

Near-Optimal Hybrid Analog and Digital Precoding for Downlink mmWave Massive MIMO Systems

Linglong Dai¹, Xinyu Gao¹, Jinguo Quan², and Shuangfeng Han³, and Chih-Lin I³

¹Tsinghua National Laboratory for Information Science and Technology (TNList),

Department of Electronic Engineering, Tsinghua University, Beijing, China

²Division of Information Science & Technology, Shenzhen Graduate School, Tsinghua University, Shenzhen, China

³Green Communication Research Center, China Mobile Research Institute, Beijing 100053, China

Abstract—Millimeter wave (mmWave) massive MIMO can achieve orders of magnitude increase in spectral and energy efficiency, and it usually exploits the hybrid analog and digital precoding to overcome the serious signal attenuation induced by mmWave frequencies. However, most of hybrid precoding schemes focus on the full-array structure, which involves a high complexity. In this paper, we propose a near-optimal iterative hybrid precoding scheme based on the more realistic sub-array structure with low complexity. We first decompose the complicated capacity optimization problem into a series of ones easier to be handled by considering each antenna array one by one. Then we optimize the achievable capacity of each antenna array from the first one to the last one by utilizing the idea of successive interference cancellation (SIC), which is realized in an iterative procedure that is easy to be parallelized. It is shown that the proposed hybrid precoding scheme can achieve better performance than other recently proposed hybrid precoding schemes, while it also enjoys an acceptable computational complexity.

I. INTRODUCTION

The mergence of millimeter-wave (mmWave) and massive multiple-input multiple-output (MIMO) is regarded as a promising technique for future 5G wireless communication systems [1], since it can provide orders of magnitude increase both in the available bandwidth and the spectrum efficiency [2], [3]. On one hand, the short wavelength associated with high frequencies of mmWave enables a large antenna array to be arranged in a small size. On the other hand, the large antenna array in massive MIMO can provide sufficient antenna gains to compensate for the serious signal attenuation induced by mmWave frequencies (e.g., rainfall effect and oxygen absorption [1]) by exploiting the precoding technique to concentrate the signal in a specific direction.

However, the precoding schemes for massive MIMO are mainly performed in the baseband domain, where a fully digital precoder is utilized to eliminate the interferences by controlling both the amplitude and phase of transmitted signals. The fully digital precoding scheme, such as dirty paper precoding (DPC) [4], can achieve the optimal performance, but requires an expensive radio frequency (RF) chain (including digital-to-analog converter, up converter, and etc.) for every antenna. As a result, the total hardware complexity and energy consumption of digital precoding becomes a serious problem due to the increasing number of antennas, specially for mmWave communications where the number of antennas at the base station (BS) can be huge (e.g. 256 antennas can

be used [1]). To solve this problem, the hybrid precoding technique combining the digital precoding and the analog precoding has been proposed to significantly reduce the number of RF chains, is proposed [5]. As it can reduce both the hardware complexity and the energy consumption without obvious performance loss, it is considered as an essential technique for realistic mmWave massive MIMO systems [6]. However, how to jointly design a near-optimal low-complexity digital and analog precoding is a challenging problem to be solved.

Recently, a hybrid precoding scheme, where the analog precoder is selected from the antenna array response vectors while the digital precoder is chosen from a predefined codebook [7], is proposed to achieve the satisfying performance, but it suffers from a high searching complexity. More recently, another scheme named spatially sparse precoding is proposed in [8]. By formulating the capacity optimization problem as a sparse approximation problem, this scheme can approach the close-optimal capacity, but it involves very high computational complexity. Moreover, both of the schemes in [7] and [8] are designed for a full-array structure where each RF chain is connected to all BS transmit antennas, which makes the analog precoder too complicated in hardware. For a more realistic hybrid precoding structure where each RF chain is connected to only a small set of transmit antennas, i.e., the sub-array structure, [9] proposes a simple precoding scheme for a single data stream transmission, where the analog precoder and the digital precoder are selected from a predefined candidate set. By searching the optimal pair of analog and digital precoders, this scheme can achieve satisfying performance. For the multiple data streams transmission scenario, [10] proposes an iterative scheme where only the analog precoder is employed. It is simple and easy to be implemented in hardware, but suffers a non-negligible performance loss especially at low signal-to-noise (SNR) regions.

