Scalable Nonlinear Multiuser Detection for mmWave Massive MIMO

Mohammed Abdelghany, Maryam Eslami Rasekh, *Student, IEEE*, and Upamanyu Madhow, *Fellow, IEEE*

***Abstract*— We consider uplink multi-user reception for a mmWave picocell with a large number of base station antennas and a proportionally large number of users. Receiver complexity can become a bottleneck for scaling such a system, necessitating development of more computationally efficient algorithms. Due to the spatial sparsity of mmWave channels, transforming the received signal into beamspace concentrates the power of each user into a small number of dimensions. Our prior work takes advantage of this fact to develop a *local* LMMSE estimator that relies on a small window of the beamspace signal for demodulating each user's signal. In this paper, we propose to layer nonlinear interference cancellation on top of the local LMMSE receiver: while most interference is suppressed linearly by the local LMMSE receiver, for each user, a small number of strongly interfering users are handled by interference cancellation after suitably whitening the local LMMSE output. The method provides reliable demodulation at higher load factor (defined as number of users divided by the number of antennas) than enabled by linear interference suppression alone, at order of magnitude lower complexity than standard interference cancellation.**

***Keywords*—Low-complexity multiuser detection, beamspace, successive interference cancellation, local LMMSE.**

I. Introduction

As millimeter wave (mmWave) technology advances, fully-digital front ends are becoming feasible, allowing massive base station arrays to communicate with tens or hundreds of users simultaneously using multiuser MIMO. We investigate here how to address the bottleneck in receiver complexity as the system grows in the number of array elements, N and number of simultaneous users, K, fixing the *load factor* $\beta = \frac{K}{N}$. For sparse mmWave channels, our prior work shows that it is possible to exploit the concentration of each user's energy in beamspace (i.e., after taking the discrete Fourier transform across antennas) for linear interference suppression. Specifically, we show in [1] that effective linear interference suppression is provided by the "local LMMSE" receiver which demodulates each user from a small window of the signal vector in beamspace. The required size of the window depends on the load factor and the minimum user separation, but does not scale with N. In this paper, we show that low-complexity nonlinear interference cancellation can be layered on top of such a local LMMSE receiver, enabling reliable demodulation at higher load factors.

The key idea is as follows. For any given "desired" user, the local LMMSE receiver effectively suppresses most of the interference except for a small number of users which are "nearby" (in beamspace). Therefore, the output of the local LMMSE receiver can be treated as a virtual MIMO system with a small number of users and colored noise. After noise whitening, we apply interference cancellation to this smaller system to enhance the reliability of demodulation for the desired user. We apply this second stage of interference cancellation for each user separately. The parameters of our approach are the size of the window used for local LMMSE reception in beamspace, and the maximum number of interferers cancelled for each user in the second stage. Our approach retains the gains in computational efficiency obtained from the sparsity of the mmWave channel, while allowing us to push the system to higher load factors than is possible with linear interference suppression.

Related Work: Multiuser detection (MUD) for MIMO has a rich history [2], with many recent works focussing on complexity reduction motivated by massive MIMO. Matrix inversion in high dimensions is a bottleneck for linear interference suppression, and proposed complexity reduction techniques include Newton iteration [3], Neumann series expansion (NSE) [4], the Gauss-Seidel method [5], [6], and Cholesky decomposition [7]. The correlation structure of the matrix can be further exploited to reduce complexity. For example, [8] exploits a tridiagonal structure for the Wishart matrix in a VLSI implementation. Complexity reduction techniques for nonlinear MUD include [9], wherein a sphere decoder is selectively applied by leveraging a linear detector, and [10], where approximate message passing is applied to reduce sphere decoding complexity. A survey of massive MIMO detection techniques can be found in [11]. Unfortunately, these and other existing techniques are not easily scaled up to the system sizes that we consider here.

