# Enabling Efficient Communications Over Millimeter-wave Massive MIMO Channels Using Hybrid Beamforming

by

Ahmed Wagdy Shaban

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2020

**Examining Committee Membership**

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

| | |
|---|---|
| External Examiner: | Halim Yanikomeroglu<br>Professor, Dept. of Systems and Computer Engineering<br>Carleton University |
| Supervisor: | Oussama Damen<br>Professor, Dept. of Electrical and Computer Engineering<br>University of Waterloo |
| Internal-External Member: | Stephen A. Vavasis<br>Professor, Dept. of Combinatorics and Optimization<br>University of Waterloo |
| Internal Member: | Guang Gong<br>Professor, Dept. of Electrical and Computer Engineering<br>University of Waterloo |
| Internal Member: | Patrick Mitran<br>Professor, Dept. of Electrical and Computer Engineering<br>University of Waterloo |

## Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

The use of massive multiple-input multiple-output (MIMO) over millimeter wave (mmWave) channels is the new frontier for fulfilling the exigent requirements of next-generation wireless systems and solving the wireless network impending crunch. Massive MIMO systems and mmWave channels offer larger numbers of antennas, higher carrier frequencies, and wider signaling bandwidths. Unleashing the full potentials of these tremendous degrees of freedom (dimensions) hinges on the practical deployment of those technologies. Hybrid analog and digital beamforming is considered as a stepping-stone to the practical deployment of mmWave massive MIMO systems since it significantly reduces their operating and implementation costs, energy consumption, and system design complexity. The prevalence of adopting mmWave and massive MIMO technologies in next-generation wireless systems necessitates developing agile and cost-efficient hybrid beamforming solutions that match the various use-cases of these systems. In this thesis, we propose hybrid precoding and combining solutions that are tailored to the needs of these specific cases and account for the main limitations of hybrid processing. The proposed solutions leverage the sparsity and spatial correlation of mmWave massive MIMO channels to reduce the feedback overhead and computational complexity of hybrid processing.

Real-time use-cases of next-generation wireless communication, including connected cars, virtual-reality/augmented-reality, and high definition video transmission, require high-capacity and low-latency wireless transmission. On the physical layer level, this entails adopting near capacity-achieving transmission schemes with very low computational delay. Motivated by this, we propose low-complexity hybrid precoding and combining schemes for massive MIMO systems with partially and fully-connected antenna array structures. Leveraging the disparity in the dimensionality of the analog and the digital processing matrices, we develop a two-stage channel diagonalization design approach in order to reduce the computational complexity of the hybrid precoding and combining while maintaining high spectral efficiency. Particularly, the analog processing stage is designed to maximize the antenna array gain in order to avoid performing computationally intensive operations such as matrix inversion and singular value decomposition in high dimensions. On the other hand, the low-dimensional digital processing stage is designed

to maximize the spectral efficiency of the systems. Computational complexity analysis shows that the proposed schemes offer significant savings compared to prior works where asymptotic computational complexity reductions ranging between $80\%$ and $98\%$. Simulation results validate that the spectral efficiency of the proposed schemes is near-optimal where in certain scenarios the signal-to-noise-ratio (SNR) gap to the optimal fully-digital spectral efficiency is less than $1$ dB.

On the other hand, integrating mmWave and massive MIMO into the cellular use-cases requires adopting hybrid beamforming schemes that utilize limited channel state information at the transmitter (CSIT) in order to adapt the transmitted signals to the current channel. This is so mainly because obtaining perfect CSIT in frequency division duplexing (FDD) architecture, which dominates the cellular systems, poses serious concerns due to its large training and excessive feedback overhead. Motivated by this, we develop low-overhead hybrid precoding algorithms for selecting the baseband digital and radio frequency (RF) analog precoders from statistically skewed DFT-based codebooks. The proposed algorithms aim at maximizing the spectral efficiency based on minimizing the chordal distance between the optimal unconstrained precoder and the hybrid beamformer and maximizing the signal to interference noise ratio for the single-user and multi-user cases, respectively. Mathematical analysis shows that the proposed algorithms are asymptotically optimal as the number of transmit antennas goes to infinity and the mmWave channel has a limited number of paths. Moreover, it shows that the performance gap between the lower and upper bounds depends heavily on how many DFT columns are aligned to the largest eigenvectors of the transmit antenna array response of the mmWave channel or equivalently the transmit channel covariance matrix when only the statistical channel knowledge is available at the transmitter. Further, we verify the performance of the proposed algorithms numerically where the obtained results illustrate that the spectral efficiency of the proposed algorithms can approach that of the optimal precoder in certain scenarios. Furthermore, these results illustrate that the proposed hybrid precoding schemes have superior spectral efficiency performance while requiring lower (or at most comparable) channel feedback overhead in comparison with the prior art.

**Acknowledgements**

All praise is due to Allah, the most merciful, the most compassionate, the creator, sustainer, and nourisher.

I would like to express my sincere gratitude to my parents Mr. Wagdy Shaban, Mrs. Naema Hassan and my soul-mates Ola Eltantawy, Radwa Shaban and Nada Shaban for their unconditional love, sacrifices, continuous support and unlimited encouragement. I would like to dedicate my deep and sincere appreciation to Ola Eltantawy for her endless understanding and unbearable sacrifices during the years of the Ph.D.

I would like to express my deepest regards and thanks to my others mothers Mrs. Elsayedah Hassan, Mrs. Wadeda Eleshmawy, and Mrs. Atyat Hassan. I wish they could have witnessed this finally happening. Special thanks to the members of my extended family, Mr. Mohamed Abdelhady, Dr. Rehab Abdelhady, Dr. Bahgat Eltantawy, Dr. Hala Nagaty and Dr. Ayman Eltantawy and Mr. Alaa Eltantawy for their continuous support and prayers.

I would like to express appreciation to my supervisor Prof. Oussama Damen for providing me with the opportunity, having the patience to allow me incubate and develop skills, and inspiring me to pursue success not only in academic endeavors but also in others aspects of life. I would like to thank him for sharpening my critical thinking skills and learning me accepting people's critiques in a positive way and act upon them to evolve and enhance my perspective.

I would like to thank Prof. Guang Gong, Prof. Patrick Mitran, and Prof. Stephen A. Vavasis for devoting their valuable time and intellect to oversee my progress, and for the fruitful critiques and observations during the proposal and the seminar. I would like also to thank Prof. Halim Yanikomeroglu for graciously agreeing to join the committee as the external examiner, and for allocating some of his precious time and effort to examine the thesis.

I feel fortunate to have many friends in Waterloo where their support and humor have alleviated the homesickness and their kindness bring part of the warmth of my family and Egypt to my daily life. Many thanks are to you, Ali Mahmoud, Mahmoud Allam, Mohamed Nassar, Ayman Eltaliawy, and Yehia Beltagy.

Warmest thanks to my dear friend Dr. Ahmed Hamza Abouelenein (the voice of wisdom) for the support and the guidance during the Ph.D.

Special thanks to my colleagues Kazem Izadinasab, Meysam Shahrbaf Motlagh and Subhajit Majhi for their support and encouragement.

Special thanks to my old but gold and childhood friends Mahmoud Sadek, Ahmed Yehia, Abdelrahman El Khoudary, Mohamed Hafez, Ahmed Farrag, and Ali Abdelhafiz.

Finally, I would like to render special thanks to Prof. Tamer Kattab and Dr. Ahmed El Shafie for their support and Prof. Essam Sourour for giving me the opportunity to start the academic research career.

## Dedication

To my beloved family especially my *father*.

# Table of Contents

# List of Tables

# List of Figures

# Abbreviations

| | |
|---|---|
| 3GPP | 3rd Generation Partnership Project |
| 5G | Fifth Generation |
| 6G | Sixth Generation |
| ADC | Analog to Digital Converter |
| BBU | Baseband Unit |
| C-RAN | Cloud/Centralized Radio Access Network |
| CSIR | Channel State Information at the Receiver |
| CSIT | Channel State Information at the Transmitter |
| DAC | Digital to Analog Converter |
| DoF | Degrees of Freedom |
| FDD | Frequency Division Duplex |
| i.i.d. | Independent Identically Distributed |
| JSDM | Joint Spatial Division Multiplexing |
| MIMO | Multiple-Input Multiple-Output |
| MMSE | Minimum Mean Squared Error |
| mmWave | Millimeter-wave |
| NR | New Radio |
| OFDM | Orthogonal Frequency Division Multiplexing |

| | |
|---|---|
| RRH | Remote Radio Head |
| SINR | Signal to Interference plus Noise Ratio |
| SLNR | Signal to Leakage plus Noise Ratio |
| SNR | Signal to Noise Ratio |
| TDD | Time Division Duplexing |
| ULA | Uniform Linear Array |
| ZF | Zero Forcing |

# Nomenclature

| | |
|---|---|
| $\mathbf{D}_M$ | The $M \times M$ DFT matrix |
| $\mathbf{x}$ | Boldface small letters denote vectors |
| $\mathbf{X}$ | Boldface capital letters denote matrices |
| $\mathbf{x}^T$, $\mathbf{X}^T$ | Transpose of $\mathbf{x}$, $\mathbf{X}$ |
| $\mathbf{x}^H$, $\mathbf{X}^H$ | Conjugate (or Hermitian) transpose of $\mathbf{x}$, $\mathbf{X}$ |
| $\mathbf{X}^{-1}$ | Inverse of matrix $\mathbf{X}$ |
| $\|\|\mathbf{x}\|\|_1$, $\|\|\mathbf{X}\|\|_1$ | $\ell_1$ norm of $\mathbf{x}$, $\mathbf{X}$ |
| $\|\|\mathbf{x}\|\|_2$, $\|\|\mathbf{X}\|\|_2$ | $\ell_2$ norm of $\mathbf{x}$, $\mathbf{X}$ |
| $\|\|\mathbf{X}\|\|_F$ | Frobenius norm of $\mathbf{X}$ |
| $\|\|\mathbf{X}\|\|_\infty$ | $\ell_\infty$ norm of $\mathbf{X}$ |
| $\det \mathbf{X}$ | Determinant of matrix $\mathbf{X}$ |
| Trace$(\mathbf{X})$ | Trace of matrix $\mathbf{X}$ |
| $x_{i,j}$ | $(i, j)$-th entry of matrix $\mathbf{X}$ |
| $\mathbf{x}_\star$ | The optimal solution |
| $\angle\mathbf{x}$, $\angle\mathbf{X}$ | The phases of the entries of $\mathbf{x}$, $\mathbf{X}$ |
| $\mathbb{C}^{N\times M}$ | The space of complex matrices of size $N \times M$ |
| $\mathbb{U}^{N\times M}$ | The space of complex matrices that have unit magnitude entries |
| $\arg\max_X\{\}$ | Return the indices of the largest $X$ elements in the set $\{\}$ |

| | |
|---|---|
| $\mathbf{I}_M$ | Identity matrix of dimension $M$ |
| $\log(\cdot)$ | Logarithm in base $2$ |
| $\mathcal{O}(\cdot)$ | Big O notation |
| $\mathcal{N}(0, \sigma_n^2)$ | Gaussian noise with zero mean and $\sigma_n^2$ variance |

# Chapter 1

# Introduction

The flourishing of radio frequency (RF) wireless communications stands out as one of the most momentous technological phenomena in the history of communication. With the exponential increase in the utilization of wireless communications in our daily life, there is an ever-growing necessity for developing wireless systems in terms of architecture, requirements, and applications [33, 39, 117]. Additionally, the majority of the evolving applications such as ultra-high-definition video streaming, augmented/virtual reality, car-to-x communications and not to mention the promising internet of things are data-heavy ones [117]. Ultimately, all these applications are envisioned to communicate together seamlessly within one gigantic wireless cellular network [39]. This requires ultra-high data rates and huge bandwidths to accommodate the foreseen end-users' demands with an acceptable quality of service. However, the RF bandwidth allocated to the current cellular networks is limited, lies between $300$ kHz and $3$ GHz, thereby becoming ever-congested. This, in turn, casts doubts on the ability of the current cellular systems to cope with those varied services and sophisticated applications, sheds light on the key factors for the future network infrastructure, and triggers the necessity for a potent alternative to the current network architecture and a viable complement to the crowded sub-$6$ GHz RF spectrum.

## 1.1 Next-generation Networks

The forthcoming wireless communication systems such as the fifth-generation (5G) and the sixth-generation (6G) cellular networks solve the dilemma of the ever-congested spectrum mainly by synergistically employing three fundamental concepts, namely, spectral aggregation, network densification strategies and cloud/centralized radio access network (C-RAN) [18,19,39,51,117]. Spectral aggregation is realized by enabling communication over higher parts of the electromagnetic spectrum which draws much attention to the exploitation of the millimeter-wave (mmWave) spectrum which extends up to 300 GHz. This leads to potentially aggregating non-contiguous fragments of bandwidth, i.e., sub-6 GHz and mmWave bands [20,110]. This, in turn, yields a dramatic increase in achievable data rates of users due to the huge bandwidth offered by mmWave bands [110]. Network densification is realized by increasing the number of antennas per unit area by packing a huge number of antenna elements into cells, i.e., utilizing the massive multiple-input multiple-output (MIMO) technology at terminals [59, 74], and shrinking the cells' areas to enable frequency reuse across a large geographic area and harvest benefits of small cells [19]. This leads to a dramatic increase in the system capacity due to the huge beamforming and multiplexing gains of massive MIMO, and the power gain and high frequency reuse factor of small cells. However, all these benefits come at the price of raising network construction costs and site acquisition difficulty, and increasing inter-cell interference. In order to reap the potential benefits of network densification while alleviating its challenges/shortcomings, the C-RAN architecture has been adopted in the 5G new radio (NR) [18–20]. The C-RAN is to split the base station processing between two main units, namely, the remote radio head (RRH) and baseband unit (BBU) where RRHs are distributively located across cells whereas BBUs are collocated at a centralized site like a data center forming a BBU pool or a could. The centralized BBU deployment reduces the number of required devices and power supplies at sites, lowers costs for rentals, energy, and equipment rooms [19]. Moreover, centralized BBUs cooperation allows for joint signal processing and resource management to alleviate the impact of inter-cell interference and improve the performance at cell edges. However, these potential benefits come at the price of increasing the required capacity of the fronthaul networks. Hence, mmWave fronthaul is de-

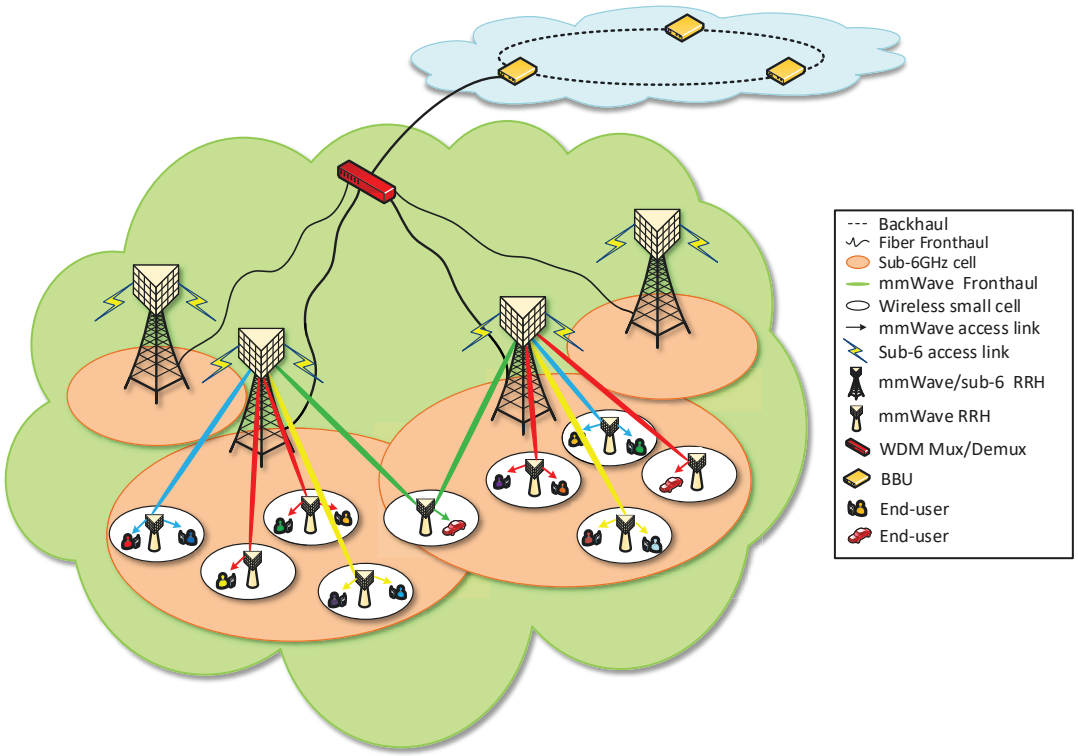


Figure 1.1: The utiliziation of mmWave, massive MIMO and CRAN in 5G/6G wireless networks.

sirable for CRAN architecture, as it is cost-effective, flexible, and easier to deploy compared to wired fronthauls [32, 104, 127]. This shows that the integration of three concepts into 5G networks is essential since each one enhances the presence of the others. Moreover, it shows that massive MIMO, mmWave, and two-stage processing are the pillars of the physical layer of 5G networks. A schematic diagram of the integration of these concepts into 5G networks is depicted in Fig. 1.1.

## 1.2 Physical Layer Challenges for Next-generation networks

The mmWave channel is endowed with huge unlicensed bandwidth, extends roughly from $30$ GHz to $300$ GHz [91]. The currently utilized mmWave bands are $28-38$ GHz, $57-64$ GHz, and $70-90$ GHz which offer channels with $500$ MHz, $2$ GHz and $4$ GHz of bandwidth, respectively, whereas most sub-$6$ GHz bands support channels with $50$ MHz of bandwidth [45, 91]. Therefore, mmWave systems can easily support a data rate of $10$ Gbps while the current wireless systems are limited to $1$ Gbps [26]. The feasibility of achieving a maximum data rate of $7$ Gbps has been demonstrated by the IEEE 802.11ad standard developed for $60$ GHz mmWave indoor communications. Despite all these advantages, the mmWave channel has some limitations. The most critical one is the insufficient radio link budget (considerable power loss) due to path loss combined with atmospheric absorption [103]. In particular, $27-55$ dB/Km power penalty for $60$ GHz system versus the current systems operating at frequencies below $6$ GHz [26]. Moreover, the large bandwidth of mmWave channels makes them more susceptible to frequency selectivity and inter-symbol interference.

The small wavelengths of mmWave bands facilitate making use of the massive MIMO concept where a few hundreds of radiators (antenna elements) can be tightly packed into a few centimeters antenna array. This huge number of antenna elements at each terminal can be exploited to increase the received signal power ratio by providing high antenna array gain (beamforming gain) that compensates for the severe power loss [59, 74]. Additionally, it can be utilized to cast a narrow communication beam which in turn decreases the delay spread of the channel, thus increasing the coherence bandwidth of the mmWave channels. Nonetheless, these great benefits of the usage of massive MIMO technology in mmWave bands come at the price of huge implementation and operational costs to the extent that it is considered a non-viable solution. This is so since massive MIMO systems typically adopt the fully-digital beamforming RF architecture. The expensive cost of this RF architecture is due to the usage of a large number of radio frequency (RF) chains, as many as the number of antenna elements, where each chain consists of a digital to analog converter (DAC), a mixer and a power amplifier at the transmitter side while analog to digital converter (ADC), mixer and low noise amplifier at the receiver side [45, 82].

These electronic components are costly, power-hungry, and have sophisticated circuits designs since they are working at high frequencies compared to the ones working at sub-6 bands. It is worth mentioning that the high speed and high precision DACs are the main factors of this problem due to the high sampling rate corresponding to the huge bandwidth of mmWave bands.

## 1.3 Hybrid Beamforming Solutions

A natural way to overcome this problem is to reduce the number of RF chains attached to the antenna array while keeping the same number of radiators (antenna elements) per array. This has motivated researchers to investigate hybrid analog and digital beamforming that uses fewer RF chains while utilizing the interesting properties of mmWave channels.

The hybrid beamforming[1] is a special MIMO architecture that has been developed primarily to significantly reduce the number of RF chains required for beamforming while offering comparable performance to the fully-digital one [4, 45, 82, 110]. It compensates for the loss in the digital processing gain, resulting from the use of a few numbers of RF chains, by the analog processing gain, ensuing from making use of a large number of RF phase shifters (much cheaper than RF chains). In this way, the hybrid beamforming managed to considerably reduce the cost and hardware complexity of massive MIMO systems and make them realizable in mmWave bands. The hybrid beamforming consists of two different consecutive beamforming stages that are being performed at two different frequencies, namely, the baseband and carrier frequencies. This combined structure naturally fits C-RAN architecture where the analog beamforming can be performed at RRHs whereas the digital beamforming can be carried out at BBUs [104, 127]. This shows that hybrid beamforming is one of the main cornerstones of the physical layer of 5G networks [4, 117].

[1] In the sequel, we mean by beamforming is either precoding at the transmitter side or combining at the receiver side, depending on the context

## 1.4 Research Objective

Hybrid beamformer can be mathematically written as a product of two different matrices i.e., analog and digital ones. The analog precoder/combiner, i.e., $\mathbf{F}_{RF}/\mathbf{W}_{RF}$, is a high-dimensional matrix that is applied in the RF domain through a network of phase shifters and various RF interconnections architectures. This RF hardware constraint is captured mathematically by restricting the entries of the analog beamformers $\mathbf{F}_{RF}/\mathbf{W}_{RF}$ to have unit magnitudes. On the other hand, the digital precoder/combiner is a low-dimensional matrix, i.e., $\mathbf{F}_{BB}/\mathbf{W}_{BB}$, which is performed in the baseband domain and has no constraints on its entries. However, both beamformers are coupled with a power constraint. As a result, the hybrid beamforming problem is to jointly design the hybrid precoder $\mathbf{F}_{RF}\mathbf{F}_{BB}$ and combiner $\mathbf{W}_{RF}\mathbf{W}_{BB}$ in order to optimize certain performance metrics such as spectral efficiency and minimum mean square error while considering the following three factors: (i) the constant gain constraints of the analog beamformers, (ii) the different RF hardware architectures such as fully or partially-connected structures, and (iii) the limited number of RF chains.

Unfortunately, solving such problems under these RF hardware constraints requires handling non-convex joint optimization over those four matrices where its global optima are intractable in most cases [4, 45, 82]. Securing sub-optimal solutions is a challenging problem and has been always the main focus of hybrid beamforming literature (a comprehensive discussion is available in Chapter 2). More importantly, these sub-optimal solutions should be obtained with very low computational complexity since we are dealing with a high dimensional problem and the majority of 5G applications are real-time ones that do not tolerate computational delays. Additionally, the hybrid beamforming inherits the same problems of the conventional beamforming such as the dependency of its performance on channel knowledge. Particularly, the highest gains of hybrid beamforming come at the cost of providing all terminals in the system with infinite precision and instantaneous channel state information (CSI) [4, 45, 82]. Reducing the required CSI for hybrid beamforming at the transmitter constitutes another major challenge to hybrid beamforming.

Typically, acquiring channel state information at terminals is a training-based estimation problem. In time division duplexing (TDD) systems, the channel state information at the re-

ceiver (CSIR) is obtained by letting the base station send training sequence (reference signals/ downlink pilots), known at the receiver beforehand, to the receiver and based on the received reference signals the receiver estimates the downlink channel using channel estimation algorithms. Similarly, exploiting the channel reciprocity in TDD systems, the channel state information at the transmitter (CSIT) is obtained by making use of the uplink pilots transmitted by the receiver(s). Estimating the mmWave channel in the forward and reverse direction for TDD systems is not as hard and costly as sub-$6$ GHz bands. This is mainly because the mmWave channels are sparse, thereby estimating fewer parameters, and consequently fewer pilots (reference signals) are needed. For instance, a $16 \times 64$ mmWave channel that has channel paths $P = 12$ and employs uniform linear arrays (ULA) at both sides has $36$ parameters to be estimated (one angle of arrival, departure, and channel gain for each propagation path) instead of $1024$ parameters required for massive MIMO channels operating in sub-$6$ GHz bands. Owing to the intrinsic sparsity of mmWave channels, many efficient CSI estimation algorithms have been developed based on compressed sensing techniques and have been shown to accurately estimate the channel with low training overhead [136, 149]. Therefore, the main challenge for $5$G TDD-based systems is to reduce the computational complexity of hybrid beamforming solutions.

On the other hand, in frequency division duplex (FDD) where uplink and downlink channels are widely separated in frequency, acquiring CSIT is a two-stage process. First, the base station sends downlink pilot signals. Then, the receiver estimates the downlink channel (similar to TDD in the forward direction) and feeds it back to the base station. This creates a huge data overhead in the reverse channel. Therefore, the CSIT acquisition is considered to be more challenging in FDD systems (the dominant system architecture of the current wireless cellular systems) [45]. This makes the main challenge for $5$G FDD systems is to develop low overhead and low complexity hybrid beamforming solutions.

### 1.4.1 Hybrid Beamforming for TDD systems

Real-time use-cases of 5G, including connected cars, virtual-reality/augmented-reality, and high definition video transmission, require high-capacity and low-latency wireless transmission [96].

On the physical layer level, this entails adopting near capacity-achieving transmission schemes with very low computational delay [96]. However, the majority of hybrid processing schemes in the literature invoke computationally exigent operations such as matrix decomposition [15, 57, 66, 84, 87, 90, 106, 111, 139, 145], matrix inversion [15, 120], gradient descent methods [90, 106, 111, 139] and matrix determinant calculation [34, 145]. Typically, these complex operations result from two design factors. First, designing the high-dimensional analog processing based on optimizing sophisticated performance metrics such as spectral efficiency [84, 120, 148], mutual information [15, 90, 106] and mean square error [66, 87]. Second, the overhead computations of the fully-digital precoder/combiner which is a prerequisite step for many hybrid processing schemes [15, 57, 66, 84, 87, 90, 106, 111, 139, 145], especially, the ones considered in the first approach. Therefore, adding these huge overhead computations to the complexity of optimizing such sophisticated metrics, intensifies the computational complexity of the hybrid processing problem and makes it impractical in high-dimensional applications such as massive MIMO communications.

### 1.4.2 Hybrid Beamforming for FDD Systems

Huge research efforts have been exerted to alleviate the channel feedback overhead by relaxing the perfect (infinite precision and instantaneous) CSIT assumption. In one direction, many works have adopted the statistical channel information, particularly the spatial channel covariance matrix, to design the analog precoder, and the perfect channel information to design the digital precoder (e.g., [2, 3, 52, 95] and references therein). This lessens the problem by reducing the required training and the corresponding feedback overhead since (i) the channel statistics of most wireless applications are slowly varying and (ii) the dimensions of the effective channel (the channel after the analog precoding) are significantly reduced. However, this direction requires optimistic conditions on the channel such as spatial user grouping, very small angular spread, uniformity of the covariance matrix across users and/or subcarriers, and the stationarity of the angles of arrival and departure in order to achieve near-optimal spectral efficiency performance. In another direction, the limited feedback hybrid precoding has been considered

where the channel or the analog precoder is selected from a predefined finite set of codewords (e.g., [8,10,15,21,35,44,68,77,115,129,144] and references therein). This reduces the feedback overhead dramatically where only a small number of bits are required to indicate the preferred codeword. However, existing research works in this approach invoke computational intensive beamforming algorithms that either perform exhaustive search [15] or require complex iterative processing such as coordinate descent algorithm [44], Tabu search [35] and cross-entropy optimization method [21], and utilize inefficient codebooks such as Hadamard codebooks [115] and fixed parts of DFT matrices [129].

## 1.5 Research Contributions

### 1.5.1 Hybrid Beamforming for TDD systems

To address the problem of high computational complexity of hybrid beamforming, we propose a unified low-complexity approach to design the hybrid precoder and combiner for both fully and partially-connected antenna array structures. This is in contrast to prior works in [15, 84, 90, 106, 111] which suit only the fully-connected structure, and prior works in [34, 66] which have been developed mainly for partially-connected ones. The proposed design approach is developed based on the channel diagonalization concept while taking the computational complexity and the RF hardware constraints into consideration. However, this approach diagonalizes the massive MIMO channel in two stages. Particularly, the analog precoder and combiner are jointly designed to maximize the antenna array gain (a low complexity metric) of the massive MIMO channel. This is in contrast to the majority of hybrid beamforming schemes where the analog precoder/combiner is either designed to iteratively maximize sophisticated performance metrics such as the spectral efficiency [34, 76, 90, 111, 120, 139, 148] and the MSE [57, 66] or developed to decompose the fully-digital precoder/combiner based on computationally intensive operations such as singular-value decomposition (SVD) and gradient descent methods in the high-dimensional domain [15, 57, 66, 84, 87, 90, 106, 111, 139, 145]. The digital precoder and combiner are jointly designed to maximize the spectral efficiency thereby eliminating the

residual interference (by-product of the analog beamforming) between data streams. The optimal solutions are shown to diagonalize the effective channel and the noise covariance (due to the two stages of combining). This approach is developed in two algorithms that accommodate for the differences in the architecture of both RF structures. The contributions of this work are summarized as follows:

- For the fully-connected antenna array structure, first, we derive a heuristic upper bound on the antenna array gain of the massive MIMO channel. Based on this bound, we develop a non-iterative hybrid processing algorithm that exploits the large number of antenna elements at both sides aiming at hardening the diagonal elements of the effective channel after the analog processing. Second, in Proposition 3.2.1, we calculate the antenna array gain due to the analog processing where it is shown to increase unboundedly with the number of antenna elements . Moreover, we show in Corollary 3.2.1.1 that the ratio of the means of the off-diagonal elements of the effective channel to the means of the diagonal ones goes to zeros as the number of antenna elements grows unboundedly. This ensures that the proposed scheme maintains high spectral efficiency at high SNR.

- For the partially-connected antenna array structure, the massive MIMO channel can be divided into low-dimensional sub-channels, i.e., it has a virtual block structure. First, we exploit this virtual block structure to formalize the high-dimensional multi-stream hybrid processing problem as a set of independent single-stream low-dimensional sub-problems which aim at maximizing the antenna array gains of the sub-channels on the diagonal. Second, we develop a low-complexity hybrid processing scheme that utilizes a simple alternating optimization technique to iteratively maximize the antenna array gains of those sub-channels. Thirdly, in order to evaluate the efficacy of the proposed iterative procedure, we derive upper and lower bounds that tightly bound the antenna array gain of the proposed scheme in Proposition 3.2.3.

- We extend the proposed hybrid processing approach to frequency-selective massive MIMO-OFDM systems. We drive a lower bound on the sum of antenna array gains across all

subcarriers. Based on this bound, we show that the analog processing problem for the frequency-selective channels can be converted into a frequency-flat hybrid processing one, and thereby, utilizing the proposed schemes for the latter one with minor modifications.

- We also extend the proposed hybrid processing approach to multi-user broadcast channels with various RF structures.
- We provide a comprehensive computational complexity analysis for the proposed schemes and the prominent prior works [15, 34, 66, 90, 106, 145, 148]. Objective computational complexity comparisons have been made in Tables 3.2 and 3.3. Although, we consider the computational complexity of the SVD algorithm in [43] (which is in favor of the prior works), the proposed schemes have a significant complexity saving compared to prior works as depicted in Figs.3.9 and 3.10.
- Extensive numerical results are presented which verify the near-optimal spectral efficiency performance of the proposed schemes in both rich and sparse scattering environments. Moreover, it is shown numerically that the proposed schemes are as robust as the optimal fully-digital ones against channel estimation/quantization errors. This shows that the proposed schemes have competitive performance under imperfect channel state information assumptions.

### 1.5.2 Hybrid Beamforming for FDD systems

To address these shortcomings of hybrid beamforming prior works for FDD systems, we propose a novel approach to design the hybrid precoder based on leveraging the second-order statistics and propagation properties of the mmWave channel aiming mainly at decreasing the feedback overhead in FDD systems. The proposed approach is embodied in two main hybrid precoding schemes, i.e., Algorithm 3 and Algorithm 4, developed for single and multi-user cases, respectively. Moreover, we provide different variants of Algorithm 3 and Algorithm 4 which lend themselves to various channel knowledge frameworks such as limited feedback channels without statistical information and mixed-CSIT.

The contributions of this work compared to the prior art are summarized as follows:

- For the single-user case, we present a simple hybrid precoding design which is based on minimizing the chordal distance between the fully-digital precoder and the hybrid one. In contrast to the exhaustive search [15], complex gradient descent-based algorithms [21, 35, 44] and iterative greedy algorithms [10, 68, 144] developed in prior works, we propose low complexity and non-iterative hybrid beamforming algorithms (Algorithm 3, and its variants). Moreover, opposed to the channel independent (fixed) codebooks such as beamsteering and Hadamard codebooks used in the majority of prior art [10, 21, 44, 68, 77, 115], we utilize skewed DFT codebooks that vary with the channel statistics. This is devised in order to finely quantize the local neighborhood around the statistically preferred directions of the dominant eigenvectors and thereby enhancing the spectral efficiency performance.

- For the multi-user case, we present a hybrid precoding design that is based on maximizing the signal to interference plus noise ratio of each user aiming at maximizing the sum-rate of the network. In contrast to the prior works [2, 3, 8, 44, 52, 77, 95], the analog precoder is designed in a distributive manner that lends itself to the distributed nature of multi-user networks. Another distinguishing feature of Algorithm 4 from prior works in [8, 21, 44, 115] is that Algorithm 4 can be applied to the case when the number of assigned users is less than that of the RF chains. However, the algorithms in [8, 21, 44, 115] are applicable only when the number of users equals to the number of RF chains where they assign each user to one RF chains.

- We derive lower and upper bounds on the average mutual information for the DFT-based hybrid precoding. The proposed bounds suggest selecting DFT codewords that are aligned to the directions of the largest eigenvectors of the transmit antenna array response matrix (or the covariance matrix, depending on the type of the channel knowledge availability). Leveraging the properties of the mmWave channels, we demonstrate in Corollary 4.4.3.1 that the proposed schemes are asymptotically optimal. Numerical results validate the near-optimal spectral efficiency performance of the proposed schemes and their superiority over prior works.

## 1.6 Thesis Outline

The rest of this thesis is organized as follows. Chapter 2 presents a literature survey on hybrid beamforming. First, in Section 2.1, we list out the main differences between analog, digital and hybrid beamforming. Second, in Section 2.2, we provide an overview on the different RF structures of hybrid beamforming and we survey the different hybrid beamforming approaches for TDD and FDD systems.

In Chapter 3, Section 3.1 presents the system model of single-user mmWave massive MIMO TDD systems with partially-connected and fully-connected hybrid RF structures. Section 3.2 addresses the problem formulation and the proposed hybrid beamforming solutions. In Section 3.3 we extend the proposed solution to work in frequency selective channels. In Sections 3.4, we investigate different extensions of the proposed solutions to various multi-user broadcast setups. In Section 3.5, we provide a thorough computational complexity analysis and comparisons of the proposed solutions and prior works. Finally, Simulation results are presented in Section 3.6.

In Chapter 4, Section 4.1 describes the system model of the single user mmWave massive MIMO systems under different CSIT assumptions. Section 4.2 presents the problem description and the proposed algorithms. Section 4.3 presents the proposed algorithms for the K-user multi-user channel. In Section 4.4, we provide mutual information analysis that bounds the performance of the proposed schemes. Finally, Section 4.5 presents some discussions on the proposed framework, followed by simulation results, to investigate and corroborate the competitive performance of the proposed work.

Finally, concluding remarks are drawn and future research directions are discussed in Chapter 5.

# Chapter 2

# Literature Review

In this chapter, we provide a concise yet comprehensive survey on the advancements in hybrid beamforming over mmWave bands. We recognize the two influential drivers of the majority of hybrid beamforming works, namely, the RF architecture and the availability of the channel knowledge at the transmitter. Accordingly, we provide two taxonomies correspondingly. Furthermore, we point out the advantages and disadvantages of the different channel availability models and RF architectures. Before proceeding to the hybrid beamforming literature, we briefly review the conventional beamforming techniques and state their main differences from the hybrid beamforming.

## 2.1 Conventional Beamforming

The conventional digital precoder/combiner is performed in the baseband domain before the digital to analog converter at the transmitter side or after the analog to digital converter at the receiver side using digital signal processors. This allows for complete control over the magnitudes and the phases of the entries of the precoding/combining matrices. As a result, fully-digital beamforming is the embodiment of most multi-stream single-user and multi-user precoding and combining techniques. However, the hardware complexity of the fully-digital precoding and

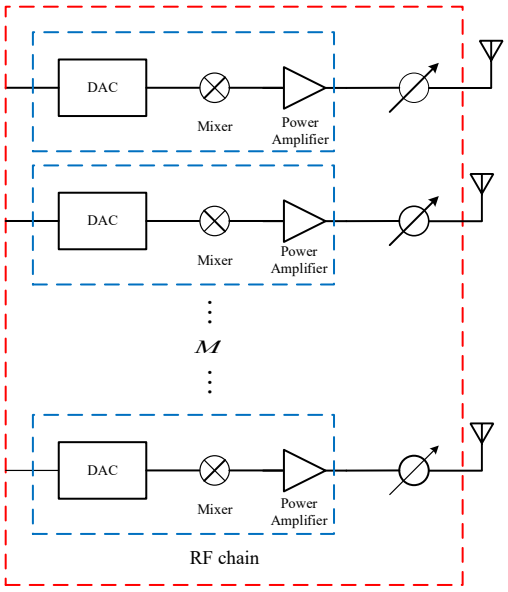


Figure 2.1: Fully-digital beamforming architecture

combining grows linearly with the number of RF chains. These RF chains include expensive and power-hungry mixed-signal devices such as ADCs and DACs which dramatically increase the implementation cost of fully-digital beamforming in mmWave systems. A schematic diagram of the fully-digital precoder is depicted in Fig. 2.1. Moreover, the fully-digital beamforming requires perfect knowledge of all the channel matrix's entries in order to calculate the optimal beamforming weights that create favorable propagation conditions [59, 74]. Throughout different parts of the thesis, we review the most prominent fully-digital beamforming schemes such as eigen-beamforming, conjugate and zero-forcing beamforming, channel block diagonalization one [126, 143].

On the other hand, the analog precoder/combiner is subjected to different hardware constraints such as quantized angles (phases) and constant gains (magnitudes) of the beamforming weights. These weights are designed to shape and direct the transmit or receive beams towards the dominant propagation directions [9]. However, those constraints hinder the ability of analog beamforming to cast multiple beams where the inter-beam interference is the limiting factor due

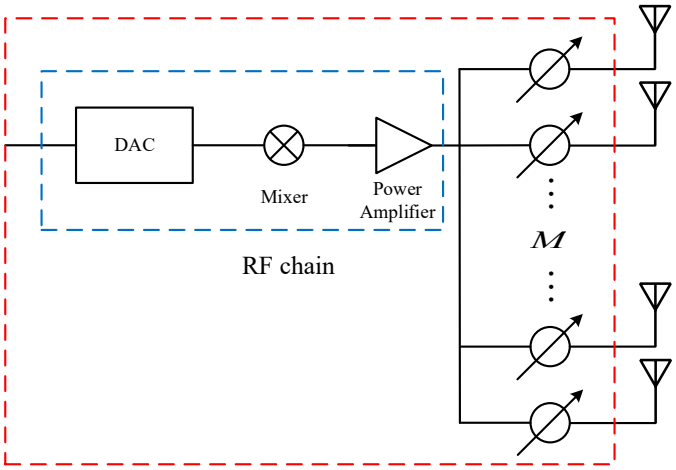


Figure 2.2: Analog beamforming architecture

to the wide sidelobes and overlapped nulls. As a result, the usage of analog beamforming is limited to single-stream transmission or beam steering applications such as sector level sweep, beam refinement, and beam tracking [56]. We note here that the complexity of the analog beamforming increases with the number of utilized phase shifters. A schematic diagram of the analog precoder is depicted in Fig. 2.2.

