Efficient Beamspace Downlink Precoding for mmWave Massive MIMO

Mohammed Abdelghany*, Upamanyu Madhow* and Antti Tölli†

* *Department of ECE, University of California at Santa Barbara, Santa Barbara, CA 93106 USA*

† *Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014, Finland*

e-mail: mabdelghany@ucsb.edu, madhow@ece.ucsb.edu, antti.tolli@oulu.fi

***Abstract*—We investigate efficient downlink precoding for all-digital downlink mmWave massive MIMO, with the number of users scaling with the number of antennas. The iterative computations required for optimal linear precoding are a severe bottleneck as the number of antennas increases, with the computational complexity per iteration scaling cubically with the number of antennas. In this paper, we propose a near-optimal linear precoding algorithm that exploits the sparsity of mmWave channels, employing a beamspace decomposition which limits the spatial channel seen by each user to a small window which does not scale with the number of antennas. This drastically reduces the complexity of computing the precoder, with complexity per iteration scaling linearly with the number of users, and makes it feasible to scale the system up to hundreds of antennas as considered in this paper.**

***Keywords*—Low-complexity Precoding, Downlink Multiuser MIMO, Downlink Interference Suppression.**

I. Introduction

We investigate linear transmit precoding for all-digital millimeter wave (mmWave) massive MIMO cellular *downlink* with a large number N of base station antennas, and with the number of simultaneously served users K scaling with N: we set $K = \beta N$, where β is termed the *load factor.* This complements our earlier work [1]–[3] in which we explore the feasibility, efficacy, and challenges of the *uplink* in such a system. Specifically, we had shown in [2] that the signal processing for uplink receive beamforming could be vastly simplified with beamspace techniques that exploit the sparsity of the mmWave channel. In this paper, we demonstrate that beamspace techniques may have an even greater impact in terms of accomplishing downlink precoding with reasonable complexity as K and N get large. The problem of linear downlink precoding involves two tasks, power allocation subject to a total budget at the base station, and beamforming for interference suppression across users. Such power allocation is not possible on the uplink: a mobile might use power control and not use the entirety of its power budget, but it cannot transfer this power to another user. However, optimal linear downlink precoding can be mapped [4], [5] to a *virtual uplink* problem with analogous power control and beamforming steps.

Contribution: Optimal downlink precoding is typically accomplished by iterative optimization, with computational complexity scaling as $O(KN^2)$, or $O(N^3)$ in the scaling regime of interest. This is clearly infeasible for the regimes of interest to us: at mmWave frequencies, hundreds of base station antennas can be packed into compact form factors, which opens up the capability to support a correspondingly large number of simultaneous users in each base station sector using spatial multiplexing. In this paper, we propose precoding in beamspace, exploiting the sparsity of the spatial channel from the base station to each mobile user. Under our model, the channel vector for each user in beamspace spans a few spatial frequency bins, and the optimal beamformer for a given user is well approximated over a window in beamspace whose size W does not scale with the number of base station antennas. The computational complexity of the resulting algorithm is $O(KW^2)$, which is linear in the number of users/antennas, and can therefore scale to the regimes of interest to us. Our numerical results illustrate the drastic reduction in complexity, and show that, for a computational budget which yields near-optimal performance with the proposed scheme, the performance of the standard approach to computing the precoder exhibits significantly poorer performance (e.g., 6 dB worse SINR) because of the small number of iterations that can be run within that computational budget.

Related Work: The transmit precoding problem can be posed as minimizing the total transmitted power at the base station, subject to each user attaining a desired SINR. The duality between this problem and that of receive beamforming problem was pointed out in [4], [5], and used to provide an iterative algorithm that converges to the optimal solution, assuming that a feasible solution exists. Discussion of feasibility within this duality framework was included in [6]

An alternative formulation of transmit precoding is to maximize the minimum SINR across users. In this form, the problem is always feasible, and can be solved by considering fixed point iterations for normalized transmit beamforming vectors and power allocations [7]. This is the approach adopted in this paper as we seek to exploit spatial channel sparsity in beamspace.