In this paper, we consider the mmWave massive MIMO systems with the sub-array structure for hybrid precoding and propose an iterative hybrid precoding scheme to achieve the near-optimal performance with low complexity. In particular, we decompose the complicated capacity optimization problem into a series of ones easier to be handled by considering each antenna array one by one, which means that the precoding for each antenna array connected to a specific RF chain will be

considered separately instead of jointly. For the first antenna array, we use the digital precoder to control the amplitude, and the analog precoder to adjust the phase, and then the capacity optimization problem for the first antenna array will be solved with low complexity. After optimizing the achievable capacity of the first antenna array, we can utilize the idea of successive interference cancellation (SIC) to eliminate the contribution of the first antenna array from the total capacity expression, and then optimize the achievable capacity of the second antenna array. We repeat such procedure until the last antenna array is considered. Analysis and simulation results verify that the proposed hybrid precoding scheme can approach the ideal capacity with low complexity.

The rest of the paper is organized as follows. Section II briefly introduces the mmWave massive MIMO system model. Section III specifies the proposed near-optimal low-complexity precoding scheme, together with the complexity analysis. The simulation results of capacity performance are shown in Section IV. Finally, conclusions are drawn in Section V.

Notation: Lower-case and upper-case boldface letters denote vectors and matrices, respectively; $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, and $\det(\cdot)$ denote the transpose, conjugate transpose, inversion, and determinant of a matrix, respectively; $\|\cdot\|_2$ denotes the 2-norm of a vector; $|\cdot|$ denote the absolute operator; $\text{Re}\{\cdot\}$ and $\text{Im}\{\cdot\}$ denote the real part and imaginary part of a complex number, respectively; $\mathbb{E}(\cdot)$ denotes the expectation; Finally, \mathbf{I}_N is the $N \times N$ identity matrix.

II. SYSTEM MODEL

Fig. 1 compares the architecture of the fully digital precoding in conventional massive MIMO and the hybrid analog/digital precoding in mmWave massive MIMO. In this paper, we consider the mmWave massive MIMO with the sub-array structure for hybrid precoding as shown in Fig. 1 (b), where the BS is equipped with NM antennas but only N independent RF chains to simultaneously transmit N data streams for the K receive antennas of user. The N data streams in the baseband are firstly precoded by an $N \times N$ digital precoder \mathbf{D} , and then pass through N RF chains. After that, each of the N data streams is precoded again by an $M \times 1$ analog precoder \mathbf{a}_i ($i = 1, \dots, N$) before transmission, where the analog precoder is usually realized by a phase shifter [11], i.e., all elements of \mathbf{a}_i have the same amplitude of one but different phases. After analog precoding, each data stream is finally transmitted by M antennas connected to the corresponding RF chain. Thus, the received $K \times 1$ signal vector $\mathbf{y} = [y_1, \dots, y_K]^T$ at the user side can be presented as

$$\mathbf{y} = \rho \mathbf{H} \mathbf{A} \mathbf{D} \mathbf{s} + \mathbf{n} = \rho \mathbf{H} \mathbf{P} \mathbf{s} + \mathbf{n}, \quad (1)$$

where ρ is the average received power, $\mathbf{s} = [s_1, \dots, s_N]^T$ presents the transmitted signal vector in the baseband, and usually $\mathbb{E}(\mathbf{s}\mathbf{s}^H) = \frac{1}{N} \mathbf{I}_N$ is assumed for the normalized signal power [8]. $\mathbf{P} = \mathbf{A} \mathbf{D}$ presents the hybrid precoding matrix of size $NM \times N$. To reduce the complexity in hardware, the digital precoding matrix \mathbf{D} can be designed as a diagonal matrix [12], and the i th diagonal elements $d_{i,i}$ can be a

real constant. That means the digital precoder can be simply realized by an amplifier. \mathbf{A} presents the $NM \times N$ analog precoding matrix as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{a}_2 & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{a}_N \end{bmatrix}_{NM \times N} \quad (2)$$

where each row of \mathbf{A} has only one nonzero entry due to the sub-array structure of the hybrid precoding scheme as shown in Fig. 1 (b). $\mathbf{H} \in \mathbb{C}^{K \times NM}$ denotes the channel matrix based on the Saleh-Valenzuela model [13] widely used for mmWave communications. Finally, $\mathbf{n} = [n_1, \dots, n_K]^T$ is the additive white Gaussian noise (AWGN) vector, whose entries follow the independent and identically distribution (i.i.d.) $\mathcal{CN}(0, \sigma^2)$.

