Energy-Efficient Hybrid Analog and Digital Precoding for MmWave MIMO Systems With Large Antenna Arrays

Xinyu Gao, *Student Member, IEEE*, Linglong Dai, *Senior Member, IEEE*, Shuangfeng Han, *Member, IEEE*, Chih-Lin I, *Senior Member, IEEE*, and Robert W. Heath Jr., *Fellow, IEEE*

***Abstract*—Millimeter wave (mmWave) MIMO will likely use hybrid analog and digital precoding, which uses a small number of RF chains to reduce the energy consumption associated with mixed signal components like analog-to-digital components not to mention baseband processing complexity. However, most hybrid precoding techniques consider a fully connected architecture requiring a large number of phase shifters, which is also energy-intensive. In this paper, we focus on the more energy-efficient hybrid precoding with subconnected architecture, and propose a successive interference cancelation (SIC)-based hybrid precoding with near-optimal performance and low complexity. Inspired by the idea of SIC for multiuser signal detection, we first propose to decompose the total achievable rate optimization problem with nonconvex constraints into a series of simple subrate optimization problems, each of which only considers one subantenna array. Then, we prove that maximizing the achievable subrate of each subantenna array is equivalent to simply seeking a precoding vector sufficiently close (in terms of Euclidean distance) to the unconstrained optimal solution. Finally, we propose a low-complexity algorithm to realize SIC-based hybrid precoding, which can avoid the need for the singular value decomposition (SVD) and matrix inversion. Complexity evaluation shows that the complexity of SIC-based hybrid precoding is only about 10% as complex as that of the recently proposed spatially sparse precoding in typical mmWave MIMO systems. Simulation results verify that SIC-based hybrid precoding is near-optimal and enjoys higher energy efficiency than the spatially sparse precoding and the fully digital precoding.**

***Index Terms*—MIMO, mmWave communications, hybrid precoding, energy-efficient, 5G.**

Manuscript received July 27, 2015; accepted December 8, 2015. Date of publication March 31, 2016; date of current version May 11, 2016. This work was supported in part by the International Science and Technology Cooperation Program of China (Grant 2015DFG12760), in part by the National Natural Science Foundation of China (Grant 61571270 and Grant 61271266), in part by the Beijing Natural Science Foundation (Grant 4142027), and in part by the Foundation of Shenzhen Government. This work was presented in part at the IEEE International Conference on Communications (ICC), London, U.K., June 2015.

X. Gao and L. Dai are with the Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Beijing 100084, China (e-mail: xy-gao14@mails.tsinghua.edu.cn; daill@tsinghua.edu.cn).

S. Han and C.-L. I are with the Green Communication Research Center, China Mobile Research Institute, Beijing 100053, China (e-mail: hanshuangfeng@chinamobile.com; icl@chinamobile.com).

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2016.2549418

I. Introduction

THE integration of millimeter-wave (mmWave) and massive multiple-input multiple-output (MIMO) technique can achieve orders of magnitude increase in system throughput due to larger bandwidth [1] and higher spectral efficiency [2], which makes it promising for future 5G wireless communication systems [3]. On one hand, massive MIMO with a very large antenna array (e.g., 256 antennas) at the base station (BS) can simultaneously serve a set of users through the use of precoding [4]. It has been theoretically proved that massive MIMO can achieve orders of magnitude increase in spectral efficiency, since it can provide more multi-user gain [2]. On the other hand, mmWave with high frequencies enables such large antenna array in massive MIMO to be packed in small physical dimension [5]. Furthermore, the large antenna array can also provide sufficient array gain by precoding [6], [7] to overcome the free-space pathloss of mmWave signals and establish links with satisfying signal-to-noise ratio (SNR) [8].

For MIMO in conventional cellular frequency band (e.g., 2–3 GHz), precoding is entirely realized in the digital domain to cancel interference between different data streams. For a conventional digital precoding, each antenna requires a dedicated energy-intensive radio frequency (RF) chain (including digital-to-analog converter, up converter, etc.), whose energy consumption is a large part (about 250 mW per RF chain [9]) of the total energy consumption at mmWave frequencies due to the wide bandwidth. If the conventional digital precoding is applied in mmWave massive MIMO system with a large number of antennas, the corresponding large number of RF chains will bring high energy consumption, e.g., 16 W is required by a mmWave massive MIMO system with 64 antennas. To solve this problem, the hybrid analog and digital precoding has been proposed [10]. The key idea is to divide the conventional digital precoder into a small-size digital precoder (realized by a small number of RF chains) to cancel interference and a large-size analog precoder (realized by a large number of analog phase shifters (PSs)) to increase the antenna array gain. In this way, hybrid precoding can reduce the number of required RF chains without obvious performance loss, which makes it enjoy a much higher energy efficiency than digital precoding [10].

The existing hybrid precoding schemes can be divided into two categories. The first category of hybrid precoding based on spatially sparse precoding was proposed in [11]–[13], which formulated the achievable rate optimization problem as a sparse

approximation problem and solved it by the orthogonal matching pursuit (OMP) algorithm [14] to achieve the near-optimal performance. The second category of hybrid precoding based on codebooks was proposed in [15]–[17], which involved an iterative searching procedure among the predefined codebooks to find the optimal hybrid precoding matrix. However, these algorithms are all designed for the hybrid precoding with the fully-connected architecture, where each RF chain is connected to all BS antennas via PSs. As the number of BS antennas is very large (e.g., 256 as considered in [11]), the fully-connected architecture requires thousands of PSs, which may bring three additional limitations: 1) it consumes more energy for excitation like the giant phased array radar [18]; 2) it requires more energy to compensate for the insertion loss of PS [18]; 3) it involves higher computational complexity, which will also bring more energy consumption [19]. In contrast, the hybrid precoding with the sub-connected architecture, where each RF chain is connected to only a subset of BS antennas, can significantly reduce the number of required PSs. Therefore, the sub-connected architecture is expected to be more energy-efficient and easier to be implemented for mmWave MIMO systems. Unfortunately, designing the optimal hybrid precoding with the sub-connected architecture is still a challenging problem [10], [20], since such architecture changes the constraints on the original problem of hybrid precoding with the fully-connected architecture.

In this paper, we propose a successive interference cancelation (SIC)-based hybrid precoding with sub-connected architecture. The contributions of this paper can be summarized as follows.

1) Inspired by the idea of SIC derived for multi-user signal detection [21], we propose to decompose the total achievable rate optimization problem with non-convex constraints into a series of simple sub-rate optimization problems, each of which only considers one sub-antenna array. Then, we maximize the achievable sub-rate of each sub-antenna array one by one until the last sub-antenna array is considered.
2) We prove that maximizing the achievable sub-rate of each sub-antenna array is equivalent to seeking a precoding vector which has the smallest Euclidean distance to the unconstrained optimal solution. Based on this fact, we can easily obtain the optimal precoding vector for each sub-antenna array.
3) We further propose a low-complexity algorithm to realize the SIC-based precoding, which avoids the need for singular value decomposition (SVD) and matrix inversion. Complexity evaluation shows that the complexity of SIC-based precoding is only about 10% as complex as that of the spatially sparse precoding [11] in typical mmWave MIMO systems. Simulation results verify that the proposed SIC-based hybrid precoding is near-optimal and enjoys higher energy efficiency than the spatially sparse precoding [11] and the fully digital precoding.

It is worth pointing out that to the best of the authors' knowledge, our work in this paper is the first one that considers the hybrid precoding design with sub-connected architecture.

The rest of the paper is organized as follows. Section II briefly introduces the system model of mmWave MIMO.

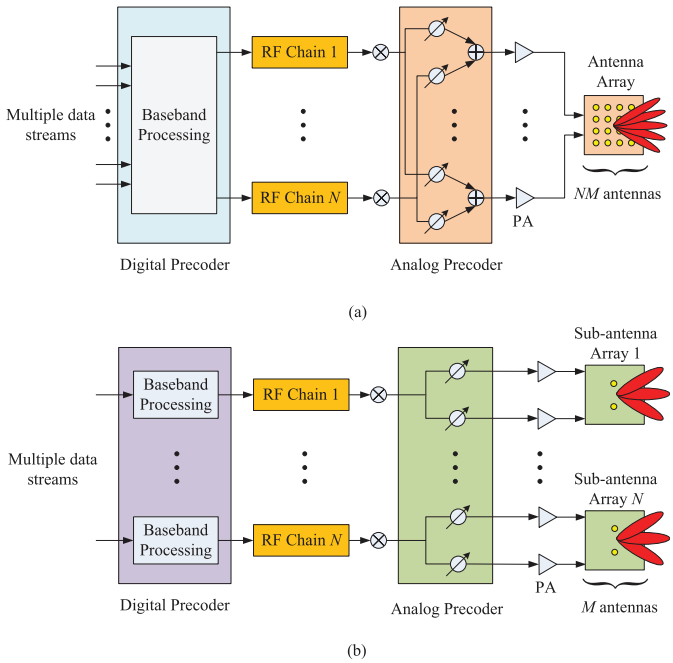

Fig. 1. Two typical architectures of the hybrid precoding in mmWave MIMO systems: (a) Fully-connected architecture, where each RF chain is connected to all BS antennas; (b) Sub-connected architecture, where each RF chain is connected to only a subset of BS antennas.

Section III specifies the proposed SIC-based hybrid precoding, together with the complexity evaluation. The simulation results of the achievable rate and energy efficiency 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 $|\cdot|$ denote the transpose, conjugate transpose, inversion, and determinant of a matrix, respectively; $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the l_1- and l_2-norm of a vector, respectively; $\|\cdot\|_F$ denotes the Frobenius norm of a matrix; $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ 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 illustrates two typical architectures for hybrid precoding in mmWave MIMO systems, i.e., the fully-connected architecture as shown in Fig. 1 (a) and the sub-connected architecture as shown in Fig. 1 (b). In both cases the BS has NM antennas but only N RF chains. From Fig. 1, we observe that the sub-connected architecture will likely be more energy-efficient, since it only requires NM PSs, while the fully-connected architecture requires N^2M PSs. To fully achieve the spatial multiplexing gain, the BS usually transmits N independent data streams to users employing K receive antennas [10].