It is worth noting that the connections between various forms of the transmit precoding problem are discussed in [8], where the authors also provide fast algorithms to approach local optima which are globally optimum under sufficiently weak interference.

Notation: We use lowercase bold letters for vectors, and uppercase bold letters for matrices. The notation $\mathbf{x} = [x_i]_{i=0}^{I}$ represents column vector $\mathbf{x}$ of length I and its elements are denoted by x_i. For a matrix, we use $\mathbf{X} = [x_{i,j}]_{i=0,j=0}^{I,J}$. If

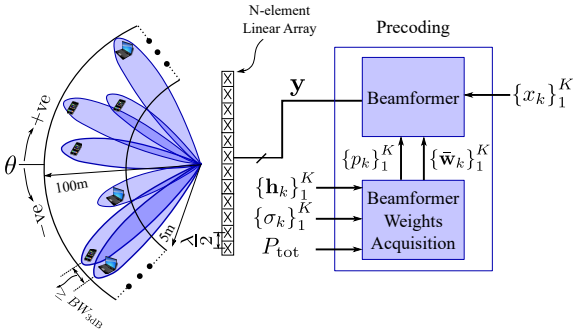


Figure 1: Downlink massive MIMO system model.

the size of the vector or the matrix can be inferred from the context, we write $\mathbf{X} = [x_{i,j}]_{i,j}$ for simplicity. $\{.\}_{k=1}^K$ denotes a list of K scalars, vectors or matrices.

II. The Downlink Precoding Problem

Consider the downlink system depicted in Fig. 1. The base station employs a linear array with N elements to simultaneously serve $K = \beta N$ mobile users. We assume that each mobile can perform ideal receive beamforming towards the base station, and include the gain due to such spatial matched filtering into the spatial channel $\mathbf{h}_k$ from the base station to mobile k, $k = 1, ..., K$.

Linear Precoding: The linear precoder at the base station allocates power p_k to mobile k, and employs beamforming direction $\{\bar{\mathbf{w}}_k\}$ (normalized to unit norm), so that the transmitted signal is given by

$$\mathbf{y} = \sum_{i=1}^{K} \bar{\mathbf{w}}_i \sqrt{p_i} x_i, \tag{1}$$

where x_k is the k^{th} user symbol. Thence, the k^{th} user's equipment receives

$$z_k = \mathbf{h}_k^H \bar{\mathbf{w}}_k \sqrt{p_k} x_k + \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{h}_k^H \bar{\mathbf{w}}_i \sqrt{p_i} x_i + n_k, \tag{2}$$

where n_k is additive white Gaussian noise (AWGN) with variance σ_k^2.

In Fig. 1, the weights acquisition block computes the power allocation and beamforming directions, given the mobile users' channel vectors, $\{\mathbf{h}_k\}$, and receiver noise variances, $\{\sigma_k^2\}$, along with the total power budget, P_{tot}. The beamformer block performs the actual precoding (1) using the computed weights.

SINR: The signal-to-interference-plus-noise ratio (SINR) of the k^{th} user is given by

$$SINR_k = \frac{|\mathbf{h}_k^H \bar{\mathbf{w}}_k|^2 p_k}{\sigma_k^2 + \sum\limits_{\substack{i=1 \\ i \neq k}}^{K} |\mathbf{h}_k^H \bar{\mathbf{w}}_i|^2 p_i}. \tag{3}$$

The SINR is a widely used performance measure because, under a Gaussian approximation for the interference-plus-noise, it provides an excellent approximation for the bit error rate (BER) (e.g., see [9] for the closely related problem of uplink multiuser detection), as well as for the achievable data rate.

Channel Model: We assume that the channel between the base station and any mobile is dominated by a single path, so that, for a linear array, the $N \times 1$ spatial channel for the k^{th} mobile is given by

$$\mathbf{h}_k = A_k \left[1 \; e^{j\Omega_k} \; e^{j2\Omega_k} \; \dots \; e^{j(N-1)\Omega_k}\right]^\intercal, \tag{4}$$

where Ω_k is the spatial frequency and $|A_k|$ the channel amplitude for the path.