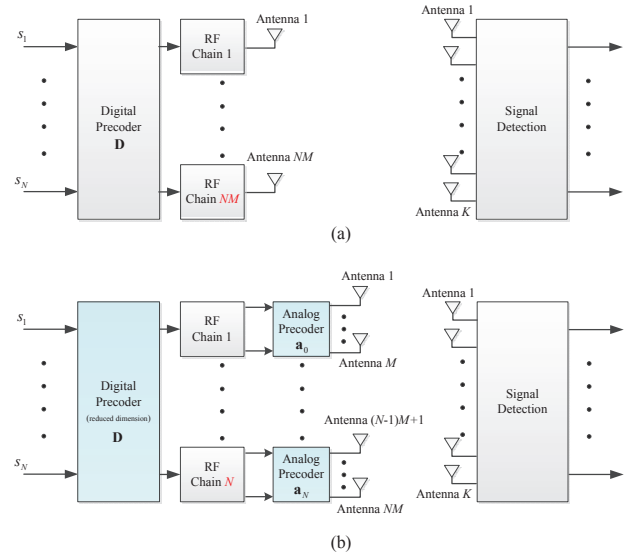


Fig. 1. Architecture comparison: (a) Fully digital precoding in conventional massive MIMO; (b) Hybrid analog/digital precoding in mmWave massive MIMO with sub-array structure.

It is known that mmWave channel will no longer obey the conventional Rayleigh fading due to the limited scatters [5], [14]. In this paper, we consider the geometric Saleh-Valenzuela channel model more appropriate for mmWave communications, where the channel matrix \mathbf{H} can be presented as [13]

$$\mathbf{H} = \sqrt{\frac{N^2 M}{L}} \sum_{l=1}^L \alpha_l \Lambda_r(\phi_l^r) \Lambda_t(\phi_l^t) \mathbf{f}_r(\phi_l^r) \mathbf{f}_t^H(\phi_l^t), \quad (3)$$

where L is the number of channel paths corresponding to limited scatters, and we usually have $L \leq N$ for mmWave communication systems. $\alpha_l \in \mathbb{C}$ is the gain of the l th path which follows the Rayleigh distribution. ϕ_l^t and ϕ_l^r are the azimuth angles of departure and arrival (AoDs/AoAs), respectively. $\Lambda_t(\phi_l^t)$ and $\Lambda_r(\phi_l^r)$ denote the transmit antenna array gain at a specific AoD and AoA, respectively.

For simplicity but without loss of generality, $\Lambda_t(\phi_l^t)$ and $\Lambda_r(\phi_l^r)$ can be set as one, which corresponds that both the receive and transmit antennas are assumed to be omnidirectional [13]. Finally, $\mathbf{f}_t(\phi_l^t)$ and $\mathbf{f}_r(\phi_l^r)$ are the antenna array response vectors which heavily depends on the antenna array structure at the BS and the users, respectively. When both the transmit antenna array and the receive antenna array are the widely used uniform linear arrays (ULAs), we have

$$\mathbf{f}_r(\phi_l^r) = \alpha \left[1, e^{j \frac{2\pi}{\lambda} d \sin(\phi_l^r)}, \dots, e^{j(N-1) \frac{2\pi}{\lambda} d \sin(\phi_l^r)} \right]^T, \quad (4)$$

$$\mathbf{f}_t(\phi_l^t) = \beta \left[1, e^{j \frac{2\pi}{\lambda} d \sin(\phi_l^t)}, \dots, e^{j(NM-1) \frac{2\pi}{\lambda} d \sin(\phi_l^t)} \right]^T, \quad (5)$$

where $\alpha = \frac{1}{\sqrt{N}}$ and $\beta = \frac{1}{\sqrt{NM}}$ are the normalized factors, λ denotes the wavelength of the signal, d is the distance between two adjacent antenna elements. Note that in this paper, we only consider azimuth angles of the AoAs and AoDs without considering elevation angles, which means we will focus on the 2-D hybrid precoding scheme. The extensive design of 3D precoding scheme is also possible and will be left for further work.

III. NEAR-OPTIMAL HYBRID ANALOG/DIGITAL PRECODING

In this section, we propose a near-optimal hybrid analog/digital precoding scheme by the joint design of analog and digital precoders for the mmWave massive MIMO system as illustrated in Fig. 1 (b). The complexity analysis is also provided to show its advantage over conventional schemes.

A. The decomposition of the capacity optimization problem

The final aim of precoding is to maximize the achievable channel capacity of MIMO systems, which can be expressed as [8]

$$R = \log_2 \left(\left| \mathbf{I}_N + \frac{\rho}{\sigma^2} \mathbf{H} \mathbf{P} \mathbf{P}^H \mathbf{H}^H \right| \right). \quad (6)$$

Based on the system model (1) as described in Section II, as the digital precoding matrix \mathbf{D} is a diagonal matrix of real elements, there remains two constraints for the design of the hybrid precoding matrix \mathbf{P} : i) \mathbf{P} should be a block diagonal matrix similar to the form of \mathbf{A} as shown in (2); ii) All the non-zero elements of the i th column of \mathbf{P} has the same amplitude $d_{i,i}$, the i th diagonal element of the digital precoding matrix \mathbf{D} .