## 2.2 Hybrid Beamforming

To overcome the expensive hardware costs and complexity of the fully-digital beamforming and the shortcomings of the analog one, the hybrid beamforming has been introduced as an effective solution to this performance-cost (structure complexity) trade-off, physically, a compromise between the number of the utilized RF chains and phase shifters. The hybrid beamforming consists of a low-dimensional digital beamformer, where the number of RF chains scales linearly with the number of transmitted streams, followed by a high-dimensional analog beamformer, where the number of phase shifters scales linearly with the number of antenna elements. The hybrid precoding starts by digitally processing $N_s$ data streams at the baseband using the digital pre-

coder, then it continues by up-converting these signals to the carrier frequency using $M_{rf}$ RF chains, where each RF chain consists of a digital-to-analog converter, mixer, and power amplifier; finally, these RF signals are processed using a network of phase shifters before being on-air through an $M$-element antenna array. This system is of interest to massive mmWave applications when $N_s \leq M_{rf} \ll M$. Due to this combined structure, the hybrid beamformer reaps the benefits of both analog and digital beamforming where it supports multiple beams/streams with much reduction in the operational cost. A generic block diagram of a hybrid beamforming transmitter is shown in Fig. 2.3. The interconnections (wiring) between the RF chains and the antenna array ports through the network of phase shifters can have different forms that cast various RF structures. The hybrid beamforming has two main RF structures, namely, full-complexity (fully-connected/full-array) and reduced-complexity (partially-connected/sub-array). These two structures represent two extreme points among the various RF structures of hybrid beamforming wherein the former each RF chain is connected to all antenna elements whereas in the latter each RF chain is connected to disjoint groups of antenna elements. Between these two extreme structures, there are a lot of different RF connections such as overlapped sub-array, dynamic sub-array, sub-array with switches. This gives the hybrid beamforming higher degrees of freedom than fully-digital and analog beamforming in trading performance for hardware complexity and vice versa.

### 2.2.1 Taxonomy Based on The RF Architecture

#### Fully-connected RF structure

In the fully-connected structure, each RF chain is connected to all antenna elements' ports through a network of phase shifters which require a total of $N_{rf}M$ phase shifters and RF paths, and $N_{rf}$ $M$-branch power splitters and combiners. This results in high power consumption (in comparison to other RF structures) due to a large number of phase shifters and insertion losses of these elements in addition to the increase of the mutual coupling between RF paths. A schematic diagram of the fully-connected RF structure is depicted in Fig. 2.4. However, it has been shown

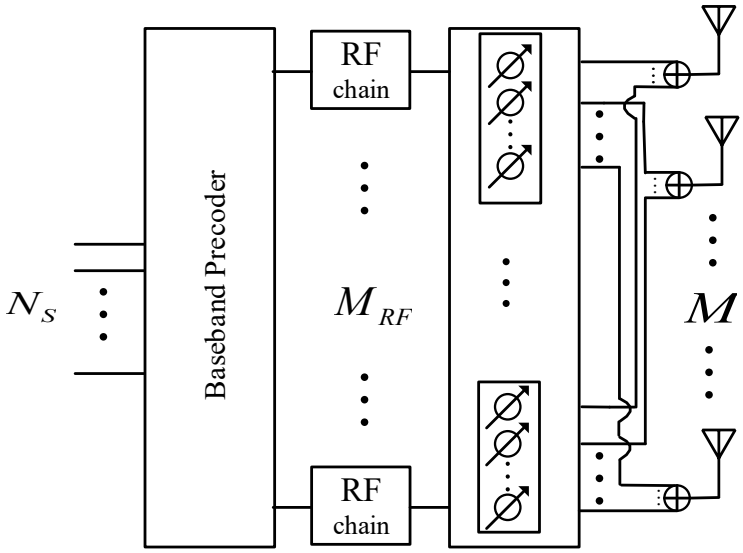


Figure 2.3: Hybrid beamforming architecture

that the fully-connected structure has the highest spectral efficiency performance since it offers the highest beamforming gain among other RF structures [8, 15, 16, 44, 52, 57, 66, 76, 84, 87, 90, 95, 106, 111, 120, 139, 145, 148]. Moreover, it has been shown that the fully-connected structure achieves the fully-digital performance when the number of RF chains as many as twice the number of the transmitted streams [120, 137]. Further, in sparse environments such as mmWave, while utilizing compressed sensing algorithms, it achieves a near-optimal spectral efficiency performance with a very fewer number of RF chains [15, 111].

**Partially-connected RF structure**

The partially-connected RF structure has been introduced aiming at enhancing the energy efficiency, reducing hardware complexity, and lowering the computational complexity of the hybrid beamforming problems. In the partially-connected structure, RF chains are connected to disjoint subsets of antenna elements, specifically, each RF chain is connected to $M_{sa} = M/M_{rf}$ [16, 25, 34, 48, 76, 139, 145, 148]. Compared to the fully-connected, this decreases the number of

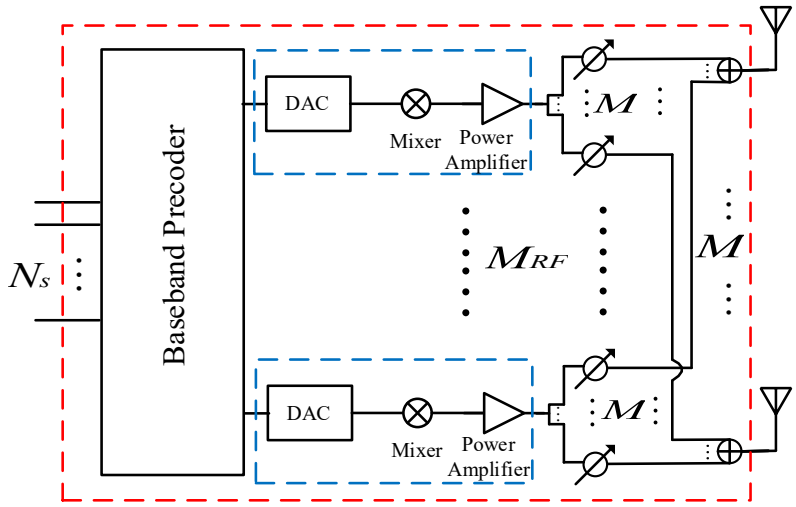


Figure 2.4: Fully-connected RF architecture

phase shifters from $MM_{rf}$ to $M$ and the number of RF combiners from $M$ to $0$ while the number of RF splitters remains constant but with $M_{rf}$ branches instead of $M$. A schematic diagram of the partially-connected RF structure is depicted in Fig. 2.5. In this way, the partially-connected structure reduces power consumption and this in certain cases leads to enhancing the energy efficiency of the system [34]. Moreover, this RF connection gives the analog precoder/ combiner a block diagonal structure, which, in turn, is utilized to decrease the computational complexity of the hybrid beamforming problem by virtually dividing the massive MIMO channel into sub-channels with smaller dimensions [34, 148]. Mainly for these two reasons and since the main motive behind introducing hybrid beamforming to massive systems is to reduce the installation cost and power consumption, the partially-connected hybrid precoding/combining structure is considered to be the practical setup. Unfortunately, this practicality comes with a significant spectral efficiency performance loss. For instance, given the same numbers of RF chains and antenna elements, it has been conjectured that the spectral efficiency gap is $N_{rf} \log_2 N_{rf}$ b/sec/Hz since the beamforming gain is decreased by $\frac{1}{N_{rf}}$ per sub-array. However, this performance loss can be compensated by either increasing the number of antenna elements per sub-array [82].

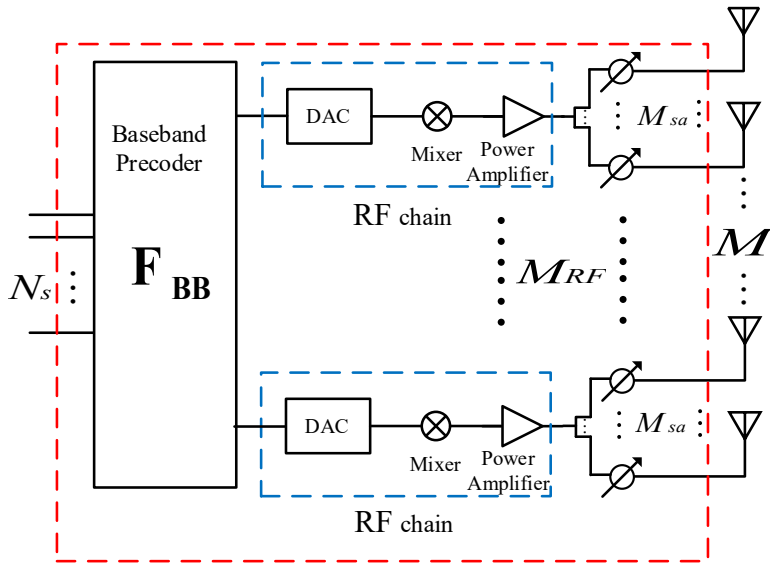


Figure 2.5: Partially-connected RF architecture

**Overlapped and Dynamic RF structures**

Between these two structures, numerous RF structures have been proposed to offer a better compromise between the hardware complexity/energy efficiency and spectral efficiency. The most successful ones are the overlapped sub-array and dynamic sub-array structures. The overlapped sub-array structure has been proposed in order to compensate for the beamforming gain reduction associated with the sub-array structure [30, 123, 134, 142]. This is realized by creating an overlap between adjacent sub-arrays and this, in turn, allows for two or more antenna elements to be connected to one RF chain. A larger number of antennas connected to one RF chain result in a higher beamforming gain but with an increased hardware complexity in terms of phase shifters. However, there are no rigorous guidelines to quantify how much overlap should be created to achieve spectral efficiency performance comparable to the fully-connected structure. Moreover, this overlapping between sub-arrays destroys the block diagonal structure of the analog precoder/combiner and consequently increases the computational complexity of the hybrid beamforming problem compared to the partially-connected structure.

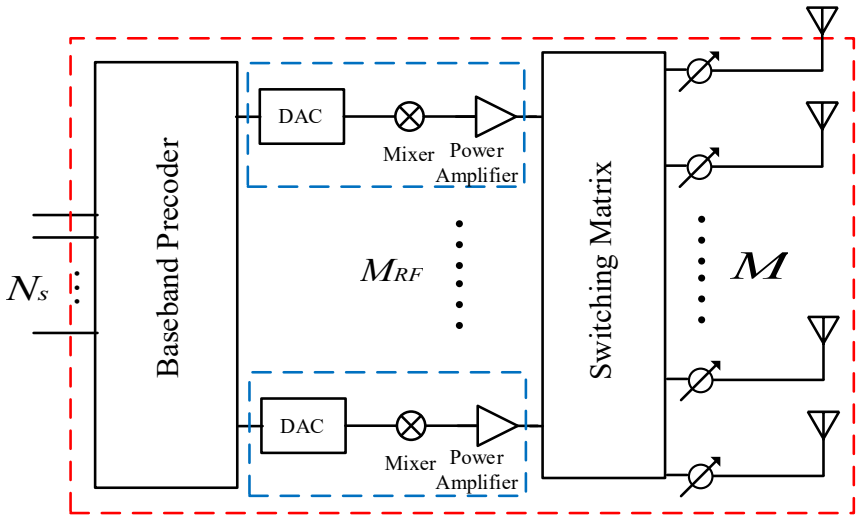


Figure 2.6: Dynamic subarray RF architecture

In the dynamic sub-array structure, each RF chain is connected to all antenna elements through a network of $MN_{rf}$ switches and phase shifters. However, at any given moment, each RF chain is connected to only $L \ll M$ antenna elements where only $L$ switches are closed and the others are opened, i.e., selecting $L$ switches and phase shifters per RF chain [52, 53, 55, 92, 153]. A schematic diagram of the dynamic subarray RF structure is depicted in Fig. 2.6. That is, the dynamic sub-array structure has higher hardware complexity than the fully-connected structure and similar energy efficiency as the partially-connected one. The dynamic sub-array structure achieves near-optimal performance using the exhaustive search antenna partitioning. Many research works have considered the greedy *soft* antenna selecting algorithms in order to decrease the computational complexity of the exhaustive search while maintaining comparable spectral efficiency performance to the fully-connected structure [92]. To sum it up, the dynamic sub-array structure achieves a near-optimal tradeoff between spectral and energy efficiency at the price of higher hardware and computational complexity.

In addition to the aforementioned RF structures, there are various ones that use RF switches instead of phases shifters [12, 99]. We excluded the discussion of all the hybrid beamforming

schemes where the phase shifters are replaced by switchers since they are basically either pure antenna selection schemes or binary phase quantization that dramatically reduce the beamforming gain produced by the analog beamforming.

### 2.2.2 Taxonomy Based on The Availability of The Channel Knowledge at The Transmitter

The channel state information at the transmitter plays a key role in the literature on analog, digital, and hybrid beamforming. It has been shown that the availability of CSIT can significantly widen the degrees of freedom (DoF) region (maximum multiplexing gain) of many wireless networks [150]. Under the full CSIT assumption, where the transmitter has global and instantaneously perfect CSI, the wireless networks enjoy the widest DoF region. On the other hand, in the total lack of CSIT the DoF region of most wireless networks collapse to the narrowest region, where its corner points are simply achievable by time-division multiplexing between users [150].

#### Perfect channel knowledge

Extensive research works have demonstrated the effectiveness of the hybrid beamforming in achieving a near-optimal high spectral efficiency performance under the perfect CSIT assumption while considering the aforementioned RF hardware constraints. This is encouraged by the interesting properties of massive MIMO and mmWave channels which make channel estimation simpler than at sub-6 GHz. Particularly, in TDD systems, the channel can be efficiently estimated at both sides with much less information overheads, compared to the rich scattering environment, utilizing its angular sparsity, spatial correlation, and reciprocity [79, 152]. Further, in many indoor applications where the surrounding environment is relatively static, the communication channel is plausible to be assumed a quasi-static, i.e., the channel remains constant for a long period compared to the transmission time. In this case, the instantaneous availability of CSIT can be realized in real life. Moreover, this implies that the performance of hybrid processing solutions with imperfect channel knowledge is very close to the ones assuming perfect channel

knowledge. As a result, the perfect CSIT assumption, in some applications and communication scenarios can be considered to be a practical assumption.

The prior art of hybrid beamforming with perfect CSIT can be summarized into two main approaches. The first, which includes the majority of hybrid beamforming schemes, is to design the analog and the digital beamformers based on approximating the optimal fully-digital beamformer by the hybrid ones [15, 57, 66, 84, 87, 90, 106, 111, 139, 145]. In essence, the analog and the digital processing matrices are designed such that their matrix product is nearly equal to the optimal fully-digital processing matrix in accordance with a certain measure such as the Euclidean distance or performance metric such as mutual information. This approach was developed in [15] where Ayach *et al.* exploited the sparsity of the mmWave channel to formalize the spectral efficiency optimization of the hybrid processing as a sparse reconstruction problem and solve it algorithmically based on an efficient compressed sensing algorithm, namely, the orthogonal matching pursuit (OMP) [15]. Utilizing the richness of the compressed sensing literature, several sparse reconstruction algorithms have been adopted, depending on the sparsity of the mmWave channel, in order to decrease the complexity of solving the optimization problem of approximating the fully-digital beamformer while considering different performance metrics such as spectral efficiency and minimum mean square error (MSE) [15, 57, 87, 111]. However, the performance of all these schemes degrades severely in rich scattering environments.

Considering the same approach of approximating the fully-digital solutions which were rebranded into constrained matrix decomposition or factorization, numerous alternating optimization techniques have been employed aiming at enhancing the performance of hybrid processing in both rich and sparse scattering environments [66, 90, 106, 139, 145]. Particularly, in [145], for sub-array antenna structures, the analog precoding problem, given the digital precoder, is formulated as a vector approximation problem using phase rotation whereas the digital precoding problem, given the analog precoder, is formulated as the Euclidean distance minimization problem with the second norm equality constraint and solved based on semi-definite relaxation. Similarly, for fully-connected antenna structures, given the analog precoder, the digital precoder is obtained as the solution of the orthogonal Procrustes matrix approximation problem [41], whereas the analog precoder, given the digital precoder, is obtained by extracting the phases of the least square

solution. In [78], the power factorization technique has been adopted aiming to trade-off spectral efficiency for decreasing the computational complexity of calculating the digital precoder in [145]. Leveraging the first-order Taylor expansion to the complex exponential function of the phase increment of the entries of the analog beamformer, Ni *et al.* showed that the optimal precoder/combiner matrix can be efficiently decomposed into the hybrid processing matrices [90]. Specifically, Ni *et al.* managed to relax the non-convex constraints of the analog processing matrix and thereby, given the digital processing matrix, finding the analog processing matrix as the solution of a convex quadratic programming problem; whereas, given the analog processing matrix, the digital processing matrix is the least-squares solution. In [106], Rajashekar *et al.* obtained the analog and the digital processing matrices alternately and iteratively as the phases of the entries of the optimal processing matrix and the least square solution, respectively. However, most of these alternating optimization techniques result in computationally complicated gradient-based iterative schemes.

On the other hand, the second approach is to decouple the design of analog and digital processing. Particularly, the analog processing is designed to optimize certain performance metrics. Then, the digital processing is designed based on the effective channel, i.e., the channel after the analog processing. For instance, in [16, 34, 76, 120, 148], the analog processing is designed to iteratively maximize the spectral efficiency or the mutual information of the system by successively canceling-out the inter-stream interference. This precoding strategy has been proposed by Pi in [102] for multi-stream analog precoding under per-antenna power constraints where he developed, based on deriving Karush-Khun-Tucker necessary and sufficient conditions for optimality, an iterative transmit beamforming algorithm which can securely converge to the optimal solutions. However, these optimal solutions do not adhere to the hardware constraints of the analog beamforming in massive MIMO systems. In [16], for the sub-array antenna structures, El Ayach *et al.* modified the algorithm in [102] to fit the hardware constraints of the analog precoder. In [34], Gao *et al.* integrated the mutual maximization algorithms into [16, 102] a successive interference cancellation strategy into calculate the hybrid precoding. They obtained the analog precoder based on the phases of the left singular vector of the augmented interference matrix while the digital precoder is a diagonal matrix where the diagonal elements are designed

to minimize the Euclidean distance between the optimal precoder and the analog precoder. Following the decoupling approach also, in [98], Payami *et al.* designed the analog processing stage based on extracting the phases of the largest $M_{rf}$ left and right singular vectors of the massive MIMO channel whereas the digital processing stage is based on the SVD of the effective channels. Moreover, in [84], Molu *et al.* designed the analog processing matrix to maximize the power of the desired signals while minimizing the inter-stream interference. This was realized by maximizing the correlation between the columns of the analog processing matrix and the optimal processing matrix. Between these two approaches, several heuristic approaches have been proposed such as the ones which are based on particle swarm optimization [11] and deep learning [62].

**Limited feedback and statistical CSI**

With the inevitable need for communicating in dynamic environments such as urban areas, the assumption of instantaneous and accurate CSIT becomes unreasonable and impractical. Acquiring such amount of information in dynamic environments while considering the signaling overhead excessively decreases the overall data rate available for users especially in FDD architecture which dominates most current wireless communication systems. The signaling overhead in FDD is attributed to the downlink training and uplink feedback. This implies that the signaling overhead in FDD systems is more severe than in TDD and might be prohibitive if no countermeasures are considered. The main solutions to this immoderate signaling overhead are either to limit the amount of information signaled to the transmitter through quantizing the feedback information in amplitude (finite precision) and/or make it less frequently updated by utilizing the long term channel statistics in beamforming. We note that the channel state information required for hybrid beamforming techniques in FDD systems can be divided into two pieces of information: the first one is required for the digital baseband processing and the second one is required for the analog RF processing. Huge research efforts have been made to alleviate the feedback overhead problem by relaxing the perfect CSIT assumption and can be categorized in the following directions.

In one direction, the limited feedback hybrid precoding has been considered where the channel is quantized or the analog precoder is selected from a predefined finite set of codewords [8, 10, 15, 44, 77, 122, 130, 133, 144]. The limited (finite precision) feedback approach is mainly a codebook-based system. In this system, the transmitter(s) and receiver (s) agree on a common codebook of precoding matrices or beamformers (codewords) designed according to certain criteria. The receiver, based on a certain metric (a function of the channel), selects the preferred codeword and sends its index back to the transmitter(s). Consequently, the design of an efficient limited feedback system is reduced to the design of codebooks that match the selection criteria in order to provide the minimum average distortion measured by certain metrics. This decreases the feedback overhead dramatically where only a small number of bits are required to indicate the preferred codeword. The first codebook introduced to the mmWave system was in [133]. In [133] the authors considered a single-user point to point MIMO channel where both the transmitter and receiver are equipped with an $M$-element uniform linear array antenna. The codebooks are fixed and designed for a phased antenna array implementing only specific four phase shifts per element $(0^\circ, 90^\circ, 180^\circ, 270^\circ)$ without amplitude adjustment. Each beamformer specifies a certain pattern (directions). The codebooks span the entire space, which is $360^\circ$ around the terminal. The codebooks are generated symmetrically, i.e., they are generated so that all beam patterns (directions) have the same gain. Also, the beamformers of the codebooks are orthogonal to each other, so that multiple beams can be generated simultaneously without large interference to each other. The codebook matrix of an $M$-element uniform linear phased array with $n$-bit phase resolution is defined by $W(M, K)$, where $K$ is the number of patterns (direction) generated. $w(m, k) = j^{\frac{((m-1)(k-1)-K/2)}{2^n/4}}$ is an entry of $W(M, K)$. One main limitation of this codebook (2-bit phase quantization) that when the array size increases, the power of side lobe radiation tends to rise drastically because it is optimally designed for a $4$-element array, and restricted to a maximum of $4$ symmetrical patterns. However, there is no guarantee that these codebooks optimize certain performance criteria. As an extension to [15], Alkhateeb *et al.*, in [10], considered the design of a hybrid beamforming transceiver for SU MIMO millimeter-wave channel with partial channel knowledge at the transmitter. The authors considered approximating the optimal precoder with hybrid precoders and finding the hybrid precoders

algorithmically using OMP but based on a fixed codebook generated by quantizing the transmit antenna array response uniformly over an angle of departure $\in [0, 2\pi)$. This used codebook is widely known as a beamsteering codebook. Unfortunately, the design of the beamsteering codebooks is subject to practical limitations: (i) it is very difficult to design non-overlapping beam patterns with the quantized phase shifters, and (ii) it is hard to apply the beam steering codebooks for non-ULAs due to the lack of intuition about their beam patterns in the design of analog-only beamforming. Other schemes in [77,144] are very similar to the ones in [10,68] expect the digital precoders of the former are designed to minimize the mean squared error while those of the latter are designed to maximize the spectral efficiency. Moreover, in [44] a codebook-based hybrid beamforming scheme was proposed where the analog precoder is selected from beamsteering codebooks using a MOSEK-based algorithm where its selection metric is based on maximizing the sum-rate. In [6] Alkhateeb *et al.* introduced a multi-level variant beam-width codebook-based hybrid beamforming scheme. The proposed codebook is a hierarchical one and consists of $s$ levels. Each level contains beamforming vectors with a certain beam-width that decreases with the number of levels $s$. For instance, if $s = 2$, the codebook consists of two levels: the first level consists of two beam patterns each having width of $\pi$ while the second level consists of four beam patterns each having a beamwidth of $\pi/2$. Despite the similarity of this codebook structure to the codebooks in [50, 133] which also have multiple levels, the authors in [6] adopted a different methodology for defining each beamforming vector and the associated beamwidth and a new technique for realizing these vectors using a hybrid architecture. The main advantage of this scheme is the additional digital processing layer that adds more degrees of freedom to the beamforming design problem which can be utilized to obtain better characteristics in the beamforming patterns. The codebook is designed such that the optimal fully-digital beamforming matrix is the one that minimizes the Euclidean distance between the desired direction and the antenna response to a certain channel realization. Then approximating the optimal fully-digital one to hybrid beamforming matrices by a modified version of the Orthogonal Matching Pursuit, or OMP algorithm [97]. In [122], the authors developed a codebook design tailored specifically to consider the strong directivity of the millimeter-wave channel and the hybrid beamforming architecture. The codebooks are designed to minimize the mean square error (MSE) between

the beamformer and the ideal beam pattern. Due to the sparse nature of the millimeter-wave channel, the MSE minimization problem is solved using the OMP algorithm. The ideal beam patterns have constant gain of $\frac{2^B}{M}$ with equal beamwidths $v_q = [-\pi + \frac{2\pi}{2^B}(q-1), -\pi + \frac{2\pi}{2^B}(q)]$ for $q \in \{1, \cdots, 2^B\}$. Once again the OMP has been used to find the analog and the baseband precoders but with MSE as a cost function instead of the Euclidean distance. However, the authors limit their analysis to the single-stream transmission and in this case, their codebook provides a higher data rate than the codebook in [6].

In another direction, many works have adopted the statistical channel information, particularly the spatial channel covariance matrix, to design the analog precoder while the perfect channel information to design the digital precoder [2, 3, 52, 63, 85, 95, 138]. This approach uses the fact that the different links often share the same fading distribution. In particular, statistical approaches are effective in situations where the channel has some form of structure, such as having a large mean component (e.g., a large Rician K-factor) or strong correlation (either in space, time, or frequency). In such channel structures, statistical beamforming provides a marginal performance loss compared to beamforming techniques that use the instantaneous channel realization over the long run. Indeed, the statistical knowledge is very effective in massive MIMO channel and mmWave bands due to the spatial correlation and small angular dispersion. Generally, in hybrid beamforming literature, the statistical CSIT is to assume that the transmitter has perfect knowledge about the statistical averages of the channel (most importantly the transmit covariance matrix) while perfect channel knowledge of the effective channel (the channel after the analog precoding). This lessens the problem by reducing the required training and the corresponding feedback overhead since (i) the channel statistics of most wireless applications are slowly varying and (ii) the dimension of the effective channel is significantly reduced.

In [2, 3, 85], the authors introduced a hybrid beamforming scheme, called joint spatial division multiplexing (JSDM), in order to reduce the overhead signaling by designing the analog RF beamformer based on the second-order channel statistics, i.e., the covariance matrices of the different users. They considered the broadcast channel with $K$ single-antenna users as a set-up for their scheme. They assumed that the channel is spatially correlated with known channels covariance matrices. The basic idea of JSDM is to partition the users into groups where users

with similar covariance matrices are huddled together in the same group while maintaining orthogonality between the groups. The orthogonality between groups is enforced by selecting the analog RF beamforming matrix for group $g$ to be in the null space of the augmented matrix of the eigenvectors of the covariance matrices of the other groups. In other words, the analog RF beamforming is designed to eliminate the inter-group interference by employing the well-known block diagonalization technique but based on the statistical average, i.e., the analog beamforming matrix is a function of the second-order statistics of the channel matrix. A necessary condition for completely eliminating inter-group interference is the span of the columns of the analog beamforming matrix is a subset of the span of the orthogonal complement of the augmented matrix of the eigenvectors of the covariance matrices of the other groups. On the other hand, the digital baseband precoding is designed to eliminate the inter-user interference between users in the same group using linear precoders, i.e., zero-forcing. Even though the latter stage requires perfect CSI, there is a significant reduction in the overall overhead signaling thanks to users partitioning in JSDM and per-group processing, i.e., users send local channel information about their groups. The main limitations of JSDM are the difficulty of maintaining orthogonality between a large number of groups and the complexity of users selection and or grouping algorithms [2].

To alleviate the shortcomings of JSDM, in [63], the authors considered a multi-user broadcast channel with $M$-antenna serving $K$ $N$-antenna users. The authors managed to relax the orthogonality conditions of JSDM by allowing a controlled level of interference between groups. The residual interference can be reduced effectively with the help of the multiple antennas available at each receiver. Similar to JSDM, the analog beamformer at the base station is designed to alleviate the interference between groups based on the second-order channel statistics by applying the well-known block diagonalization scheme but with relaxed orthogonality constraints. On the other hand, the analog combiners for each group of users are designed to reduce the residual inter-group interference (due to the relaxation of the orthogonality condition) by maximizing the received intra-group signal to inter-group interference plus noise ratio. The optimal analog combiner that maximizes the signal to interference plus noise ratio. It can be readily found by solving a generalized Rayleigh quotient problem. The digital combiners are designed to maximize the net rate of each user by using the MMSE detector. Finally, the digital precoders at the

base station are designed to maximize the conditional average sum data rate by formalizing the problem as a weighted average mean square error minimization and solve it algorithmically. It is worth mentioning that the JSDM scheme and some of its modifications in [63, 85, 138], do not consider the hardware constraints associated with analog beamforming.

In [67], the constant gain RF constraint has been considered in the hybrid beamforming design, and the user-grouping concept of JSDM is extended to a more general case where different numbers of RF chains are dynamically assigned to different groups. Although the hybrid processing techniques in [3, 67] are limited to the assumption that users can be divided into groups such that users in each group share the same covariance matrix, a general hybrid precoding technique is proposed without the users grouping in [64]. In [64], each analog precoding vector is constructed from each user's single dominant eigenvector of the covariance matrix. In contrast to JSDM where the analog precoder is designed to remove the inter-group interference, in [95], the analog precoder is designed to maximize the average of the signal to leakage plus noise ratio (SLNR). Particularly, the analog beamforming vector(s) of each user is given by extracting the phases of the solution(s) of a generalized Rayleigh quotient problem of its covariance matrix and other users' covariance matrices. On the other hand, the digital precoder is designed based on the effective channel after the analog precoding and it is given as the regularized zero-forcing solution.

# Chapter 3

# Hybrid Beamforming Schemes for TDD Systems

## 3.1 System and Channel Models

### 3.1.1 System Model

We consider a point-to-point massive MIMO system with $M$-transmit and $N$-receive antenna elements. The transmitter sends $N_s$ independent data streams $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ to the receiver where both are equipped with limited number of RF chains, i.e., $M_{rf} \ll M$ and $N_{rf} \ll N$, respectively. As a result, the hybrid beamforming structure is considered at both sides. This point-to-point massive MIMO model is of interest due to its wide-range applicability in many recent applications such as high definition video streaming, virtual/augmented-reality, connected cars and links between base stations [20]. In hybrid precoding structures, the data vector $\mathbf{s}$ is pre-processed by two different precoders at the transmitter, namely, the baseband (digital) precoder $\mathbf{F}_{BB} \in \mathbb{C}^{M_{rf} \times N_s}$ and the RF (analog) precoder $\mathbf{F}_{RF} \in \mathbb{U}^{M \times N_{rf}}$. As a result, the received signal vector $\mathbf{y} \in \mathbb{C}^{N \times 1}$ is

given by:

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{n}, \tag{3.1}$$

where $\rho$ is the average received signal power and $\mathbf{n} \in \mathbb{C}^{N\times 1}$ is the additive white Gaussian noise vector at the receiver which has independent identically distributed (i.i.d.) elements with zero mean and variance $\sigma^2$. $\mathbf{H} \in \mathbb{C}^{N\times M}$ is the fading channel matrix. Similarly, the hybrid combining structure at the receiver is implemented by post-processing the received vector $\mathbf{r}$ by two different combiners, namely, the RF (analog) combiner $\mathbf{W}_{RF} \in \mathbb{U}^{N\times N_{rf}}$ and the baseband (digital) combiner $\mathbf{W}_{BB} \in \mathbb{C}^{N_{rf}\times N_s}$. Thus, the processed received signal $\tilde{\mathbf{s}} \in \mathbb{C}^{N_s\times 1}$ is given by:

$$\tilde{\mathbf{s}} = \sqrt{\rho}\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{n}, \tag{3.2}$$

We note here that both RF precoder and combiner are implemented by analog phase shifters with constant gain amplifiers, therefore, their entries have a constant norm, i.e., $|f_{RF,ij}| = \frac{1}{\sqrt{M}}$ and $|w_{RF,ij}| = \frac{1}{\sqrt{N}}$, respectively. On the other hand, both the baseband precoder and combiner are implemented in the digital domain and have only power constraints. The total power is normalized such that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \frac{1}{N_s}\mathbf{I}_{\mathbf{N_s}}$, $\| \mathbf{F}_{RF}\mathbf{F}_{BB} \|_F^2 = N_s$ and $\mathbb{E}[||\mathbf{H}||_F] = NM$.

Moreover, we consider two hybrid processing structures, namely, fully and partially-connected structures. The main difference between both structures is the architecture of the RF precoder and combiner. Particularly, in the fully-connected structure, each RF chain is connected to all antenna elements through a network of phase shifters and constant gain power amplifiers as depicted in Fig. 3.1. On the other hand, in the partially-connected one, RF chains are connected to disjoint subsets of antenna elements; specifically, each RF chain is connected to $M_{sa} = M/M_{rf}$ or $N_{sa} = N/N_{rf}$ antenna elements at the transmitter and receiver, respectively, as depicted in Fig. 3.2. Further, we assume, as in many exiting works (e.g., [15, 66, 84, 90, 106, 145]), perfect channel state information (CSI) is available at both the transmitter and receiver. In TDD systems, it has been shown that mmWave massive MIMO channels can be efficiently and accurately obtained at both sides with much less information overhead, compared to the Rayleigh channel

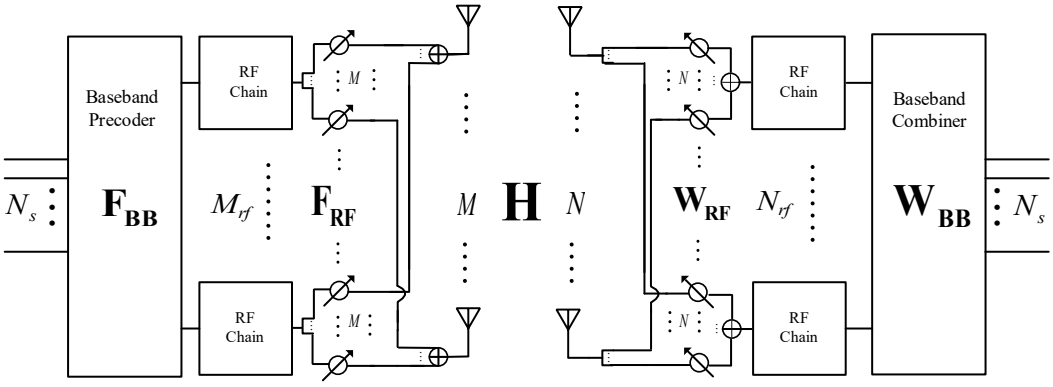


Figure 3.1: Hybrid precoding and combining with fully-connected antenna structure

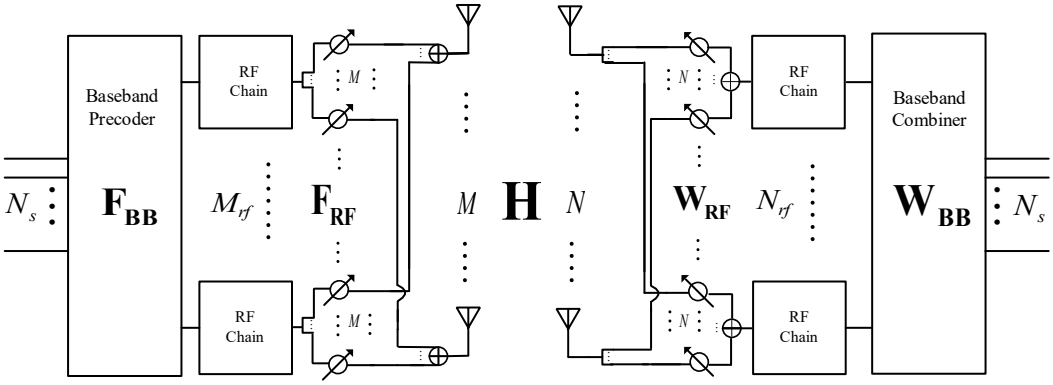


Figure 3.2: Hybrid precoding and combining with partially-connected antenna structure

(or sub-6 GHz), utilizing its sparsity, spatial correlation and reciprocity (e.g. [79, 152] and the references therein). On the other hand, in FDD systems, solutions are obtained at the receiver based on perfect CSIT and then they are quantized using efficient codebooks and are fedback to the transmitter [15, 116]. However, in this chapter we limit our discussion to TDD systems.

### 3.1.2 Channel Models

We consider two channel models that describe two relatively different fading environments, rich and sparse scattering environments. For the former, we adopt the i.i.d. Rayleigh channel model

where all the entries of the fading channel $h_{Rij} \backsim \mathcal{CN}(0,1),\ 1 \leq i \leq N$ and $1 \leq j \leq M$. For the latter, we consider the clustered channel model which is widely adopted in the literature of millimeter wave systems [15, 57, 66, 84, 87, 90, 106, 111, 139, 145]. The clustered channel $\mathbf{H}_{Cl}$ is given as a sum of all propagation paths that are scattered in $N_{cl}$ clusters with each cluster contributing $N_{ray}$ rays. As a result, the channel matrix is [15,57,66,84,87,90,106,111,139,145]:

$$\mathbf{H}_{Cl} = \sqrt{\frac{MN}{N_{cl}N_{ray}}} \sum_{i,l}^{N_{cl}N_{ray}} \alpha_{il}\mathbf{a}_r(\theta_{il}^r)\mathbf{a}_t^H(\theta_{il}^t), \tag{3.3}$$

where $\alpha_{il}$ are i.i.d. $\mathcal{CN}(0,1)$ depicting the complex gains of the $l^{\text{th}}$ ray in the $i^{\text{th}}$ cluster, $\mathbf{a}_t(\theta_{il}^t)$ is the transmit antenna array response vector for a given angle of departure $\theta_{il}^t$, and $\mathbf{a}_r(\theta_{il}^r)$ is the receive antenna array response vector for a given angle of arrival $\theta_{il}^r$.

## 3.2 Problems' Formulation and Proposed Algorithms for the Single-user Case

Our main goal is to develop simple and unified approach for the design of the hybrid precoder $\mathbf{F}_{RF}\mathbf{F}_{BB}$ and combiner $\mathbf{W}_{RF}\mathbf{W}_{BB}$ for massive MIMO systems with different RF structures. Moreover, a major concern of this work lies in significantly reducing the computational complexity of obtaining those four matrices since most recent point-to-point applications are real-time ones that require low computational delays. Towards that goal while considering the computational complexity concern, we tackle the hybrid processing problem from a different perspective.

### 3.2.1 The Design Rationale

Considering the spectral efficiency as a performance metric, the hybrid beamforming problem is given by:

$$\begin{aligned} \max_{\mathbf{F}_{RF},\mathbf{F}_{BB},\mathbf{W}_{RF},\mathbf{W}_{BB}} \quad & \log\det(\mathbf{I}_{N_s} + \mathbf{R}_{nBB}^{-1}\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^H\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{W}_{RF}\mathbf{W}_{BB}), \\ \text{s.t.} \quad & |f_{RF_{ij}}| = 1 \ \ \forall i,j, \quad |w_{RF_{ul}}| = 1 \ \ \forall u,l, \\ & \text{trace}(\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^H) \leq p \end{aligned} \tag{3.4}$$

where $\mathbf{R}_{nBB} = \sigma_n^2\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{W}_{RF}\mathbf{W}_{BB}$ is the noise covariance matrix after at the baseband after the two-stage hybrid combining. Finding the optimal solutions of this problem is extremely challenging and intractable for different RF structures due to mainly the non-convex RF hardware constraints in (3.4) [15, 34, 90, 106, 120, 145, 148]. Without considering these RF constraints, the optimal solutions come readily as $\mathbf{F}_{RF\star} = \bar{\mathbf{V}}_{N_s}$, $\mathbf{F}_{BB\star} = \mathbf{\Gamma}$ and $\mathbf{W}_{RF\star} = \bar{\mathbf{U}}_{N_s}$, and $\mathbf{W}_{BB\star} = \mathbf{I}_{N_s}$, where $\bar{\mathbf{V}}_{N_s}$, and $\bar{\mathbf{U}}_{N_s}$ are the $N_s$ right and left singular vectors associated with the largest singular values of $\mathbf{H}$ and $\mathbf{\Gamma}$ is a diagonal matrix that contains the power fraction of each stream on its diagonal. This has been established the in traditional (small-scale) MIMO systems literature where it has been shown that optimal linear precoding and combining strategies, which optimize different spectral efficiency and mean square error, are based on diagonalizing the channel (converting the MIMO channel into a set of parallel SISO channels) by removing the off-diagonal elements of the MIMO channel using different channel decomposition techniques such QR decomposition and SVD [114]. However, the high dimensionality of the mmWave massive MIMO systems and the hardware constraints on the analog precoder and combiner make the computational burden of these traditional techniques impractical, thereby soaring the computational complexity of their hybrid beamforming approximations [15, 57, 66, 84, 87, 90, 106, 111, 139, 145].