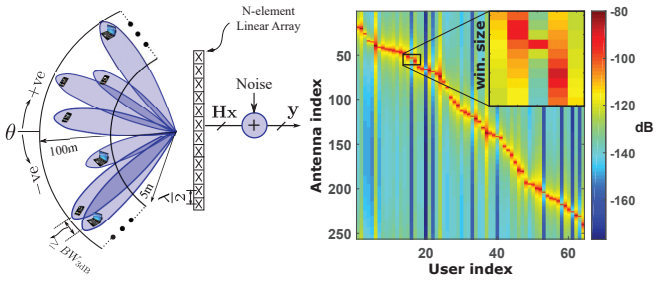
In our prior work [1], we show that, as long as we pay the $\mathcal{O}(N \log N)$ price of a spatial fast Fourier transform (FFT), linear interference suppression for each user can be accomplished at complexity that does not scale with system size, assuming that the mmWave channel is concentrated in beamspace. In the present paper, we show that similar conclusions hold for nonlinear multiuser detection as well, enabling us to push the system load factor further up without sactificing link reliability.

Notation: We use lowercase bold letters for vectors, and uppercase bold letters for matrices. The notation $\mathbf{x} = [x_i]_{i=1}^{I}$ represents column vector $\mathbf{x}$ of length I and its elements are denoted by x_i. For a matrix, we use $\mathbf{X} = [x_{i,j}]_{i=1,j=1}^{I,J}$. $\{.\}_{k=1}^{K}$ denotes a list of K scalars, vectors or matrices. The identity matrix is denoted by $\mathbf{I}$ and $\mathbf{0}_M$ is a column vector which consists of M zeros.

II. System Model

We consider the uplink MIMO system depicted in Fig. 1 (a). The base station employs a linear array with N elements to simultaneously serve $K = \beta N$ mobile users, where β is the system load factor. Each mobile transmits a single data

M. Abdelghany, M. E. Rasekh, and U. Madhow are with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: mabdelghany@ucsb.edu; rasekh@ece.ucsb.edu; madhow@ece.ucsb.edu).



(a) MIMO system model. (b) mmWave beamspace channel matrix.

Figure 1. (a) Uplink massive MIMO system model. (b) The sparsity of single-path channel in beamspace.

stream and uses an antenna array to perform ideal transmit beamforming towards the base station.

Channel Model: We assume that a single path dominates the channel between the base station and any mobile. Such a model is well suited for mmWave channels as it has been experimentally validated in a typical university campus at 60 GHz [12]. Therefore, the channel matrix is of the form $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \dots \mathbf{h}_K]$, where $\mathbf{h}_k$ is the $N \times 1$ spatial channel for the k^{th} mobile, $\mathbf{h}_k = \alpha_k \left[1\, e^{j\Omega_k}\, e^{j2\Omega_k} \dots e^{j(N-1)\Omega_k}\right]^{\mathsf{T}}$. Here, Ω_k is the spatial frequency (corresponding to the angle of arrival) and α_k is the complex channel amplitude of the path of user k.

The single path channel has a concentrated structure in the discrete spatial frequency domain, or "beamspace", as shown for a typical example in Fig. 1 (b). We define the beamspace channel matrix as $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1 \dots \tilde{\mathbf{h}}_K]$, where $\tilde{\mathbf{h}}_k = \mathcal{DFT}(\mathbf{h}_k)$ and $\mathcal{DFT}(.)$ is the discrete Fourier transform (DFT) operator. The received signal vector in the original spatial domain is given by $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{x}$ is the $K \times 1$ vector of users' symbols, $\mathbb{E}(|x_k|^2) = 1$, and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is additive white Gaussian noise (AWGN). By taking the DFT of this vector, we obtain the beamspace signal model, $\tilde{\mathbf{y}} = \tilde{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{n}}$, where $\tilde{\mathbf{n}} = \mathcal{DFT}(\mathbf{n}) \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$. Our goal is to perform MUD by estimating the transmitted symbol vector, $\mathbf{x}$, from this beamspace signal, which is a sufficient statistic for estimating $\mathbf{x}$. We describe in the following section some conventional approaches for MUD.

III. Conventional MIMO Detectors

In this paper, we assume that channel estimation is error-free and focus our discussion on the MUD, which consists of two blocks: a preprocessor and a demodulator. Based on the estimated channel matrix and noise variance, the preprocessor provides the demodulator with filters and parameters required to decode users' symbols from the received signal vector. Conventional multiuser reception techniques include the following.

