Gina Cody School of Engineering and Computer Science

Department of Electrical and Computer Engineering

Hybrid Precoding in Millimeter-Wave Massive MIMO Systems

Pranav Jha

Student ID: 40081750

Ph.D. Comprehensive Examination Report

Supervisor

Dr. Wei-Ping Zhu

February 2020

Abstract

Millimeter-wave (mm-Wave) multiple-input multiple-output (MIMO) has been considered as a promising technique in improving the overall throughput of the future 5G wireless networks due to the large available bandwidth in the mm-Wave spectrum and higher spectrum efficiency of massive MIMO systems. However, mm-Wave signals suffer from free-space path loss and require sufficient antenna array gain or beamforming gain to minimize this effect. The array gain can be achieved by precoding (beamforming) in massive MIMO systems with large antenna array which provides higher multi-user gain and establishes links with a reasonable signal-to-noise ratio (SNR). In addition, mm-Wave signals enable such large antenna array in massive MIMO to be packed in the small physical dimension. For MIMO, conventionally, precoding is done in baseband, however, due to the baseband hardware complexity and power consumption related to the mixed analog-to-digital components (ADCs), digital precoding is prohibitive in mm-Wave massive MIMO systems. This makes hybrid precoding where the precoder processing is divided between analog and digital domains, a promising solution for reducing the hardware complexity and energy consumption. In this report, we investigate the hybrid precoding designs for mm-Wave massive MIMO systems proposed in two recent research papers. In the first paper [1], authors have proposed a low-complex solution to the hybrid precoding problem where the total achievable rate optimization problem has been divided into a series of sub-rate optimization problems. In the second paper, [2], a deep learning (DL) enabled mm-Wave massive MIMO framework for hybrid precoding is proposed. Specifically, deep neural network (DNN) is adopted to extract the features of the communication model, in which different layers of the network can process specific functions. This report reviews these two papers in detail and briefly discusses their results and limitations along with the possible future research directions.

Contents

1 Introduction

1.1 Dawn of Millimeter-Wave Massive MIMO

Millimeter-Wave massive MIMO is a promising candidate technology for exploring new frontiers for future 5G networks. It benefits from the combination of large available bandwidth in mm-Wave frequency bands and high antenna gains achievable with massive MIMO antenna arrays. With enhanced energy and spectral efficiencies, increased reliability, compactness, flexibility, and improved overall system capacity, mm-Wave massive MIMO is expected to address the challenges of the explosively growing mobile data demand. For mm-Wave massive MIMO systems, maximum benefits can be achieved when different transceiver antenna pairs experience independently fading channel coefficients. This is realizable when the antenna element's spacing is at least 0.5λ, where λ is the wavelength of the signal. Since λ reduces with the increasing carrier frequency, a higher number of elements in antenna arrays of the same physical dimension can be realized at mm-Wave than at μ-Wave frequencies. At mm-Wave frequencies, the dimensions of antenna elements, as well as the inter-antenna spacing, become incredibly small due to their dependence on wavelength. This makes it possible to pack a large number of antenna elements in a physically limited space, thereby enabling massive MIMO antenna array, not only at the base stations (BSs) but also at the user equipments (UEs) [3]. However, the maximum numbers of antennas under consideration by the 3^{rd} Generation Partnership Project (3GPP) at 70 GHz are 1024 and 64 for the BSs and UEs, respectively. As for the RF chains, the maximum numbers are 32 and 8 for the BSs and UEs, respectively [4].

1.1.1 Need of Hybrid Precoding

Conventional MIMO arrays at low-frequencies tend to use less than ten antennas [5] and have all the signal processing performed in the baseband, which requires that each antenna is connected to its own radio frequency (RF) chain devices (e.g. power amplifiers (PA), low noise amplifiers (LNA), analog-to-digital converters (ADC)). For massive MIMO, however, the cost and power consumption of the RF chain, especially in high frequencies, and the space occupied by all these devices, despite the small size of the antennas, will likely prevent the systems from using a

complete RF chain per antenna [6]. To overcome these issues, a promising solution is the hybrid precoding which uses much less RF chains than the number of antennas in the array. In the hybrid precoding, the low-dimensional transmitted signal is digitally precoded in the baseband domain, up-converted, and processed in the RF domain in order to produce the high-dimensional transmitted waveforms.The hybrid precoding combines the high array gain, provided by the use of all antenna elements, with the multiplexing gain, provided by the use of multiple RF chains. The additional digital/baseband processing allows to reduce interference, to support multi-stream and multi-user transmissions. Furthermore, with proper design, hybrid precoding can provide near-optimal performance compared to the fully-digital precoding, with a much lower number of RF chains, and consequently, lower complexity and power consumption.

Despite these advantages, the hybrid precoding design has some significant challenges: the PS network impose a constant-modulus and phase quantization constraints, leading to non-convex, combinatorial, NP-hard design problems; and the coupling between analog and digital beamformers adds non-linearity to the problem [7]. To address these problems, a spatially sparse hybrid precoding scheme which is the first low-complexity sub-optimal design in a mm-Wave massive MIMO system was proposed in [8]. Here, the authors have assumed that the mm-Wave channel is sparse in the angular domain and with this assumption, the hybrid precoding/combining design problem was formulated as a sparse approximation problem, and a variant of the matching pursuit algorithm was developed to efficiently design the analog/digital precoding and combining matrices. This design illustrated that hybrid precoding can effectively achieve performance gains comparable to digital baseband solutions while requiring much less hardware complexity, which makes it a promising precoding solution for mm-Wave systems.

Based on the spatially sparse hybrid precoding, several hybrid precoding schemes have been proposed. In [9], a low-complexity version of the spatially sparse hybrid precoding is proposed. The main contributions of this work include the derivation and integration of a matrix-inversion-bypass OMP algorithm to eliminate the matrix inversion operations and development of a specific precoding reconstruction algorithm for the hardware implementation by considering the mm-Wave channel properties. In [10], a modified spatially sparse hybrid precoding is proposed for the mm-Wave massive MIMO system with partial channel knowledge, where the BS and the

user only know their own local angles of arrival (AoAs). In [11], the spatially sparse hybrid precoding is combined with channel estimation, and a multi-resolution codebook is designed to estimate the AoA/AoD. The precoding performances of these schemes are limited and can be improved further by incorporating deep learning (DL) technique into the mm-Wave massive MIMO systems. We will discuss this in detail in the following sections.

1.1.2 Deep Learning for Wireless Physical Layer

Recently, machine learning (ML), especially DL, has been successfully applied for improving the performance of baseband signal processing algorithms [12], including channel decoding, channel estimation and detection, and so on. DL has gained much interest recently for the solution of many challenging problems such as speech recognition, visual object recognition, and language processing [13–15]. The DL is an extraordinarily remarkable technology for handling explosive data and addressing complicated nonlinear problems. It has been proved that DL is an excellent tool to deal with complex non-convex problems and high computational issues. Deep learning has several advantages such as low computational complexity when solving optimization-based or combinatorial search problems and the ability to extrapolate new features from a limited set of features contained in a training dataset [14]. By using the training dataset (pre-obtained or online) to learn the actual signal model and refine the parameters of algorithms via DL, mm-Wave systems may be able to overcome hardware constraints and imperfections in terms of signal processing. Moreover, DL can optimize the non-convex sum-rate optimization problem of hybrid precoding in mm-Wave massive MIMO systems following its remarkable performance when applied to massive MIMO based on the further decomposition of the baseband precoding matrix.

Basic idea of DL: Here, we provide an intuitive insight into the DL by illustrating the basic idea behind this simple yet powerful technique.