Such a model is well suited for mmWave channels for several reasons:

- Typical surfaces (e.g., roads, concrete walls) look rougher at small carrier wavelengths. Hence a significant portion of the energy from a reflection is scattered. Thus, mmWave channels are typically comprised of a small number of dominant paths.
- The relative delay between different paths is large (relative to the symbol interval) for the large signaling bandwidths at mmWave bands. Gathering the energy across a large number of symbols using an appropriately designed space-time filter is computationally complex. Hence a reasonable design is to focus spatial beams along a single dominant path.
- For a large antenna array, beamforming along a given path significantly attenuates other paths, so that they can be safely neglected post-beamforming.

A. Problem Formulation

We consider here the max-min fair formulation of the precoding optimization problem. Thus, precoding weights acquisition block calculates the beamforming directions, $\{\bar{\mathbf{w}}_k\}$, and the power allocation, $\{p_k\}$, by solving the following optimization problem:

$$\gamma_o = \max_{\bar{\mathbf{w}}_k, p_k \forall k} \quad \min_k \; SINR_k \tag{5a}$$

$$\text{s.t.} \quad \sum_{i=1}^{K} p_i \leq P_{\text{tot}}, \tag{5b}$$

$$\|\bar{\mathbf{w}}_k\|_2 = 1 \;\; \forall k, \tag{5c}$$

$$p_k \geq 0 \;\; \forall k. \tag{5d}$$

This problem can be cast as a generalized eigenvalue problem and is always feasible [7].

After defining suitable Lagrange multipliers λ_k, the optimality conditions for problem (5) can be formulated as follows,

$$\mathbf{h}_k^H \left(\mathbf{I} + \sum_{i=1}^{K} \frac{\lambda_i}{\sigma_i^2} \mathbf{h}_i \mathbf{h}_i^H \right)^{-1} \mathbf{h}_k \frac{\lambda_k}{\sigma_k^2} = \frac{\gamma_o}{1+\gamma_o} \;\; \forall k, \tag{6}$$

$$\sum_{i=1}^{K} \lambda_i = P_{\text{tot}}, \tag{7}$$

$$\lambda_k \geq 0 \;\; \forall k. \tag{8}$$

Algorithm 1 Fixed point iteration to find optimal λ_k [7]

Input: $\{\mathbf{h}_k\}$, $\{\sigma_k^2\}$, and P_{tot}
Output: $\{\lambda_k\}$

1: initialize $\lambda_k = P_{\text{tot}}/K$
2: **repeat**
3: set $\mathbf{B} = \left(\mathbf{I} + \sum_i \mathbf{h}_i \mathbf{h}_i^H \lambda_i/\sigma_i^2\right)$ (III)
4: set $\mathbf{G} = \mathbf{B}^{-1}$ (IV)
5: set $q_k = \mathbf{h}_k^H \mathbf{G} \mathbf{h}_k \lambda_k/\sigma_k^2$ (V)
6: set $\bar{\lambda}_k = \lambda_k/q_k$
7: set $\lambda_k = P_{\text{tot}}\lambda_k/\sum_i \bar{\lambda}_i$
8: **until** q_k are all equal $\forall k$.

As a consequence, the beamforming directions can be written as follows,

$$\bar{\mathbf{w}}_k = \frac{\left(\mathbf{I} + \sum_{i=1}^K \frac{\lambda_i}{\sigma_i^2}\mathbf{h}_i\mathbf{h}_i^H\right)^{-1}\mathbf{h}_k}{\left\|\left(\mathbf{I} + \sum_{i=1}^K \frac{\lambda_i}{\sigma_i^2}\mathbf{h}_i\mathbf{h}_i^H\right)^{-1}\mathbf{h}_k\right\|_2}, \tag{9}$$

and the power vector, $\mathbf{p} = [p_1, \ldots, p_K]^\intercal$, can be evaluated by solving the following system of linear equations,

$$\left(\frac{1+\gamma_o}{\gamma_o}\mathbf{I} - \left[\frac{|\mathbf{h}_i^H\bar{\mathbf{w}}_j|^2}{|\mathbf{h}_i^H\bar{\mathbf{w}}_i|^2}\right]_{i=1,j=1}^{K,K}\right)\mathbf{p} = \left[\frac{\sigma_i^2}{|\mathbf{h}_i^H\bar{\mathbf{w}}_i|^2}\right]_{i=1}^{K}. \tag{10}$$

It is evident that the Lagrange multipliers, λ_k, play a critical role in solving the optimization problem posed in (5). Hence, all solution approaches revolve around finding optimal (or suboptimal) values of the Lagrange multipliers λ_k.