In the sub-connected architecture as shown in Fig. 1 (b), N data streams in the baseband are precoded by the digital precoder $\mathbf{D}$. In cases where complexity is a concern, $\mathbf{D}$ can be further specialized to be a diagonal matrix as $\mathbf{D} = \mathrm{diag}\,[d_1, d_2, \ldots, d_N]$, where $d_n \in \mathbb{R}$ for $n = 1, 2, \ldots, N$ [10]. Then the role of $\mathbf{D}$ essentially performs some power allocation. After passing through the corresponding RF chain, the

digital-domain signal from each RF chain is delivered to only M PSs [22] to perform the analog precoding, which can be denoted by the analog weighting vector $\bar{\mathbf{a}}_n \in \mathbb{C}^{M\times 1}$, whose elements have the same amplitude $1/\sqrt{M}$ but different phases [22]. After the analog precoding, each data stream is finally transmitted by a sub-antenna array with only M antennas associated with the corresponding RF chain. Then, the received signal vector $\mathbf{y} = [y_1, y_2, \ldots, y_K]^T$ at the user in a narrowband system[1] can be presented as

$$\mathbf{y} = \sqrt{\rho}\mathbf{HADs} + \mathbf{n} = \sqrt{\rho}\mathbf{HPs} + \mathbf{n}, \tag{1}$$

where ρ is the average received power; $\mathbf{H} \in \mathbb{C}^{K\times NM}$ denotes the channel matrix, $\mathbf{A}$ is the $NM \times N$ analog precoding matrix comprising N analog weighting vectors $\{\bar{\mathbf{a}}_m\}_{m=1}^N$ as

$$\mathbf{A} = \begin{bmatrix} \bar{\mathbf{a}}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{a}}_2 & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \bar{\mathbf{a}}_N \end{bmatrix}_{NM\times N}, \tag{2}$$

$\mathbf{s} = [s_1, s_2, \ldots, s_N]^T$ represents the transmitted signal vector in the baseband. In this paper, we assume the widely used Gaussian signals [10]–[13], [15]–[17] with normalized signal power $\mathbb{E}(\mathbf{ss}^H) = \frac{1}{N}\mathbf{I}_N$, while the practical system with finite-alphabet inputs [23], [24] will be also briefly discussed in Section IV. $\mathbf{P} = \mathbf{AD}$ presents the hybrid precoding matrix of size $NM \times N$, which satisfies $\|\mathbf{P}\|_F \le N$ to meet the total transmit power constraint [11]. Finally, $\mathbf{n} = [n_1, n_2, \ldots, n_N]^T$ is an additive white Gaussian noise (AWGN) vector, whose entries follow the independent and identical distribution (i.i.d.) $\mathcal{CN}(0, \sigma^2)$.

It is known that mmWave channel $\mathbf{H}$ will not likely follow the rich-scattering model assumed at low frequencies due to the limited number of scatters in the mmWave prorogation environment [3]. In this paper, we adopt the geometric Saleh-Valenzuela channel model to embody the low rank and spatial correlation characteristics of mmWave communications [10]–[13], [15]–[17], [25] as

$$\mathbf{H} = \gamma \sum_{l=1}^{L} \alpha_l \Lambda_r\left(\phi_l^r, \theta_l^r\right) \Lambda_t\left(\phi_l^t, \theta_l^t\right) \mathbf{f}_r\left(\phi_l^r, \theta_l^r\right) \mathbf{f}_t^H\left(\phi_l^t, \theta_l^t\right), \tag{3}$$

where $\gamma = \sqrt{\frac{NMK}{L}}$ is a normalization factor, L is the number of effective channel paths corresponding to the limited number of scatters, and we usually have $L \le N$ for mmWave communication systems. $\alpha_l \in \mathbb{C}$ is the gain of the lth path. ϕ_l^t (θ_l^t) and ϕ_l^r (θ_l^r) are the azimuth (elevation) angles of departure and arrival (AoDs/AoAs), respectively. $\Lambda_t\left(\phi_l^t, \theta_l^t\right)$ and $\Lambda_r\left(\phi_l^r, \theta_l^r\right)$ denote the transmit and receive antenna array gain at a specific AoD and AoA, respectively. For simplicity but without loss of generality, $\Lambda_t\left(\phi_l^t, \theta_l^t\right)$ and $\Lambda_r\left(\phi_l^r, \theta_l^r\right)$ can be set as one within the range of AoDs/AoAs [26]. Finally, $\mathbf{f}_t\left(\phi_l^t, \theta_l^t\right)$ and $\mathbf{f}_r\left(\phi_l^r, \theta_l^r\right)$ are the antenna array response vectors depending on the antenna array structures at the BS and the user, respectively. For the uniform linear array (ULA) with U elements, the array response vector can be presented as [18]

$$\mathbf{f}_{\text{ULA}}(\phi) = \frac{1}{\sqrt{U}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin(\phi)}, \ldots, e^{j(U-1)\frac{2\pi}{\lambda}d\sin(\phi)}\right]^T, \tag{4}$$

where λ denotes the wavelength of the signal, and d is the antenna spacing. Note that here we abandon the subscripts $\{t, r\}$ in (3) and we also do not include θ since the ULA response vector is independent of the elevation angle. Additionally, when we consider the uniform planar array (UPA) with W_1 and W_2 elements ($W_1 W_2 = U$) on horizon and vertical, respectively, the array response vector can be given by [18]

$$\mathbf{f}_{\text{UPA}}(\phi, \theta) = \frac{1}{\sqrt{U}}\Big[1, \ldots, e^{j\frac{2\pi}{\lambda}d(x\sin(\phi)\sin(\theta)+y\cos(\theta))}, \ldots, e^{j\frac{2\pi}{\lambda}d((W_1-1)\sin(\phi)\sin(\theta)+(W_2-1)\cos(\theta))}\Big]^T, \tag{5}$$

where $0 \le x \le (W_1 - 1)$ and $0 \le y \le (W_2 - 1)$.

III. SIC-Based Hybrid Precoding for mmWave MIMO Systems

In this section, we propose a low-complexity SIC-based hybrid precoding to achieve the near-optimal performance. The evaluation of computational complexity is also provided to show its advantages over current solutions.

A. Structure of SIC-based hybrid precoding

In this paper, we aim to maximize the total achievable rate R of mmWave MIMO systems[2], while other criteria such as the max-min fairness criterion [27] are also of interest. Specifically, R can be expressed as [11]

$$R = \log_2\left(\left|\mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{HPP}^H\mathbf{H}^H\right|\right). \tag{6}$$

According to the system model (1) in Section II, since the hybrid precoding matrix $\mathbf{P}$ can be represented as $\mathbf{P} = \mathbf{AD} = \text{diag}\{\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_N\} \cdot \text{diag}\{d_1, \ldots, d_N\}$, there are three constraints for the design of $\mathbf{P}$:

Constraint 1: $\mathbf{P}$ should be a block diagonal matrix similar to the form of $\mathbf{A}$ as shown in (2), i.e., $\mathbf{P} = \text{diag}\{\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_N\}$, where $\bar{\mathbf{p}}_n = d_n\bar{\mathbf{a}}_n$ is the $M \times 1$ non-zero vector of the nth column $\mathbf{p}_n$ of $\mathbf{P}$, i.e., $\mathbf{p}_n = \left[\mathbf{0}_{1\times M(n-1)},\ \bar{\mathbf{p}}_n,\ \mathbf{0}_{1\times M(N-n)}\right]^T$;

Constraint 2: The non-zero elements of each column of $\mathbf{P}$ should have the same amplitude, since the digital precoding matrix $\mathbf{D}$ is a diagonal matrix, and the amplitude of non-zero elements of the analog precoding matrix $\mathbf{A}$ is fixed to $1/\sqrt{M}$;

Constraint 3: The Frobenius norm of $\mathbf{P}$ should satisfy $\|\mathbf{P}\|_F \le N$ to meet the total transmit power constraint, where N is the number of RF chains equal to the number of transmitted data streams.

Unfortunately, these non-convex constraints on $\mathbf{P}$ make maximizing the total achievable rate (6) very difficult to be solved.

[1] While mmWave systems are expected to be broadband as in prior work [3], the narrowband system can be regarded as a reasonable first step. The extension to broadband system is an interesting topic.

[2] The maximization sum-rate criterion can also suppress the interference as much as possible, and the mathematical quantification of such interference will be an important topic for future work.

However, based on the special block diagonal structure of the hybrid precoding matrix $\mathbf{P}$, we observe that the precoding on different sub-antenna arrays are independent. This inspires us to decompose the total achievable rate (6) into a series of sub-rate optimization problems, each of which only considers one sub-antenna array.

In particular, we can divide the hybrid precoding matrix $\mathbf{P}$ as $\mathbf{P} = \left[\mathbf{P}_{N-1}\ \mathbf{p}_N\right]$, where $\mathbf{p}_N$ is the Nth column of $\mathbf{P}$, and $\mathbf{P}_{N-1}$ is an $NM \times (N-1)$ matrix containing the first $(N-1)$ columns of $\mathbf{P}$. Then, the total achievable rate R in (6) can be rewritten as

$$\begin{aligned} R &= \log_2\left(\left|\mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}\mathbf{P}^H\mathbf{H}^H\right|\right) \\ &= \log_2\left(\left|\mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\left[\mathbf{P}_{N-1}\ \mathbf{p}_N\right]\left[\mathbf{P}_{N-1}\ \mathbf{p}_N\right]^H\mathbf{H}^H\right|\right) \\ &= \log_2\left(\left|\mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_{N-1}\mathbf{P}_{N-1}^H\mathbf{H}^H \right.\right. \\ &\qquad \left.\left. + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{p}_N\mathbf{p}_N^H\mathbf{H}^H\right|\right) \\ &\overset{(a)}{=} \log_2\left(\left|\mathbf{T}_{N-1}\right|\right) + \log_2\left(\left|\mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{T}_{N-1}^{-1}\mathbf{H}\mathbf{p}_N\mathbf{p}_N^H\mathbf{H}^H\right|\right) \\ &\overset{(b)}{=} \log_2\left(\left|\mathbf{T}_{N-1}\right|\right) + \log_2\left(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_N^H\mathbf{H}^H\mathbf{T}_{N-1}^{-1}\mathbf{H}\mathbf{p}_N\right), \end{aligned} \tag{7}$$

where (a) is obtained by defining the auxiliary matrix $\mathbf{T}_{N-1} = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_{N-1}\mathbf{P}_{N-1}^H\mathbf{H}^H$, and (b) is true due to the fact that $|\mathbf{I} + \mathbf{X}\mathbf{Y}| = |\mathbf{I} + \mathbf{Y}\mathbf{X}|$ by defining $\mathbf{X} = \mathbf{T}_{N-1}^{-1}\mathbf{H}\mathbf{p}_N$ and $\mathbf{Y} = \mathbf{p}_N^H\mathbf{H}^H$. Note that the second term $\log_2\left(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_N^H\mathbf{H}^H\mathbf{T}_{N-1}^{-1}\mathbf{H}\mathbf{p}_N\right)$ on the right side of (7) is the achievable sub-rate of the Nth sub-antenna array, while the first term $\log_2\left(|\mathbf{T}_{N-1}|\right)$ shares the same form as (6). This observation implies that we can further decompose $\log_2\left(|\mathbf{T}_{N-1}|\right)$ using the similar method in (7) as

$$\log_2\left(|\mathbf{T}_{N-2}|\right) + \log_2\left(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_{N-1}^H\mathbf{H}^H\mathbf{T}_{N-2}^{-1}\mathbf{H}\mathbf{p}_{N-1}\right).$$

Then, after N such decompositions, the total achievable rate R in (6) can be presented as

$$R = \sum_{n=1}^{N}\log_2\left(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_n^H\mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{p}_n\right), \tag{8}$$

where we have $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$ and $\mathbf{T}_0 = \mathbf{I}_N$. From (8), we observe that the total achievable rate optimization problem can be transformed into a series of sub-rate optimization problems of sub-antenna arrays, which can be optimized one by one[3]. After that, inspired by the idea of SIC for multi-user signal detection [21], we can optimize the achievable sub-rate of the first sub-antenna array and update the matrix $\mathbf{T}_1$. Then, the similar method can be utilized to optimize the achievable sub-rate of the second sub-antenna array. Such procedure will be executed until the last sub-antenna array is considered. Fig. 2 shows the diagram of the proposed SIC-based hybrid precoding. Next, we will discuss how to optimize the achievable sub-rate of each sub-antenna array.