LMMSE reception is the optimal *linear* MUD method. It provides the best combination of zero-forcing interference suppression and matched filtering which are optimal at high and low SNR, respectively. The LMMSE beamformer estimates the vector of transmitted symbols as $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$, with the optimal receive matrix, W, which is computed by the preprocessor based on channel state information as

$$\mathbf{W} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + \sigma^2\mathbf{I})^{-1} = (\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I})^{-1}\mathbf{H}^H.$$

It should be noted that the second equality is more computationally efficient to calculate as it requires inverting a smaller matrix (since $K \leq N$). The computational complexity of LMMSE is $\mathcal{O}(\beta N^2)$ for beamforming and $\mathcal{O}(\beta^2 N^3)$ for computing $\mathbf{W}$, which does not scale well with the number of antennas. Furthermore, even though it provides near-optimal performance in low-loaded systems, its performance diminishes once the load factor, β, exceeds $\frac{1}{4}$, especially when the near-far power disparity is large or some channels are very close in spatial frequency. In these conditions, nonlinear techniques can provide significant performance gains.

Interference cancellation (IC) is the most intuitive and well-known nonlinear MUD method. After decoding a user's *digital* symbol, this receiver calculates the interference of that user on other channels and subtracts it from the original observations before a second demodulation. This can be done successively (SIC) [13], starting from the strongest user to the weakest, or in parallel (PIC) [14]. The former is advantageous in terms of performance, especially when variation in channel strength is large among users, but entails higher delay and is not easily parallelizable. An efficient implementation of SIC is the V-BLAST algorithm that admits a complexity of $\mathcal{O}(N^3)$ [15].

In the next section, we describe our proposed MUD which combines linear techniques in beamspace with nonlinear interference cancellation to provide a scalable receiver design.

IV. Scalalbe Nonlinear Multiuser Detection

The beamspace local LMMSE receiver developed in our prior work [1] takes advantage of channel concentration in spatial frequency domain to perform linear estimation of each user's symbol using a small window of the beamspace signal vector, $\tilde{\mathbf{y}}$. This windowing approach significantly reduces the computational burden of the linear detector; however, the limited dimensionality of observations limits its interference suppression capability, and linear techniques, in general, are very suboptimal at high load factor or when trying to detect users that are close in spatial frequency and have highly correlated channels. Nonlinear techniques are effective in these cases, but their complexity can become a bottleneck for massive MIMO systems. To facilitate a scalable nonlinear detector, we augment the beamspace local LMMSE receiver with a user-centric *virtual MIMO system* that models the cross-interaction of nearby neighbors in beamspace. Nonlinear detection is possible on this smaller virtual system, especially as the number of significant interferers for any given user remains relatively constant as the system is scaled up in size. In this section, we describe the stages of this proposed approach and determine the computational complexity of each stage.

A. Local LMMSE

The initial local LMMSE stage (summarized in Algorithm 1) carries out a lightweight linear estimation of users' symbols by transferring the received signal vector to beamspace via an FFT operation, and then limiting the observation window used for the m^{th} user to a small number (W_1) of FFT bins around the m^{th} user's spatial frequency. This window is chosen for each user such that the resulting mean squared error (MSE) is minimized, as described in Algorithm 1. Using these limited

dimensions, the local LMMSE receiver suppresses interference produced by other users via linear projection, and provides the *local LMMSE estimate*,

$$\bar{x}_m = \mathbf{w}_m^H \tilde{\mathbf{y}}, \tag{1}$$

where $\mathbf{w}_m$ is the local LMMSE filter for the m^{th} user (obtained by Algorithm 1) which contains W_1 nonzero entries. It is worth noting that there are only W_1 nonzero complex multiplication operations in (1). The estimates $\{\bar{x}_k\}_{k=1}^K$ serve as observations for the next stage of processing.