Unfortunately, the non-convex constraints on \mathbf{P} make the optimization problem (6) very difficult to be solved. To this end, as illustrated in Fig. 2, we propose to decompose the difficult optimization problem (6) into a series of sub-optimal problems much easier to be solved. In particular, by considering each antenna array connected to each RF chain one by one, we can optimize the achievable capacity of the first antenna array by assuming that all the other antenna arrays are closed. After that, we can eliminate the contribution of the first antenna array from (6), and then optimize the achievable capacity of the second antenna array. Such similar procedure will be executed until the last antenna array is considered. The

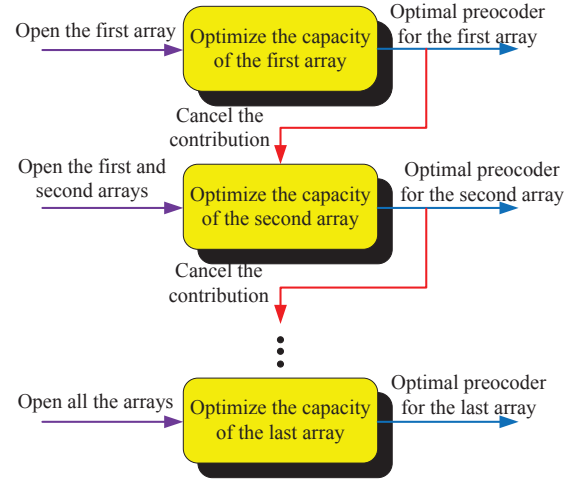


Fig. 2. Diagram of the proposed iterative hybrid analog/digital precoding.

proposed iterative hybrid analog/digital precoding scheme will be described in detail in the next subsection.

B. Near-optimal hybrid analog/digital precoding

According to the analysis above, the achievable capacity of the first antenna array can be expressed by

$$C_1 = \log_2 \left(1 + \frac{\rho}{\sigma^2} \mathbf{p}_1^H \mathbf{H}^H \mathbf{H} \mathbf{p}_1 \right), \quad (7)$$

where $\mathbf{p}_1 = d_{1,1} \mathbf{b}_1$ is the first column of the hybrid precoding matrix \mathbf{P} (i.e., the precoder vector for the first antenna array), \mathbf{b}_1 is the first column of the analog precoding matrix \mathbf{A} containing the $M \times 1$ non-zero vector \mathbf{a}_1 in (2) and the $(N-1)M \times 1$ zero vector, i.e., $\mathbf{b}_1 = [\mathbf{a}_1 \mathbf{0}]_{NM \times 1}^T$. Let $\mathbf{G} = \mathbf{H}^H \mathbf{H}$, since only the first M elements of \mathbf{p}_1 are non-zeros, the term $\mathbf{p}_1^H \mathbf{G} \mathbf{p}_1$ which is required to be optimized in (7) can be rewritten as

$$\mathbf{p}_1^H \mathbf{G} \mathbf{p}_1 = \mathbf{p}_1^H \mathbf{S} \mathbf{p}_1, \quad (8)$$

where \mathbf{S} is a sub-matrix of \mathbf{G} which only keeps the rows and columns of \mathbf{G} from 1 to M . Define the ordered singular value decomposition (SVD) of the Hermitian matrix \mathbf{S} as $\mathbf{S} = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^H$, where $\mathbf{\Sigma}$ is a $M \times M$ diagonal matrix containing the singular values of \mathbf{S} in a decreasing order, and \mathbf{V} is an $M \times M$ unitary matrix. It is known that the optimal solution to the objective (7) can be obtained by [15]

$$\mathbf{p}_{1,opt} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{0} \end{bmatrix}_{NM \times 1}, \quad (9)$$

where the $M \times 1$ vector \mathbf{v}_1 is the first column of \mathbf{V} . However, based on the constraints as we mentioned in Section III-A, we cannot directly choose \mathbf{p}_1 as $\mathbf{p}_{1,opt}$ since the elements of $\mathbf{p}_{1,opt}$ don't obey the constraint of constant amplitude. For this reason, we propose to explore a choice of \mathbf{p}_1 that is sufficiently close to $\mathbf{p}_{1,opt}$ by minimizing the mean-square-error (MSE) between \mathbf{p}_1 and $\mathbf{p}_{1,opt}$ (or equivalently $d_{1,1} \mathbf{a}_1$ and \mathbf{v}_1) under