[3]Note that different precoding orders of sub-antenna arrays will lead to the same performance, since the total achievable rate R in (8) can be exactly represented by the summation of the sub-rate of each sub-antenna array without any performance loss.

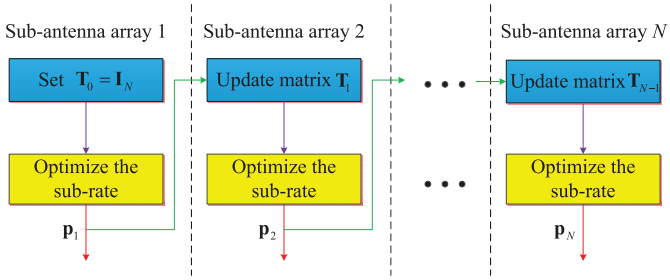

Fig. 2. Diagram of the proposed SIC-based hybrid precoding.

B. Solution to the Subrate Optimization Problem

In this subsection, we focus on the sub-rate optimization problem of the nth sub-antenna array, which can be directly applied to other sub-antenna arrays. According to (8), the sub-rate optimization problem of the nth sub-antenna array by designing the nth precoding vector $\mathbf{p}_n$ can be stated as

$$\mathbf{p}_n^{\text{opt}} = \arg\max_{\mathbf{p}_n \in \mathcal{F}} \log_2\left(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_n^H\mathbf{G}_{n-1}\mathbf{p}_n\right), \tag{9}$$

where $\mathbf{G}_{n-1}$ is defined as $\mathbf{G}_{n-1} = \mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}$, $\mathcal{F}$ is the set of all feasible vectors satisfying the three constraints described in Section III-A. Note that the nth precoding vector $\mathbf{p}_n$ only has M non-zero elements from the $(M(n-1)+1)$th one to the (Mn)th one. Therefore, the sub-rate optimization problem (9) can be equivalently written as

$$\bar{\mathbf{p}}_n^{\text{opt}} = \arg\max_{\bar{\mathbf{p}}_n \in \bar{\mathcal{F}}} \log_2\left(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H\bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\right), \tag{10}$$

where $\bar{\mathcal{F}}$ includes all possible $M \times 1$ vectors satisfying *Constraint 2* and *Constraint 3*, $\bar{\mathbf{G}}_{n-1}$ of size $M \times M$ is the corresponding sub-matrix of $\mathbf{G}_{n-1}$ by only keeping the rows and columns of $\mathbf{G}_{n-1}$ from the $(M(n-1)+1)$th one to the (Mn)th one, which can be presented as

$$\bar{\mathbf{G}}_{n-1} = \mathbf{R}\mathbf{G}_{n-1}\mathbf{R}^H = \mathbf{R}\mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{R}^H, \tag{11}$$

where $\mathbf{R} = \left[\mathbf{0}_{M\times M(n-1)}\ \mathbf{I}_M\ \mathbf{0}_{M\times M(N-n)}\right]$ is the corresponding selection matrix.

Define the singular value decomposition (SVD) of the Hermitian matrix $\bar{\mathbf{G}}_{n-1}$ as $\bar{\mathbf{G}}_{n-1} = \mathbf{V}\Sigma\mathbf{V}^H$, where Σ is an $M \times M$ diagonal matrix containing the singular values of $\bar{\mathbf{G}}_{n-1}$ in a decreasing order, and $\mathbf{V}$ is an $M \times M$ unitary matrix. It is known that the optimal unconstrained precoding vector of (10) is the first column $\mathbf{v}_1$ of $\mathbf{V}$, i.e., the first right singular vector of $\bar{\mathbf{G}}_{n-1}$ [11]. However, according to the constraints mentioned in Section III-A, we cannot directly choose $\bar{\mathbf{p}}_n^{\text{opt}}$ as $\mathbf{v}_1$ since the elements of $\mathbf{v}_1$ do not obey the constraint of same amplitude (i.e., *Constraint 2*). To find a feasible solution to the sub-rate optimization problem (10), we need to further convert (10) into another form, which is given by the following **Proposition 1**.

Proposition 1: The optimization problem (10)

$$\bar{\mathbf{p}}_n^{\text{opt}} = \arg\max_{\bar{\mathbf{p}}_n \in \bar{\mathcal{F}}} \log_2\left(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H\bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\right)$$

is equivalent to the following problem

$$\bar{\mathbf{p}}_n^{\text{opt}} = \arg\min_{\bar{\mathbf{p}}_n \in \bar{\mathcal{F}}} \|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2, \tag{12}$$

where $\mathbf{v}_1$ is the first right singular vector of $\bar{\mathbf{G}}_{n-1}$.

Proof: See Appendix A. ■

Proposition 1 indicates that we can find a feasible precoding vector $\bar{\mathbf{p}}_n$, which is sufficiently close (in terms of Euclidean distance) to the optimal but unpractical precoding vector $\mathbf{v}_1$, to maximize the achievable sub-rate of the nth sub-antenna array. Since $\bar{\mathbf{p}}_n = d_n \bar{\mathbf{a}}_n$ according to (1), the target $\|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2$ in (12) can be rewritten as

$$\begin{aligned}
&\|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2 \\
&= (\mathbf{v}_1 - d_n\bar{\mathbf{a}}_n)^H (\mathbf{v}_1 - d_n\bar{\mathbf{a}}_n) \\
&= \mathbf{v}_1^H\mathbf{v}_1 + d_n^2\bar{\mathbf{a}}_n^H\bar{\mathbf{a}}_n - 2d_n\text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right) \\
&\overset{(a)}{=} 1 + d_n^2 - 2d_n\text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right) \\
&= \left(d_n - \text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)\right)^2 + \left(1 - \left[\text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)\right]^2\right),
\end{aligned} \tag{13}$$

where (a) is obtained based on the facts that $\mathbf{v}_1^H\mathbf{v}_1 = 1$ and $\bar{\mathbf{a}}_n^H\bar{\mathbf{a}}_n = 1$, since $\mathbf{v}_1$ is the first column of the unitary matrix $\mathbf{V}$ and each element of $\bar{\mathbf{a}}_n$ has the same amplitude $1/\sqrt{M}$.

From (13), we observe that the distance between $\bar{\mathbf{p}}_n$ and $\mathbf{v}_1$ consists of two parts. The first one is $\left(d_n - \text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)\right)^2$, which can be minimized to zero by choosing $d_n = \text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)$. The second one is $\left(1 - \left[\text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)\right]^2\right)$, which can be minimized by maximizing $\left|\text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)\right|$. Note that both $\bar{\mathbf{a}}_n$ and $\mathbf{v}_1$ have a fixed power of one, i.e., $\mathbf{v}_1^H\mathbf{v}_1 = 1$ and $\bar{\mathbf{a}}_n^H\bar{\mathbf{a}}_n = 1$. Therefore, the optimal $\bar{\mathbf{a}}_n^{\text{opt}}$ to maximize $\left|\text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right)\right|$ is

$$\bar{\mathbf{a}}_n^{\text{opt}} = \frac{1}{\sqrt{M}} e^{j\text{angle}(\mathbf{v}_1)}, \tag{14}$$

where $\text{angle}(\mathbf{v}_1)$ denotes the phase vector of $\mathbf{v}_1$, i.e., each element of $\bar{\mathbf{a}}_n^{\text{opt}}$ shares the same phase as the corresponding element of $\mathbf{v}_1$[4]. Accordingly, the optimal choice of d_n^{opt} is

$$d_n^{\text{opt}} = \text{Re}\left(\mathbf{v}_1^H\bar{\mathbf{a}}_n\right) = \frac{1}{\sqrt{M}}\text{Re}\left(\mathbf{v}_1^H e^{j\text{angle}(\mathbf{v}_1)}\right) = \frac{\|\mathbf{v}_1\|_1}{\sqrt{M}}. \tag{15}$$

Based on (14) and (15), the optimal solution $\bar{\mathbf{p}}_n^{\text{opt}}$ to the optimization problem (12) (or equivalently (10)) can be obtained by

$$\bar{\mathbf{p}}_n^{\text{opt}} = d_n^{\text{opt}}\bar{\mathbf{a}}_n^{\text{opt}} = \frac{1}{M}\|\mathbf{v}_1\|_1 e^{j\text{angle}(\mathbf{v}_1)}. \tag{16}$$

It is worth pointing out that $\mathbf{v}_1$ is the first column of the unitary matrix $\mathbf{V}$, each element v_i of $\mathbf{v}_1$ (for $i = 1, \ldots, M$) has the amplitude less than one. Therefore, we have $\left\|\bar{\mathbf{p}}_n^{\text{opt}}\right\|_2^2 \leq 1$.

[4]It is worth pointing out that the analog precoding vector $\bar{\mathbf{a}}_n$ can be also restricted to a DFT vector to save the overhead of quantization for limited feedback systems [11]. However, since more constraints are set on the design of analog precoding, such scheme may lead to some performance loss compared to the proposed one (14).

Note that for all sub-antenna arrays, the optimal solution $\bar{\mathbf{p}}_n^{\text{opt}}$ for $n = 1, 2, \ldots, N$ have a similar form. Thus, we can conclude that

$$\left\|\mathbf{P}^{\text{opt}}\right\|_F^2 = \left\|\text{diag}\left\{\bar{\mathbf{p}}_1^{\text{opt}}, \ldots, \bar{\mathbf{p}}_N^{\text{opt}}\right\}\right\|_F^2 \leq N, \tag{17}$$

which demonstrates that the total transmit power constraint (*Constraint 3*) is satisfied.

After we have acquired $\bar{\mathbf{p}}_n^{\text{opt}}$ for the nth sub-antenna array, the matrices $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$ (8) and $\bar{\mathbf{G}}_n = \mathbf{R}\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}\mathbf{R}^H$ (11) can be updated. Then, the method described above for the nth sub-antenna array can be reused again to optimize the achievable sub-rate of the $(n+1)$th sub-antenna array. To sum up, solving the sub-rate optimization problem of the nth sub-antenna array consists of the following three steps.