B. User-centric whitened virtual MIMO

In the second stage, we create a small virtual MIMO system in order to obtain a better estimate of its symbol as follows. The virtual MIMO system for user m is obtained by taking the set of $W_2 - 1$ nearest users (in beamspace) and forming the set $\mathcal{J}_m$ of users, as shown in Fig. 2. The measurement vector for this system is denoted as

$$\begin{aligned}\mathbf{z}_m &= [\bar{x}_k]_{k\in\mathcal{J}_m} = \dot{\mathbf{W}}_m^H(\tilde{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{n}}) \\ &= \underbrace{\dot{\mathbf{W}}_m^H \tilde{\mathbf{H}}_{\mathcal{J}_m}}_{\mathbf{B}_m} \mathbf{x}_{\mathcal{J}_m} + \underbrace{\dot{\mathbf{W}}_m^H(\tilde{\mathbf{H}}_{\mathcal{J}_m^c}\mathbf{x}_{\mathcal{J}_m^c} + \tilde{\mathbf{n}})}_{\mathbf{i}_m},\end{aligned} \tag{2}$$

where $\dot{\mathbf{W}}_m = [\mathbf{w}_k]_{k\in\mathcal{J}_m}$, $\tilde{\mathbf{H}}_{\mathcal{J}_m} = [\tilde{\mathbf{h}}_\ell]_{\ell\in\mathcal{J}_m}$, $\tilde{\mathbf{H}}_{\mathcal{J}_m^c} = [\tilde{\mathbf{h}}_\ell]_{\ell\notin\mathcal{J}_m}$, $\mathbf{x}_{\mathcal{J}_m} = [x_\ell]_{\ell\in\mathcal{J}_m}$, and $\mathbf{x}_{\mathcal{J}_m^c} = [x_\ell]_{\ell\notin\mathcal{J}_m}$. We treat the interference from users that are not in $\mathcal{J}_m$ as noise, and compute the overall noise covariance matrix of (2) as

$$\mathbf{\Sigma}_m = \mathbb{E}[\mathbf{i}_m \mathbf{i}_m^H] = \dot{\mathbf{W}}_m^H(\tilde{\mathbf{H}}_{\mathcal{J}_m^c}\tilde{\mathbf{H}}_{\mathcal{J}_m^c}^H + \sigma^2\mathbf{I})\dot{\mathbf{W}}_m. \tag{3}$$

Since the FFT taps used for different users in $\mathcal{J}_m$ are likely to overlap, this distortion is "colored". We whiten each virtual MIMO system by computing

$$\bar{\mathbf{z}}_m = \mathbf{\Sigma}_m^{-\frac{1}{2}} \mathbf{z}_m.$$

The whitening filter can be computed efficiently using a Cholesky decomposition [16]. This yields the following model for the m^{th} whitened virtual MIMO system:

$$\bar{\mathbf{z}}_m = \mathbf{A}_m \mathbf{x}_{\mathcal{J}_m} + \mathbf{n}_m, \tag{4}$$

where the effective channel seen by the users in $\mathcal{J}_m$ becomes

$$\mathbf{A}_m = \mathbf{\Sigma}_m^{-\frac{1}{2}} \mathbf{B}_m,$$

and the "noise" (which includes interference due to users in $\mathcal{J}_m^c$) in the virtual MIMO system is white: $\mathbf{n}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.

C. Nonlinear MUD for the virtual MIMO systems

We may now apply any nonlinear MUD to the W_2-dimensional virtual MIMO system centered at user m to get an estimate of its symbol, x_m. We report results for LMMSE-SIC operating on each virtual MIMO system. For simplicity, we describe it for a generic whitened virtual MIMO system of the form:

$$\bar{\mathbf{z}} = \mathbf{A}\mathbf{x} + \mathbf{n},$$

where we drop the subscripts from (4), and number the users from 1 to W_2.

The SIC demodulator consists of W_2 successive stages. In the first stage, it receives the observation vector $\bar{\mathbf{z}}$ and linearly

Algorithm 1 Local LMMSE Preprocessing.