The simplest deep learning (DL) model is a linear model, which is expressed as

$$f(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{n} v_i w_i, \tag{1.1}$$

where $\mathbf{w}$ represents the weight of the network, and this model is designed to take n input values

as $\{v_1, v_2, ..., v_n\}$. Also, $f(\cdot)$ represents the output of the model, which can classify two different branches by identifying whether $f(\mathbf{v}, \mathbf{w})$ is positive or negative. Then, motivated by the basic idea of more computational units that can facilitate intelligent interaction manners, many works develop new models that consist of multiple units and are capable of dealing with linear and non-linear problems. In the deep learning area, deep neural network (DNN) is considered as one of the most popular generative models. As a multilayer processor, the DNN is capable of dealing with many non-convex and non-linear issues. Also, the multilayer perceptron mechanism and special training policy promote the DNN to be a commendable tool to leverage the sparsity characteristics of the mm-Wave massive MIMO. DNN is designed as a neural network with many hidden layers. In particular, there are multiple neurons implementing in each hidden layer, as well as an output with the weighted sum of these neurons operated by a nonlinear function. In order to realize recognition and representation operation, the DNN is processed by activation function. In general, the Sigmoid $\sigma(\cdot)$ function and the rectified linear unit ReLU$(\cdot)$ function are the most universal choices in the nonlinear operation, given as

$$\sigma(a) = \frac{1}{1 + e^{-\alpha}}, \ \ \text{ReLU}(a) = \max(0, a) \tag{1.2}$$

where a is denoted as the argument of the function. Lately, large breeds of DL architectures are based on rectified linear unit ReLU. The mapping between the input $\mathbf{v}$ and the output $\mathbf{o}$ of the DNN of a massive MIMO system is obtained as

$$\mathbf{o} = f(\mathbf{v}, \mathbf{w}) = f^{(n-1)}(f^{(n-2)}(\cdots f^{1}(\mathbf{v}))), \tag{1.7}$$

where n represents the number of layers in the neural network. These layers can process specific activation functions and realize corresponding mapping relationship.

1.2 Millimeter-Wave Massive MIMO System Model

1.2.1 Antenna Array

There are three types of antenna array architecture that have evolved over time as fully-digital, fully-analog, and hybrid analog-digital architecture. A fully-digital implementation employs dedicated RF front-end and digital baseband per antenna, which for the mm-Wave massive MIMO is prohibitively costly and practically infeasible due to tight space constraints. The fully-analog array, on the other hand, uses only one RF chain with multiple analog phase shifters (PSs). It has a simple hardware structure but suffers from poor system performance. Also, it has low antenna gain, as only the phases of the signals, but not their amplitudes can be controlled.

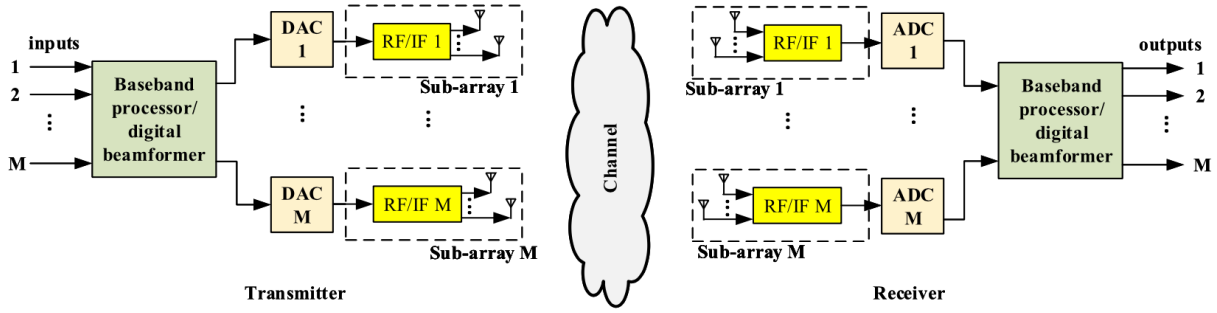

Figure 1.1: Architecture of hybrid antenna array system.

The more feasible and practical approach, according to research trends, is the massive hybrid array which consists of multiple analog sub-arrays with their own respective digital chains [16]. In the massive hybrid array architecture, antenna elements are grouped into analog sub-antenna arrays. In this architecture, only one PS is dedicated to a single antenna element and all other components are shared by all antenna elements in each sub-array. Each sub-array is fed with only one digital input and outputs only one digital signal, and all digital signals from all the sub-arrays are jointly processed in a digital processor. Overall, this hybrid structure, shown in Fig. 1.1 [3], significantly reduces the cost, number of required hardware components and system complexity, and the performance is roughly comparable with the optimal but costly and unfeasible fully-digital architecture [1].

In terms of structure, antenna arrays are typically designed as either uniform linear array (ULA) or uniform planar array (UPA). The array response vector for ULA with U elements is

given as

$$\mathbf{f}_{\mathrm{ULA}}(\phi) = \frac{1}{\sqrt{U}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin(\phi)}, \ldots, e^{j(U-1)\frac{2\pi}{\lambda}d\sin(\phi)}\right]^T. \tag{1.8}$$

where λ is the wavelength of the signal and d is the antenna spacing. In addition, the array response vector for UPA with W_1 and W_2 elements on horizontal and vertical, respectively is given as

$$\mathbf{f}_{\mathrm{UPA}}(\phi,\theta) = \frac{1}{\sqrt{U}}\left[1, \ldots, e^{j\frac{2\pi}{\lambda}d(x\sin(\phi)\sin(\theta)+y\cos(\theta))}, \ldots, e^{j\frac{2\pi}{\lambda}d((W_1-1)\sin(\phi)\sin(\theta)+(W_2-1)\cos(\theta))}\right]^T, \tag{1.9}$$

where $0 \leq x \leq (W_1 - 1)$ and $0 \leq y \leq (W_2 - 1)$. Considering UPAs are of interest in mm-Wave beamforming because they (1) yield smaller antenna array dimensions; (2) facilitate packing more antenna elements in a reasonably sized array; and (3) enable beamforming in the elevation domain (also known as 3D beamforming) [16].

1.2.2 Channel Model

The high free-space path loss is a characteristic of mm-Wave propagation, leading to limited spatial selectivity or scattering. On the other hand, the large tightly packed antenna arrays are characteristics of mm-Wave transceivers, leading to high levels of antenna correlation. This feature of tightly packed arrays in sparse scattering environments makes many of the statistical fading distributions used in traditional MIMO analysis inaccurate for mm-Wave channel modeling [16]. For this reason, a narrowband channel representation, based on the Saleh-Valenzuela model has been considered, which allows to accurately capture characteristics in mm-Wave channels. Using this channel model, the channel matrix $\mathbf{H}$ is assumed to be a sum of the contributions of L propagation paths. Therefore, the discrete-time narrowband channel $\mathbf{H}$ can be written as

$$\mathbf{H} = \gamma \sum_{l=1}^{L} \alpha_l \Lambda_r(\phi_l^r, \theta_l^r)\Lambda_t(\phi_l^t, \theta_l^t)\mathbf{f}_r(\phi_l^r, \theta_l^r)\mathbf{f}_t^H(\phi_l^t, \theta_l^t), \tag{1.10}$$

where $\gamma = \sqrt{\frac{NMK}{L}}$ represents the normalization factor, L represents the limited number of scatters where $L \leq N$ for mm-Wave communication systems, $\alpha_l \in \mathbb{C}$ represents the gain of l-th path, $\phi_l^t(\theta_l^t)$ and $\phi_l^r(\theta_l^r)$ represent the azimuth (elevation) angles of departure and arrival

(AoDs/AoAs), respectively, $\Lambda_t(\phi_l^t, \theta_l^t)$ and $\Lambda_r(\phi_l^r, \theta_l^r)$ represent the transmit and receive antenna array gain at a specific AoD and AoA, respectively, and $\mathbf{f}_r(\phi_l^r, \theta_l^r)$ and $\mathbf{f}_t^H(\phi_l^t, \theta_l^t)$ represent the antenna array response vectors depending on the antenna array structures at the BS and the the user, respectively.

Notations: Lower-case and upper-case boldface letters represent vectors and matrices, respectively; $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$ and $|\cdot|$ represent transpose, conjugate transpose, inversion and determinant of a matrix, respectively; $\|\cdot\|_1, \|\cdot\|_2$ and $\|\cdot\|_F$ represent l_1-norm, l_2-norm of a vector and Frobenius norm of a matrix, respectively; $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ represent Real and Imaginary part of a complex number, respectively; $\mathbb{E}(\cdot)$ and $\mathbf{I}_N$ represent Expectation and $N \times N$ identity matrix, respectively.