Building upon the channel diagonalization concept while taking seriously the computational complexity into consideration, we propose to solve the problem in (3.4) by diagonalizing $\mathbf{H}$ over two successive stages that account for the differences between analog and digital beamforming.

Particularly, in the analog processing stage, the massive MIMO channel matrix is encapsulated by the analog precoding and combining matrices aiming to exploit the huge antenna array gain by co-phasing the channel vectors associated with the RF chains. Thanks to the large number of antenna elements and the channel co-phasing techniques (developed in Algorithms 1 and 2) at both sides, the diagonal elements of the channel matrix after the analog processing stage (referred to as the effective channel) are hardened, i.e., "concentrated" around relatively large values. Hence, the main target of the analog processing is to create a big difference between the magnitudes of the diagonal and off-diagonal elements of the effective channel in a way to make it as close as possible to the diagonal channel. This ensures that the spectral efficiency loss due to the analog beamforming is as close as possible to zero. We note here that it is impossible to diagonalize the channel completely due to the constant magnitude constraint on the entries of the analog processing and the limited number of RF chains [15, 120]. We show in Section IV that this stage has significantly low computational complexity compared to prior works. Then, in the second stage, the digital precoder and combiner are jointly designed to cancel out the residual interference between data streams by completely removing the off-diagonal elements of the effective channel and allocate the streams' powers in order to maximize the spectral efficiency of the system.

### 3.2.2 Analog Processing

#### Analog processing for fully-connected structures

Abstracting the digital processing, the effective channel after the analog precoding and combining, $\tilde{\mathbf{H}} = \mathbf{W}_{RF}^{H}\mathbf{H}\mathbf{F}_{RF}$ can be written as:

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{w}_1^H\mathbf{H}\mathbf{f}_1 & \mathbf{w}_1^H\mathbf{H}\mathbf{f}_2 & \cdots & \mathbf{w}_1^H\mathbf{H}\mathbf{f}_{N_{rf}} \\ \mathbf{w}_2^H\mathbf{H}\mathbf{f}_1 & \mathbf{w}_2^H\mathbf{H}\mathbf{f}_2 & \cdots & \mathbf{w}_2^H\mathbf{H}\mathbf{f}_{N_{rf}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{w}_{N_{rf}}^H\mathbf{H}\mathbf{f}_1 & \mathbf{w}_{N_{rf}}^H\mathbf{H}\mathbf{f}_2 & \cdots & \mathbf{w}_{N_{rf}}^H\mathbf{H}\mathbf{f}_{M_{rf}} \end{bmatrix}, \tag{3.5}$$

where $\mathbf{f}_i$ and $\mathbf{w}_j$ are the $i^{\text{th}}$ and $j^{\text{th}}$ columns of $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$, respectively, and we drop the subscript RF here for simplicity of exposition. Since we aim at diagonalizing the massive MIMO channel over two stages, this stage are concerned with maximizing the squared magnitudes of the diagonal elements of the effective channel or equivalently maximizing the antenna array gain of the massive MIMO channel provided by the analog processing stage. The antenna array gain of MIMO is defined as the gain in the average received signal power due to the precoding/combining strategy, however, in multi-stream schemes, it is the sum of the average received power of the *independent* streams. Accordingly, assuming that $N_{rf} = M_{rf}$, the joint analog precoding-combining design problem is formalized as:

$$\begin{aligned} \underset{\mathbf{F}_{RF},\mathbf{W}_{RF}}{\text{maximize}} \quad & \sum_{n=1}^{N_{rf}} |\mathbf{w}_n^H \mathbf{H} \mathbf{f}_n|^2 \\ \text{subject to} \quad & |f_{ij}| = \frac{1}{\sqrt{M}} \;\; \forall i,j, \quad \text{rank}(\mathbf{F}_{RF}) = N_{rf} \\ & |w_{ul}| = \frac{1}{\sqrt{N}} \;\; \forall u,l, \quad \text{rank}(\mathbf{W}_{RF}) = N_{rf}, \end{aligned} \tag{3.6}$$

We note here that the rank constraints is defined, in engineering sense, such that the mathematical rank is satisfied with low condition number, i.e., well-conditioned matrix in order to provide the required multiplexing gain [132]. This problem is a non-convex optimization problem where securing global optimal solutions is extremely difficult. This is mainly due to the constant magnitudes constraints of the analog beamforming and rank constraints [15, 57, 66, 84, 87, 90, 106, 111, 139, 145]. Accordingly, instead of maximizing the cost function in (3.6), we resort to maximizing an upper bound on (3.6) in order to find near-optimal closed-form solutions of the precoder and combiner. First, we start by rewriting the cost function in (3.6) in terms of the entries of $\mathbf{W}_{RF}, \mathbf{H}$ and $\mathbf{F}_{RF}$ as $\sum_{n=1}^{N_{rf}} |\mathbf{w}_n^H \mathbf{H} \mathbf{f}_n|^2 = \sum_{n=1}^{N_{rf}} |\sum_{i=1}^{N}\sum_{j=1}^{M} w_{in}^* h_{ij} f_{jn}|^2$. This can be also rewritten as $\sum_{n}^{N_{rf}} |w_{nn}^* \sum_{j=1}^{M} h_{nj} f_{jn} + \sum_{i=1, i\neq n}^{N} w_{in}^* \sum_{j=1}^{M} h_{ij} f_{jn}|^2$ by extracting $n^{\text{th}}$ term of the first summation. Using Cauchy-Schwarz inequality, i.e., $|\sum_{i=1}^{q} u_i v_i^*|^2 \leq \sum_{j=1}^{q} |u_j|^2 \sum_{k=1}^{q} |v_k|^2$, and

putting $v = 1$, and $q = 2$, one gets:

$$\sum_{n=1}^{N_{rf}} |\mathbf{w}_n^H \mathbf{H}\mathbf{f}_n|^2 \leq \sum_{n=1}^{N_{rf}} 2(|w_{nn}^* \sum_{j=1}^{M} h_{nj} f_{jn}|^2 + |\sum_{i=1, i\neq n}^{N} w_{in}^* \sum_{j=1}^{M} h_{ij} f_{jn}|^2) \tag{3.7}$$

Based on this upper bound, we formalize the analog precoding and combining problem as:

$$\begin{aligned} \underset{\mathbf{F}_{RF}, \mathbf{W}_{RF}}{\text{maximize}} \quad & \sum_{n=1}^{N_{rf}} |w_{nn}^* \sum_{j=1}^{M} h_{nj} f_{jn}|^2 + \sum_{n=1}^{N_{rf}} |\sum_{i=1, i\neq n}^{N} w_{in}^* \sum_{j=1}^{M} h_{ij} f_{jn}|^2 \\ \text{subject to} \quad & |f_{ij}| = \frac{1}{\sqrt{M}} \quad \forall i, j, \quad \text{rank}(\mathbf{F}_{RF}) = N_{rf} \\ & |w_{ul}| = \frac{1}{\sqrt{N}} \quad \forall u, l, \quad \text{rank}(\mathbf{W}_{RF}) = N_{rf} \end{aligned} \tag{3.8}$$

Although this problem inherits the same issues of the one in (3.6), this formulation allows for finding sub-optimal solution by treating it as a multi-objective optimization problem [27]. This is achievable based on separately maximizing the first term in (3.8) from the second term and both have optimal closed-form solutions. In particular, the optimization problem defined in (3.8) can be divided into two different optimization problems that are solved successively aiming at securing closed-form sub-optimal solutions to (3.8). As a result, the first problem can be formulated as:

$$\begin{aligned} \underset{\mathbf{F}_{RF}, \mathbf{W}_{RF}}{\text{maximize}} \quad & \sum_{n=1}^{N_{rf}} |w_{nn}^* \sum_{j=1}^{M} h_{nj} f_{jn}|^2 \\ \text{subject to} \quad & f_{jn} = \frac{1}{\sqrt{M}} e^{j\theta_{jn}}, \quad 1 \leq j \leq M, \quad 1 \leq n \leq N_{rf}, \\ & w_{in} = \frac{1}{\sqrt{N}} e^{j\phi_{in}}, \quad 1 \leq i \leq N, \quad 1 \leq n \leq N_{rf}. \end{aligned} \tag{3.9}$$

where $\theta_{jn}$ and $\phi_{in} \in [0, 2\pi)$. Fortunately, the cost function in (3.9) is a sum of independent positive terms, and thereby, each term is maximized individually. The maximum values of these

terms are $\max_{\boldsymbol{\theta}_n,\boldsymbol{\phi}_n}(\frac{|e^{j\phi_{nn}}\sum_{j=1}^{M}h_{nj}e^{j\theta_{jn}}|}{\sqrt{MN}}) = \frac{\sum_{j=1}^{M}|h_{nj}|}{\sqrt{MN}}, 1 \leq n \leq N_{rf}$. This comes readily by setting $\theta_{jn} = \angle(h_{nj}^*), 1 \leq j \leq M$ and $1 \leq n \leq N_{rf}$ while setting $\phi_{nn}$ to any arbitrary phase since $|ab| = |a||b|$ and $|w_{nn}| = 1$. Although the rank constraint on $\mathbf{F}_{RF}$ has been relaxed in (3.9), its optimal solution, $\mathbf{F}_{RF\star}$ almost surely has full rank (satisfy the rank constraint). That is, the $M \times N_{rf}$ optimal phases $\theta_{ij_\star}, \forall, i, j$, which maximize (3.9), are extracted from the $M \times N$ i.i.d. randomly distributed entries of $\mathbf{H}$ thereby rending $\mathbf{F}_{RF}$ full rank. On the other hand, the second problem is:

$$\begin{aligned} \underset{\mathbf{F}_{RF},\mathbf{W}_{RF}}{\text{maximize}} \quad & \sum_{n=1}^{N_{rf}} |\sum_{i=1,i\neq n}^{N} w_{in}^* \sum_{j=1}^{M} h_{ij}f_{jn}|^2 \\ \text{subject to} \quad & f_{jn} = \frac{1}{\sqrt{M}}e^{j\theta_{jn}}, \quad 1 \leq j \leq M, \quad 1 \leq n \leq N_{rf}, \\ & w_{in} = \frac{1}{\sqrt{N}}e^{j\phi_{in}}, \quad 1 \leq i \leq N, \quad 1 \leq n \leq N_{rf}. \end{aligned} \tag{3.10}$$

Given the solutions of (3.9), closed-form solutions of (3.10) come readily as $\phi_{in} = \angle(\sum_{j=1}^{M} h_{ij}f_{jn}), \forall i \neq n$. Similar to the solutions of (3.9), the solutions of (3.10), i.e., $\phi_{in_\star}, \forall i \neq n$ render $\mathbf{W}_{RF}$ full rank. The reason is that some of these phases are pair-wise independent and the other are uncorrelated. This is based on $\mathbb{E}\{e^{j(\psi_{ij}-\theta_{jn})}e^{j\psi_{ij}}\} = \mathbb{E}\{e^{j\psi_{ij}}\mathbb{E}\{e^{j(\psi_{ij}-\theta_{jn})}|e^{j\psi_{ij}}\}\} = 0$ where if $\theta_{ij}$ or $\psi_{ij}$ are uniformly distributed in $[0, 2\pi)$, then their sum modulo $2\pi$ is also uniformly distributed in this interval. These two steps are summarized in Algorithm 1.

**Proposition. 3.2.1.** *The antenna array gain per RF chain of Algorithm 1 increases unboundedly with $N$ where the mean and the variance of the effective channels' diagonal elements are asymtotically given as $(\sqrt{\frac{M\pi}{4N}} + \sqrt{\frac{N\pi}{4}})$ and $(\frac{4-\pi}{4} + \frac{4-\pi}{4N})$, respectively.*

**Algorithm 1** Maximizing Antenna Array Gain Per RF-chain

**Require:** $\mathbf{H}$

1: **while** $n \leq N_{rf}$

2: $f_{jn} = \frac{1}{\sqrt{M}} \frac{h_{nj}^*}{|h_{nj}|}, \quad \forall 1 \leq j \leq M$

3: $w_{in}^* = \frac{1}{\sqrt{N}} \frac{\sum_{j=1}^{M} h_{ij} f_{jn}}{|\sum_{j=1}^{M} h_{ij} f_{jn}|}, \quad 1 \leq i \leq N, \forall i \neq n$

4: $n = n + 1$

5: **end**

6: **return** $\mathbf{F}_{RF} = [\mathbf{f}_1, \cdots, \mathbf{f}_{N_{rf}}]$, $\mathbf{W}_{RF} = [\mathbf{w}_1, \cdots, \mathbf{w}_{N_{rf}}]$

*Proof.* The $n^{th}$ diagonal element of the effective channel is given by:

$$\tilde{h}_{nn} = \frac{1}{\sqrt{NM}} \sum_{i=1}^{N} e^{-j\angle(\sum_{p=1}^{M} h_{ip} e^{-j\angle h_{np}})} \sum_{j=1}^{M} h_{ij} e^{-j\angle h_{nj}} \tag{3.11}$$

$$= \frac{1}{\sqrt{NM}} \sum_{j=1}^{M} |h_{nj}| + \frac{1}{\sqrt{NM}} \sum_{i=1, i\neq n}^{N} |\sum_{j=1}^{M} h_{ij} e^{-j\angle h_{nj}}| \tag{3.12}$$

Using the Lindeberg-Lévy central limit theorem where, for i.i.d. complex Gaussian vector of length $M$ and distributed $\backsim \mathcal{CN}(0,1)$, $\sqrt{M}\left(\frac{1}{M}\angle \mathbf{h}_k{}^H \mathbf{h}_k - \frac{\pi}{4}\right) \xrightarrow{d} \mathcal{N}(0, \frac{4-\pi}{4})$, the distribution of the first term in (3.12) converges to a Gaussian distribution such that

$$\sqrt{\frac{M}{N}} \left( \frac{1}{M} \sum_{j=1}^{M} |h_{nj}| - \sqrt{\frac{\pi}{4}} \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{4-\pi}{4}\right),$$

and $\frac{1}{\sqrt{M}} \sum_{j=1}^{M} h_{ij} e^{-j\angle h_{nj}} \xrightarrow{d} \mathcal{N}(0,1)$. Further, $\sqrt{N}\left(\frac{1}{N} \sum_{i=1, i\neq n}^{N} z_i - \sqrt{\frac{\pi}{4}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{4-\pi}{4}\right)$, where $z_i = |\sum_{j=1}^{M} h_{ij} e^{-j\angle h_{nj}}|$ has a Rayleigh distribution with $\sigma = 1/2$. Moreover, the distribution

of the second term in (3.12) converges to a Gaussian distribution as well but with different mean and variance, i.e.,

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1,i\neq n}^{N}|\frac{1}{\sqrt{M}}\sum_{j=1}^{M}h_{ij}e^{-j\angle h_{nj}}|-\sqrt{\frac{\pi}{4}}\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{4-\pi}{4}\right).$$

Therefore, the mean and the variance of the $n^{th}$ diagonal element are asymptotically given by $\left(\sqrt{\frac{M\pi}{4N}}+\sqrt{\frac{N\pi}{4}}\right)$ and $\left(\frac{4-\pi}{4}+\frac{4-\pi}{4N}\right)$, respectively. Consequently, the antenna array gain, defined as the gain in the average received signal power, is asymptotically huge where $\mathbb{E}\left(|\tilde{h}_{nn}|^2\right)=\left(\sqrt{\frac{M\pi}{4N}}+\sqrt{\frac{N\pi}{4}}\right)^2+\frac{4-\pi}{4}+\frac{4-\pi}{4N}$.

■

This shows that the diagonal elements are hardened, i.e., concentrated around a relatively high value as both $N$ and $M$ grow unboundedly. Moreover, it highlights that the means of the diagonal elements of the effective channel scales linearly as $\sqrt{N}$. This assures that Algorithm 1 concentrates the diagonal elements of the effective channel around relatively large values, and thereby, the first goal is achieved.

**Corollary. 3.2.1.1.** *The ratio of the means of the off-diagonal elements to the means of the diagonal ones goes to zeros as $M$ and $N$ grow unboundedly while $\frac{M}{N}<1$, i.e., $\frac{\mathbb{E}\{\tilde{h}_{lp}\}}{\mathbb{E}\{\tilde{h}_{nn}\}}\rightarrow 0$ as $M\rightarrow\infty$ and $\frac{M}{N}<1$, thereby concentrating diagonal elements and dispersing the off-diagonal ones.*

*Proof.* Proposition 3.2.1 shows that the means of the diagonal entries of the effective channel scale linearly with $\sqrt{N}$ given Algorithm 1. Additionally, it can be shown that the means of the off-diagonal entries of the effective channel do not scale with neither $M$ nor $N$ but they are asymptotically upper bounded by a fixed small number, i.e., $\sqrt{\frac{\pi M}{4N}}$. Particularly, the $(l,p), p\neq l$

off-diagonal entry of the effective channel given Algorithm 1 is:

$$\tilde{h}_{l,p} = \frac{1}{\sqrt{NM}} \sum_{i=1}^{N} \sum_{j=1}^{M} h_{ij} e^{-j\angle h_{lj}} e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} h_{iu} e^{-j\angle pu}\right)} \tag{3.13}$$

Given that $h_{ij} = |h_{ij}|e^{j\psi_{ij}}$ where $|h_{ij}|$ has a Rayleigh distribution and $\psi_{ij}$ has a uniform one, the $(l,p), p \neq l$ off-diagonal entry of the effective channel is rewritten as:

$$\begin{aligned} \tilde{h}_{lp} =& \frac{1}{\sqrt{NM}} \sum_{i\neq p,l}^{N} \sum_{j=1}^{M} |h_{ij}| e^{-j(\psi_{ij}-\psi_{lj})} e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} |h_{iu}| e^{-j(\psi_{iu}-\psi_{pu})}\right)} \\ &+ \frac{1}{\sqrt{NM}} \sum_{j=1}^{M} |h_{pj}| e^{-j(\psi_{pj}-\psi_{lj})} + \sum_{j=1}^{M} |h_{lj}| e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} |h_{lu}| e^{-j(\psi_{lu}-\psi_{pu})}\right)} \end{aligned} \tag{3.14}$$

Accordingly, the mean of the the $(l,p), p \neq l$ off-diagonal entry of the effective channel is given by:

$$\begin{aligned} \mathbb{E}\{\tilde{h}_{lp}\} \overset{(a)}{=}& \frac{1}{\sqrt{NM}} \sum_{i\neq p,l}^{N} \sum_{j=1}^{M} \mathbb{E}\{|h_{ij}| e^{-j(\psi_{ij}-\psi_{lj})} e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} |h_{iu}| e^{-j(\psi_{iu}-\psi_{pu})}\right)}\} \\ &+ \frac{1}{\sqrt{NM}} \sum_{j=1}^{M} \mathbb{E}\{|h_{pj}| e^{-j(\psi_{pj}-\psi_{lj})}\} + \sum_{j=1}^{M} \mathbb{E}\{|h_{lj}| e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} |h_{lu}| e^{-j(\psi_{lu}-\psi_{pu})}\right)}\} \\ \overset{(b)}{=}& \frac{1}{\sqrt{NM}} \sum_{i\neq p,l}^{N} \sum_{j=1}^{M} \mathbb{E}\{e^{j\psi_{lj})}\} \mathbb{E}\Big\{|h_{ij}| e^{-j(\psi_{ij}-\psi_{lj})} e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} |h_{iu}| e^{-j(\psi_{iu}-\psi_{pu})}\right)}\Big\} \\ &+ \frac{1}{\sqrt{NM}} \sum_{j=1}^{M} \mathbb{E}\{e^{j\psi_{lj})}\} \mathbb{E}\{|h_{pj}| e^{-j\psi_{pj}}\} \\ &+ \frac{1}{\sqrt{NM}} \sum_{j=1}^{M} \mathbb{E}\Big\{|h_{lj}|\ \mathbb{E}\{e^{-j\angle\left(\frac{1}{\sqrt{M}} \sum\limits_{u=1}^{M} |h_{lu}| e^{-j(\psi_{lu}-\psi_{pu})}\right)} |\ |h_{lj}|\}\Big\} \\ \overset{(c)}{<}& \sqrt{\frac{\pi M}{4N}} \end{aligned} \tag{3.15}$$

Eq.(3.15.a) is due to the linearity of the $\mathbb{E}\{\}$ operator. Eq. (3.15.b) follows from the independence of the phases and magnitudes of the complex Gaussian random variable and the total law of expectation. Eq. (3.15.c) is upped bounded by $\sqrt{\frac{\pi M}{4N}}$ since the first and the second terms are zeros where $\psi_{lj}$ is uniformly distributed $[0, 2\pi)$ and thereby, $\mathbb{E}\{e^{j\psi_{lj}}\} = 0$ whereas the third term is upper bounded by bounding $\mathbb{E}\{e^{-j\angle\left(\frac{1}{\sqrt{M}}\sum_{u=1}^{M}|h_{lu}|e^{-j(\psi_{lu}-\psi_{pu})}\right)}|\ |h_{lj}|\} \leq 1$ where $\mathbb{E}\{|h_{ij}|\} = \sqrt{\frac{\pi}{4}}$. Therefore, using Algorithm 1, the means of the diagonal elements of the effective channel asymptotically approach $\mathbb{E}\{\tilde{h}_{nn}\} = \sqrt{\frac{M\pi}{4N}} + \sqrt{\frac{N\pi}{4}}$ whereas the means of the off-diagonal elements of the effective channel are bounded by $\mathbb{E}\{\tilde{h}_{lp}\} \leq \sqrt{\frac{\pi M}{4N}}$, thereby $\frac{\mathbb{E}\{\tilde{h}_{lp}\}}{\mathbb{E}\{\tilde{h}_{nn}\}} \leq \frac{1}{1+\sqrt{\frac{N^2}{M}}} \to 0$ as $M \to \infty$ and $\frac{M}{N} < 1$.

■

**Remark. 3.2.2.** ***Starting with the less constrained side:*** *Algorithm 1 consists of two main steps where the order of performing these steps dictates how the effective array gain scales with $M$ and $N$. The procedure order in Algorithm 1 results in effective array gain that scales with $N$ which makes this procedure beneficial when $N > M$, i.e., in the uplink transmission. On the other hand, reversing the order of these two steps by starting with co-phasing columns of the analog combiner to the channel columns, i.e., $w_{jn} = \frac{1}{\sqrt{N}}\frac{h^*_{jn}}{|h_{jn}|}, \forall 1 \leq j \leq N$, then proceeding to the second step of calculating the analog precoder, i.e., $f^*_{in} = \frac{1}{\sqrt{M}}\frac{\sum_{j=1}^{N} w^*_{nj}h_{ji}}{|\sum_{j=1}^{N} w^*_{nj}h_{ji}|},\ 1 \leq i \leq N, \forall i \neq n$, gives the distribution of the $n^{th}$ diagonal element $\xrightarrow{d} \mathcal{N}\left(\sqrt{\frac{N\pi}{4M}} + \sqrt{\frac{M\pi}{4}}, \frac{4-\pi}{4} + \frac{4-\pi}{4M}\right)$. This makes reversing the order of these two steps is beneficial when $M > N$, i.e., in the downlink transmission. Therefore, in our simulation, we run the two main steps in Algorithm 1 in succession according to the following rules. In case of $N > M$, the analog precoder is obtained first then the analog combiner is calculated. On the other hand, in case of $M > N$, the analog combiner is calculated first by co-phasing columns of the analog combiner to the channel columns, similar to step 2 in*

*Algorithm 1 but on column-wise instead of row-wise, then the analog precoder is calculated. In summary, starting with the less constrained side allows for harvesting higher antenna array gain from the other side.*

**Analog processing for partially-connected antenna array structures**

Due to the fewer number of antenna elements connected to the RF chains in the sub-array structure, both the analog precoder and combiner have block diagonal structures. Particularly, $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$ are sparse and contain vectors of sizes $M_{sa} \times 1$ and $N_{sa} \times 1$, respectively, on their diagonal whereas the rest entries are zeros. With a slight abuse of notation, we denote the non-zero part of the $i^{th}$ column of the analog precoding matrix $\mathbf{F}_{RF}$ by $\mathbf{f}_i$ with size $M_{sa} \times 1$, while the non-zero part of $i^{th}$ column of the analog combining matrix $\mathbf{W}_{RF}$ is denoted by $\mathbf{w}_i$ and has a size $N_{sa} \times 1$. Accordingly, the effective channel after analog beamforming is:

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{w}_1^H \mathbf{H}_{11} \mathbf{f}_1 & \cdots & \mathbf{w}_1^H \mathbf{H}_{1M_{rf}} \mathbf{f}_{M_{rf}} \\ \mathbf{w}_2^H \mathbf{H}_{21} \mathbf{f}_1 & \cdots & \mathbf{w}_2^H \mathbf{H}_{2M_{rf}} \mathbf{f}_{M_{rf}} \\ \vdots & \vdots & \vdots \\ \mathbf{w}_{N_{rf}}^H \mathbf{H}_{N_{rf}1} \mathbf{f}_1 & \cdots & \mathbf{w}_{N_{rf}}^H \mathbf{H}_{N_{rf}M_{rf}} \mathbf{f}_{M_{rf}} \end{bmatrix}, \tag{3.16}$$

where $\mathbf{H}_{ij}$ for $1 \leq i \leq N_{rf}$ and $1 \leq j \leq M_{rf}$ is the $(i,j)^{th}$ block of the channel matrix $\mathbf{H}$ and has a size $N_{sa} \times M_{sa}$. Interestingly, the block diagonal structure of the analog precoder and combiner virtually divides the massive MIMO channel $\mathbf{H}$ into $M_{rf} N_{rf}$ blocks "sub-channels" of size $N_{sa} \times M_{sa}$, each. Similar to the fully-connected case, here, the analog processing is concerned with maximizing the sum of sub-antenna array gains, i.e., the sum of the squared norms of the effective

sub-channel gains per sub-array. The joint analog precoding and combining problem is given by:

$$\begin{aligned} \underset{\mathbf{F}_{RF},\mathbf{W}_{RF}}{\text{maximize}} \quad & \sum_{i}^{N_{rf}} |\mathbf{w}_i^H \mathbf{H}_{ii} \mathbf{f}_i|^2 \\ \text{subject to} \quad & |(\mathbf{f}_i)_j| = \frac{1}{\sqrt{M_{sa}}} \;\; \forall i, 1 \leq j \leq M_{sa}, \;\; \text{rank}(\mathbf{F}_{RF}) = N_{rf} \\ & |(\mathbf{w}_i)_l| = \frac{1}{\sqrt{N_{sa}}} \;\; \forall i, 1 \leq l \leq N_{sa}, \;\; \text{rank}(\mathbf{W}_{RF}) = N_{rf} \end{aligned} \tag{3.17}$$

Due to the virtual block structure of $\tilde{\mathbf{H}}$, the summands in (3.17) are non-negative functions of different sub-channels, i.e., $\mathbf{H}_{ii}, 1 \leq i \leq N_{rf}$, thereby being separable. Moreover, the block diagonal structures of $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$ render the rank constraints are always satisfied for any given solution expect the all zeros one, thereby dropping these constraints. Therefore, due to the separability of the cost function and constraints, the optimization problem in (3.17) can be solved by separately solving $N_{rf}$ independent sub-problems. Accordingly, the design problem of joint analog precoding and combining for multi-stream systems is reduced to $N_{rf}$ independent single-stream joint analog precoding and combining problems. These problems are:

$$\begin{aligned} \underset{\boldsymbol{\phi}_i,\boldsymbol{\theta}_i}{\text{maximize}} \quad & |\mathbf{w}_i^H \mathbf{H}_{ii} \mathbf{f}_i|^2 \\ \text{subject to} \quad & \mathbf{f}_i = \frac{1}{\sqrt{M_{sa}}} e^{j\boldsymbol{\theta}_i}, \;\; \mathbf{w}_i = \frac{1}{\sqrt{N_{sa}}} e^{j\boldsymbol{\phi}_i}, \end{aligned} \tag{3.18}$$

to find $N_{rf}$ analog precoding and combining pairs, i.e., $(\mathbf{f}_i, \mathbf{w}_i)$ for $1 \leq i \leq N_{rf}$, where $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_i$ are vectors of length $M_{sa}$ and $N_{sa}$, respectively, and their entries are real numbers $\in [0, 2\pi]$. These optimization problems are non-convex and challenging to secure optimal solutions mainly due to the non-convex RF constraints and the cost function is invariant under phase rotation.Motivated by the excellent performance of alternating optimization techniques in the literature of hybrid and digital beamforming [120, 131, 145] and driven by the need to low-complexity solutions, we solve the optimization problem defined in (3.18) algorithmically using alternating optimization. The joint analog precoding and combining strategy is implemented iteratively in two alternating steps. First, given $\mathbf{f}_i$, we solve (3.18) for $\mathbf{w}_i$. Interestingly, given $\mathbf{f}_i$

and while relaxing the constant magnitude entries constraint on the combiner to the unit second norm constraint (local convexification), the optimal combiner is $\mathbf{w}_{\star i} = \frac{\mathbf{H}_{ii}\mathbf{f}_i}{\|\mathbf{H}_{ii}\mathbf{f}_i\|_2}$ [73]. After that, we impose the unit magnitude entries constraint by dividing each entry of the optimal combiner, obtained from the relaxed problem, by its magnitude, i.e., $\frac{(\mathbf{w}_{\star i})_j}{|(\mathbf{w}_{\star i})_j|}, \forall 1 \leq j \leq N_{sa}$. We note that this method of imposing the unit magnitude constraint by dividing the entries of the optimal solution of the relaxed problem by their norms is widely adopted in the literature of hybrid beamforming [34, 90, 120, 145] since this solution has the minimum Euclidean distance to the optimal solution of the relaxed problem. Similarly, in the second step, given the updated value of $\mathbf{w}_i$, and relaxing the unit magnitude entries constraints, the optimal beamformer is $\mathbf{f}_{\star i} = \frac{\mathbf{H}_{ii}^H\mathbf{w}_i}{\|\mathbf{H}_{ii}\mathbf{f}_i\|_2}$. Thereby, the updated values of the beamformer are $\frac{(\mathbf{f}_{\star i})_j}{|(\mathbf{f}_{\star i})_j|}, \forall 1 \leq j \leq M_{sa}$. Finally, we keep iterating between these two steps until satisfying certain stopping criterion or reaching to maximum number of iterations, i.e., $K^{\max}$. We summarize these two steps in Algorithm 2.

**Algorithm 2** Iterative Maximization of Antenna Array Gain per Sub-Array

**Require:** $\mathbf{H}$, $\mathbf{F}_{RF}^{(0)}$, $K^{\max}$
1: **While** $i \leq N_{rf}$
2: $k = 0$,
3: **Repeat**
4: $k = k + 1$
5: $(\mathbf{w}_i^{(k)})_j = \frac{1}{\sqrt{N_{sa}}} \frac{(\mathbf{H}_{ii}\mathbf{f}_i^{(k-1)})_j}{|(\mathbf{H}_{ii}\mathbf{f}_i^{(k-1)})_j|}, \quad \forall 1 \leq j \leq N_{sa}$
6: $(\mathbf{f}_i^{(k)})_j = \frac{1}{\sqrt{M_{sa}}} \frac{(\mathbf{H}_{ii}^H\mathbf{w}_i^{(k)})_j}{|(\mathbf{H}_{ii}^H\mathbf{w}_i^{(k)})_j|}, \quad \forall 1 \leq j \leq M_{sa}$
7: **Until** stopping criterion is satisfied or $k = K^{\max}$
8: **End**
9: **return** $\mathbf{F}_{RF} = [\mathbf{f}_1, \cdots, \mathbf{f}_i, \cdots, \mathbf{f}_{N_{rf}}], \mathbf{W}_{RF} = [\mathbf{w}_1, \cdots, \mathbf{w}_i, \cdots, \mathbf{w}_{N_{rf}}]$

In order to evaluate the performance of the proposed analog precoding and combining procedure, we derive upper and lower bounds on the antenna array gain per sub-array, i.e., the maximum of the cost function in (3.18).

**Proposition. 3.2.3.** *The antenna array gain per sub-channel for a MIMO system with analog*

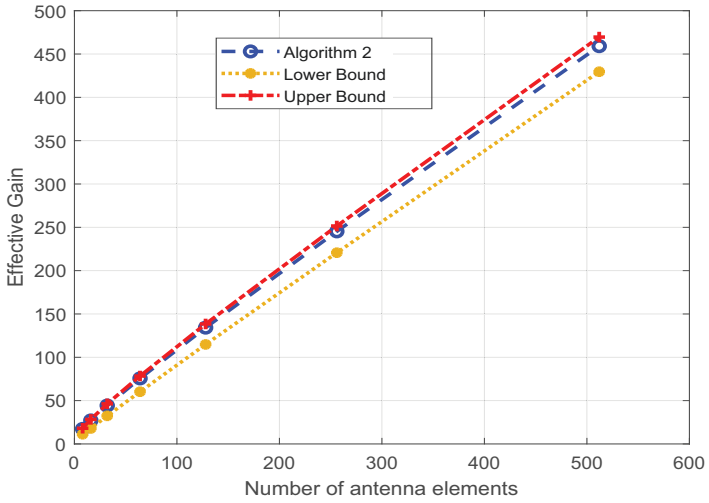


Figure 3.3: Upper and lower bounds on the effective gain of Algorithm 2

*precoding and combining, and partially-connected antenna array structure is bounded by:*

$$\max\left(\frac{\|\mathbf{H}_{ii}\|_{\infty}^{2}}{M_{sa}}, \frac{\|\mathbf{H}_{ii}\|_{1}^{2}}{N_{sa}}\right) < \frac{|e^{-j\boldsymbol{\phi}_i}\mathbf{H}_{ii}e^{j\boldsymbol{\theta}_i}|^2}{N_{sa}M_{sa}} < \min\left(\frac{\sigma_1^2\|\mathbf{u}_1\|_{1}^{2}}{N_{sa}}, \frac{\sigma_1^2\|\mathbf{v}_1\|_{1}^{2}}{M_{sa}}\right), \tag{3.19}$$

*where $\sigma_1$ is the largest singular value of the matrix $\mathbf{H}_{ii}$, $\mathbf{u}_1$ and $\mathbf{v}_1$ are the right and left singular vector associated with $\sigma_1$, respectively.*

*Proof.* please refer to Appendix A.1. ■

To the best of the authors' knowledge, these bounds on the joint equal gain transmission and combiner, have not been reported in the literature, although some pieces of the proof have been available for a long time and they have been studied by different researchers such as Love *et al.* [73] and Tsai [131]. The tightness of the upper and lower bounds, and the performance evaluation of the proposed analog procedure are shown in Fig. 3.3 where antenna array gain per sub-channel of Algorithm 2 along with the obtained bounds are drawn when increasing the number of antenna elements over a Rayleigh channel. Fig. 3.3 shows that the antenna array gain of the proposed analog procedure in Algorithm 2 is tightly bounded by the derived bounds.

**Remark. 3.2.4.** ***Extension to different numbers of RF chains at both sides:*** *For the more general case when $N_{rf} \neq M_{rf}$, i.e., $\tilde{\mathbf{H}}$ is a rectangular matrix and both algorithms run over the entries of the main diagonal of $\tilde{\mathbf{H}}$ in order to jointly calculate the beamforming pairs $(\mathbf{f}_n, \mathbf{w}_n)$ for $n \leq \min(N_{rf}, M_{rf})$. This leaves $\max(N_{rf}, M_{rf}) - \min(N_{rf}, M_{rf})$ beamforming vectors, either $\mathbf{f}_i$ or $\mathbf{w}_i$ for $\min(N_{rf}, M_{rf}) < i \leq \max(N_{rf}, M_{rf})$ are not determined, depending on either $M_{rf} > N_{rf}$ or $N_{rf} > M_{rf}$, respectively. There are many options for calculating these beamforming vectors. For instance, a trivial solution is obtained by setting these beamforming vectors to zeros (turning off the extra RF chains for saving power), i.e., with loss of generality when $M_{rf} > N_{rf}$, $\mathbf{f}_i = \mathbf{0}$ for $N_{rf} < i \leq M_{rf}$. This solution forces the extra $\max(N_{rf}, M_{rf}) - \min(N_{rf}, M_{rf})$ rows or columns of the rectangular effective channel (represent interference) to be zero vectors. Indeed, this solution is equivalent to the setup where $N_{rf} = M_{rf}$ and they have the same performance. From channel diagonalization perspective, this would be a good solution since it removes interference due to the extra columns/rows and increase the transmitted power pumped into the main diagonal entries. However, the analog beamforming stage (performed by Algorithm 1 and 2) is followed by a digital beamforming stage that can be utilized to remove the interference. Consequently, these unused RF chains can be utilized efficiently to provide higher array gains even though they might cause interference between streams. This made possible by co-phasing the $\max(N_{rf}, M_{rf}) - \min(N_{rf}, M_{rf})$ beamforming vectors (precoder/combiner) to $\max(N_{rf}, M_{rf}) - \min(N_{rf}, M_{rf})$ channel (row/column) vectors. This solution gives a higher spectral efficiency performance than the trivial solution (turning off the extra RF chains) and enhances the spectral efficiency proportionally with $\max(N_{rf}, M_{rf}) - \min(N_{rf}, M_{rf})$. To see this, with out loss of generality, let us assume that $M_{rf} > N_{rf}$ therefore $\tilde{\mathbf{H}} = [\mathbf{A} \;\; \mathbf{B}]$ where $\mathbf{A}$ is an $N_{rf} \times N_{rf}$ matrix and $\mathbf{B}$ is $N_{rf} \times (M_{rf} - N_{rf})$ matrix. The singular values of $\tilde{\mathbf{H}}$ are square roots of the eigenvalues of*

$$\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H = \begin{bmatrix} \mathbf{A}^H\mathbf{A} & \mathbf{A}^H\mathbf{B} \\ \mathbf{B}^H\mathbf{A} & \mathbf{B}^H\mathbf{B} \end{bmatrix}.$$

*According to the Cauchy interlacing theorem, if the eigenvalues of $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H$ are $\gamma_1 \leq \cdots \leq \gamma_{M_{rf}}$ and the eigenvalues of $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H$ after removing $M_{rf} - N_{rf}$ rows and vectors, i.e., $\mathbf{A}^H\mathbf{A}$ are*

$\alpha_1 \leq \cdots \leq \alpha_{N_{rf}}$ *then for all* $j \leq N_{rf}$,

$$\gamma_j \leq \alpha_j \leq \gamma_{M_{rf}-N_{rf}+j}.$$

*This means that the* $N_{rf}$ *largest eigenvalues of* $\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H$ *are larger or at least equal those of* $\mathbf{A}^H\mathbf{A}$ *and thereby their associated singular values. This implies that the spectral efficiency as a function of* $\tilde{\mathbf{H}}$ *resulting from the proposed solution is higher than the spectral efficiency resulting from the trivial solution or when* $N_{rf} = M_{rf}$, *i.e.,* $\mathcal{R}(\tilde{\mathbf{H}}) \geq \mathcal{R}(\mathbf{A})$. *This is verified in Fig. 3.15 where we used Algorithm 2 to calculate* $(\mathbf{w}_n, \mathbf{f}_n)$ *pairs for all* $n \leq N_{rf}$ *and the scheme of co-phasing the extra* $(M_{rf} - N_{rf})$ *preocder vectors to* $(M_{rf} - N_{rf})$ *channel rows next to the* $N_{rf}$ *rows utilized by Algorithm 2.*

### 3.2.3 Digital Processing

Here, given the analog precoder and combiner, we drive the optimal digital precoder and combiner that jointly maximize the spectral efficiency of the system. However, the digital stage can be designed to optimize more practical performance metrics such as symbol mean square error and minimum distance between symbol hypotheses [114]. This can be integrated readily into our framework since the analog stage is built based on the channel diagonalization concept which is the core machinery for optimizing all these performance metrics. Given the analog precoder and combiner, the spectral efficiency optimization problem in (3.4) can be written as:

$$\begin{aligned} \max_{\mathbf{F}_{BB},\mathbf{W}_{BB}} \quad & \log\det\left(\mathbf{I}_{N_s} + \left(\mathbf{W}_{BB}^H\mathbf{R}_{nRF}\mathbf{W}_{BB}\right)^{-1}\mathbf{W}_{BB}^H\tilde{\mathbf{H}}\mathbf{F}_{BB}\mathbf{F}_{BB}^H\tilde{\mathbf{H}}^H\mathbf{W}_{BB}\right), \\ \text{s.t.} \quad & \text{trace}(\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^H) \leq p \end{aligned} \tag{3.20}$$

where $\mathbf{R}_{nRF} = \sigma_n^2\mathbf{W}_{RF}^H\mathbf{W}_{RF}$ and $\tilde{\mathbf{H}} = \mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}$. This problem has closed-form optimal solutions as given in the following proposition.