Input: $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_k]_{k=1}^K \in \mathcal{C}^{N\times K}$, σ^2, N, and m
Output: $\mathbf{w}_m$
Parameter: W_1 (window size)

1: set $\mathbf{G} = \tilde{\mathbf{H}}\tilde{\mathbf{H}}^H + \sigma^2\mathbf{I}$ {Compute the covariance matrix}
2: **for** $i = 1$ **to** $N - W_1 + 1$ **do**
3: set $\mathbf{R}_i = \left[G_{\ell,n}\right]_{\ell=i,n=i}^{i+W_1-1,i+W_1-1}$
4: set $\dot{\mathbf{h}}_i = \left[\tilde{H}_{\ell,m}\right]_{\ell=i}^{i+W_1-1}$
5: set $MSE_i = 1 - \dot{\mathbf{h}}_i^H \mathbf{R}_i^{-1} \dot{\mathbf{h}}_i$
6: **end for**
7: set $i^* = \arg\min_i MSE_i$ {Optimal window location}
8: set $\mathbf{w}_m^H = [\mathbf{0}_{i^*-1}^\intercal, \dot{\mathbf{h}}_{i^*}^H \mathbf{R}_{i^*}^{-1}, \mathbf{0}_{N-i^*-W_1+1}^\intercal]$
9: set $\mathbf{w}_m \leftarrow \frac{\mathbf{w}_m}{\mathbf{w}_m^H \tilde{\mathbf{h}}_m}$ {Remove the estimation bias}

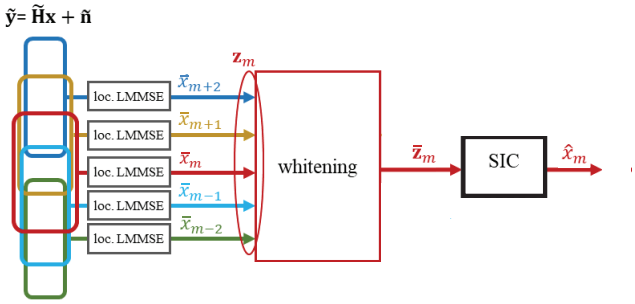


Figure 2. Proposed MUD scheme for one virtual MIMO system.

projects it on $\mathbf{v}_1$ (see the description of the preprocessing in Algorithm 2) to decode user ℓ_1's symbol, i.e., $\dot{x}_{\ell_1} = \mathbf{v}_1^H \bar{\mathbf{z}}$. Then, using a constellation demapper, the demodulator retrieves the original constellation symbol $\hat{x}_{\ell_1}$ from the estimate $\dot{x}_{\ell_1}$. The SIC then subtracts its effect from the observation vector to get $\bar{\mathbf{z}}^{(1)} = \bar{\mathbf{z}} - [A_{n,\ell_1}]_{n=1}^{W_2} \hat{x}_{\ell_1}$. In the next step, the same process is applied on $\bar{\mathbf{z}}^{(1)}$ to decode user ℓ_2's symbol, and so on.

The nonlinear demodulator needs the order of users $\mathcal{L} = \{\ell_k\}_{k=1}^{W_2}$ and the projection vectors $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^{W_2}$ from the preprocessing step. As shown in algorithm 2, preprocessing starts by computing the MSE of each user's estimate (step 3), picks the user with the highest SINR (step 4) and computes its projection vector (step 5-6). It then removes that user's channel vector from the channel matrix $\mathbf{A}$ (step 8). The acquisition repeats this procedure W_2 times until it completely computes $\mathcal{L}$ and $\mathcal{V}$.

D. Computational complexity

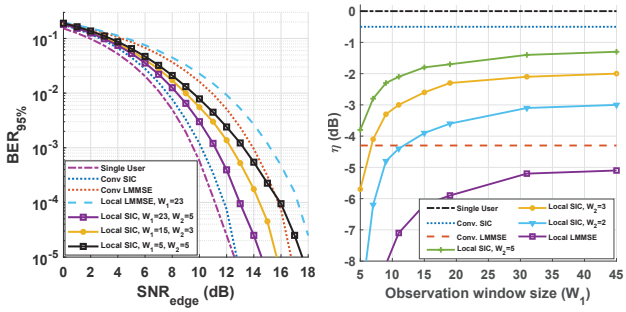
Algorithm 1 describes the preprocessing stage of the local LMMSE block. The algorithm starts with computing the covariance matrix $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H + \sigma^2\mathbf{I}$ (step 1), and then searches for the location of the optimum observation window to minimize the MSE for each user (steps 2-7). The algorithm then forms the local LMMSE projection vector (steps 8-9).