Step 1: Execute the SVD of $\bar{\mathbf{G}}_{n-1}$ to obtain $\mathbf{v}_1$;

Step 2: Let $\bar{\mathbf{p}}_n^{\text{opt}} = \frac{1}{M}\|\mathbf{v}_1\|_1 e^{j\text{angle}(\mathbf{v}_1)}$ as the optimal solution to the current nth sub-antenna array;

Step 3: Update matrices $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$ and $\bar{\mathbf{G}}_n = \mathbf{R}\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}\mathbf{R}^H$ for the next $(n+1)$th sub-antenna array.

Note that although we can obtain the optimal solution $\bar{\mathbf{p}}_n^{\text{opt}}$ by the method above, we need to compute the SVD of $\bar{\mathbf{G}}_{n-1}$ (*Step 1*) and the matrix $\bar{\mathbf{G}}_n$ (*Step 3*) involving the matrix inversion of large size, which leads to high computational complexity as well as high energy consumption for computation [19]. To this end, next we will propose a low-complexity algorithm to obtain $\bar{\mathbf{p}}_n^{\text{opt}}$ without the complicated SVD and matrix inversion.

C. Low-Complexity Algorithm to Obtain the Optimal Solution

We start by considering how to avoid the SVD involving high computational complexity as well as a large number of divisions, which are difficult to be implemented in hardware. We observe from *Step 1* that the SVD of $\bar{\mathbf{G}}_{n-1}$ does not need to be computed to acquire $\mathbf{\Sigma}$ and $\mathbf{V}$, as only the first column $\mathbf{v}_1$ of $\mathbf{V}$ is enough to obtain $\bar{\mathbf{p}}_n^{\text{opt}}$. This observation inspires us to exploit the simple power iteration algorithm [28], which is used to compute the largest eigenvalue and the corresponding eigenvector of a diagonalizable matrix. Since $\bar{\mathbf{G}}_{n-1}$ is a Hermitian matrix, it follows that: 1) $\bar{\mathbf{G}}_{n-1}$ is also a diagonalizable matrix; 2) The singular values (right singular vectors) of $\bar{\mathbf{G}}_{n-1}$ are the same as the eigenvalues (eigenvectors). Therefore, the power iteration algorithm can be also utilized to compute $\mathbf{v}_1$ as well as the largest singular value Σ_1 of $\bar{\mathbf{G}}_{n-1}$ with low complexity.

More specifically, as shown by the pseudo-code in **Algorithm 1**, the power iteration algorithm starts with an initial solution $\mathbf{u}^{(0)} \in \mathbb{C}^{M\times 1}$, which is usually set as $[1, 1, \ldots, 1]^T$ without loss of generality [28]. In each iteration, it first computes the auxiliary vector $\mathbf{z}^{(s)} = \bar{\mathbf{G}}_{n-1}\mathbf{u}^{(s-1)}$ (s is the number of iterations) and then extracts the element of $\mathbf{z}^{(s)}$ having the largest amplitude as $m^{(s)}$. After that, $\mathbf{u}^{(s)}$ is updated as $\mathbf{u}^{(s)} = \frac{\mathbf{z}^{(s)}}{m^{(s)}}$ for the next iteration. The power iteration algorithm will stop until the number of iterations reaches the predefined number S. Finally, $m^{(S)}$ and $\mathbf{u}^{(S)}/\left\|\mathbf{u}^{(S)}\right\|_2$ will be output as the largest singular value Σ_1 and the first right singular vector $\mathbf{v}_1$ of $\bar{\mathbf{G}}_{n-1}$, respectively.

Algorithm 1. Power iteration algorithm

Input: (1) $\bar{\mathbf{G}}_{n-1}$;
(2) Initial solution $\mathbf{u}^{(0)}$;
(3) Maximum number of iteration S
for $1 \le s \le S$
1) $\mathbf{z}^{(s)} = \bar{\mathbf{G}}_{n-1}\mathbf{u}^{(s-1)}$
2) $m^{(s)} = \arg\max_{z_i^{(s)}} \left| z_i^{(s)} \right|$
3) **if** $1 \le s \le 2$
$n^{(s)} = m^{(s)}$
else
$n^{(s)} = \frac{m^{(s)}m^{(s-2)} - \left(m^{(s-1)}\right)^2}{m^{(s)} - 2m^{(s-1)} + m^{(s-2)}}$
end if
4) $\mathbf{u}^{(s)} = \frac{\mathbf{z}^{(s)}}{n^{(s)}}$
end for
Output: (1) The largest singular value $\Sigma_1 = n^{(S)}$
(2) The first singular vector $\mathbf{v}_1 = \frac{\mathbf{u}^{(S)}}{\left\|\mathbf{u}^{(S)}\right\|_2}$

According to [28], we know that

$$m^{(s)} = \Sigma_1 \left[1 + \mathcal{O}\left(\left(\frac{\Sigma_2}{\Sigma_1} \right)^s \right) \right], \tag{18}$$

where Σ_2 is the second largest singular value of $\bar{\mathbf{G}}_{n-1}$. From (18), we conclude that $m^{(s)}$ will converges to Σ_1 as long as $\Sigma_1 \neq \Sigma_2$. Similarly, when $\Sigma_1 \neq \Sigma_2$, $\mathbf{u}^{(s)}/\left\|\mathbf{u}^{(s)}\right\|_2$ will also converge to $\mathbf{v}_1$, i.e.,

$$\lim_{s\to\infty} m^{(s)} = \Sigma_1, \quad \lim_{s\to\infty} \frac{\mathbf{u}^{(s)}}{\left\|\mathbf{u}^{(s)}\right\|_2} = \mathbf{v}_1. \tag{19}$$

Although the power iteration algorithm is convergent, its convergence rate may be slow if $\Sigma_1 \approx \Sigma_2$ based on (18). To solve this problem, we propose to utilize the Aitken acceleration method [29] to further increase the convergence rate of the power iteration algorithm. Specifically, we can compute

$$\begin{cases} n^{(s)} = m^{(s)}, & \text{for } 1 \le s \le 2, \\ n^{(s)} = \frac{m^{(s)}m^{(s-2)} - \left(m^{(s-1)}\right)^2}{m^{(s)} - 2m^{(s-1)} + m^{(s-2)}}, & \text{for } 2 < s \le S. \end{cases} \tag{20}$$

Then, $\mathbf{u}^{(s)}$ and Σ_1 will be correspondingly changed to $\mathbf{u}^{(s)} = \frac{\mathbf{z}^{(s)}}{n^{(s)}}$ and $\Sigma_1 = n^{(S)}$, respectively.

Next, we will focus on how to reduce the complexity to compute the matrices $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$ and $\bar{\mathbf{G}}_n = \mathbf{R}\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}\mathbf{R}^H$, which involve the complicated matrix-to-matrix multiplication and matrix inversion of large size. In particular, with some standard mathematical manipulations, the computation of $\bar{\mathbf{G}}_n$ can be significantly simplified as shown by the following **Proposition 2**.

Proposition 2: The matrix $\bar{\mathbf{G}}_n = \mathbf{R}\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}\mathbf{R}^H$, where $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$, can be simplified as

$$\bar{\mathbf{G}}_n \approx \bar{\mathbf{G}}_{n-1} - \frac{\frac{\rho}{N\sigma^2}\Sigma_1^2\mathbf{v}_1\mathbf{v}_1^H}{1 + \frac{\rho}{N\sigma^2}\Sigma_1}, \tag{21}$$

where $\Sigma_1$and $\mathbf{v}_1$ are the largest singular value and first right singular vector of $\bar{\mathbf{G}}_{n-1}$, respectively.

Proof: See Appendix B. ■

Proposition 2 implies that we can simply exploit Σ_1 and $\mathbf{v}_1$ that have been obtained by **Algorithm 1** as described above to update $\bar{\mathbf{G}}_n$, which only involves one vector-to-vector multiplication instead of the complicated matrix-to-matrix multiplication and matrix inversion. Note that the evaluation of computational complexity will be discussed in detail in Section III-E.

D. Summary of the Proposed SIC-Based Hybrid Precoding

Based on the discussion so far, the pseudo-code of the proposed SIC-based hybrid precoding can be summarized in **Algorithm 2**, which can be explained as follows. The proposed SIC-based hybrid precoding starts by computing the largest singular value Σ_1 and first right singular vector $\mathbf{v}_1$ of $\bar{\mathbf{G}}_{n-1}$, which is achieved by **Algorithm 1**. After that, according to Section III-B, the optimal precoding vector for the nth sub-antenna array can be obtained by utilizing $\mathbf{v}_1$. Finally, based on **Proposition 2**, $\bar{\mathbf{G}}_n$ can be updated with low complexity for the next iteration. This procedure will be executed until the last (Nth) sub-antenna array is considered. Finally, after N iterations, the optimal digital, analog, and hybrid precoding matrices $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{P}$ can be obtained, respectively.

Algorithm 2. SIC-based hybrid precoding

Input: $\bar{\mathbf{G}}_0$
for $1 \le n \le N$
1) Compute $\mathbf{v}_1$ and Σ_1 of $\bar{\mathbf{G}}_{n-1}$ by **Algorithm 1**
2) $\bar{\mathbf{a}}_n^{\text{opt}} = \frac{1}{\sqrt{M}} e^{j\text{angle}(\mathbf{v}_1)}$, $d_n^{\text{opt}} = \frac{\|\mathbf{v}_1\|_1}{\sqrt{M}}$,
$\bar{\mathbf{p}}_n^{\text{opt}} = \frac{1}{M}\|\mathbf{v}_1\|_1 e^{j\text{angle}(\mathbf{v}_1)}$ (14)-(16)
3) $\bar{\mathbf{G}}_n = \bar{\mathbf{G}}_{n-1} - \frac{\frac{\rho}{N\sigma^2}\Sigma_1^2\mathbf{v}_1\mathbf{v}_1^H}{1+\frac{\rho}{N\sigma^2}\Sigma_1}$ (**Proposition 2**)
end for
Output: (1) $\mathbf{D} = \text{diag}\left\{d_1^{\text{opt}}, \ldots, d_N^{\text{opt}}\right\}$
(2) $\mathbf{A} = \text{diag}\left\{\bar{\mathbf{a}}_1^{\text{opt}}, \ldots, \bar{\mathbf{a}}_N^{\text{opt}}\right\}$
(3) $\mathbf{P} = \mathbf{A}\mathbf{D}$

It is worth pointing out that the idea of SIC-based hybrid precoding can be also extended to the combining at the user following the similar logic in [11]. When the number of RF chains at the BS is smaller than that at the user, we first compute the optimal hybrid precoding matrix $\mathbf{P}$ according to **Algorithm 2**, where we assume that the combining matrix $\mathbf{Q} = \mathbf{I}$. Then, given the effective channel matrix $\mathbf{HP}$, we can similarly obtain the optimal hybrid combining matrix $\mathbf{Q}$ by referring to **Algorithm 2**, where the input $\bar{\mathbf{G}}_0$ and the optimal unconstrained solution $\mathbf{v}_1$ should be correspondingly replaced. Conversely, when the number of RF chains at the BS is larger than that at the user, we can assume $\mathbf{P} = \mathbf{I}$ and obtain the optimal hybrid combining matrix $\mathbf{Q}$. After that, the optimal precoding matrix $\mathbf{P}$ can be acquired given the effective channel matrix $\mathbf{QH}$. Additionally, to further improve the performance, we can combine the above method with the "Ping-pong" algorithm [22], which involves an iteration procedure between the BS and the user, to jointly seek the optimal hybrid precoding and combining matrices pair. Further discussion about hybrid combining will be left for future work.