**Proposition. 3.2.5.** *Let the eigenvalue decomposition to* $(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\tilde{\mathbf{H}}^H\mathbf{R}_{n_{RF}}^{-1}\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}\bar{\mathbf{V}}^H$*, then the optimal digital precoder and combiner that maximize (3.20) are given by:*

$$\begin{aligned}\mathbf{F}_{BB\star} &= (\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\bar{\mathbf{V}}_{N_s}\mathbf{\Gamma}\\ \mathbf{W}_{BB\star} &= \mathbf{R}_{n_{RF}}^{-1}\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\bar{\mathbf{V}}_{N_s}\mathbf{\Gamma},\end{aligned} \tag{3.21}$$

*where* $\bar{\mathbf{V}}_{N_s}$ *is the* $N_s$ *largest eigenvectors of* $\bar{\mathbf{V}}$*, and* $\mathbf{\Gamma}$ *is an* $N_s \times N_S$ *diagonal matrix that contains the power fractions of all transmitted streams on its diagonal,*

$$|\gamma_{ii}|^2 = \Big(\frac{p + \sum_k^{N_s}\lambda_{kk}^{-1}}{N_s} - \frac{1}{\lambda_{ii}}\Big)^+, i = 1, \cdots, N_s. \tag{3.22}$$

*Proof.* To decouple the total power constraint in (3.20), one uses a change of variables $\mathbf{F}_{BB} = \big(\mathbf{F}_{RF}^H\mathbf{F}_{RF}\big)^{-\frac{1}{2}}\mathbf{F}_{bb}$. This change of variables results in the following problem:

$$\begin{aligned}\max_{\mathbf{F}_{BB},\mathbf{W}_{BB}} \quad & \log\det\big(\mathbf{I}_{N_s} + \big(\mathbf{W}_{BB}^H\mathbf{R}_{nRF}\mathbf{W}_{BB}\big)^{-1}\mathbf{W}_{BB}^H\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\mathbf{F}_{bb}\mathbf{F}_{bb}^H(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\tilde{\mathbf{H}}^H\mathbf{W}_{BB}\big),\\ \text{s.t.} \quad & \text{trace}(\mathbf{F}_{bb}\mathbf{F}_{bb}^H) \le p\end{aligned} \tag{3.23}$$

This problem is equivalent to (3.20) since the change of variables $\psi(\mathbf{F}_{bb}) = (\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\mathbf{F}_{bb}$ is a one-to-one mapping function where $\mathbf{F}_{RF}$ is assumed to be full rank, thereby $\psi^{-1}(\psi(\mathbf{F}_{bb})) = \mathbf{F}_{bb}$. The cost function in (3.23) can be written as:

$$\begin{aligned}\mathcal{R} &= \log\det\big(\mathbf{I}_{N_s} + \mathbf{F}_{bb}^H(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\tilde{\mathbf{H}}^H\mathbf{R}_{nRF}^{-\frac{1}{2}}\,\mathbf{\Xi}_{\mathbf{R}_{n_{RF}}^{\frac{1}{2}}\mathbf{W}_{BB}}\,\mathbf{R}_{nRF}^{-\frac{1}{2}}\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\mathbf{F}_{bb}\big)\\ &\le \log\det\big(\mathbf{I}_{N_s} + \mathbf{F}_{bb}^H(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\tilde{\mathbf{H}}^H\mathbf{R}_{nRF}^{-1}\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\mathbf{F}_{bb}\big),\end{aligned} \tag{3.24}$$

where $\mathbf{\Xi}_{\mathbf{R}_{n_{RF}}^{\frac{1}{2}}\mathbf{W}_{BB}} = \mathbf{R}_{n_{RF}}^{\frac{1}{2}}\mathbf{W}_{BB}(\mathbf{W}_{BB}^H\mathbf{R}_{nRF}\mathbf{W}_{BB})^{-1}\mathbf{W}_{BB}^H\mathbf{R}_{n_{RF}}^{\frac{1}{2}}$. $\mathbf{\Xi}_{\mathbf{R}_{n_{RF}}^{\frac{1}{2}}\mathbf{W}_{BB}} \preceq \mathbf{I}_{N_{RF}}$ since it is the orthogonal projector onto the range of $\mathbf{R}_{n_{RF}}^{\frac{1}{2}}\mathbf{W}_{BB}$. The upper bound on $\mathcal{R}$ is achievable if $\mathbf{R}_{n_{RF}}^{\frac{1}{2}}\mathbf{W}_{BB} = \mathbf{R}_{n_{RF}}^{-\frac{1}{2}}\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\mathbf{F}_{bb}$. This yields $\mathbf{W}_{BB\star} = \mathbf{R}_{n_{RF}}^{-1}\tilde{\mathbf{H}}(\mathbf{F}_{RF}^H\mathbf{F}_{RF})^{-\frac{1}{2}}\mathbf{F}_{bb}$. Substituting with $\mathbf{W}_{BB\star}$ in (3.23), one gets the water-filling solution obtained from maximizing the mutual information bound in (3.24) over $\mathbf{F}_{bb}$, i.e., $\mathbf{F}_{bb\star} = \bar{\mathbf{V}}_{N_s}\mathbf{\Gamma}$ [114]. ■

We note here that the solutions in (3.21) are not given by the simple SVD of $\tilde{\mathbf{H}}$ due to the coupling between analog and digital precoding in the total power constraint and the contribution of analog and digital combiners in coloring the Gaussian noise. Moreover, the solutions in (3.21) lead to diagonalizing the $\tilde{\mathbf{H}}$ (removing its dispersed off-diagonal elements with small means) and the noise covariance matrix $\mathbf{W}_{BB}^{H}\mathbf{R}_{n_{RF}}\mathbf{W}_{BB}$. This is in contrast to the digital beamforming in [15,66,84,90,106,139,145] where only power normalization techniques are considered instead of accounting for the joint power constraint between analog and digital precoder. Moreover, this is in contrast to the digital combining in [15, 34, 120] where MMSE filter is considered (or decoupling the transmitter and receiver designs) instead of the joint precoding and combining design.

## 3.3 Extensions to Multi-User cases

The most adopted multi-user case in both theoretical works and real-life applications is the broadcast channel. We consider three different RF architectures of the $K$-user MIMO broach channel which show the different extreme points of the RF connections. The first is where the base station and all users sub-array RF structures. The second is where the base station considers the sub-array RF structure and all users have fully-connected RF structures. The third is where all network's terminals have RF structures.

### 3.3.1 System Model

In the $K$-user MIMO broadcast setup, we assume that the base station (transmitting terminal) has $M$ -element antenna and $KM_{rf} < M$ RF chains while each user $k$ , $\in \{1, \cdots, K\}$ is equipped (receiving terminal) with an $N_k$-element antenna and $N_{rf,k} < N_k$ RF chains. RF chains at any terminal can be either connected to all antenna elements or to disjoint groups of antenna elements constructing either fully or partially-connected RF structure, respectively. We assume that the base station sends $Ns < M_{rf}$ independent data streams to support each user with $Ns < N_{rf,k}$

stream per channel use. Since the number of RF chains at both sides is limited, we considered the hybrid beamforming architecture at both sides. Particularly, the base station pre-processes an $KNs \times 1$ data vector $\mathbf{s}$ at baseband using an $KM_{rf} \times KNs$ digital precoder $\mathbf{F}_{BB}$ and then pre-processes them at RF using an $M \times KM_{rf}$ analog precoder $\mathbf{F}_{RF}$. As a result, the transmitted data vector is:

$$\mathbf{x} = \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} = \sum_{k=1}^{K} \mathbf{F}_{RF}\mathbf{F}_{BB,k}\mathbf{s}_k, \tag{3.25}$$

where $\mathbf{F}_{BB} = [\mathbf{F}_{BB,1}, \cdots, \mathbf{F}_{BB,K}]$ and $\mathbf{s} = [\mathbf{s}_1^T, \cdots, \mathbf{s}_K^T]^T$. Accordingly, the received signal at user $k$ is:

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{F}_{RF}\mathbf{F}_{BB,k}\mathbf{s}_k + \mathbf{H}_k \sum_{l=1,l\neq k}^{K} \mathbf{F}_{RF}\mathbf{F}_{BB,l}\mathbf{s}_l + \mathbf{n}_k, \tag{3.26}$$

where $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$ the channel matrix of user $k$ and $\mathbf{n}_k \in \mathbb{C}^{N_k \times 1}$ is the white Gaussian noise vector.

Analogous to the transmitter side, each user post-processes the received vector at RF using an $N_k \times N_{rf,k}$ analog combiner $\mathbf{W}_{RF,k}$ and then post-processes it at baseband using an $N_{rf,k} \times N_s$ digital combiner, i.e., $\mathbf{W}_{BB,k}$, and therefore, the processed received signal vector at user $k$ is:

$$\tilde{\mathbf{s}}_k = \overbrace{\mathbf{W}_{BB,k}^H\mathbf{W}_{RF,k}^H\mathbf{H}_k\mathbf{F}_{RF}\mathbf{F}_{BB,k}\mathbf{s}_k}^{\text{Desired signal of user k}} + \underbrace{\mathbf{W}_{BB,k}^H\mathbf{W}_{RF,k}^H\mathbf{H}_k \sum_{l=1,l\neq k}^{K} \mathbf{F}_{RF}\mathbf{F}_{BB,l}\mathbf{s}_l}_{\text{Inter-user interference}} + \underbrace{\mathbf{W}_{BB,k}^H\mathbf{W}_{RF,k}^H\mathbf{n}_k}_{\text{Gaussian noise at baseband}}, \tag{3.27}$$

A generic block diagram of the $K$-user MIMO broadcast channel with hybrid beamforming structure is depicted in Fig. 3.4.

The spectral efficiency of user $k$ assuming Gaussian signaling is:

$$\mathcal{R}_k = \log\det(\mathbf{I}_{N_s} + \mathbf{R}_{i+n,k}^{-1}\mathbf{W}_{BB,k}^H\mathbf{W}_{RF,k}^H\mathbf{H}_k\mathbf{F}_{RF}\mathbf{F}_{BB,k}\mathbf{F}_{BB,k}^H\mathbf{F}_{RF}^H\mathbf{H}_k^H\mathbf{W}_{RF,k}\mathbf{W}_{BB,k}), \tag{3.28}$$

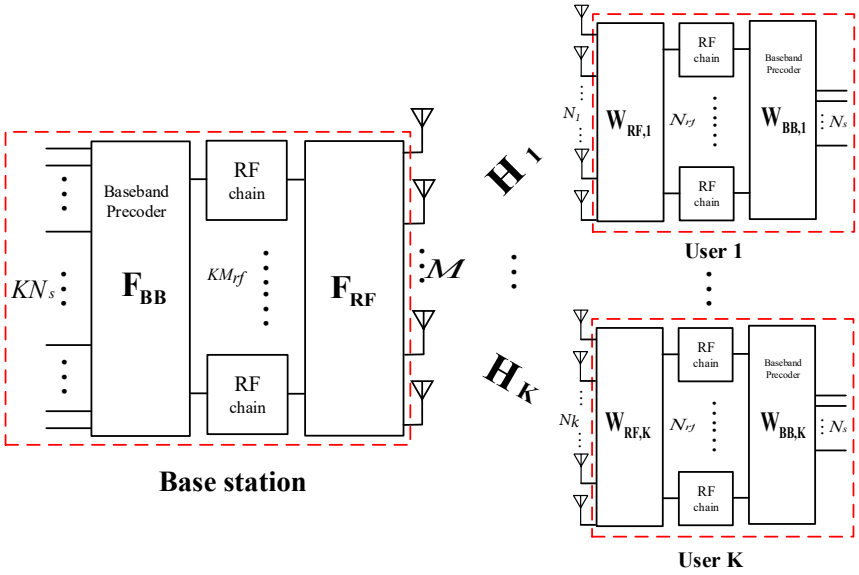


Figure 3.4: K-user MIMO broadcast channel with hybrid beamforming structure.

where $\mathbf{R}_{i+n,k} = \mathbf{W}_{BB,k}^H \mathbf{W}_{RF,k}^H \mathbf{H}_k \big( \sum_{l=1, l\neq k}^{K} \mathbf{F}_{RF} \mathbf{F}_{BB,l} \mathbf{F}_{BB,l}^H \mathbf{F}_{RF}^H \big) \mathbf{H}_k^H \mathbf{W}_{RF,k} \mathbf{W}_{BB,k} + \sigma_n^2 \mathbf{W}_{BB,k}^H \mathbf{W}_{RF,k}^H \mathbf{W}_{RF,k} \mathbf{W}_{BB,k}$ is the interference plus noise covariance matrix of user $k$. Similar to the single-user case, we consider maximizing the sum of the spectral efficiencies under total average power constraint and the RF constraints of analog precoders and combiners where the problem of interest is formalized as:

$$\begin{aligned} \max_{\mathbf{F}_{BB}, \mathbf{W}_{BB,k}} \quad & \sum_{k=1}^{K} \mathcal{R}_k \\ \text{s.t.} \quad & |f_{RF(i,j)}| = \frac{1}{\sqrt{M_{sa}}} \ \forall i,j, \quad |w_{RF,k(u,l)}| = \frac{1}{\sqrt{N_{sa}}} \ \forall u,l,k \\ & \sum_{k} \text{trace}(\mathbf{F}_{RF} \mathbf{F}_{BB,k} \mathbf{F}_{BB,k}^H \mathbf{F}_{RF}^H) \leq p, \end{aligned} \tag{3.29}$$

This problem is non-convex and it is extremely challenging to find its optimal solutions due to the difficult nature of the RF constraints. We are interested in securing good sub-optimal

and low complexity solutions to this problem under different RF structures. Before proceeding to the proposed solutions, it is worth mentioning that the optimal linear fully-digital solution that maximizes (3.29) while relaxing the RF hardware constraints is the block diagonalization scheme [125]. The block diagonalization scheme diagonalizes the augmented matrix of the multi-user channels, i.e., $\mathbf{H}_{aug} = [\mathbf{H}_1^T, \mathbf{H}_2^T, \cdots, \mathbf{H}_K^T]^T$, over two stages. The first stage is to remove the inter-user interference by transmitting each user's signals over the common null space of other users' channels which leads to block diagonalizing $\mathbf{H}_{aug}$. This stage requires calculating the SVD of a $(K-1)N \times M$ matrix for $K$ times. The second is to remove the inter-stream interference at each user by diagonalizing each block on the main diagonal of $\mathbf{H}_{aug}$ which leads to completely converting $\mathbf{H}_{aug}$ into $KN_s$ parallel channels. This stage requires to calculate the SVD of an $N \times N_s$ matrix for $K$ times. However, adopting this scheme in mmWave massive MIMO systems is highly avoided due to its high computational complexity and inapplicability to hybrid structures.

Taking our cue from the block diagonalization scheme, the proposed solutions are based on diagonalizing the channel over two stages where the analog stage is aiming at condensing the elements on the block diagonal of the augmented matrix of the multi-user effective channels through harvesting the antenna array gain offered by the massive MIMO channel whereas the digital one aims at completely canceling out the inter-stream and inter-user interference and ultimately maximizing the sum of the users' achievable rates.

### 3.3.2 Analog Processing

Similar to the single-user case, the main purpose of the analog processing is to harden/condense the diagonal elements of the blocks on the diagonal of the augmented matrix of the effective channels after the analog process, i.e., $\tilde{\mathbf{H}}_{aug} = \mathbf{W}_{RF}^H \mathbf{H}_{aug} \mathbf{F}_{RF}$ by utilizing appropriate co-phasing (adding coherently using phase-only control) techniques while dispersing the elements of the off-diagonal blocks. However, the various RF structures of analog processing results in different structures of $\tilde{\mathbf{H}}_{aug}$ that require different treatments. In the following subsections, we show how to handle the various RF structures to achieve that goal.

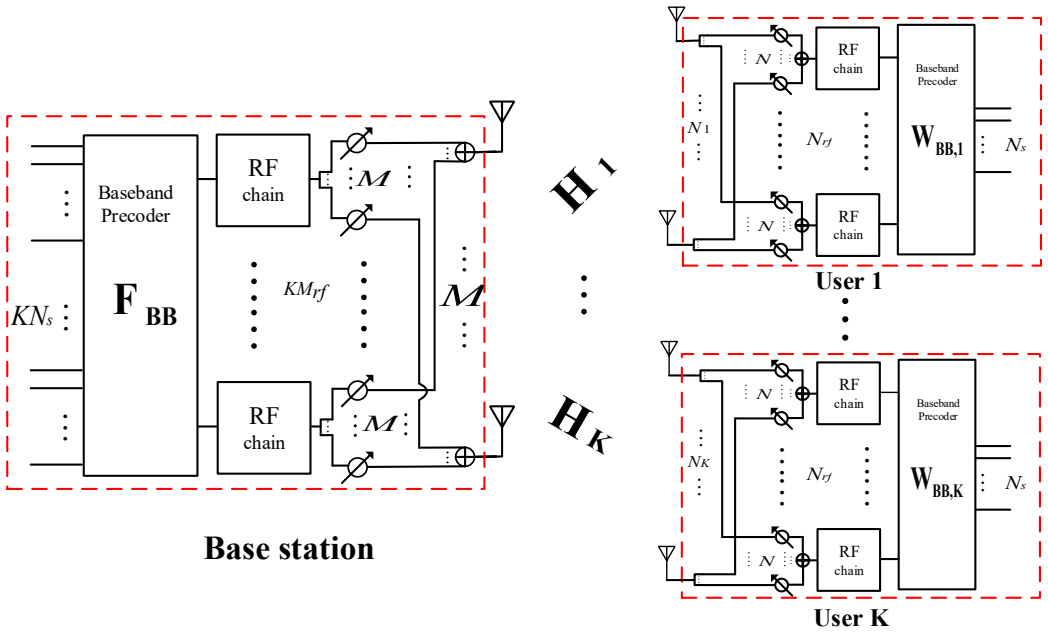


Figure 3.5: K-user MIMO broadcast channel with fully-connected RF structure.

**Fully-connected RF structures at all terminals**

In this setup, all the network's terminals (base station and $K$ users) consider the fully-connected RF structure where each RF chain is connected to all the antenna elements as indicated in Fig. 3.5.

Abstracting the digital processing, one can write the augmented matrix of the multi-user

effective channels after analog processing as:

$$\tilde{\mathbf{H}}_{aug} = \begin{bmatrix} \mathbf{W}_{RF,1}^H & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{RF,2}^H & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_{RF,K}^H \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_K \end{bmatrix} \begin{bmatrix} \mathbf{F}_{RF,1}, \mathbf{F}_{RF,2}, \cdots, \mathbf{F}_{RF,K} \end{bmatrix} \tag{3.30}$$

$$= \begin{bmatrix} \mathbf{W}_{RF,1}^H \mathbf{H}_1 \mathbf{F}_{RF,1} & \mathbf{W}_{RF,1}^H \mathbf{H}_1 \mathbf{F}_{RF,2} & \cdots & \mathbf{W}_{RF,1}^H \mathbf{H}_1 \mathbf{F}_{RF,K} \\ \mathbf{W}_{RF,2}^H \mathbf{H}_2 \mathbf{F}_{RF,1} & \mathbf{W}_{RF,2}^H \mathbf{H}_2 \mathbf{F}_{RF,2} & \cdots & \mathbf{W}_{RF,2}^H \mathbf{H}_2 \mathbf{F}_{RF,K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{RF,K}^H \mathbf{H}_K \mathbf{F}_{RF,1} & \mathbf{W}_{RF,K}^H \mathbf{H}_K \mathbf{F}_{RF,2} & \cdots & \mathbf{W}_{RF,K}^H \mathbf{H}_K \mathbf{F}_{RF,K} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{H}}_1 \\ \tilde{\mathbf{H}}_2 \\ \vdots \\ \tilde{\mathbf{H}}_K \end{bmatrix}$$

Due to the distributed nature of the reception processing at users, the augmented analog combiner $\mathbf{W}_{RF}$ has a block diagonal structure as shown in (3.30). This allows for separating the multi-user design problem of the analog processing to $K$ independent lower dimensional single-user analog processing problems. Particularly, we formalize the analog processing problems for $k = 1, \cdots, K$ as:

$$\begin{aligned} \underset{\mathbf{F}_{RF,k}, \mathbf{W}_{RF,k}}{\text{maximize}} \quad & \sum_{n=1}^{N_{rf}} |\mathbf{w}_{RF,k_{(n)}}^H \mathbf{H}_k \mathbf{f}_{RF,k_{(n)}}|^2 \\ \text{subject to} \quad & |f_{RF,k_{(i,j)}}| = \frac{1}{\sqrt{M}} \quad \forall i,j, \quad \text{rank}(\mathbf{F}_{RF,k}) = N_{rf} \\ & |w_{RF,k_{(u,l)}}| = \frac{1}{\sqrt{N}} \quad \forall u,l, \quad \text{rank}(\mathbf{W}_{RF,k}) = N_{rf} \end{aligned} \tag{3.31}$$

where $\mathbf{f}_{RF,k_{(n)}}$ and $\mathbf{w}_{RF,k_{(n)}}$ are the $n^{\text{th}}$ columns of $\mathbf{F}_{RF,k}$, and $\mathbf{W}_{RF,k}$, respectively. Fortunately, these problems have the same formulation of the single-user analog processing problem in (3.6) and can be solved in parallel to obtain $(\mathbf{F}_{RF,k}, \mathbf{W}_{RF,k})$ for $k = 1, \cdots K$ using Algorithm 1.

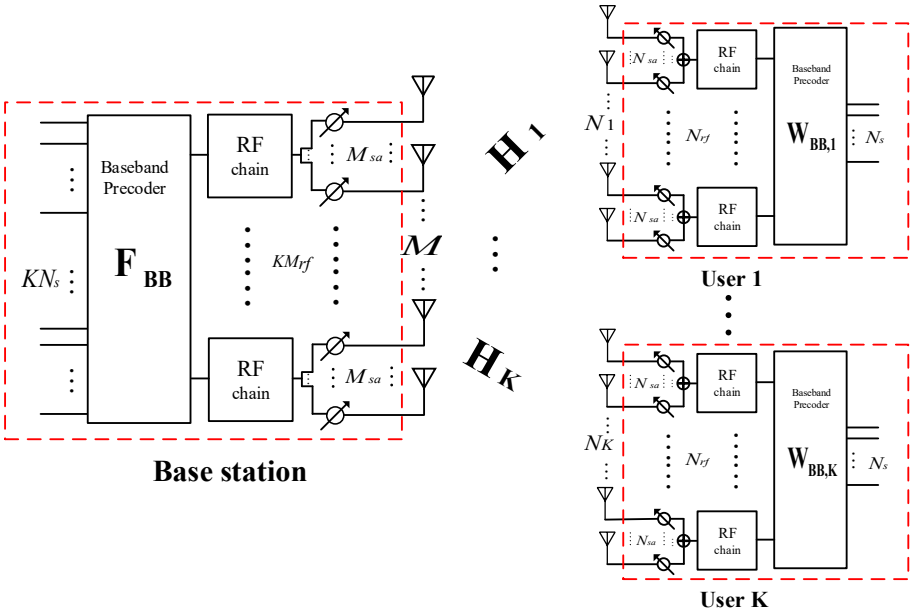


Figure 3.6: K-user MIMO broadcast channel with partially-connected RF structure.

**Partially-connected RF structures at all terminals**

In this setup, all the network's terminals (base station and $K$ users) consider the partially-connected RF structure where each RF chain is either connected to $M_{sa}$ antenna elements at base station or $N_{sa}$ antenna elements at users as depicted in Fig. 3.6.

Since the partially-connected RF structure is considered at all terminals, the analog precoder and combiners have block diagonal structure. This naturally divides $\mathbf{H}_{aug}$ into $K^2N_{rf}M_{rf}$ sub-blocks of size $N_{sa} \times M_{sa}$ where each block of $\tilde{\mathbf{H}}_{aug}$ in (3.30) is divided into $N_{rf}M_{rf}$ sub-blocks. This directly allows for separating the multi-user analog processing problem to $KN_{rf}$ independent single-user single stream analog processing problems. Particularly, we formalize the analog processing problems $n = 1, \cdots, KN_{rf}$ as:

$$\begin{aligned} &\underset{\mathbf{f}_{RF(n)},\mathbf{w}_{RF(n)}}{\text{maximize}} \quad |\mathbf{w}_{RF(n)}^{H}\mathbf{H}_{(n,n)}\mathbf{f}_{RF(n)}|^2 \\ &\text{subject to} \quad |\mathbf{f}_{RF(i,n)}| = \frac{1}{\sqrt{M_{sa}}}\forall i,\ |\mathbf{w}_{RF(j,n)}| = \frac{1}{\sqrt{N_{sa}}}\forall j, \end{aligned} \tag{3.32}$$

where $\mathbf{H}_{(n,n)}$ is the $n^{\text{th}}$ sub-block on the diagonal of the $\mathbf{H}_{aug}$ and $\mathbf{f}_{RF(n)}, \mathbf{w}_{RF(n)}$ are non-zero parts of the $n^{\text{th}}$ columns of the $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$, respectively. Fortunately, this problem has the same formulation of the analog processing of single-user sub-array structure in (3.18), thereby solving it using Algorithm 2 to obtain the beamforming pairs $(\mathbf{f}_{RF(n)}, \mathbf{w}_{RF(n)})\ \forall k$.

**Mixed partially and fully-connected RF structure**

In this subsection, we limit out discussion to the setup where the base station has partially-connected RF structure while users have fully-connected ones, as indicated in Fig. 3.7, since other setups come readily based on this setup.

The sub-array RF structure of the analog precoder divides each user's channel matrix into $M_{rf}$ sub-blocks of size $N \times M_{sa}$, i.e., $\mathbf{H}_k = [\mathbf{H}_{k(1)}, \mathbf{H}_{k(2)}, \cdots, \mathbf{H}_{k(M_{rf})}]$, or equivalently, divides each block of $\tilde{\mathbf{H}}_{aug}$ in (3.30) into $M_{rf}$ sub-blocks. Therefore, the $(k,k)$ block of $\tilde{\mathbf{H}}_{aug}$ in (3.30) can be written as:

$$\mathbf{W}_{RF,k}^{H}\mathbf{H}_k\mathbf{F}_{RF,k} = \begin{bmatrix} \mathbf{w}_{RF,k_{(1)}}^{H}\mathbf{H}_{k(1)}\mathbf{f}_{RF,k_{(1)}} & \cdots & \mathbf{w}_{RF,k_{(1)}}^{H}\mathbf{H}_{k(M_{rf})}\mathbf{f}_{RF,k_{(M_{rf})}} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{RF,k_{(N_{rf})}}^{H}\mathbf{H}_{k(1)}\mathbf{f}_{RF,k_{(1)}} & \cdots & \mathbf{w}_{RF,k_{(N_{rf})}}^{H}\mathbf{H}_{k(M_{rf})}\mathbf{f}_{RF,k_{(M_{rf})}} \end{bmatrix} \tag{3.33}$$

Similar to previous formulations, the analog processing problems are given by:

$$\begin{aligned} \underset{\mathbf{f}_{RF,k_{(n)}},\mathbf{w}_{RF,k_{(n)}}}{\text{maximize}} \quad & |\mathbf{w}_{RF,k_{(n)}}^{H}\mathbf{H}_{k(n)}\mathbf{f}_{RF,k_{(n)}}|^2 \\ \text{subject to} \quad & |\mathbf{f}_{RF,k_{(i,n)}}| = \frac{1}{\sqrt{M_{sa}}}\forall i,\ |\mathbf{w}_{RF,k_{(j,n)}}| = \frac{1}{\sqrt{N}}\forall j, \end{aligned} \tag{3.34}$$

Similar to problems in (3.18) and (3.32), this problem can be solved using Algorithm 2 to obtain the beamforming pairs $(\mathbf{f}_{RF,k_{(n)}}, \mathbf{w}_{RF,k_{(n)}})\forall n, k$.

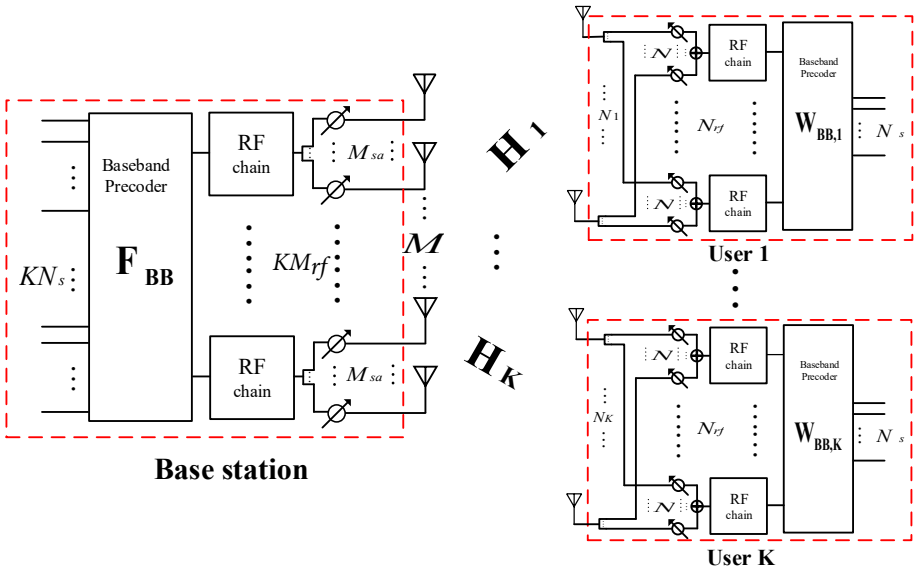


Figure 3.7: K-user MIMO broadcast channel with mixed partially and fully-connected RF structures.

### 3.3.3 Digital Processing

After the analog processing stage, the augmented matrix of the effective channels $\tilde{\mathbf{H}}_{aug}$ has much smaller dimensions compared to $\mathbf{H}_{aug}$. Moreover, digital processing has only power constraints where digital precoder and combiners have variable magnitude and phase beamforming weights (entries). Hence, the digital processing stage can be designed based on the conventional fully-digital schemes on traditional MIMO literature.

Given the analog precoder and combiners, the processed received signal is given as:

$$\tilde{\mathbf{s}}_k = \overbrace{\mathbf{W}_{BB,k}^H \tilde{\mathbf{H}}_k \mathbf{F}_{BB,k} \mathbf{s}_k}^{\text{Desired signal of user k + Inter-stream interference}} + \underbrace{\mathbf{W}_{BB,k}^H \tilde{\mathbf{H}}_k \sum_{l=1, l\neq k}^{K} \mathbf{F}_{BB,l} \mathbf{s}_l}_{\text{Inter-user interference}} + \underbrace{\mathbf{W}_{BB,k}^H \mathbf{W}_{RF,k}^H \mathbf{n}_k}_{\text{Gaussian noise at baseband}}, \tag{3.35}$$

This equation shows that in contrast to the single-user case where analog processing alleviates

the effect of inter-stream interference, it does not harness the inter-user interference (the main detrimental factor in multi-user case). This necessitates performing the digital processing in two stages that remove the inter-user interference and inter-stream interference successively while maximizing the sum spectral efficiencies of users.

Using the block-diagonalization strategy to completely remove the inter-user interfere, the digital precoder of user $k$, $\mathbf{F}_{BB,k}$ is designed to lie in the null spaces of other users' effective channels, i.e., $\tilde{\mathbf{H}}_i\mathbf{F}_{BB,k}, \forall i \neq k$. Let $\tilde{\mathbf{H}}_{k/} = [\tilde{\mathbf{H}}_1^T, \cdots, \tilde{\mathbf{H}}_{k-1}^T, \tilde{\mathbf{H}}_{k+1}^T, \cdots, \tilde{\mathbf{H}}_K^T]^T$, then $\mathbf{F}_{BB,k} \in$ Nul $\tilde{\mathbf{H}}_{k/}$. This comes readily based on the SVD of $\tilde{\mathbf{H}}_{k/} = \mathbf{U}_{k/}\mathbf{\Sigma}_{k/}[\mathbf{V}_{k/}^1, \mathbf{V}_{k/}^0]$ such that $\mathbf{F}_{BB,k} \in$ Col $\mathbf{V}_{k/}^{(0)}$ $\forall k$, i.e., $\mathbf{F}_{BB,k} = \mathbf{V}_{k/}^{(0)}\tilde{\mathbf{F}}_{BB,k}$ where $\mathbf{V}_{k/}^{(0)}$ is the collection of the singular vectors that are associated with zero singular values. When applying these solutions to $\tilde{\mathbf{H}}_{aug}$, they completely remove the inter-user interference such that the augmented matrix of the effective channel after the first digital precoding stage is a block diagonal matrix and given as:

$$\tilde{\tilde{\mathbf{H}}}_{aug} = \begin{bmatrix} \tilde{\mathbf{H}}_1^H\mathbf{V}_{1/}^{(0)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{H}}_2^H\mathbf{V}_{2/}^{(0)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{H}}_K^H\mathbf{V}_{K/}^{(0)} \end{bmatrix} = \begin{bmatrix} \tilde{\tilde{\mathbf{H}}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\tilde{\mathbf{H}}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\tilde{\mathbf{H}}}_K \end{bmatrix} \tag{3.36}$$

This converts the $K$-user MIMO broadcast channels in to $K$ parallel MIMO single-user channels with a joint total average power constraint. Therefore, the spectral efficiency maximization problem in (3.29) is finally simplified to:

$$\begin{aligned} \max_{\tilde{\mathbf{F}}_{BB,k}, \mathbf{W}_{BB,k}} \quad & \sum_{k=1}^{K} \log\det\left(\mathbf{I}_{N_s} + \left(\mathbf{W}_{BB,k}^H\mathbf{R}_{nRF}\mathbf{W}_{BB,k}\right)^{-1}\mathbf{W}_{BB,k}^H\tilde{\tilde{\mathbf{H}}}_k\tilde{\mathbf{F}}_{BB,k}\tilde{\mathbf{F}}_{BB,k}^H\tilde{\tilde{\mathbf{H}}}_k\mathbf{W}_{BB,k}\right), \\ \text{s.t.} \quad & \sum_{k} \text{trace}(\mathbf{F}_{RF}\mathbf{F}_{BB,k}\mathbf{F}_{BB,k}^H\mathbf{F}_{RF}^H) \leq p, \end{aligned} \tag{3.37}$$

where $\mathbf{R}_{nRF,k} = \sigma_n^2\mathbf{W}_{RF,k}^H\mathbf{W}_{RF,k}$. This problem has a closed-form optimal solutions as given in the following proposition.

**Proposition. 3.3.1.** *Let* $\overline{\tilde{\tilde{\mathbf{V}}}_{kN_s}}$ *is the* $N_s$ *largest eigenvectors of*

$$\big(\mathbf{V}_{k/}^{(0)^H}\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{V}_{k/}^{(0)}\big)^{-\frac{1}{2}}\tilde{\tilde{\mathbf{H}}}_k^H\mathbf{R}_{n_{RF}}^{-1}\tilde{\tilde{\mathbf{H}}}_k\big(\mathbf{V}_{k/}^{(0)^H}\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{V}_{k/}^{(0)}\big)^{-\frac{1}{2}} = \overline{\tilde{\tilde{\mathbf{V}}}_k}\,\overline{\tilde{\tilde{\mathbf{\Lambda}}}_k}\,\overline{\tilde{\tilde{\mathbf{V}}}_k}^H,$$

*then the optimal digital precoders and combiners that maximize (3.37) are given as:*

$$\begin{aligned}\tilde{\mathbf{F}}_{BB,k_\star} &= (\mathbf{V}_{k/}^{(0)^H}\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{V}_{k/}^{(0)})^{-\frac{1}{2}}\overline{\tilde{\tilde{\mathbf{V}}}_{kN_s}}\mathbf{\Gamma}_k\\ \mathbf{W}_{BB,k_\star} &= \mathbf{R}_{n_{RF,k}}^{-1}\tilde{\tilde{\mathbf{H}}}_k\big(\mathbf{V}_{k/}^{(0)^H}\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{V}_{k/}^{(0)}\big)^{-\frac{1}{2}}\overline{\tilde{\tilde{\mathbf{V}}}_{kN_s}}\mathbf{\Gamma}_k,\end{aligned} \tag{3.38}$$

*where*

$$\mathbf{\Gamma} = \begin{bmatrix}\mathbf{\Gamma}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{\Gamma}_2 & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Gamma}_K\end{bmatrix},$$

*is an* $KN_s \times KN_s$ *diagonal matrix that contain the power fractions of all transmitted streams on its diagonal such that its* $i^{th}$ *diagonal entries is given as:*

$$|\gamma_{ii}|^2 = \Big(\frac{p+\sum_k^{N_s}\lambda_{kk}^{-1}}{N_s} - \frac{1}{\lambda_{ii}}\Big)^+, i = 1, \cdots, KN_s, \tag{3.39}$$

*and* $\lambda_{ii}$ *is the* $i^{th}$ *diagonal entry of*

$$\mathbf{\Lambda} = \begin{bmatrix}\mathbf{\Lambda}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_2 & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Lambda}_K\end{bmatrix}.$$

*Proof.* The proof comes readily following the same approach that we used in the single user-case in Proposition 3.2.5. ■

## 3.4 Extensions to mmWave Wideband MIMO-OFDM channels

In contrast to the previous sections where we consider narrowband massive MIMO systems, here, we extend our investigation to the hybrid beamforming designs for wideband systems. Many envisioned applications, which utilize mmWave massive MIMO systems, are expected to operate over broadband channels which are typically characterized as frequency-selective channels [82]. This necessitates considering frequency-selectivity mitigation techniques such as orthogonal frequency division multiplexing (OFDM). Considering OFDM modulation results in a new challenge to the design of hybrid beamformers where common analog precoder and combiner (post-IFFT and pre-FFT processing, respectively) are shared across all subcarriers whereas digital precoders and combiners are different among subcarriers. With this consideration, the received discrete base band signal at subcarrier $u$ after OFDM modulation, demodulation and hybrid processing is represented by [7, 66, 100, 121, 145]:

$$\tilde{\mathbf{s}}[u] = \sqrt{\rho}\mathbf{W}_{BB}^{H}[u]\mathbf{W}_{RF}^{H}\mathbf{H}[u]\mathbf{F}_{RF}\mathbf{F}_{BB}[u]\mathbf{s}[u] + \mathbf{W}_{BB}^{H}[u]\mathbf{W}_{RF}^{H}\mathbf{n}[u], \tag{3.40}$$

where $u \in \{0, \cdots, U\}$, $U$ is the total number of subcarries and $\mathbf{H}[u]$ is the frequency domain channel matrix for the $u^{\text{th}}$ subcarrier. Following the same rationale of the proposed design approach for the narrowband hybrid processing in previous sections, the analog beamforming design in MIMO-OFDM system is formulated as:

$$\begin{aligned} \underset{\mathbf{F}_{RF},\mathbf{W}_{RF}}{\text{maximize}} \quad & \frac{1}{U}\sum_{u}^{U}\sum_{n=1}^{N_{rf}} |\mathbf{w}_n^{H}\mathbf{H}[u]\mathbf{f}_n|^2 \\ \text{subject to} \quad & \mathbf{F}_{RF} \in \mathbb{U}^{M\times N_{rf}}, \mathbf{W}_{RF} \in \mathbb{U}^{N\times N_{rf}} \end{aligned} \tag{3.41}$$

Obtaining optimal solutions for this problem is challenging since the problem is non convex even for narrowband scenario. By comparing the problem in (3.41) with that ones in (3.6) and

(3.17) for the narrowband scenarios, it is clear that they bear similarities except that the analog precoder and combiner have to be optimized commonly for all subcarriers. Tackling this difficulty, we derive a lower bound on the objective function in (3.41). The lower bound comes readily using Jensen's inequality as:

$$\frac{1}{U}\sum_{u}^{U}\sum_{n=1}^{N_{rf}}|\mathbf{w}_n^H\mathbf{H}[u]\mathbf{f}_n|^2 \overset{(a)}{\geq} \sum_{n=1}^{N_{rf}}|\mathbf{w}_n^H\tilde{\mathbf{H}}\mathbf{f}_n|^2 \tag{3.42}$$

where $\tilde{\mathbf{H}} = \frac{1}{U}\sum_{u}^{U}\mathbf{H}[k]$ and (3.42) is due to the convexity of $|\mathbf{w}_n^H\mathbf{H}[u]\mathbf{f}_n|^2$ in terms of $\mathbf{H}[u]$ such that for any convex function $f()$, if $\sum_i \beta_i = 1$, then $\sum_i \beta_i f(\mathbf{X}_i) \geq f(\sum_i \beta_i \mathbf{X}_i)$. Accordingly, we propose to design the analog for the frequency-selective channels such that it maximizes the lower bound in (3.42) and this gives

$$\begin{aligned} &\underset{\mathbf{F}_{RF},\mathbf{W}_{RF}}{\text{maximize}} \quad \sum_{n=1}^{N_{rf}}|\mathbf{w}_n^H\tilde{\mathbf{H}}\mathbf{f}_n|^2 \\ &\text{subject to} \quad \mathbf{F}_{RF}\in\mathbb{U}^{M\times N_{rf}}, \mathbf{W}_{RF}\in\mathbb{U}^{N\times N_{rf}} \end{aligned} \tag{3.43}$$

This optimization problem is exactly the same as the frequency-flat analog beamforming problem in (3.6) except that, here, $\tilde{\mathbf{H}} = \frac{1}{U}\sum_{u}^{U}\mathbf{H}[u]$. Consequently, the proposed techniques for the fully-connected and sub-array frequency-flat analog beamforming in Section 1.2 are applicable to those of frequency-selective analog beamforming, thereby employing Algorithm 1 and Algorithms 3.2.1 to calculate the fully-connected and subarray analog beamformer for the frequency selective scenario by passing $\tilde{\mathbf{H}}$ instead of $\mathbf{H}$ to their procedures. It is worth mentioning that transforming the analog precoding design problem for frequency-selective channels into analog precoding design problem for flat-fading channels has been considered in many prior works [7, 66, 100, 121, 145] and showed competitive performance under different metrics such as spectral efficiency [7, 121, 145] and MMSE [66, 145] based on utilizing a different transformation such as $\frac{1}{U}\sum_u \mathbf{H}[u]^H\mathbf{H}[u]$ [7, 66, 121, 145].

