Compressive Sensing with Applications to Millimeter-wave Architectures

Hadi Ghauch*, Taejoon Kim†, Carlo Fischione*, Mikael Skoglund*
*School of Electrical Engineering and Computer Science (EECS), KTH, Stockholm, 10044, Sweden
†Department of EECS, The University of Kansas, Lawrence, KS 66045, USA
Email: ghauch@kth.se, taejoon@ku.edu, carlofi@kth.se, skoglund@kth.se

***Abstract*—To make the system available at low-cost, millimeter-wave (mmWave) multiple-input multiple-output (MIMO) architectures employ analog arrays, which are driven by a limited number of radio frequency (RF) chains. One primary challenge of using large hybrid analog-digital arrays is that the digital baseband cannot directly access the signal to/from each antenna. To address this limitation, recent research has focused on retransmissions, iterative precoding, and subspace decomposition methods. Unlike these approaches that exploited the channel's low-rank, in this work we exploit the sparsity of the received signal at both the transmit/receive antennas. While the signal itself is de facto dense, it is well-known that most signals are sparse under an appropriate choice of basis. By delving into the structured compressive sensing (CS) framework and adapting them to variants of the mmWave hybrid architectures, we provide methodologies to recover the analog signal at each antenna from the (low-dimensional) digital signal. Moreover, we characterizes the minimal numbers of measurement and RF chains to provide this recovery, with high probability. We discuss their applications to common variants of the hybrid architecture. By leveraging the inherent sparsity of the received signal, our analysis reveals that a hybrid MIMO system can be "turned into" a fully digital one: the number of needed RF chains increases logarithmically with the number of antennas.**

I. Introduction

With the growing demand for extreme broadband, the operating frequency of 5G radio is steadily shifting upward to the mmWave band. Despite its widespread use in mmWave communications [1], [2], [3], the hybrid analog-digital architecture suffers from the fact that one does not have access to the signal at the transmit/receiver antennas. This fundamental problem has several implications: a) Unlike conventional (fully digital) MIMO systems, the receive signal at the antennas cannot be directly manipulated. b) The received signal can be digitally processed only after the application of the analog combiner. Thus, the channel output can be observed only through a low-dimensional projection, since the number of radio frequency (RF) chains, r, is drastically smaller than the number of receive antennas, N.

To circumvent this limitation previous approaches had recourse to schemes such as, i) repetition [4], i.e., transmitting the same signal several times and combining it each time with a different filter at the receiver (see Fig 1). Evidently, this procedure is wasteful of the channel resources due to the large communication overhead that scales as $\mathcal{O}(N/r)$. ii) Hybrid precoding algorithms that approximate the fully digital projection by a cascade of analog and digital filters. These algorithms are inherently iterative, with high computational complexity that scales with $\mathcal{O}(r^3)$, and lack convergence guarantees [5]. iii) Low-rank subspace decomposition that leverages the restricted isometry property of the channel sounding signals [6]. These approaches thus result in large communication overhead. Providing low overhead is critical to mmWave MIMO systems suffering from the sheer scarcity of channel coherence resources. It is now well understood that a received signal - at the output mmWave MIMO channel - is dense in its spatial domain, but becomes sparse in a properly transformed space [7]. Promising directions have recently been proposed in the context of structured compressive sensing (CS) [8].

In this work, we pay heed to the framework of structured CS [9], [8]. Unlike the previous works [4], [6], which rely on the sparsity of the channel's eigenmodes, in this work we exploit the sparsity of the signal at the transmit/receive antennas. Our insight is to leverage the inherent sparsity of the signal in the transformed space, to recover the received signal with a minimal number of measurements. We theoretically investigate and compare the minimal number of measurements for three known recovery methods. We discuss their application to variants of the hybrid architecture, namely, the fully connected architecture, and the electromagnetic lens architecture. Finally, we argue how the proposed method can be used to tradeoff computational complexity and communication overhead. Our numerical results show the proposed approach can recover the N-dimensional signal at the MS antennas, with $\approx N/8$ measurements.

