remark 3: It may be advantageous in some cases to impose the additional constraint that $\mathbf{F}_{\mathrm{BB}}$ be unitary. Unitary precoders can be more efficiently quantized and are thus more attractive in limited feedback systems. With this additional constraint, (17) can be solved again via Algorithm 1 by replacing the least squares solution for $\mathbf{F}_{\mathrm{BB}}$ in step 7, by the solution to the corresponding orthogonal Procrustes problem [73]. This is given by $\mathbf{F}_{\mathrm{BB}} = \hat{\mathbf{U}}\hat{\mathbf{V}}^*$ where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are unitary matrices defined by the singular value decomposition of $\mathbf{F}_{\mathrm{RF}}^*\mathbf{F}_{\mathrm{opt}}$, i.e.,



(a) Beam Pattern of Optimal Beamforming Vector

(b) Beam Pattern with Proposed Solution

(c) Beam Pattern of Optimal Steering Vector

Fig. 2. Beam pattern generated a 256-element square array in an example channel realization with 6 rays (or equivalently 6 clusters with 0 angular spread) using (a) optimal unconstrained beamforming, (b) the proposed sparse precoding solution with 4 RF chains, and (c) the beam steering vector in the channel's dominant physical direction. The proposed algorithm is shown to result in beam patterns that closely resemble the patterns generated by optimal beamforming; this beam pattern similarity will ultimately result in similar spectral efficiency. For illustration purposes, the channel's angle spread is set to 0° in this figure.

$\mathbf{F}_{\mathrm{RF}}^{*}\mathbf{F}_{\mathrm{opt}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^{*}$ [73].

Remark 4: In the limit of large antenna arrays ($N_{\mathrm{t}}, N_{\mathrm{r}} \rightarrow \infty$) in very poor scattering environments for which $N_{\mathrm{cl}}N_{\mathrm{ray}} = o(\min(N_{\mathrm{t}}, N_{\mathrm{r}}))$, the results of [19] indicate that simple RF-only beam steering becomes optimal, i.e., it becomes optimal to simply transmit each stream along one of the N_{s} most dominant vectors $\mathbf{a}_{\mathrm{t}}(\phi_{i\ell}^{\mathrm{t}},\ \theta_{i\ell}^{\mathrm{t}})$. For arrays of practical sizes, however, Section VI shows that there can be significant gains from more involved precoding strategies such as the one presented in this section.

IV. Practical Millimeter Wave Receiver Design

In Section III, we abstracted receiver-side processing and focused on designing practical mmWave precoders that maximize mutual information. Effectively, we assumed that the mmWave receiver can optimally decode data using its N_{r}-dimensional received signal. Such a decoder can be of prohibitively high complexity in multi-antenna systems, making lower-complexity receivers such as the commonly used linear MMSE receiver more appealing for practical implementation. In fact, in mmWave architectures such as the one shown in Fig. 1, such optimal decoders are *impossible* to realize since received signals *must* be linearly combined in the analog domain before any detection or decoding is performed.

In this section, we address the problem of designing linear combiners for the mmWave receiver in Fig. 1, which uses both analog and digital processing before detection. Assuming the hybrid precoders $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ are fixed, we seek to design hybrid combiners $\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}$ that minimize the mean-squared-error (MSE) between the transmitted and processed received signals. The combiner design problem can therefore be stated as

$$\begin{aligned}(\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}}, \mathbf{W}_{\mathrm{BB}}^{\mathrm{opt}}) = \underset{\mathbf{W}_{\mathrm{RF}},\ \mathbf{W}_{\mathrm{BB}}}{\arg\min}\ & \mathbb{E}\left[||\mathbf{s} - \mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbf{y}||_2^2\right], \\ \text{s.t.}\quad & \mathbf{W}_{\mathrm{RF}} \in \mathcal{W}_{\mathrm{RF}},\end{aligned} \tag{19}$$

where $\mathcal{W}_{\mathrm{RF}}$ is the set of feasible RF combiners, i.e., $\mathcal{W}_{\mathrm{RF}}$ is the set of $N_{\mathrm{r}} \times N_{\mathrm{r}}^{\mathrm{RF}}$ matrices with constant-gain phase-only entries. In the absence of any hardware limitations that restrict the set of feasible linear receivers, the exact solution to (19) is well known [74] to be

$$\begin{aligned}
&\mathbf{W}_{\mathrm{MMSE}}^{*}\\
&=\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{-1}\\
&=\frac{\sqrt{\rho}}{N_{\mathrm{s}}}\mathbf{F}_{\mathrm{BB}}^{*}\mathbf{F}_{\mathrm{RF}}^{*}\mathbf{H}^{*}\left(\frac{\rho}{N_{\mathrm{s}}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^{*}\mathbf{F}_{\mathrm{RF}}^{*}\mathbf{H}^{*}+\sigma_{\mathrm{n}}^{2}\mathbf{I}_{N_{\mathrm{r}}}\right)^{-1}\\
&\overset{(a)}{=}\frac{1}{\sqrt{\rho}}\left(\mathbf{F}_{\mathrm{BB}}^{*}\mathbf{F}_{\mathrm{RF}}^{*}\mathbf{H}^{*}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}+\frac{\sigma_{\mathrm{n}}^{2}N_{\mathrm{s}}}{\rho}\mathbf{I}_{N_{\mathrm{s}}}\right)^{-1}\mathbf{F}_{\mathrm{BB}}^{*}\mathbf{F}_{\mathrm{RF}}^{*}\mathbf{H}^{*},
\end{aligned}\tag{20}$$

where (a) follows from applying the matrix inversion lemma. Just as in the precoding case, however, this optimal unconstrained MMSE combiner $\mathbf{W}_{\mathrm{MMSE}}^{*}$ need not be decomposable into a product of RF and baseband combiners $\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}$ with $\mathbf{W}_{\mathrm{RF}}\in\mathcal{W}_{\mathrm{RF}}$. Therefore $\mathbf{W}_{\mathrm{MMSE}}^{*}$ cannot be realized in the system of Fig. 1. Further, just as in the precoding case, the complex non-convex constraint $\mathbf{W}_{\mathrm{RF}}\in\mathcal{W}_{\mathrm{RF}}$ makes solving (19) analytically impossible and algorithmically non-trivial. To overcome this difficulty, we leverage the methodology used in [50], [51] to find linear MMSE estimators with complex structural constraints.