The complexity of the local LMMSE beamformer is $\mathcal{O}(\beta W_1 N)$ for demodulation (performed on a per-symbol

Algorithm 2 SIC Preprocessing.

Input: $\mathbf{A} \in \mathcal{C}^{W_2 \times W_2}$, and W_2
Output: $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^{W_2}$ and $\mathcal{L} = \{\ell_k\}_{k=1}^{W_2}$

set $\mathcal{M} = \{1, 2, \ldots, W_2\}$
for $i = 1$ **to** W_2 **do**
  set $\mathbf{B} = \left(\mathbf{A}^H \mathbf{A} + \mathbf{I}\right)^{-1}$
  set $n = \arg\min_k B_{kk}$ {Find the user with minimum MSE}
  set $\mathbf{v}_i^H = [B_{nk}]_{k=1}^{W_2} \mathbf{A}^H$
  set $\mathbf{v}_i \leftarrow \frac{\mathbf{v}_i}{\mathbf{v}_i^H [A_{mn}]_{m=1}^{W_2}}$ {Remove the estimation bias}
  set $\ell_i = \mathcal{M}_n$
  set $[A_{mn}]_{m=1}^{W_2} = [\,]$ {Remove the n^{th} column}
  set $\mathcal{M}_n = [\,]$ {Remove the n^{th} entry}
end for



(a) $\beta = 1/2$. (b) Target BER 10^{-3}.

Figure 3. (a) The BER achieved by at least 95% of the users for different window sizes. (b) The efficiency of the proposed scheme relative to the conventional SIC.

scale), and $\mathcal{O}(\beta W_1 N^2)$ for preprocessing (performed on scale of channel coherence time). The most computationally expensive part of this step is computing the Gram matrix $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H$. Initially, it has a computational complexity of $\mathcal{O}(\beta N^3)$, however, the algorithm uses the elements on the matrix diagonal band only, reducing the computational complexity to $\mathcal{O}(\beta W_1 N^2)$.

The complexity of the whitening process is $\mathcal{O}(W_2^2 \beta N)$. For preprocessing, computing the overall noise covariance matrix, described in (3), dominates the computational complexity. Notice that the matrix $\dot{\mathbf{W}}_m$ has only $W_1 \times W_2$ nonzero elements. Hence, computing $\dot{\mathbf{W}}_m^H \tilde{\mathbf{H}}_{\mathcal{J}_m^c}$ and the covariance matrix $\mathbf{\Sigma}_m$ incur a computational complexity of $\mathcal{O}(W_2 W_1 \beta N)$ and $\mathcal{O}(W_2^2 \beta N)$ per virtual MIMO system, respectively. Since we have K virtual system, the total computational complexity of the whitening preprocessor becomes $\mathcal{O}(W_2(W_1+W_2)\beta^2 N^2)$. The total computational complexity of the nonlinear MUD step is $\mathcal{O}(W_2^2 \beta N)$ for detection and $\mathcal{O}(W_2^3 \beta N)$ for preprocessing.

Thus, the overall complexity of our proposed algorithm is dominated by $\sim \mathcal{O}(N \log N)$ for demodulation (which is performed on a symbol by symbol basis) and $\sim \mathcal{O}(N^2)$ for preprocessing (which is repeated on a time scale of channel coherence time), assuming window sizes are small.

V. Results

We consider the MIMO system illustrated in Fig. 1 (a) with a carrier frequency of 140 GHz. We select the number of antennas at the base station to be $N = 256$ according to the link budget calculation described in [17]. All numerical simulations are conducted at load factor $\beta = 1/2$, unless otherwise stated. The sector field of view is restricted to $-\pi/3 \leq \theta \leq \pi/3$ radians. The users are placed uniformly in the coverage area, at a distance of at least 5 m and at most 100 m from the base station.

While the user terminals are placed randomly in our simulations, we enforce a minimum separation between them in spatial frequency to avoid irrecoverable excessive interference. As shown in Fig. 1 (a), the minimum spatial frequency between any two users $\Delta\Omega_{\text{min}}$ is at least half the 3 dB beamwidth (BW_{3dB}), i.e., $\Delta\Omega_{\text{min}} = \frac{2.783}{N}$ radians [18]. We assume that the base station can serve users that are closer than this threshold in different time or frequency resource blocks.