Notation: In the following, we use bold upper-case letters to denote matrices, and bold lower-case letters denote vectors. For a given matrix $\boldsymbol{A}$, $[\boldsymbol{A}]_{i,j}$ denotes the element (i,j) in $\boldsymbol{A}$, $\mathrm{tr}(\boldsymbol{A})$ denotes its trace, $\|\boldsymbol{A}\|_F^2$ its Frobenius norm, and $\boldsymbol{A}^\dagger$ its conjugate transpose. $[\boldsymbol{A}]_{\mathcal{R}}$ denotes the submatrix formed by taking rows indexed by $\mathcal{R}$. We let $\{n\} \triangleq \{1, ..., n\}$, and $\boldsymbol{I}$ denotes identity matrix. For set $\mathcal{X}$, $|\mathcal{X}|$ denotes its cardinality.

II. System Model

Assume a single user MIMO system with M and N antennas at the base station (BS) and mobile station (MS), respectively, where each is equipped with r RF chains, and sends d independent data streams. We assume that $d \leq r \leq$

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

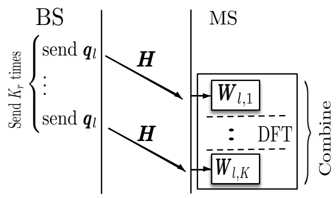

Fig. 1. Repetition-aided (RAID) echoing for the hybrid analog-digital architecture

$\min(M,N)$ and $r \ll N$. The downlink (DL) received signal is given by,

$$\tilde{\boldsymbol{x}} = \boldsymbol{W}^\dagger \boldsymbol{y}_R = \boldsymbol{W}^\dagger \boldsymbol{H}\boldsymbol{F}\boldsymbol{x}_T + \boldsymbol{W}^\dagger \boldsymbol{n}_R \tag{1}$$

where $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ is the complex channel, assumed to be slowly block-fading, $\boldsymbol{F} \in \mathbb{C}^{M\times r}$ is the analog precoder, $\boldsymbol{W} \in \mathbb{C}^{N\times r}$ the analog combiner, $\boldsymbol{y}_R$ the N-dimensional signal at the MS antennas, $\boldsymbol{x}_T$ is the transmit signal, and $\boldsymbol{n}_R$ is the zero-mean AWGN noise at the MS, with $\mathbb{E}[\|\boldsymbol{n}_R\|_2^2] = \sigma_R^2$. Note that $_T$ and $_R$ subscripts denote quantities at the BS and MS, respectively. Both the analog precoder and combiner are constrained to have constant modulus elements which model phase shifters. The above assumptions are widespread in the mmWave MIMO literature [10]. We ignore the digital precoder and combiner without affecting our results.

A. Motivation

Unlike conventional (fully digital) MIMO systems, the transmit/receive signal at the antennas cannot be directly manipulated. Our earlier approach for mitigating this fun issue was to use simple repetitions/retransmissions [4]: Retransmit the same precoded signal K_R times at the BS, and each time combine with columns of a DFT matrix at the MS (refer to Fig. 1). This essentially turns a hybrid analog-digital MIMO link into a conventional one. However, this comes at the cost of N/r (resp. M/r) retransmissions in the downlink (resp. uplink). The added (but potentially unnecessary) communication overhead is due to the restriction that we observe the N-dimensional channel output only via an r-dimensional observation (where $r \ll N$).

In this work, we aim to use the framework of Compressive Sensing (CS), to significantly reduce this repetition overhead. More specifically, we wish to recover the signal at the MS antennas, $\boldsymbol{y}_A$, from the limited number of measurements provided by the digital output $\boldsymbol{y}_D$. We will exploit a well-known property, namely, that the signal at the antennas is sparse in some basis $\boldsymbol{\Psi}$: Although this observation has been established [7] the problem considered here has not been addressed. We thus investigate several recovery methods and analytically compare the minimum number of retransmission that each requires, their recovery guarantees, and their computational complexity. In contrast to the asymptotic guarantees offered by CS, we look at exact dependence between the minimum number of measurements and the recovery probability.