the constraint of constant amplitude. Specifically, the MSE function can be presented as

$$\begin{aligned} & \mathbb{E} \left\{ \|\mathbf{p}_{1,opt} - \mathbf{p}_1\|_2^2 \right\} \\ &= \mathbb{E} \left\{ \|\mathbf{v}_1 - d_{1,1} \mathbf{a}_1\|_2^2 \right\} \\ &= \mathbb{E} \left\{ (\mathbf{v}_1 - d_{1,1} \mathbf{a}_1)^H (\mathbf{v}_1 - d_{1,1} \mathbf{a}_1) \right\} \\ &= \mathbb{E} \left\{ \mathbf{v}_1^H \mathbf{v}_1 - d_{1,1} (\mathbf{a}_1^H \mathbf{v}_1 + \mathbf{v}_1^H \mathbf{a}_1) + d_{1,1}^2 \mathbf{a}_1^H \mathbf{a}_1 \right\}. \end{aligned} \quad (10)$$

According to facts that the elements of $M \times 1$ non-zero vector \mathbf{a}_1 only have different phase and $d_{1,1}$ is a real scalar, a feasible and intuitive choice of \mathbf{a}_1 is

$$\mathbf{a}_1 = e^{j\angle(\mathbf{v}_1)}, \quad (11)$$

where $\angle(\mathbf{v}_1)$ denotes the phase vector of \mathbf{v}_1 , which means that the element of \mathbf{a}_1 will share the same phase as the corresponding element of \mathbf{v}_1 with amplitude of one. Substituting (11) into (10), we can observe that the MSE function (10) will be simplified to a quadratic function that only depends on $d_{1,1}$ as

$$\begin{aligned} & \mathbb{E} \left\{ \|\mathbf{p}_{1,opt} - \mathbf{p}_1\|_2^2 \right\} \\ &= \mathbb{E} \left\{ \mathbf{v}_1^H \mathbf{v}_1 - d_{1,1} (\mathbf{a}_1^H \mathbf{v}_1 + \mathbf{v}_1^H \mathbf{a}_1) + d_{1,1}^2 \mathbf{a}_1^H \mathbf{a}_1 \right\} \\ &= 1 + \mathbb{E} \left\{ M d_{1,1}^2 - d_{1,1} \left(e^{j\angle(\mathbf{v}_1^H)} \mathbf{v}_1 + \mathbf{v}_1^H e^{j\angle(\mathbf{v}_1)} \right) \right\}. \end{aligned} \quad (12)$$

where $\mathbf{a}_1^H \mathbf{a}_1 = \|e^{j\angle(\mathbf{v}_1)}\|_2^2 = M$, $\mathbf{v}_1^H \mathbf{v}_1 = 1$ since \mathbf{v}_1 is a column of unitary matrix \mathbf{V} . To minimize (12), it is easy to obtain the best choice of $d_{1,1}$ as

$$d_{11} = \frac{e^{j\angle(\mathbf{v}_1^H)} \mathbf{v}_1 + \mathbf{v}_1^H e^{j\angle(\mathbf{v}_1)}}{2M} = \frac{\text{Re} [\mathbf{v}_1^H e^{j\angle(\mathbf{v}_1)}]}{M}. \quad (13)$$

Combining (11) and (13), we can obtain the precoder vector \mathbf{p}_1 for the first antenna array sufficiently close to $\mathbf{p}_{1,opt}$ as

$$\mathbf{p}_1 = \begin{bmatrix} \frac{\text{Re} [\mathbf{v}_1^H e^{j\angle(\mathbf{v}_1)}]}{M} \times e^{j\angle(\mathbf{v}_1)} \\ \mathbf{0} \end{bmatrix}_{NM \times 1}. \quad (14)$$

After acquiring the precoder vector for the first antenna array, we can utilize the idea of SIC to eliminate the contribution of the first antenna array from the total channel capacity (6), and then focus on the optimal design of the second precoder vector, and so on. Specifically, when the precoder vectors for the first m antenna arrays have been obtained, by exploiting the Sherman-Morrison determinant identity [10], the total channel capacity R after adding the contribution of the $(m+1)$ th antenna array should be

$$R = \log_2(|\mathbf{T}_m|) + \log_2 \left\{ 1 + \frac{\rho}{\sigma^2} \mathbf{p}_{m+1}^H \mathbf{H}^H \mathbf{T}_m^{-1} \mathbf{H} \mathbf{p}_{m+1} \right\}, \quad (15)$$