B. Fixed Point Iterations for Optimal Precoding

We review the method proposed in [7] for tackling the optimization problem (5). This provides a benchmark for optimal precoding for general channel models, as well as a basis for our proposed beamspace approach tailored to sparse channels. The optimality condition (6) can be rewritten as follows:

$$\lambda_k = \frac{\gamma_o}{1+\gamma_o}\frac{\sigma_k^2}{\mathbf{h}_k^H\left(\mathbf{I} + \sum_{i=1}^K \frac{\lambda_i}{\sigma_i^2}\mathbf{h}_i\mathbf{h}_i^H\right)^{-1}\mathbf{h}_k} \ \forall k, \tag{11}$$

which motivates using a fixed-point iteration method to find the optimal Lagrange multipliers, $\{\lambda_k\}$. The scaling of the fixed point depends on γ_o, which is the max-min SINR solution to the optimization problem, and is therefore unknown. Thus, fixed point iterations are interleaved with a scaling step based on the total power constraint (7). The resulting algorithm, whose convergence is proved in [7], is summarized as Algorithm 1: one fixed point iteration (steps 5 and 6) is followed by imposing the total power constraint (step 7), repeated until convergence to within some tolerance of the optimality condition (6).

Most prior evaluations of optimal precoding focus on a relatively small number of antennas. As we increase the number of antennas, the computational complexity for attaining convergence becomes excessive. In order to compare our low-complexity beamspace technique with the state of the art, we consider terminating Algorithm 1 after a fixed number of iterations based on a computational budget. The resulting Lagrange multipliers are suboptimal, and the optimality condition (6) is not necessarily satisfied. We can still compute the normalized beamforming directions (9) using these suboptimal Lagrange multipliers, but the power allocation (10) cannot be used, since we do not know γ_o. Instead, we fix the suboptimal beamforming directions $\bar{\mathbf{w}}_k$, and solve an optimal power allocation problem as follows to obtain a benchmark for comparison:

$$\max_{p_k \forall k} \ \min_k \ \frac{|\mathbf{h}_k^H\bar{\mathbf{w}}_k|^2 p_k}{\sigma_k^2 + \sum\limits_{\substack{i=1\\ i\neq k}}^K |\mathbf{h}_k^H\bar{\mathbf{w}}_i|^2 p_i} \tag{12a}$$

$$\text{s.t.} \quad \sum_{i=1}^K p_i \leq P_{\text{tot}}, \tag{12b}$$

$$p_k \geq 0 \ \forall k. \tag{12c}$$

Once again, the optimization problem (12) is always feasible and admits a fixed point solution that satisfies

$$\tilde{\mathbf{p}} = \left(\left[\frac{|\mathbf{h}_i^H\bar{\mathbf{w}}_j|^2}{|\mathbf{h}_i^H\bar{\mathbf{w}}_i|^2}\right]_{i=1,j=1}^{K,K} - \mathbf{I}\right)\mathbf{p} + \left[\frac{\sigma_i^2}{|\mathbf{h}_i^H\bar{\mathbf{w}}_i|^2}\right]_{i=1}^{K}, \tag{13}$$

$$p_k = \tilde{p}_k \frac{P_{\text{tot}}}{\sum_i \tilde{p}_i}. \tag{14}$$

Computational Complexity: The complexity of Algorithm 1 is dominated by the steps labeled (III), (IV), and (V). The computational complexity *per iteration* for these steps is calculated as follows.

- (III): The complexity of computing matrix $\mathbf{B} \in \mathcal{C}^{N\times N}$ is $\mathcal{O}(KN^2)$.
- (IV): The matrix inversion can be carried out efficiently using Cholesky decomposition [10], whose complexity is $\mathcal{O}(N^3)$.
- (V): The complexity of this step is $\mathcal{O}(KN^2)$.