We start by reformulating the problem in (19) by expanding MSE as follows

$$\begin{aligned}
\mathbb{E}&\left[\|\mathbf{s}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbf{y}\|_{2}^{2}\right]\\
&=\mathbb{E}\left[(\mathbf{s}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbf{y})^{*}(\mathbf{s}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbf{y})\right]\\
&=\mathbb{E}\left[\mathrm{tr}\left((\mathbf{s}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbf{y})\left(\mathbf{s}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbf{y}\right)^{*}\right)\right]\\
&=\mathrm{tr}\left(\mathbb{E}\left[\mathbf{s}\mathbf{s}^{*}\right]\right)-2\Re\left\{\mathrm{tr}\left(\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\right\}\\
&\quad+\mathrm{tr}\left(\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right).
\end{aligned}\tag{21}$$

We now note that since the optimization problem in (19) is over the variables $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}$, we can add any term that is independent of $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}$ to its objective function without changing the outcome of the optimization. Thus, we choose to add the constant term $\mathrm{tr}\left(\mathbf{W}_{\mathrm{MMSE}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{MMSE}}\right)-\mathrm{tr}\left(\mathbb{E}\left[\mathbf{s}\mathbf{s}^{*}\right]\right)$ and minimize the equivalent objective function

$$\begin{aligned}
\mathcal{J}&(\mathbf{W}_{\mathrm{RF}},\mathbf{W}_{\mathrm{BB}})\\
&=\mathrm{tr}\left(\mathbf{W}_{\mathrm{MMSE}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{MMSE}}\right)\\
&\quad-2\Re\left\{\mathrm{tr}\left(\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\right\}\\
&\quad+\mathrm{tr}\left(\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\\
&\overset{(a)}{=}\mathrm{tr}\left(\mathbf{W}_{\mathrm{MMSE}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{MMSE}}\right)\\
&\quad-2\Re\left\{\mathrm{tr}\left(\mathbf{W}_{\mathrm{MMSE}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\right\}\\
&\quad+\mathrm{tr}\left(\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\\
&=\mathrm{tr}\left(\left(\mathbf{W}_{\mathrm{MMSE}}^{*}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\right)\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\right.\\
&\qquad\qquad\left.\times\left(\mathbf{W}_{\mathrm{MMSE}}^{*}-\mathbf{W}_{\mathrm{BB}}^{*}\mathbf{W}_{\mathrm{RF}}^{*}\right)^{*}\right)\\
&=\left\|\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{1/2}\left(\mathbf{W}_{\mathrm{MMSE}}-\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\right\|_{F}^{2},
\end{aligned}\tag{22}$$

where (a) follows from rewriting the second term as $\mathrm{tr}\left(\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)=\mathrm{tr}(\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{-1}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}})$ and using the fact that $\mathbf{W}_{\mathrm{MMSE}}^{*}=\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{-1}$ which implies that $\mathrm{tr}\left(\mathbb{E}\left[\mathbf{s}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)=\mathrm{tr}\left(\mathbf{W}_{\mathrm{MMSE}}^{*}\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)$. As a result of (22), the MMSE estimation problem is equivalent to finding hybrid combiners that solve

$$\begin{aligned}
&\left(\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}},\mathbf{W}_{\mathrm{BB}}^{\mathrm{opt}}\right)=\\
&\qquad\underset{\mathbf{W}_{\mathrm{RF}},\ \mathbf{W}_{\mathrm{BB}}}{\arg\min}\ \left\|\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{\frac{1}{2}}\left(\mathbf{W}_{\mathrm{MMSE}}-\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}\right)\right\|_{F}\\
&\qquad\ \text{s.t.}\quad\mathbf{W}_{\mathrm{RF}}\in\mathcal{W}_{\mathrm{RF}},
\end{aligned}\tag{23}$$

which amounts to finding the projection of the unconstrained MMSE combiner $\mathbf{W}_{\mathrm{MMSE}}$ onto the set of hybrid combiners of the form $\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}$ with $\mathbf{W}_{\mathrm{RF}}\in\mathcal{W}_{\mathrm{RF}}$. Thus, the design of MMSE receivers for the mmWave system of interest closely resembles the design of its hybrid precoders. Unlike in the precoding case however, the projection now is not with respect to the standard norm $\|\cdot\|_{F}^{2}$ and is instead an $\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]$-weighted Frobenius norm. Unfortunately, as in the case of the precoding problem in (15), the non-convex constraint on $\mathbf{W}_{\mathrm{RF}}$ precludes us from practically solving the projection problem in (23). The same observations that allowed us to leverage the structure of mmWave channels to solve the precoding problem in Section III, however, can be translated to the receiver side to solve the combiner problem as well. Namely, because of the structure of clustered mmWave channels, near-optimal receivers can be found by further constraining $\mathbf{W}_{\mathrm{RF}}$ to have columns of the form $\mathbf{a}_{\mathrm{r}}(\phi,\theta)$ and instead solving

$$\begin{aligned}
&\widetilde{\mathbf{W}}_{\mathrm{BB}}^{\mathrm{opt}}=\\
&\quad\underset{\widetilde{\mathbf{W}}_{\mathrm{BB}}}{\arg\min}\quad\left\|\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{1/2}\mathbf{W}_{\mathrm{MMSE}}-\mathbb{E}\left[\mathbf{y}\mathbf{y}^{*}\right]^{1/2}\mathbf{A}_{\mathrm{r}}\widetilde{\mathbf{W}}_{\mathrm{BB}}\right\|_{F},\\
&\quad\ \text{s.t.}\quad\left\|\mathrm{diag}\left(\widetilde{\mathbf{W}}_{\mathrm{BB}}\widetilde{\mathbf{W}}_{\mathrm{BB}}^{*}\right)\right\|_{0}=N_{\mathrm{r}}^{\mathrm{RF}}
\end{aligned}\tag{24}$$

where $\mathbf{A}_{\mathrm{r}}=\left[\mathbf{a}_{\mathrm{r}}(\phi_{1,1}^{\mathrm{r}},\theta_{1,1}^{\mathrm{r}}),\ \ldots,\ \mathbf{a}_{\mathrm{t}}(\phi_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{r}},\theta_{N_{\mathrm{cl}},N_{\mathrm{ray}}}^{\mathrm{r}})\right]$ is an $N_{\mathrm{r}}\times N_{\mathrm{cl}}N_{\mathrm{ray}}$ matrix of array response vectors and $\widetilde{\mathbf{W}}_{\mathrm{BB}}$ is an $N_{\mathrm{cl}}N_{\mathrm{ray}}\times N_{\mathrm{s}}$ matrix; the quantities $\mathbf{A}_{\mathrm{r}}$ and $\widetilde{\mathbf{W}}_{\mathrm{BB}}$ act as auxiliary variables from which we obtain $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}$ in a manner similar to Section III.[6] As a result, the MMSE estimation problem is again equivalent to the problem of sparse signal recovery with multiple measurement vectors and can thus be solved via the orthogonal matching pursuit concept used in Section III. For completeness the pseudo code is given in Algorithm 2.

Remark 5: This section relaxes the perfect-receiver assumption of Section III and proposes practical methods to find low-complexity linear receivers. The design of precoders and combiners, however, remains decoupled as we have assumed that the precoders $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ are fixed while designing $\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}$ (and that receivers are optimal while designing $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$). This decoupled approach simplifies mmWave transceiver design, and will be shown to perform well in Section VI, however, some simple "joint decisions" may be both practical and beneficial. For example, consider the case where a receiver only has a single RF chain and thus is restricted to applying a single response vector $\mathbf{a}_{\mathrm{r}}(\phi,\theta)$. In such a situation, designing $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ to radiate power in $N_{\mathrm{t}}^{\mathrm{RF}}$ different directions may lead to a loss in actual received power (since the receiver can only form a beam in one direction). As

[6] As noted in Section III the receiver need not know the exact angles $(\phi_{i\ell}^{\mathrm{r}},\theta_{i\ell}^{\mathrm{r}})$ and can instead use any set of representative azimuth and elevation angles of arrival to construct the matrix of basis vectors $\mathbf{A}_{\mathrm{r}}$.

