Practical Precoding Design for Modern Multiuser MIMO Communications

by

Le Liang
B.Eng., Southeast University, 2012

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Practical Precoding Design for Modern Multiuser MIMO Communications

by

Le Liang
B.Eng., Southeast University, 2012

Supervisory Committee

Dr. Xiaodai Dong, Supervisor
(Department of Electrical and Computer Engineering)

Dr. T. Aaron Gulliver, Departmental Member
(Department of Electrical and Computer Engineering)

Supervisory Committee

Dr. Xiaodai Dong, Supervisor
(Department of Electrical and Computer Engineering)

Dr. T. Aaron Gulliver, Departmental Member
(Department of Electrical and Computer Engineering)

Abstract

The use of multiple antennas to improve the reliability and capacity of wireless communication has been around for a while, leading to the concept of multiple-input multiple-output (MIMO) communications. To enable full MIMO potentials, the precoding design has been recognized as a crucial component. This thesis aims to design multiuser MIMO precoders of practical interest to achieve high reliability and capacity performance under various real-world constraints like inaccuracy of channel information acquired at the transmitter, hardware complexity, etc. Three prominent cases are considered which constitute the mainstream evolving directions of the current cellular communication standards and future 5G cellular communications. First, in a relay-assisted multiuser MIMO system, heavily quantized channel information obtained through limited feedback contributes to noticeable rate loss compared to when perfect channel information is available. This thesis derives an upper bound to characterize the system throughput loss caused by channel quantization error, and then develops a feedback quality control strategy to maintain the rate loss within a bounded range. Second, in a massive multiuser MIMO channel, due to the large array size, it is difficult to support each antenna with a dedicated radio frequency chain, thus making high-dimensional baseband precoding infeasible. To address this challenge, a low-complexity hybrid precoding scheme is designed to divide the precoding into two cascaded stages, namely, the low-dimensional baseband precoding and the high-dimensional phase-only processing at the radio frequency domain. Its performance

is characterized in a closed form and demonstrated through computer simulations. Third, in a mmWave multiuser MIMO scenario, smaller wavelengths make it possible to incorporate excessive amounts of antenna elements into a compact form. However, we are faced with even worse hardware challenges as mixed signal processing at mmWave frequencies is more complex and power consuming. The channel sparsity is taken advantage of in this thesis to enable a simplified precoding scheme to steer the beam for each user towards its dominant propagation paths at the radio frequency domain only. The proposed scheme comes at significantly reduced complexity and is shown to be capable of achieving highly desirable performance based on asymptotic rate analysis.

Acknowledgement

There are a number of people I wish to thank for making my experience as a graduate student at the University of Victoria as exciting and rewarding as it is. Foremost, I would like to thank my supervisor Professor Xiaodai Dong for her consistent guidance and support throughout my research and study in the past three years. She is such a nice person and friend, from whom I have benefited tremendously in many aspects, including both career and life. I am also very grateful to Professor Wei Xu from Southeast University, who also served as my supervisor during my undergraduate study. His continual guidance and close collaboration over the past years have always helped me push towards the right direction in research endeavors. I would also like to thank Professor Wu-Sheng Lu for delivering such great courses as Engineering Optimization in a way that can be understood so easily and meanwhile deeply.

I am also profoundly indebted to my friends and roommates Leyuan Pan and Yongyu Dai for so many stimulating discussions and so much fun during the time we spent together in the lovely house. I also wish to express my heartfelt gratitude to all members of our research group, from whom I have learned a lot. In particular, I would like to thank Yi Shi, Weiheng Ni, Zheng Xu, Wanbo Li, Guang Zeng, Yiming Huo, Ping Cheng, Tong Xue, Farnoosh Talaei for many inspiring discussions. Weiheng contributed a lot to the proof in Chapter 3. Leyuan and Yongyu provided invaluable help in developing the analysis presented in Chapter 4. Finally, I express my sincere gratitude to all my friends, both inside and outside of UVic, whose friendship makes my three years in the beautiful city of Victoria so wonderful and unforgettable. This thesis is dedicated to my family whose support and love have always been a source of courage and confidence for me.

I acknowledge the Natural Sciences and Engineering Research Council of Canada and the University of Victoria Graduate Awards program for providing financial support for my Master Studies.

Le Liang

Contents

List of Figures

Acronyms

BD block diagonalization.

BS base station.

CS compressed sensing.

CSI channel state information.

DPC dirty paper coding.

FDD frequency division duplexing.

LAN local area network.

LTE-A long term evolution-advanced.

MIMO multiple-input multiple-output.

mmWave millimeter wave.

OFDM orthogonal frequency division multiplexing.

PA power amplifier.

RF radio frequency.

RS relay station.

SINR signal-to-noise-plus-interference ratio.

SNR signal-to-noise ratio.

SVD singular value decomposition.

WLAN wireless local area network.

WPAN wireless personal area network.

ZF zero forcing.

Chapter 1

Introduction

Academic research efforts and industrial practices on the use of multiple antennas at the transmitter and/or the receiver in wireless communication systems have been around for years under the terminology of multiple-input multiple-output (MIMO). In recent years, the MIMO technology has been successfully integrated into a number of well-established standards, e.g., the 4G long term evolution-advanced (LTE-A) cellular communication system, local area network (LAN) standards 802.11n, etc. It is also considered as an integral part of future 5G communications, which is envisioned to be able to provide 1000x increase in network throughput and provide better coverage [1].

With multiple antennas at the two communication ends, the transmitted signals arrive at the receiver side after passing through different propagation paths and are incident upon different antennas mounted at the receiver. Consider the case where there is only one stream ready to be communicated over the link. Two interesting things to notice. One is at the receiver side, we have received multiple copies of the same transmitted waveform, each from an aggregation of a number of replicas obtained from different paths. Coherent addition of the multiple received copies would significantly increase the received signal-to-noise ratio (SNR) of the original transmitted signals. The other point to make is that in a MIMO system, the transmitter side is also capable of forming the signal beams to some specific directions so that different delayed versions of transmitted waveforms will end up in coherent addition at the receiver. Then the received SNR can be further improved. Exploiting multiple antennas in this way is mainly to harvest the so-called diversity gain, which is helpful to improve communication reliability by reducing received bit/symbol error rates. Another merit of MIMO is the capability to multiplex more data streams to be sent

over the pairs of links between multiple antennas at both sides through precoding and combining at the transmitter and receiver, respectively. Thus the name multiplexing gain is termed to characterize such benefits, which will substantially increase the communication capacity of the system under, however, some assumptions. The most common assumption made is there exists a rich scattering environment to make the channel full rank, which may not always be the case in practice unfortunately. The single communication pair discussed so far is often referred to as single-user MIMO or point-to-point MIMO.

If we have multiple receivers trying to communicate with a single transmitter equipped with multiple antennas, then it's ended up in a multiuser MIMO scenario. This is the typical case in a cellular system where the base station (BS) is responsible for communicating with a multitude of users. Generally, this communication scenario is of more practical interest while at the same time is fundamentally different from single-user MIMO. First, the multiple users access the BS through the same propagation medium using the same frequency and time resources [2] without being able to collaborate among themselves either to transmit or receive. This will inevitably cause interference to each other and the system is essentially interference limited if it is not handled properly. Second, since the many users tend to be at different locations, they are more likely to see totally different channels from the BS. In this case, the aggregated MIMO system composed of the transmitter and multiple users is in a better position to reap multiplexing gains as the channel now has better chance to be of full rank. The unfavorable channel conditions can thus be overcome given the angular separation of users exceeds the Rayleigh resolution of the transmit array. Lots of research efforts in recent years have been invested in understanding the information-theoretical capacity of multiuser MIMO channels and designing efficient channel coding and signal processing algorithms to pursue the capacity limit.

Generally, in cellular communication systems, we usually want the complicated signal processing to be done at the more capable BS rather than power-limited user ends. Hence, precoding design at the transmitter is more of practical significance to facilitate high rate communications in current and future cellular systems. In the meanwhile, the real-world implementation of MIMO communications always comes at various practical constraints and is often compromised in performance as compared to theoretical studies under some perfect assumptions. For instance, to perform efficient transmit precoding, the BS needs to get access to the downlink channel state information (CSI), which is usually estimated at the receiver side based on the pilot

symbols sent at the beginning of each coherent time period by the BS. This estimated CSI is often quantized and then sent back to the BS over a finite-rate link. Let alone the estimation inaccuracy, the quantization error of CSI will be detrimental to the system and will contribute to residual interuser interference even if, e.g., zero-forcing precoding is performed at the BS intended to null out interference among users.

This thesis aims at designing efficient precoding schemes for next-generation wireless cellular systems to maximize the benefits of multiple antennas under various practical constraints. Three mainstream multiuser MIMO scenarios are considered and methods to address practical challenges have been proposed herein. In the rest of this chapter, practical motivations for the thesis are described in greater detail, and the main results are outlined with notations used throughout the thesis introduced in the end.

1.1 Motivation

The relay technology is proposed to extend cell coverage in a multiuser environment and improve the performance especially for cell-boundary users. Data streams from the BS are first transmitted to a relay station (RS) and then forwarded to remote users which might be located far away from the BS. As understood from numerous existing literature [4]–[10], in multi-antenna relay channels, good knowledge of CSI at the transmitter is important to achieve multiplexing gains promised by the MIMO technology. Normally, the channel between the BS and RS is estimated at the RS and then sent back to the BS for precoding purposes. The channels from the RS to a multiple of users are accordingly estimated at the user sides and then fed back to the RS.