TABLE I
COMPLEXITY COMPARISON

	Number of Multiplications	Number of Divisions
SIC-based hybrid precoding	$\mathcal{O}\left(M^2\left(NS+K\right)\right)$	$\mathcal{O}\left(2NS\right)$
Spatially sparse precoding [11]	$\mathcal{O}\left(N^4M+N^2L^2+N^2M^2L\right)$	$\mathcal{O}\left(2N^3\right)$

E. Complexity Evaluation

In this subsection, we provide the complexity evaluation of the proposed SIC-based hybrid precoding in terms of the required numbers of complex multiplications and divisions. From **Algorithm 2**, we observe that the complexity of SIC-based hybrid precoding comes from the following four parts:

1) The first one originates from the computation of $\bar{\mathbf{G}}_0 = \mathbf{R}\mathbf{H}^H\mathbf{H}\mathbf{R}^H$ according to (11). Note that $\mathbf{R}$ is a selection matrix and $\mathbf{H}$ has the size $K \times NM$. Therefore, this part involves KM^2 times of multiplications without any division.
2) The second one is from executing **Algorithm 1**. It is observed that in each iteration we need to compute a matrix-to-vector multiplication $\mathbf{z}^{(s)} = \bar{\mathbf{G}}_{n-1}\mathbf{u}^{(s-1)}$ together with the Aitken acceleration method (20). Therefore, we totally require $S\left(M^2+2\right)-4$ and $(2S-2)$ times of multiplications and divisions, respectively.
3) The third one stems from acquiring the optimal solution $\bar{\mathbf{p}}_n^{\text{opt}}$ in step 2 of **Algorithm 2**. We find that this part is quite simple, which only needs 2 times of multiplications without any division, since $\mathbf{v}_1$ has been obtained and $\frac{1}{\sqrt{M}}$ is a fixed constant.
4) The last one comes from the update of $\bar{\mathbf{G}}_n$. According to **Proposition 2**, we know that this part mainly involves a outer product $\mathbf{v}_1\mathbf{v}_1^H$. Thus, it requires M^2 times of multiplications with only one division.

To sum up, the proposed SIC-based hybrid precoding approximately requires $\mathcal{O}\left(M^2\left(NS+K\right)\right)$ times of multiplications and $\mathcal{O}\left(2NS\right)$ times of divisions. Table I provides the complexity comparison between SIC-based hybrid precoding and the recently proposed spatially sparse precoding [11], which requires $\mathcal{O}\left(N^4M+N^2L^2+N^2M^2L\right)$ times of multiplications and $\mathcal{O}\left(2N^3\right)$ times of divisions. Here, L is the number of effective channel paths as defined in (3). Considering the typical mmWave MIMO system with $N=8$, $M=8$, $K=16$, $L=3$ [11], we observe that the complexity of SIC-based hybrid precoding is about 4×10^3 times of multiplications and 10^2 times of divisions, where we set $S=5$ (note that $S\geq5$ is usually sufficient to guarantee the performance, which is verified through intensive simulations). By contrast, the complexity of the spatially sparse precoding is about 5×10^4 times of multiplications and 10^3 times of divisions. Therefore, the proposed SIC-based hybrid precoding enjoys much lower complexity, which is only about 10% as complex as that of the spatially sparse precoding.

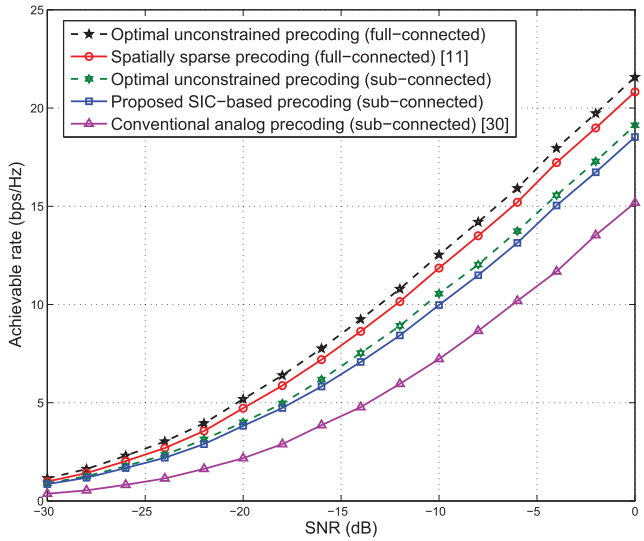

Fig. 3. Achievable rate comparison for an $NM \times K = 64 \times 16$ $(N=8)$ mmWave MIMO system.

IV. SIMULATION RESULTS

In this section, we provide the simulation results of the achievable rate and energy efficiency to evaluate the performance of the proposed SIC-based hybrid precoding. We compare the performance of SIC-based hybrid precoding with the recently proposed spatially sparse precoding [11] and the optimal unconstrained precoding based on the SVD of the channel matrix, which are both with fully-connected architecture. Additionally, we also include the conventional analog precoding [30] and the optimal unconstrained precoding (i.e., $\bar{\mathbf{p}}_n^{\text{opt}} = \mathbf{v}_1$) which are both with sub-connected architecture as benchmarks for comparison.

The simulation parameters are described as follows. We generate the channel matrix according to the channel model [31] described in Section II. The number of effective channel paths is $L=3$ [11]. The carrier frequency is set as 28GHz [15]. Both the transmit and receive antenna arrays are ULAs with antenna spacing $d=\lambda/2$. Since the BS usually employs the directional antennas to eliminate interference and increase antenna gain [3], the AoDs are assumed to follow the uniform distribution within $\left[-\frac{\pi}{6},\frac{\pi}{6}\right]$. Meanwhile, due to the random position of users, we assume that the AoAs follow the uniform distribution within $[-\pi,\pi]$, which means the omni-directional antennas are adopted by users. Furthermore, we set the maximum number of iterations $S=5$ to run **Algorithm 2**. Finally, SNR is defined as $\frac{\rho}{\sigma^2}$.

Firstly, we consider the perfect channel state information (CSI) scenario. Fig. 3 shows the achievable rate comparison in mmWave MIMO system, where $NM \times K = 64 \times 16$ and the number of RF chains is $N=8$. We observe from Fig. 3 that the proposed SIC-based hybrid precoding outperforms the conventional analog precoding with sub-connected architecture in whole simulated SNR range. Meanwhile, Fig. 3 also verifies the near-optimal performance of SIC-based hybrid precoding, since it can achieve about 99% of the rate achieved by the optimal unconstrained precoding with sub-connected architecture.

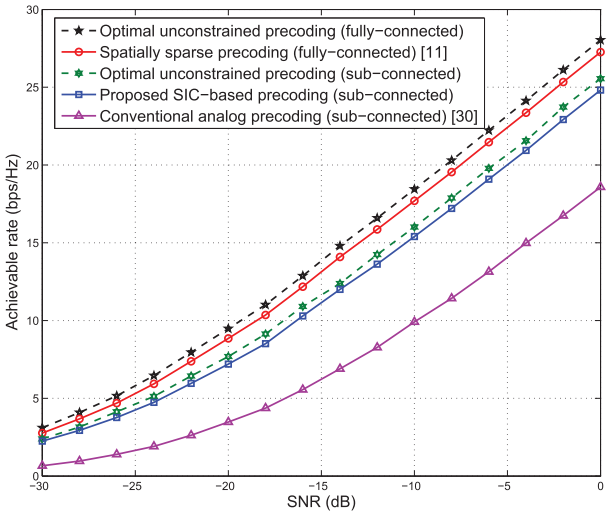

Fig. 4. Achievable rate comparison for an $NM \times K = 128 \times 32$ ($N = 16$) mmWave MIMO system.

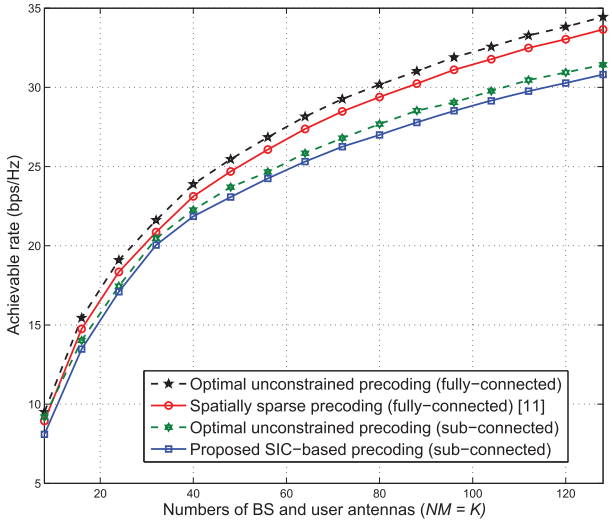

Fig. 5. Achievable rate comparison against the numbers of BS and user antennas ($NM = K$), where $N = 8$ and SNR = 0 dB.

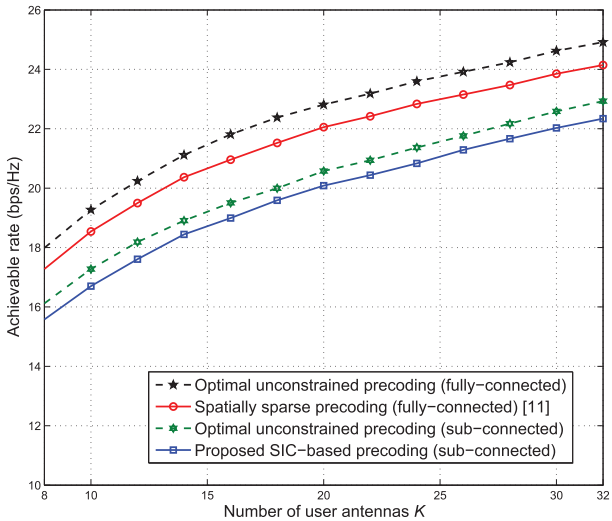

Fig. 6. Achievable rate comparison against the number of user antennas K, where $NM = 64$, $N = 8$, and SNR = 0 dB.