Algorithm 2 Spatially Sparse MMSE Combining via Orthogonal Matching Pursuit

```
Require: $\mathbf{W}_{\mathrm{MMSE}}$
$\mathbf{W}_{\mathrm{RF}} = \text{Empty Matrix}$
$\mathbf{W}_{\mathrm{res}} = \mathbf{W}_{\mathrm{MMSE}}$
for $i \leq N_{\mathrm{r}}^{\mathrm{RF}}$ do
    $\mathbf{\Psi} = \mathbf{A}_r^* \mathbb{E}\left[\mathbf{y}\mathbf{y}^*\right] \mathbf{W}_{\mathrm{res}}$
    $k = \arg\max_{\ell=1,\ \ldots,\ N_{\mathrm{cl}} N_{\mathrm{ray}}} \left(\mathbf{\Psi}\mathbf{\Psi}^*\right)_{\ell,\ell}$
    $\mathbf{W}_{\mathrm{RF}} = \left[\mathbf{W}_{\mathrm{RF}} | \mathbf{A}_r^{(k)}\right]$
    $\mathbf{W}_{\mathrm{BB}} = \left(\mathbf{W}_{\mathrm{RF}}^* \mathbb{E}\left[\mathbf{y}\mathbf{y}^*\right] \mathbf{W}_{\mathrm{RF}}\right)^{-1} \mathbf{W}_{\mathrm{RF}}^* \mathbb{E}\left[\mathbf{y}\mathbf{y}^*\right] \mathbf{W}_{\mathrm{MMSE}}$
    $\mathbf{W}_{\mathrm{res}} = \frac{\mathbf{W}_{\mathrm{MMSE}} - \mathbf{W}_{\mathrm{RF}} \mathbf{W}_{\mathrm{BB}}}{\|\mathbf{W}_{\mathrm{MMSE}} - \mathbf{W}_{\mathrm{RF}} \mathbf{W}_{\mathrm{BB}}\|_F}$
end for
return $\mathbf{W}_{\mathrm{RF}}$, $\mathbf{W}_{\mathrm{BB}}$
```

a result, it is beneficial to account for the limitations of the more-constrained terminal when designing either precoders or combiners. To do so, we propose to run Algorithms 1 and 2 in succession according to the following rules

$$N_{\mathrm{t}}^{\mathrm{RF}} < N_{\mathrm{r}}^{\mathrm{RF}} \left\{ \begin{array}{l} \text{1. Solve for } \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \text{ using Algorithm 1.} \\ \text{2. Given } \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\text{, solve for } \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}} \\ \quad \text{using Algorithm 2.} \end{array} \right.$$

$$N_{\mathrm{t}}^{\mathrm{RF}} > N_{\mathrm{r}}^{\mathrm{RF}} \left\{ \begin{array}{l} \text{1. Solve for } \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}} \text{ using Algorithm 2} \\ \quad \text{assuming } \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} = \mathbf{F}_{\mathrm{opt}}. \\ \text{2. Solve for } \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \text{ for the effective} \\ \quad \text{channel } \mathbf{W}_{\mathrm{BB}}^*\mathbf{W}_{\mathrm{RF}}^*\mathbf{H}. \end{array} \right. \tag{25}$$

In summary, starting with the more constrained side, the hybrid precoder or combiner is found using Algorithm 1 or 2. Then, given the output, the remaining processing matrix is found by appropriately updating the effective mmWave channel.

Finally, we note that while the numerical results of Section VI indicate that this decoupled approach to mmWave transceiver design yields near-optimal spectral efficiency, a more direct joint optimization of $(\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}, \mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}})$ is an interesting topic for future investigation. Similarly, while we have solved the sparse formulation of the precoding and combining problems via orthogonal matching pursuit, the problems in (17) and (24) can be solved by leveraging other algorithms for simultaneously sparse approximation [47].

V. Limited Feedback Spatially Sparse Precoding

Section III implicitly assumed that the transmitter has perfect knowledge of $\mathbf{H}$ and is thus able to calculate $\mathbf{F}_{\mathrm{opt}}$. Since such transmitter channel knowledge may not be available in practical systems, we propose to fulfill this channel knowledge requirement via limited feedback [20], [52]–[54]. Namely, we assume that the receiver (i) acquires perfect knowledge of $\mathbf{H}$, (ii) calculates $\mathbf{F}_{\mathrm{opt}}$ and a corresponding hybrid approximation $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$, then (iii) feeds back information about $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}$ to the transmitter. Since hybrid precoders are naturally decomposed into an RF and baseband component, we propose to quantize $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$ separately while exploiting the mathematical structure present in each of them.

A. Quantizing the RF Precoder

Recall that the precoder $\mathbf{F}_{\mathrm{RF}}$ calculated Section III has $N_{\mathrm{t}}^{\mathrm{RF}}$ columns of the form $\mathbf{a}_{\mathrm{t}}(\phi, \theta)$. Therefore, $\mathbf{F}_{\mathrm{RF}}$ admits a natural parametrization in terms of the $N_{\mathrm{t}}^{\mathrm{RF}}$ azimuth and elevation angles that it uses. Thus, $\mathbf{F}_{\mathrm{RF}}$ can be efficiently encoded by quantizing its $2N_{\mathrm{t}}^{\mathrm{RF}}$ free variables. For simplicity, we propose to uniformly quantize the $N_{\mathrm{t}}^{\mathrm{RF}}$ azimuth and elevation angles using N_ϕ and N_θ bits respectively. Therefore, the quantized azimuth and elevation angles are such that

$$\mathcal{C}_\phi = \left\{ \phi_{\min}^{\mathrm{t}} + \frac{\phi_{\mathrm{range}}^{\mathrm{t}}}{2^{N_\phi+1}},\ \phi_{\min}^{\mathrm{t}} + \frac{3\phi_{\mathrm{range}}^{\mathrm{t}}}{2^{N_\phi+1}},\ \ldots,\ \phi_{\max}^{\mathrm{t}} - \frac{\phi_{\mathrm{range}}^{\mathrm{t}}}{2^{N_\phi+1}} \right\}$$
$$\mathcal{C}_\theta = \left\{ \theta_{\min}^{\mathrm{t}} + \frac{\theta_{\mathrm{range}}^{\mathrm{t}}}{2^{N_\theta+1}},\ \theta_{\min}^{\mathrm{t}} + \frac{3\theta_{\mathrm{range}}^{\mathrm{t}}}{2^{N_\theta+1}},\ \ldots,\ \theta_{\max}^{\mathrm{t}} - \frac{\theta_{\mathrm{range}}^{\mathrm{t}}}{2^{N_\theta+1}} \right\} \tag{26}$$