That being said, even if we assume that the downlink CSI is perfectly estimated with no errors at the RS and users, respectively, the CSI obtained at the transmitters will have finite precision as the quantization error is inevitable if the limited feedback technology is employed to transfer the channel knowledge from receivers to transmitters. Studies from [6] and [7] have shown that CSI mismatch will cause severe interuser interference and make the communication rates achieved by the system eventually saturate at high SNR despite the use of zero forcing (ZF) precoding at the transmitters in a broadcast channel. In fact, this is also the case in a relay-assisted multiuser MIMO channel as shown in Fig. 1.1 where optimal point-to-point MIMO-based precoding and combining is performed at the BS and RS along with

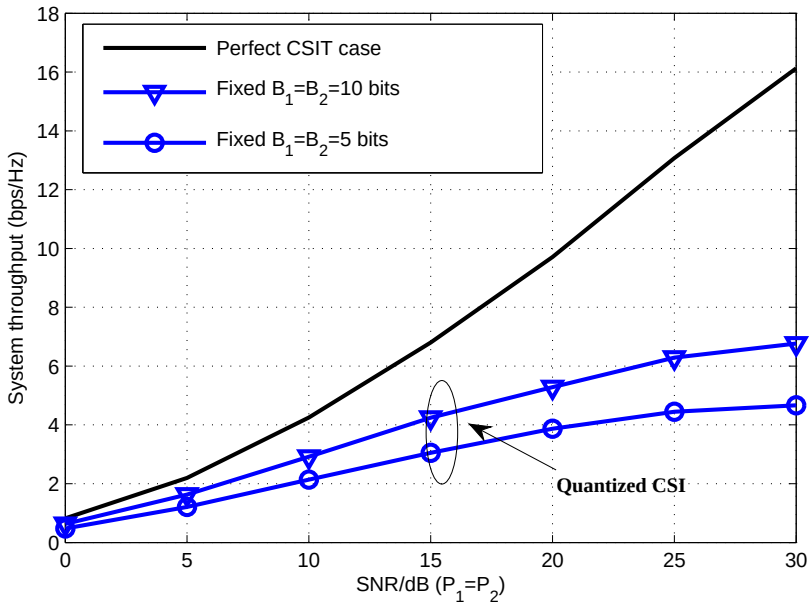

Figure 1.1: Illustration of rate saturation at high SNR with quantized CSI in a multi-antenna relay channel with $M = 6, N_t = 4, N_r = 2$, and $K = \frac{N_t}{N_r} = 2$.

block-diagonalization ZF precoding applied at the RS. P_1 and P_2 are the transmit powers at the BS and RS, respectively while B_1 and B_2 are the numbers of feedback bits of the BS-RS link and RS-User channels. M, N_t and N_r stand for the number of antennas at the BS, RS, the users, respectively. K is the number of users. It is observed that the the system throughput achieved by the limited feedback strategy suffers increasing rate loss compared to the perfect CSI at the transmitters (CSIT) case and is eventually saturating at high SNR. Therefore, it is of great significance to understand the adverse effects of the CSI quantization errors on the system performance and quantify the rate loss associated with a particular feedback precision setting if possible. Given these understandings, we might be able to come up with new approaches to compensate for the performance loss with, e.g., increased feedback quality or other enhanced measures. This motivates the work of the first chapter in this thesis.

Another popular trend of modern wireless communications is to equip the BS with a large array of antennas to facilitate high throughput given a limited spectrum. In a multiuser environment with rich scattering, installing more antennas at the BS would always help [28]. Greater numbers of antennas would enable the system to climb out of noise and simultaneously serve a multiplicity of terminals. Waveforms are directed towards intended terminals with more accuracy and less interference is

caused to adjacent terminals due to the asymptotic orthogonality of channels seen from different terminals with increasing antenna numbers. In addition, more antennas mean more beamforming gain. The system is capable of exploiting such beamforming gain to considerably increase the received power at serviced terminals. Thus greater throughput can potentially be supported in such systems with simple linear precoding schemes such as ZF precoding.

However, to reap such promising gains from mounting more antennas at the BS, we are still faced with several critical challenges, with a stark one being the substantially increased hardware complexity given current implementation of MIMO technologies. The de facto implementation is to perform all the necessary precoding at the digital baseband on incoming complex symbols, which is then transformed to the radio frequency (RF) waveform after passing through digital-to-analog converters (D/A), mixers, power amplifier (PA), etc. This implementation method thus requires a dedicated RF chain to support such operations for each antenna element before the precoded waveform is transmitted over the air. With a modest number of antennas, such as at most 8 in LTE-A systems, this implementation is reasonable. But when the antenna size scales large, as currently investigated in lots of literature under the terminology of massive MIMO [28]–[30], this hardware complexity becomes formidable. Therefore, a low-complexity implementation of efficient precoding schemes is needed to address such issues before we can tangibly benefit from the massive number of antennas at the BS. Apart from this, many practitioners in the industry are also concerned about the potentially very big physical size of antennas if a very large antenna array is indeed deployed at the BS. Fortunately, the emerging study of communications at the millimeter wave (mmWave) frequencies gives a good answer to this problem as the dramatic decrease of the carrier wavelength of mmWave bands can easily enable incorporation of large arrays of antennas into a very compact form. Better still, large chunk of underutilized spectrum at the mmWave frequencies provides another perspective of facilitating high speed communications, i.e., use of wider wireless spectrum. Nevertheless, communication at mmWave frequencies comes with a lot of different characteristics than its low frequency counterparts, such as higher mixed signal processing costs, sparse channel multipath components, etc. Hence it is needed to design efficient precoding schemes specialized to the mmWave communications on top of just trying to exploit the excessive numbers of antennas as considered in the massive multiuser MIMO scenario at lower frequencies. These observations and analysis facilitate the works of Chapters 3 and 4.

1.2 Overview of Thesis

This thesis focuses on understanding various practical constraints in modern multiuser MIMO systems and designing efficient precoding schemes to maximize the capacity gains provided by the MIMO technology. Three different yet inherently related multiuser MIMO scenarios are considered in the main chapters, namely, limited feedback-based precoding for relay-assisted multiuser MIMO in Chapter 2, low-complexity precoding for massive MIMO communications in Chapter 3, and sparse precoding design for RF chain-limited mmWave MIMO communications in Chapter 4.

In Chapter 2, we study the multi-antenna MIMO relay downlink channel with limited feedback CSI from both BS-RS and RS-User links. Data streams from the BS are first transmitted to an intermediate RS with singular value decomposition-based precoding and receiver combining at the BS and RS, respectively. Block diagonalization precoding is then applied at the RS to forward the received signals to the remote multi-antenna users. An upper bound is derived to characterize the system throughput loss due to CSI quantization error. Then the thesis proposes to scale feedback bits of both links with respect to the transmit power at the BS and RS to maintain a bounded rate loss. The proposed scaling law provides insights into the interplay of the transmit powers and feedback quality and gives useful guidelines for practical feedback design in wireless relay systems.

Chapter 3 begins with the observation that hardware complexity is more of an issue when massive amounts of antennas are mounted at the BS to facilitate high speed communications. To address this practical challenge, it proposes a low-complexity hybrid precoding scheme to approach the performance of the traditional baseband ZF precoding (referred to as full-complexity ZF), which is considered a virtually optimal linear precoding scheme in massive multiuser MIMO systems. The proposed hybrid precoding scheme, named phased-ZF (PZF), essentially applies phase-only control at the RF domain and then performs a low-dimensional baseband ZF precoding based on the effective channel seen from baseband. Heavily quantized RF phase control up to 2 bits of precision is also considered and shown to incur very limited degradation.

In Chapter 4, it is shown that although the mmWave band is a suitable candidate for realizing massive MIMO benefits, it comes at costs and requires specialized design. Apart from the aforementioned hardware complexity issues (particularly RF chain limitations), the most striking feature of communication over mmWave frequencies is the very sparse scattering nature of the channel, which is fundamentally differ-

ent than its low frequency counterparts. We take into consideration of the channel sparsity and the RF chain limitations present in large mmWave multiuser systems and propose to approach the desirable yet infeasible full-complexity ZF precoding by essentially steering the beam towards each user's strongest path, termed MUBS, through analog processing at the RF domain. Theoretical analysis is also developed to validate the desirable performance of the proposed MUBS scheme. Based on this, two enhancements are further made by exploiting a hybrid precoding structure, which are shown to achieve closer performance to the pure baseband implementation of ZF precoding but with substantially reduced hardware complexity.

The last chapter summarizes key points in the thesis and gives concluding remarks.

1.3 Notations

The following notation is used throughout this thesis: The bold upper case letters are used to denote matrices, e.g., $\mathbf{X}$, $\mathbf{Y}$. Lower case letters are used to denote vectors, e.g., $\mathbf{x}$, $\mathbf{y}$. An n-dimensional identity matrix is denoted by $\mathbf{I}_N$. For a matrix $\mathbf{A}$, $\mathbf{A}^T$, $\mathbf{A}^H$, $|\mathbf{A}|$ and $\mathrm{tr}(\mathbf{A})$ return its transpose, conjugate transpose, determinant and trace, respectively. $\mathbb{E}[\cdot]$ is the expectation operator. $\mathbb{C}$ denotes the complex space. $\Re(\cdot)$ and $\Im(\cdot)$ return the real and imaginary parts of a complex number, respectively. The notation $y = o(x)$ is equivalent to $\lim_{x\to\infty} \frac{y}{x} = 0$.

Chapter 2

Relay-Assisted Multiuser Precoding with Limited Feedback

In traditional multi-antenna broadcast channels, linear precoding schemes including zero-forcing (ZF) beamforming and block diagonalization (BD) precoding are known to exploit available degrees of freedom. They can perform measurably close to the dirty paper coding (DPC) technique which is capacity-achieving but of very high complexity [3], [4]. However, to achieve full multiplexing gain, accurate channel state information at the transmitter (CSIT) is a prerequisite [5], [6] while in practice full CSIT is hardly available especially for frequency division duplexing (FDD) systems. Limited feedback can be an effective solution to this challenge where each mobile feeds back a finite number of bits regarding its channel instantiation to the transmitter to provide partial channel state information (CSI). Intensive research has been conducted in [6] and [7] on the performance of the limited feedback scheme in multiple-input multiple-output (MIMO) broadcast channels.