2 Problem Statement

The non-trivial task of designing hybrid precoders in mm-Wave massive MIMO systems results from the optimization of a performance metric, as the spectral efficiency or the total achievable rate. As mentioned earlier, the hybrid precoders involve the coupling between their analog and baseband counterparts, making the problem highly nonlinear. Moreover, the analog precoder and combiner rely on a phase-shifters network, which imposes a constant modulus constraint in their components, also turning the problem non-convex. To solve this design problem of optimal hybrid precoder, researchers have provided different methods in recent years and achieved near-optimal performance. However, these works are based on conventional mathematical means such as the singular value decomposition (SVD) and the Geometric mean decomposition (GMD), have high computational complexity and are too weak to exploit the sparsity statistics of the mm-Wave massive MIMO channel. Consequently, traditional low-complexity schemes are realized at the cost of degrading the hybrid precoding performance of the systems. Hence, to further improve the performance and decrease the complexity of hybrid precoding in mm-Wave massive MIMO systems, it is necessary to develop new techniques to design the best hybrid precoders. In light of these considerations, we aim to

- Develop a low-complex energy-efficient solution for the hybrid precoder design problem taking into account all the practical design constraints.
- Further consider the application of deep learning (DL) to develop optimal hybrid precoder with much reduced computational complexity.

For this purpose, in this report, we investigate two papers from the hybrid precoding design literature. In the first paper, the authors have proposed an efficient hybrid precoding design by using a sub-connected antenna array architecture and optimize the achievable sub-rate of each sub-antenna array, successively. On the other hand, in the second paper, the authors have considered a fully-connected antenna array architecture and proposed a deep learning-enabled massive MIMO framework for effective hybrid precoding.

3 Hybrid Precoding in Millimeter-Wave Massive MIMO System

In this section, we provide an overview of the paper [1]: X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath,"Energy-efficient hybrid analog and digital precoding for mm-wave mimo systems with large antenna arrays", *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 998–1009, 2016. This paper is chosen due to its relevant contribution in demonstrating 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 and 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. In the following subsections, we describe in more details the research problem, solution methodology and results of this paper.

3.1 Problem Formulation

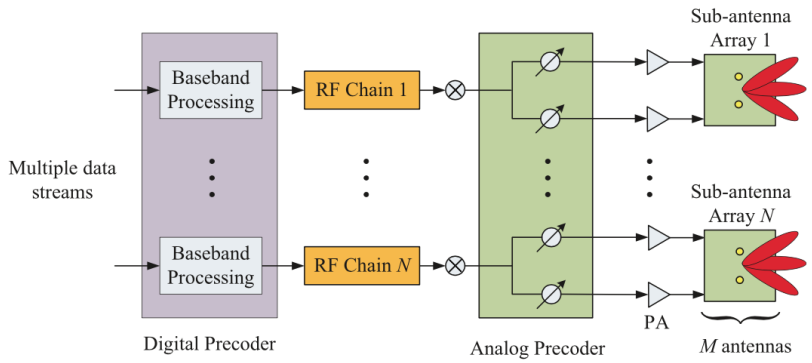

Figure 3.1: Sub-connected architecture for hybrid precoding for an $NM \times K = 64 \times 16$ $(N = 8)$ mm-Wave MIMO system.

In the work [1], the authors have considered a sub-connected architecture for hybrid precoding in mm-Wave massive MIMO systems as shown in Fig. 3.1 with NM base station (BS) antennas and N RF chains. Each RF chain is associated to M phase shifters (PSs) and each of these M PSs are further connected to a sub-antenna array with only M antennas which is a subset of the total BS antennas. The BS transmits N independent data streams to users with K receive antennas and these N data streams in the baseband are precoded by a digital precoder **D** where

$\mathbf{D} = \text{diag}[d_1, d_2, \ldots, d_N]$, is a diagonal matrix and $d_n \in \mathbb{R}$ for $n = 1, 2, \ldots, N$. The digitally precoded signal passes through the corresponding RF chain and reaches to M PSs associated with each RF chain to perform analog precoding. This can be represented by the analog weighting vector $\bar{\mathbf{a}}_n = \mathbb{C}^{M\times 1}$, elements of which is of same magnitude $1/\sqrt{M}$ but different phases. After the analog precoding, each data stream is finally transmitted by the corresponding sub-antenna array associated to each RF chain and the received signal vector $\mathbf{y} = [y_1, y_2, \ldots, y_k]^T$ at the user in a narrowband system is given as

$$\mathbf{y} = \sqrt{\rho}\mathbf{HADs} + \mathbf{n} = \sqrt{\rho}\mathbf{HPs} + \mathbf{n}, \tag{3.1}$$

where ρ represents the average received power; $\mathbf{H} \in \mathbb{C}^{K\times NM}$ represents the channel matrix based on the well-known Saleh-Valenzuela (SV) channel model, $\mathbf{A}$ represents the $NM \times N$ analog precoding matrix which comprises N analog weighting vectors $\{\bar{\mathbf{a}}_m\}_{m=1}^{N}$ given as

$$\begin{bmatrix} \bar{\mathbf{a}}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{a}}_2 & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \bar{\mathbf{a}}_N \end{bmatrix}_{NM\times N}, \tag{3.2}$$

$\mathbf{s} = [s_1, s_2, \ldots, s_N]^T$ is the transmitted signal vector in the baseband. In this work, the widely used Gaussian signals with normalized signal power $\mathbb{E}(\mathbf{ss}^H) = \frac{1}{N}\mathbf{I}_N$ is used. The hybrid precoding matrix of size $NM\times N$ is given by $\mathbf{P} = \mathbf{AD} = \text{diag}\{\bar{\mathbf{a}}_1, ..., \bar{\mathbf{a}}_N\}\cdot\text{diag}\{d_1, ..., d_N\}$ and satisfies $\|\mathbf{P}\|_F \leq N$ to achieve the total transmit power constraint. Lastly, $\mathbf{n} = [n_1, n_2, \ldots, n_N]^T$ is an independent and identically distributed (i.i.d.) additive white Gaussian noise (AWGN) $\mathcal{CN}(0, \sigma^2)$ vector.

In this work, the aim is to maximize the total achievable rate R of mm-Wave MIMO systems where R is given as

$$R = \log_2\left(\left|\mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{HPP}^H\mathbf{H}^H\right|\right). \tag{3.3}$$

As seen in (3.1), $\mathbf{P}$ can be represented as $\mathbf{P} = \mathbf{AD}$ where both $\mathbf{A}$ and $\mathbf{D}$ are the diagonal matrices, the design of $\mathbf{P}$ is depends on three constraints which are given as:

1. *Constraints 1:* $\mathbf{P}$ should be a block diagonal matrix similar to A as shown in (3.2). This represents $\mathbf{P} = \text{diag}\{\bar{\mathbf{p}}_1, \ldots, \bar{\mathbf{p}}_N\}$, where $\bar{\mathbf{p}_n} = d_n\bar{\mathbf{a}_n}$ is the $M \times 1$ non-zero vector of the

n-th column $\mathbf{p}_n$ of P, i.e., $p_n = [\mathbf{0}_{1\times M(n-1)},\ \bar{\mathbf{p}_n},\ \mathbf{0}_{1\times M(N-n)}]^T$.

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

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

These constraints on **P** are non-convex and so, the solution for the minimization of the total achievable rate in (3.3) is intractable. The diagonal structure of hybrid precoding matrix **P** suggests that the precoding of different sub-antenna arrays are independent, so the total achievable rate (3.3) can be decomposed into a series of sub-rate optimization problems for each sub-antenna array individually.