Fig. 4 compares the achievable rate in mmWave MIMO system with $NM \times K = 128 \times 32$ and $N = 16$, where we observe similar trends as those from Fig. 3. More importantly, Fig. 3 and Fig. 4 show that the performance of SIC-based hybrid precoding is also close to the spatially sparse precoding and the optimal unconstrained precoding with fully-connected architecture. For example, when SNR = 0 dB, our method can achieve more than 90% of the rate achieved by the near-optimal spatially sparse precoding in both simulated mmWave MIMO configurations. Considering the low computational complexity of the proposed SIC-based hybrid precoding as analyzed before, we further conclude that SIC-based hybrid precoding can achieve much better trade-off between the performance and computational complexity.

Fig. 5 provides a achievable rate comparison in mmWave MIMO systems against the numbers of BS and user antennas, where $NM = K$, the number of RF chains is fixed to $N = 8$, and SNR = 0 dB. We find that the performance of the proposed SIC-based hybrid precoding can be improved by increasing the number of BS and user antennas, which involves much lower energy consumption than increasing the number of energy-intensive RF chains [18].

Fig. 6 shows the achievable rate comparison against the numbers of user antennas K, where $NM = 64$, $N = 8$, and SNR = 0 dB. We imply from Fig. 6 that the performance loss of SIC-based hybrid precoding due to the sub-connected architecture can be compensated by increasing the number of user antennas K. For example, the achievable rate of SIC-based hybrid precoding when $K = 30$ is the same as that of the spatially sparse precoding when $K = 20$. Note that in this case, the required number of PSs of SIC-based hybrid precoding is $NM = 64$, while for the spatially sparse precoding, the number of required PSs is $N^2M = 512$. That means much energy can saved by SIC-based hybrid precoding, which will be also verified by simulation results later. In contrast, the cost of increasing the number of user antennas K will be negligible since the energy consumption of user antenna is usually small [18].

Next we evaluate the impact of imperfect CSI on the proposed SIC-based hybrid precoding. The estimated channel matrix (imperfect CSI) $\hat{\mathbf{H}}$ can be modeled as [4]

$$\hat{\mathbf{H}} = \xi \mathbf{H} + \sqrt{1 - \xi^2}\mathbf{E}, \tag{22}$$

where $\mathbf{H}$ is the actual channel matrix, $\xi \in [0, 1]$ presents the CSI accuracy, and $\mathbf{E}$ is the error matrix with entries following the distribution i.i.d. $\mathcal{CN}(0, 1)$. Fig. 7 shows the achievable rate comparison for an $NM \times K = 64 \times 16$ ($N = 8$) mmWave MIMO system, where the perfect CSI and the imperfect CSI with different ξ scenarios are considered. We observe that the proposed SIC-based hybrid precoding is not sensitive to the CSI accuracy. For example, the achievable rate of SIC-based hybrid precoding when $\xi = 0.9$ is quite close to that in the perfect CSI scenario, where the SNR gap is about 1 dB. Even when the CSI accuracy is quite poor (i.e., $\xi = 0.5$), SIC-based hybrid precoding with imperfect CSI can still achieve more than 88% of the rate in the perfect CSI scenario. Additionally, Fig. 8 shows

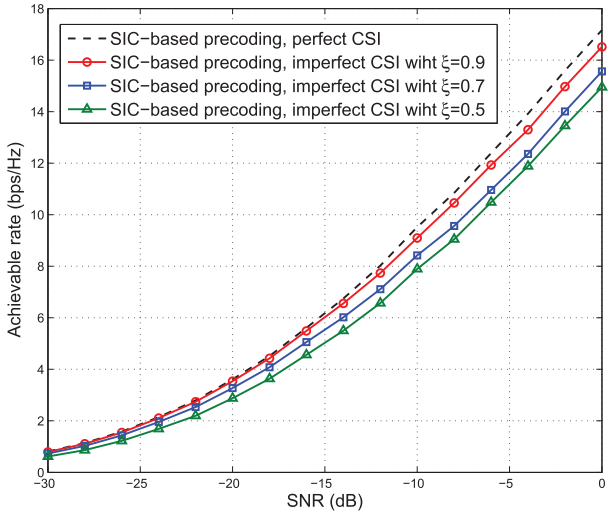

Fig. 7. Impact of imperfect CSI on SIC-based hybrid precoding for an $NM \times K = 64 \times 16$ ($N = 8$) mmWave MIMO system.

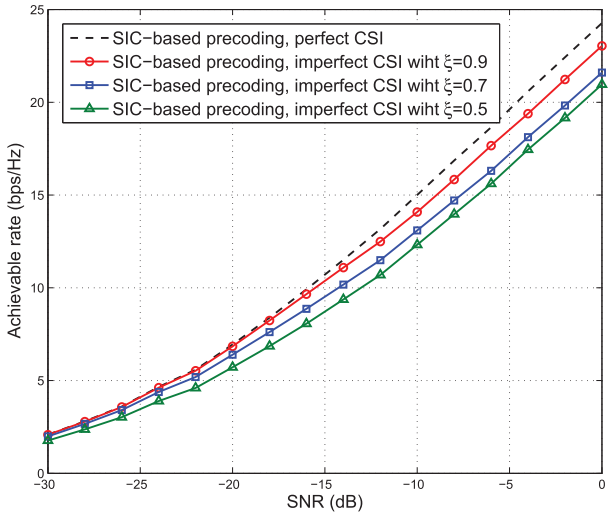

Fig. 8. Impact of imperfect CSI on SIC-based hybrid precoding for an $NM \times K = 128 \times 32$ ($N = 16$) mmWave MIMO system.

the achievable rate comparison for an $NM \times K = 128 \times 32$ ($N = 16$) mmWave MIMO system, where similar conclusions as those from Fig. 7 can be derived.

After that, we will also evaluate the proposed SIC-based hybrid precoding in practical systems with finite-alphabet signals instead of ideal Gaussian signals. Here, we also aim to maximize the achievable sum-rate, since it has been shown in [23, Section IV] that maximizing the achievable sum-rate is an excellent criterion for precoding, and it also has direct impact on the coded bit error rate performance. For the finite-alphabet signals $\tilde{\mathbf{s}}$, whose values are taken from a practical constellation Q, the achievable rate $\tilde{R}$ can be presented as [23]

$$\tilde{R} = N\log_2|Q| - \frac{1}{|Q|^N}\sum_{m=1}^{|Q|^N}\mathbb{E}_{\tilde{\mathbf{n}}}\left\{\log_2\sum_{k=1}^{|Q|^N} e^{-u_{m,k}}\right\}, \quad (23)$$

where N is the number of RF chains (also the number of transmitted data streams), $|Q|$ is the cardinality of Q, $\mathbb{E}_{\tilde{\mathbf{n}}}$ denotes the

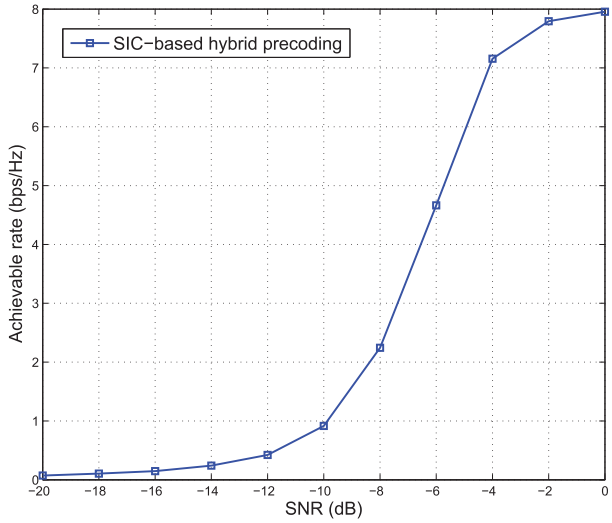

Fig. 9. Achievable rate of SIC-based hybrid precoding with finite-alphabet inputs, where $N = 8$, $NM = K = 64$, and BSPK is adopted.

expectation with respect to the noise vector $\tilde{\mathbf{n}}$, $u_{m,k}$ is defined as

$$u_{m,k} = \frac{\left\|\mathbf{HP}\left(\tilde{\mathbf{s}}_m - \tilde{\mathbf{s}}_k\right) + \tilde{\mathbf{n}}\right\|_2^2 - \left\|\tilde{\mathbf{n}}\right\|_2^2}{\tilde{\sigma}^2}, \quad (24)$$

$\mathbf{H}$ and $\mathbf{P}$ are the channel matrix and precoding matrix, respectively, $\tilde{\sigma}^2$ is the power of noise, and $\tilde{\mathbf{s}}_m$ is a possible signal vector with N elements taking values from Q.

From (23) we know that the achievable rate of practical signaling is quite different from ideal Gaussian signaling, where the upper bound is determined by $N\log_2|Q|$ [24], [32]. Fig. 9 shows the achievable rate of SIC-based hybrid precoding with finite-alphabet inputs, where $NM \times K = 64 \times 16$, $N = 8$, and the simple BPSK modulation ($|Q| = 2$) is considered as an example. We can observe that as the SNR becomes large, the proposed SIC-based hybrid precoding can also achieve the performance quite close to the upper bound $N\log_2|Q| = 8$bits.

In the end, we evaluate the energy efficiency of the proposed SIC-based hybrid precoding. Based on the energy consumption model in [33], [34], the energy efficiency η can be defined as

$$\eta = \frac{R}{P_{\text{total}}} = \frac{R}{P_{\text{t}} + N_{\text{RF}}P_{\text{RF}} + N_{\text{PS}}P_{\text{PS}}}\ (\text{bps/Hz/W}), \quad (25)$$

where $P_{\text{total}} \triangleq P_{\text{t}} + N_{\text{RF}}P_{\text{RF}} + N_{\text{PS}}P_{\text{PS}}$ is the total energy consumption, P_{t} is the transmitted energy, P_{RF} is the energy consumed by RF chain, P_{PS} is the energy consumed by PS (including the energy for the excitation and the energy for the compensation of insertion loss [18]), N_{RF} and N_{PS} are the numbers of required RF chains and PSs, respectively.