B. Measurement Model

In this part, we are concerned with recovering the desired signal, $\boldsymbol{H}\boldsymbol{F}\boldsymbol{x}_T$, despite the reduced number of measurements (i.e. limited number of RF chains), and the limitation of additive noise. Consider the following series of repetitions, where the BS sends the same (precoded) signal $\boldsymbol{H}\boldsymbol{F}\boldsymbol{x}_T$, K_R times, and the received signal is combined with a different analog combiner $\boldsymbol{W}^{(l)}, l \in \{K_R\}$,

$$\boldsymbol{y}_D^{(l)} = \boldsymbol{W}^{(l)^\dagger}(\boldsymbol{H}\boldsymbol{F}\boldsymbol{x}_T + \boldsymbol{n}_R^{(l)}) = \boldsymbol{W}^{(l)^\dagger}(\boldsymbol{y}_S + \boldsymbol{n}_R^{(l)}), \tag{2}$$

where $\boldsymbol{y}_S$ is the desired part of the received signal at the MS antennas. The intuition behind this measurement model is that r observations are obtained through the analog combiner (since all of the r RF chains are activated), during each of the K_R transmissions. We then combine these transmission to estimate the desired signal $\boldsymbol{y}_S$, from the limited number of measurements. Our proposal is to exploit recent tools in structured compressive sensing (CS) to drastically reduce the above communication overhead. Though initially conceived for sampling band-limited signals [9], the ideas have been successfully applied to MRI compression/reconstruction, image processing, optics, ADC design, etc. [8]. A common take home message from the plethora of applications, is that most signals are sparse, under the right basis.

While $\boldsymbol{y}_S$ is dense, it is sparse in some basis $\boldsymbol{\Psi}$. This implies that there exists a basis $\boldsymbol{\Psi} \in \mathbb{C}^{N\times N}$, such that $\boldsymbol{y}_S = \boldsymbol{\Psi}\boldsymbol{s}$, where $\|\boldsymbol{s}\|_0 \leq L_R$ and L_R is the sparsity level. For instance, most time-domain signals which pass through a mmWave MIMO link have are sparse in the frequency domain (i.e., when $\boldsymbol{\Psi}$ is a DFT matrix) [7]. It is known that frequency domain transforms (e.g., DFT, DCT), compact the signal energy in a small set of frequencies, thereby resulting in a sparse signal; see examples in [8]. We validate this insight in Section V. Letting $Q \triangleq rK_R$ be the total number of measurements, we rewrite (2) as

$$\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{s} + \tilde{\boldsymbol{n}}, \tag{3}$$

where $\boldsymbol{z} \triangleq [\boldsymbol{y}_D^{(1)^T}, \cdots, \boldsymbol{y}_D^{(K_R)^T}]^T$ denotes the aggregate $Q\times 1$ measurement vector, $\tilde{\boldsymbol{n}} = [\tilde{\boldsymbol{n}}_R^{(1)^T}, \cdots, \tilde{\boldsymbol{n}}_R^{(K_R)^T}]^T$ the effective $Q\times 1$ noise vector, $\boldsymbol{\Phi} = [(\boldsymbol{W}^{(1)^\dagger})^T, \cdots, (\boldsymbol{W}^{(K_R)^\dagger})^T]^T$ the $Q \times N$ is the sensing matrix. Thus, the sparse signal can be recovered using the Basis Pursuit with Inequality Constraints:
(BPIC) : $\boldsymbol{s}_{\text{BPIC}} = \underset{\boldsymbol{s}}{\text{argmin}}\ \|\boldsymbol{\Psi}\boldsymbol{s}\|_1$ s. t. $\|\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{s}\|_2 \leq \epsilon$.
Note that, the number of measurements, Q, increases with r and K_R. While the number of RF chains, r, corresponds to the implementation complexity, the number of transmissions, K_R, denotes the signaling complexity: Evidently, lower values are preferred, i.e., $Q = rK_R \ll N$.

III. Proposed Method for Recovery

We first employ known recovery methods in the CS literature, namely, recovery from Random Frequency Measurements (RFM), recovery using Subsampled Incoherent Bases (SIB), and recovery from Bernoulli Random Matrix (RM). They are also adapted to a regime compatible with mmWave assumptions, where N, M can be arbitrarily large but r, K_R, L are small; see Section IV for a discussion of their application to the fully connected and the electromagnetic lens architecture.