In recent years, relay assistance has attracted an upsurge of attention for its capability to extend the radio range and enhance capacity in broadband cellular networks [8]. The relay stations (RS) receive signals from the base stations (BS) and then amplify and forward (AF), or decode and forward (DF) the signals to remote users located in the cell boundaries. The received signal-to-noise-plus-interference ratio (SINR) of cell boundary users can thus be effectively strengthened. Analysis on the MIMO relay channel has been made in [9] from a general information theoretic perspective, and [10] proposes an efficient relaying strategies with linear processing at the single fixed relay to support multiuser transmission in MIMO relay broadcast

channels. Notably a method based on singular value decomposition (SVD) of the BS-RS channel combined with ZF-DPC coding at the relay is developed in [10] which achieves good performance as compared to the AF relay capacity upper bound derived therein. The capacity performance of linear beamforming schemes in multi-relay MIMO channels is considered in [11] when the relay-to-destination CSI is not perfectly known at the relay nodes. Optimization of the linear processing operator at the relay is conducted in [12] with partial relay selection considered for a point-to-point MIMO relay scenario. Furthermore, robust linear beamforming design for both single and multiple relays is studied in [13], and the problem of joint source and relay design for multiuser MIMO nonregenerative relay networks is considered in [14], [15].

Since perfect CSIT is difficult to obtain in practice, application of the limited feedback strategy in two-hop multi-antenna relay channels has been extensively studied in [16]–[21]. Joint precoding design at the BS and RS has been investigated in [16] for both full CSIT and limited feedback MIMO relay systems. Limited feedback-based adaptive resource allocation and subcarrier pairing are considered in [17] for orthogonal frequency division multiplexing (OFDM) relay channels with one source node transmitting to one destination node with a single relay assistance. The problem of beamforming codebook design is then addressed in [18] and the paper further proposes a modified quantized scheme requiring less feedback bits than traditional strategies. Robust design against quantization degradation of linear beamforming for MIMO relay broadcast channels with limited feedback is considered in [19], [20]. Notably in [21], a suboptimal structured linear precoding scheme based on limited feedback is proposed with its capacity performance analyzed in details.

However, each user in [21] is equipped with a single antenna which critically restricts the system throughput. In this thesis, we generalize the single-antenna to multi-antenna scenario, which enables each user to accommodate multiple data streams simultaneously. The BD precoding scheme is applied in this case because each user can coordinate its own multiple streams and thus interference cancellation is only performed among different users, but not necessarily different antenna elements at each user. In this way, more degrees of freedom are directed to increase the signal strength. The major contributions of this chapter are two-fold.

1. Exploiting the linear precoding scheme proposed in [21], we analyze the system performance of multi-antenna AF relay broadcast channels with limited feedback CSI from both two-hop links. Notably, the random matrix quantization (RMQ) method is applied to enable downlink CSI quantization and feedback

facilitating SVD-based and BD precoding at the BS and RS, respectively. A closed-form upper bound is then derived to characterize the rate loss of the limited feedback-based precoding scheme relative to the full CSIT case.

2. According to the derived upper bound, this thesis proposes a feedback quality control strategy to maintain a bounded rate loss. Studies in [7] and [21] have developed effective feedback scaling strategies to bound the rate loss for BD-based MIMO broadcast channels and single-antenna receiver relay broadcast channels, respectively. We generalize both works to multi-antenna receiver AF relay broadcast channels and propose to scale the feedback size B_1 of the BS-RS link and the size B_2 of the RS-User link according to

$$B_1 = \frac{M-1}{3} P_2(\text{dB}) - (M-1)\log_2\left(M + \frac{N_t}{P_1}\right) + (M-1)\log_2 \frac{2T}{N_t\left(b^{\frac{1}{N_r}} - 1\right)},$$

$$B_2 = \frac{N_r(N_t - N_r)}{3} P_2(\text{dB}) + N_r(N_t - N_r)\log_2 \frac{2A}{N_t\left(b^{\frac{1}{N_r}} - 1\right)}$$

where P_1 and P_2 are power constraints at the BS and RS, respectively. M, N_t, and N_r are the corresponding numbers of antennas at the BS, RS, and each user. The parameter b concerns the predetermined rate loss gap. T and A follow from (2.26) and (2.27), respectively. The proposed scaling strategy gives insights into the inherent relationship between system transmit power and the feedback sizes, as well as provides guidelines for limited feedback design in practical relay-assisted multiuser MIMO systems.

The rest of this chapter is organized as follows. The system model is presented in Section 2.1. Section 2.2 provides some preliminaries useful for the remaining throughput analysis. Section 2.3 derives an upper bound for the rate loss of the focused system and then proposes a feedback quality control strategy to maintain a bounded throughput loss. Conclusions are finally drawn in Section 2.4.

2.1 System Model

A multi-antenna AF relay broadcast channel is illustrated in Fig. 2.1. In this scenario, the BS and RS have M and N_t antennas respectively while each user is equipped with N_r antennas capable of receiving multiple data streams simultaneously. It is assumed

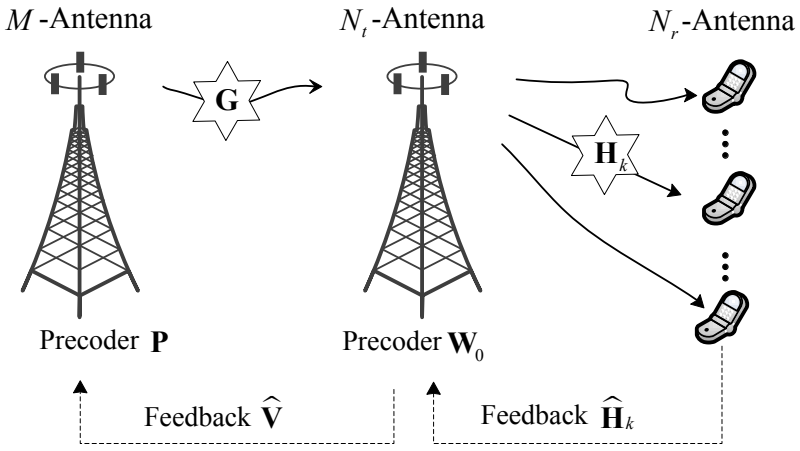

Figure 2.1: System model of the multi-antenna AF relay channel.

that the transmitted symbol vector $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ is normalized so that $\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right] = \mathbf{I}_{N_t}$. Note that we ignore the direct link between the BS and each remote user for very severe large-scale path loss. Then the signal received at user k is characterized by

$$\mathbf{y}_k = \sqrt{\rho_1\rho_2}\mathbf{H}_k^H\mathbf{W}_0\mathbf{G}\mathbf{P}\mathbf{x} + \sqrt{\rho_2}\mathbf{H}_k^H\mathbf{W}_0\mathbf{n} + \mathbf{z}_k \tag{2.1}$$

where $\mathbf{G} \in \mathbb{C}^{N_t \times M}$ is the channel matrix from the BS to RS while the channel between user k and the RS is represented by $\mathbf{H}_k^H \in \mathbb{C}^{N_r \times N_t}$. Both channels follow the Rayleigh distribution with their entries assumed to be independent and identically distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance. $\mathbf{P} \in \mathbb{C}^{M \times N_t}$ and $\mathbf{W}_0 \in \mathbb{C}^{N_t \times N_t}$ are precoding matrices at the BS and RS, respectively. $\mathbf{n}$ and $\mathbf{z}_k$ are additive Gaussian noise with zero mean and unit variance. We assume here the number of BS antennas is no less than that of the RS, i.e., $M \geq N_t$. It is also assumed that $K = \frac{N_t}{N_r}$ $(K \geq 2)$ so that no user selection is considered. Power scaling factors ρ_1 and ρ_2 are defined to meet the power constraints at the BS and RS, respectively and given by

$$\rho_1 = \frac{P_1}{\mathbb{E}\left[\mathrm{tr}(\mathbf{P}\mathbf{P}^H)\right]} \tag{2.2}$$

and

$$\rho_2 = \frac{P_2}{\mathbb{E}\left[\rho_1 \mathrm{tr}(\mathbf{W}_0\mathbf{G}\mathbf{P}\mathbf{P}^H\mathbf{G}^H\mathbf{W}_0^H) + \mathrm{tr}(\mathbf{W}_0\mathbf{W}_0^H)\right]} \tag{2.3}$$

where P_1 and P_2 are the corresponding power constraints at the BS and RS.

2.1.1 Linear Precoding with Perfect CSIT

Optimal joint design of the precoding matrices $\mathbf{P}$ and $\mathbf{W}_0$ to maximize the relay system capacity is yet to be found. In this thesis, we exploit the structured source and relay precoding scheme proposed in [21] and extend it to the multi-antenna user scenario by peforming BD for downlink precoding at the RS while the SVD-based precoding at the BS remains.

In fact, the sum capacity of the two-hop relay system is restricted by the minimum one of the BS-RS and RS-User links. The first link is essentially a point-to-point MIMO channel, where the SVD-based precoding is known as optimal. For the second link, it is a broadcast channel in essence, which motivates the capacity-achieving DPC design. Actually, the scheme of SVD-based precoding at the BS in conjunction with DPC at the RS is shown in [10] to achieve good performance as compared to the AF relay capacity upper bound, where ZF-DPC is used in place of traditional DPC. However, since DPC or ZF-DPC is highly complex to implement and hard to incorporate the limited feedback strategy which will be explored later, we prefer to employ such simple linear ZF methods as BD at the RS. The linear ZF schemes, while generally suboptimal, are known to achieve the same asymptotic sum capacity as that of DPC for MIMO broadcast channels [22].