where we recall that $[\phi_{\min}^{\mathrm{t}},\ \phi_{\max}^{\mathrm{t}}]$ and $[\theta_{\min}^{\mathrm{t}},\ \theta_{\max}^{\mathrm{t}}]$ are the sectors over which $\Lambda_{\mathrm{t}}(\phi, \theta) \neq 0$. Further, for some configurations of $([\phi_{\min}^{\mathrm{t}},\ \phi_{\max}^{\mathrm{t}}], [\theta_{\min}^{\mathrm{t}},\ \theta_{\max}^{\mathrm{t}}], N_\phi, N_\theta)$, one may be able to further constrain the angles in $\mathcal{C}_\phi$ and $\mathcal{C}_\theta$ to correspond to orthogonal beams as in [33]–[36]. The receiver can then quantize $\mathbf{F}_{\mathrm{RF}}$ by simply selecting the entries of $\mathcal{C}_\phi$ and $\mathcal{C}_\theta$ that are closest in Euclidean distance to $\mathbf{F}_{\mathrm{RF}}$'s angles. Alternatively, as stated in Remark 2, Algorithm 1 can be run directly using the $N_{\mathrm{t}} \times 2^{N_\phi + N_\theta}$ matrix of "quantized response vectors"

$$\mathbf{A}_{\mathrm{t}}^{\mathrm{quant}} = \left[\mathbf{a}_{\mathrm{t}}(\phi_1^{\mathrm{t}}, \theta_1^{\mathrm{t}}),\ \ldots,\ \mathbf{a}_{\mathrm{t}}(\phi_i^{\mathrm{t}}, \theta_\ell^{\mathrm{t}}),\ \ldots,\ \mathbf{a}_{\mathrm{t}}(\phi_{2^{N_\phi}}^{\mathrm{t}}, \theta_{2^{N_\theta}}^{\mathrm{t}}) \right], \tag{27}$$

and the index of the selected angles can be fed back to the transmitter. While this latter approach has higher search complexity, it has the advantage of (i) "jointly quantizing" all $2N_{\mathrm{t}}^{\mathrm{RF}}$ angles, and (ii) automatically matching the baseband precoder $\mathbf{F}_{\mathrm{BB}}$ to the quantized angles.

B. Quantizing the Baseband Precoder

To efficiently quantize $\mathbf{F}_{\mathrm{BB}}$, we begin by highlighting its mathematical structure in mmWave systems of interest. Namely, we note that for systems with large antenna arrays, we typically have that $\mathbf{F}_{\mathrm{RF}}^*\mathbf{F}_{\mathrm{RF}} \approx \mathbf{I}_{N_{\mathrm{t}}^{\mathrm{RF}}}$. When coupled with Approximation 1, we have that $\mathbf{F}_{\mathrm{BB}}^*\mathbf{F}_{\mathrm{BB}} \approx \mathbf{I}_{N_{\mathrm{s}}}$, i.e., $\mathbf{F}_{\mathrm{BB}}$ is approximately unitary. In fact, $\mathbf{F}_{\mathrm{BB}}$ can be made exactly unitary as discussed in Remark 3. Further, we recall that the spectral efficiency expression in (3) is invariant to $N_{\mathrm{s}} \times N_{\mathrm{s}}$ unitary transformations of the baseband precoder. Therefore, $\mathbf{F}_{\mathrm{BB}}$ is a subspace quantity that can be quantized on the Grassmann manifold [52], [53]. Suitable codebooks for $\mathbf{F}_{\mathrm{BB}}$ can be designed using Lloyd's algorithm on a training set of baseband precoders and using the chordal distance as a distance measure [75]. Since such codebook construction is well-studied in the literature on limited feedback MIMO, we omit its details for brevity and refer the reader to [76, Section IV] for an in-depth description of the process.

VI. Simulation Results

In this section, we present simulation results to demonstrate the performance of the spatially sparse precoding algorithm presented in Section III when combined with the sparse MMSE combining solution presented in Section IV. We model the propagation environment as a $N_{\mathrm{cl}} = 8$ cluster environment with $N_{\text{ray}} = 10$ rays per cluster with Laplacian distributed azimuth and elevation angles of arrival and departure [29], [61]. For simplicity of exposition, we assume all clusters are of equal power, i.e., $\sigma^2_{\alpha,i} = \sigma^2_{\alpha}$ $\forall i$, and that the angle spread at both the transmitter and receiver are equal in the azimuth and elevation domain, i.e., $\sigma^{\text{t}}_{\phi} = \sigma^{\text{r}}_{\phi} = \sigma^{\text{t}}_{\theta} = \sigma^{\text{r}}_{\theta}$. Since outdoor deployments are likely to use sectorized transmitters to decrease interference and increase beamforming gain, we consider arrays of directional antenna elements with a response given in (5) [8], [9]. The transmitter's sector angle is assumed to be 60°-wide in the azimuth domain and 20°-wide in elevation [8]. In contrast, we assume that the receivers have relatively smaller antenna arrays of omni-directional elements; this is since receivers must be able to steer beams in any direction since their location and orientation in real systems is random. The inter-element spacing d is assumed to be half-wavelength. We compare the performance of the proposed strategy to optimal unconstrained precoding in which streams are sent along the channel's dominant eigenmodes. When perfect channel knowledge is available to the transmitter, we use the channel's actual angles of departure $(\phi^{\text{t}}_{i\ell}, \theta^{\text{t}}_{i\ell})$ as input to Algorithm 1, i.e., the algorithm is implemented as described in the main discussion Section III. We also compare with a simple beam steering solution in which data streams are steered onto the channel's best propagation paths and received along the best corresponding receive paths [17], [19].[7] For fairness, the same total power constraint is enforced on all precoding solutions and signal-to-noise ratio is defined as $\text{SNR} = \frac{\rho}{\sigma_n^2}$. Finally, all reported results are averaged over 5000 random channel realizations.

Fig. 3 shows the spectral efficiency achieved in a 64×16 system with square planar arrays at both transmitter and receiver. For the proposed precoding strategy, both transmitter and receiver are assumed to have four transceiver chains with which they transmit $N_{\text{s}} = 1$ or 2 streams. Fig. 3 shows that the proposed framework achieves spectral efficiencies that are essentially equal to those achieved by the optimal unconstrained solution in the case $N_{\text{s}} = 1$ and are within a small gap from optimality in the case of $N_{\text{s}} = 2$. This implies that the proposed strategy can very accurately approximate the channel's dominant singular vectors as a combination of four steering vectors. When compared to traditional beam steering, Fig. 3 shows that there is a non-negligible improvement to be had from more sophisticated precoding strategies in mmWave systems with practical array sizes. To explore performance in mmWave systems with larger antenna arrays, Fig. 4 plots the performance achieved in a 256×64 system with $N^{\text{RF}}_{\text{t}} = N^{\text{RF}}_{\text{r}} = 6$ RF chains. Fig. 4 shows that the proposed precoding/combining solution achieves almost-perfect performance in both $N_{\text{s}} = 1$ and $N_{\text{s}} = 2$ cases. Further, we note that although beam steering is expected to be optimal in the limit of large arrays, as discussed in Remark 4, the proposed solution still outperforms beam steering by approximately 5 dB in this larger mmWave system.