A. Recovery from Random Frequency Measurements

This method is based on the seminal work of Candes [11] - adapted to the problem in question - to obtain non-uniform recovery guarantees, which provide recovery probability for the sparse signal $\boldsymbol{s}$. Let $\mathcal{R}$ be a set of indexes selected uniformly at random from $\{N\}$, where $\mathcal{R} \subseteq \{N\}$, and $|\mathcal{R}| = Q$. Consider the following method where each measurement is a row picked at random from a $N \times N$ DFT matrix, $\boldsymbol{D}_N$. Formally, the quantities in (3) are defined as: $\boldsymbol{\Phi} \triangleq Q^{-1/2}[\boldsymbol{D}_N^{\dagger}]_{\mathcal{R}}$, and $\tilde{\boldsymbol{n}} \triangleq \boldsymbol{\Phi}\boldsymbol{n}$. It is simple to verify that $\tilde{\boldsymbol{n}}$ is zero-mean AWGN, with $\tilde{\sigma}^2 = \mathbb{E}[\|\tilde{\boldsymbol{n}}\|_2^2] = N\sigma_R^2/Q$, and $\mathbb{E}[\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}^{(r)^{\dagger}}] = (N\sigma_R^2/Q)\boldsymbol{I}_Q$. Thus, $\boldsymbol{s}$ is recovered by solving the known LASSO [11],

$$\boldsymbol{s}_{\text{LASSO}} = \underset{\boldsymbol{s}}{\operatorname{argmin}} \ \|\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{s} - \boldsymbol{z}\|_2^2 + \lambda\tilde{\sigma}^{(r)}\|\boldsymbol{\Psi}\boldsymbol{s}\|_1.$$

The recovery probability, γ, is thus related to Q as follows.

Proposition 1 (Random Frequency Measurements (RFM))**.** *Given an arbitrary fixed sparse signal $\boldsymbol{s}$, pick the number of measurements Q as,*

$$Q \geq c(1+\beta)L_R \log(N) \tag{4}$$

Then with probability at least

$$\gamma \triangleq 1 - 6(N^{-1} + e^{-\beta}) \tag{5}$$

the solution of the LASSO with $\lambda = 10\sqrt{\log N}$, is such that $\|\boldsymbol{s}^{\star} - \boldsymbol{s}_{\text{LASSO}}\|_2^2 \leq$ $\text{polylog}(N)\frac{L_R}{N}\tilde{\sigma}^2$ where c is a positive numerical constant.

Refer to [11][Theorem 1.2].

Given Proposition 1, we substitute β in (4) with its expression. After simple manipulations we write (5) as,

$$\beta = -\log(N(1-\gamma) - 6) + \log(6N) \geq \log(6N) \geq \log(N)$$

We then plug the above lowerbound on β to derive another lowerbound on Q in (4),

$$Q \geq c(1+\log(N))L_R \log(N) \approx cL_R \log^2(N) \triangleq T_{\text{RFM}} \tag{6}$$

B. Recovery from Subsampled Incoherent Bases

Another design paradigm in CS is to jointly design $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ to be maximally incoherent, by minimizing their mutual coherence $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi})$ [12]. Note that, the minimum value of $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi})$ is achieved when $\boldsymbol{\Psi}$ is an orthonormal DFT basis, and $\boldsymbol{\Phi}$ is the standard basis [8].

Proposition 2 (Subsampled Incoherent Bases)**.** *Let $\mathcal{R}$ be a uniformly randomly selected subset of $\{N\}$, where $\mathcal{R} \subseteq \{N\}$ and $|\mathcal{R}| = Q$. We subsample $\boldsymbol{\Phi}$ as follows: $\boldsymbol{\Phi} \triangleq [\boldsymbol{I}_N]_{\mathcal{R}}$. Given constants γ and δ, with $0 \leq \gamma, \delta \leq 1$ and let Q be such that,*

$$Q \geq C\delta^{-2}L_R \max\left(\log^3(L_R)\log(N), \log(\gamma^{-1})\right) \triangleq T_{\text{SIB}}$$

where C is independent of all other constants. Thus, the sparse signal can be recovered from the BPIC solution, with probability greater than γ.