According to the structured precoding scheme, we apply SVD to the channel $\mathbf{G}$ and get

$$\mathbf{G} = \mathbf{U}\left[\mathbf{\Sigma}\ \mathbf{0}\right]\left[\mathbf{V}\,\mathbf{V}_0\right]^H \tag{2.4}$$

where $\mathbf{V}$ comprises the first N_t right singular vectors and $\mathbf{V}_0$ holds the last $(M - N_t)$ right singular vectors. Then precoders $\mathbf{P}$ and $\mathbf{W}_0$ at the BS and RS are given, respectively, by

$$\mathbf{P} = \mathbf{V} \text{ and } \mathbf{W}_0 = \mathbf{W}\mathbf{U}^H \tag{2.5}$$

where $\mathbf{W} \in \mathbb{C}^{N_t \times N_t}$ is designed based on the BD criterion [4]. It suggests that the kth column block matrix $\mathbf{W}_k \in \mathbb{C}^{N_t \times N_r}$ satisfies $\mathbf{H}_j^H \mathbf{W}_k = \mathbf{0}$, for all $j \neq k$, i.e., $\mathbf{W}_k$ is chosen in the nullspace of the concatenation matrix $[\mathbf{H}_1, \cdots, \mathbf{H}_{k-1}, \mathbf{H}_{k+1}, \cdots, \mathbf{H}_K]^H$ with each column normalized to identity. While complete diagonalization, i.e., ZF

beamforming can meet this interference cancellation requirement, it is not a good choice in this case since each user is capable of coordinating the processing of its own signal outputs [4]. Hence, the block effective diagonal method is preferred to cancel co-channel interference (CCI) for different users instead of different antenna elements at each user.

After applying the linear precoding scheme at the BS and RS, multiuser interference can be fully suppressed with K decoupled data channels created between the BS and mobile users. From (2.1), it gives a per user spectral efficiency [4]

$$R_{CSIT} = \frac{1}{2}\mathbb{E}\left[\log_2 \frac{\left|\mathbf{I}+\rho_2 \mathbf{H}_k^H \mathbf{W}_k \mathbf{W}_k^H \mathbf{H}_k + \rho_1\rho_2 \mathbf{H}_k^H \mathbf{W}_k \mathbf{\Sigma}_k^2 \mathbf{W}_k^H \mathbf{H}_k\right|}{\left|\mathbf{I}+\rho_2 \mathbf{H}_k^H \mathbf{W}_k \mathbf{W}_k^H \mathbf{H}_k\right|}\right] \tag{2.6}$$

where the factor $\frac{1}{2}$ is applied because symbols are transmitted over two time slots. The matrix $\mathbf{\Sigma}_k \in \mathbb{C}^{N_r \times N_r}$ is the kth block in the diagonal matrix $\mathbf{\Sigma}$, i.e., $\mathbf{\Sigma} = \text{diag}\{\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_k, \cdots, \mathbf{\Sigma}_K\}$ where $\text{diag}\{\cdots\}$ returns a diagonal matrix with given elements on its diagonal positions. With the linear precoding scheme, we obtain the power scaling factors ρ_1 and ρ_2 from (2.2) and (2.3) as

$$\rho_1 = \frac{P_1}{N_t} \tag{2.7}$$

and

$$\begin{aligned} \rho_2 &= \frac{P_2}{\text{tr}\left(\mathbf{W}^H \mathbf{W} \mathbb{E}_{\mathbf{G}}\left[\rho_1 \mathbf{\Sigma}^2 + \mathbf{I}\right]\right)} \\ &= \frac{P_2}{\left(\frac{P_1 M}{N_t} + 1\right)\text{tr}\left(\mathbf{W}^H \mathbf{W}\right)} \qquad (2.8) \\ &= \frac{P_2}{P_1 M + N_t} \qquad (2.9) \end{aligned}$$

where(2.8) utilizes $\mathbb{E}_{\mathbf{G}}\left[\mathbf{\Sigma}^2\right] = M\mathbf{I}_{N_t}$ because $\mathbf{\Sigma}^2$ comprises the unordered eigenvalues of a Wishart matrix $\mathbf{G}\mathbf{G}^H$[21] and the last equality holds as each column of $\mathbf{W}$ is normalized.

2.1.2 Linear Precoding with Quantized CSI Feedback

To fully cancel the multiuser interference, the structured source and relay precoding scheme requires the transmitters to have perfect downlink channel information, i.e.,

full knowledge of the BS-RS channel $\mathbf{G}$, or equivalently $\mathbf{V}$ in this case, at the BS and full knowledge of the RS-User channel $\mathbf{H}_k$ at the RS. In practice, downlink CSI can be obtained through channel estimation at the receivers, which is then fed back to the transmitters through uplink channels. In reality, due to the limited capacity of uplink channels, only a finite number of bits regarding the channel knowledge can be sent, which motivates us to exploit the limited feedback strategy.

With the limited feedback scheme, the BS-RS channel $\mathbf{G}$ is accurately estimated at the RS and each column of the orthogonal matrix $\mathbf{V}$ resulting from the SVD of $\mathbf{G}$ is quantized into B_1 bits according to a prescribed codebook $\mathcal{C}_1$ at the RS side. The finite bits are then fed back to the BS through the uplink channel. Similarly, the RS-User channel $\mathbf{H}_k$ is estimated and quantized into B_2 bits according to another predetermined codebook $\mathcal{C}_2$ at the user k, and then sent back to the RS through the uplink channel.

To quantize the matrix $\mathbf{V}$, the RS employs random vector quantization (RVQ) [6] to quantize each column vector $\mathbf{v}_j \in \mathbb{C}^{M\times 1}, (j = 1, 2, \ldots, N_t)$ of $\mathbf{V}$ according to a codebook $\mathcal{C}_1$[1] and then concatenate the quantized vectors to form the quantization matrix $\hat{\mathbf{V}}$. The codebook $\mathcal{C}_1$, known to both the BS and RS, consists of 2^{B_1} unit-norm vectors $\{\mathbf{f}_1, \ldots, \mathbf{f}_{2^{B_1}}\}$ in $\mathbb{C}^{M\times 1}$. Each element of $\mathcal{C}_1$ is isotropically and independently drawn from $\mathfrak{g}_{M,1}$, where $\mathfrak{g}_{M,N}$ is the Grassmann manifold and defined as the set of all N dimensional subspaces passing through the origin of an M dimensional space [23]. The quantization of the column vector $\mathbf{v}_j$, say $\hat{\mathbf{v}}_j$, is selected from $\mathcal{C}_1$ according to

$$\hat{\mathbf{v}}_j = \arg\max_{\mathbf{f}\in\mathcal{C}_1} |\mathbf{v}_j^H \mathbf{f}|^2. \tag{2.10}$$

Then the precoding matrix at the BS is given by

$$\mathbf{P} = \hat{\mathbf{V}}. \tag{2.11}$$

The user k employs RMQ [7], [24] to quantize the channel $\mathbf{H}_k$ according to a codebook $\mathcal{C}_2$, which is known to the RS as well. The codebook $\mathcal{C}_2$ consists of 2^{B_2} matrices $\{\mathbf{M}_1, \ldots, \mathbf{M}_{2^{B_2}}\}$ in $\mathbb{C}^{N_t\times N_r}$ and they are uniformly and independently distributed over $\mathfrak{g}_{N_t,N_r}$. The quantization of $\mathbf{H}_k$, i.e. $\hat{\mathbf{H}}_k$, is chosen from $\mathcal{C}_2$ according

[1]It is intuitively natural to perform RMQ on the matrix $\mathbf{V}$, but it is not suitable in this case. Because the minimum chordal distance defined in [7, Eq. (3)] only requires the spatial direction of column spaces spanned by the two matrices to be close enough and it is invariant under unitary rotation. In this sense, the RMQ version $\hat{\mathbf{V}}_M$ cannot reach the goal of diagonalizing $\mathbf{V}^H\hat{\mathbf{V}}_M = \mathbf{I}$ even when the feedback size is infinitely large. For this reason the RMQ method is not used here.

to

$$\hat{\mathbf{H}}_k = \arg \min_{\mathbf{M} \in \mathcal{C}_2} d^2(\mathbf{H}_k, \mathbf{M}) \tag{2.12}$$

where $d(\mathbf{H}_k, \mathbf{M})$ is the distance metric. Here we use the chordal distance defined as [23] [24]

$$d(\mathbf{H}_k, \mathbf{M}) = \sqrt{\sum_{i=1}^{N_r} \sin^2 \theta_i} \tag{2.13}$$

where θ_i's are the principal angles between the two subspaces spanned by $\mathbf{H}_k$ and $\mathbf{M}$.

Then the precoder at the RS is given by

$$\mathbf{W}_0 = \hat{\mathbf{W}} \mathbf{U}^H \tag{2.14}$$

where we assume the RS can perfectly estimate the channel $\mathbf{G}$ and thus $\mathbf{U}$ is known by the RS itself even in this limited feedback case. The matrix $\hat{\mathbf{W}}$ is the concatenation of $\hat{\mathbf{W}}_k$ which results from performing BD precoding on the channel quantization $\hat{\mathbf{H}}_k$.