In this paper, we use the practical values $P_{\text{RF}} = 250$ mW [9], $P_{\text{PS}} = 1$ mW [18], and $P_{\text{t}} = 1$ W (about 30 dBm) in a small cell transmission scenario [35], since mmWave is more likely to be applied in small cells. Fig. 10 shows the energy efficiency comparison against the number of RF chains N, where SNR = 0 dB, $NM = K = 64$ ($N = 1, 2, 4, \ldots, 64$ to ensure that M is an integer). We observe that both the conventional spatially sparse precoding and the proposed SIC-based precoding can achieve higher energy efficiency than the optimal

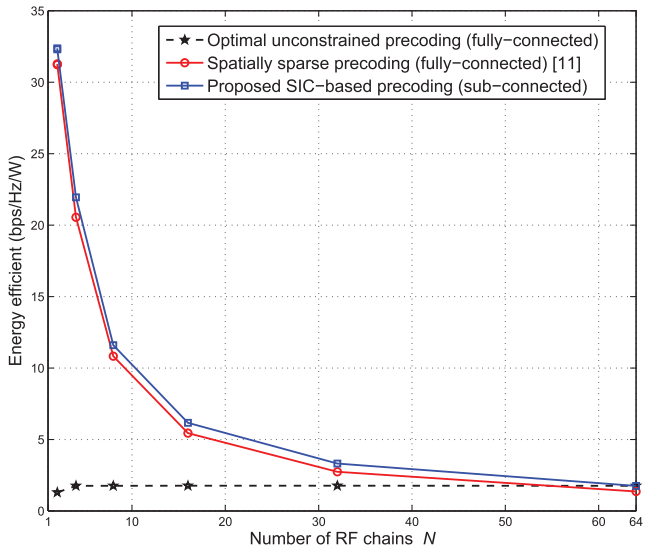

Fig. 10. Energy efficiency comparison against the numbers of RF chains N, where $NM = K = 64$.

unconstrained precoding (also known as the fully digital precoding), especially when the number of RF chains N is limited (e.g., $N < 30$). Besides, we also observe that the proposed SIC-based precoding is more energy efficient than the conventional spatially sparse precoding.

V. Conclusions

In this paper, we proposed a SIC-based hybrid precoding with sub-connected architecture for mmWave MIMO systems. We first showed that the total achievable rate optimization problem with non-convex constraints can be decomposed into a series of sub-rate optimization problems, each of which only considers one sub-antenna array. Then, we proved that the sub-rate optimization problem of each sub-antenna array can be solved by simply seeking a precoding vector sufficiently close to the unconstrained optimal solution. Finally, a low-complexity algorithm was proposed to realize SIC-based precoding without the complicated SVD and matrix inversion. Complexity evaluation showed that the complexity of the proposed SIC-based hybrid precoding is only about 10% of that of the recently proposed spatially sparse precoding with fully-connected architecture in typical mmWave MIMO system. Simulation results verified the near-optimal performance and high energy efficiency of the proposed SIC-based hybrid precoding. Our further work will focus on the limited feedback scenario, where the angles of PSs are quantified.

Appendix A
Proof of Proposition 1

Define the target of the optimization problem (10) as

$$R_n = \log_2\left(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H \bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\right), \tag{26}$$

and the SVD of $\bar{\mathbf{G}}_{n-1}$ as $\bar{\mathbf{G}}_{n-1} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^H$. Then, by separating the matrices $\mathbf{\Sigma}$ and $\mathbf{V}$ into two parts:

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{bmatrix}, \quad \mathbf{V} = [\mathbf{v}_1\ \mathbf{V}_2], \tag{27}$$

R_n in (26) can be rewritten as

$$\begin{aligned} R_n &= \log_2\left(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H \bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\right) \\ &= \log_2\left(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H \mathbf{V}\mathbf{\Sigma}\mathbf{V}^H\bar{\mathbf{p}}_n\right) \\ &= \log_2\Bigg(1 + \frac{\rho}{N\sigma^2} \\ &\qquad \times \bar{\mathbf{p}}_n^H [\mathbf{v}_1\ \mathbf{V}_2]\begin{bmatrix} \Sigma_1 & 0 \\ 0 & \mathbf{\Sigma}_2 \end{bmatrix}[\mathbf{v}_1\ \mathbf{V}_2]^H\bar{\mathbf{p}}_n\Bigg) \\ &= \log_2\Big(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H \mathbf{v}_1\Sigma_1\mathbf{v}_1^H\bar{\mathbf{p}}_n \\ &\qquad + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H \mathbf{V}_2\mathbf{\Sigma}_2\mathbf{V}_2^H\bar{\mathbf{p}}_n\Big). \end{aligned} \tag{28}$$

Since we aim to find a vector $\bar{\mathbf{p}}_n$ sufficiently "close" to $\mathbf{v}_1$, it is reasonable to assume that $\bar{\mathbf{p}}_n$ is approximately orthogonal to the matrix $\mathbf{V}_2$, i.e., $\bar{\mathbf{p}}_n^H\mathbf{V}_2 \approx \mathbf{0}$ [11]. Then, (28) can be simplified as

$$\begin{aligned} R_n &\approx \log_2\left(1 + \frac{\rho\Sigma_1}{N\sigma^2}\bar{\mathbf{p}}_n^H \mathbf{v}_1\mathbf{v}_1^H\bar{\mathbf{p}}_n\right) \\ &\overset{(a)}{=} \log_2\left(1 + \frac{\rho\Sigma_1}{N\sigma^2}\right) \\ &\qquad + \log_2\left(1 - \left(1 + \frac{\rho\Sigma_1}{N\sigma^2}\right)^{-1}\frac{\rho\Sigma_1}{N\sigma^2}\left(1 - \bar{\mathbf{p}}_n^H \mathbf{v}_1\mathbf{v}_1^H\bar{\mathbf{p}}_n\right)\right) \\ &\overset{(b)}{\approx} \log_2\left(1 + \frac{\rho\Sigma_1}{N\sigma^2}\right) + \log_2\left(\bar{\mathbf{p}}_n^H \mathbf{v}_1\mathbf{v}_1^H\bar{\mathbf{p}}_n\right) \end{aligned} \tag{29}$$

where (a) is obtained by using the formula $\mathbf{I} + \mathbf{X}\mathbf{Y} = (\mathbf{I} + \mathbf{X})\left(\mathbf{I} - (\mathbf{I} + \mathbf{X})^{-1}\mathbf{X}(\mathbf{I} - \mathbf{Y})\right)$ [11], where we define $\mathbf{X} = \frac{\rho\Sigma_1}{N\sigma^2}$ and $\mathbf{Y} = \bar{\mathbf{p}}_n^H \mathbf{v}_1\mathbf{v}_1^H\bar{\mathbf{p}}_n$; (b) is valid by employing the high SNR approximation [36], i.e.,

$$\left(1 + \frac{\rho\Sigma_1}{N\sigma^2}\right)^{-1}\frac{\rho\Sigma_1}{N\sigma^2} \approx 1. \tag{30}$$

From (29), we observe that maximizing R_n is equivalent to maximizing $\bar{\mathbf{p}}_n^H \mathbf{v}_1\mathbf{v}_1^H\bar{\mathbf{p}}_n = \left\|\bar{\mathbf{p}}_n^H \mathbf{v}_1\right\|_2^2$, the square of inner product between two vectors $\bar{\mathbf{p}}_n$ and $\mathbf{v}_1$. Note that $\mathbf{v}_1$ is a fixed vector. Therefore, exploring a vector $\bar{\mathbf{p}}_n$, which has the largest projection on $\mathbf{v}_1$, will lead to the smallest Euclidean distance to $\mathbf{v}_1$ as well. Based on this fact, we conclude that the optimization problem (10) is equivalent to the following problem

$$\bar{\mathbf{p}}_n^{\text{opt}} = \arg\min_{\bar{\mathbf{p}}_n \in \bar{\mathcal{F}}} \|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2. \tag{31}$$

Appendix B
Proof of Proposition 2

We first consider the matrix $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$, which should be inversed to compute $\bar{\mathbf{G}}_n$ (11). By partitioning $\mathbf{P}_n$ as $\mathbf{P}_n = \left[\mathbf{P}_{n-1}\ \mathbf{p}_n\right]$, $\mathbf{T}_n$ can be rewritten as

$$\begin{aligned} \mathbf{T}_n &= \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H \\ &= \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\left[\mathbf{P}_{n-1}\ \mathbf{p}_n\right]\left[\mathbf{P}_{n-1}\ \mathbf{p}_n\right]^H\mathbf{H}^H \\ &= \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_{n-1}\mathbf{P}_{n-1}^H\mathbf{H}^H + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{p}_n\mathbf{p}_n^H\mathbf{H}^H \\ &= \mathbf{T}_{n-1} + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{p}_n\mathbf{p}_n^H\mathbf{H}^H. \end{aligned} \tag{32}$$

Then, by utilizing the Sherman-Morrison formula [[28], Eq 2.1.4]

$$\left(\mathbf{A}+\mathbf{u}\mathbf{v}^T\right)^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}, \tag{33}$$

$\mathbf{T}_n^{-1}$ can be presented as

$$\begin{aligned}\mathbf{T}_n^{-1}&=\left(\mathbf{T}_{n-1}+\frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{p}_n\mathbf{p}_n^H\mathbf{H}^H\right)^{-1}\\&=\mathbf{T}_{n-1}^{-1}-\frac{\frac{\rho}{N\sigma^2}\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{p}_n\mathbf{p}_n^H\mathbf{H}^H\mathbf{T}_{n-1}^{-1}}{1+\frac{\rho}{N\sigma^2}\mathbf{p}_n^H\mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{p}_n}.\end{aligned} \tag{34}$$

Substituting (34) into $\mathbf{G}_n=\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}$, we have

$$\begin{aligned}\mathbf{G}_n&=\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}\\&=\mathbf{H}^H\left(\mathbf{T}_{n-1}^{-1}-\frac{\frac{\rho}{\sigma^2}\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{p}_n\mathbf{p}_n^H\mathbf{H}^H\mathbf{T}_{n-1}^{-1}}{1+\frac{\rho}{\sigma^2}\mathbf{p}_n^H\mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{p}_n}\right)\mathbf{H}\\&=\mathbf{G}_{n-1}-\frac{\frac{\rho}{\sigma^2}\mathbf{G}_{n-1}\mathbf{p}_n\mathbf{p}_n^H\mathbf{G}_{n-1}}{1+\frac{\rho}{\sigma^2}\mathbf{p}_n^H\mathbf{G}_{n-1}\mathbf{p}_n}.\end{aligned} \tag{35}$$

Then, according to (11), $\bar{\mathbf{G}}_n$ can be obtained by

$$\begin{aligned}\bar{\mathbf{G}}_n&=\mathbf{R}\mathbf{G}_n\mathbf{R}^H\\&=\mathbf{R}\left(\mathbf{G}_{n-1}-\frac{\frac{\rho}{N\sigma^2}\mathbf{G}_{n-1}\mathbf{p}_n\mathbf{p}_n^H\mathbf{G}_{n-1}}{1+\frac{\rho}{N\sigma^2}\mathbf{p}_n^H\mathbf{G}_{n-1}\mathbf{p}_n}\right)\mathbf{R}^H\\&=\bar{\mathbf{G}}_{n-1}-\frac{\frac{\rho}{N\sigma^2}\bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\bar{\mathbf{p}}_n^H\bar{\mathbf{G}}_{n-1}}{1+\frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H\bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n}.\end{aligned} \tag{36}$$