The authors have divided the hybrid precoding matrix **P** as $\mathbf{P} = [\mathbf{P}_{N-1}\mathbf{p}_N]$, with $\mathbf{p}_N$ be the N-th column of **P** and $\mathbf{P}_{N-1}$ be a $NM \times (N-1)$ matrix which contains first $N-1$ columns of **P**. Then, the total achievable rate R in (3.3) is rewritten as

$$R = \log_2(|\mathbf{T}_{N-1}|) + \log_2\Big(\Big|1 + \frac{\rho}{N\sigma^2}\mathbf{p}_N^H\mathbf{H}^H\mathbf{T}_{N-1}^{-1}\mathbf{H}\mathbf{p}_N\Big|\Big), \tag{3.4}$$

The second term $\log_2\Big(\Big|1 + \frac{\rho}{N\sigma^2}\mathbf{p}_N^H\mathbf{H}^H\mathbf{T}_{N-1}^{-1}\mathbf{H}\mathbf{p}_N\Big|\Big)$ in (3.4) is the achievable sub-rate of the N-th sub-antenna array, while the first term $\log_2(|\mathbf{T}_{N-1}|)$ shares the same form as (3.3). This implication helps in further decomposition of $\log_2(|\mathbf{T}_{N-1}|)$ and after N such decomposition, the total achievable rate R in (3.3) is given as

$$R = \sum_{n=1}^{N} \log_2\Big(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_n^H\mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{p}_n\Big), \tag{3.5}$$

where $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$ and $\mathbf{T}_0 = \mathbf{I}_N$. These series of sub-rate optimization problems of sub-antenna arrays can be optimized one by one. The use of SIC for multiuser detection [17] enables to optimize the achievable sub-rate of each sub-antenna array and update the corresponding matrix **T**, successively as shown in Fig. 3.2 [1].

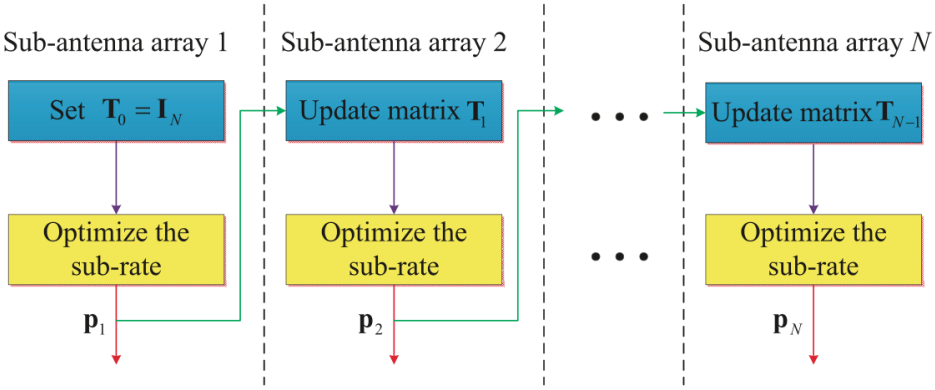

Figure 3.2: Proposed SIC-based hybrid precoding

3.2 Solution Methodology

3.2.1 Solution to the sub-rate optimization problem

According to (3.5), by designing the n-th precoding vector $\mathbf{p}_n$, the sub-rate optimization problem of the n-th sub-antenna array is given as

$$\mathbf{p}_n^{\text{opt}} = \arg\max_{\mathbf{p}_n \in \mathcal{F}} \log_2\left(1 + \frac{\rho}{N\sigma^2}\mathbf{p}_n^H \mathbf{G}_{n-1}\mathbf{p}_n\right), \tag{3.6}$$

where $G_{n-1} = \mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}$, $\mathcal{F}$ is the set of all feasible vectors which satisfy the three constraints mentioned in Section 3.2. The n-th 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 in (3.6) can be written as

$$\bar{\mathbf{p}}_n^{\text{opt}} = \arg\max_{\bar{\mathbf{p}}_n \in \bar{\mathcal{F}}} \log_2\left(1 + \frac{\rho}{N\sigma^2}\bar{\mathbf{p}}_n^H \bar{\mathbf{G}}_{n-1}\bar{\mathbf{p}}_n\right), \tag{3.7}$$

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

$$\bar{\mathbf{G}}_{n-1} = \mathbf{R}\mathbf{G}_{n-1}\mathbf{R}^H = \mathbf{R}\mathbf{H}^H\mathbf{T}_{n-1}^{-1}\mathbf{H}\mathbf{R}^H, \tag{3.8}$$

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

The singular value decomposition (SVD) of the Hermitian matrix $\bar{\mathbf{G}}_{n-1}$ is defined as $\bar{\mathbf{G}}_{n-1} =$

$\mathbf{V}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{\Sigma}$ is an $M \times M$ diagonal matrix which contains the singular values of $\bar{\mathbf{G}}_{n-1}$ in a decreasing order and $\mathbf{V}$ is an $M \times M$ unitary matrix. The optimal unconstrained precoding vector of (3.7) is the first column $\mathbf{v}_1$ of $\mathbf{V}$, i.e., the first right singular vector of $\bar{\mathbf{G}}_{n-1}$ [8]. However, according to the constraints mentioned in Section 3.2, the authors have noted that $\bar{\mathbf{p}}_n^{\text{opt}}$ cannot be chosen directly as v_1 since the elements of $\mathbf{v}_1$ do not obey the constraint of same amplitude (i.e. Constraint 2). Hence, to find a feasible solution to the sub-rate optimization problem (3.7), the authors have converted (3.7) into another equivalent form as

$$\bar{\mathbf{p}}_n^{\text{opt}} = \arg\min_{\bar{\mathbf{p}}_n \in \bar{\mathcal{F}}} \|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2 , \tag{3.9}$$

where $\mathbf{v}_1$ is the first right singular vector of $\bar{\mathbf{G}}_{n-1}$. This indicates that it is possible to find a feasible precoding vector $\bar{\mathbf{p}}_n$, which is sufficiently close to the optimal but unpractical precoding vector $\mathbf{v}_1$, to maximize the achievable sub-rate of the n-th sub-antenna array. Since $\bar{\mathbf{p}}_n = d_n \bar{\mathbf{a}}_n$ according to (3.1), the target $\|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2$ in (3.9) is rewritten as

$$\|\mathbf{v}_1 - \bar{\mathbf{p}}_n\|_2^2 = \Big(d_n - \text{Re}\big(\mathbf{v}_1^H \bar{\mathbf{a}}_n\big)\Big)^2 + \Big(1 - \big[\text{Re}\big(\mathbf{v}_1^H \bar{\mathbf{a}}_n\big)\big]^2\Big), \tag{3.10}$$

which is obtained considering 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 $1/\sqrt{M}$. The optimal $\bar{\mathbf{a}}_n^{\text{opt}}$ to maximize $|\text{Re}(\mathbf{v}_1^H \bar{\mathbf{a}}_n)|$ is

$$\bar{\mathbf{a}}_n^{\text{opt}} = \frac{1}{\sqrt{M}} e^{j\text{angle}(\mathbf{v}_1)}, \tag{3.11}$$

where $\text{angle}(\mathbf{v}_1)$ represents the phase vector of $\mathbf{v}_1$. Correspondingly, the optimal choice of d_n^{opt} is

$$d_n^{\text{opt}} = \text{Re}\Big(\mathbf{v}_1^H \bar{\mathbf{a}}_n\Big) = \frac{\|\mathbf{v}_1\|_1}{\sqrt{M}}. \tag{3.12}$$

Based on (3.11) and (3.12), the optimal solution $\bar{\mathbf{p}}_n^{\text{opt}}$ to the optimization problem (3.9) is given as

$$\bar{\mathbf{p}}_n^{\text{opt}} = d_n^{\text{opt}} \bar{\mathbf{a}}_n^{\text{opt}} = \frac{1}{M} \|\mathbf{v}_1\|_1 \, e^{j\text{angle}(\mathbf{v}_1)}. \tag{3.13}$$

As $\mathbf{v}_1$ 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 which implies $\|\bar{\mathbf{p}}_n^{\text{opt}}\|_2^2 \leq 1$ and the optimal solution $\bar{\mathbf{p}}_n^{\text{opt}}$ for $n = 1, 2, \ldots, N$ have a similar form. So, it can be concluded that

$$\left\|\mathbf{P}^{\text{opt}}\right\|_F^2 = \left\|\text{diag}\left\{\bar{\mathbf{p}}_1^{\text{opt}}, \ldots, \bar{\mathbf{p}}_N^{\text{opt}}\right\}\right\|_F^2 \leq N, \tag{3.14}$$

which satisfies the total transmit power constraint (*Constraint 3*).