$$\mathbf{T}_m = \mathbf{I}_N + \frac{\rho}{\sigma^2} \mathbf{H} \mathbf{P} (1:m) \mathbf{P} (1:m)^H \mathbf{H}^H, \quad (16)$$

where $\mathbf{P} (1:m)$ denotes the first m columns of \mathbf{P} and $\mathbf{p}_{m+1} = d_{m+1,m+1} \mathbf{b}_{m+1}$ is the $(m+1)$ th column of \mathbf{P} which corresponds to the precoder vector for the $(m+1)$ th antenna array. We can observe that the first term on the right side of (15) is independent of \mathbf{p}_{m+1} . Therefore, maximizing (15) as

a function of \mathbf{p}_{m+1} equals to maximizing the second term of (15) on the right side. Note that this term and (7) share a similar form. Therefore, if we redefine the matrix \mathbf{G} as

$$\mathbf{G} = \mathbf{H}^H \mathbf{T}_m^{-1} \mathbf{H}, \quad (17)$$

we can utilize the similar method when we consider \mathbf{p}_1 in (7) to obtain a near-optimal precoder vector \mathbf{p}_{m+1} for the $(m+1)$ th antenna array. After \mathbf{p}_{m+1} has been obtained, the precoder vector for the rest antenna arrays can be also acquired similarly.

Input: (1) Number of RF chains N ;
 (2) Number of antennas for one RF chain M ;
 (3) Channel matrix \mathbf{H}
Initialization: $\mathbf{P} = \mathbf{0}_{NM \times N}$
for $1 \leq m \leq N$
 1) Update matrix \mathbf{G} :
 $\mathbf{G} \leftarrow \mathbf{H}^H \mathbf{T}_m^{-1} \mathbf{H}$;
 2) Acquire the effective sub-matrix of \mathbf{G} :
 $\mathbf{S} \leftarrow \mathbf{G}(M(m-1) : Mm, M(m-1) : Mm)$;
 3) SVD:
 $\mathbf{v}_1 \leftarrow$ The first right-singular vector of \mathbf{S} ;
 4) The analog precoder:
 $\mathbf{a}_m \leftarrow e^{j\angle(\mathbf{v}_1)}, \mathbf{b}_m \leftarrow \begin{bmatrix} \mathbf{0}_{M(m-1) \times 1} \\ \mathbf{a}_m \\ \mathbf{0}_{M(N-m) \times 1} \end{bmatrix}$;
 5) The digital precoder:
 $d_{m,m} \leftarrow \frac{\text{Re} [\mathbf{v}_1^H \mathbf{a}_m]}{M}$;
 6) The total precoder:
 $\mathbf{p}_m \leftarrow d_{m,m} \mathbf{b}_m$;
end for
Output: (1) $\mathbf{A} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N]$;
 (2) $\mathbf{D} = \text{diag}(d_{1,1}, d_{2,2}, \dots, d_{N,N})$;
 (3) $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N]$.

Algorithm 1: Iterative hybrid analog/digital precoding

To sum up, the pseudo-code of the proposed iterative hybrid analog/digital precoding scheme is described in **Algorithm 1**. Note that this scheme can be guaranteed to achieve a near-optimal performance, which is close to the upper-bound achieved by the optimal scheme when we choose $\mathbf{p}_m = \mathbf{v}_1$ without considering the constraint of constant amplitude, as will be verified later by the simulation results in Section IV.

C. Complexity analysis

In this subsection, we compare the computational complexity of the proposed iterative hybrid analog/digital precoding and the fully analog precoding in [10]. Since both of them need to compute (17) in each iteration, we compare their complexity after we have obtained the matrix \mathbf{G} in terms of the required number of multiplications [16].

It can be found from **Algorithm 1** that the complexity of the proposed scheme except computing (17) comes from four parts. The first one is the SVD of the $M \times M$ matrix \mathbf{S} to obtain \mathbf{v}_1 . There are several methods to realize SVD such

as the one based on QR decomposition, or the one based on Jacobi eigenvalue algorithm [15], and the corresponding complexity is $\mathcal{O}(M^3)$ [15]. The second one is to obtain the analog precoder \mathbf{a}_m . Since \mathbf{a}_m only presents the phase of \mathbf{v}_1 with amplitude of one, this part involves no extra computation. The third one comes from the digital precoder $\mathbf{d}_{m,m}$. This part requires one multiplication of the $1 \times M$ vector \mathbf{v}_1^H and the $M \times 1$ vector \mathbf{a}_m . Therefore, the complexity is $\mathcal{O}(M^2)$. The last one originates from computing the hybrid precoding matrix \mathbf{p}_m . Since there are only M non-zero elements in \mathbf{p}_m , the complexity of this part is $\mathcal{O}(M)$.