III. Proposed Beamspace Solution

We define the beamspace representation of the channel matrix as $\bar{\mathbf{H}} = [\mathcal{DFT}(\mathbf{h}_1), \ldots, \mathcal{DFT}(\mathbf{h}_K)]$ where $\mathcal{DFT}(.)$ is the discrete Fourier transform (DFT) operator. We plot the magnitude of $\bar{\mathbf{H}}$ in Fig. 2, which makes evident the sparsity of single-path channels in beamspace. As shown in our prior work [2], for such channel models, operating in beamspace can significantly reduce the complexity of uplink multiuser detection. Given downlink-uplink duality and the iterative nature of optimization for downlink precoding, we expect even greater savings in complexity in our present setting.

We describe the proposed beamspace optimization algorithm, depicted in Algorithm 2, as follows. We assume here that we have access to estimates of the $N \times 1$ channel vectors, $\{\mathbf{h}_k\}$, and hence account for the complexity of taking

DFT to go to beamspace. This process could potentially be avoided by use of channel estimation techniques that utilize beamspace techniques up front (e.g., the use of reciprocity, and uplink techniques such as those in [2]).

1) **Computing the DFT of the channel vectors:** The DFT is used to transform each channel vector $\mathbf{h}_k$ from the antenna domain to the beam domain to get $\bar{\mathbf{h}}_k$ evaluated as follows,

$$\bar{h}_{ki} = \sum_{n=1}^{N} h_{kn} e^{-j2\pi(n-1)(i-1)/N}. \tag{15}$$

Using the fast Fourier transform (FFT) algorithm [11], the complexity of this step becomes $\mathcal{O}(KN\log(N))$.

2) **Energy detection:** The energy distribution of the channel vector in beamspace is concentrated around its spatial frequency. Because we do not know the spatial frequency beforehand, we search for a window of size W that contains most of the channel energy. The use of a sliding window for this purpose incurs $\mathcal{O}(N)$ complexity per user.

For a given user, after finding the window that holds most of its channel energy, it is convenient to define two "synthetic" channels in beamspace: a truncated $W \times 1$ channel $\tilde{\mathbf{h}}_k$ centered on the chosen window for user k, and an approximated $N \times 1$ channel $\hat{\mathbf{h}}_k$ obtained by filling in zeros around the window.

3) **Computing Lagrange multipliers:** We use steps similar to Algorithm 1 to calculate Lagrange multipliers, but with a drastic reduction of complexity by using synthetic channels in beamspace.

- We use the approximated channel vectors, each containing only W nonzero elements, to compute the matrix $\mathbf{B}$. As a consequence, the complexity of this step decreases to $\mathcal{O}(KW^2)$ instead of $\mathcal{O}(KN^2)$ per iteration, where $W \ll N$.
- In step (V) of Algorithm 1, we replace the original channel vector with the approximated ones. For each user, only the inverse of a small $W \times W$ block inside $\mathbf{B}$, denoted by $\mathbf{G}_k$, needs to be computed in step (IV): compare step (IV) in Algorithm 1, where we invert the entire matrix $\mathbf{B}$, with that in Algorithm 2, where we invert K blocks of size $W \times W$. Thus, the complexity of step (IV) is reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(KW^3)$.
- Finally, the complexity of step (V) is automatically reduced from $\mathcal{O}(KN^2)$ to $\mathcal{O}(KW^2)$.