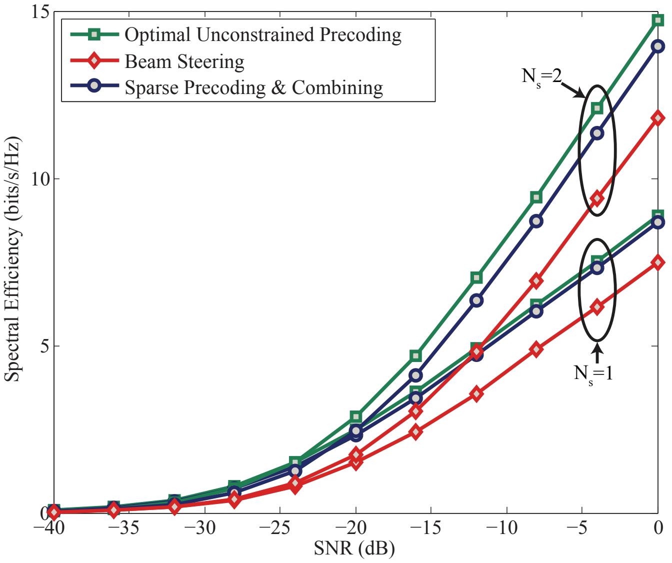


Fig. 3. Spectral Efficiency achieved by various precoding solutions for a 64×16 mmWave system with planar arrays at the transmitter and receiver. The propagation medium is a $N_{\text{cl}} = 8$ cluster environment with $N_{\text{ray}} = 10$ and an angular spread of 7.5°. Four RF chains are assumed to be available for sparse precoding and MMSE combining.

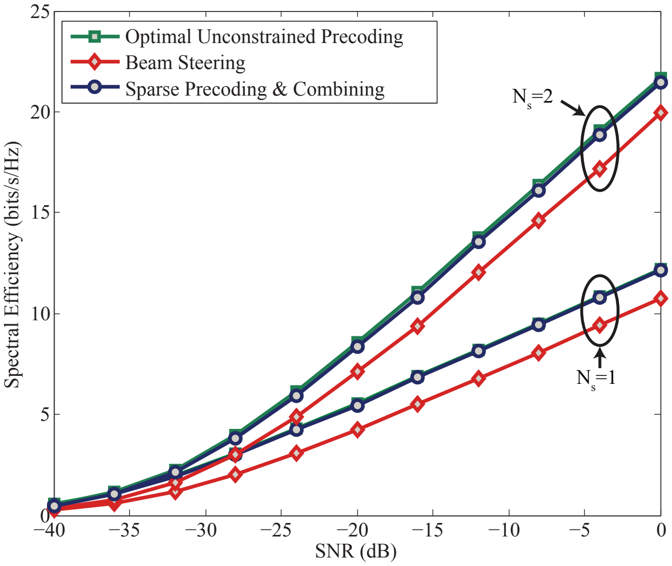


Fig. 4. Spectral Efficiency achieved in a 256×64 mmWave system with planar arrays at the transmitter and receiver. Channel parameters are set as in Fig. 3. Six RF chains are available for sparse precoding and combining.

While Section III focused on the design of fixed-rank precoders with equal power allocation across streams, the same framework can be applied to systems in which N_{s} is determined dynamically and streams are sent with unequal power. This configuration allows us to compare the rates achieved by the proposed precoding/combining framework to the mmWave channel's waterfilling capacity. To do so, Algorithm 1 is simply set to approximate $\mathbf{F}_{\text{opt}} = \mathbf{V}\mathbf{\Gamma}$ where $\mathbf{\Gamma}$ is a diagonal matrix resulting from the waterfilling power allocation. Fig. 5 demonstrates the performance achieved when Algorithms 1 and 2 are used to approximate the channel's

[7] Note that, when $N_{\text{s}} > 1$, the best propagation paths in terms of spectral efficiency may not be the ones with the highest gains. This is since, with no receiver baseband processing, different paths must be sufficiently separated so as they do not interfere. In this case, the best paths for transmission and reception are chosen to maximize mutual information via a costly exhaustive search over a very fine-grain set of beam steering directions. Further, when power allocation is considered in Fig. 5, the same waterfilling power allocation is applied to the beam steering solution.

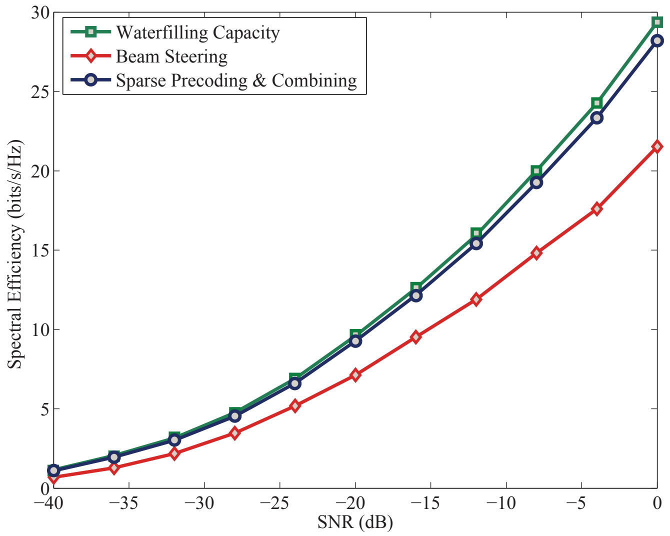


Fig. 5. This figure compares the spectral efficiency achieved when rank adaptation and unequal power allocation is allowed in 256×64 system with $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 4$. It is shown that sparse precoding and combining can approach the performance of an unconstrained capacity-achieving (waterfilling) precoder. The figure also demonstrates large gains over a beam steering strategy in which streams are sent along different physical directions with a similar unequal power allocation.

capacity-achieving precoders and combiners in a 256×64 mmWave system with $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 4$. Fig. 5 shows that the proposed framework allows systems to approach channel capacity and provides large gains over simple beam steering. Since the multiplexing gain of the mmWave system is limited by $N_{\mathrm{s}} \leq \min\{N_{\mathrm{t}}^{\mathrm{RF}}, N_{\mathrm{r}}^{\mathrm{RF}}\}$, capacity cannot be approached at very high SNR when the optimal N_{s} exceeds $\min\{N_{\mathrm{t}}^{\mathrm{RF}}, N_{\mathrm{r}}^{\mathrm{RF}}\}$. Fig. 5 indicates, however, that even at an SNR of 0 dB where we observe that $N_{\mathrm{s}} = 3$ streams are sent over most channel realizations, the proposed strategy is still within a small gap from capacity. Finally, we note that although the derivation leading up to (14) does not account for unequal power allocation across streams, Fig. 5 indicates that Algorithm 1 is nevertheless a sensible approach to designing such precoders.