Now, when the $\bar{\mathbf{p}}_n^{\text{opt}}$ is obtained for the n-th sub-antenna array, the matrices $\mathbf{T}_n = \mathbf{I}_K + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{P}_n\mathbf{P}_n^H\mathbf{H}^H$ (3.5) and $\bar{\mathbf{G}}_n = \mathbf{R}\mathbf{H}^H\mathbf{T}_n^{-1}\mathbf{H}\mathbf{R}^H$ (3.8) can be updated. Then, the same method can be applied again to optimize the achievable sub-rate of the $(n+1)$-th sub-antenna array.

3.2.2 Low-complexity algorithm

Based on the observation that the SVD of $\bar{\mathbf{G}}_{n-1}$ need not to be computed to achieve $\mathbf{\Sigma}$ and $\mathbf{V}$, since the first column $\mathbf{v}_1$ of $\mathbf{V}$ is sufficient to obtain $\bar{\mathbf{p}}_n^{\text{opt}}$, a power iteration algorithm [18], which is used to compute the largest eigenvalue and the corresponding eigenvector of a diagonalizable matrix can be realized and the calculation can be simplified as

$$\bar{\mathbf{G}_n} \approx \bar{\mathbf{G}}_{n-1} - \frac{\frac{\rho}{N\sigma^2}\Sigma_1^2\mathbf{v}_1\mathbf{v}_1^H}{1 + \frac{\rho}{N\sigma^2}\Sigma_1}, \tag{3.15}$$

where Σ_1 and $\mathbf{v}_1$ are the largest singular value and first right singular vector of $\bar{\mathbf{G}}_{n-1}$, respectively.

Thus, the obtained Σ_1 and $\mathbf{v}_1$ can be exploited to update $\bar{\mathbf{G}_n}$, where only one vector-to-vector multiplication is involved. The pseudo-code of the proposed SIC-based hybrid precoding is summarized in **Algorithm 1**.

Algorithm 1: SIC-based hybrid precoding

Input : $\bar{\mathbf{G}}_0$

for $1 \leq n \leq N$ **do**

  1) Compute $\mathbf{v}_1$ and Σ_1 of $\bar{\mathbf{G}}_{n-1}$

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

    $\bar{\mathbf{p}}_n^{\text{opt}} = \frac{1}{M}\|\mathbf{v}_1\|_1 \, e^{j\text{angle}(\mathbf{v}_1)}$ (3.11)-(3.13)

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

end

Output: (1) $\mathbf{D} = \text{diag}\{d_1^{\text{opt}}, \ldots, d_N^{\text{opt}}\}$

(2) $\mathbf{A} = \text{diag}\{\bar{\mathbf{a}}_1^{\text{opt}}, \ldots, \bar{\mathbf{a}}_N^{\text{opt}}\}$

(3) $\mathbf{P} = \mathbf{AD}$

3.3 Numerical Results

In this section, simulation results for the achievable rate and energy efficiency to evaluate the performance of the proposed SIC-based hybrid precoding have been provided and the performance is compared to the fully-connected architecture of spatially sparse precoding [8] and the optimal unconstrained precoding based on the SVD of the channel matrix. The sub-connected architecture of conventional analog precoding [19] and the optimal unconstrained precoding (i.e., $\bar{\mathbf{p}}_n^{\text{opt}} = \mathbf{v}_1$) is also considered as benchmarks for comparison. The parameters for simulation are as follows: The channel matrix has been generated based on the channel model [11] described in Section 3.1, the number of effective channel paths is considered as $L = 3$ [8] and the carrier frequency is set as 28 GHz [20]. Both the transmit and receive antenna arrays are ULAs with antenna spacing $d = \lambda/2$. The AoDs and the AoAs are assumed to follow the uniform distribution within $\big[-\frac{\pi}{6}, \frac{\pi}{6}\big]$ and $\big[-\pi, \pi\big]$, respectively [21]. The maximum number of iterations is set as $S = 5$ to run the **Algorithm 1**. Lastly, the SNR is defined as $\frac{\rho}{\sigma^2}$.

Fig. 3.3 shows the achievable rate comparison in a mm-Wave MIMO system, where $NM \times K = 64 \times 16$ and the number of RF chains is $N = 8$. From Fig. 3.3, it can be observed that the proposed SIC-based hybrid precoding is superior to the conventional analog precoding with sub-connected architecture in the whole simulated SNR range. Also, the near-optimal performance of SIC-based hybrid precoding is verified as it reaches around 99% of the rate achieved by the optimal unconstrained precoding with sub-connected architecture [19]. More importantly, Fig. 3.3 shows that the performance of SIC-based hybrid precoding is close to the spatially sparse

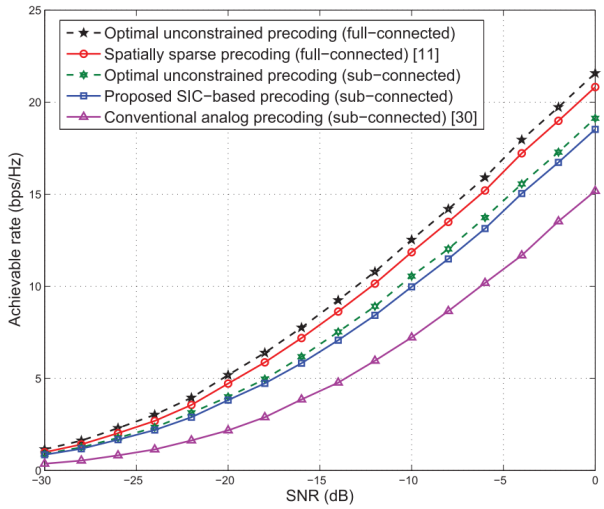

Figure 3.3: Achievable rate comparison of an $NM \times K = 64 \times 16$ ($N = 8$) mm-Wave MIMO system.

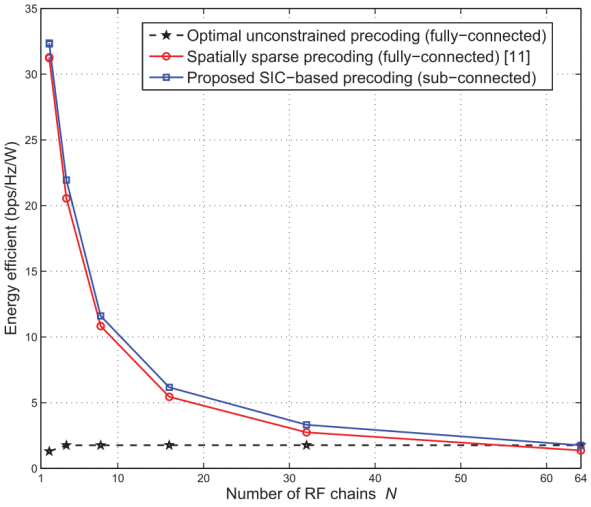

Figure 3.4: Energy efficiency comparison against the numbers of RF chains N, where $NM = K = 64$.

precoding and the optimal unconstrained precoding with fully connected architecture. Fig. 3.4 shows the energy efficiency comparison against the number of RF chains N, where SNR $= 0$ dB, $NM = K = 64$. Both the conventional spatially sparse precoding and the proposed SIC-based precoding can achieve higher energy efficiency than the optimal unconstrained precoding (also known as the fully digital precoding), especially when the number of RF chains N is limited. Besides, the proposed SIC- based precoding is more energy-efficient than the conventional spatially sparse precoding.