IV. Results

We consider the system depicted in Fig. 1, with number of antennas fixed at $N = 256$. The field of view for the sector is restricted to $-\pi/3 \leq \theta \leq \pi/3$. The users are uniformly distributed inside a region bordered by a minimum and a maximum distance away from the base station, $R_{\text{min}} = 5$ m and $R_{\text{max}} = 100$ m, respectively. While the user terminals are placed randomly in our simulations, we enforce a minimum separation in spatial frequency between any two users in order not to incur excessive interference, arbitrarily choosing it as half the 3 dB beamwidth: $\Delta\Omega_{\text{min}} = \frac{2.783}{N}$ [12]. BW_{3dB} in

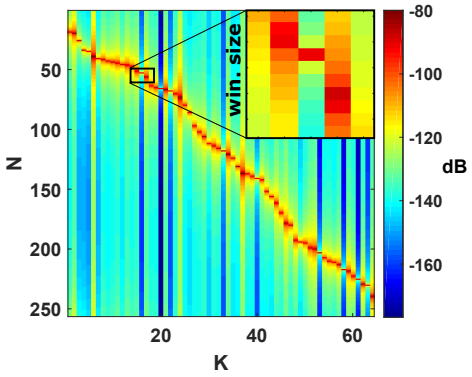


Figure 2: Sparsity of single-path channel in beamspace.

Algorithm 2 Proposed beamspace approach to find near-optimal λ_k

Input: $\{\mathbf{h}_k\}$, $\{\sigma_k^2\}$, W and P_{tot}
Output: $\{\lambda_k\}$

1: set $\bar{\mathbf{h}}_k = \mathcal{FFT}(\mathbf{h}_k)$ (I)
2: set $\ell_k = \arg\max_{\ell} \sum_{i=\ell}^{\ell+W-1} |\bar{h}_{ki}|^2$ (II)
3: set $\tilde{\mathbf{h}}_k = [\bar{h}_{ki}]_{i=\ell_k}^{\ell_k+W-1}$
4: set $\hat{\mathbf{h}}_k = \left[\mathbf{0}_{1\times(\ell_k-1)}\ \tilde{\mathbf{h}}_k^{\intercal}\ \mathbf{0}_{1\times(N-\ell_k-W+1)}\right]^{\intercal}$
5: initialize $\lambda_k = P_{\text{tot}}/K$
6: **repeat**
7: set $\mathbf{B} = [b_{ij}]_{i,j} = \left(\mathbf{I} + \sum_i \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^H \lambda_i/\sigma_i^2\right)$ (III)
8: set $\mathbf{G}_k = \left([b_{ij}]_{i=\ell_k,j=\ell_k}^{\ell_k+W-1,\ell_k+W-1}\right)^{-1}$ (IV)
9: set $q_k = \tilde{\mathbf{h}}_k^H \mathbf{G}_k \tilde{\mathbf{h}}_k \lambda_k/\sigma_k^2$ (V)
10: set $\bar{\lambda}_k = \lambda_k/q_k$
11: set $\lambda_k = P_{\text{tot}}\bar{\lambda}_k/\sum_i \bar{\lambda}_i$
12: **until** q_k are all equal $\forall k$.

Fig. 1 stands for the 3 dB beamwidth. We assume that users with similar spatial frequency can be served in different time or frequency resource blocks.

We measure link quality by the outage probability at a target uncoded BER of 10^{-3} for QPSK, which corresponds to a target SINR of 9.8 dB for each downlink user.

We define the SNR_{edge} as the SNR that would be attained by a single user at the cell edge (100 m away from the base station) if the entire power budget of the base station were directed at that user. For free space propagation and ideal beamforming at both ends, we have

$$SNR_{\text{edge}} = \frac{NMG_{\text{t}}G_{\text{r}}}{L_{100m}\sigma^2}P_{\text{tot}}, \tag{16}$$

where L_{100m} is the free space path loss incurred at 100 m away from the base station, M is the number of elements in the mobile's array, σ^2 is the noise variance in the mobile (which is identical in all mobiles), and G_{t} and G_{r} are the

transmit and receive element gain, respectively.

Precoding Efficiency: Fig. 3 (a) shows the 5^{th} percentile of the minimum SINR across different channel realization, namely $SINR_{\text{min}}$, versus the power budget represented in SNR_{edge}. That is, $SINR_{\text{min}}$ is defined such that $\mathbb{P}(\min(SINR) \leq SINR_{\text{min}}) = 5\%$.