We do not deploy any power control scheme in our simulations, and hence the near-far effect between users prevails. We assume that the base station has perfect knowledge of the channel state information (CSI). We measure link quality by the outage probability at a target uncoded BER of 10^{-3} for QPSK. This BER requires SNR of about 10 dB for a single AWGN link, which becomes the target SINR at the output of the multiuser detector for an edge user. We use the single user scenario as a benchmark, and compare between four MUD schemes: conventional LMMSE, conventional SIC, local LMMSE, and the proposed scheme, which we refer to in figures by "Local SIC."

Performance and efficiency: Fig. 3 (a) depicts the bit error rate that 95% of users in the cell achieve as a function of the SNR of the edge user, which is defined as $\text{SNR}_{\text{edge}} = \frac{N P_{\text{tx}} |\alpha_{100}|^2}{\sigma^2}$ where $|\alpha_{100}|^2$ is the free-space path loss at 100 m and P_{tx} is the transmitted power of user devices.

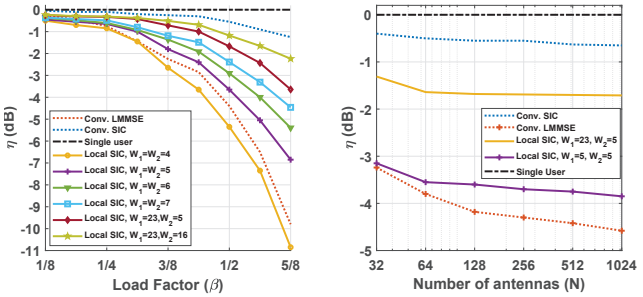
Fig. 3 (b) shows the efficiency of each MUD scheme compared to single user performance at target uncoded BER of 10^{-3}. We define the efficiency η as the ratio between the transmit power levels required for single user operation and multiuser operation (with a given MUD scheme) achieving the target BER, i.e.,

$$\eta = \frac{\text{SNR}_{\text{edge}}(\text{Single User})}{\text{SNR}_{\text{edge}}(\text{MUD})},$$

where $\text{SNR}_{\text{edge}}(\text{Single User})$ and $\text{SNR}_{\text{edge}}(\text{MUD})$ are the SNR levels required for the edge user to achieve the target BER in single user and multiuser scenarios, respectively.

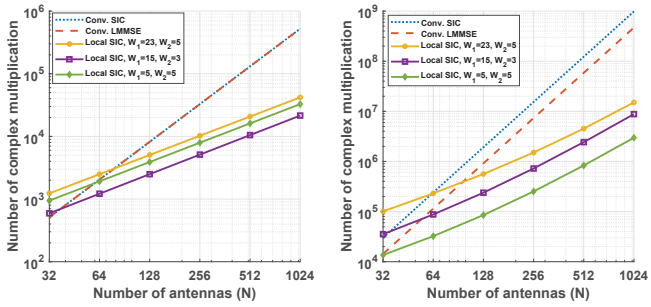
Scalability: Fig. 4 (a) depicts efficiency relative to single-user performance as a function of load factors for different MUD schemes and window sizes. The performance gap between the different MUDs and the single-user baseline increases as the load factor increases. Fig. 4 (b) reports these trends as a function of array size. It is clear that the efficiency of the proposed MUD is almost constant, regardless of the number of elements. Therefore, for maintaining the desired performance, window sizes do not need to be scaled with the number of antennas.

Computational complexity: We categorize the complexity of the MUD into beamforming and preprocessing complexities. Fig. 5 (a) and (b) demonstrate the number of complex multiplications required to carry out each MUD scheme. The FFT dominates the proposed scheme's beamforming complexity



(a) MUD efficiency versus β. (b) MUD efficiency versus N.

Figure 4. The efficiency of different configurations of the proposed MUD versus (a) the load factor and (b) the number of antenna elements.



(a) Beamformer complexity. (b) Preprocessing complexity.