4 Hybrid Precoding in Deep Learning based Millimeter-Wave Massive MIMO System

The second paper we analyze in this report is [2]: H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive mimo for hybrid precoding", *IEEE Transactions on Vehicular Technology*, vol. 68,no. 3, pp. 3027–3032, 2019. The choice of this paper was due to its focus on revealing that the DNN can facilitate the hybrid precoding dedicated by its super-excellent recognition and representation abilities and can leverage the spatial statistics of the large antenna systems in mm-Wave massive MIMO system. The following subsections describe in detail about the research problem, solution methodology and results.

4.1 Problem Formulation

In [2], a typical mm-Wave massive MIMO system is considered where a BS is designed with a uniform linear array (ULA) of N_t transmit antennas and user with N_r receive antennas. The BS sends N_s independent data streams to the user assuming that there is no information available on any of the communication links. It is also assumed that the BS and the user have N_t^{RF} and N_r^{RF} RF chains, respectively, which meet the $N_s \leq N_t^{RF} \leq N_t$ and $N_t \leq N_r^{RF} \leq N_r$ requirements [8]. Once again, the authors have used the well-known Saleh-Valenzuela (SV) channel model where the channel matrix is $\mathbf{H} \in \mathbb{C}^{N_t \times N_r}$. A hybrid precoder is the combination of a high-dimensional analog precoder $\mathbf{D}_A \in \mathbb{C}^{N_t \times N_t^{RF}}$ and a low-dimensional digital precoder $\mathbf{D}_D \in \mathbb{C}^{N_t^{RF} \times N_s}$ and is represented as $\mathbf{D} = \mathbf{D}_A\mathbf{D}_D \in \mathbb{C}^{N_t \times N_s}$. Hence, the transmitted signal $\mathbf{x}$ is given as

$$\mathbf{x} = \mathbf{D}\mathbf{s} = \mathbf{D}_A\mathbf{D}_D\mathbf{s}, \tag{4.1}$$

where $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ represents the source signal with normalized power $E[\mathbf{s}\mathbf{s}^H] = \mathbf{I}_{N_s}$ and to satisfy the constraint of transmit power, it is assumed that $\text{tr}\{\mathbf{D}\mathbf{D}^H\} \leq N_s$ [22]. The received signal vector is given as

$$\begin{aligned} \mathbf{y} &= \mathbf{B}^H\mathbf{H}\mathbf{x} + \mathbf{B}^H\mathbf{n} \\ &= (\mathbf{B}_D^H\mathbf{B}_A^H)\mathbf{H}\mathbf{D}_A\mathbf{D}_D\mathbf{s} + \mathbf{B}_D^H\mathbf{B}_A^H\mathbf{n}, \end{aligned} \tag{4.2}$$

where $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_s})$ represents the AWGN noise, $\mathbf{B}_D \in \mathbb{C}^{N_r^{RF} \times N_s}$ and $\mathbf{B}_A \in \mathbb{C}^{N_r \times N_r^{RF}}$ represent a digital combiner and a analog combiner, respectively and $\mathbf{B}^H = \mathbf{B}_D^H \mathbf{B}_A^H$ represents a hybrid combiner. The analog precoder is always installed using phase shifters (PSs) and hence, imposing the constraints on the elements of $\mathbf{D}_A$ and $\mathbf{B}_A$ as

$$|\{\mathbf{D}_A\}_{i,j}| = \frac{1}{\sqrt{N_t}}, \;\; |\{\mathbf{B}_A\}_{i,j}| = \frac{1}{\sqrt{N_r}}. \tag{4.3}$$

In the next section, we will analyze the proposed novel precoding framework which employs the deep neural network (DNN) and exploits the sparsity present in the mm-Wave channel to improve the performance of the hybrid precoding.

4.2 Solution Methodology

4.2.1 DNN Learning Framework

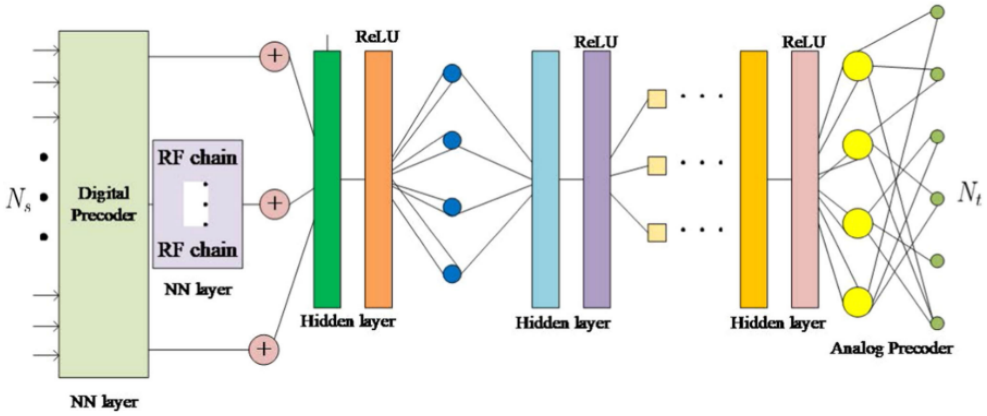

Figure 4.1: Deep learning-based mm-Wave hybrid precoding framework.

The proposed DNN framework to realize hybrid precoding is shown in Fig. 4.1 where the input layer is a fully-connected (FC) layer with 128 neurons for capturing features of the input data. In this layer, the length of each training sequence is determined by its dimension. The next two hidden layers are also FC layers with 400 neurons and 256 neurons, respectively and used for processing encoding operation. A 200 neurons AWGN noise layer is also considered for mixing distortion. Subsequently, the next two hidden layers with 128 neurons and 64 neurons, respectively, are considered for decoding and the last layer represents an output layer deployed

to generate expected output signals of the network. The authors have used the ReLU function as the activation function of the input layer and the hidden layers and designed a special activation function to enforce the power constraint in the output layer as

$$f(\mathbf{s}) = \min(\max(\mathbf{s}, 0), N_s). \tag{4.4}$$

4.2.2 Learning policy

To simplify the mapping relation of the hybrid precoding, GMD method is employed to decompose the complex mm-Wave massive MIMO channel matrix and $\mathbf{H}$ is formulated by

$$\mathbf{H} = \mathbf{WQR}^H = [\mathbf{W}_1, \mathbf{W}_2] \begin{bmatrix} \mathbf{Q}_1 & * \\ \mathbf{0} & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1^H \\ \mathbf{R}_2^H \end{bmatrix}, \tag{4.5}$$

where both the combiner $\mathbf{W}_1 \in \mathbb{C}^{N_r \times N_s}$ and the precoder $\mathbf{R}_1 \in \mathbb{C}^{N_t \times N_s}$ are semi-unitary matrices, $\mathbf{Q}_1 \in \mathbb{C}^{N_s \times N_s}$ is an upper triangular matrix, while $*$ is an arbitrary matrix which can be neglected. The largest N_s singular values are formulated as $q_{i,i} = (\delta_1, \delta_2, ..., \delta_{N_s})^{\frac{1}{N_s}} \in \bar{\mathbf{q}}, \forall i$, where $q_{i,j}$ represent the elements in matrix $\mathbf{Q}_1$. The received signal is given as

$$\begin{aligned} \mathbf{y} &= \mathbf{B}^H\mathbf{Hx} + \mathbf{B}^H\mathbf{n} \\ &= \mathbf{W}_1^H\mathbf{HR}_1\mathbf{s} + \mathbf{W}_1^H\mathbf{n} \\ &= \mathbf{Q}_1\mathbf{s} + \mathbf{W}_1^H\mathbf{n}. \end{aligned} \tag{4.6}$$

To train the hybrid precoder to realize precoding, the loss function is given as

$$\begin{aligned} loss &= \|\mathbf{R}_1 - \mathbf{R}_A\mathbf{R}_D\|_F \\ &= \sqrt{\textstyle\sum_{i=1}^{\min\{N_t, N_s\}}} \delta_i^2(\mathbf{R}_1 - \mathbf{R}_A\mathbf{R}_D), \end{aligned} \tag{4.7}$$

where $\mathbf{R}_A$ and $\mathbf{R}_D$ represent the GMD-based analog and digital precoder, respectively. In addition, $\delta_i(\mathbf{R}_1 - \mathbf{R}_A\mathbf{R}_D)$ represent the singular values of matrix $(\mathbf{R}_1 - \mathbf{R}_A\mathbf{R}_D)$. In (4.7), the constraints $|\{\mathbf{R}_A\}_{i,j}| = \frac{1}{\sqrt{N_t}}$ and $\text{tr}(\mathbf{R}_A\mathbf{R}_D\mathbf{R}_D^H\mathbf{R}_A^H) \leq N_s$ need to be satisfied. Next, to construct

an autoencoder, the authors have employed the DNN framework as

$$\mathbf{R}_1 = f(\mathbf{R}_A \mathbf{R}_D; \Omega), \tag{4.8}$$

where Ω represent the dataset of the samples and $f(\cdot)$ represents the mapping relation for which, the detailed training procedure is provided as follows.