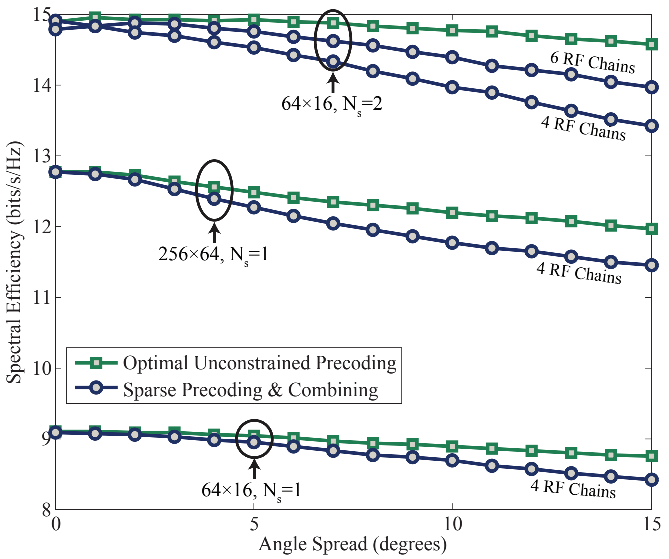


Fig. 6. Spectral Efficiency vs. Angle Spread in a number of different mmWave system configuration at an SNR of 0 dB. For simplicity of exposition, we assume that the angle spread is such that $\sigma_{\phi}^{\mathrm{t}} = \sigma_{\phi}^{\mathrm{r}} = \sigma_{\theta}^{\mathrm{t}} = \sigma_{\theta}^{\mathrm{r}}$. It is shown that as angle spread increased, and scattering becomes richer, the performance of the proposed algorithm degrades. However, the rate gap remains below 10% at a significant angle spread of 15°. For more reasonable angle spreads of around 5°, the rate gap is negligible.

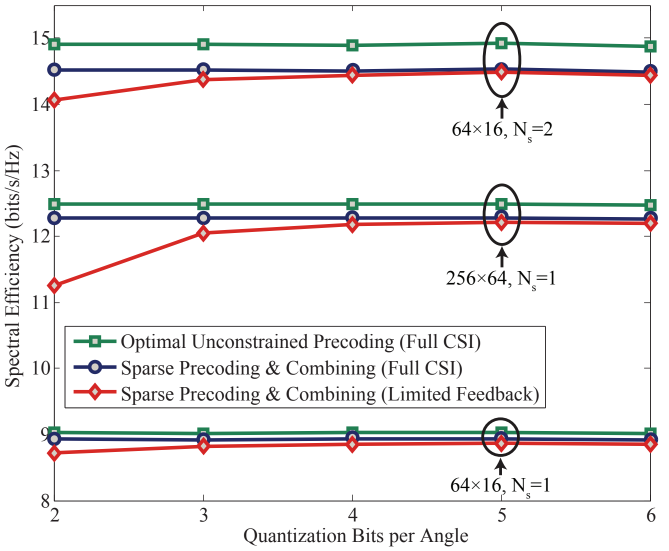


Fig. 7. Spectral Efficiency vs. Quantization Bits per Angle different mmWave system configurations, all with $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 4$, at an SNR of 0 dB in a channel with an azimuth and elevation angular spread of 7.5°. For simplicity of exposition, we assume that $N_{\phi} = N_{\theta}$ and an baseband precoder codebook of 4 bits in the $N_{\mathrm{s}} = 1$ case and 6 bits in the $N_{\mathrm{s}} = 2$ case. The figure indicates that for the considered array sizes, 3 bits per angle is often enough to achieve almost-perfect performance.

The proposed precoding/combining framework leverages the mathematical structure of large mmWave channels with relatively limited scattering. To examine performance in propagation environments with varying levels of scattering, Fig. 6 plots spectral efficiency as a function of the channel's angle spread for a number of mmWave system configurations. Fig. 6 indicates that when the angle spread is low, i.e., the scattering is rather limited, the performance of the proposed algorithm is within a small gap from the performance of unconstrained precoding. As angle spread increases, the rates achieved by the proposed solutions slowly degrade. However, Fig 6 indicates that in the two $N_{\mathrm{s}} = 1$ cases shown, the rate gap remains below 10% at a significant angle spread of 15° and is negligible for more reasonable angle spreads of around 5°. In the case of $N_{\mathrm{s}} > 1$ with smaller arrays, spectral efficiency degrades more rapidly with angle spread. This can be seen by examining the 64×16 system with $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 4$ and $N_{\mathrm{s}} = 2$. If possible, the effect of increased scattering can be mitigated by increasing the number of RF chains at the mmWave terminals which enables them to generate more flexible precoders/combiners. This can be seen by examining the same 64×16 system with $N_{\mathrm{t}}^{\mathrm{RF}} = N_{\mathrm{r}}^{\mathrm{RF}} = 6$.

Finally, we examine the performance of the proposed precoding strategy in systems without channel state information at the transmitter. For this performance characterization, we assume that the receiver calculates $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}$ with full knowledge of the channel and feeds back their parameters as described in Section V. We assume that the receiver uses four and six bits to quantize $\mathbf{F}_{\rm BB}$ in the case of $N_{\rm s}=1$ and $N_{\rm s}=2$ respectively, and constructs codebooks as described in Section V-B. The receiver uses a variable number of bits to quantize the azimuth and elevation angles used in $\mathbf{F}_{\rm RF}$. For simplicity of exposition, we assume that $N_\phi = N_\theta$. Fig. 7 indicates that similar performance can be expected in limited feedback systems and that the performance degradation due to quantization is limited. Namely, Fig. 7 indicates that no more than 3 bits are needed to quantize each steering angle in practical systems, and even 2 bits yields almost-perfect performance for a 64×16 systems with $N_{\rm s}=1$. In general the number of bits needed to properly quantize the steering angles grows slowly with array size since larger arrays generate narrower beams and require finer steering. Since beam width is inversely proportional to the antenna array dimensions, a reasonable rule-of-thumb is to add 1 bit per azimuth (or elevation) steering angle whenever the array's width (or height) doubles. Fig. 7 is promising as it indicates that it takes no more than 20 bits to quantize a 64×1 precoder and about 22 bits for a 64×2 precoder. When considering the fact that practical mmWave systems will use twenty to fifty times more antennas compared to traditional MIMO systems, which use about 4 to 6 bits of feedback [52], we see that exploiting spatial sparsity in precoding helps dramatically compress feedback and keep its overhead manageable.

VII. Conclusion

In this paper we considered single user precoding and combing in mmWave systems where traditional MIMO solutions are made infeasible by the heavy reliance on RF precoding. By leveraging the structure of realistic mmWave channels, we developed a low hardware-complexity precoding solution. We formulated the problem of mmWave precoder design as a sparsity-constrained signal recovery problem and presented an algorithmic solution using orthogonal matching pursuit. We showed that the same framework can be applied to the problem of designing practical MMSE combiners for mmWave systems. We showed that the proposed precoders can be efficiently quantized and that the precoding strategy is well-suited for limited feedback systems. Finally, we presented numerical results on the performance of spatially sparse mmWave processing and showed that it allows systems to approach their theoretical limits on spectral efficiency. Future work related to such mmWave precoding includes relaxing the assumptions made throughout this paper such as (i) perfect channel state information at the receiver, (ii) knowledge of the antenna array structure, and (iii) the specialization to narrowband channels.

References

[1] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE Wireless Commun.*, vol. 17, no. 3, pp. 10–22, 2010.

[2] D. Lopez-Perez, I. Guvenc, G. De La Roche, M. Kountouris, T. Q. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, 2011.

[3] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, 2011.

[4] S. Yong and C. Chong, "An overview of multigigabit wireless through millimeter wave technology: potentials and technical challenges," *EURASIP J. Wireless Commun. Netw.*, vol. 2007, no. 1, pp. 50–50, 2007.

[5] R. Daniels and R. W. Heath, Jr., "60 GHz wireless communications: emerging requirements and design recommendations," *IEEE Veh. Technol. Mag.*, vol. 2, no. 3, pp. 41–50, 2007.