Refer to [8] for proof.

Note that for a high recovery probability ($\gamma \rightarrow 1$), the minimum number of measurements becomes,

$$T_{\text{SIB}} \approx C\delta^{-2}L_R \log^3(L_R)\log(N) \tag{7}$$

Note that, algorithmically, SIB and RFM are equivalent (implemented using randomly selected DFT measurements). However, the result in differing minimum number of measurements, which is the main interest of this work.

C. Recovery from Random Matrices (RM)

We use another recovery strategy based on random matrices. While the previous method ensures that $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are incoherent, random constructions of $\boldsymbol{\Psi}$ result in low values of $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi})$. For instance, when the entries in $\boldsymbol{\Phi}$ are independent identically distributed Rademacher, then $\boldsymbol{\Phi}$ is incoherent with any basis $\boldsymbol{\Psi}$, with high probability. These constructions are desirable as they imply that the sparsity basis need not be known.

Proposition 3 (Rademacher Matrices (RM) [13])**.** *Let $\boldsymbol{\Phi}$ in (3) denote a normalized Rademacher matrices, i.e., $\mathbb{P}\left[[\boldsymbol{\Phi}]_{k,j} = \frac{+1}{\sqrt{Q}}\right] = \mathbb{P}\left[[\boldsymbol{\Phi}]_{k,j} = \frac{-1}{\sqrt{Q}}\right] = \frac{1}{2}$. Moreover, the number of measurements Q satisfies,*

$$Q \geq C_\beta \delta^{-2}\left(L_R \log(N/L_R) + \log(\gamma^{-1})\right) \triangleq T_{\text{RM}}$$

Thus, with probability greater than γ, the BPIC solution is such that $\|\boldsymbol{s}^{\star} - \boldsymbol{s}_{BPIC}\|_2 \leq a_1/\sqrt{Q}\|\boldsymbol{s}^{\star} - \boldsymbol{s}_L\|_1 + a_2$, where $\boldsymbol{s}_L$ is the best L-sparse approximation for $\boldsymbol{s}^{\star}$.

Refer to [13] for proof

For $\gamma \rightarrow 1$ (recovery w.h.p.), the minimum number of measurements for this method is approximated as,

$$T_{\text{RM}} \approx C_\beta \delta^{-2}L_R \log(N/L_R) \tag{8}$$

It becomes clear at this stage that recovery with Radamacher matrices requires the minimum number of samples, among all others. Although this particular method is more complex, it can be implemented with analog hardware by randomly setting the phase of each element in $\boldsymbol{W}$ to $\{+\pi, -\pi\}$, and normalizing the output of the combiner by $1/\sqrt{Q}$.

D. Summary of Approach

We summarize the main steps of our approach. In the DL, the transmitter retransmits the (precoded) signal K_R times, which is combined with a different analog filter (by activating all r RF chains each transmission). The receiver constructs the components of the measurement model in (3), and employs any of the above methods to recover signal at its antennas, $\boldsymbol{y}_A$. Focusing on the ''best recovery" method, RM, hereafter, the minimum number of measurements is given by,

$$T_{\text{RM}} \triangleq rK_R = C_R L_R \log(N/L_R) \tag{9}$$

The same CS based recovery method can be applied to an uplink communication leading to the minimum number of measurements at the transmitter, U_{RM},

$$U_{\text{RM}} \triangleq rK_T = C_T L_T \log(M/L_T) \tag{10}$$

where K_T and L_T are the number of retransmissions and sparsity level of the signal, in the uplink, respectively.

IV. Application to MmWave Array Architecture

A. MmWave Array Architectures

We discuss the practical issues for implementing these recovery methods, for widely used variants of the architecture.