The matrices $\mathbf{R}_A$ and $\mathbf{R}_D$ are initialized as empty matrices after which the random data sequences are generated in the DNN. The DNN is trained with the input data sequences as per different channel conditions and $\mathbf{R}_A$ and $\mathbf{R}_D$ is updated. Correspondingly, the physical AoA θ_p^r and AoD θ_p^t is generated randomly and the bias between $\mathbf{R}_1$ and $\mathbf{R}_A \mathbf{R}_D$ is obtained from the output layer of the DNN. Thus, the training set Ω is achieved which consists of the structural feature of the mm-Wave massive MIMO model and the input data sequences and output data of the DNN. The authors have employed stochastic gradient descent (SGD) algorithm with momentum to process the loss function given as

$$\mathbf{R}_A^{j+1} = \mathbf{R}_A^j + v, \tag{4.9}$$

$$\mathbf{R}_D^{j+1} = \mathbf{R}_D^j + v, \tag{4.10}$$

where v represents the velocity for facilitating the gradient element and j represents the iteration. $\mathbf{R}_A^0$ and $\mathbf{R}_D^0$ are assumed to be the randomly generated initial solutions for $\mathbf{R}_A$ and $\mathbf{R}_D$, respectively. The update procedure of v can be given as

$$\begin{aligned} v &= \alpha v - \epsilon g \\ &= \alpha v - \epsilon \frac{1}{N} \bigtriangledown_{\mathbf{R}_A, \mathbf{R}_D} \sqrt{\textstyle\sum_{i=1}^{\min\{N_t, N_s\}} \delta_i^2 (\mathbf{R}_1 - \mathbf{R}_A \mathbf{R}_D)}, \end{aligned} \tag{4.11}$$

where α represents the momentum parameter and ϵ represents the learning rate. Also, g and N represent the gradient element and the number of samples, respectively. The learning framework for hybrid precoding is summarized below in **Algorithm 2**.

Algorithm 2: DNN-based hybrid precoding

Input : The physical AoA θ_p^r and AoD θ_p^t, environment simulator.
Output: Optimized precoder $\mathbf{R}_1$.

Initialization: The amount of iteration is initialized as $j = 0$ and the weight is $\omega = 0$. Meanwhile, initialize error threshold as $\tau = 10^{-7}$. Furthermore, set $\mathbf{R}_A = \mathbf{0}$ and $\mathbf{R}_D = \mathbf{0}$.
Product a series of training sequences. Also, θ_p^r and θ_p^t are generated randomly.
Construct the proposed DNN framework.
Process the environment simulator to simulate wireless channel with noise.
while *error* $\geq \tau$
    Train the DNN by processing the SGD with momentum according to (4.9), (4.10), (4.11).
    Update $\mathbf{R}_A$ and $\mathbf{R}_D$.
    Obtain the bias between $\mathbf{R}_1$ and $\mathbf{R}_A\mathbf{R}_D$ from the output layer of the network.
end
return: Optimized precoder $\mathbf{R}_1$.

Table 4.1 provides the complexity comparison of the DL-based scheme to SIC-based scheme [1] and the spatially sparse hybrid precoding [8] . Here, for simplicity of the notations in all of these three works [1, 2, 8], let us assume that N_s independent input data streams are transmitted by N_t transmit antennas through L effective channel paths to K users, each with N_r receive antennas. The advantage of the proposed DL-based hybrid precoding scheme [2] is that it has the lowest computational complexity compared to the SIC-based hybrid precoding [1] and SVD-based spatially sparse precoding hybrid precoding [8].

Table 4.1: Complexity comparison

	Number of Multiplications	**Number of Divisions**
DL-based scheme [2]	$\mathcal{O}(N_s N_t^2)$	$\mathcal{O}(L^2)$
SIC-based hybrid precoding [1]	$\mathcal{O}(N_t^2(N_s N_r + K))$	$\mathcal{O}(2N_r N_s)$
Spatially sparse precoding [8]	$\mathcal{O}(N_r^4 N_t + N_r^2 L^2 + N_r^2 N_t^2 L)$	$\mathcal{O}(2N_r^3)$

4.3 Numerical Results

In this section, the performance of the proposed DL-based mm-Wave massive MIMO scheme is presented. Without loss of generality, the mm-Wave channel model has been generated as described in Section 3.1 with $P = 3$ at 28 GHz and the angles are generated randomly in the domain of $\{-\pi/2, \pi/2\}$. The ray-tracing simulator [23] is introduced to generate channel measurements

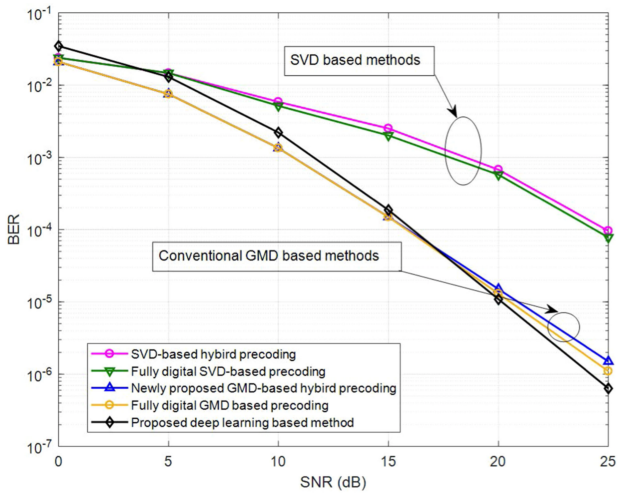

Figure 4.2: BER versus SNR.

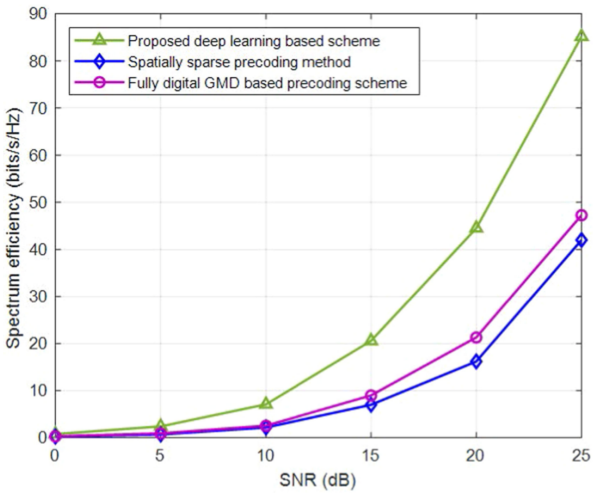

Figure 4.3: Spectrum efficiency versus SNR.

such as the AoAs and the AoDs and to construct and process the DNN framework, *Keras* is used, which is an open-source neural network library written in Python and designed to enable fast experimentation with deep neural networks. The model is trained for 45k iterations with 0.001 learning rate. Also, the momentum of 0.85 and weight decay of 0.0001 are introduced, and there are 500k samples and 20k samples in the training set and the testing set, respectively.