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

[7] C. Doan, S. Emami, D. Sobel, A. Niknejad, and R. Brodersen, "Design considerations for 60 GHz CMOS radios," *IEEE Commun. Mag.*, vol. 42, no. 12, pp. 132–140, 2004.

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

[9] G. Hendrantoro, R. Bultitude, and D. Falconer, "Use of cell-site diversity in millimeter-wave fixed cellular systems to combat the effects of rain attenuation," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 602–614, 2002.

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

[11] E. Torkildson, B. Ananthasubramaniam, U. Madhow, and M. Rodwell, "Millimeter-wave MIMO: wireless links at optical speeds," in *Proc. 2006 Allerton Conf. Commun., Control Comput.*

[12] A. Valdes-Garcia, S. T. Nicolson, J.-W. Lai, A. Natarajan, P.-Y. Chen, S. K. Reynolds, J.-H. C. Zhan, D. G. Kam, D. Liu, and B. Floyd, "A fully integrated 16-element phased-array transmitter in SiGe BiCMOS for 60-GHz communications," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2757–2773, 2010.

[13] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, 2004.

[14] A. F. Molisch, M. Z. Win, Y.-S. Choi, and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1759–1772, 2005.

[15] A. Gorokhov, D. A. Gore, and A. J. Paulraj, "Receive antenna selection for MIMO spatial multiplexing: theory and algorithms," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2796–2807, 2003.

[16] Z. Xu, S. Sfar, and R. S. Blum, "Analysis of MIMO systems with receive antenna selection in spatially correlated Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 251–262, 2009.

[17] J. Wang, Z. Lan, C.-W. Pyo, T. Baykas, C.-S. Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, *et al.*, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, 2009.

[18] A. F. Molisch and X. Zhang, "FFT-based hybrid antenna selection schemes for spatially correlated MIMO channels," *IEEE Commun. Lett.*, vol. 8, no. 1, pp. 36–38, 2004.

[19] O. El Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "The capacity optimality of beam steering in large millimeter wave MIMO systems," in *Proc. 2012 IEEE International Workshop Signal Process. Advances Wireless Commun.*, pp. 100–104.

[20] D. Love and R. W. Heath, Jr., "Equal gain transmission in multiple-input multiple-output wireless systems," *IEEE Trans. Commun.*, vol. 51, no. 7, pp. 1102–1110, July 2003.

[21] P. Sudarshan, N. Mehta, A. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, Dec. 2006.

[22] X. Zheng, Y. Xie, J. Li, and P. Stoica, "MIMO transmit beamforming under uniform elemental power constraint," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5395–5406, 2007.

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

[24] F. Gholam, J. Vía, and I. Santamaría, "Beamforming design for simplified analog antenna combining architectures," *IEEE Trans. Veh. Technol.*, vol. 60, no. 5, pp. 2373–2378, 2011.

[25] Z. Pi, "Optimal transmitter beamforming with per-antenna power constraints," in *Proc. 2012 IEEE International Conf. Commun.*, pp. 3779–3784.

[26] X. Zhang, A. Molisch, and S.-Y. 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.

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

[28] P. Smulders and L. Correia, "Characterisation of propagation in 60 GHz radio channels," *Electron. Commun. Eng. J.*, vol. 9, no. 2, pp. 73–80, 1997.

[29] H. Xu, V. Kukshya, and T. Rappaport, "Spatial and temporal characteristics of 60-GHz indoor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, 2002.

[30] Q. Spencer, B. Jeffs, M. Jensen, and A. Swindlehurst, "Modeling the statistical time and angle of arrival characteristics of an indoor multipath channel," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 347–360, Mar. 2000.

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

[32] P. Xia, S.-K. Yong, J. Oh, and C. Ngo, "A practical SDMA protocol for 60 GHz millimeter wave communications," in *Proc. 2008 Asilomar Conf. Signals, Syst. Comput.*, pp. 2019–2023.

[33] A. Sayeed and N. Behdad, "Continuous aperture phased MIMO: basic theory and applications," in *Proc. 2010 Allerton Conf. Commun., Control, Comput.*, pp. 1196–1203.

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

[35] A. M. Sayeed and N. Behdad, "Continuous aperture phased MIMO: a new architecture for optimum line-of-sight links," in *Proc. 2011 IEEE International Symp. Antennas Propagation*, pp. 293–296.

[36] G. H. Song, J. Brady, and A. Sayeed, "Beamspace MIMO transceivers for low-complexity and near-optimal communication at mm-wave frequencies," in *Proc. 2013 IEEE International Conf. Acoustics, Speech, Signal Process.*

[37] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," to appear in *Proc. 2013 IEEE Global Telecommun. Conf.*

[38] I. Markovsky and S. Van Huffel, "Overview of total least-squares methods," *Signal Process.*, vol. 87, no. 10, pp. 2283–2302, 2007.

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

[40] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[41] L. Rebollo-Neira and D. Lowe, "Optimized orthogonal matching pursuit approach," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 137–140, 2002.

[42] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[44] O. El Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems," in *Proc. 2012 IEEE International Conf. Commun.*, pp. 3724–3729.

[45] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, 2005.

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

[47] J. A. Tropp, "Algorithms for simultaneous sparse approximation—part II: convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

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

[49] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[50] T. Michaeli and Y. Eldar, "Constrained linear minimum MSE estimation," Oct. 2007. Available: http://webee.technion.ac.il/publication-link/index/id/439

[51] T. Michaeli and Y. Eldar, "Constrained nonlinear minimum MSE estimation," in *Proc. 2008 IEEE International Conf. Acoustics, Speech, Signal Process.*, pp. 3681–3684.

[52] D. Love and R. W. Heath, Jr., "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2967–2976, Aug. 2005.

[53] D. Love, R. W. Heath, Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[54] J. C. Roh and B. D. Rao, "Design and analysis of MIMO spatial multiplexing systems with quantized feedback," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 2874–2886, 2006.

[55] 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, 2010.

[56] D. Ramasamy, S. Venkateswaran, and U. Madhow, "Compressive adaptation of large steerable arrays," in *Proc. 2012 Inf. Theory Applications Workshop*, pp. 234–239.

[57] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *2013 Proc. Inf. Theory Applications Workshop.*

[58] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, 2003.

[59] IEEE 802.15 WPAN Millimeter Wave Alternative PHY Task Group 3c. Available: www.ieee802.org/15/pub/TG3c.html,Sept.,2011.

[60] V. Raghavan and A. M. Sayeed, "Sublinear capacity scaling laws for sparse MIMO channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 345–364, 2011.

[61] A. Forenza, D. Love, and R. W. Heath, Jr., "Simplified spatial correlation models for clustered MIMO channels with different array configurations," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1924–1934, 2007.

[62] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-GHz mesh networks: the case for rethinking medium access control," *IEEE/ACM Trans. Netw.*, vol. 19, no. 5, pp. 1513–1527, 2011.

[63] C. Balanis, *Antenna Theory*. Wiley, 1997.

[64] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: a unified framework for convex optimization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2381–2401, 2003.

[65] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, 2006.