## 3.5 Computational Complexity Analysis

In this section, we provide comprehensive computational complexity analysis of the proposed algorithms, i.e., Algorithms 1, and 2 followed by the digital stage, and those of the most prominent prior works in Tables 3.2 and 3.3 in terms of the number of floating points operations (flops). We follow suit the approach of [40] in counting the number of flops. However, before proceeding to our computational complexity analysis, we discuss the complexity of computing the largest $N_s \ll \min(N, M)$ right and left singular vectors (known as truncated (partial) SVD). This is because that the truncated SVD is widely adopted in the literature on hybrid beamforming where it is a prerequisite step to the majority of hybrid solutions [82].

### 3.5.1 Computational Complexity of SVD

The canonical way of computing the truncated (partial) SVD is to compute the full SVD and then retain the first $k$ singular values and vectors. The full SVD of an $N \times M$ matrix is computed numerically by a two-step procedure where the matrix is first reduced to a bidiagonal matrix and then compute the SVD of the bidiagonal matrix. The first step requires $\mathcal{O}(M^2N)$ floating-point operations (flops) and the second one takes $\mathcal{O}(N^2)$ flops (assuming $M > N$). This makes the total cost of the full SVD is $\mathcal{O}(M^2N)$ flops. The exact computational complexity of calculating the full SVD depends on the utilized algorithm and the required parts of the SVD. For instance, the SVD requires $4MN^2 + 8N^3$ flops for computing $\mathbf{V}$ (required for obtaining the optimal beamformer) [40]. Extensive research efforts have been exerted aiming at providing less complex partial SVD. All these efforts furnish different sub-optimal solutions. These algorithms can be listed in two main approaches based on the nature of their algorithmic procedures, namely, deterministic and randomized approaches.

The majority of deterministic approximate partial SVD algorithms are obtained by computing partial QR factorization and postprocess the factors [58] such as truncated pivoted QR decomposition (variations of orthogonal triangularization by Householder transformation) [40] and rank revealing QR factorization [40, 42]. These techniques require $\mathcal{O}(MNk)$ flops. Effi-

cient algorithms, which are based on Krylov subspace methods such as Lanczos and Arnoldi techniques, can be used to compute partial SVD for large sparse matrices. Although these algorithms have computational complexity lower than or comparable to the partial QR-based ones, they are less robust and require complete reorthogonalization (or performing only local orthogonalization at every Lanczos steps) in order to maintain the orthogonality between the computed Lanczos vectors [17]. The most effective alternatives to the partial SVD among all these deterministic algorithms are the ones based on rank revealing QR (RRQR) factorization where their matrix approximations are almost as good as those derived from truncated full SVD ones [58]. The computational complexity of these algorithms is slightly larger than the pivoted QR-based and Krylov subspace algorithms. The exact computation complexity RRQR-based partial SVD in [42] is $4MNk + (3M^2 + 2N^2)k - 2(M + N)k + (3N + 4M)k^2 + 9k^3$. Several algorithms have been developed to compute RRQR and strong RRQR aiming at gaining more accuracy and robustness. The value of their approximation errors depends on the values of functions bounded by low-degree polynomials in $k$ and $N$. The accuracy and complexities of different algorithms are tabulated in [42].

On the other hand, several randomized algorithms compute the truncated SVD. These algorithms have been shown to provide approximations that have tolerable errors for some applications such as data mining and principal component analysis (PCA) [58]. However, the additive approximations' errors are prone to be arbitrarily large for the matrices with sufficiently large spectral and Frobenius norms. Particularly, the most accurate one has an approximation error as high as $11\sqrt{k(M+N)}\sigma_{k+1}$. Unfortunately, the accuracy of these schemes is sometimes inadequate in many applications [43]. The main idea of these randomized algorithms is to first construct a low-dimensional subspace that captures the action of the matrix then restricts the matrix to this subspace and compute a standard SVD of the reduced matrix. In this way, the computation is subdivided between two stages where, in the first, simple methods such as random columns/rows selection and random projection. The computational complexity of these techniques can be as low as $\mathcal{O}(MN\log(k))$. This comes at the price of higher approximation errors, especially where they have additive errors that could be arbitrarily large for matrices with sufficiently large $||\mathbf{A}||_F^2$ or $||\mathbf{A}||_2^2$ [43]. Generally speaking, the ones that are based on random

projection have smaller approximation errors than those that are based on random subset selection [43]. The main idea of random projection algorithms is to capture the subspace of the matrix being factorized, i.e., $\mathbf{A}$, by multiplying $\mathbf{A}$ by a random matrix $\mathbf{\Omega}$ where its entries are drawn from random distributions such as Gaussian, uniform, and Bernoulli. One prominent random projection scheme is the one utilizes the Gaussian random matrix $\mathbf{\Omega}$ developed in [43] which is called randomized SVD. The exact computational complexity of the randomized SVD is $2(2q+2)MNk + 4N^2k + (2M+10N)k^2 + \frac{25}{3}k^3$ flops for computing $q$ power iterations in large dimensions, one QR decomposition and one SVD in much smaller dimensions [43]. The average error in computing the partial SVD using this algorithm is bounded by [43]:

$$\mathbb{E}||\mathbf{A} - \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^*||_2 \leq \sigma_{k+1} + \left[1 + 4\sqrt{\frac{2\min\{M,N\}}{k-1}}\right]^{\frac{1}{2q+1}}\sigma_{k+1} \tag{3.44}$$

A comparison between the most prominent schemes is presented in Table 3.1 and Fig. 3.8. Although the truncated SVD is the most accurate low-rank approximation, its high computational complexity is a huge burden in high dimensions applications. On the other hand, the randomized SVD algorithms have the lowest computational complexity for calculating the truncated SVD. However, they may not perform well in some cases, especially, when the singular values of the matrix being factorized decay slowly [43]. Furthermore, due to their random nature, these methods are not exact and any error control they can offer, such as power iteration, is probabilistic with a non-negligible probability of failure in some cases [58]. In contrast, the RRQR-based partial SVD algorithm provides accurate approximation in comparison to the randomized schemes and tolerable computational complexity in comparison to the truncated SVD. Therefore, we recommend using the RRQR-based partial SVD algorithm in massive MIMO applications in order to reap both gains, i.e., high accuracy and medium complexity. However, in our complexity analysis, we adopt the computational complexity cost of the partial SVD based on a randomized approach [43]. This means that we give the advantage to the prior works of hybrid beamforming schemes over the proposed ones since calculating the partial SVD based on the randomized approach has the lowest computation complexity. We note that the calculating partial SVD is an essential step for the majority of hybrid beamforming techniques and we totally avoid it in the

proposed schemes.

| Scheme | Pros | Cons | Number of flops | Example $16 \times 64$, $N_s = 3$ |
|---|---|---|---|---|
| Truncated SVD [40] | Exact and Robust | high complexity | $4m^2n + 8mn^2 + 9n^3$ | 430080 |
| RRQR-based [58] | very accurate | medium complexity | $4MNk + (3M^2 + 2N^2)k - 2(M+N)k$ $+(3N + 4M)k^2 + 9k^3$ | 53187 |
| Randomized [43] | low complexity | low accuracy | $2(2q+2)MNk + 4N^2k$ $+(2M + 10N)k^2 + \frac{23}{3}k^3$ | 41583 |

Table 3.1: Computational complexity comparison of partial (trunctated) SVD algorithms where $k = N_s$ and $q = 2$.

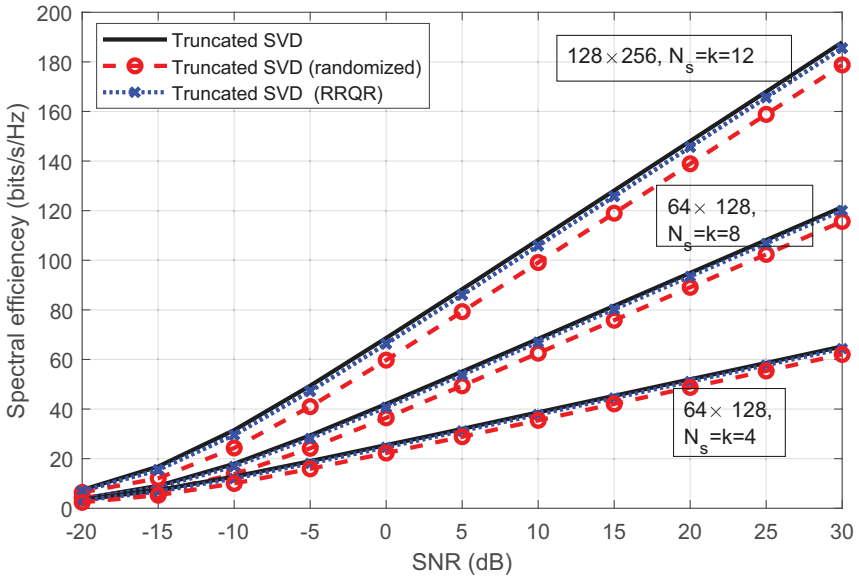


Figure 3.8: Spectral efficiency comparison of different truncated SVD algorithms

### 3.5.2 Computational Complexity of Hybrid Beamfroming Schemes

In our analysis, we assume real arithmetic operations require 1 flop while complex addition and multiplication require 2 and 6 flops [40, p. 36], respectively. Moreover, we assume that the complex versions of matrix computation algorithms require about four times the number of flops as the real ones since the complex versions are fairly straightforward driven from the real

ones and involve approximately as many complex multiplications as additions [40, p. 256]. For instance, the SVD of an $N \times M$ dense matrix requires $4M^2N+8MN^2+9N^3$ flops for computing $\mathbf{U}$, $\sigma$, and $\mathbf{V}$, and consequently, its complex version executed in $16M^2N+32MN^2+36N^3$ flops [40, p. 243]. In our computational complexity analysis, we consider the following calculations:

- An approximate truncated (partial) SVD of complex large dimensions low rank $N \times M$ matrix requires $48MNN_s + 16N^2N_s - (8M + 6N)N_s + (24M + 32N)N_s^2 + \frac{92}{3}N_s^3$ flops using the randomized SVD algorithm developed in [43]. This calculation is made exact by counting the complex computations of randomized SVD procedure [43, p. 9]. Although, this algorithm is less accurate than classical methods of truncating deterministic SVD algorithms, its computational complexity is much smaller than the ones of classical methods [43, 58]. However, this gives the advantage to the prior works that requires the SVD of the massive MIMO channel as a perquisite step for their hybrid beamforming procedures [15, 34, 66, 84, 90, 106, 120, 145, 148].

- Multiplication of $N \times M$ and $M \times N_{rf}$ complex matrices requires $(8M - 2)NN_{rf}$ flops.

- Full SVD of complex low dimensions dense $N_{rf} \times M_{rf}$ matrix is executed in $16M_{rf}^2N_{rf} + 32M_{rf}N_{rf}^2 + 36N_{rf}^3$ flops [40, p. 439].

- The Moore-Penrose pseudo inverse and the inverse of complex matrices are calculated based of QR decomposition such that $\mathbf{A}^\dagger = \mathbf{R}^{-1}\mathbf{Q}^H$. The Householder QR of a $N \times M$ complex matrix requires at least $8MN^2 - \frac{8}{3}N^3$ for computing $\mathbf{R}$ and $16MNN_s - 8(M + Ns)N^2 + \frac{16}{3}N^3$ for computing $N_s$ orthonormal columns from $\mathbf{Q}$ (thin QR) [40].

Based on these computations, we calculate and list out the complexity of computing the main steps of many hybrid beamforming schemes in Tables 3.2 and 3.3. These comparison tables give deeper understanding of the computations that dominate the complexities of the hybrid beamforming schemes.

Fig. 3.9 shows the asymptotic behavior of the computation complexity of the state-of-the-art schemes for the fully-connected structure, depicted in Table 3.2, along with the proposed scheme,

| Scheme | Operation | | Number of flops | $16 \times 64, N_s = 4$ $K = N_{rf} = 6$ |
|---|---|---|---|---|
| Proposed scheme fully-connected | Preprocessing | N/A | 0 | 84128 |
| | Analog TX | Step 2 Algorithm 1 | $6MN_{rf}$ | |
| | Analog RX | Step 3 Algorithm 1 | $M(8N-2)N_{rf} + 6NN_{rf}$ | |
| | Digital | SVD($\mathbf{H}$) | $16M_{rf}^2N_{rf} + 32M_{rf}N_{rf}^2 + 36N_{rf}^3$ | |
| Algorithm [84] | Preprocessing | SVD($\mathbf{H}$) | $48MNN_s + 16N^2N_s - (8M+6N)N_s$ $+(24M+32N)N_s^2 + \frac{92N_s^3}{3}$ | 299450 |
| | Analog | arctan & Multi. | $12MM_{rf} + 12NN_{rf}$ | |
| | Digital | SVD($\mathbf{H}$) | $16M_{rf}^2N_{rf} + 32M_{rf}N_{rf}^2 + 36N_{rf}^3$ | |
| Algorithm [145] | Pre-processing | SVD($\mathbf{H}$) | $48MNN_s + 16N^2N_s - (8M+6N)N_s$ $+(24M+32N)N_s^2 + \frac{92N_s^3}{3}$ | 459050 |
| | Analog | arctan & Mat. Multi. | $K\big(MN_{rf}(8M_s-2) + 6MM_{rf}$ $NN_{rf}(8N_s-2) + 6NN_{rf}\big)$ | |
| | Digital TX | SVD($\mathbf{F}_{opt}^H\mathbf{F}_{RF}$) | $K\big(16M_{rf}^2N_s + 32M_{rf}N_s^2 + 36N_s^3\big)$ | |
| | Digital RX | SVD($\mathbf{W}_{opt}^H\mathbf{W}_{RF}$) | $K\big(16N_{rf}^2N_s + 32N_{rf}N_s^2 + 36N_s^3\big)$ | |
| | Post-processing | Normalization | $MNs(8M_{rf}-2) + NNs(8N_{rf}-2)$ | |
| Algorithm [15] | Preprocessing | SVD($\mathbf{H}$) | $48MNN_s + 16N^2N_s - (8M+6N)N_s$ $+(24M+32N)N_s^2 + \frac{92N_s^3}{3}$ | 1064800 |
| | Analog | Mat. Multi. | $K\big((8N-2)LN_s + (8M-2)LN_s\big)$ | |
| | | sorting | $K\big(M_{rf}\log M + N_{rf}\log N\big)$ | |
| | Digital | Pseudo-inverse | $K\big(24MM_{rf}^2 - 2M_{rf}^2 - \frac{12M_{rf}^3}{3}$ $+24NN_{rf}^2 - 2N_{rf}^2 + \frac{12N_{rf}^3}{3}\big)$ | |
| | | Normalization | $K\big(MNs(8M_{rf}-2) + NNs(8N_{rf}-2)\big)$ | |
| Algorithm [90] | Preprocessing | SVD($\mathbf{H}$) | $48MNN_s + 16N^2N_s - (8M+6N)N_s$ $+(24M+32N)N_s^2 + \frac{92N_s^3}{3}$ | 823750 |
| | Analog | Convex QP | $K\big(\frac{4M_{rf}^3}{3}M + \frac{4N_{rf}^3}{3}N\big)$ | |
| | Digital | Least Squares | $K\big(24MM_{rf}^2 - 2M_{rf}^2 - \frac{12M_{rf}^3}{3}$ $+24NN_{rf}^2 - 2N_{rf}^2 + \frac{12N_{rf}^3}{3}\big)$ | |
| | Postprocessing | Normalization | $MNs(8M_{rf}-2) + NNs(8N_{rf}-2)$ | |
| Algorithm [106] | Preprocessing | SVD($\mathbf{H}$) | $48MNN_s + 16N^2N_s - (8M+6N)N_s$ $+(24M+32N)N_s^2 + \frac{92N_s^3}{3}$ | 804040 |
| | Analog | arctan & Mat. Multi. | $K\big(MN_s(8M_{rf}-2) + 6MM_{rf}$ $+NN_s(8N_{rf}-2) + 6NN_{rf}\big)$ | |
| | | Mat. Inverse | $K\big(\frac{32M_{rf}^3}{3} + \frac{32N_{rf}^3}{3}\big)$ | |
| | Digital | Least Squares | $K\big(24MM_{rf}^2 - 2M_{rf}^2 - \frac{12M_{rf}^3}{3}$ $+24NN_{rf}^2 - 2N_{rf}^2 + \frac{12N_{rf}^3}{3}\big)$ | |

Table 3.2: Computational complexity comparison of different hybrid processing schemes with fully-connected antenna structures

| Scheme | Operation | | Number of flops | $32 \times 128, N_s = 6$ $K = N_{rf} = 8$ |
|---|---|---|---|---|
| Proposed scheme sub-array | Pre-processing | N/A | 0 | 111360 |
| | Analog TX | Step 5 Algorithm 2 | $K\big(M_{sa}(8N_{sa}-2)M_{rf}+6M_{sa}M_{rf}\big)$ | |
| | Analog RX | Step 6 Algorithm 2 | $K\big(N_{sa}(8M_{sa}-2)N_{rf}+6N_{sa}N_{rf}\big)$ | |
| | Digital | SVD($\tilde{\mathbf{H}}$) | $16M_{rf}^2N_{rf}+32M_{rf}N_{rf}^2+36N_{rf}^3$ | |
| Algorithm [34] [148] | Pre-processing | N/A | 0 | 588646 |
| | Analog | Mat. Inversion | $K\big(\frac{32M_{rf}^3}{3}+\frac{32N_{rf}^3}{3}\big)$ | |
| | | Mat. Multiplication | $K\big(2MM_{rf}(8N_{rf}-2)+2NM_{rf}(8N_{rf}-2)\big)$ | |
| | | partial SVD | $K(16M_{sa}N_{sa}^2+23N_{sa}^2+16N_{sa}M_{sa}^2)$ | |
| | Digital | Full SVD($\tilde{\mathbf{H}}$) | $16M_{rf}^2N_{rf}+32M_{rf}N_{rf}^2+36N_{rf}^3$ | |
| Algorithm [145] | Pre-processing | SVD ($\mathbf{H}$) | $48MNN_s+16N^2N_s-(8M+6N)N_s$ $+(24M+32N)N_s^2+\frac{92N_s^3}{3}$ | 18869924 |
| | Analog | Mat. Multi. | $K\big(MN_s(8M_{rf}-2)+NN_s(8N_{rf}-2)\big)$ | |
| | Digital | SDR | $K\big(\frac{16}{3}M_{rf}^3N_{rf}+8N_{rf}^3M_{rf}\big)$ [75] | |

Table 3.3: Complexity comparison of hybrid processing schemes with sub-array structures.

shown in red dotted lines with circles. Fig. 3.9 illustrates that the computational complexity of the proposed algorithm for the fully-connected structure is significantly lower than the ones of prior art schemes. For instance, it is only $20\%$ of the computational complexity of the one that has the lowest computational complexity reported in the literature which is shown in yellow dotted line with triangles.

Fig. 3.10 shows the asymptotic behavior of the computational complexity of the proposed scheme for sub-array structures, shown in blue dotted lines with circles, in comparison with the ones of the prior art schemes that are depicted in Table 3.3. It is clear that the complexity of the proposed scheme is significantly lower compared to the state-of-the-art schemes in [145] and [34, 148] (less than $20$ and $50$ times, respectively).

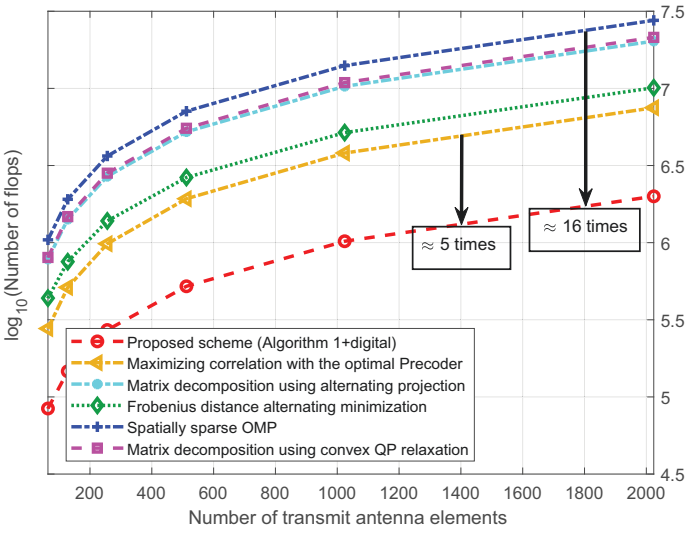


Figure 3.9: Asymptotic Computational complexity comparison of hybrid beamforming schemes with fully-connected RF structures

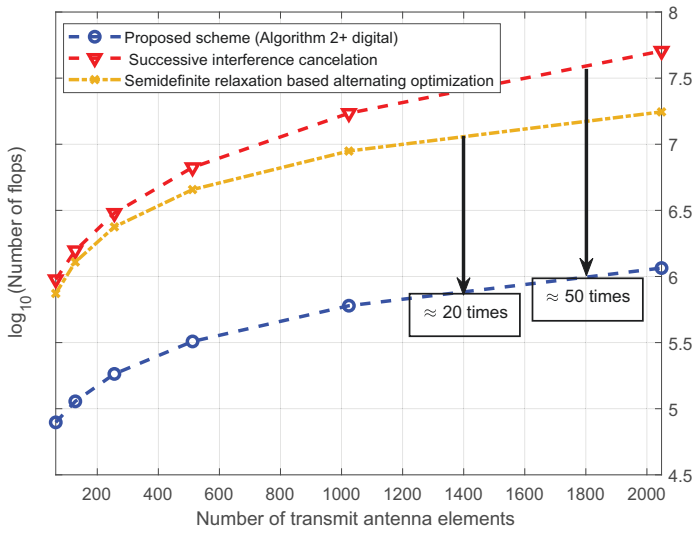


Figure 3.10: Asymptotic Computational complexity comparison of hybrid beamforming schemes with Partially-connected RF structures

Finally, we note that the computation complexity of the $K$-user hybrid processing techniques

with different RF structures is calculated based on the ones of the single-user cases. Particularly, the complexity of the analog processing in the multi-user case is $K$ complexity of Algorithm 1 or Algorithm 2 depending on the used RF structure. On the other hand, the complexity of digital processing in the multi-user case is $K$ times the complexity of single-user ones in addition to the complexity of $K$ SVD of a $(K-1)N_{rf} \times M_{rf}$ matrix. We also note that the computation complexity of wideband hybrid processing techniques with $U$-point OFDM is calculated based on the ones of narrowband hybrid beamforming schemes. Particularly, the complexity of the analog processing in wideband scenarios is the complexity of the analog processing in narrowband scenarios in addition to the complexity of computing $\frac{1}{U}\sum_u \mathbf{H}[u]^H\mathbf{H}[u]$. On the other hand, the complexity of digital processing in wideband scenarios is $U$ times the complexity in narrowband ones.

## 3.6 Simulation Results

In this section, we evaluate the performance of the proposed schemes presented in the aforementioned section in terms of the spectral efficiency. All the plotted curves in Fig. 3.11 - Fig. 3.16 are the spectral efficiency averaged over $10^5$ channel realizations. In these simulations, we consider both propagation environments discussed in Section 1.1. For the sparse scattering environment, we consider $M$-element and $N$-element uniform linear arrays (ULA) at both the transmitter and receiver, respectively. Moreover, we assume that the angles of departure and arrival are i.i.d. randomly Laplacian distributed. The means of the angles of departure and arrival are assumed to be uniformly-distributed between $0$ and $2\pi$, and the angular spreads are assumed to be constant within the cluster and has values of $7.5^\circ$ [15, 84, 90]. Further, we set the number of iterations equal to the number of RF chains in all the iterative algorithms in [34, 90, 106, 120, 145] to be fairly compared with compressed sensing based schemes such as the spatially sparse algorithm in [15] which has fixed number of iterations equals to $N_{rf}$. Furthermore, the initial points for all iterative schemes have been chosen randomly, where the phases of the entries are uniformly distributed. As a result, the spectral efficiency of some iterative hybrid processing schemes might

be slightly less than what has been reported in the literature.

In Figs. 3.11 and 3.12, we evaluate the spectral efficiency of five different processing schemes with fully-connected antenna array structure; namely, the optimal SVD-based fully-digital solution, shown in black solid line, the proposed scheme, shown in red dotted line with circles, spatially sparse hybrid precoding based on OMP [15], shown in blue dotted line with stars, iterative spectral efficiency maximization [120], shown in violet dotted lines with squares, Frobenius distance alternating minimization [145], shown in green dotted line with rhombuses, and lastly maximizing correlation with the optimal precoder [84], shown in yellow dotted line with triangles.

The spectral efficiency performance of Algorithm 1 is compared in Fig. 3.11 against the most prominent hybrid processing schemes and optimal fully-digital scheme when $N_s = 2$ and $4$ data streams are transmitted over $64 \times 16$ Rayleigh MIMO system with $N_{rf} = 3$ and $6$ RF chains, respectively. Fig. 3.11 shows that the spectral efficiency of the proposed scheme comes very close to those of prior works and with a marginal SNR gap to the optimal scheme. Moreover, it outperforms the scheme which is based on maximizing correlation with optimal precoder [84]. Despite increasing the number of RF chains and the number of the data streams, i.e., $6$ and $4$, respectively, the spectral efficiency achieved by the proposed scheme is almost doubled while the SNR gap is marginally increased. This marginal increase in the SNR gap to the optimal precoder can be compensated by using more RF chains. We note here that the spatially sparse hybrid processing scheme [15] is excluded from this figure since it is mainly developed to work in sparse environment such as mmWave channels where it suffers from significant spectral efficiency loss over rich scattering environments [84, 90]. Moreover, the hybrid processing schemes in [90] and [106] are excluded from Figs. 3.11 and 3.12 for the sake of clarity and better presentation since their spectral efficiency curves come on top of the other.

In Fig. 3.12, the spectral efficiency of all the hybrid processing schemes is evaluated over a $64 \times 16$ millimeter wave channel. Fig. 3.12 shows that there is a slight degradation in the spectral efficiency gap between the proposed scheme and the optimal scheme in comparison to Fig. 3.11. This degradation is expected since the correlation of the spatially sparse channel, simulated in

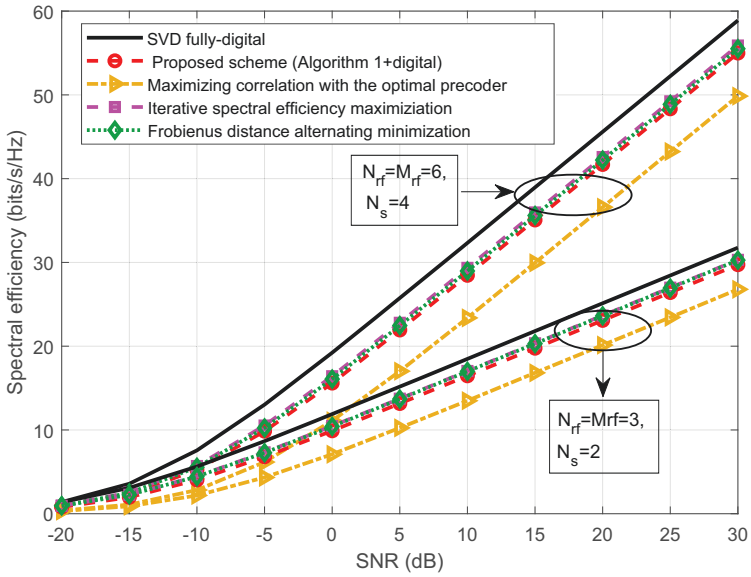


Figure 3.11: Spectral efficiency comparison of fully-connected hybrid processing schemes over $64 \times 16$ Rayleigh fading channel

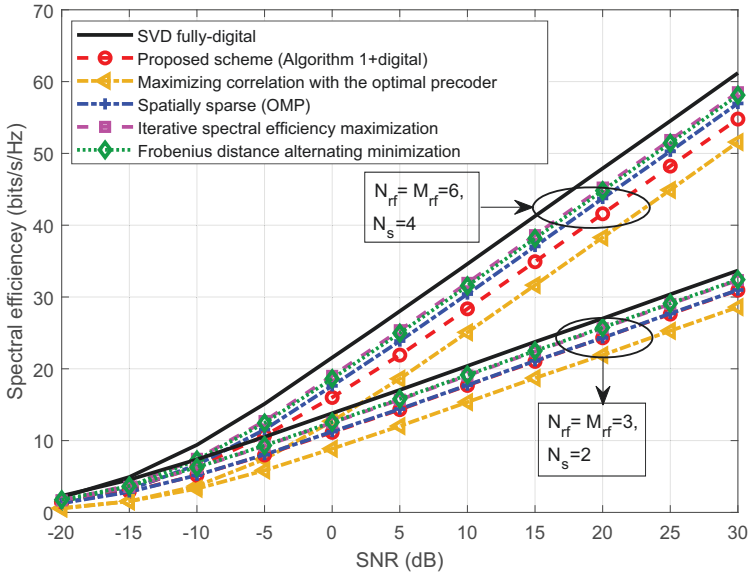


Figure 3.12: Spectral efficiency comparison of fully-connected hybrid processing schemes over $64 \times 16$ mmWave with $N_{cl} = 8$, $N_{ray} = 5$

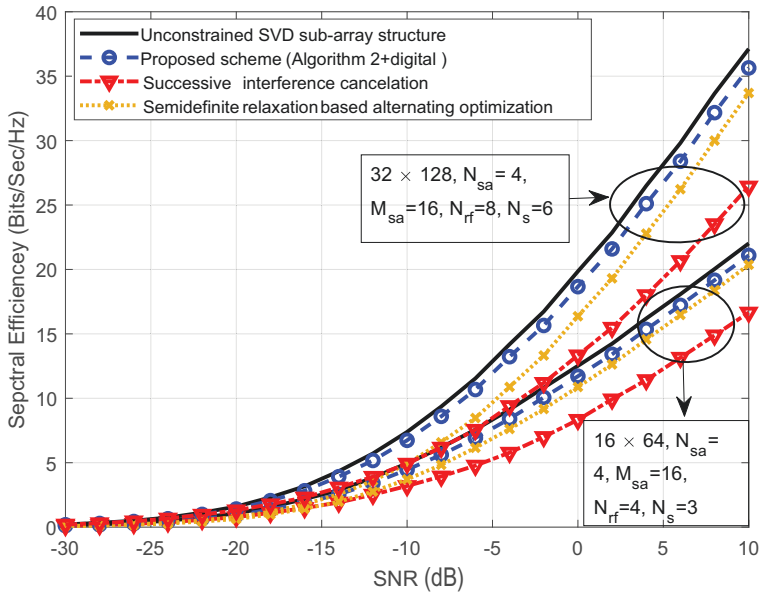


Figure 3.13: Spectral efficiency comparison of partially-connected hybrid processing schemes over mmWave with $N_{cl} = 4$, $N_{ray} = 5$

Fig. 3.12, degenerates the hardening property of massive MIMO channels (stated in Remark 1).

In Fig. 3.13 and 3.14, we evaluate the spectral efficiency of four different precoding and combining schemes with sub-array antenna structure; namely, the optimal unconstrained SVD based hybrid precoding and combining, shown in black solid line, the proposed scheme, shown in blue dashed line with circles, successive interference cancellation [34, 148], shown in red dotted line with triangles, semidefinite relaxation based alternating optimization, shown in yellow dotted lines with dots [145]. Fig. 3.13 shows that our proposed scheme, i.e., Algorithm 2 followed by the digital stage, outperforms the ones based on successive interference cancellation [34, 148], semidefinite relaxation and alternating optimization [145]. Fig. 3.14 illustrates that the proposed scheme maintains its superior performance over the Rayleigh channel. Moreover, the spectral efficiency loss of the proposed scheme due to the hardware constraints over both environments is marginal compared to the unconstrained SVD solution (less then 1 dB).

In Fig. 3.15, we evaluate the spectral efficiency of the proposed scheme when the number

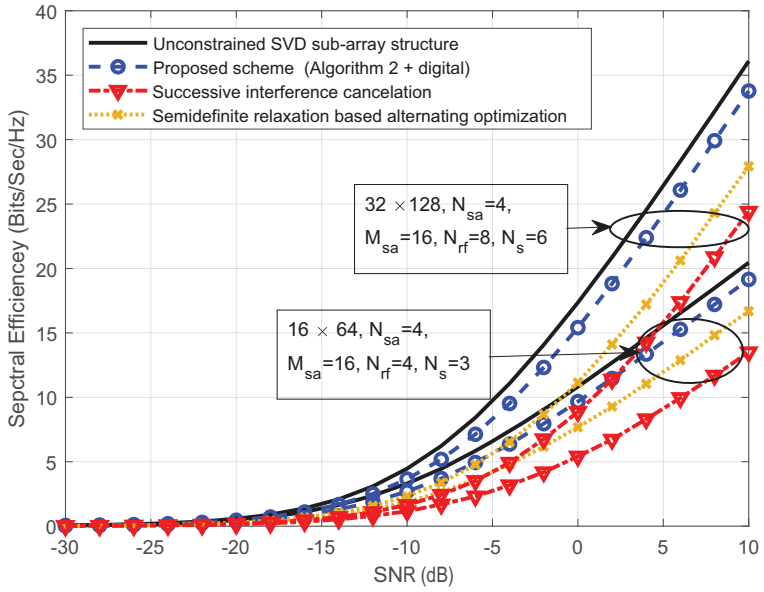


Figure 3.14: Spectral efficiency comparison of partially-connected hybrid processing schemes over Rayleigh fading channel

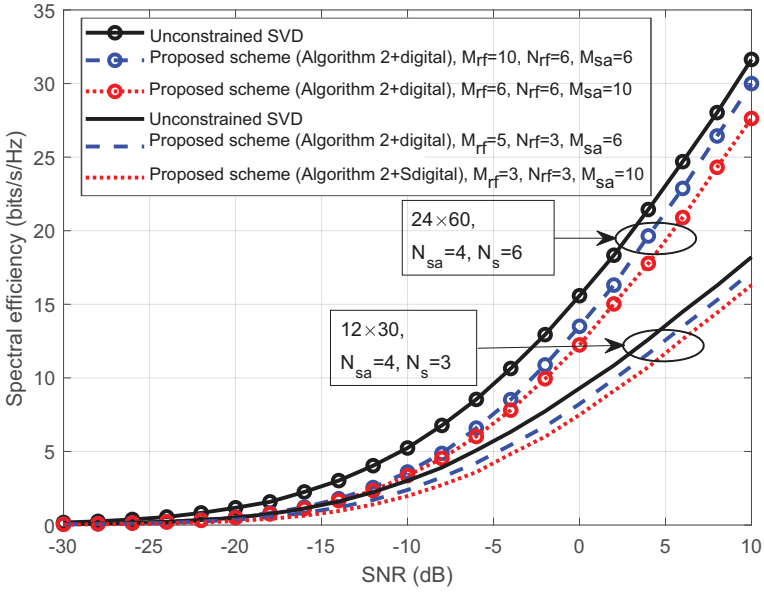


Figure 3.15: Studying the effect of increasing the number of RF chains at one side, i.e., $M_{rf} > N_{rf}$, over a millimeter-wave channel with $N_{cl} = 5$, $N_{ray} = 6$ and

of RF chains at both sides are different (without loss of generality $M_{rf} > N_{rf}$). Fig. 3.15 illustrates that there is an improvement in the spectral efficiency of the proposed scheme when $M_{rf} > N_{rf}$ compared to $M_{rf} = N_{rf}$, particularly, there is a $0.5$ dB gain in the spectral efficiency when increasing $M_{rf}$ from 3 to 5 (shown in dotted red and dashed blue, respectively). However, this improvement is more pronounced at higher dimensions, specifically, more than 1 dB gain (shown in dotted red and dashed blue with circles). We note that, in Fig. 3.15 where $M > N$, i.e., $M = 60$ or 30, and $N = 24$ or 12, we assume that $M_{rf} > N_{rf}$ since it is more reasonable to assume that the side that has larger number of antenna elements has larger number of RF chains not the reverse.