Implementation in the Fully Connected Hybrid Architecture: RFM is particularly well suited for a direct implementation to the fully connected hybrid architecture, where every RF chain is connected to all antennas. After generating the random sequence $\mathcal{R}$, the receiver successively selects columns of analog combiner from columns of a DFT matrix indexed by $\mathcal{R}$ (r columns at a time): $\boldsymbol{W}^{(l)} =$

$1/\sqrt{Q}\left[\boldsymbol{d}_{(l-1)r+1},\cdots,\boldsymbol{d}_{lr}\right]$, $\forall l \in \{1,\cdots,K_R\}$. Evidently, the constant modulus constraint on $\boldsymbol{W}$ is satisfied.

Implementation in the EM Lens Array Architecture: At the receiver, the EM lens is modeled as a cascade of a codebook ($N \times N$ DFT matrix), $\boldsymbol{D}_R$, followed by a selecion matrix, $\boldsymbol{S}_R$, [14]. Note that SIB is directly applicable to this architecture: The codebook acts as a sparsifying basis, i.e., $\boldsymbol{D}_R = \boldsymbol{\Psi}$. Moreover, the selection matrix, $\boldsymbol{S}_R$, consists of randomly selected columns from an identity matrix, $\boldsymbol{I}_N$ (chosen successively from random sequence, $\mathcal{R}$, r at a time). Formally speaking, $\boldsymbol{y}_D = \boldsymbol{S}_R^\dagger \boldsymbol{D}_R^\dagger \boldsymbol{y}_A$.

Other implementation issues: We note that all recovery methods involved a randomization step. This is easily implemented with pseudo-random number generators, which are widely used in CDMA.

B. Overhead Reduction for Subspace Estimation

In our earlier work [4], we proposed a Subspace Estimation and Decomposition (SED) algorithm, a used retransmission scheme (similar to that of Fig. 1) that allowed its operation in a hybrid MIMO architecture. Moreover, quantified the resulting communication (signaling) overhead: $\Omega_{SED} \propto (M+N)/r$. However, when employing the proposed CS-based recovery in conjunction with SED, the communication overhead reduces to $\Omega_{CS+SED} \propto (\ L_T \log(M/L_T) + L_R \log(N/L_R)\ )$. Thus the overhead ratio η is given by

$\eta = (\ L_T \log(M/L_T) + L_R \log(N/L_R)\ )r/(M+N)$. Moreover, as N, M grow large

$\eta \to (L_T \log M + L_R \log N\ )r/(M+N) \to 0$.

This implies that the overhead ratio resulting from using CS (in conjunction with SED), asymptotically decays to zero, as $N, M \to \infty$. Here we assume that the sparsity levels of signals, L_R, L_T, are fixed as N, M tend to large. This assumption is reasonable since L_T, L_R depend on the channel (number of paths, frequency, etc), irrespective of N, M. These results are thus relevant of mmWave MIMO.

C. Implications of Proposed Approach

Flexible Resource Allocation: The approach provides us flexibility to trade-off communication overhead with complexity (more RF chains). We exemplify with the two extremes:

One-shot recovery. A special case where the number of RF chains is large enough such that one transmission is needed (no repetitions, $K_R = K_T = 1$). This may correspond to a cellular uplink communication, where the increased number of RF chains can be handled at the BS. Then the conditions in (10), (9) imply that the minimum number of RF chains at the BS and MS are $C_T L_T \log(M/L_T)$ and $C_R L_R \log(N/L_R)$, receptively. Interestingly, this also implies that a hybrid MIMO link with $\mathcal{O}(\log(N))$ (resp. $\mathcal{O}(\log(M))$) RF chains at the receiver (resp transmitter) is equivalent to a fully one, as $N, M \to \infty$.