To sum up, the overall complexity of the proposed iterative hybrid precoding scheme in each iteration is $\mathcal{O}(M^3)$.

In contrast, the computational complexity of the fully analog precoding scheme [10] in each iteration is $\mathcal{O}(NM^2)$, which means the computational complexity of the proposed scheme is comparable with that of the fully analog precoding scheme. Additionally, the process of each iteration of the fully analog precoding scheme is hard to be parallelized [10], while our method can be easily implemented in a parallel manner, which indicates that lower hardware complexity can be achieved by the proposed iterative hybrid analog/digital precoding scheme.

IV. SIMULATION RESULTS

To evaluate the performance of the proposed iterative hybrid analog/digital precoding scheme, we provide the simulation results of the achievable channel capacity. Here we also provide the performance of the recently proposed fully analog precoding scheme [10] based on the sub-array structure and the optimal precoder utilizing the precoding vector $\mathbf{p}_m = \mathbf{v}_1$ as the benchmark for comparison. The simulation parameters are described as follows. The carrier frequency is set as 28GHz [1], and two typical mmWave massive MIMO configurations with $NM \times K = 128 \times 16$ ($M = 8$) and $NM \times K = 128 \times 32$ ($M = 4$) are considered, respectively. Both the transmit and receive antenna arrays are ULAs with antenna spacing $d = \lambda/2$. We generate the channel matrix according to the channel model [13] described in Section II. The AoAs/AoDs are assumed to follow the uniform distribution within $[0, 2\pi]$. The number of scattering propagation paths is set as $L = 10$.

Fig. 3 shows the capacity comparison between the fully analog precoding scheme [10] and the proposed iterative hybrid precoding scheme when $NM \times K = 128 \times 16$ ($M = 8$). Note that [10] also proposes a multimode strategy, where each antenna array can provide additional gain for the existing data stream instead of transmitting a new independent data stream. However, since the performance gain of this strategy is negligible at high SNR (> 5 dB) [10], here we only consider that each antenna array transmits an independent data stream. We can observe from Fig. 3 that the proposed iterative hybrid precoding scheme outperforms the fully analog one in the whole SNR range, and the performance gap becomes more obvious with the increasing SNR. For example, when SNR = 10 dB, the capacity gap between the proposed scheme and the conventional one is about 3 bps/Hz, while when SNR =

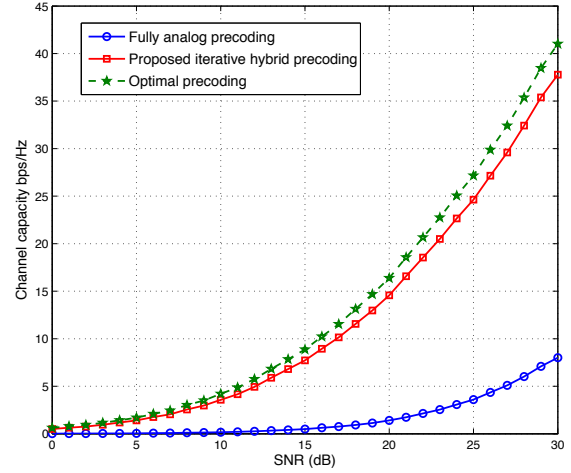


Fig. 3. Capacity comparison for an $NM \times K = 128 \times 16$ ($M = 8$) mmWave massive MIMO system.

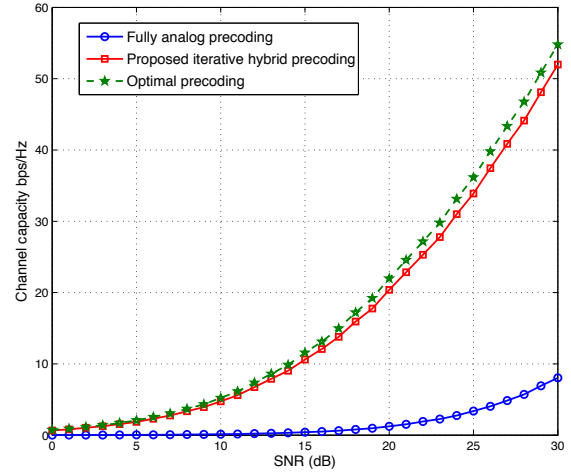


Fig. 4. Capacity comparison for an $NM \times K = 128 \times 32$ ($M = 4$) mmWave massive MIMO system.