**Remark. 3.6.1.** ***Effect of imperfect channel state information:*** *In practical scenarios, there are some channel errors due to channel estimation and/or feedback (quantization) errors. Therefore, it is important to evaluate the sensitivity of the proposed schemes to channel errors. Since channel estimation and feedback techniques are not discussed in this work, we assume a general framework of channel estimation/quantization errors for the simulation. In particular, we assume that the channel errors are additive i.i.d $\mathcal{CN}(0, \sigma_e^2)$ where $\sigma_e^2$ is the variance of the channel errors [79]. We note here that this model captures the effect of the imperfect channel knowledge at both the transmitter and receiver since we jointly design the hybrid precoder and combiner based on the erroneous (estimated) channel $\hat{\mathbf{H}} = \mathbf{H} - \mathbf{E}$ where $\mathbf{E}$ is the error matrix. In Fig. 3.16, we show the spectral efficiency of the unconstrained SVD scheme and Algorithm 2 while assuming imperfect channel state information with different error variances, i.e., $\sigma_e^2 = 0.05$ and $0.1$, respectively. The solid lines represent the spectral efficiency of both schemes with accurate channel state information while the dash lines and dotted lines present the corresponding performance with imperfect channel state information with error variances $\sigma_e^2 = 0.05$ and $0.1$, respectively. As expected, the spectral efficiency performance of both schemes degrades as the variance of channel errors increases. However, the spectral efficiency loss of Algorithm 2 is approximately equal to the one of the unconstrained SVD scheme in both $32 \times 128$ and $16 \times 64$ mmWave channels, with different channel errors' variances. This implies that most of conventional error mitigation techniques can be applied to the proposed schemes.*

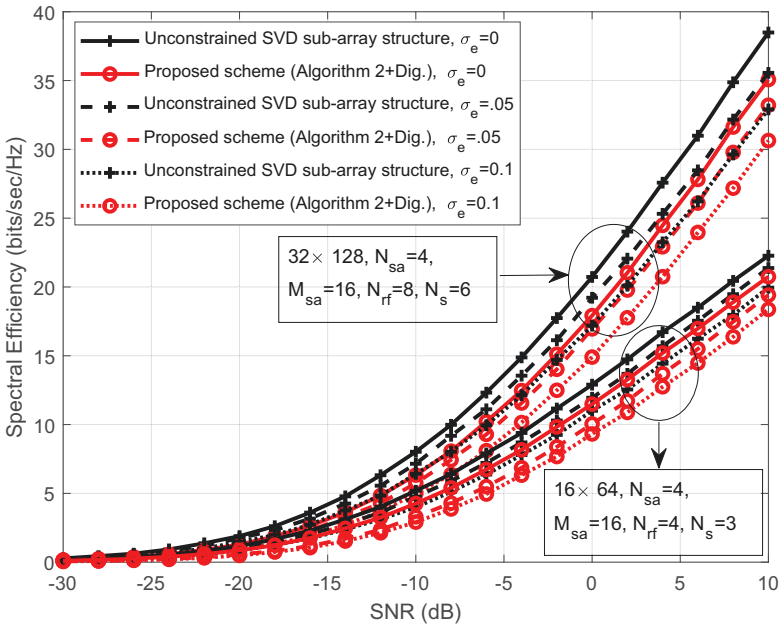


Figure 3.16: The effect of channel errors on the spectral efficiency of Algorithm 2

Fig. 3.17 shows that the proposed approach achieves high spectral efficiency in the multi-user case while maintaining low computational complexity in comparison to prior works in [66, 88].

Fig. 3.18 shows that the proposed approach maintains high spectral efficiency performance over wideband frequency selective channels in comparison to prior works in [66, 121, 145].

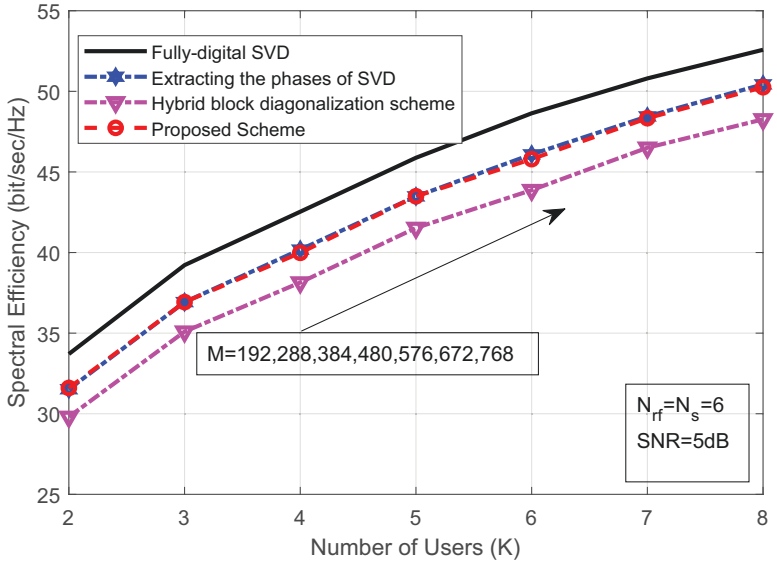


Figure 3.17: Spectral efficiency comparison of hybrid processing schemes for K-user MIMO Broadcast channel over mmWave channels with $N_{cl} = 4, N_{ray} = 5$.

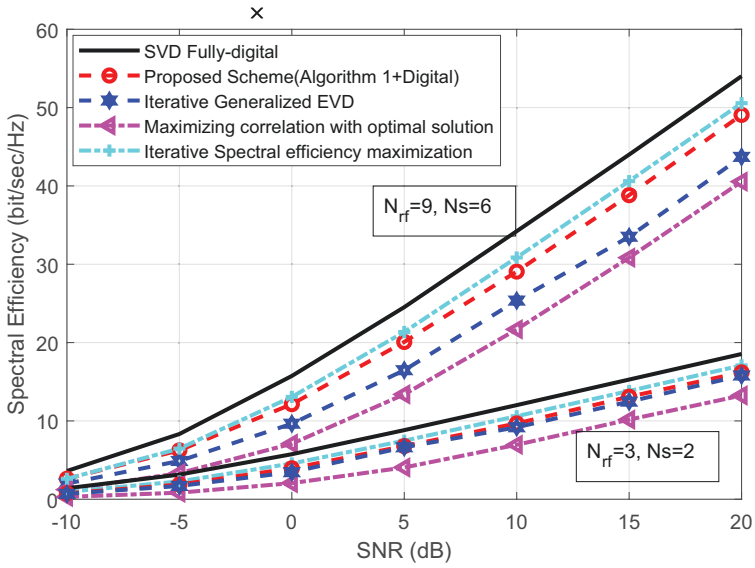


Figure 3.18: Spectral efficiency comparison of hybrid processing schemes for wideband Rayleigh fading channels

# Chapter 4

# Hybrid Beamforming Schemes for FDD Systems

## 4.1 System Model

We consider a massive MIMO FDD system with a limited number of RF chains. In this system, we present two communication scenarios, namely, single-user and multi-user massive MIMO scenarios. In the single-user scenario, the transmitter and the receiver are equipped with $M$-transmit and $N$-receive antenna elements. This point-to-point massive MIMO model is of interest due to its wide-range applicability in many recent applications such as high definition video streaming, virtual-reality/augmented-reality, connected cars and links between base stations [20]. The transmitter sends $N_s$ independent data streams to the receiver with the help of $N_{rf} \ll \min(M, N)$ RF chains. Due to the use of a limited number of RF chains, the hybrid beamforming structure is considered. In particular, we assume that the data vector, $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$, is pre-processed by two different precoding matrices. As a result, the received signal is given by:

$$\mathbf{r} = \sqrt{\rho}\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{n}, \tag{4.1}$$

where $\mathbf{F}_{BB} \in \mathbb{C}^{N_{rf} \times N_s}$ is the baseband precoder and $\mathbf{F}_{RF} \in \mathbb{C}^{M \times N_{rf}}$ is the RF one. $\rho$ is the average signal to noise ratio, and $\mathbf{n}$ is the additive white Gaussian noise vector, with i.i.d entries $\sim \mathcal{CN}(0, \sigma_n^2)$. Generally, the fading channel matrix, i.e., $\mathbf{H} \in \mathbb{C}^{N \times M}$ in mmWave bands, is spatially correlated due to the usage of large-scale and densely-packed phased antenna arrays at both sides [3, 8, 15, 95] and sparse in the angular domain due to the limited scattering nature of mmWave bands [15, 49]. This is modeled mathematically in literature on mmWave channels by either the Kronecker or the clustered channel models. The Kronecker correlation model describes the stochastic spatial correlation evolution of each channel realization. As a result, the channel is given by [2, 3]:

$$\mathbf{H}_{Kr} = \mathbf{R}_r^{1/2} \mathbf{H}_w \mathbf{R}_t^{T/2}, \tag{4.2}$$

where $\mathbf{R}_r = \mathbb{E}[\mathbf{H}_{Kr} \mathbf{H}_{Kr}^H]$ is the receive correlation matrix, $\mathbf{R}_t = \mathbb{E}[\mathbf{H}_{Kr}^H \mathbf{H}_{Kr}]$ is the transmit correlation matrix and $\mathbf{H}_w \in \mathbb{C}^{N \times M}$ has entries are i.i.d. Gaussian distributed. On the other hand, the clustered channel model expresses the channel as a function of its spatial parameters where the channel matrix is given as [8, 15]:

$$\mathbf{H}_{Cl} = \sqrt{\frac{MN}{N_{cl} N_{ray}}} \sum_{i,l}^{N_{cl} N_{ray}} \alpha_{il} \mathbf{a}_r(\theta_{il}^r) \mathbf{a}_t^H(\theta_{il}^t), \tag{4.3}$$

$$= \mathbf{A}_r \mathbf{G} \mathbf{A}_t^H, \tag{4.4}$$

where $\alpha_{il}$ denotes the complex gains of the $l^{\text{th}}$ ray in the $i^{\text{th}}$ cluster, with $N_{cl}$ clusters, each contributing $N_{ray}$ rays such that the total number of rays/paths is $P = N_{cl} N_{ray}$. The vector $\mathbf{a}_t(\theta_{il}^t)$ is the transmit antenna array response vector of length $M$ for a given angle of departure $\theta_{il}^t$, and $\mathbf{a}_r(\theta_{il}^r)$ is the receive antenna array response vector of length $N$ for a given angle of arrival $\theta_{il}^r$. Eq.(4.4) is the augmented matrix representation of the clustered channel model where $\mathbf{A}_t$ and $\mathbf{A}_r$ are the augmented transmit and receive antenna array response matrices and $\mathbf{G}$ is a diagonal channel contains the normalized complex gains $\alpha_{il} \sqrt{\frac{MN}{N_{cl} N_{ray}}}$. Although the two models are different, they bear some similar statistical properties under certain propagation assumptions.

Mainly, the transmit and the receive correlation matrices are separable where $\mathbb{E}\{\mathbf{H}_{Kr}^H\mathbf{H}_{Kr}\} = \mathbf{R}_t$ and $\mathbb{E}\{\mathbf{H}_{Kr}\mathbf{H}_{Kr}^H\} = \mathbf{R}_r$, and $\mathbb{E}\{\mathbf{H}_{Cl}^H\mathbf{H}_{Cl}\} = \mathbb{E}_{\mathbf{A}_t}\{\mathbf{A}_t\mathbf{A}_t^H\}$ and $\mathbb{E}\{\mathbf{H}_{Cl}\mathbf{H}_{Cl}^H\} = \mathbb{E}_{\mathbf{A}_r}\{\mathbf{A}_r\mathbf{A}_r^H\}$ for both the Kronecker and the clustered channel models, respectively under the uncorrelated scattering (channel gain paths, angles of departure and arrival are mutually independent) and equal power of channel paths assumptions (typical assumptions on the literature of mmWave channels [2, 3, 15, 95]). Given these assumptions, both models are being used interchangeably in hybrid precoding over mmWave channels literature where the Kronecker model is preferred in theoretical analysis (especially statistical ones) and the clustered model is utilized in numerical simulations by setting $\mathbf{R}_t = \mathbb{E}_{\mathbf{A}_t}\{\mathbf{A}_t\mathbf{A}_t^H\}$ (see [2, 95]). We follow the same approach and hereafter we drop the subscript of $\mathbf{H}$ for the sake of notation simplicity. The total power is normalized such that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \frac{1}{N_s}\mathbf{I}_{N_s}$, $\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2 = N_s$ and $\mathbb{E}[||\mathbf{H}||_F] = NM$. Similarly, the receiver processes the received vector $\mathbf{r}$ by two different combining matrices:

$$\tilde{\mathbf{s}} = \sqrt{\rho}\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{n}, \tag{4.5}$$

where $\mathbf{W}_{RF} \in \mathbb{C}^{N\times N_{rf}}$ and $\mathbf{W}_{BB} \in \mathbb{C}^{N_{rf}\times N_s}$ are the RF and baseband combiners, respectively. We note here that both RF (analog) precoder and combiner are implemented by analog phase shifters with constant amplitude amplifiers, therefore, their entries have a constant norm.

The multi-user scenario is obtained directly from the point-to-point one by considering minor modifications. We assume that the base station communicates with $K$ single-antenna users and sends an independent date stream to each user. Therefore, the hybrid beamforming structure is implemented at the transmitter side only. The vector of the received signals is given by (4.1) with a minor change in the dimensions since $N = Ns = K$. Further, the augmented channel matrix is given by $\mathbf{H}^H = [\mathbf{h}_1, \cdots, \mathbf{h}_K]$, where $\mathbf{h}_k$ is the channel vector of the $k^{th}$ user.

We assume that the channel is known at the receiver(s), i.e., CSIR. CSIR is commonly adopted in all the current wireless standards, and it will also be implemented in mmWave-based standards (e.g., IEEE 802.11ad). The CSIR is not only utilized by the proposed schemes but it is also an essential requirement for many signal reception processes such as signal detection and hybrid (or digital) combining at the receiver(s) [5]. Owing to the intrinsic sparsity of millimeter

wave channels, many efficient two-stage CSI estimation algorithms have been developed based on compressed sensing techniques. This is in contrast to the single-stage techniques which are utilized in massive MIMO [36, 105, 107]. We consider two different types of short-term CSIT:

- **Limited feedback CSIT:** The transmitter has a finite rate (quantized) knowledge about the channel throughout a limited capacity feedback channel. Particularly, we assume that both the transmitter and the receiver agree on two predefined codebooks, $\mathcal{C}_{BB}$ and $\mathcal{C}_{RF}$, one for the digital stage and the other for the analog one.

- **Mixed (partial) CSIT:** The transmitter has two types of CSIT; a finite rate feedback knowledge for the massive MIMO channel and a perfect knowledge for the effective channel, i.e., $\mathbf{HF}_{RF}$. This assumption is practical when the communication channel is quasi-static and the number of transmitted stream is small. When the channel varies slowly with respect to the transmission rate, this allows for estimating the channel accurately at the receiver and feeding it back to the transmitter with a negligible rate overhead compared to the information rate. This assumption is widely considered in both multi-user and single-user cases (e.g., [2, 3, 6, 8, 10, 44, 52, 129]).

In addition to these two types of CSI which provide the base station with short-term channel updates, we assume, as in many other prior works [3, 52, 95], that the long-term second-order channel statistics are available at terminals. We note that the fading channel of many applications are *locally* wide sense stationary over time where the channel statistics remain constant for a very long period of time [108]. Hence, the channel covariance can be estimated very accurately at the base station without requiring frequent training. For instance, the channel covariance matrices can be known to the base station through efficient covariance estimation techniques for the hybrid structure such as the ones based on compressed sensing techniques (e.g., [94] and references therein). However, covariance estimation is out of the scope of this paper. We point out that, the combination of long and short-term channel state information creates different channel knowledge availability scenarios. We refer to the ones that have long-term channel knowledge as statistically-aided scenarios. We note that $\mathbf{F}_{RF}$ is designed based on either only limited feedback

CSIT or limited feedback CSIT aided with statistical information such as $\mathbf{R}_t$. This is different from prior work [15] where $\mathbf{F}_{RF}$ is designed based on the perfect channel realization, [2,3,52,95] where it depends only on the second-order channel statistics and [8,10,15,21,35,44,68,115,129] where it is constructed based on fixed predefined codebooks that do not change with the long-term channel statistics.

## 4.2 Single-user Problem Formulation and Proposed Algorithms

The canonical limited feedback problem in the literature of the fully-digital beamforming is to design a codebook that match a certain selection criteria in order to minimize an average distortion measure [69]. Similarly, the limited feedback hybrid precoding problem in mmWave literature is a codebook-based problem but with additional constraints on the analog precoder [8, 10, 15, 21, 35, 44, 68, 115, 129]. As a result, it boils down to how to define a codeword selection criterion and to design codebooks that naturally fit these constraints while utilizing the characteristics of mmWave channels. Here, we formalize the limited feedback precoding as a codebook-based subspace approximation problem while exploiting the sparsity and the spatial correlation of mmWave massive MIMO channels. Our design and problem formulation are described as follows. We start our design by discussing the selection metric.

### 4.2.1 Selection Metric

Since the massive MIMO and mmWave technologies are primarily meant to dramatically increase the capacity of 5G networks, the ergodic capacity, spectral efficiency or mutual information are reasonable selection metrics. However, these selection metrics have high computational complexity where they require computationally intensive calculations such as matrix determinant and inverse. This makes them impractical in high-dimensional applications. Exploiting the sparsity of mmWave channels, it has been shown that the loss in the mutual information due to approximating the optimal precoder, i.e., $\mathbf{F}_{opt}$, by the hybrid precoder, $\mathbf{F}_{RF}\mathbf{F}_{BB}$ is dictated

by the squared chordal distance between them, i.e., $\left\|\mathbf{F}_{opt}\mathbf{F}_{opt}^{H} - \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^{H}\mathbf{F}_{RF}^{H}\right\|_{F}^{2}$ [15]. The chordal distance is a subspace distance that measures the geodesic distance between two subspaces on the Grassmannian manifold [29]. We choose the chordal distance as a selection metric for two reasons. First, minimizing the chordal distance between the hybrid and optimal precoders directly minimizes the mutual information loss due to the hybrid structure [15]. Second, since it is a subspace distance, it suits the codebook-based precoding problem formulation as a subspace approximation problem [69]. The optimal precoder is given by the largest $N_s$ right singular vectors of $\mathbf{H}$, i.e., $\mathbf{F}_{opt} = \bar{\mathbf{V}}_{Ns}$ where $\mathbf{H} = \mathbf{U}_{\mathbf{H}}\mathbf{\Sigma}_{\mathbf{H}}\mathbf{V}_{\mathbf{H}}^{H}$ is the SVD of the channel matrix.

### 4.2.2 Codebook Design

The design of the codebook should account for two main properties of mmWave channels, i.e., the spatial correlation and the angular sparsity. Contrary to the spatially independent channel where its eigenvectors are isotropically distributed, the dominant eigenvectors of the spatially correlated channel point to certain preferred directions [70]. Moreover, the angular sparsity makes the channel tends to have a few numbers of dominant eigenvectors [15]. Considering both observations, we design the codebooks as follows. We consider a DFT-based codebook where its bases are drawn from a DFT matrix of size $M \times M$. The $M \times M$ DFT matrix $\mathbf{D}_M =$

$$\frac{1}{\sqrt{M}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^{2} & \cdots & \omega^{(M-1)} \\ 1 & \omega^{2} & \omega^{4} & \cdots & \omega^{2(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{(M-1)} & \omega^{2(M-1)} & \cdots & \omega^{(M-1)(M-1)} \end{bmatrix} \tag{4.6}$$

where $\omega = e^{-j\frac{2\pi}{M}}$. Considering all the combinations of $N_{rf}$ columns of the DFT matrix in (4.6), we can construct the RF codebook, $\mathcal{C}_{RF} \subseteq \mathcal{D}(M, N_{rf})$, where $\mathcal{D}(M, N_{rf})$ is the set of $M \times N_{rf}$ matrices, which has $\binom{M}{N_{rf}}$ matrices of size $M \times N_{rf}$. We note here that each matrix in $\mathcal{D}(M, N_{rf})$ represents an $N_{rf}$-dimensional subspace of $\mathbb{U}^{M\times N_{rf}}$ where $\mathbb{U}^{M\times N_{rf}}$ denotes the space of $M \times N_{rf}$ matrices that have constant magnitudes entries.

This codebook choice suits the angular sparsity of the mmWave channel since it divides the angular space into $M$ orthogonal beams which have $M$ distinctive angular directions. However, this does not account for the effect of the spatial correlation on the directivity of the subspaces of the eigenvectors of the channel. The main function of the RF codebook is to efficiently approximate these subspaces, therefore, the RF codebooks have to be tilted towards the subspace of the dominant eigenvectors of the channel, i.e., the column space of the optimal precoder. This is realized by multiplying $i^{\text{th}}$ codeword in $\mathcal{D}(M, N_{rf})$, i.e., $\big\{\mathcal{D}(M, N_{rf})\big\}_i, 1 \leq i \leq \binom{M}{N_{rf}}$, from the left by the transmit correlation matrix $\mathbf{R}_t$. This results in the skewed RF codebook, i.e., $\mathcal{C}_{RF,skewed}$

$$= \Big\{ \frac{\mathbf{R}_t\big\{\mathcal{D}(M, N_{rf})\big\}_1}{||\mathbf{R}_t\big\{\mathcal{D}(M, N_{rf})\big\}_1||}, \frac{\mathbf{R}_t\big\{\mathcal{D}(M, N_{rf})\big\}_2}{||\mathbf{R}_t\big\{\mathcal{D}(M, N_{rf})\big\}_2||}, \dots \Big\}.$$

This linear transformation makes the subspace of each codeword of the skewed RF codebook lives in the subspace of the optimal precoder. This, in turn, allows the skewed RF codebook to finely quantize the local neighborhood around the statistically preferred directions of the dominant eigenvectors, and thereby, leveraging efficiently the spatial correlation of mmWave channels.

We note that there are many good practical codebooks in the literature. For example q-bit resolution beam codebook for an $M$-element ULA developed in [56] and the codebooks in IEEE 802.15.3c [38] and wireless personal area networks (WPAN) operating in 60 GHz frequency band [13]. Although these codebooks are designed to simplify the hardware implementation cost and reduce the power consumption, they suffer from beam gain loss in some beam directions. In contrast, the DFT codebooks provide uniform maximum gain in all directions.

We point out that directly designing codebooks under constant magnitude entries constraint is extremely difficult and results in intractable optimization problems. This conclusion has been established in both fully-digital [71] and hybrid beamforming [46] literature. Instead, research works resort to either use predefined codebooks that have phases only entries (adhere to the RF hardware constraints) such as the beamsteering codebooks [8, 15, 21, 35, 44] and Hadamard codebooks [115] or quantizing sub-optimal solutions which are obtained by imposing these constraints on the optimal unconstrained precoder as in [66, 106, 145]. However, majority of

research works follow the former approach due to its superior performance compared to the latter one. In comparison to Hadamard codebooks, the DFT codebooks provides finer quantization, and thereby, having better performance. Moreover, compared to beamsteering codebooks, the columns of DFT codebooks are orthogonal. The orthogonality between columns is favorable in transmitting multiple streams (spatial multiplexing mode) and in dividing spanned spaces into orthogonal subspaces. On the contrary, the beamsteering codebooks are obtained by uniformly quantizing the angle of departure of the transmit antenna array response vector, i.e., $\mathbf{a}_t(\phi)$. For instance, for uniform linear array, $\mathbf{a}_t(\phi) = \frac{1}{\sqrt{M}}\begin{bmatrix}1 & e^{jkd\sin\phi} & \dots & e^{jkd(M-1)\sin\phi}\end{bmatrix}^H$, where $d$ is distance between two consecutive antenna elements, and $k = \frac{2\pi}{\lambda}$ is the wave number. Quantizing the angle of departure uniformly using $B = 2^q$ points, i.e., $\phi = \frac{2\pi u}{B}, u \in \{0, 1, \cdots, B-1\}$, leads to a codebook given by: $\mathbf{a}_t(\frac{2\pi u}{B}) = \frac{1}{\sqrt{M}}\begin{bmatrix}1 & e^{jkd\sin(\frac{2\pi u}{B})} & \dots & e^{jkd(M-1)\sin(\frac{2\pi u}{B})}\end{bmatrix}^H$. When $B \leq M$, the columns of the beamsteering codebooks are not orthogonal due to the periodicity of the sine function. Moreover, the columns of the beamsteering codebook are not necessarily asymptotically orthogonal. In particular, for some $u$ and $k \in \{0, 1, \cdots, B-1\}, \sin(\frac{2\pi u}{B}) - \sin(\frac{2\pi k}{B}) = \frac{1}{M}$; checking for the orthogonality between the columns of an $M \times M$ beamsteering matrix, i.e., $|\mathbf{a}_t^H(\frac{2\pi u}{M})\mathbf{a}_t(\frac{2\pi k}{M})|$, one has

$$\begin{aligned}\left|\mathbf{a}_t^H(\frac{2\pi u}{M})\mathbf{a}_t(\frac{2\pi k}{M})\right| &= \frac{1}{M}\left|\frac{1-e^{j\pi(\sin(\frac{2\pi k}{B})-\sin(\frac{2\pi u}{B}))M}}{1-e^{i\pi(\sin(\theta_k)-\sin(\theta_j))}}\right| \\ &= \frac{1}{M}\left|\frac{1-e^{j\pi}}{1-e^{j\pi/M}}\right| \\ &\to \frac{2}{\pi} \neq 0, M \to \infty\end{aligned}$$

In practice, $M$ is large but still finite and thus there is a non-negligible probability that there are $u$ and $k$ such that $\sin(\frac{2\pi u}{B}) - \sin(\frac{2\pi k}{B}) \leq \frac{1}{M}$. This directly makes the performance of the DFT codebook (utilized in the proposed algorithms) better than the beamsteering codebooks (used in [8, 10, 15, 44, 68]) as shown in the numerical results section.

In contrast to the RF precoding matrix, the baseband precoder $\mathbf{F}_{BB}$ does not have hardware constraints, i.e., $\mathbf{F}_{BB} \in \mathbb{C}^{N_{rf} \times N_s}$ and it can be implemented digitally. Moreover, it has relatively

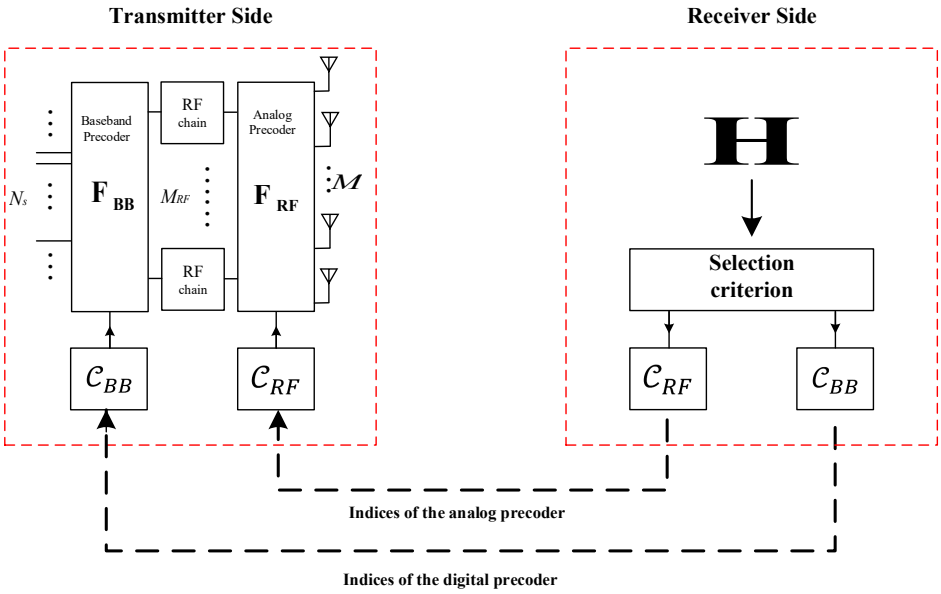


Figure 4.1: A limited feedback system for single user hybrid beamforming structure.

small dimensions of the same order of regular MIMO systems. This gives higher degrees of freedom in designing the digital codebook, i.e., $\mathcal{C}_{BB}$. Such codebook design is well-studied in the literature on limited feedback regular MIMO [71]. The optimal yet theoretical codebooks for quantizing $\mathbf{F}_{BB}$ are the ones that are based on Grassmannian codebooks. However, constructing an optimal Grassmannian codebook is a challenging problem [71]. In practice, it is preferable to use easily constructed and structured codebooks such as Hadamard [115], QPSK alphabet-based [109], and DFT codebooks. We consider the DFT codebooks in quantizing $\mathbf{F}_{BB}$ since they have nested structures, are easily constructed, and provide finer quantization than Hadamard and QPSK alphabet-based codebooks.

Finally, a schematic diagram of the limited feedback system for hybrid beamforming structure is depicted in Fig. 4.1.

### 4.2.3 Problem Formulation and Proposed Algorithms

We consider minimizing the chordal distance between the unconstrained optimal precoder, given by the $N_s$ right singular vectors of the channel matrix, i.e., $\mathbf{F}_{opt} = \bar{\mathbf{V}}_{N_s}$, and the hybrid beamforming matrix, i.e., $\mathbf{F}_{RF}\mathbf{F}_{BB}$, such that these matrices are selected from $\mathcal{C}_{RF}$ and $\mathcal{C}_{BB}$, respectively. As a result, the limited feedback precoding problem for the mmWave channel with hybrid precoding structure is:

$$\begin{aligned} \underset{\mathbf{F}_{RF},\mathbf{F}_{BB}}{\arg\min} \quad & \| \mathbf{F}_{opt}\mathbf{F}_{opt}^{H} - \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^{H}\mathbf{F}_{RF}^{H} \|_F^2 \\ \text{subject to} \quad & 1- \quad \mathbf{F}_{RF} \in \mathcal{C}_{RF,skewed} \\ & 2- \quad \mathbf{F}_{BB} \in \mathcal{C}_{BB} \end{aligned} \tag{4.7}$$

Unfortunately, the optimization problem in (4.7) is non-convex as a result of the combinatorial nature of the constraints. Hence, finding its global optimal solution requires prohibitive complexity, and thereby, in practice an efficient sub-optimal solution is preferred specially in high dimension applications. Contrary to the traditional way of the exhaustive search that has been considered in [15, 73] and complex optimization algorithms [21, 35, 44] and inspired by the greedy selections algorithms developed in [10, 68, 115], we solve the optimization problem in (4.7) algorithmically in two steps. This made possible by first approximating the subspace of the optimal precoder (finding its bases) and then finding the best linear combination of these bases that makes the hybrid precoder as close as possible to the optimal precoder. The precoding algorithm starts by projecting the optimal precoder on an $M \times M$ $\mathbf{R}_t\mathbf{D}_M$ and selects the $N_{rf}$ vectors along which the optimal precoder has the maximum projection (measured by the dot product of the columns of the optimal precoder and the columns of the DFT matrix). After identifying these vectors, i.e., $\mathbf{F}_{RF}$, the algorithm proceeds to find the $N_s$ linear combinations of these vectors along which the optimal precoder has the maximum projection, i.e., $\mathbf{F}_{BB}$. This makes it possible by projecting the $N_{rf} \times N_{rf}$ DFT matrix, the bases of the baseband codebook $\mathcal{C}_{BB}$, on the column space of $\mathbf{F}_{RF}$ to obtain $N_{rf}$ linear combinations; out of these linear combinations, we select the $N_s$ vectors along which the optimal precoder has the maximum projection again.

This strategy is to select $\mathbf{F}_{RF}$ and $\mathbf{F}_{BB}$ such that the $N_s$ columns of the hybrid precoder have smallest angles with the columns of the optimal precoder. These two steps are summarized in the following algorithm.

**Algorithm 3** Statistically-Aided Maximum Projection

1: **Input**: $\mathbf{F}_{opt}$, $\mathbf{D}_M$, $\mathbf{D}_{N_{rf}}$, $\mathbf{R}_t$
2: $\mathbf{Y} = (\mathbf{R}_t\mathbf{D}_M)^H\mathbf{F}_{opt}$
3: $Indices_{RF} = \arg\max_{N_{rf}}\{diag(\mathbf{Y}\mathbf{Y}^H)\}$
4: $\mathbf{F}_{RF} = \angle\big(\mathbf{R}_t\mathbf{D}_M(:, Indices_{RF})\big)$
5: $\mathbf{Z} = \left(\mathbf{D}_{N_{rf}}^H\mathbf{F}_{RF}^H\mathbf{F}_{RF}\mathbf{D}_{N_{rf}}\right)^{-1}(\mathbf{F}_{RF}\mathbf{D}_{N_{rf}})^H\mathbf{F}_{opt}$
6: $Indices_{BB} = \arg\max_{N_s}\{diag(\mathbf{Z}\mathbf{Z}^H)\}$
7: $\mathbf{F}_{BB} = \mathbf{D}_{N_{rf}}(:, Indices_{BB})$
8: **Return**: $Indices_{RF}$, $Indices_{BB}$

We note here that the linear transformation of the DFT codebook by the spatial correlation matrix allows for finer quantization to the vicinity of the optimal precoder. However, this results in analog precoders that do not adhere to the RF hardware since the DFT codebooks are adapted by the statistical correlation matrix. Therefore, we apply the hardware constraints on the analog precoder of the skewed codebooks using the phase extraction technique, i.e., step 4 in Algorithm 3. This procedure has been widely considered in the literature on hybrid beamforming [3, 15, 52, 66, 95] since it is the solution that has the shortest Euclidean distance to the unconstrained solution [34].

We point out that all the calculations of Algorithm 3 and its variants are performed at the receiver side aiming mainly to reduce the feedback overhead while exploiting the available CSIR (a prerequisite for other reception processing). Moreover, calculating $\mathbf{F}_{opt}$, which is based on the SVD of $\mathbf{H}$, does not require overhead of calculations since computing the SVD of $\mathbf{H}$ is a prerequisite for most of fully-digital or hybrid combining techniques [15]. We note that Algorithm 3 requires (i) limited feedback channel and (ii) statistical information about the spatial correlation. Moreover, Algorithm 3 can be modified in order to fit into different frameworks, such as the unavailability of statistical correlation matrix at the transmitter, and the presence of quasi-static

channels.

***Algorithm 3, Variant 1:*** When the statistical correlation information is not available at the transmitter or the channel is statistically uncorrelated, Algorithm 3 is easily modified by just replacing $\mathbf{R}_t$ by the identity matrix. This allows the constructed $\mathbf{F}_{RF}$ to have the constant magnitude entries and thereby there is no need to use $\angle$ operator in Step 4 in Algorithm 3 where $\mathbf{F}_{RF} = \mathbf{D}_M(:, Indices_{RF})$. Moreover, it preserves the orthogonality between the DFT columns, and hence, there is no need for the matrix inverse operation of Step 5 in Algorithm 3, i.e., $\mathbf{Z} = (\mathbf{F}_{RF}\mathbf{D}_{N_{rf}})^H\mathbf{F}_{opt}$.

***Algorithm 3, Variant 2:*** The second modification is based on utilizing the mixed CSIT instead of the limited one in order to improve the performance of Algorithm 3. The mixed CSIT assumes that the receiver sends the baseband precoder to the transmitter instantaneously and with infinite precision while the analog precoder is available to the transmitter with finite-precision (few bits). In particular, one selects the columns of $\mathbf{F}_{RF}$ from a statistically skewed DFT codebook and solves for $\mathbf{F}_{BB}$ as the least square solution such that selected columns have maximum projections on the subspace spanned by the optimal unconstrained (fully-digital) precoder. This permits the digital precoder to have entries with variable magnitude and phase, which improves the performance at the cost of increasing the feedback overhead. The hybrid beamforming procedure based on mixed CSIT is summarized in the following two steps. First, constructing the analog precoder using the same maximum projection procedure of Algorithm 3 (from line 1 to line 4). Second, given the analog precoder, and instead of executing the last part of Algorithm 3, one obtains $\mathbf{F}_{BB}$ by minimizing the chordal distance, in (4.7), while relaxing the second constraint. As a result, the baseband precoder is $\mathbf{F}_{BB} = \frac{\left(\mathbf{F}_{RF}^H\mathbf{F}_{RF}\right)^{-1}\mathbf{F}_{RF}^H\mathbf{F}_{opt}}{||\left(\mathbf{F}_{RF}^H\mathbf{F}_{RF}\right)^{-1}\mathbf{F}_{RF}^H\mathbf{F}_{opt}||_F}$. Finally, when Variant 1 and Variant 2 are combined, i.e., in the lack of statistical channel information and the availability of mixed CSIT, we call it *Algorithm 3, Variant 3*.

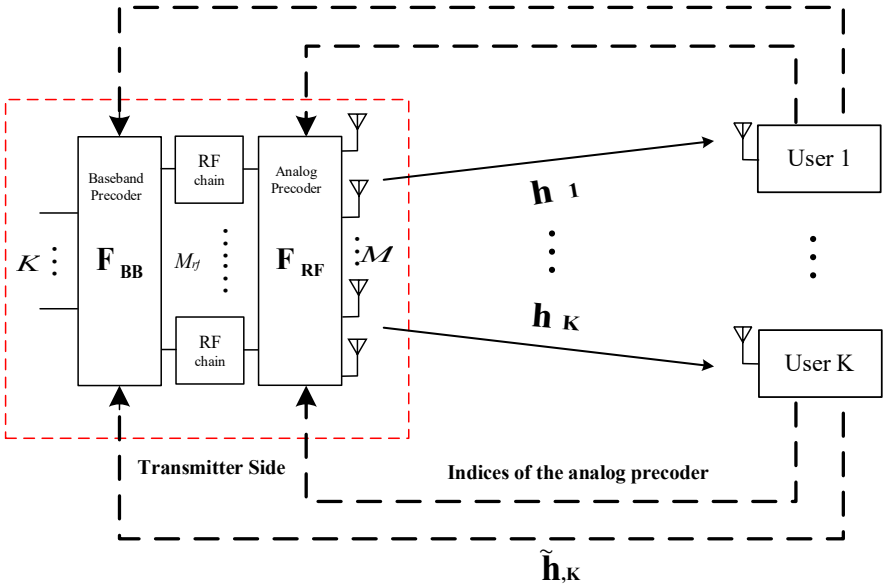


Figure 4.2: A limited feedback system for multi-user hybrid beamforming structure.

## 4.3 Multi-user Problem Formulation and Proposed Algorithms

In this section, we consider the design of the codebook-based hybrid precoding for the downlink multi-user (MU) multiple-input single-output (MISO) broadcast (BC) system in which the base station is equipped with an $M$-element antenna array and a limited number of RF chains, and serves $K$ single-antenna users where $M \gg N_{rf} \geq K$. A schematic diagram of the limited feedback system for the K-user MISO BC channel with hybrid beamforming structure is depicted in Fig. 4.2.

Before we proceed with the problem formulation, we highlight the main differences between MU MISO and SU MIMO systems that drive us to treat the problem differently. First, maximizing the sum-rate of the $K$-user MISO BC requires sophisticated and computationally intensive *non-linear* operations such as dirty paper coding [135] and its optimal solution does not adhere to the RF constraints on the analog precoder. Second, since all users (receivers) are separated, where no joint processing of their signals can be done at receivers, approximating sub-optimal

linear precoding schemes, such as zero forcing [61], minimum mean square error and generalized eigenvector beamforming [112], is not applicable in our framework. This is mainly because all these solutions require either global CSIR about all users or user cooperation which entail huge training and feedback overheads. Although these schemes are designed to maximize signal to interference noise ratio (SINR), they achieve sum-rates within a fixed SNR gap of the network capacity [61, 112].

The received signal of the single-antenna user $k$ is:

$$y_k = \mathbf{h}_k^H \mathbf{F}_{RF} \mathbf{f}_{BB,k} s_k + \sum_{i \neq k}^{K} \mathbf{h}_k^H \mathbf{F}_{RF} \mathbf{f}_{BB,i} s_i + n_k, \tag{4.8}$$

where $\mathbf{f}_{BB,k}$ is the baseband precoder vector for user $k$, i.e., the $k^{th}$ column of $\mathbf{F}_{BB}$ and $\mathbf{h}_k^H$ is the channel row vector of length $M \times 1$ for user $k$ where $\mathbf{h}_k = \mathbf{R}_{t,k}^{1/2} \mathbf{h}_{w,k}$ and $\mathbf{h}_{w,k}$ has i.i.d Gaussian distributed entries. Moreover, we assume that the users have different spatial correlation matrices $\mathbf{R}_{t,k} = \mathbb{E}[\mathbf{h}_k \mathbf{h}_k^H]$. Considering an equal power normalized transmission strategy, the achievable rate of user $k$ is:

$$r_k = \log_2 \left(1 + \text{SINR}_k\right) \tag{4.9}$$

$$\text{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{F}_{RF} \mathbf{f}_{BB,k}|^2}{\sum\limits_{i \neq k} |\mathbf{h}_k^H \mathbf{F}_{RF} \mathbf{f}_{BB,i}|^2 + K\sigma_k^2}, \tag{4.10}$$

where $\text{SINR}_k$ is the signal to interference noise ratio of user $k$. Therefore, the sum-rate of the $K$-user broadcast channel is given as $R_{sum} = \sum_k r_k$.