[66] J. M. Lee, *Introduction to Smooth Manifolds*. Springer, 2012, vol. 218.

[67] T. Figiel, J. Lindenstrauss, and V. D. Milman, "The dimension of almost spherical sections of convex bodies," *Acta Mathematica*, vol. 139, no. 1, pp. 53–94, 1977.

[68] N. Kwak, "Principal component analysis based on l1-norm maximization," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008.

[69] 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.

[70] A. S. Lewis and J. Malick, "Alternating projections on manifolds," *Mathematics Operations Research*, vol. 33, no. 1, pp. 216–234, 2008.

[71] R. Escalante and M. Raydan, *Alternating Projection Methods*. Society for Industrial and Applied Mathematics, 2011, vol. 8.

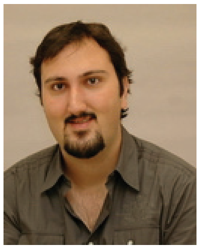
[72] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Rev.*, vol. 38, no. 3, pp. 367–426, 1996.

[73] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. Oxford University Press, 2004, vol. 3.

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

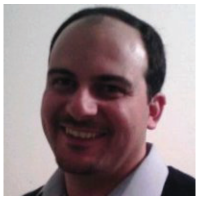
[75] R. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, no. 2, pp. 4–29, Apr. 1984.

[76] S. Zhou and B. Li, "BER criterion and codebook construction for finite-rate precoded spatial multiplexing with linear receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1653–1665, 2006.

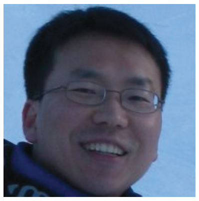


Omar El Ayach (S'08, M'13) is currently a Senior Systems Engineer at Qualcomm Research in San Diego, CA. He received his M.S. and Ph.D. in Electrical and Computer Engineering degree from The University of Texas at Austin in 2010 and 2013, respectively. Before joining the University of Texas, he received his B.E. degree in Computer and Communications Engineering from the American University of Beirut, Lebanon, in 2008. He was an intern at the University of California, Berkeley in the Summer of 2007 and an intern at Samsung Research America - Dallas in the Summers of 2011 and 2012. His research interests are in the broad area of network sciences, signal processing and information theory. In the context of wireless communication, his interests are in MIMO systems, interference management, and mmWave communication.

Sridhar Rajagopal (M98, SM09) is currently a Sr. Staff Engineer at Samsung Research America Dallas. He received his M.S. and Ph.D. degrees in Electrical and Computer Eng. from Rice University. He has previously worked at Nokia Research Center and at WiQuest Communications, and has contributed to multiple communication standards. His research interests are in algorithms and architectures for short-range, high throughput and low power technologies, mmWave and optical wireless communication.



Shadi Abu-Surra is currently a Staff Engineer at Samsung Research America Dallas. He received M.S. in Electrical and Computer Engineering from the Jordan University of Science and Technology in 2004 and his Ph.D. degree in Electrical and Computer Engineering from the University of Arizona in 2009. His research interests are in coding theory, LDPC code design and decoder architectures, cellular communication, millimeter wave and 60 GHz communication.



Zhouyue Pi (SM'13) is a Senior Director at Samsung Research America in Dallas, Texas, where he leads the Emerging Technology Lab doing research in next generation mobile devices, smart home solutions, and medical diagnostic technologies. Before joining Samsung in 2006, he was with Nokia Research Center in Dallas and San Diego, where he worked on 3G wireless standardization and modem development for systems such as 3GPP2 1xEV-DV Revision C and D, 1xEV-DO Revisions A and B, and UMB. In2006 2009, he was a main contributor to Samsung's 4G standardization efforts in 3GPP LTE and IEEE 802.16m (mobile WiMAX evolution), and to IEEE 802.11ad for 60 GHz communication also commonly referred to as Wireless Gigabit (WiGig). In 2009 2012, he pioneered mm-wave mobile communication and led the development of the world's first baseband and RF system prototype that demonstrated the feasibility of mobile communication in frequency as high as 28 GHz. He has authored more than 20 technical papers in internationally circulated journals and premier conferences and is the inventor of more than 140 patents and applications. He holds a B.E. degree from Tsinghua University (with honor), a M.S. degree from the Ohio State University, and an MBA degree from Cornell University (with distinction). He is a Senior Member of IEEE.

Robert W. Heath, Jr. (S'96– M'01–SM'06–F'11) received the B.S. and M.S. degrees from the University of Virginia, Charlottesville, VA, in 1996 and 1997 respectively, and the Ph.D. from Stanford University, Stanford, CA, in 2002, all in electrical engineering. From 1998 to 2001, he was a Senior Member of the Technical Staff then a Senior Consultant at Iospan Wireless Inc, San Jose, CA where he worked on the design and implementation of the physical and link layers of the first commercial MIMO-OFDM communication system. Since January 2002, he has been with the Department of Electrical and Computer Engineering at The University of Texas at Austin where he is a Cullen Trust for Higher Education Endowed Professor, and is Director of the Wireless Networking and Communications Group. He is also President and CEO of MIMO Wireless Inc. and Chief Innovation Officer at Kuma Signals LLC. His research interests include several aspects of wireless communication and signal processing: limited feedback techniques, multihop networking, multiuser and multicell MIMO, interference alignment, adaptive video transmission, manifold signal processing, and millimeter wave communication techniques.

Dr. Heath has been an Editor for the IEEE Transactions on Communication, an Associate Editor for the IEEE Transactions on Vehicular Technology, lead guest editor for an IEEE Journal on Selected Areas in Communications special issue on limited feedback communication, and lead guest editor for an IEEE Journal on Selected Topics in Signal Processing special issue on Heterogenous Networks. He currently serves on the steering committee for the IEEE Transactions on Wireless Communications. He was a member of the Signal Processing for Communications Technical Committee in the IEEE Signal Processing Society and is a former Chair of the IEEE COMSOC Communications Technical Theory Committee. He was a technical co-chair for the 2007 Fall Vehicular Technology Conference, general chair of the 2008 Communication Theory Workshop, general co-chair, technical co-chair and co-organizer of the 2009 IEEE Signal Processing for Wireless Communications Workshop, local co-organizer for the 2009 IEEE CAMSAP Conference, technical co-chair for the 2010 IEEE International Symposium on Information Theory, the technical chair for the 2011 Asilomar Conference on Signals, Systems, and Computers, general chair for the 2013 Asilomar Conference on Signals, Systems, and Computers, founding general co-chair for the 2013 IEEE GlobalSIP conference, and is technical co-chair for the 2014 IEEE GLOBECOM conference.

Dr. Heath was a co-author of best student paper awards at IEEE VTC 2006 Spring, WPMC 2006, IEEE GLOBECOM 2006, IEEE VTC 2007 Spring, and IEEE RWS 2009, as well as co-recipient of the Grand Prize in the 2008 WinTech WinCool Demo Contest. He was co-recipient of the 2010 and 2013 *EURASIP Journal on Wireless Communications and Networking* best paper awards and the 2012 *Signal Processing Magazine* best paper award. He was a 2003 Frontiers in Education New Faculty Fellow. He is also a licensed Amateur Radio Operator and is a registered Professional Engineer in Texas.