30 dB, the capacity gap increases to 30 bps/Hz. What's more, Fig. 3 also shows that the proposed scheme is near-optimal, since the required SNR gap between the optimal precoding and the proposed one to achieve the same capacity is within 1 dB.

Fig. 4 shows the capacity comparison in the mmWave massive MIMO system with $NM \times K = 128 \times 32$ ($M = 4$). Comparing Fig. 3 and Fig. 4, we can find that with the increasing number of RF chains (i.e., N is increased from 16 to 32), the performance of the fully analog precoding becomes worse. For example, for the $NM \times K = 128 \times 16$ system, when SNR = 30 dB, the fully analog scheme can achieve 20% of the optimal capacity, while for the $NM \times K = 128 \times 32$ system, it can only achieve 16% of the optimal capacity. In contrast, the proposed iterative precoding scheme can achieve 88% and 96% of the optimal capacity for the $NM \times K = 128 \times 16$ and $NM \times K = 128 \times 32$ systems, respectively. This can

be explained by the fact that the performance improvement obtained by the digital precoder becomes obvious with the increasing number of RF chains, and the fully analog precoder system will suffer from a non-negligible performance loss. More importantly, we can also conclude that when the number of antennas connected to each RF chain (i.e., M) decreases, the proposed scheme will be more close to the optimal precoding scheme.

V. CONCLUSIONS

In this paper, we propose a near-optimal hybrid analog/digital precoding scheme with low complexity for mmWave massive MIMO systems. By considering each antenna array separately and exploiting the idea of SIC, we decompose the complicated capacity optimization problem into a series sub-optimal problems much easier to be solved, and realize the proposed scheme in an iterative way which can be implemented in a parallel manner. It shows that the computational complexity of the proposed scheme is $\mathcal{O}(M^3)$, which is comparable with the conventional fully analog precoding scheme. Simulation results verify that the performance of the proposed scheme is close to the optimal capacity, especially when M is relatively small (e.g., $M = 4$). Our further work will focus on extending the proposed hybrid precoding scheme from 2D scenario to 3D scenario, where the elevation angles of the AoAs/AoDs will be also taken into account.

ACKNOWLEDGMENTS

This work was supported by National Key Basic Research Program of China (Grant No. 2013CB329203), National High Technology Research and Development Program of China (Grant No. 2014AA01A704), National Nature Science Foundation of China (Grant Nos. 61271266 and 61201185), Science and Technology Foundation for Beijing Outstanding Doctoral Dissertation Supervisor (Grant No. 20121000303), and Foundation of Shenzhen government.

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 Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
- [2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
- [3] H. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2012.
- [4] M. H. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, Mar. 1983.
- [5] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.
- [6] Z. Xiao, "Suboptimal spatial diversity scheme for 60 GHz millimeter-wave WLAN," *IEEE Commun. Lett.*, vol. 17, no. 9, pp. 1790–1793, Sep. 2013.
- [7] T. Kim, J. Park, J.-Y. Seol, S. Jeong, J. Cho, and W. Roh, "Tens of Gbps support with mmwave beamforming systems for next generation communications," in *Proc. IEEE Global Communications Conference (GLOBECOM'13)*, Dec. 2013, pp. 3685–3690.
- [8] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
- [9] T. Kim, C. Kim, and J.-Y. Seol, "A low complexity hybrid beamforming algorithm for multi-beam diversity transmission," to appear in *IEEE Global Communications Conference (GLOBECOM'14)*, Dec. 2014.
- [10] O. El Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *Proc. IEEE Global Communications Conference (GLOBECOM'13)*, Dec. 2013, pp. 3476–3480.
- [11] F. Gholam, J. Via, and I. Santamaría, "Beamforming design for simplified analog antenna combining architectures," *IEEE Trans. Veh. Technol.*, vol. 60, no. 5, pp. 2373–2378, May 2011.
- [12] S. Han, C.-L. I, Z. Xu, and S. Wang, "Reference signals design for hybrid analog and digital beamforming," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1191–1193, Jul. 2014.
- [13] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," to appear in *IEEE J. Sel. Top. Signal Process.*
- [14] Z. Gao, L. Dai, and Z. Wang, "Structured compressive sensing based superimposed pilot design in downlink large-scale MIMO systems," *Electron. Lett.*, vol. 50, no. 12, pp. 896–898, Jun. 2014.
- [15] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012.
- [16] X. Gao, L. Dai, Y. Ma, and Z. Wang, "Low-complexity near-optimal signal detection for uplink large-scale MIMO systems," *Electron. Lett.*, vol. 50, no. 18, pp. 1326–1328, Aug. 2014.