Assuming no interference between the users, if the base station power budget is allocated equally between K edge users, then each user would attain an SINR of SNR_{edge}/K. Using this as the benchmark against which we compare the minimum SINR attained by our precoding scheme, the precoding efficiency η is defined as

$$\eta = \frac{SINR_{\text{min}}}{SNR_{\text{edge}}/K}. \tag{17}$$

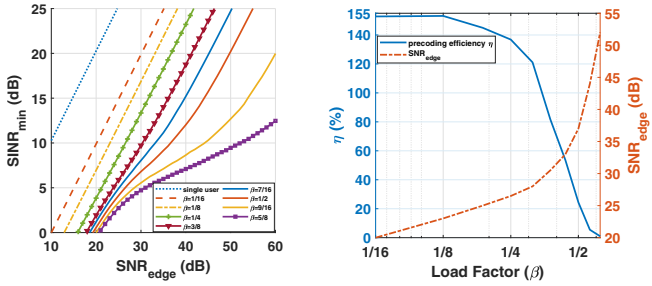
As shown in Fig. 3 (b), the efficiency can exceed 100% at low load factor β, since the base station can transfer power from nearby users to edge users to enhance the minimum SINR, and the noise enhancement due to interference suppression on the virtual uplink is small. As the load factor increases, the loss in SINR due to interference suppression becomes more significant, and efficiency drops below 100%.

Feasibility of Target SINR: We evaluate this using the same system settings as in our prior work on uplink design [1]: $M = 16$, $G_t = G_r = 3$ dBi, $L_{100m} = 115$ dB and $\sigma^2 = -70$ dBm. For a given SNR_{edge}, the resulting link budget requires a total transmitted power of $P_{\text{tot}} = SNR_{\text{edge}}(\text{dB}) + 3$ dBm. The required emitted power for the power amplifier (PA) driving each antenna is a factor of N smaller, or 24 dB smaller for $N = 256$, and is therefore given by $P_{\text{PA}} = SNR_{\text{edge}}(\text{dB}) - 21$ dBm. The required SNR_{edge} corresponding to attaining the target SINR of 9.8 dB with 5% outage is obtained by simulations and shown in Fig. 3 (b). For $\beta = 1/2$, $SNR_{\text{edge}} = 37$ dB, corresponding to $P_{\text{tot}} = 40$ dBm and $P_{\text{PA}} = 16$ dBm. Such a PA specification is difficult to obtain with low-cost CMOS technologies (CMOS PA designs of up to 11 dBm have been reported in [13]), and may require more expensive alternatives such as InP technology [14]. On the other hand, if we reduce the load factor to $\beta = 1/4$, we obtain $P_{\text{tot}} = 30$ dBm and $P_{\text{PA}} = 6$ dBm, which can be comfortably attained in CMOS.

Complexity and Performance: Table I lists the computational complexity, in terms of number of multiplication and addition operations, of the computationally expensive steps, labeled by Roman numerals, in algorithms 1 and 2. The table clearly brings out the big savings in complexity due to the proposed beamspace algorithm. Of course, the proposed algorithm incurs the additional cost of going to beamspace (steps I and II). However, these steps are required only once per channel realization, whereas the other steps (III, IV, V) are invoked on every iteration. Furthermore, as noted earlier, we may be able to fold steps I and II into channel estimation algorithms operating in beamspace.

Fig. 4 (a) depicts, for different load factors, the multiplication operations count for both the conventional and the proposed algorithm to achieve the same performance versus the number of elements in the base station array. It is evident that the difference in complexity is at least one order of magnitude, even for a relatively small $N = 16$.

Fig. 4 (b) illustrates the performance gap between the conventional and the proposed algorithm if the computational budget is limited to that of a single iteration of the conventional algorithm. As shown, the beamspace algorithm achieves higher SINR (by 6 dB) while using only one-fifth of hardware resources.



(a) The 5^{th} percentile of the minimum SINR.

(b) Feasibility and Efficiency.