Optimizing the sum-rate and other sophisticated performance metrics such as energy efficiency and bit error rate while considering the RF hardware constraints results in notorious non-convex and sparse problem formulations where a series of convex relaxations and approximations are performed to secure satisfactory sub-optimal solutions [21, 35, 44]. These solutions are obtained based on a series of sophisticated tangled iterative algorithms that require either modern or classic convex solvers such as MOSEK and interior point, respectively, where their convergence

depends highly on the initial point.

Given the intractability and impracticality associated with applying the prior art of fully-digital [135] and hybrid beamforming [21, 35, 44] methodologies in massive MIMO mmWave systems, we tackle the design problem differently. Particularly, the proposed scheme is developed based on maximizing the signal to interference plus noise ratio (4.10) over two successive stages using non-iterative and low-complexity procedure; for instance, one iteration of the algorithm in [44] has a computational complexity of $((M^6 + 64)K^3 + 6K^2M^2)$ [44] while Algorithm 4 has a computational complexity of $((4M^2K + 8MK + 9K^3 + 2MK^2))$ where its complexity is dominated by the Moore-Penrose pseudo-inverse [24]. We show that this approach achieves higher sum-rate than prior art in [21, 35, 44] and comparable to fully-digital scheme with perfect CSIT.

This is based on leveraging the property of spatially correlated channels that the channel vectors of different users exist in different subspaces identified by the statistical covariance matrices and point to specific preferred directions. This can be utilized to relieve the global perfect channel knowledge assumption, consolidate the separability of users in the space, and enhance the performance of the hybrid precoding schemes. Particularly, exploiting the directivity of subspaces of users' channels, each user designs its analog precoder vector(s) selfishly to maximize its signal strength while ignoring the interference. Its aim is to decrease the feedback overhead and for CSIT where this step requires only local statistical knowledge of its own channel vector at the receiver and limited feedback channel. In the second stage, the digital precoder is designed based on the effective channel, i.e., $\mathbf{H}^H\mathbf{F}_{RF}$, in order to cancel out the inter-user interference.

Therefore, the analog precoding problems, while abstracting the digital processing, are given by:

$$\begin{aligned} &\underset{\mathbf{f}_{RF,k}}{\arg\max} \quad |\mathbf{h}_k^H \mathbf{f}_{RF,k}|^2, \qquad 1 \leq k \leq K \\ &\text{subject to} \quad \mathbf{f}_{RF,k} \in \mathcal{C}_{RF,skewed,k}, \qquad 1 \leq k \leq K, \end{aligned} \tag{4.11}$$

where $\mathbf{f}_{RF,k}$ is the $k$ column of $\mathbf{F}_{RF}$ and $\mathcal{C}_{RF,skewed,k}$ is the analog codebook of user $k$ skewed by its covariance matrix. The problem defined in (4.11) is a typical codebook-based precoding problem. Owing to the distributed nature of the multi-user networks where neither user coop-

eration to select the analog precoder columns is assumed nor global CSIT and only collocated signal processing are required, the columns of the analog precoder are selected in a distributed manner from the available codebooks. We solve (4.11) for each user individually where the channel matching metric is utilized to select the best codeword(s). However, in mmWave bands, one often finds users that have one or more common scatters/clusters. This implies that one or more users may choose the same codeword which results in a near singular or rank deficient analog precoding matrix. We propose here a selection strategy that avoid such situations. In particular, each user selects the best $L$ codewords that match its channel vector from the predefined codebook and feeds them back to the base station. Then, the base station constructs the RF precoder matrix by selecting the preferred $K$ different codewords out of the total $K \times L$ codewords received from all $K$ receivers. Each of the preferred $K$ codewords(vectors) is corresponding to one respective user. Particularly, each user's receiver transmits a set of indices indicating the best $L$ columns in descending order (or ascending order) such that the first (or last) element in the set of indices indicates the column along which the channel vector has the maximum projection. In the case where two or more receivers share one or more codewords, the base station selects the next best codeword(s) out of the $L$ codewords (i.e., using the next entry in the set of indices), such that all the columns of $\mathbf{F}_{RF}$ are different codewords. Hence, the base station can select $K$ DFT columns corresponding to respective $K$ MU receivers, to form the RF precoder matrix as $\mathbf{F}_{RF} = [\mathbf{f}_{RF,1}, \mathbf{f}_{RF,2}, \cdots, \mathbf{f}_{RF,K}]$. This selection strategy is devised to enforce the full rank constraint on $\mathbf{F}_{RF}$ which is required to achieve the highest multiplexing gain offered by the network. We note here that $L$ is identified empirically since it is a function of the number of common clusters/scatters between users and the location of users.

We also note that this precoding strategy can be extended to the case where there are more RF chains than the number of users. When $2K \geq N_{rf} > K$, the base station selects the columns of the analog precoder in two rounds. First, it constructs the first $K$ columns of $\mathbf{F}_{RF}$, similar to the previous strategy, by selecting the first entries of all users' sets, i.e., the indices of the columns that have the largest projection, such that they are different. Then, for the remaining $N_{rf} - K$ columns of $\mathbf{F}_{RF}$, the base station selects the second entries of only $N_{rf} - K$ user's sets.

On the other hand, the digital precoder problem is given based on the effective channel as:

$$\begin{aligned} \underset{\mathbf{F}_{BB}}{\arg\max} \quad & \frac{|\tilde{\mathbf{h}}_k^H \mathbf{f}_{BB,k}|^2}{\sum_{i \neq k} |\tilde{\mathbf{h}}_k^H \mathbf{f}_{BB,i}|^2 + K\sigma_n^2}, \quad 1 \leq k \leq K, \\ \text{subject to} \quad & ||\mathbf{F}_{RF}\mathbf{F}_{BB}||_F^2 \leq K \end{aligned} \tag{4.12}$$

Contrarily to the analog precoding, the digital precoder has an optimal solution given by zero-forcing [125]. Having determined the RF precoder matrix $\mathbf{F}_{RF}$, the base station then determines the baseband precoder matrix $\mathbf{F}_{BB}$ based on the effective channel after applying the RF precoder matrix. Particularly, each user's receiver feeds back its estimated effective channel vector $\tilde{\mathbf{h}}$. The baseband precoder matrix $\mathbf{F}_{BB}$ is determined by the base station to be the zero-forcing solution based on these effective channel vectors, represented by $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{F}_{RF}$, such that $\mathbf{F}_{BB} = \tilde{\mathbf{H}}^H(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H)^{-1}$. We note here that the effective channel has small dimensions in comparison to the massive MIMO channel. This significantly reduces the channel training and feedback required for estimating and acquiring the effective channel. These two steps are summarized in Algorithm 4.

We note that Algorithm 4 requires mixed CSIT; specifically (i) limited feedback channels and (ii) statistical information about the spatial correlation for constructing the analog precoder and (iii) perfect knowledge about the low dimensional effective channel. Similar to Algorithm 3 Variant 1, Algorithm 4 can be modified in order to accommodate for the unavailability of the transmit covariance matrix at the transmitter. We refer to this modification as Algorithm 4 Variant 1.

**Algorithm 4** Distributed Statistically-Aided Maximum Projection Multi-User

1: **Input**: $\mathbf{H}$, $\{\mathbf{R}_{t,1}, \cdots, \mathbf{R}_{t,K}\}$, $\mathbf{D}_M$
2: **While** $k \leq K$
$\mathbf{Y}_k = (\mathbf{R}_{t,k}\mathbf{D}_M)^H \mathbf{h}_k$
$Indices_{RF,k} = \arg\max_{L}\{diag(\mathbf{Y}_k \mathbf{Y}_k^H)\}$
**End while**
3: The base station selects $K$ indices out of the set of indices, i.e., $\{Index_{RF,1}, \cdots, Index_{RF,K}\}$ and constructs $\mathbf{F}_{RF}$ based on extracting the phases of the corresponding vectors of these indices multiplied by the corresponding covariance matrices.
4: The base station applies $\mathbf{F}_{RF}$ in the downlink to allow for estimating the effective channel at the different users.
5: Each user feeds back its estimated effective channel vector $\tilde{\mathbf{h}}$ to the base station .
6: $\mathbf{F}_{BB} = \mathbf{F}_{RF}^H \mathbf{H}(\mathbf{H}^H \mathbf{F}_{RF} \mathbf{F}_{RF}^H \mathbf{H})^{-1}$
7: $\mathbf{F}_{BB} = \dfrac{\sqrt{K}\mathbf{F}_{BB}}{\mathbf{F}_{RF}\mathbf{F}_{BB}}$
8: **Return** $\mathbf{F}_{RF}$ and $\mathbf{F}_{BB}$

## 4.4 Bounds And Asymptotic Analysis of the Achievable Rate of The Proposed Schemes

In order to evaluate the performance of Algorithm 3, Algorithm 4 and their variants, we consider the mutual information of the analog precoding stage (the common stage among the proposed schemes) as a performance metric. This is made possible by abstracting the digital precoding and receiver side processing while considering an equal power transmission strategy. We start by providing lower and upper bounds on the mutual information of any DFT codebook-based analog precoding strategy. Then, we show that the proposed schemes are asymptotically optimal as the number of transmit antennas $M$ goes to infinity and the millimeter wave channel has a limited number of paths, i.e., $P<M$.

**Proposition. 4.4.1.** *The mutual information of the mmWave channel with hybrid precoding structure at the transmitter, where the analog precoder is selected from an $M \times M$ DFT matrix, i.e.,*

*$I_{RF} = \log\det\left(\mathbf{I}_{N_{rf}} + \frac{\rho}{\sigma_n^2 N_s}\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{H}\mathbf{F}_{RF}\right)$, is bounded by (4.13) shown at the top of the next page.*

$$\sum_{i=1}^{N_{rf}} \log\left(1 + \frac{\rho}{\sigma_n^2 N_s}\lambda_{M-N_{rf}+i}(\mathbf{A}_t)\lambda_i(\mathbf{G}^H\mathbf{A}_r^H)\right) \leq I_{RF} \leq \sum_{i=1}^{N_{rf}} \log\left(1 + \frac{\rho}{\sigma_n^2 N_s}\lambda_i(\mathbf{A}_t)\lambda_i(\mathbf{G}^H\mathbf{A}_r^H)\right). \tag{4.13}$$

*In Eq. (4.13), $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_M(\mathbf{A})$ are the eigenvalues of the matrix $\mathbf{A}\mathbf{A}^H$ in a descending order.*

*Proof.* We start with the mutual information of the channel when using the analog beamformer at the transmitter:

$$\begin{aligned}
R &= \log\det\left(\mathbf{I}_{N_{rf}} + \frac{\rho}{\sigma_n^2 N_s}\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{H}\mathbf{F}_{RF}\right) \\
&\overset{(a)}{=} \log\det\left(\mathbf{I}_{N_{rf}} + \frac{\rho}{\sigma_n^2 N_s}\mathbf{F}_{RF}^H\mathbf{A}_t\mathbf{G}^H\mathbf{A}_r^H\mathbf{A}_r\mathbf{G}\mathbf{A}_t^H\mathbf{F}_{RF}\right) \\
&\overset{(b)}{=} \log\det\left(\mathbf{I}_{N_{rf}} + \frac{\rho}{\sigma_n^2 N_s}\mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^H\mathbf{G}^H\mathbf{A}_r^H\mathbf{A}_r\mathbf{G}\mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^T\mathbf{U}_\mathbf{X}^H\right) \\
&\overset{(c)}{=} \log\det\left(\mathbf{I}_{N_{rf}} + \frac{\rho}{\sigma_n^2 N_s}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^H\mathbf{G}^H\mathbf{A}_r^H\mathbf{A}_r\mathbf{G}\mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^T\right) \\
&\overset{(d)}{=} \log\det\left(\mathbf{I}_{N_{rf}} + \frac{\rho}{\sigma_n^2 N_s}\mathbf{\Sigma}_\mathbf{X}^2\mathbf{V}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}^2\mathbf{V}_\mathbf{Y}^H\right) \\
&\overset{(e)}{=} \sum_{i=1}^{N_{rf}} \log(1 + \frac{\rho}{\sigma_n^2 N_s}[\mathbf{\Sigma}_\mathbf{X}]_{i,i}^2[\mathbf{\Sigma}_\mathbf{Y}]_{i,i}^2)
\end{aligned} \tag{4.14}$$

where (a) is due to the use of the augmented matrix representation of the channel in (4.4) and (b) is based on considering the SVD of $\mathbf{X} = \mathbf{F}_{RF}^H\mathbf{A}_t = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^H$. Eq. (c) is due to using $\log\det(\mathbf{I} + \mathbf{A}\mathbf{B}) = \log\det(\mathbf{I} + \mathbf{B}\mathbf{A})$ and $\mathbf{U}_\mathbf{X}\mathbf{U}_\mathbf{X}^H = \mathbf{I}$, i.e., the invariance of the mutual information formula under unitary transformation. Eq. (d) is based on considering the SVD of $\mathbf{Y} = \mathbf{A}_r\mathbf{G}\mathbf{V}_\mathbf{X} = \mathbf{U}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^H$. Eq. (e) comes from (d) by using $\lambda_i(\mathbf{\Sigma}_\mathbf{X}^2) = [\mathbf{\Sigma}_\mathbf{X}^2]_{i,i}$, $\lambda_i(\mathbf{U}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^H) \leq \lambda_i(\mathbf{\Sigma}_\mathbf{Y}) = [\mathbf{\Sigma}_\mathbf{X}^2]_{i,i}$ based on the eigenvalue interlacing property where equality holds since $\mathbf{V}_\mathbf{Y}$ is an $M \times M$ unitary matrix [147], $\det(\mathbf{\Sigma}) = \prod_i^{N_{rf}}[\mathbf{\Sigma}]_{i,i}$ and $\log(\prod_i) = \sum_i \log()$.

Bounding (e) requires bounding the non-negative elements of $\mathbf{\Sigma_X}$ where the effect of precoding is only captured in $\mathbf{\Sigma_X}$. The elements of $\mathbf{\Sigma_X^2}$ correspond to the eigenvalue decomposition of $\mathbf{F}_{RF}^H \mathbf{A}_t \mathbf{A}_t^H \mathbf{F}_{RF}$ where $\mathbf{F}_{RF}$ is a part of an $M \times M$ DFT matrix, i.e., $\mathbf{F}_{RF} = \overline{\mathbf{D}}_{N_{rf}}$ and $\overline{\mathbf{D}}_{N_{rf}}$ denotes a collection of $N_{rf}$ columns selected from an $M \times M$ DFT matrix. Since the $\overline{\mathbf{D}}_{N_{rf}}$ is a semi-unitary matrix, i.e., $\overline{\mathbf{D}}_{N_{rf}}^H \overline{\mathbf{D}}_{N_{rf}} = \mathbf{I}_{N_{rf}}$ and using Lemma 3.3.1 in [47], the eigenvalues of the product, i.e., $\lambda_i(\overline{\mathbf{D}}_{N_{rf}}^H \mathbf{A}_t)$ can be bounded by:

$$\lambda_{M-N_{rf}+i}(\mathbf{A}_t) \leq \lambda_i(\overline{\mathbf{D}}_{N_{rf}}^H \mathbf{A}_t) \leq \lambda_i(\mathbf{A}_t), \tag{4.15}$$

where $\lambda_i(\overline{\mathbf{D}}_{N_{rf}}^H \mathbf{A}_t) = [\mathbf{\Sigma_X}]_{i,i}^2$. Substituting these bounds in (e), (4.13) readily follows. ■

Similar bounds can be obtained as a function of the transmit correlation matrix $\mathbf{R}_t$ instead of $\mathbf{A}_t$ by using (4.2) instead of (4.4).

**Remark. 4.4.2.** ***SNR Gap:*** *Indeed, there is a wide SNR gap between the upper and lower bounds in (4.13) and it is more pronounced at small values of $N_{rf}$ as shown in Fig. 3.19 subplots (a) and (b). This is expected since the proposed bounds are valid for any DFT-based hybrid precoding scheme. However, this SNR gap reduces gradually with increasing $N_{rf}$ till reaching to zero, i.e., both bounds coincide, at $N_{rf} = M$ as shown in Fig. 3.19 (c) (subplot (c) shows the SNR gaps between different schemes and the upper bound). The wide SNR gap between the upper and lower bounds at small values of $N_{rf}$ suggests that the DFT columns of $\mathbf{F}_{RF}$ should be carefully selected to achieve the upper bound while the small SNR gap at large values of $N_{RF}$ hints that no sophisticated selection is required and just fixed or random selection is sufficient.*

**Remark. 4.4.3.** ***Design insight:*** *The upper bound in (4.13) is achieved when all the selected DFT columns are in the same directions of the largest $N_{rf}$ eigenvectors of $\mathbf{A}_t$. On the other hand, the lower bound in (4.13) is achieved when none of the selected DFT columns are pointing toward any of the largest $N_{rf}$ eigenvectors' directions of $\mathbf{A}_t$. This suggests that efficient DFT-based hybrid precoding schemes should consider aligning the DFT columns of $\mathbf{F}_{RF}$ to the egienvectors of $\mathbf{A}_t$ and its efficiency is measured by how close its mutual information curve to the upper*

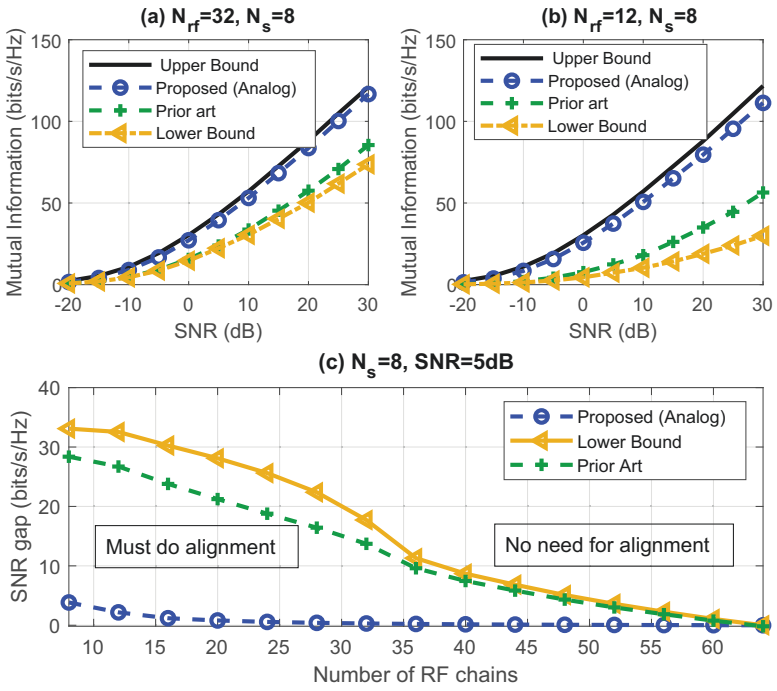


Figure 4.3: Bounds on mutual information and SNR gap of DFT-based hybrid beamforming schemes over $16 \times 64$ mmWave channel with $N_{cal} = 5$ and $N_{ray} = 6$.

*bound. On the other hand, inefficient DFT-based hybrid precoding schemes do not consider any alignment technique, and thereby, their performance is close to the lower bound. For instance, comparing subplots (a) and (b) in Fig. 3.19, one finds the mutual information curves of the proposed analog precoder, denoted by (Maximum Projection), outperform the ones of the prior art scheme developed in [129] for different values of $N_{rf}$. This is because, in the proposed analog strategy, the $i^{th}$ DFT column is selected to be aligned to the $i^{th}$ eigenvector of $\mathbf{R}_t$ whereas, in [129], the columns of $\mathbf{F}_{RF}$ are either randomly selected from an $M \times M$ DFT matrix, or are represented as the first $N_{rf}$ columns of an $M \times M$ DFT matrix.*

**Corollary. 4.4.3.1.** *For a large uniform linear antenna array where $M \to \infty$ in a limited scattering mmWave channel, defined in (4.3), and $P<M$, the proposed analog precoding strategy (the common procedure among the proposed algorithms) achieves the upper bound on the mutual*

*information in (4.13)*

*Proof.* For large $M$ ($M \to \infty$ and $P < M$), the columns of $\mathbf{A}_t$ are asymptotically orthogonal and have unit norms [31]. Moreover, $\mathbf{A}_t$ and the optimal precoder, given by the largest right singular vectors of the channel matrix, i.e., $\mathbf{F}_{opt} = \overline{\mathbf{V}}_{N_s}$, span the same subspace. Therefore, the channel matrix representation in (4.4) converges to its SVD [31, Lemma 2], i.e., $\mathbf{A}_t$ and $\mathbf{A}_r$ converge to $\mathbf{V_H}$ and $\mathbf{U_H}$, respectively. Considering ULA at the base station, the columns $\mathbf{A}_t$ have the same structure as the DFT matrix; hence the optimal precoder, i.e., $\mathbf{F}_{opt}$, has a DFT structure as well [3]. Consequently, the dot product between the $M \times M$ DFT matrix and the optimal digital precoder is sufficient to select the DFT columns that are perfectly aligned to the optimal precoder vectors, i.e., the largest eigenvectors of $\mathbf{A}_t$. Since $\mathbf{F}_{RF}$ in all the proposed algorithms is constructed based on these $N_{rf}$ DFT columns, its columns are asymptotically aligned to the eigenvectors of $\mathbf{A}_t$ as well. Consequently, and based on [47, Lemma 3.3.1], the proposed analog procedure asymptotically achieves the upper bound in (4.13). ■

## 4.5 Simulation Results

In this section, we evaluate the performance of the proposed algorithms, Algorithm 3 and Algorithm 4, and their variants. All these hybrid beamforming schemes are compared with the prominent prior works, mentioned in Section I, in terms of spectral efficiency over mmWave bands. Since, we are mainly concerned with decreasing the channel feedback overhead, we limit the application of the proposed algorithms to the transmitter side. For the single-user case, the hybrid precoder is obtained by either Algorithm 3 or its variants (depending on the available CSIT) while the hybrid combiner, i.e., $\mathbf{W}_{RF}\mathbf{W}_{BB}$, is obtained by approximating the dominant $N_s$ left singular vectors of the channel, i.e., $\bar{\mathbf{U}}_{N_s}$, using the procedure described in [15, equation (16)-(18)]. On the other hand, for multi-user case, since all users are equipped with single antennas, there is no combining available at the users' sides.

### 4.5.1 Simulation Setups

We consider the clustered channel model, described in (3.3) where the complex channel gains, i.e., $\alpha_{il}$, are i.i.d $\sim \mathcal{CN}(0, \sigma^2_{h,il})$ and $\sigma^2_{h,il}$ are randomly generated from an exponential distribution and normalized such that $\sum_{li} \sigma^2_{n,il} = 1$ [2, 15, 95]. Moreover, we assume that the total number of paths/rays, $P$, is sufficiently larger than the number of transmitted streams per user, i.e., $P = N_{ray}N_{cl} > N_s$ and $P > 1$ for single and multi-user cases, respectively. For the single-user case, the transmitter is equipped with an antenna array with $64$ elements, and the receiver is equipped with an antenna array with $16$ elements. On the other hand, for the multi-user case the base station is equipped with $64$ or $32$ elements and the number of user varies between $4$ and $20$. Moreover, we assume that the number of users $K$ is smaller than the number of antenna elements $M$. Thus, there is no need to consider any users' scheduling or opportunistic selection strategy and instead, we serve all users. In both cases, we consider ULA where the transmit/receive antenna array responses of ULA with half wave length element spacing and $M$-element is given by:

$$\mathbf{a}_{t/r}(\theta^{t/r}_{il}) = \frac{1}{\sqrt{M}}[1, e^{j\pi \sin(\theta^{t/r}_{il})}, \cdots, e^{j(M-1)\pi \sin(\theta^{t/r}_{il})}]^H \tag{4.16}$$

The angles of departure, $\theta^t_{il}$, and arrival $\theta^r_{il}$, are drawn form Laplace distributions with means $\overline{\theta}^t_i$ and $\overline{\theta}^r_i$ and angular spread $\Delta_\theta$ of $7.5^\circ$ [15, 95]. Accordingly, the $(m, n)$ entry of the transmit correlation matrix of user $k$, i.e., $\mathbf{R}_{t,k}$, for ULA with half wave length element spacing is [2, 3, 95]:

$$[\mathbf{R}_{t,k}]_{m,n} = \frac{1}{N_{cl}} \sum_{i=1}^{N_{cl}} \frac{1}{\sqrt{2}\Delta_\theta} \int_{\bar{\theta}^t_i - \pi}^{\bar{\theta}^t_i + \pi} e^{-\frac{\sqrt{2}}{\Delta_\theta}|\psi - \bar{\theta}^t_i| - j\pi(m-n)sin\psi} d\psi. \tag{4.17}$$

### 4.5.2 Spectral efficiency performance evaluation

From the rich literature on hybrid beamforming, we choose the most prominent and relevant prior art schemes as benchmarks in order to evaluate the efficacy of the proposed schemes. The

first benchmark solves the hybrid beamforming problem algorithmically based on an efficient compressed sensing algorithm, namely, orthogonal matching pursuit (OMP) [15] where the analog precoder is represented as linear combinations of the columns of $\mathbf{A}_t$. This scheme requires perfect CSIT since $\mathbf{A}_t$ changes with each channel realization. Therefore, this benchmark works as an upper bound to its exhaustive search-based limited feedback version in [15] and all hybrid beamforming schemes that utilize mixed or limited feedback CSIT. We call this benchmark by spatially sparse OMP. The second also utilized the OMP algorithm, however, the analog beamformer is obtained based on the beamsteering codebook whereas the digital precoder is designed to eliminate the inter-user interference based on the zero-forcing approach [10, 68]. We call this benchmark by codebook-based spatially sparse [10, 68]. This scheme requires mixed CSIT and it represents the prior works that utilizes beamsteering codebooks and compressed sensing greedy algorithms [8, 68, 77, 115, 144]. We note here that the schemes in [77, 144] are very similar to the ones in [10, 68] expect the digital precoders of the former are designed to minimize the mean squared error while those of the latter are designed to maximize the spectral efficiency. The third is also a codebook-based hybrid beamforming scheme where the analog precoder is selected from beamsteering codebooks using a MOSEK-based algorithm where its selection metric is based on maximizing the sum-rate. We refer to this scheme as codebook-based sum-rate maximization [44]. This scheme also requires mixed CSIT and it represents the prior works that utilize complex iterative processing [21, 35]. The fourth is developed in [95] where the analog precoders of the users are jointly designed at the base station to maximize the signal to leakage and noise ratio (SLNR) based on the second-order statistical channel knowledge of all users; specifically, the analog precoder is given as the largest $K$ eigenvectors of the sum of the transmit covariance matrices of the users. Moreover, the digital precoder is designed, based on perfect effective channel knowledge, to minimize the SINR. We refer to this benchmark as SLNR-based statistical beamforming [95]. The fifth is widely known as the joint spatial division multiplexing scheme (JSDM [2, 3, 52]). The basic idea of JSDM is to partition the user population into groups where users with similar covariance matrices are grouped together while maintaining orthogonality between the groups. The analog RF beamforming is designed to reduce the inter-group interference by employing the well-known block diagonalization technique using only statistical

knowledge where the analog beamforming is the augmented matrix of the largest eigenvectors of the covariance matrix of the users' groups. On the other hand, the digital baseband precoding is designed to eliminate the inter-user interference between users in the same group using linear precoders based on perfect channel knowledge.

In Figs. 4.4 and 4.5, we plot the spectral efficiency of five different hybrid beamforming schemes for the single-user case, namely, spatially sparse hybrid precoding based on OMP [15], shown in red solid line with circles, Algorithm 3, shown in sky blue solid line, Algorithm 3 Variation 1 (Alg3.Var1), shown in solid blue line with squares, Algorithm 3 Variation 2 (Alg3.Var2), shown in green solid line with plus signs, Algorithm 3 Variation 3 (Alg3.Var3), shown in yellow solid line with crosses and the codebook-based spatially sparse hybrid precoding scheme in [10], shown in solid violet line with diamonds in addition to the fully-digital SVD based precoder, shown in solid black line.

Fig.4.4 shows that Alg3.Var2 and Alg3.Var3 have almost the same performance as the spatially sparse OMP-based scheme which requires full CSIT [15] while Alg3.Var2 and Alg3.Var3 require statistically-aided mixed CSIT and mixed CSIT, respectively. This shows the efficiency of Alg3.Var2 and Alg3.Var3 in reducing the feedback overhead (where only knowledge about the low-dimensional effective channel is required) while having marginal spectral efficiency degradation. Moreover, Fig.4.4 demonstrates that Algorithm 3 and Alg3.Var1 both outperform the performance of the codebook-based spatially sparse benchmark [10]. This, in turn, significantly reduces the required channel knowledge to achieve the same high spectral efficiency performance since Algorithm 3 and Alg3.Var1 require limited feedback CSIT while the scheme in [10] requires mixed CSIT. We note here that the impact of the statistical knowledge on increasing spectral efficiency is marginal where the curves of Alg3.Var2 and Alg3.Var3, and Algorithm 3 and Alg3.Var1 are on top of each other.

In order to evaluate the impact of the statistical channel knowledge on the spectral efficiency, in Fig. 4.5, we simulated all these schemes in the same set-up of Fig. 4.4 except, codebooks' lengths in this set-up are decreased from $64$ to $32$. Fig.4.5 shows that even though there is some performance loss due to the reduction in codebooks' lengths, the statistically-aided schemes

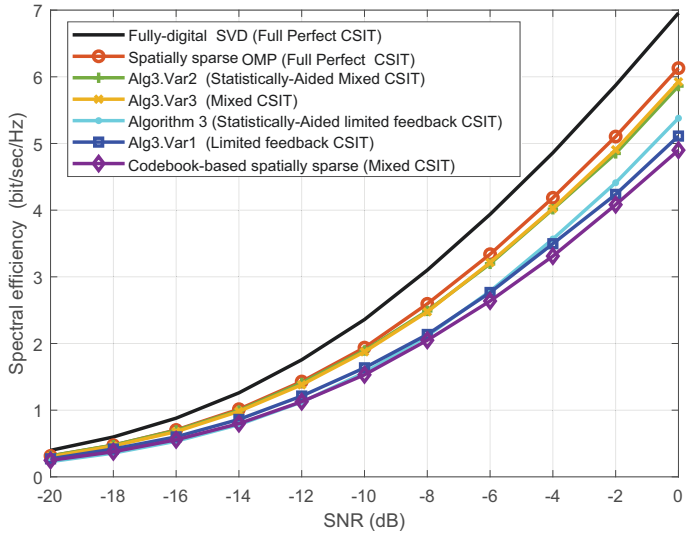


Figure 4.4: Spectral efficiency comparison of different hybrid beamforming schemes over $16 \times 64$ mmWave channel with $N_{rf} = 3, N_s = 2, N_{cl} = 4$ and $N_{ray} = 3$ where the codebooks' lengths are $64$.

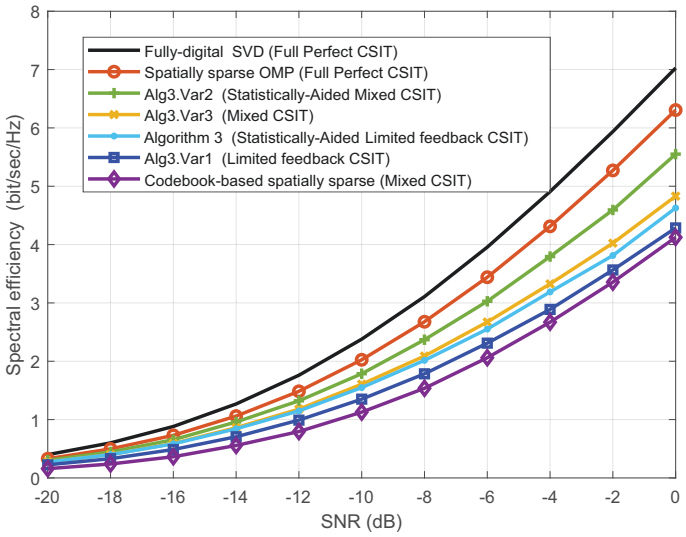


Figure 4.5: Spectral efficiency comparison of different hybrid beamforming schemes over $16 \times 64$ mmWave channel with $N_{rf} = 3,\ N_s = 2, N_{cl} = 4$ and $N_{ray} = 3$ where the codebooks' lengths are $32$.

generally outperform the ones without statistical knowledge. Moreover, in comparison with Fig. 4.4, Fig. 4.5 illustrates that the spectral efficiency loss of the statistically aided schemes, i.e., Algorithm 3 and Alg3.Var2, due to the reduction in codebooks' lengths is less severe than loss of the ones without statistical knowledge, i.e., Alg3.Var1 and Alg3.Var3. Particularly, the required SNR gaps that compensate for these losses are $2$ dB and $4$ dB, respectively. Further, from Fig.4.4, it is clear that the statistical channel knowledge has a marginal impact on the proposed hybrid beamforming schemes since there is almost no performance gap between the ones that are statistically-aided ones and the others. Thus, we can infer from Figs. 4.4 and 4.5 that the impact of the statistical knowledge on increasing the spectral efficiency performance diminishes by increasing codebooks' lengths or equivalently the degradation in the spectral efficiency performance due to the lack of statistical knowledge can be compensated by increasing the rate of the feedback channels.

In Figs. 4.6, 4.7, and 4.8, we plot the spectral efficiency of five different hybrid beamforming schemes for the multi-user case, namely, Algorithm 4, shown in blue solid line with triangles, Algorithm 4 Variation 1 (Alg4.Var1), shown in red solid line with circles, codebook-based sum-rate maximization [44], shown in violet solid line with points, SLNR-based statistical beamforming [95], shown in solid green line with crosses, and JSDM [2, 3, 52], shown in yellow solid line with diamonds in addition to the fully-digital zero-forcing precoding scheme, shown in solid black line. It is worth mentioning that all the simulated hybrid beamforming schemes require perfect knowledge of the effective channel after the analog precoding while the last two benchmarks require only second order statistical knowledge for constructing the analog precoders (none codebook-based).

Fig. 4.6 illustrates that the spectral efficiency performance of the proposed schemes for the $K$-user MISO broadcast channel outperforms all the benchmarks that utilize different channel knowledge. Particularly, there are SNR gaps of $3$ dB, $4$ dB and $8$ dB between the performance of Algorithm 4 and SLNR-based statistical beamforming, codebook-based sum-rate maximization and JSDM. We note here that although JSDM has been shown to be asymptotically optimal under certain channel conditions, it suffers from severe performance degradation as depicted in Fig. 4.6. This is mainly because it requires a sophisticated user grouping strategy and utilizes the

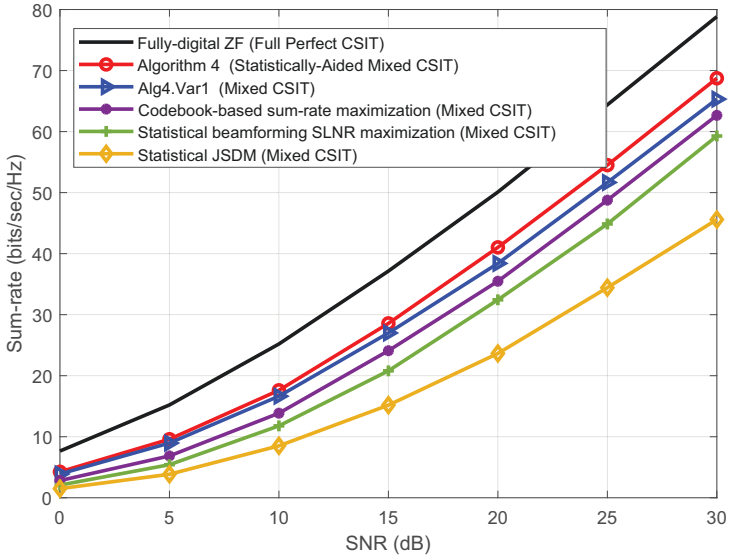


Figure 4.6: Sum-rate comparison of different hybrid beamforming schemes over 10-user MISO BC mmWave channel with $M = 32$ , $N_{cl} = 3$ and $N_{ray} = 2$, and $L = 2$ for Algorithm 4 and Alg4.Var1.

orthogonality between the groups of users. However, here, we consider more realistic channel models where they are characterized by multiple scattering clusters ($N_{cl} = 3$), and where these clusters may overlap.

In Fig. 4.7, the proposed schemes maintain their superior spectral efficiency performance compared to the prior arts even when the number of users increased from $10$ to $16$ while increasing the number of transmitting elements from $32$ to $64$ in comparison to Fig. 4.6. Moreover, Fig. 4.7 shows that the impact of the statistical knowledge on increasing the spectral efficiency of the proposed schemes (e.g., compare the curves of Algorithm 4 and Alg4.Var1) is more pronounced at higher channel dimensions. This is expected since with higher channel dimensions the statistical knowledge becomes more beneficial in directing the information towards users. We note here the performance of the proposed schemes starts to floor at higher SNR.

In Fig. 4.8, we study the effect of increasing the number of users $K$ while fixing the number of transmitting elements to $64$. From Fig. 4.8, one can observe that the prior art schemes, which

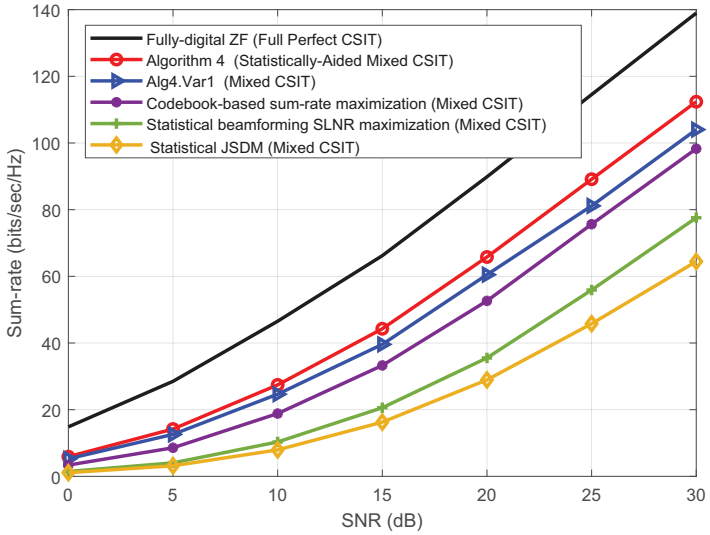


Figure 4.7: Sum-rate comparison of different hybrid beamforming schemes over 16-user MISO BC mmWave channel with $M = 64$, $N_{cl} = 3$ and $N_{ray} = 2$, and $L = 2$ for Algorithm 4 and Alg4.Var1.

utilize only the transmit covariance matrix in the analog beamforming such as JSDM [2, 3, 52] and statistical beamforming SLNR maximization scheme [95], suffer from significant sum-rate flooring or even plunging when $K$ is increased gradually in contrast with the codebook-based schemes such as the proposed schemes and codebook-based sum-rate maximization [44]. Moreover, this illustrates that the proposed schemes maintain their superior sum-rate performance over the entire range of $K$ in comparison with benchmarks. Comparing the sum-rate curves of Algorithm 4 and Alg4.Var1, one deduces that the significance of exploiting the statistical knowledge in enhancing the sum-rate increases with increasing the number of the users, for instance, at $K = 8$, both curves are on top of each other while, at $K = 20$, there is a sum-rate gap of 13 bits/s/Hz in favor to Algorithm 4 (the statistically-aided one). This is mainly due to the considerable role of statistical knowledge in decreasing the inter-user interference by separating users in space based on their covariance matrices.

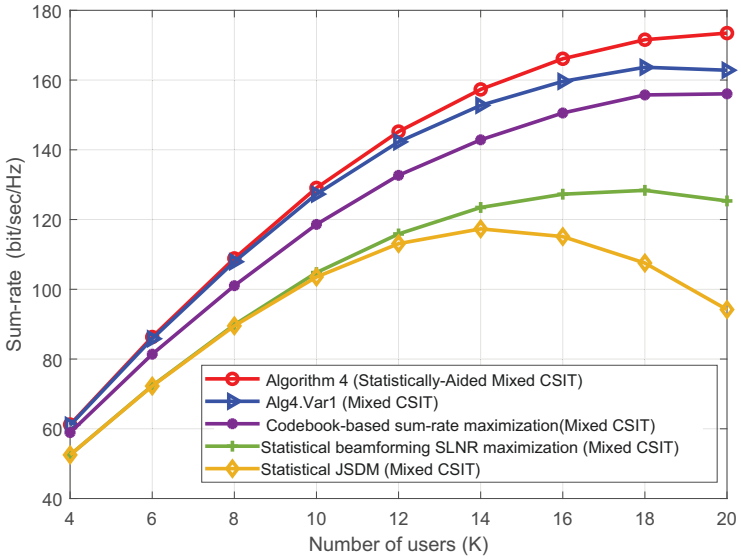


Figure 4.8: Sum-rate comparison of different hybrid beamforming schemes over $K$-user MISO BC mmWave channel with $M = 64$, $N_{cl} = 3$ and $N_{ray} = 2$, and $L = 2$ for Algorithm 4 and Alg4.Var1.