With the limited feedback-based precoding at both the BS and RS, we can obtain a per user rate as shown in (2.15). Comparison of the full CSIT rate in (2.6) and limited feedback-based rate in (2.15) shows that CSI quantization leads to residual interuser interference and substantially degrades the system capacity performance as a result. Based on this observation, we are well motivated to analyze the effects of channel quantization and come up with strategies to control the rate loss for CSI limited feedback systems.

$$R_{QUANT} = \frac{1}{2}\mathbb{E}\left[\log_2 \frac{\left|\mathbf{I} + \rho_2 \mathbf{H}_k^H \hat{\mathbf{W}} \hat{\mathbf{W}}^H \mathbf{H}_k + \rho_1 \rho_2 \mathbf{H}_k^H \hat{\mathbf{W}} \mathbf{\Sigma} \mathbf{V}^H \hat{\mathbf{V}} \hat{\mathbf{V}}^H \mathbf{V} \mathbf{\Sigma}^H \hat{\mathbf{W}}^H \mathbf{H}_k\right|}{\left|\mathbf{I} + \rho_2 \mathbf{H}_k^H \hat{\mathbf{W}} \hat{\mathbf{W}}^H \mathbf{H}_k + \rho_1 \rho_2 \sum_{j \neq k} \mathbf{H}_k^H \hat{\mathbf{W}} \mathbf{\Sigma} \mathbf{V}^H \hat{\mathbf{V}}_j \hat{\mathbf{V}}_j^H \mathbf{V} \mathbf{\Sigma}^H \hat{\mathbf{W}}^H \mathbf{H}_k\right|}\right] \tag{2.15}$$

2.2 Background and Preliminary Calculations

In this section, we will explore some background knowledge and preliminary findings which are useful for the throughput analysis coming up in the next few sections of

this chapter.

2.2.1 Random Vector Quantization

According to [6], the column vector $\mathbf{v}_j$ of the orthogonal matrix $\mathbf{V}$ can be decomposed as a weighted sum of its quantization $\hat{\mathbf{v}}_j$ and another vector $\mathbf{s}$ isotropically distributed in the nullspace of $\hat{\mathbf{v}}_j$:

$$\mathbf{v}_j = \sqrt{1-Z}\hat{\mathbf{v}}_j + \sqrt{Z}\mathbf{s} \tag{2.16}$$

where Z represents the quantization error, which is the minimum of 2^{B_1} Beta$(M-1,1)$ random variables [6]. We define the expectation of error Z averaged over both the random codebooks and Rayleigh fading distribution as $\mathbb{E}[Z] \triangleq \epsilon$. It follows [23, Eq. (13)]

$$\epsilon \approx \frac{M-1}{M} 2^{-\frac{B_1}{M-1}}. \tag{2.17}$$

With the above results, we come to the following lemma motivated by [21, Lemma 2]. But first, we define $\hat{\mathbf{V}}_k$ as the quantization of the kth column block matrix, i.e., $\hat{\mathbf{V}} = \left[\hat{\mathbf{V}}_1, \cdots, \hat{\mathbf{V}}_k, \cdots, \hat{\mathbf{V}}_K\right]$ where $\hat{\mathbf{V}}_k \in \mathbb{C}^{M\times N_r}$.

Lemma 1. *The quantization $\hat{\mathbf{V}}_j$ and the matrix $\mathbf{V}$ follow the equation*

$$\begin{aligned}\sum_{j=1,j\neq k}^{K} \mathbb{E}\left[\mathbf{V}^H \hat{\mathbf{V}}_j \hat{\mathbf{V}}_j^H \mathbf{V}\right] = \operatorname{diag}\Bigg\{ &\left(1 - \frac{M-N_t+N_r}{M-1}\epsilon\right)\mathbf{I}_{N_r}, \\ \cdots, \underbrace{\left(\frac{N_t-N_r}{M-1}\epsilon\right)\mathbf{I}_{N_r}}_{k\text{th diagonal block}}, &\left(1 - \frac{M-N_t+N_r}{M-1}\epsilon\right)\mathbf{I}_{N_r}, \cdots \Bigg\}.\end{aligned} \tag{2.18}$$

Proof. For the kth column block matrix $\hat{\mathbf{V}}_k$ of the quantization matrix $\hat{\mathbf{V}}$, we have

$$\mathbb{E}\left[\mathbf{V}^H \hat{\mathbf{V}}_k \hat{\mathbf{V}}_k^H \mathbf{V}\right] = \sum_{m=1}^{N_r} \mathbb{E}\left[\mathbf{V}^H \hat{\mathbf{v}}_{k,m} \hat{\mathbf{v}}_{k,m}^H \mathbf{V}\right] \tag{2.19}$$

where $\hat{\mathbf{v}}_{k,m}$ stands for the mth column of the matrix $\hat{\mathbf{V}}_k$, i.e., the $((k-1)N_r+m)$th column of $\hat{\mathbf{V}}$.

Meanwhile, according to [21, Lemma 2], for the vector $\hat{\mathbf{v}}_j$, which is the jth column

of the matrix $\hat{\mathbf{V}}$, we have the following equality

$$\mathbb{E}\left[\mathbf{V}^H\hat{\mathbf{v}}_j\hat{\mathbf{v}}_j^H\mathbf{V}\right] = \mathrm{diag}\left\{\frac{1}{M-1}\epsilon,\ldots,\underbrace{(1-\epsilon)}_{j\text{th element}},\frac{1}{M-1}\epsilon,\ldots\right\}. \tag{2.20}$$

Henceforth, substituting (2.20) into (2.19), we further obtain

$$\begin{aligned}\mathbb{E}\left[\mathbf{V}^H\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^H\mathbf{V}\right] = \mathrm{diag}\Bigg\{&\frac{N_r\epsilon}{M-1}\mathbf{I}_{N_r},\ldots,\\ &\underbrace{(1-\frac{M-N_r}{M-1}\epsilon)}_{k\text{th block element}}\mathbf{I}_{N_r},\ldots,\frac{N_r\epsilon}{M-1}\mathbf{I}_{N_r}\Bigg\}.\end{aligned} \tag{2.21}$$

Then we sum up the $K-1$ elements and obtain the final result

$$\begin{aligned}\sum_{j=1,j\neq k}^{K}\mathbb{E}\left[\mathbf{V}^H\hat{\mathbf{V}}_j\hat{\mathbf{V}}_j^H\mathbf{V}\right] = \mathrm{diag}\Bigg\{&\left(1-\frac{M-N_t+N_r}{M-1}\epsilon\right)\mathbf{I}_{N_r},\\ \cdots,&\underbrace{\left(\frac{N_t-N_r}{M-1}\epsilon\right)\mathbf{I}_{N_r}}_{k\text{th diagonal block}},\left(1-\frac{M-N_t+N_r}{M-1}\epsilon\right)\mathbf{I}_{N_r},\cdots\Bigg\}.\end{aligned} \tag{2.22}$$

□

2.2.2 Random Matrix Quantization

According to [7], the orthonormal basis $\tilde{\mathbf{H}}_k$ of the channel instantiation $\mathbf{H}_k$ can be decomposed as a weighted sum of its quantization $\hat{\mathbf{H}}_k \in \mathbb{C}^{N_t\times N_r}$, and a matrix $\mathbf{S}_k \in \mathbb{C}^{N_t\times N_r}$ in the left nullspace of $\hat{\mathbf{H}}_k$, i.e.,

Lemma 2. *[7] The orthonormal basis $\tilde{\mathbf{H}}_k$ and its quantization $\hat{\mathbf{H}}_k$, obtained according to* (2.12)*, are related in the equation*

$$\tilde{\mathbf{H}}_k = \hat{\mathbf{H}}_k\mathbf{X}_k\mathbf{Y}_k + \mathbf{S}_k\mathbf{Z}_k \tag{2.23}$$

where $\mathbf{X}_k \in \mathbb{C}^{N_r\times N_r}$ is a unitary matrix isotropically distributed over $\mathfrak{g}_{N_r,N_r}$. $\mathbf{Z}_k \in \mathbb{C}^{N_r\times N_r}$ and $\mathbf{Y}_k \in \mathbb{C}^{N_r\times N_r}$ are both upper triangular matrices with positive diagonal elements, satisfying $\mathbf{Y}_k^H\mathbf{Y}_k = \mathbf{I}_{N_r} - \mathbf{Z}_k^H\mathbf{Z}_k$. Similar to Z in (2.16)*, the matrix $\mathbf{Z}_k$ here concerns the quantization error in $d^2(\mathbf{H}_k,\hat{\mathbf{H}}_k) = \mathrm{tr}(\mathbf{Z}_k^H\mathbf{Z}_k)$.*

The quantization error averaged over both the random codebooks and channel fading distribution associated with the channel $\mathbf{H}_k$ is given by

$$D \triangleq \mathbb{E}\left[d^2(\mathbf{H}_k, \hat{\mathbf{H}}_k)\right] \leq \overline{D} \tag{2.24}$$

where $\overline{D}$ is given by [23, Theorem 4] and in practice can be closely approximated by [7, Eq. (16)]

$$\overline{D} \approx \frac{AN_r}{N_t} 2^{-\frac{B_2}{T}} \tag{2.25}$$

where

$$T = N_r(N_t - N_r) \tag{2.26}$$

and

$$A = \frac{N_t}{N_r} \frac{\Gamma\left(\frac{1}{T}\right)}{T} \left(\frac{1}{T!} \prod_{i=1}^{N_r} \frac{(N_t - i)!}{(N_r - i)!}\right)^{-\frac{1}{T}}. \tag{2.27}$$

2.2.3 A Useful Matrix Inequality

In this subsection, we give a useful matrix inequality for the throughput analysis afterwards.