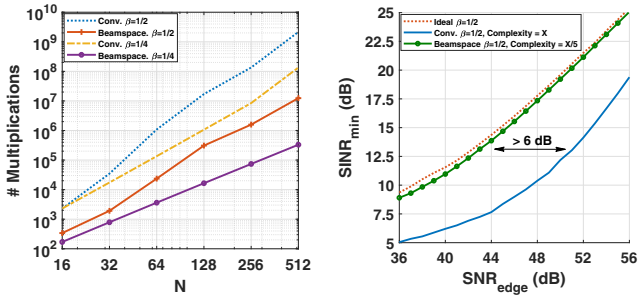
Figure 3: (a) The solution to the optimization problem (5) for different power budgets and system load factors. (b) The power budget required to achieve minimum SINR of ~ 10 dB along with the precoding efficiency at various system load factors.

Table I: The approximate number of multiplications and additions in the conventional [7] and the proposed beamspace algorithm to find nearly-optimal values of Lagrange multipliers λ_k. W and J denote the window size and the number of iterations.

Step	# Multiplications		# Additions	
	Conventional	Beamspace	Conventional	Beamspace
I	0	$\frac{KN}{2}(log_2(N) - 1)$	0	$KNlog_2(N)$
II	0	KN	0	$2KN$
III	KN^2J	KW^2J	KN^2J	KW^2J
IV	$\frac{N^3}{2}J$	$K\frac{W^3}{2}J$	$\frac{N^3}{2}J$	$K\frac{W^3}{2}J$
V	KN^2J	KW^2J	KN^2J	KW^2J

V. Conclusion

We have demonstrated the drastic complexity reduction in computing optimal downlink linear precoding weights via beamspace techniques exploiting spatial channel sparsity. Conventional iterative techniques, which are required for general channel models, require a complexity per iteration which is cubic in the number of antennas, while the proposed beamspace algorithm requires linear complexity per iteration. Coupled with our prior work [2] showing the efficacy of beamspace techniques for uplink multiuser detection, it is clear that beamspace techniques are a powerful tool for supporting truly massive MIMO in the mmWave and THz bands, since



(a) Computational Complexity. (b) Performance.

Figure 4: (a) Comparison of the number of multiplication operations in the conventional and the beamspace algorithm as the number of elements in base station increases. (b) The beamspace algorithm needs less than one-fourth of the budget power to achieve the same minimum SINR.

they are naturally matched to the channel sparsity characteristic of these bands. Another conclusion from this work and [2], as well as from related work on hardware-constrained design [1], [3], is that operating at lower load factors provides significant advantages as we scale up the number of antennas. Ongoing work focuses on developing a comprehensive system design and signal processing framework around these concepts.

Acknowledgment

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, A. A. Farid, U. Madhow, and M. J. W. Rodwell, "Towards all-digital mmWave massive MIMO: Designing around non-linearities," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct 2018, pp. 1552–1557.

[2] 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)*, July 2019, pp. 1–5.

[3] M. E. Rasekh, M. Abdelghany, U. Madhow, and M. Rodwell, "Phase noise analysis for mmWave massive MIMO: a design framework for scaling via tiled architectures," in *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, March 2019, pp. 1–6.

[4] F. Rashid-Farrokhi, K. J. R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1437–1450, Oct 1998.

[5] E. Visotsky and U. Madhow, "Optimum beamforming using transmit antenna arrays," in *1999 IEEE 49th Vehicular Technology Conference (Cat. No.99CH36363)*, vol. 1, May 1999, pp. 851–856 vol.1.

[6] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, Jan 2004.

[7] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 161–176, Jan 2006.

[8] C. W. Tan, M. Chiang, and R. Srikant, "Maximizing sum rate and minimizing MSE on multiuser downlink: Optimality, fast algorithms and equivalence via max-min SINR," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6127–6143, 2011.

[9] H. V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE transactions on Information theory*, vol. 43, no. 3, pp. 858–871, 1997.

[10] 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.

[11] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

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

[13] D. Simic and P. Reynaert, "A 14.8 dBm 20.3 dB power amplifier for D-band applications in 40 nm CMOS," in *2018 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, June 2018, pp. 232–235.

[14] T. B. Reed, M. Rodwell, Z. Griffith, P. Rowell, A. Young, M. Urteaga, and M. Field, "A 220 GHz InP HBT solid-state power amplifier MMIC with 90mW POUT at 8.2dB compressed gain," in *2012 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, Oct 2012, pp. 1–4.