In Fig. 4.2 [2], the BER performance of the deep learning (DL)-based proposed scheme is shown where it is compared against the SVD-based spatially sparse hybrid precoding [8], fully digital SVD-based precoding, fully GMD-based precoding and the new GMD-based precoding [22]. The deep neural network (DNN) architecture exploits the structural information of the mm-Wave massive MIMO system which, in turn, makes the DL-based precoding superior to other conventional precoding schemes. This implies that the DL can be useful in solving the existing non-convex optimization problem in hybrid precoding. In Fig. 4.3 [2], the spectrum efficiency is shown for the proposed DL-based hybrid precoding, the spatially sparse precoding [8] and the fully digital GMD-based precoding. From Fig. 4.3, it can be observed that the proposed DL-based hybrid precoding scheme again achieves better performance compared to other schemes. Additionally, when the SNR increases, the performance gap of the DL-based scheme and the other schemes becomes larger, implying the further superiority of the proposed method.

5 Critical Review and Future Work

In general, both the papers have addressed the relevant problem of improving the hybrid precoding performance in mm-Wave massive MIMO systems with innovative solutions, however, on one hand, first paper focuses on energy-efficiency of the system and considers the achievable rate as the performance metric and on the other hand, the second paper focuses on spectrum-efficiency of the system considering the bit-error ratio (BER) as the performance metric. As the performance metrics are different in these two papers, their approaches towards the problem are also different.

The authors in the first paper have used a sub-connected ULA antenna array architecture, which consumes less energy, and proposed a low-complexity solution to optimize the total achievable rate. The proposed method decomposes 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 and maximizes the achievable sub-rate of each sub-antenna array by simply seeking a precoding vector sufficiently close to the unconstrained optimal solution. In the second paper, the authors have used a fully-connected ULA antenna array architecture and proposed a deep learning-based framework for mm-Wave massive MIMO to facilitate the hybrid precoding. The proposed method uses a deep neural network (DNN) as an autoencoder where the activation functions optimize multiple layers of the network and create corresponding relations to extract the sparse features of the mm-Wave channel model. Nevertheless, both methods are successful in designing the hybrid precoding scheme considering all the non-convex constraints imposed on the system yet both of these methods achieve near-optimal performances for their respective performance metrics.

We understand that hybrid precoding is an effective solution for the mm-Wave massive MIMO systems to significantly decrease the number of radio frequency (RF) chains without an apparent performance loss. However, in both of the proposed methods, the hybrid precoding algorithm assumes

- Infinite or high-resolution phase shifters (PSs) equipped at the analog domain. Obviously, the available precoding gains depend on the resolution of PSs, whereas high-resolution PSs

not only bring large precoding gains but also high complexity and power consumption.

- The number of RF chains with high-resolution DACs/ADCs in order to achieve satisfactory performances but at increasing cost and power consumption as the frequency increases.

- The channel model as a narrowband mm-Wave channel, however, the actual mm-Wave systems are likely to be wideband due to the availability of substantial bandwidth at mm-Wave frequencies [24].

Moreover, in the second paper, training data has been generated from traditional geometric mean decomposition (GMD) method, which suffers from a performance limitation and is too weak to exploit the sparsity characteristics of mm-Wave channel.

This literature review has provided different insights on the design of hybrid precoders and gave us a glance at the limitations of these structures. We believe that there is still a significant performance gap between the hybrid precoding algorithms and the optimal full-digital solution. Hence, considering all the above-mentioned issues, we suggest the following research topics:

- To address the high energy consumption problem, the usage of low-resolution PSs instead of HR-PSs [25, 26] and low-resolution DACs/ADCs [27–29] appears to be a feasible solution and can be explored further. Also, an appropriate trade-off between energy efficiency and system sum-rate for a wideband mm-Wave channel is still an open research problem.

- To solve the design of finding the best hybrid precoder and improve the performance of the system in mm-Wave massive MIMO systems, the idea of cross-entropy optimization (CEO) in deep learning solutions, which was developed initially for machine learning applications, to search the optimal analog precoder intelligently, can be utilized. The cross-entropy (CE) algorithm is a general optimization method that has been applied successfully to many NP-hard combinatorial problems [30].

6 Conclusion

In this report, we have discussed the motivations and the requirements of hybrid precoding in mm-Wave massive MIMO systems. The design of hybrid precoding is very challenging due to the complexity of the underlying hardware and frequently results in approximate sub-optimal solutions. Here, we analyzed two research papers that have proposed hybrid precoding designs for mm-Wave massive MIMO systems. These papers illustrate the wide variety of solutions that can be derived from the hybrid precoding problem and the different approaches that may be used in the design process. By studying these papers and keeping their limitations in mind, we have further proposed future research directions of this research, which mainly consist of developing more energy- or spectral- efficient and less computationally complex solutions. Further, we have concluded that to achieve better hybrid precoding performance, considering the difficulties and complexities of deducing an accurate analytical signal model that embraces all the hardware constraints and imperfections, it could be more practical and of great benefit to leverage deep learning (DL) in mm-Wave massive MIMO systems.

7 References

[1] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmwave mimo systems with large antenna arrays," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 998–1009, 2016.

[2] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive mimo for hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3027–3032, 2019.

[3] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive mimo communication for future wireless systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 836–869, 2017.

[4] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin *et al.*, "Millimeter wave communications for future mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909–1935, 2017.

[5] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive mimo: The next wireless revolution?" *IEEE Communications Magazine*, vol. 52, no. 9, pp. 56–62, 2014.

[6] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave mimo systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[7] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, "Mimo precoding and combining solutions for millimeter-wave systems," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 122–131, 2014.

[8] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave mimo systems," *IEEE transactions on wireless communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[9] Y. 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 Transactions on Signal Processing*, vol. 63, no. 2, pp. 305–317, 2014.

[10] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE transactions on wireless communications*, vol. 14, no. 11, pp. 6481–6494, 2015.

[11] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[12] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Communications*, vol. 14, no. 11, pp. 92–111, 2017.

[13] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," *MIT press*, 2016.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[15] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2010.

[16] S. Mumtaz, J. Rodriguez, and L. Dai, "Mmwave massive mimo: A paradigm for 5g," *Academic Press*, 2016.

[17] 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 transactions on signal processing*, vol. 56, no. 8, pp. 3627–3637, 2008.

[18] C. Van Loan and G. Golub, "Matrix computations," *Baltimore, MD: Johns Hopkins University Press*, vol. 3, 2012.

[19] O. El Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave mimo transmitters with multiple antenna sub-arrays," *2013 IEEE Global Communications Conference (GLOBECOM)*, pp. 3476–3480, 2013.

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

[21] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up mimo: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[22] T. Xie, L. Dai, X. Gao, M. Z. Shakir, and J. Li, "Geometric mean decomposition based hybrid precoding for millimeter-wave massive mimo," *China Communications*, vol. 15, no. 5, pp. 229–238, 2018.

[23] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5g mimo data for machine learning: Application to beam-selection using deep learning," *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018.

[24] T. Mir, M. Z. Siddiqi, U. Mir, R. Mackenzie, and M. Hao, "Machine learning inspired hybrid precoding for wideband millimeter-wave massive mimo systems," *IEEE Access*, vol. 7, pp. 62 852–62 864, 2019.

[25] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmwave massive mimo systems," *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2017.

[26] S. Gao, Y. Dong, C. Chen, and Y. Jin, "Hierarchical beam selection in mmwave multiuser mimo systems with one-bit analog phase shifters," *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, pp. 1–5, 2016.

[27] Y. Dong and L. Qiu, "Spectral efficiency of massive mimo systems with low-resolution adcs and mmse receiver," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1771–1774, 2017.

[28] J. Zhang, L. Dai, S. Sun, and Z. Wang, "On the spectral efficiency of massive mimo systems with low-resolution adcs," *IEEE Communications Letters*, vol. 20, no. 5, pp. 842–845, 2016.

[29] C. Kong, C. Zhong, S. Jin, S. Yang, H. Lin, and Z. Zhang, "Full-duplex massive mimo relaying systems with low-resolution adcs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5033–5047, 2017.

[30] R. Y. Rubinstein and D. P. Kroese, "Simulation and the monte carlo method," *John Wiley & Sons*, vol. 10, 2016.