Figure 5. Complexity comparison of the proposed scheme with other MUD techniques.

as $\mathcal{O}(N\log(N))$, whereas the preprocessing complexity is $\mathcal{O}(\beta W_1 N^2)$ which mostly pertains to computing the Gram matrix. At the cost of 1 dB performance degradation in efficiency, the proposed algorithm achieves savings in complexity by four and ten times compared to SIC in beamforming and preprocessing, respectively.

VI. Conclusions

The proposed nonlinear multiuser detection strategy leverages the sparsity of the mmWave channel in beamspace to accomplish drastic reductions in the complexity of both computing the receiver parameters (which remain unchanged over a channel coherence time) in *preprocessing,* and of per-symbol *demodulation.* Preprocessing complexity scales quadratically instead of cubically (which is the complexity of standard linear multiuser detection) with system size. Per-symbol complexity is dominated by the spatial FFT, and is $\mathcal{O}(N\log N)$, instead of $\mathcal{O}(\beta N^2)$ as with linear multiuser detection. The performance is close to that of standard interference cancellation with an order of magnitude lower complexity.

Acknowledgments

This work was supported in part by the Semiconductor Research Corporation (SRC) under the JUMP program (2018-JU-2778) and by DARPA (HR0011-18-3-0004). Use was made of the computational facilities administered by the Center for Scientific Computing at the CNSI and MRL (an NSF MRSEC; DMR-1720256) and purchased through NSF CNS-1725797.

References

[1] M. Abdelghany, U. Madhow, and A. Tölli, "Beamspace local LMMSE: An efficient digital backend for mmWave massive MIMO," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[2] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 1941–1988, 2015.

[3] C. Tang, C. Liu, L. Yuan, and Z. Xing, "High precision low complexity matrix inversion based on Newton iteration for data detection in the massive MIMO," *IEEE Communications Letters*, vol. 20, no. 3, pp. 490–493, 2016.

[4] F. Wang, C. Zhang, J. Yang, X. Liang, X. You, and S. Xu, "Efficient matrix inversion architecture for linear detection in massive MIMO systems," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 248–252.

[5] J. Zeng, J. Lin, and Z. Wang, "An improved Gauss-Seidel algorithm and its efficient architecture for massive MIMO systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 9, pp. 1194–1198, 2018.

[6] Z. Wu, Y. Xue, X. You, and C. Zhang, "Hardware efficient detection for massive MIMO uplink with parallel Gauss-Seidel method," in *2017 22nd International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[7] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, 2011.

[8] C. Zhang, X. Liang, Z. Wu, F. Wang, S. Zhang, Z. Zhang, and X. You, "On the low-complexity, hardware-friendly tridiagonal matrix inversion for correlated massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6272–6285, 2019.

[9] A. K. Sah and A. K. Chaturvedi, "An MMP-based approach for detection in large MIMO systems using sphere decoding," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 158–161, 2016.

[10] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902–915, 2014.

[11] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3109–3132, 2019.

[12] M. E. Rasekh, Z. Marzi, Y. Zhu, U. Madhow, and H. Zheng, "Noncoherent mmWave path tracking," in *Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*, 2017, pp. 13–18.

[13] K. A. Alnajjar, P. J. Smith, and G. K. Woodward, "Low complexity V-BLAST for massive MIMO," in *2014 Australian Communications Theory Workshop (AusCTW)*. IEEE, 2014, pp. 22–26.

[14] L. Fang, L. Xu, and D. D. Huang, "Low complexity iterative MMSE-PIC detection for medium-size massive MIMO," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 108–111, 2015.

[15] K. Pham and K. Lee, "Low-complexity SIC detection algorithms for multiple-input multiple-output systems," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4625–4633, 2015.

[16] A. Krishnamoorthy and D. Menon, "Matrix inversion using Cholesky decomposition," in *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2013, pp. 70–72.

[17] M. Abdelghany, A. A. Farid, U. Madhow, and M. J. Rodwell, "Towards all-digital mmWave massive MIMO: Designing around nonlinearities," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 1552–1557.

[18] C. A. Balanis, *Antenna Theory: Analysis and Design*. New York, NY, USA: Wiley-Interscience, 2005.