Lemma 3. *For any positive semi-definite matrices* $\mathbf{A}$, $\mathbf{B}$ *and positive definite matrix* $\mathbf{A}_0$, *the following inequality holds*

$$\log_2 \frac{|\mathbf{A}_0 + \mathbf{A} + \mathbf{B}|}{|\mathbf{A}_0 + \mathbf{A}|} \leq \log_2 \frac{|\mathbf{A}_0 + \mathbf{B}|}{|\mathbf{A}_0|}. \tag{2.28}$$

Proof. As $\mathbf{A}_0$ is positive definite and $\mathbf{A}$ is positive semidefinite, we have $(\mathbf{A} + \mathbf{A}_0) \succeq \mathbf{A}_0$, where $\mathbf{A}_1 \succeq \mathbf{A}_2$ means that $\mathbf{A}_1 - \mathbf{A}_2$ is positive semidefinite provided that both $\mathbf{A}_1$ and $\mathbf{A}_2$ are Hermitian matrices. Then according to [26, Corollay 7.7.4], we have

$$(\mathbf{A}_0)^{-1} \succeq (\mathbf{A} + \mathbf{A}_0)^{-1} \tag{2.29}$$

and [26, Oservation 7.7.2] gives

$$\left(\mathbf{B}^{1/2}\right)^H \mathbf{A}_0^{-1}\mathbf{B}^{1/2} \succeq \left(\mathbf{B}^{1/2}\right)^H (\mathbf{A}+\mathbf{A}_0)^{-1}\mathbf{B}^{1/2} \tag{2.30}$$

where $\mathbf{B}^{1/2}$ is the unique square root of the positive semidefinite matrix $\mathbf{B}$. Considering the following determinants, we have

$$\begin{aligned} \left|\mathbf{I}+\mathbf{A}_0^{-1}\mathbf{B}\right| &\overset{(a)}{=} \left|\mathbf{I}+\mathbf{B}^{1/2}\mathbf{A}_0^{-1}\mathbf{B}^{1/2}\right| \\ &\overset{(b)}{\geq} \left|\mathbf{I}+\mathbf{B}^{1/2}(\mathbf{A}+\mathbf{A}_0)^{-1}\mathbf{B}^{1/2}\right| \\ &= \left|\mathbf{I}+(\mathbf{A}+\mathbf{A}_0)^{-1}\mathbf{B}\right| \end{aligned} \tag{2.31}$$

where (a) uses $|\mathbf{I}+\mathbf{A}_1\mathbf{A}_2| = |\mathbf{I}+\mathbf{A}_2\mathbf{A}_1|$ for any matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ satisfying multiplying conditions. Step (b) follows from (2.30) and the fact that $\mathbf{B}^{1/2}$ is Hermitian [26, Theorem 7.2.6]. The proved inequality $\left|\mathbf{I}+\mathbf{A}_0^{-1}\mathbf{B}\right| \geq |\mathbf{I}+(\mathbf{A}+\mathbf{A}_0)^{-1}\mathbf{B}|$ directly leads to

$$\log_2 \frac{|\mathbf{A}_0+\mathbf{A}+\mathbf{B}|}{|\mathbf{A}_0+\mathbf{A}|} \leq \log_2 \frac{|\mathbf{A}_0+\mathbf{B}|}{|\mathbf{A}_0|} \tag{2.32}$$

due to the monotonically increasing property of the function $\log_2(\cdot)$ and the fact that $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ provided that $\mathbf{A}$ is invertable. □

2.3 Throughput Analysis

Due to the channel quantization, the system suffers a substantial rate loss relative to the perfect CSIT case [21]. In this section, we first derive an upper bound of the rate loss and then propose a strategy of scaling feedback bits B_1 and B_2 with increasing power at the BS and RS to maintain a bounded loss according to the derived closed-form upper bound expressions.

2.3.1 The Rate Loss Upper Bound

With the per user rates of perfect CSIT and quantized CSI cases given in (2.6) and (2.15), respectively, we upper bound the rate loss due to CSI quantization at high signal-to-noise ratio (SNR) regime in the following theorem.

Theorem 1. *At high SNR regime, the rate loss per user relative to the perfect CSIT case incurred by limited feedback in the MIMO relay downlink is upper bounded by*

$$\Delta R = R_{CSIT} - R_{QUANT} \tag{2.33}$$

$$\leq \frac{N_r}{2}\log_2\left(1+T\rho_1\rho_2 2^{-\frac{B_1}{M-1}}+A\rho_2(1+\rho_1 M)2^{-\frac{B_2}{T}}\right)+o(1) \tag{2.34}$$

where T and A follow from (2.26) *and* (2.27)*, respectively.*

Proof. By substituting the full CSIT rate (2.6) and limited feedback-based rate (2.15) in (2.33), we get the rate loss expression (2.35) after some basic manipulations. In the subsequent steps, we will analyze the two summation terms ΔR_1 and ΔR_2 in (2.35) separately and show that ΔR_1 dominates the rate loss at high SNR regime while ΔR_2 is essentially negligible in the meanwhile. Then Theorem 1 follows directly by adding the effects of ΔR_1 and ΔR_2.

$$\Delta R = \underbrace{\frac{1}{2}\mathbb{E}\left[\log_2 \frac{\left|\mathbf{I}+\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\hat{\mathbf{W}}^H\mathbf{H}_k+\rho_1\rho_2\sum_{j\neq k}\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}_j\hat{\mathbf{V}}_j^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right|}{|\mathbf{I}+\rho_2\mathbf{H}_k^H\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_k|}\right]}_{\Delta R_1}$$

$$+\underbrace{\frac{1}{2}\mathbb{E}\left[\log_2 \frac{\left|\mathbf{I}+\rho_2\mathbf{H}_k^H\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_k+\rho_1\rho_2\mathbf{H}_k^H\mathbf{W}_k\boldsymbol{\Sigma}_k^2\mathbf{W}_k^H\mathbf{H}_k\right|}{\left|\mathbf{I}+\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\hat{\mathbf{W}}^H\mathbf{H}_k+\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}\hat{\mathbf{V}}^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right|}\right]}_{\Delta R_2} \tag{2.35}$$

1. **Analysis of ΔR_1 in** (2.35)

$$\begin{aligned}
&2\Delta R_1\\
&\overset{(a)}{\leq}\mathbb{E}\left[\log_2\left|\mathbf{I}+\rho_2\sum_{j\neq k}\mathbf{H}_k^H\hat{\mathbf{W}}_j\hat{\mathbf{W}}_j^H\mathbf{H}_k+\rho_1\rho_2\sum_{j\neq k}\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}_j\hat{\mathbf{V}}_j^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right|\right]\\
&\overset{(b)}{=}\mathbb{E}\left[\log_2\left|\mathbf{I}+\left(\rho_2\sum_{j\neq k}\tilde{\mathbf{H}}_k^H\hat{\mathbf{W}}_j\hat{\mathbf{W}}_j^H\tilde{\mathbf{H}}_k\right.\right.\right.\\
&\qquad\left.\left.\left.+\rho_1\rho_2\sum_{j\neq k}\tilde{\mathbf{H}}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}_j\hat{\mathbf{V}}_j^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\tilde{\mathbf{H}}_k\right)\boldsymbol{\Lambda}_k\right|\right]
\end{aligned}$$

$$\overset{(c)}{\leq} \log_2 \left| \mathbf{I} + \rho_2 N_t \sum_{j \neq k} \mathbb{E}_{\mathbf{H}} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}}_j \hat{\mathbf{W}}_j^H \tilde{\mathbf{H}}_k \right] \right.$$
$$\left. + \rho_1 \rho_2 N_t \mathbb{E}_{\mathbf{H},\mathbf{G}} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}} \mathbf{\Sigma} \sum_{j \neq k} \mathbf{V}^H \hat{\mathbf{V}}_j \hat{\mathbf{V}}_j^H \mathbf{V} \mathbf{\Sigma}^H \hat{\mathbf{W}}^H \tilde{\mathbf{H}}_k \right] \right| \tag{2.36}$$

where (a) results from Lemma 3 and the fact that $\mathbf{H}_k^H \mathbf{W}_k \mathbf{W}_k^H \mathbf{H}_k$ and $\mathbf{H}_k^H \hat{\mathbf{W}}_k \hat{\mathbf{W}}_k^H \mathbf{H}_k$ follow the same distribution as $\mathbf{W}_k$ and $\hat{\mathbf{W}}_k$ are isotropically distributed and both are independent of $\mathbf{H}_k$. Step (b) follows from $\mathbf{H}_k \mathbf{H}_k^H = \tilde{\mathbf{H}}_k \mathbf{\Lambda}_k \tilde{\mathbf{H}}_k^H$ where $\tilde{\mathbf{H}}_k$ forms an orthonormal basis for the column space of $\mathbf{H}_k$ and $\mathbf{\Lambda}_k$ is a diagonal matrix with N_r unordered eigenvalues of $\mathbf{H}_k \mathbf{H}_k^H$ on its diagonal positions. The last step (c) follows from Jensen's inequality for the concavity of $\log_2 |\cdot|$ and $\mathbb{E}[\mathbf{\Lambda}_k] = N_t \mathbf{I}_{N_r}$ [7].

According to [7, Eq. (45)], we have

$$\mathbb{E}_{\mathbf{H}} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}}_j \hat{\mathbf{W}}_j^H \tilde{\mathbf{H}}_k \right] = \frac{D}{N_t - N_r} \mathbf{I} \tag{2.37}$$

where D follows from (2.24). This leads us to the result

$$\sum_{j \neq k} \mathbb{E}_{\mathbf{H}} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}}_j \hat{\mathbf{W}}_j^H \tilde{\mathbf{H}}_k \right] = \frac{D}{N_r} \mathbf{I}. \tag{2.38}$$

Furthermore, D can be tightly upper bounded by $\overline{D}$ from (2.24).