Single RF chain Recovery. When complexity has to remain at a minimal level (single RF chain receiver) can be tradedoff for communication overhead. Setting $r = 1$ and rewriting (10) and (9), we derive a lowerbound on the communication overhead (number of retransmissions), needed for successful

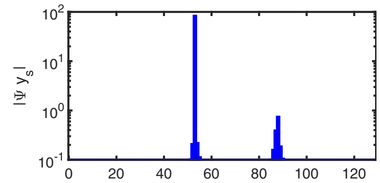

Fig. 2. Magnitude of entries in the sparse signal,$|\boldsymbol{\Psi s}|$

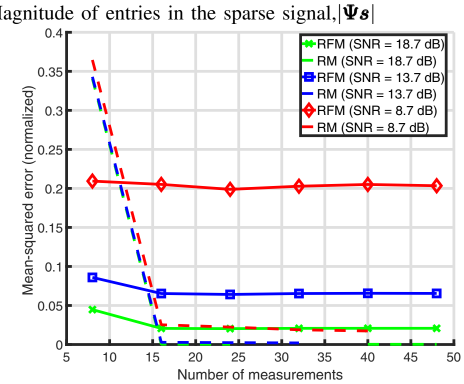

Fig. 3. Normalized MSE as a function of Q, for low/medium/high SNR

recovery of the analog signal as: $K_R = C_R L_R \log(N/L_R)$, $K_T = C L_T \log M$.

Turning a hybrid MIMO into fully-digital MIMO : The proposed approach may be used to bypass the lack of access to analog signal at the BS/MS antennas. Such a limitation is the reason that that many classical DSP techniques in MIMO (e.g., pilot-based channel estimation, fully digital precoding/combining), are not applicable to mmWave MIMO. The retransmission scheme in [4] (see Fig. 1) allowed us to recover the full (high-dimensional) signal at the receiver, with a overhead cost of N/r channel uses. With proposed method, the number of channel uses from retransmissions for DL communication become $K_R = C L_R \log(N/L_R)/r$. Thus the 'cost' of turning a hybrid MIMO link into a fully digital one is drastically reduced by using the proposed approach especially as $N \gg 1$ (which is the case in mmWave systems).

V. Numerical Results

We numerically evaluate our approach for a hybrid MIMO system, with $M = N = 128$, beamforming along the right singular vector, and no retransmissions ($Q = r$). We fix the signal at the MS and average over random realizations of $\boldsymbol{\Phi}$, for each value of Q. Fig 2 shows the magnitude of the entries in MS signal, $\boldsymbol{s}$, after applying $\boldsymbol{\Psi}$: it clearly validates our intuition as $\boldsymbol{\Psi s}$ has a sparsity level of $L_R \leq 10$. We compare the Normalized MSE (NMSE) for the RFM and RM methods for various SNR values. We observe in Fig 3 that both methods can successfully recover $\boldsymbol{s}$(with NMSE $\leq .07$) from a relatively low number of measurements, $Q \leq 16$, in medium/high SNR setting. RM is able to recover the 128-dimensional signal extremely accurately, with as little as 16 measurements, across a range of worsening SNRs. This implies that a hybrid MIMO system using RM with $Q = 16$ measurements (RF chains) can almost perfectly mimic its fully digital 128×128 MIMO.

References

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

[2] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4391–4403, 2013.

[3] R. W. Heath, N. Gonzlez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 436–453, April 2016.

[4] H. Ghauch, T. Kim, M. Bengtsson, and M. Skoglund, "Subspace estimation and decomposition for large millimeter-wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 528–542, April 2016.

[5] X. Yu, J. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 485–500, April 2016.

[6] W. Zhang, T. Kim, D. J. Love, and E. Perrins, "Leveraging the restricted isometry property: Improved low-rank subspace decomposition for hybrid millimeter-wave systems," *IEEE Transactions on Communications*, vol. 1, pp. 1–14, July 2018.

[7] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications,*, vol. 13, pp. 1499–1513, March 2014.

[8] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Transactions on Signal Processing*, vol. 59, pp. 4053–4085, Sept 2011.

[9] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Transactions on Information Theory*, vol. 56, pp. 520–544, Jan 2010.

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

[11] E. J. Candes and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, pp. 7235–7254, Nov 2011.

[12] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, pp. 2231–2242, Oct 2004.

[13] F. Krahmer and H. Rauhut, "Structured random measurements in signal processing," *CoRR*, vol. abs/1401.1106, 2014.

[14] A. Sayeed and N. Behdad, "Continuous aperture phased MIMO: Basic theory and applications," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 1196–1203, Sept 2010.