# Chapter 5

# Conclusion and Future Work

## 5.1 Concluding Remarks

In this thesis, we develop hybrid beamforming solutions that cater to next-generation wireless networks. Throughout the thesis, we lay emphasis on practical design considerations that harness the capabilities and employ the properties of the physical layer pillars of next-generation wireless networks; namely, mmWave, massive MIMO and C-RAN. The proposed solutions address the key challenges of hybrid beamforming in mmWave massive MIMO systems, namely, the hardware constraints, the common design for various RF architectures, the high computation complexity, and the overhead of the channel acquisition at the transmitter.

In Chapter 3, we considered single-user and multi-user massive MIMO systems with hybrid precoding and combining. We proposed low-complexity hybrid processing schemes for both partially and fully-connected antenna array structures based on developing a unified two-stage channel diagonalization approach. Particularly, we showed that, instead of directly maximizing the spectral efficiency of the massive MIMO channel, near-optimal and low-complexity hybrid processing schemes are obtained by maximizing the antenna array gain (beamforming gain) of the massive MIMO channel provided by the analog processing while eliminating the interference between data streams and maximizing the spectral efficiency in the digital processing stage. In

this way, contrary to the majority of prior work approaches, the computationally intensive operations such as SVD, matrix inversion, iterative gradient-decent, are implemented only in the low-dimensional digital domain while simpler computational operations such as matrix multiplication are implemented in the high-dimensional analog domain. We presented numerous numerical results on the performance of the proposed schemes. These results showed that the proposed hybrid precoding and combining schemes have marginal spectral efficiency loss compared to the optimal processing strategy while superior (or at least comparable) spectral efficiency performance in comparison with the prior art. Moreover, our analysis demonstrated that the proposed schemes have significant computation complexity saving compared to those of the-state-of-the-art. This gives the advantage to the proposed schemes to be used in the high-dimensional and real-time use-cases of millimeter wave and massive MIMO TDD systems.

In Chapter 4, we considered single-user and multiple-user MIMO hybrid (analog/digital) precoding. For FDD systems, leveraging the spatial correlation and sparsity massive MIMO systems working in mmWave bands, we developed practical and simple codebook-based hybrid precoding strategies assuming limited feedback channel or (mixed) partial channel knowledge while exploiting the statistical information of these channels. The proposed algorithms are designed efficiently to achieve high spectral efficiency while decreasing the feedback overhead by constructing the hybrid precoders based on statistically skewed DFT cookbooks. Numerical results showed that the proposed algorithms allow mmWave systems with much less channel knowledge and feedback overhead to approach the achievable rates of prior arts' schemes that require perfect and mixed channel knowledge. Moreover, numerical results illustrated the potency of assisting the limited feedback systems with statistical information where, for the single-user case, it enhances the immunity of the system's spectral efficiency against the feedback rate reduction, whereas, for the multi-user cases, it consolidates the separability of the users, thereby enhancing the sum-rate of the network. In summary, the advantages of the proposed schemes over the prior art schemes are their simplicity, low-overhead requirement, and near-optimal spectral efficiency.

## 5.2 Directions for Future Work

There are several potential directions for future research. The hybrid beamforming solutions proposed in this thesis are developed and evaluated under certain practical assumptions and setups. However, these schemes would be extended to incorporate more practical assumptions such as RF impairments and non-ideal lossless hardware. Moreover, in this thesis, we consider a single-cell setup or multi-cell one but with low-frequency reuse factor such that there is negligible inter-cell interference.

The vast majority of hybrid beamforming solutions assume using ideal RF circuit components. However, the very high frequencies of mmWave channels impose difficulties and challenges on the design of RF components such as power amplifiers, phase shifters, RF power splitters and combiners, and local oscillators. Therefore, practical RF circuits have some RF hardware impairments such as the non-linearity of power amplifiers, oscillator phase noise, insertion losses of RF power splitters, I/Q imbalance and insertion loss variation of phase shifters [81, 91, 101, 113, 154]. We should develop accurate statistical, analytical, or empirical RF circuit models that incorporate these impairments, different hybrid RF architectures, and various mmWave bands. Studying the impact of these impairments on the performance of mmWave massive MIMO systems is important since they may introduce new design trade-offs and insights that change the way we perceive the current hybrid beamforming designs. For instance, it is theoretically plausible that the fully-connected RF architecture provides higher beamforming gain than the partially-connected one, however this higher beamforming gain may be degraded when taking insertion losses of RF power splitters and phase shifters into consideration [81, 124]. In this case, the interesting trade-off between the hardware complexity and the beamforming gain vanishes and yields the partially-connected architecture an optimal choice.

Eliminating inter-cell interference in conventional cellular networks requires employing coordinated multi-point (CoMP) beamforming techniques. Despite the great theoretical gains, the 3GPP LTE (3rd Generation Partnership Project Long Term Evolution) standardization of CoMP has not achieved many practical gains [51, 128, 140]. This is mainly due to the lack of joint processing, the large channel information overhead required for coordination. Adopting mmWave

bands, massive MIMO technology, as well as new architectures such as C-RAN in the 5G NR allows for achieving the theoretical gains of CoMP transmission techniques. As a result, an extension of the proposed hybrid beamforming schemes can be accomplished to CoMP under some modifications. The proposed schemes have been developed under a sum power constraint. Since the power cannot be shared among base stations, extensions of the proposed schemes should be developed under per base station power constraint in order to fit the CoMP setup. On the other hand, the huge channel information overhead can be handled by a combination of second-ordered statistical channel knowledge and instantaneous knowledge of the effective channels.

# References


[1] IEEE standard for information technology–telecommunications and information exchange between systems–local and metropolitan area networks–specific requirements-part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 3: Enhancements for very high throughput in the 60 Ghz band. *IEEE Std 802.11ad-2012*, pages 1–628, Dec 2012.

[2] A. Adhikary, E. Al Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch. Joint spatial division and multiplexing for mm-wave channels. *IEEE J. Sel. Areas Commun.*, 32(6):1239–1255, Jun. 2014.

[3] Ansuman Adhikary, Junyoung Nam, Jae-Young Ahn, and Giuseppe Caire. Joint spatial division and multiplexing—the large-scale array regime. *IEEE Trans. Inf. Theory*, 59(10):6441–6463, Oct. 2013.

[4] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman. A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives. *IEEE Commun. Surv. Tutor.*, 20(4):3060–3097, Fourthquarter 2018.

[5] M. A. Albreem, M. Juntti, and S. Shahabuddin. Massive MIMO detection techniques: A survey. *IEEE Commun. Surv. Tutor.*, 21(4):3109–3132, Fourthquarter 2019.

[6] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath. Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE J. Sel. Topics Signal Process.*, 8(5):831–846, Oct. 2014.

[7] A. Alkhateeb and R. W. Heath. Frequency selective hybrid precoding for limited feedback millimeter wave systems. *IEEE Transactions on Communications*, 64(5):1801–1818, 2016.

[8] A. Alkhateeb, G. Leus, and R. W. Heath. Limited feedback hybrid precoding for multi-user millimeter wave systems. *IEEE Trans. Wireless Commun.*, 14(11):6481–6494, Nov. 2015.

[9] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath. Mimo precoding and combining solutions for millimeter-wave systems. *IEEE Communications Magazine*, 52(12):122–131, 2014.

[10] Ahmed Alkhateeb, Omar El Ayach, Geert Leus, and Robert W Heath. Hybrid precoding for millimeter wave cellular systems with partial channel knowledge. In *Proc. Inf. Theory Appl. Workshop*, pages 1–5, Feb. 2013.

[11] O. Alluhaibi, Q. Z. Ahmed, C. Pan, and H. Zhu. Capacity maximisation for hybrid digital-to-analog beamforming mm-wave systems. In *Proc. IEEE Global Telecommun. Conf.*, pages 1–6, 2016.

[12] Khaled Ardah, Gábor Fodor, Yuri CB Silva, Walter C Freitas, and Francisco RP Cavalcanti. A unifying design of hybrid beamforming architectures employing phase shifters or switches. *IEEE Transactions on Vehicular Technology*, 67(11):11243–11247, 2018.

[13] IEEE Standards Association et al. Ieee std 802.11 ad-2012,"part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," amendment 3: Enhancements for very high throughput in the 60 Ghz band, IEEE standard for information technology—telecommunications and information exchange between systems—local and metropolitan area networks—specific requirements. *IEEE Computer Society*, 2012.

[14] O. E. Ayach, R. W. Heath, S. Abu-Surra, S. Rajagopal, and Z. Pi. Low complexity precoding for large millimeter wave mimo systems. In *Proc. IEEE ICC*, pages 3724–3729, Jun. 2012.

[15] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath. Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans. Wireless Commun.*, 13(3):1499–1513, 2014.

[16] 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 Telecommun. Conf.*, pages 3476–3480, 2013.

[17] James Baglama and Lothar Reichel. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 27(1):19–42, 2005.

[18] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart. Networks and devices for the 5g era. *IEEE Communications Magazine*, 52(2):90–96, 2014.

[19] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer. Network densification: the dominant theme for wireless evolution into 5g. *IEEE Communications Magazine*, 52(2):82–89, 2014.

[20] T. E. Bogale and L. B. Le. Massive MIMO and mmWave for 5G wireless HetNet: Potential benefits and challenges. *IEEE Veh. Technol. Mag.*, 11(1):64–75, 2016.

[21] J. Chen. Efficient codebook-based beamforming algorithm for millimeter-wave massive MIMO systems. *IEEE Trans. Veh. Technol.*, 66(9):7809–7817, Sep. 2017.

[22] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim. Compressed sensing for wireless communications: Useful tips and tricks. *IEEE Commun. Surv. Tutor.*, 19(3):1527–1550, 2017.

[23] B. Clerckx, Y. Zhou, and S. Kim. Practical codebook design for limited feedback spatial multiplexing. In *Proc. IEEE ICC*, pages 3982–3987, May 2008.

[24] Pierre Courrieu. Fast computation of Moore-Penrose inverse matrices. *Neural Inf. Process. Lett. Rev.*, 8(2):25–29,, Aug. 2005.

[25] Linglong Dai, Xinyu Gao, Jinguo Quan, Shuangfeng Han, and I Chih-Lin. Near-optimal hybrid analog and digital precoding for downlink mmwave massive MIMO systems. In *Communications (ICC), 2015 IEEE International Conference on*, pages 1334–1339. IEEE, 2015.

[26] Robert C Daniels and Robert W Heath Jr. 60 GHz wireless communications: Emerging requirements and design recommendations. *IEEE Vehicular Technology Magazine*, 2(3), 2007.

[27] Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*, volume 16. John Wiley & Sons, 2001.

[28] C. Dehos, J. L. González, A. D. Domenico, D. Kténas, and L. Dussopt. Millimeter-wave access and backhauling: the solution to the exponential data traffic increase in 5g mobile communications systems? *IEEE Communications Magazine*, 52(9):88–95, 2014.

[29] I. S. Dhillon, Jr. R. W. Heath, T. Strohmer, and J. A. Tropp. Constructing packings in grassmannian manifolds via alternating projection. *Experimental Mathematics*, 17(1):9–35, 2008.

[30] T. Ding, Y. Zhao, and L. Li. Overlapped sub-array based hybrid precoding for mmwave massive MIMO system with lens arrays. In *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 537–541, 2019.

[31] Omar El Ayach, Robert W Heath, Shadi Abu-Surra, Sridhar Rajagopal, and Zhouyue Pi. The capacity optimality of beam steering in large millimeter wave MIMO systems. In *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, pages 100–104, Jun. 2012.

[32] G. Femenias and F. Riera-Palou. Cell-free millimeter-wave massive mimo systems with limited fronthaul capacity. *IEEE Access*, 7:44596–44612, 2019.

[33] Global Mobile Data Traffic Forecast. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. *Update*, 2017:2022, 2019.

[34] X. Gao, L. Dai, S. Han, C. I, and R. W. Heath. Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays. *IEEE J. Sel. Areas Commun.*, 34(4):998–1009, 2016.

[35] X. Gao, L. Dai, C. Yuen, and Z. Wang. Turbo-like beamforming based on Tabu search algorithm for millimeter-wave massive MIMO systems. *IEEE Trans. Veh. Technol.*, 65(7):5731–5737, Jul. 2016.

[36] Z. Gao, L. Dai, Z. Wang, and S. Chen. Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO. *IEEE Trans. Signal Process.*, 63(23):6169–6183, 2015.

[37] F. Gholam, J. Via, and I. Santamaria. Beamforming design for simplified analog antenna combining architectures. *IEEE Trans. Veh. Technol.*, 60(5):2373–2378, 2011.

[38] JP Gilb. Ieee standards 802.15. 3c?–part 15.3: wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPANs) amendment 2: millimeter-wave-based alternative physical layer extension [s]. *IEEE Computer Society, New York*, 2009.

[39] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi. Toward 6g networks: Use cases and technologies. *IEEE Communications Magazine*, 58(3):55–61, 2020.

[40] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[41] John C Gower, Garmt B Dijksterhuis, et al. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

[42] Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[43] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[44] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong. Codebook-based hybrid precoding for millimeter wave multiuser systems. *IEEE Trans. Signal Process.*, 65(20):5289–5304, Oct. 2017.

[45] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed. An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE J. Sel. Topics Signal Process.*, 10(3):436–453, 2016.

[46] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed. An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE J. Sel. Topics Signal Process.*, 10(3):436–453, Apr. 2016.

[47] Roger A Hom and Charles R Johnson. Topics in matrix analysis. *Cambridge UP, New York*, 1991.

[48] C. Hu, Y. Wang, Y. . P. Hong, and W. Chen. MMSE hybrid beamforming for weighted sum rate maximization in NOMA systems. In *Proc. 2017 IEEE Global Telecommun. Conf.)*, pages 1–6.

[49] S. Hur, S. Baek, B. Kim, Y. Chang, A. F. Molisch, T. S. Rappaport, K. Haneda, and J. Park. Proposal on millimeter-wave channel modeling for 5G cellular system. *IEEE J. Sel. Topics Signal Process.*, 10(3):454–469, Apr. 2016.

[50] Sooyoung Hur, Taejoon Kim, David J Love, James V Krogmeier, Timothy A Thomas, Amitava Ghosh, et al. Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans. Communications*, 61(10):4391–4403, 2013.

[51] Giovanni Interdonato, Emil Björnson, Hien Quoc Ngo, Pål Frenger, and Erik G Larsson. Ubiquitous cell-free massive mimo communications. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):197, 2019.

[52] J. Jin, C. Xiao, W. Chen, and Y. Wu. Channel-statistics-based hybrid precoding for millimeter-wave MIMO systems with dynamic subarrays. *IEEE Trans. Commun.*, 67(6):3991–4003, Jun. 2019.

[53] Juening Jin, Chengshan Xiao, Wen Chen, and Yongpeng Wu. Hybrid precoding in mmwave MIMO broadcast channels with dynamic subarrays and finite-alphabet inputs. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2018.

[54] N. Jindal. MIMO broadcast channels with finite-rate feedback. *IEEE Trans. Inf. Theory*, 52(11):5045–5060, Nov. 2006.

[55] Xiaorong Jing, Lianghong Li, Hongqing Liu, and Shaoqian Li. Dynamically connected hybrid precoding scheme for millimeter-wave massive MIMO systems. *IEEE Communications Letters*, 22(12):2583–2586, 2018.

[56] Junyi Wang, Zhou Lan, Chang-woo Pyo, T. Baykas, Chin-sean Sum, M. A. Rahman, Jing Gao, R. Funada, F. Kojima, H. Harada, and S. Kato. Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems. *IEEE J. Sel. Areas Commun.*, 27(8):1390–1399, Oct. 2009.

[57] M. Kim and Y. H. Lee. MSE-based hybrid RF/baseband processing for millimeter-wave communication systems in MIMO interference channels. *IEEE Trans. Veh. Technol.*, 64(6):2714–2720, 2015.

[58] N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

[59] Erik G Larsson, Ove Edfors, Fredrik Tufvesson, and Thomas L Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, 2014.

[60] D. Lee. MIMO OFDM channel estimation via block stagewise orthogonal matching pursuit. *IEEE Commun. Lett.*, 20(10):2115–2118, 2016.

[61] J. Lee and N. Jindal. Dirty paper coding vs. linear precoding for MIMO broadcast channels. In *Proc. Asilomar Conf. Signals, Syst. Comput.*, pages 779–783, Oct. 2006.

[62] Xiaofeng Li and Ahmed Alkhateeb. Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems. *arXiv preprint arXiv:1905.13212*, 2019.

[63] Zheda Li, Shengqian Han, and Andreas F Molisch. Hybrid beamforming design for millimeter-wave multi-user massive MIMO downlink. In *Communications (ICC), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

[64] L. Liang, Y. Dai, W. Xu, and X. Dong. How to approach zero-forcing under rf chain limitations in large mmwave multiuser systems? In *2014 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 518–522, 2014.

[65] L. Liang, W. Xu, and X. Dong. Low-complexity hybrid precoding in massive multiuser MIMO systems. *IEEE Wireless Commun. Lett.*, 3(6):653–656, 2014.

[66] T. Lin, J. Cong, Y. Zhu, J. Zhang, and K. Ben Letaief. Hybrid beamforming for millimeter wave systems using the mmse criterion. *IEEE Transactions on Communications*, 67(5):3693–3708, 2019.

[67] A. Liu and V. Lau. Phase only rf precoding for massive mimo systems with limited rf chains. *IEEE Transactions on Signal Processing*, 62(17):4505–4515, 2014.

[68] X. Liu, W. Zou, and S. Chen. Joint design of analog and digital codebooks for hybrid precoding in millimeter wave massive MIMO systems. *IEEE Access*, 6:69818–69825, 2018.

[69] D. J. Love and R. W. Heath. Limited feedback unitary precoding for spatial multiplexing systems. *IEEE Trans. Inf. Theory*, 51(8):2967–2976, Aug. 2005.

[70] D. J. Love and R. W. Heath. Limited feedback diversity techniques for correlated channels. *IEEE Trans. Veh. Technol.*, 55(2):718–722, Mar. 2006.

[71] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews. An overview of limited feedback in wireless communication systems. *IEEE J. Sel. Areas Commun.*, 26(8):1341–1365, Oct. 2008.

[72] D. J. Love, R. W. Heath, and T. Strohmer. Grassmannian beamforming for multiple-input multiple-output wireless systems. *IEEE Trans. Inf. Theory*, 49(10):2735–2747, Oct. 2003.

[73] David James Love and Robert W Heath. Equal gain transmission in multiple-input multiple-output wireless systems. *IEEE Trans. Commun.*, 51(7):1102–1110, 2003.

[74] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang. An overview of massive mimo: Benefits and challenges. *IEEE Journal of Selected Topics in Signal Processing*, 8(5):742–758, 2014.

[75] Zhi-Quan Luo, Wing-Kin Ma, Anthony Man-Cho So, Yinyu Ye, and Shuzhong Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Process. Mag.*, 27(3):20–34, 2010.

[76] M. Majidzadeh, J. Kaleva, N. Tervo, H. Pennanen, A. Tölli, and M. Latva-Aho. Rate maximization for partially connected hybrid beamforming in single-user MIMO systems. In *Proc. IEEE Int. Workshop on Signal Process. Advances in Wireless Commun.*, pages 1–5, 2018.

[77] J. Mao, Z. Gao, Y. Wu, and M. Alouini. Over-sampling codebook-based hybrid minimum sum-mean-square-error precoding for millimeter-wave 3D-MIMO. *IEEE Wireless Commun. Lett.*, 7(6):938–941, 2018.

[78] A. Mejri, M. Hajjaj, S. Hasnaoui, and R. Bouallegue. Power factorization based hybrid precoding design for millimeter wave MIMO transmitters using sub-array structure. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–5, 2017.

[79] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli. Massive MIMO performance with imperfect channel reciprocity and channel estimation error. *IEEE Trans. Commun.*, 65(9):3734–3749, 2017.

[80] R. Méndez-Rial, N. González-Prelcic, and R. W. Heath. Augmented covariance estimation with a cyclic approach in DOA. In *Proc. IEEE ICASSP)*, pages 2784–2788, Apr. 2015.

[81] N. N. Moghadam, G. Fodor, M. Bengtsson, and D. J. Love. On the energy efficiency of mimo hybrid beamforming for millimeter-wave systems with nonlinear power amplifiers. *IEEE Transactions on Wireless Communications*, 17(11):7208–7221, 2018.

[82] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda. Hybrid beamforming for massive MIMO: A survey. *IEEE Commun. Mag.*, 55(9):134–141, 2017.

[83] A. F. Molisch and Xinying Zhang. FFT-based hybrid antenna selection schemes for spatially correlated MIMO channels. *IEEE Communications Letters*, 8(1):36–38, Jan. 2004.

[84] M. M. Molu, P. Xiao, M. Khalily, K. Cumanan, L. Zhang, and R. Tafazolli. Low-complexity and robust hybrid beamforming design for multi-antenna communication systems. *IEEE Trans. Wireless Commun.*, 17(3):1445–1459, 2018.

[85] Junyoung Nam, Jae-Young Ahn, Ansuman Adhikary, and Giuseppe Caire. Joint spatial division and multiplexing: Realizing massive MIMO gains with limited channel state information. In *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, pages 1–6. IEEE, 2012.

[86] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta. Energy and spectral efficiency of very large multiuser MIMO systems. *IEEE Trans. Commun.*, 61(4):1436–1449, 2013.

[87] Duy HN Nguyen, Long Bao Le, Tho Le-Ngoc, and Robert W Heath. Hybrid MMSE precoding and combining designs for mmwave multiuser systems. *IEEE Access*, 5:19167–19181, 2017.

[88] W. Ni and X. Dong. Hybrid block diagonalization for massive multiuser mimo systems. *IEEE Transactions on Communications*, 64(1):201–211, 2016.

[89] W. Ni and X. Dong. Hybrid block diagonalization for massive multiuser MIMO systems. *IEEE Trans. Commun.*, 64(1):201–211, Jan. 2016.

[90] W. Ni, X. Dong, and W. Lu. Near-optimal hybrid processing for massive MIMO systems via matrix decomposition. *IEEE Trans. Signal Process.*, 65(15):3922–3933, 2017.

[91] Yong Niu, Yong Li, Depeng Jin, Li Su, and Athanasios V Vasilakos. A survey of millimeter wave communications (mmwave) for 5g: opportunities and challenges. *Wireless networks*, 21(8):2657–2676, 2015.

[92] S. Park, A. Alkhateeb, and R. W. Heath. Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems. *IEEE Transactions on Wireless Communications*, 16(5):2907–2920, 2017.

[93] S. Park and R. W. Heath. Spatial channel covariance estimation for mmWave hybrid MIMO architecture. In *Proc. Asilomar Conf. Signals, Syst. Comput.*, pages 1424–1428, Nov. 2016.

[94] S. Park and R. W. Heath. Spatial channel covariance estimation for the hybrid MIMO architecture: A compressive sensing-based approach. *IEEE Trans. Wireless Commun.*, 17(12):8047–8062, Dec. 2018.

[95] S. Park, J. Park, A. Yazdan, and R. W. Heath. Exploiting spatial channel covariance for hybrid precoding in massive MIMO systems. *IEEE Trans. Signal Process.*, 65(14):3818–3832, Jul. 2017.

[96] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai. A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Commun. Surv. Tut.*, 20(4):3098–3130, 2018.

[97] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[98] S. Payami, M. Ghoraishi, and M. Dianati. Hybrid beamforming for large antenna arrays with phase shifter selection. *IEEE Transactions on Wireless Communications*, 15(11):7258–7271, 2016.

[99] S. Payami, M. Khalily, T. H. Loh, and K. Nikitopoulos. Hybrid beamforming with switches and phase shifters over frequency-selective channels. *IEEE Wireless Communications Letters*, pages 1–1, 2020.

[100] S. Payami, M. Sellathurai, and K. Nikitopoulos. Low-complexity hybrid beamforming for massive mimo systems in frequency-selective channels. *IEEE Access*, 7:36195–36206, 2019.

[101] N. Peccarelli, B. James, R. Irazoqui, J. Metcalf, C. Fulton, and M. Yeary. Survey: Characterization and mitigation of spatial/spectral interferers and transceiver nonlinearities for 5g mimo systems. *IEEE Transactions on Microwave Theory and Techniques*, 67(7):2829–2846, 2019.

[102] Z. Pi. Optimal transmitter beamforming with per-antenna power constraints. In *2012 IEEE International Conference on Communications (ICC)*, pages 3779–3784, 2012.

[103] Zhouyue Pi and Farooq Khan. An introduction to millimeter-wave mobile broadband systems. *IEEE Communications Magazine*, 49(6), 2011.

[104] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi. Integrated access and backhaul in 5g mmwave networks: Potential and challenges. *IEEE Communications Magazine*, 58(3):62–68, 2020.

[105] C. Qi, Y. Huang, S. Jin, and L. Wu. Sparse channel estimation based on compressed sensing for massive MIMO systems. In *Proc. IEEE Int. Conf. Commun.*, pages 4558–4563, Jun. 2015.

[106] R. Rajashekar and L. Hanzo. Iterative matrix decomposition aided block diagonalization for mm-wave multiuser MIMO systems. *IEEE Trans. Wireless Commun.*, 16(3):1372–1384, 2017.

[107] X. Rao and V. K. N. Lau. Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems. *IEEE Trans. Signal Process.*, 62(12):3261–3271, 2014.

[108] Theodore S Rappaport, Robert W Heath Jr, Robert C Daniels, and James N Murdock. *Millimeter wave wireless communications*. Pearson Education, 2015.

[109] 3GPP TS 36.211 V11.1.0. Evolved universal terrestrial radio access; physical channels and modulation. Feb. 2013.

[110] W. Roh, J. 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*, 52(2):106–113, 2014.

[111] C. Rusu, R. Mèndez-Rial, N. González-Prelcic, and R. W. Heath. Low complexity hybrid precoding strategies for millimeter wave communication systems. *IEEE Trans. Wireless Commun.*, 15(12):8380–8393, 2016.

[112] M. Sadek, A. Tarighat, and A. H. Sayed. A leakage-based precoding scheme for downlink multi-user MIMO channels. *IEEE Trans. Wireless Commun.*, 6(5):1711–1721, May 2007.

[113] Murat Babek Salman and Gokhan Muzaffer Guvensen. A comprehensive performance analysis for mm-wave massive mimo hybrid beamforming under pa nonlinearities. *arXiv preprint arXiv:2006.15930*, 2020.

[114] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath. Optimal designs for space-time linear precoders and decoders. *IEEE Trans. Signal Process.*, 50(5):1051–1064, 2002.

[115] H. Seleem, A. I. Sulyman, and A. Alsanie. Hybrid precoding-beamforming design with Hadamard RF codebook for mmWave large-scale MIMO systems. *IEEE Access*, 5:6813–6823, 2017.

[116] A. W. Shaban, O. Damen, Y. Xin, and E. Au. Statistically-aided codebook-based hybrid precoding for millimeter wave channels. *IEEE Access*, 8:101500–101513, 2020.

[117] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder. 5g: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, 35(6):1201–1221, 2017.

[118] W. Shen, L. Dai, Y. Shi, Z. Gao, and Z. Wang. Massive MIMO channel estimation based on block iterative support detection. In *Proc. IEEE WCNC*, pages 1–6, 2016.

[119] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen. Large-scale convex optimization for ultra-dense cloud-ran. *IEEE Wireless Communications*, 22(3):84–91, 2015.

[120] F. Sohrabi and W. Yu. Hybrid digital and analog beamforming design for large-scale antenna arrays. *IEEE J. Sel. Topics Signal Process.*, 10(3):501–513, 2016.

[121] F. Sohrabi and W. Yu. Hybrid analog and digital beamforming for mmwave ofdm large-scale antenna arrays. *IEEE Journal on Selected Areas in Communications*, 35(7):1432–1443, 2017.

[122] J. Song, J. Choi, and D. J. Love. Codebook design for hybrid beamforming in millimeter wave systems. In *Proc. IEEE ICC*, pages 1298–1303, Jun. 2015.

[123] N. Song, T. Yang, and H. Sun. Overlapped subarray based hybrid beamforming for millimeter wave multiuser massiveMIMO. *IEEE Signal Processing Letters*, 24(5):550–554, 2017.

[124] X. Song, T. Kühne, and G. Caire. Fully-/partially-connected hybrid beamforming architectures for mmwave mu-mimo. *IEEE Transactions on Wireless Communications*, 19(3):1754–1769, 2020.

[125] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt. Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels. *IEEE Trans. Signal Process.*, 52(2):461–471, Feb. 2004.

[126] Quentin H Spencer, A Lee Swindlehurst, and Martin Haardt. Zero-forcing methods for downlink spatial multiplexing in multiuser mimo channels. *IEEE transactions on signal processing*, 52(2):461–471, 2004.

[127] R. G. Stephen and R. Zhang. Joint millimeter-wave fronthaul and ofdma resource allocation in ultra-dense cran. *IEEE Transactions on Communications*, 65(3):1411–1423, 2017.

[128] S. Sun, T. S. Rappaport, and M. Shaft. Hybrid beamforming for 5g millimeter-wave multi-cell networks. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 589–596, 2018.

[129] W. Tan, M. Matthaiou, S. Jin, and X. Li. Spectral efficiency of DFT-based processing hybrid architectures in massive MIMO. *IEEE Wireless Commun. Lett.*, 6(5):586–589, Oct. 2017.

[130] W. Tan, M. Matthaiou, S. Jin, and X. Li. Spectral efficiency of DFT-based processing hybrid architectures in massive MIMO. *IEEE Wireless Communications Letters*, 6(5):586–589, 2017.

[131] S. Tsai. Equal gain transmission with antenna selection in MIMO communications. *IEEE Trans. Wireless Commun.*, 10(5):1470–1479, 2011.

[132] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[133] Junyi Wang, Zhou Lan, Chin-Sean Sum, Chang-Woo Pyo, Jing Gao, Tuncer Baykas, Azizur Rahman, Ryuhei Funada, Fumihide Kojima, Ismail Lakkis, et al. Beamforming codebook design and performance evaluation for 60GHz wideband WPANs. In *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, pages 1–6. IEEE, 2009.

[134] Z. Wang, J. Zhu, J. Wang, and G. Yue. An overlapped subarray structure in hybrid millimeter-wave multi-user MIMO system. In *2018 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, 2018.

[135] H. Weingarten, Y. Steinberg, and S. Shamai. The capacity region of the gaussian MIMO broadcast channel. In *Proc. IEEE Int. Symp. Inf. Theory*, page 174, Jun. 2004.

[136] Z. Xiao, P. Xia, and X. Xia. Channel estimation and hybrid precoding for millimeter-wave mimo systems: A low-complexity overall solution. *IEEE Access*, 5:16100–16110, 2017.

[137] Xinying Zhang, A. F. Molisch, and Sun-Yuan Kung. Variable-phase-shift-based rf-baseband codesign for mimo antenna selection. *IEEE Transactions on Signal Processing*, 53(11):4091–4103, 2005.

[138] Yi Xu, Guosen Yue, Narayan Prasad, Sampath Rangarajan, and Shiwen Mao. User grouping and scheduling for large scale mimo systems with two-stage precoding. In *Communications (ICC), 2014 IEEE International Conference on*, pages 5197–5202. IEEE, 2014.

[139] Z. Xu, S. Han, Z. Pan, and C. I. Alternating beamforming methods for hybrid analog and digital MIMO transmission. In *Proc. IEEE International Conf. Commun.*, pages 1595–1600, 2015.

[140] Chenyang Yang, Shengqian Han, Xueying Hou, and Andreas F Molisch. How do we design comp to achieve its promised potential? *IEEE Wireless Communications*, 20(1):67–74, 2013.

[141] D. Yang, L. . Yang, and L. Hanzo. DFT-based beamforming weight-vector codebook design for spatially correlated channels in the unitary precoding aided multiuser downlink. In *Proc. IEEE ICC*, pages 1–5, May 2010.

[142] G. Yang, Y. Zhou, and W. Xia. Virtual channel optimization for overloaded MIMO systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(3):331–335, 2018.

[143] H. Yang and T. L. Marzetta. Performance of conjugate and zero-forcing beamforming in large-scale antenna systems. *IEEE Journal on Selected Areas in Communications*, 31(2):172–179, 2013.

[144] K. Ying, Z. Gao, S. Lyu, Y. Wu, H. Wang, and M. Alouini. GMD-based hybrid beamforming for large reconfigurable intelligent surface assisted millimeter-wave massive MIMO. *IEEE Access*, 8:19530–19539, 2020.

[145] Xianghao Yu, Juei-Chin Shen, Jun Zhang, and Khaled B Letaief. Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems. *IEEE J. Sel. Topics Signal Process.*, 10(3):485–500, 2016.

[146] F. Yuan, S. Han, C. Yang, Y. Zhang, G. Wang, and M. Lei. Weighted DFT codebook for multiuser MIMO in spatially correlated channels. In *Proc. IEEE VTC*, pages 1–5, May 2011.

[147] Fuzhen Zhang and Qingling Zhang. Eigenvalue inequalities for matrix product. *IEEE Transactions on Automatic Control*, 51(9):1506–1509, 2006.

[148] Z. Zhang, X. Wu, and D. Liu. Joint precoding and combining design for hybrid beamforming systems with subconnected structure. *IEEE Syst. J.*, pages 1–12, 2019.

[149] L. Zhao, D. W. K. Ng, and J. Yuan. Multi-user precoding and channel estimation for hybrid millimeter wave systems. *IEEE Journal on Selected Areas in Communications*, 35(7):1576–1590, 2017.

[150] Nan Zhao, F Richard Yu, Minglu Jin, Qiao Yan, and Victor CM Leung. Interference alignment and its applications: A survey, research issues, and challenges. *IEEE Communications Surveys & Tutorials*, 18(3):1779–1803, 2016.

[151] Z. Zhou, J. Fang, L. Yang, H. Li, Z. Chen, and R. S. Blum. Low-rank tensor decomposition-aided channel estimation for millimeter wave mimo-ofdm systems. *IEEE Journal on Selected Areas in Communications*, 35(7):1524–1538, 2017.

[152] Z. Zhou, J. Fang, L. Yang, H. Li, Z. Chen, and S. Li. Channel estimation for millimeter-wave multiuser MIMO systems via parafac decomposition. *IEEE Trans. Wireless Commun.*, 15(11):7501–7516, 2016.

[153] Yazhou Zhu, Qingchuan Zhang, and Tao Yang. Low-complexity hybrid precoding with dynamic beam assignment in mmwave OFDM systems. *IEEE Transactions on Vehicular Technology*, 67(4):3685–3689, 2017.

[154] Y. Zou, P. Zetterberg, U. Gustavsson, T. Svensson, A. Zaidi, T. Kadur, W. Rave, and G. Fettweis. Impact of major rf impairments on mm-wave communications using ofdm waveforms. In *2016 IEEE Globecom Workshops (GC Wkshps)*, pages 1–7, 2016.

# APPENDICES

# Appendix A

# Hybrid Beamforming Schemes for TDD Systems

## A.1 Poof of Proposition 3.2.3

Utilizing the singular value decomposition of the $i^{\text{th}}$ sub-channel on the diagonal, i.e., $\mathbf{H}_{ii} = \sum_l \sigma_{iil}\mathbf{u}_{iil}\mathbf{v}_{iil}^H$, and using the optimal combining vector, i.e., the left singular vector associated with the largest singular value $\mathbf{u}_1$ instead of $\frac{e^{-j\boldsymbol{\phi}_i}}{\sqrt{N_{sa}}}$, however it violates the constant magnitude of entries constraint on the analog combiner, the antenna array gain is upper bounded by:

$$\frac{|e^{-j\boldsymbol{\phi}_i}\mathbf{H}_{ii}e^{j\boldsymbol{\theta}_i}|^2}{N_{sa}M_{sa}} < |\mathbf{u}_1^H \sum_l \sigma_l \mathbf{u}_l \mathbf{v}_l^H \frac{e^{j\boldsymbol{\theta}_i}}{\sqrt{M_{sa}}}|^2 \tag{A.1}$$

$$\overset{\text{(a)}}{=} \frac{\sigma_1^2}{M_{sa}}|\mathbf{v}_1^H e^{j\boldsymbol{\theta}_i}|^2$$

$$\overset{\text{(b)}}{\leq} \frac{\sigma_1^2}{M_{sa}}\|\mathbf{v}_1\|_1^2 \tag{A.2}$$

where (A.2, a) is due to the unitary property of the singular vectors, i.e., $\mathbf{u}_l^H\mathbf{u}_j = 0, \forall j \neq l$ and $\mathbf{u}_l^H\mathbf{u}_l = 1$. Eq. (A.2, b) is obtained based on solving $\max_{\boldsymbol{\theta}_i}|\mathbf{v}_1^H e^{j\boldsymbol{\theta}_i}|^2$ by co-phasing $\mathbf{v}_1$ and $e^{j\boldsymbol{\theta}_i}$,

i.e., $\boldsymbol{\theta}_{\star i} = \angle \mathbf{v}_1$. Similarly, $\frac{|e^{-j\boldsymbol{\phi}_i}\mathbf{H}_{ii}e^{j\boldsymbol{\theta}_i}|^2}{N_{sa}M_{sa}} < \frac{\sigma_1^2}{N_{sa}}\|\mathbf{u}_1\|_1^2$ comes readily by considering the optimal precoder, i.e., $\mathbf{v}_1$ instead $\frac{e^{j\boldsymbol{\theta}_i}}{\sqrt{M_{sa}}}$. After that, one has to consider the minimum of both array gains, i.e., $\frac{\sigma_1^2}{M_{sa}}\|\mathbf{v}_1\|_1^2$ and $\frac{\sigma_1^2}{N_{sa}}\|\mathbf{u}_1\|_1^2$ since both are the antenna array gains of enhanced setups due to considering the optimal combiner and precoder, respectively, while relaxing the constant magnitudes of entries constraint. On the other hand, the lower bound is obtained by considering selection precoding/combining, instead of considering the optimal precoder/combiner used to obtain the upper bound, at one side while keeping the analog combining/precoding at the other side. Therefore, the array gain of the analog precoding and combining at both side is lower bounded by the array gain of either selection precoding and analog combining or analog precoding and selection combining strategies. Thus, the array gain of the analog precoding and combining strategy:

$$\frac{|e^{-j\boldsymbol{\phi}_i}\mathbf{H}_{ii}e^{j\boldsymbol{\theta}_i}|^2}{N_{sa}M_{sa}} > \frac{1}{\sqrt{M_{sa}}}\max_{1\leq j\leq N_{sa}}|(\mathbf{H}_{ii}e^{j\boldsymbol{\theta}_i})_j|^2 \overset{\text{(a)}}{=} \frac{\|\mathbf{H}_{ii}\|_\infty^2}{M_{sa}} \tag{A.3}$$

where (A.3, a) is obtained by co-phasing $e^{j\boldsymbol{\theta}_i}$ to the row that has the maximum $\ell_1$ norm [73]. Similarly, $\frac{|e^{-j\boldsymbol{\phi}_i}\mathbf{H}_{ii}e^{j\boldsymbol{\theta}_i}|^2}{N_{sa}M_{sa}} > \frac{\|\mathbf{H}_{ii}\|_1^2}{N_{sa}}$ comes readily by considering the selection beamforming and analog combining as the transceiver strategy. Since, both setups have degraded performance compared to the analog precoding and combining strategy, the array gain of latter strategy is lower bounded by the maximum of array gains of both former strategies.