Then we consider the second term in (2.36) and have

$$\mathbb{E}_{\mathbf{H},\mathbf{G}} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}} \mathbf{\Sigma} \sum_{j \neq k} \mathbf{V}^H \hat{\mathbf{V}}_j \hat{\mathbf{V}}_j^H \mathbf{V} \mathbf{\Sigma}^H \hat{\mathbf{W}}^H \tilde{\mathbf{H}}_k \right]$$
$$\overset{(d)}{=} \mathbb{E}_{\mathbf{H}} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}} \mathbb{E}_{\mathbf{\Sigma}} \left[\mathbf{\Sigma}^2 \right] \sum_{j \neq k} \mathbb{E}_{\mathbf{G}} \left[\mathbf{V}^H \hat{\mathbf{V}}_j \hat{\mathbf{V}}_j^H \mathbf{V} \right] \hat{\mathbf{W}}^H \tilde{\mathbf{H}}_k \right]$$
$$\overset{(e)}{=} \frac{M(N_t - N_r)\epsilon}{M-1} \mathbb{E} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}}_k \hat{\mathbf{W}}_k^H \tilde{\mathbf{H}}_k \right] + M \left(1 - \frac{M - N_t + N_r}{M-1} \epsilon \right) \sum_{j \neq k} \mathbb{E} \left[\tilde{\mathbf{H}}_k^H \hat{\mathbf{W}}_j \hat{\mathbf{W}}_j^H \tilde{\mathbf{H}}_k \right]$$
$$\overset{(f)}{=} \left(\frac{MT\epsilon}{N_t(M-1)} + \frac{DM}{N_r} - \frac{M(M - N_t + N_r)D\epsilon}{N_r(M-1)} \right) \mathbf{I} \tag{2.39}$$

where step (d) follows from the fact that the channel $\mathbf{G}$ and $\mathbf{H}$ are mutually independent and $\sum_{j=1, j \neq k}^{K} \mathbb{E} \left[\mathbf{V}^H \hat{\mathbf{V}}_j \hat{\mathbf{V}}_j^H \mathbf{V} \right]$ is diagonal. Step (e) holds by substituting (2.18) and using equality $\mathbb{E}[\mathbf{\Sigma}^2] = M \mathbf{I}_{N_t}$ because the entries of $\mathbf{\Sigma}^2$ are unordered eigenvalues of

the Wishart matrix $\mathbf{G}\mathbf{G}^H$ [21, Eq. (16)]. Note that $\tilde{\mathbf{H}}_k^H\hat{\mathbf{W}}_k\hat{\mathbf{W}}_k^H\tilde{\mathbf{H}}_k$ is matrix-variate Beta(N_r, N_t-N_r) distributed and $\mathbb{E}\left[\tilde{\mathbf{H}}_k^H\hat{\mathbf{W}}_k\hat{\mathbf{W}}_k^H\tilde{\mathbf{H}}_k\right] = \frac{N_r}{N_t}\mathbf{I}_{N_r}$ [27] as $\tilde{\mathbf{H}}_k$ and $\hat{\mathbf{W}}_k$ are isotropically and independently distributed. This fact validates step (f) with (2.38) substituted here.

After substituting (2.38) and (2.39) in (2.36), we come to

$$\Delta R_1 \leq \frac{N_r}{2}\log_2\left(1+T\rho_1\rho_2 2^{-\frac{B_1}{M-1}} + A\rho_2(1+\rho_1 M)2^{-\frac{B_2}{T}}\right) \tag{2.40}$$

by neglecting the positive definite term $\frac{M(M-N_t+N_r)D\epsilon}{N_r(M-1)}\mathbf{I}_{N_r}$ of (2.39) when performing the determinant calculation in (2.36).

2. **Analysis of ΔR_2 in** (2.35)

$$\begin{aligned}
&2\Delta R_2 \\
&\overset{(a)}{=}\mathbb{E}\left[\log_2\frac{\left|\mathbf{I}+\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}_k\hat{\mathbf{W}}_k^H\mathbf{H}_k+\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\mathbf{V}_k\mathbf{V}_k^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right|}{\left|\mathbf{I}+\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\hat{\mathbf{W}}^H\mathbf{H}_k+\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}\hat{\mathbf{V}}^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right|}\right] \\
&=\mathbb{E}\left[\log_2\left|\mathbf{I}+\left(\mathbf{I}+\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\hat{\mathbf{W}}^H\mathbf{H}_k+\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}\hat{\mathbf{V}}^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right)^{-1}\right.\right. \\
&\quad\times\left(\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\left(\mathbf{V}_k\mathbf{V}_k^H-\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^H\right)\mathbf{V}^H\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right. \\
&\quad\left.\left.\left.-\rho_2\sum_{j\neq k}\mathbf{H}_k^H\hat{\mathbf{W}}_j\hat{\mathbf{W}}_j^H\mathbf{H}_k-\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\sum_{j\neq k}\left(\mathbf{V}^H\hat{\mathbf{V}}_j\hat{\mathbf{V}}_j^H\mathbf{V}\right)\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right)\right|\right] \qquad (2.41) \\
&\overset{(b)}{\leq}\mathbb{E}\left[\log_2\left|\mathbf{I}+\left(\mathbf{I}+\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\hat{\mathbf{W}}^H\mathbf{H}_k+\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\hat{\mathbf{V}}\hat{\mathbf{V}}^H\mathbf{V}\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right)^{-1}\right.\right. \\
&\quad\left.\left.\times\left(\rho_1\rho_2\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\mathbf{V}^H\left(\mathbf{V}_k\mathbf{V}_k^H-\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^H\right)\mathbf{V}^H\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right)\right|\right] \\
&\overset{(c)}{\leq}\mathbb{E}\left[\log_2\left|\mathbf{I}+\left(\mathbf{I}+\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}\left(\mathbf{V}^H\hat{\mathbf{V}}\hat{\mathbf{V}}^H\mathbf{V}\right)\boldsymbol{\Sigma}^H\hat{\mathbf{W}}^H\mathbf{H}_k\right)^{-1}\left(\sqrt{Z_k}\mathbf{H}_k^H\hat{\mathbf{W}}\boldsymbol{\Sigma}^2\hat{\mathbf{W}}^H\mathbf{H}_k\right)\right|\right] \qquad (2.42)
\end{aligned}$$

where (a) holds because $\mathbf{H}_k^H\hat{\mathbf{W}}_k\hat{\mathbf{W}}_k^H\mathbf{H}_k^H$ and $\mathbf{H}_k^H\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_k$ follow the same distribution and both are independent of $\boldsymbol{\Sigma}_k$. Step (b) is obtained by neglecting the last two positive semi-definite interference terms in (2.41) while (c) follows both from Lemma 3 and the conclusion $\sqrt{Z_k}\mathbf{I} \succeq \left(\mathbf{V}_k\mathbf{V}_k^H-\hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^H\right)$ drawn in [21, Lemma 3] with

$\sqrt{Z_k} = \sum_{j=1}^{N_r} \sqrt{Z_{k,j}}$ where we analyze each column of $\mathbf{V}_k$ and $\hat{\mathbf{V}}_k$ separately with Z defined in (2.16).

It is easy to observe from (2.42) that the sum component ΔR_2 depends solely on the CSI quantization or, alternatively, the feedback sizes B_1 and B_2. In contrast, the component ΔR_1 from (2.40) increases as transmit power P_1 and P_2 grow larger. We may thus conclude that ΔR_1 dominates the rate loss at high SNR regime. Hence, it is justified to neglect the effect of ΔR_2 in this scenario and we write it as $o(1)$. □

We observe from Theorem 1 that the incurred rate loss is a monotonically decreasing function with respect to the feedback sizes B_1 and B_2 while it increases as the power-relevant factors ρ_1 and ρ_2 grow. This is intuitively satisfying as the rate loss shall approach zero if the feedback sizes increase to infinity. Meanwhile with fixed feedback quality, the multiuser interference deteriorates system performance even more severely with growing transmit power. In this case, the channel eventually evolves to be interference-limited [6]. Fig. 2.2 verifies the effectiveness of the derived bound for a system with $M = 8, N_t = 4$, and $N_r = 2$ in which the term $o(1)$ is neglected. Concerning the accuracy of the derived upper bound, it is shown in the figure that the derived bound consistently stays above the actual rate loss curve and gets tighter as the feedback sizes grow larger. Thereby, this closed-form upper bound provides us with a tractable tool to assist in practical system performance analysis.

Remark 1. *In practice, perfect or near-perfect CSI of the slowly varying BS-RS link is likely to be known at the BS. This is because the BS and RS are usually mounted high and stationary for most practical applications. In this case, the derived rate loss upper bound of* (2.34) *simplifies to*

$$\Delta R \leq \frac{N_r}{2} \log_2 \left(1 + A\rho_2(1 + \rho_1 M)2^{-\frac{B_2}{T}}\right) + o(1) \tag{2.43}$$

by setting B_1 *in* (2.34) *to infinity and thus cancelling the second summation term in the parenthesis.*

2.3.2 Feedback Quality Control

The aforementioned relationship between the rate loss and feedback quality as well as the transmit power constraints motivates us to increase feedback bits B_1 and B_2 when power constraints P_1 and P_2 grow so as to maintain a bounded rate loss. In this

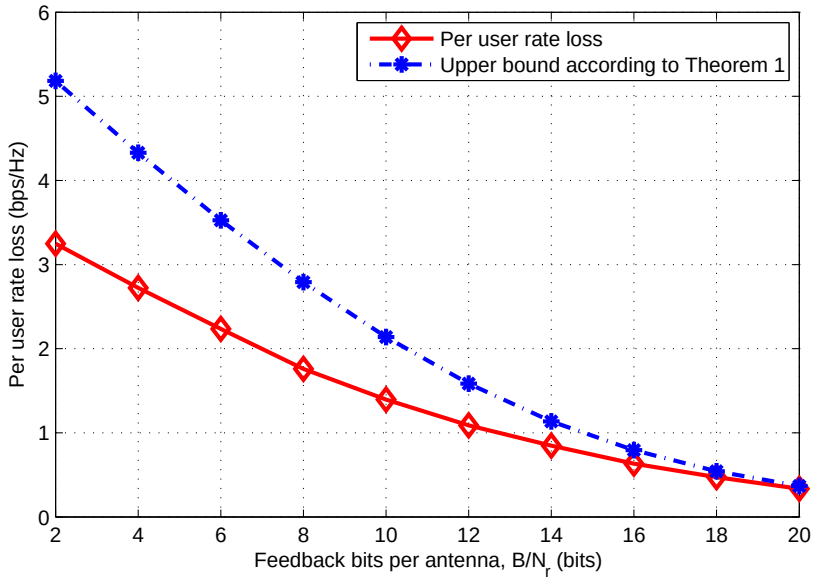

Figure 2.2: Accuracy of the derived rate loss upper bound for $M = 8, N_t = 4, N_r = 2$, and $K = \frac{N_t}{N_r} = 2$ with $P_1 = P_2 = 20$dB.

circumstance, the system is able to obtain the multiplexing gain that is an inherent property of multi-antenna schemes despite some constrained throughput loss relative to the perfect channel knowledge scenario.