Note that in Section III-B, we have obtained the precoding vector $\bar{\mathbf{p}}_n$ sufficiently close to $\mathbf{v}_1$, i.e., $\bar{\mathbf{p}}_n\approx\mathbf{v}_1$. Thus, (36) can be well approximated by replacing $\bar{\mathbf{p}}_n$ with $\mathbf{v}_1$ as

$$\begin{aligned}\bar{\mathbf{G}}_n&=\bar{\mathbf{G}}_{n-1}-\frac{\frac{\rho}{N\sigma^2}\bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\bar{\mathbf{p}}_n^H\bar{\mathbf{G}}_{n-1}}{1+\frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H\bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n}\\&\approx\bar{\mathbf{G}}_{n-1}-\frac{\frac{\rho}{N\sigma^2}\bar{\mathbf{G}}_{n-1}\mathbf{v}_1\mathbf{v}_1^H\bar{\mathbf{G}}_{n-1}}{1+\frac{\rho}{N\sigma^2}\mathbf{v}_1^H\bar{\mathbf{G}}_{n-1}\mathbf{v}_1}\\&\overset{(a)}{=}\bar{\mathbf{G}}_{n-1}-\frac{\frac{\rho}{N\sigma^2}\Sigma_1^2\mathbf{v}_1\mathbf{v}_1^H}{1+\frac{\rho}{N\sigma^2}\Sigma_1},\end{aligned} \tag{37}$$

where (a) is true due to fact that $\mathbf{v}_1^H\bar{\mathbf{G}}_{n-1}=\Sigma_1\mathbf{v}_1^H$, since $\bar{\mathbf{G}}_{n-1}$ is an Hermitian matrix.

REFERENCES

[1] T. Bai, A. Alkhateeb, and R. Heath, "Coverage and capacity of millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 70–77, Sep. 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] 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.

[4] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[5] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 136–143, Dec. 2014.

[6] M. Samimi *et al.*, "28 GHz angle of arrival and angle of departure analysis for outdoor cellular communications using steerable beam antennas in New York City," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring'13)*, May 2013, pp. 1–6.

[7] A. Alkhateeb, J. Mo, N. González-Prelcic, and R. Heath, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[8] B. Yin *et al.*, "High-throughput beamforming receiver for millimeter wave mobile communication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM'13)*, Dec. 2013, pp. 3697–3702.

[9] P. Amadori and C. Masouros, "Low RF-complexity millimeter-wave beamspace-MIMO systems by beam selection," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2212–2222, Jun. 2015.

[10] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid precoding analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[11] 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.

[12] Y. Lee, C.-H. Wang, and Y.-H. Huang, "A hybrid RF/baseband precoding processor based on parallel-index-selection matrix-inversion-bypass simultaneous orthogonal matching pursuit for millimeter wave MIMO systems," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 305–317, Jan. 2015.

[13] C.-E. Chen, "An iterative hybrid transceiver design algorithm for millimeter wave MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 285–288, Jun. 2015.

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

[15] W. Roh *et al.*, "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.

[16] 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 Commun. Conf. (GLOBECOM'13)*, Dec. 2013, pp. 3685–3690.

[17] C. Kim, J. S. Son, T. Kim, and J.-Y. Seol, "On the hybrid beamforming with shared array antenna for mmWave MIMO-OFDM systems," in *Proc. IEEE WCNC'14)*, Apr. 2014.

[18] C. A. Balanis, *Antenna Theory: Analysis and Design*. Hoboken, NJ, USA: Wiley, 2012.

[19] D. Schneider, "Could supercomputing turn to signal processors (again)?" *IEEE Spectr.*, vol. 49, no. 10, pp. 13–14, Oct. 2012.

[20] 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.

[21] Y.-C. Liang, E. Y. Cheu, L. Bai, and G. Pan, "On the relationship between MMSE-SIC and BI-GDFE receivers for large multiple-input multiple-output channels," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3627–3637, Aug. 2008.

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

[23] C. Xiao, Y. R. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3301–3314, Apr. 2011.

[24] Y. Wu, C. Xiao, X. Gao, J. D. Matyjas, and Z. Ding, "Linear precoder design for MIMO interference channels with finite-alphabet signaling," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3766–3780, Sep. 2013.

[25] F. Zhu, F. Gao, M. Yao, and H. Zou, "Joint information- and jamming-beamforming for physical layer security with full duplex base station," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6391–6401, Dec. 2014.

[26] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Proc. IEEE Inf. Theory Appl. Workshop (ITA'13)*, 2013, pp. 1–5.

[27] M. Tao and R. Wang, "Linear precoding for multi-pair two-way MIMO relay systems with max-min fairness," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5361–5370, Oct. 2012.

[28] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, USA: JHU Press, 2012.

[29] Å. Björck, "Numerical methods in matrix computations," Berlin, German: Springer, 2015.

[30] 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 Commun. Conf. (GLOBECOM'13)*, Dec. 2013, pp. 3476–3480.
[31] A. Alkhateeb, O. El Ayach, G. Leus, and R. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[32] Y. Wu, C. Xiao, Z. Ding, X. Gao, and S. Jin, "Linear precoding for finite-alphabet signaling over MIMOME wiretap channels," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2599–2612, Jul. 2012.
[33] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.
[34] C. Masouros, M. Sellathurai, and T. Rantarajah, "Computationally efficient vector perturbation using thresholded optimization," *IEEE Trans. Commun.*, vol. 61, no. 5, pp. 1880–1890, May 2013.
[35] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, "State of the art in 60-GHz integrated circuits and systems for wireless communications," *Proc. IEEE*, vol. 99, no. 8, pp. 1390–1436, Aug. 2011.
[36] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

Xinyu Gao (S'14) received the B.E. degree in communication engineering from Harbin Institute of Technology, Heilongjiang, China, in 2014. He is currently pursuing the Ph.D. degree in electronic engineering at Tsinghua University, Beijing, China. His research interests include massive MIMO and mmWave communications, with the emphasis on signal detection and precoding. He has authored several journal and conference papers published in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE ICC, IEEE GLOBECOM, etc. He was the recipient of the National Scholarship in 2015.

Linglong Dai (M'11–SM'14) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree (with the highest Hons.) from the China Academy of Telecommunications Technology (CATT), Beijing, China, in 2006, and the Ph.D. degree (with the highest Hons.) from Tsinghua University, Beijing, China, in 2011. From 2011 to 2013, he was a Postdoctoral Fellow with the Department of Electronic Engineering, Tsinghua University, where he has been an Assistant Professor since July 2013 and now an Associate Professor. His research interests include wireless communications, with a focus on multi-carrier techniques, multiantenna techniques, and multiuser techniques. He has authored more than 50 IEEE journal papers and more than 30 IEEE conference papers. He currently serves as a Co-Chair of the IEEE Special Interest Group (SIG) on Signal Processing Techniques in 5G Communication Systems. He was the recipient of the Outstanding Ph.D. Graduate of Tsinghua University Award in 2011, the Excellent Doctoral Dissertation of Beijing Award in 2012, the IEEE ICC Best Paper Award in 2013, the National Excellent Doctoral Dissertation Nomination Award in 2013, the IEEE ICC Best Paper Award in 2014, the URSI Young Scientists Award in 2014, the IEEE Transactions on Broadcasting Best Paper Award in 2015, the IEEE RADIO Young Scientists Award in 2015.

Shuangfeng Han (M'06) received the M.S. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2002 and 2006, respectively. He joined Samsung Electronics as a Senior Engineer, in 2006, working on MIMO, multi-BS MIMO, etc. Since 2012, he has been a Senior Project Manager in the Green Communication Research Center, China Mobile Research Institute. His research interests include green 5G, massive MIMO, full duplex, NOMA and EE-SE co-design. Currently, he is the Vice Chair of Wireless Technology Work Group of China's IMT-2020 (5G) promotion group.

Chih-Lin I (SM'03) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA. She has been working at multiple world-class companies and research institutes leading the R&D, including AT&T Bell Labs; AT&T HQ, ITRI of Taiwan, and ASTRI of Hong Kong. Currently, she is a China Mobile's Chief Scientist of wireless technologies and has established the Green Communications Research Center, spearheading major initiatives including key 5G technology R&D; high energy efficiency system architectures, technologies and devices; green energy; and C-RAN and soft base stations. Her research interests include green communications, C-RAN, network convergence, bandwidth refarming, EE-SE co-design, massive MIMO, and active antenna arrays. She was an elected Board Member of the IEEE ComSoc, the Chair of the ComSoc Meetings and Conferences Board, and the Founding Chair of the IEEE WCNC Steering Committee. She is currently an Executive Board Member of GreenTouch and a Network Operator Council Member of ETSI NFV. She was the recipient of the IEEE TRANSACTIONS ON COMMUNICATIONS Stephen Rice Best Paper Award and is a winner of the CCCP National 1000 Talent program.

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

Dr. Heath has been an Editor of the IEEE TRANSACTIONS ON COMMUNICATION, an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, a Lead Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS special issue on Limited Feedback Communication, and a Lead Guest Editor of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING special issue on Heterogenous Networks. He currently serves on the steering committee for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was a member of the Signal Processing for Communications Technical Committee in the IEEE Signal Processing Society and is a former Chair of the IEEE COMSOC Communications Technical Theory Committee. He was a technical Co-Chair for the 2007 Fall Vehicular Technology Conference, a General Chair of the 2008 Communication Theory Workshop, a General Co-Chair, a Technical Co-Chair and a Co-Organizer of the 2009 IEEE Signal Processing for Wireless Communications Workshop, a Local Co-Organizer for the 2009 IEEE CAMSAP Conference, a Technical Co-Chair for the 2010 IEEE International Symposium on Information Theory, the Technical Chair for the 2011 Asilomar Conference on Signals, Systems, and Computers, the General Chair for the 2013 Asilomar Conference on Signals, Systems, and Computers, a founding General Co-Chair for the 2013 IEEE GlobalSIP conference, and is a Technical Co-Chair for the 2014 IEEE GLOBECOM conference. He was a 2003 Frontiers in Education New Faculty Fellow. He is also a licensed amateur radio operator and is a registered Professional Engineer in Texas. He was the recipient of the of Best Student Paper Awards at the IEEE VTC 2006 Spring, WPMC 2006, IEEE GLOBECOM 2006, IEEE VTC 2007 Spring, and IEEE RWS 2009, as well as the corecipient of the Grand Prize in the 2008 WinTech WinCool Demo Contest. He was also the corecipient of the 2010 and 2013 *EURASIP Journal on Wireless Communications and Networking* Best Paper Awards and the 2012 *Signal Processing Magazine* Best Paper Award.