Theorem 2. *In order to bound the rate loss per user from above within* $\frac{1}{2}\log_2 b$ *at high SNR regime, it suffices to scale the feedback bits* B_1 *and* B_2 *according to*

$$B_1 = \frac{M-1}{3} P_2(\text{dB}) - (M-1)\log_2\left(M + \frac{N_t}{P_1}\right) + (M-1)\log_2 \frac{2T}{N_t\left(b^{\frac{1}{N_r}} - 1\right)} \tag{2.44}$$

and

$$B_2 = \frac{N_r(N_t - N_r)}{3} P_2(\text{dB}) + N_r(N_t - N_r)\log_2 \frac{2A}{N_t\left(b^{\frac{1}{N_r}} - 1\right)} \tag{2.45}$$

where T *and* A *follow from* (2.26) *and* (2.27)*, respectively. The notation dB means the values are expressed in decibels.*

Proof. Equating the right side of (2.34) with $\frac{1}{2}\log_2 b$ while neglecting the term $o(1)$

leads to

$$2^{-\frac{B_1}{M-1}} + \frac{A(1+\rho_1 M)}{T\rho_1} 2^{-\frac{B_2}{T}} = \frac{b^{\frac{1}{N_r}} - 1}{T\rho_1\rho_2}. \tag{2.46}$$

We take a simple but effective method, i.e., equally attributing the rate loss to the two terms, to solve (2.46) as follows

$$2^{-\frac{B_1}{M-1}} = \frac{1}{2}\frac{b^{\frac{1}{N_r}} - 1}{T\rho_1\rho_2} \tag{2.47}$$

and

$$\frac{A(1+\rho_1 M)}{T\rho_1} 2^{-\frac{B_2}{T}} = \frac{1}{2}\frac{b^{\frac{1}{N_r}} - 1}{T\rho_1\rho_2}. \tag{2.48}$$

Solving (2.47) and (2.48) yields the results (2.44) and (2.45), respectively. One can try different rate loss allocations among the two terms in (2.46) which maybe give different B_1 and B_2 scaling results. However, authors of [21] have done similar works with the conclusion that equal allocation is actually optimal in practical antenna settings. Based on this, we argue that equal rate loss allocation is reasonably near optimal in practice and is an effective method to solve the problem. □

Observation from Theorem 2 indicates that in order to maintain a bounded loss, the feedback size B_1 needs to scale in proportion to both power constraints P_1 and P_2 while B_2 only needs to increase linearly as P_2 grows. In addition, we conclude that P_1 is less influential than P_2 when the SNR is high since P_1 appears only in the denominator of the term in the logarithmic operation of (2.44). One can also find that Theorem 2 is an interesting generalization of Theorem 3 in [21] where N_r takes on the value 1 as a special case. Notably, the pre-log factor of B_2 is $\frac{N_r(N_t-N_r)}{3}$ for N_r antennas per user, or $\frac{N_t-N_r}{3}$ per antenna. This is compared to the factor of $\frac{N_t-1}{3}$ in the ZF beamforming case [21], implying less feedback bits are required in the BD-based precoding scheme than complete diagonalization to achieve the same multiplexing gain.

Remark 2. *When perfect or near-perfect CSI of the BS-RS channel can be obtained at the BS, or alternatively $B_1 = +\infty$, the proposed feedback quality control strategy*

reduces to

$$B_2 = \frac{N_r(N_t - N_r)}{3} P_2(\text{dB}) + N_r(N_t - N_r) \log_2 \frac{A}{N_t \left(b^{\frac{1}{N_r}} - 1 \right)} \tag{2.49}$$

by using the same solving method as in Theorem 2 without, however, the rate loss allocation. By comparing B_2 expressions in both (2.45) and (2.49), we find that less feedback bits are required from each user to the RS to achieve equivalent rate loss control when good CSI knowledge of the BS-RS link is available at the BS.

2.3.3 Numerical Results

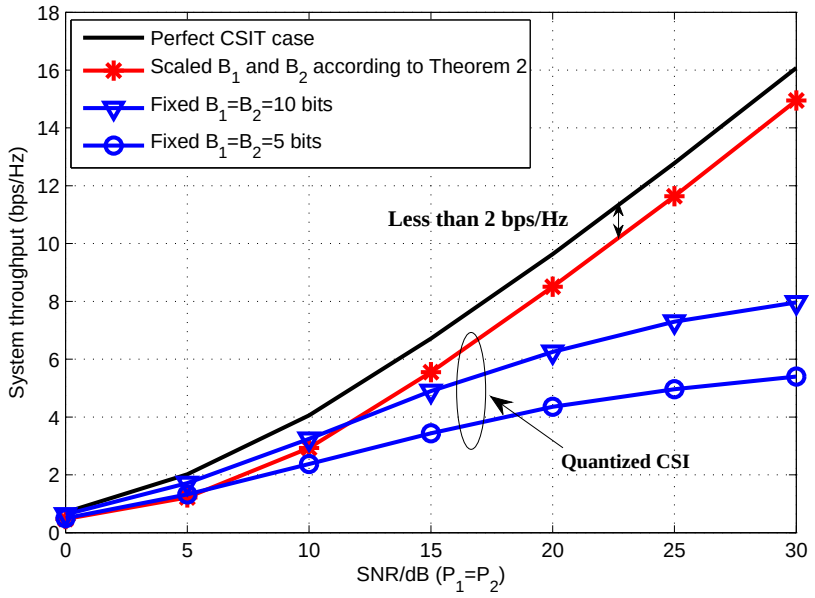

Figure 2.3: Multi-antenna relay-assisted broadcast channel with $M = 4, N_t = 4, N_r = 2$, and $K = \frac{N_t}{N_r} = 2$.

We provide numerical results for $N_t = 4, N_r = 2, K = \frac{N_t}{N_r} = 2$ and $M = 4, 6$ in Fig. 2.3 and Fig. 2.4, respectively. Our feedback quality control goal is to maintain the system throughput gap within 2 bps/Hz, i.e., $b = 4$ in (2.44) and (2.45). Note that the proposed scaling law is merely a sufficient condition, and thus it is a conservative strategy to scale according to Theorem 2 which results in a gap less than 2 bps/Hz. It is also important to note that as the feedback bits B_1 and B_2 given by Theorem 2 can be very large and the computational complexity may be unacceptable at high SNR regime, we exploit statistics of random quantization codebooks to precisely and

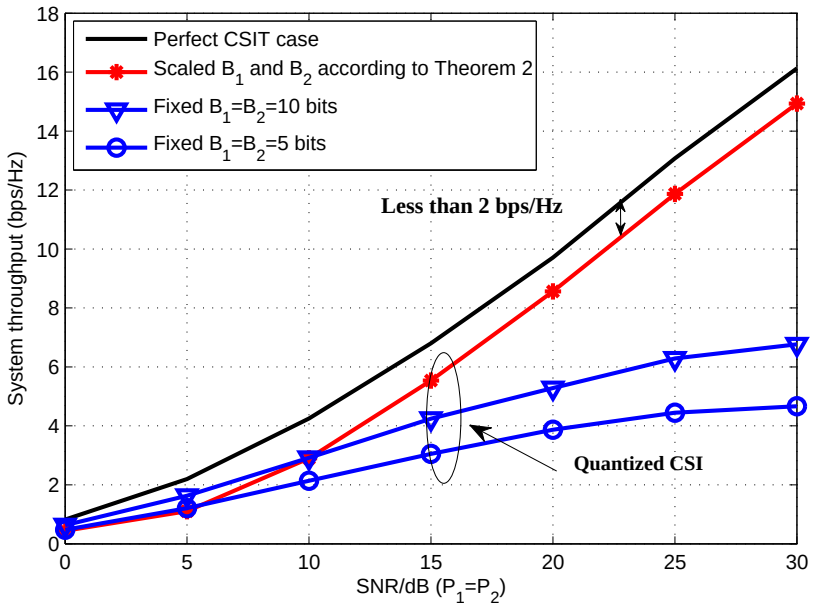

Figure 2.4: Multi-antenna relay-assisted broadcast channel with $M = 6, N_t = 4, N_r = 2$, and $K = \frac{N_t}{N_r} = 2$.

efficiently emulate the quantization process [7], [25]. Moreover, in the figures we show that with fixed feedback quality, the system throughput will eventually saturate at high SNR regime, causing an unbounded rate loss compared to the perfect CSIT system performance. One may notice, however, the fixed size cases achieve better performance than the scaled one at low SNR regime. This is because the feedback quality control scheme proposed in Theorem 2 is designed to address the issue of uncontrolled rate loss at high SNRs. Thereby we set the minimum number of CSI quantization size to 4 bits per user at very low SNRs, in order to achieve a reasonably good performance during this range. In Fig. 2.3 and 2.4, for the fixed feedback quality case, we set $B_1 = B_2 = 5$ bits and $B_1 = B_2 = 10$ bits, both of which are larger than 4 bits. That is why they perform better than the scaled case at low SNRs.

In Fig. 2.5, we present the system performance results when good knowledge about the channel of the first hop link is available at the BS compared with both hops relying on limited feedback for $M = 6, N_t = 4$, and $N_r = 2$. It is observed that the relay system achieves much better capacity performance with fixed feedback size for the RS-User link than when both hops employ limited feedback. Also, we show that with the modified feedback control strategy (2.49) intended for the perfect first link CSI scenario, the system can still maintain a fairly good bounded rate loss relative to the case with perfect CSIT for both links.

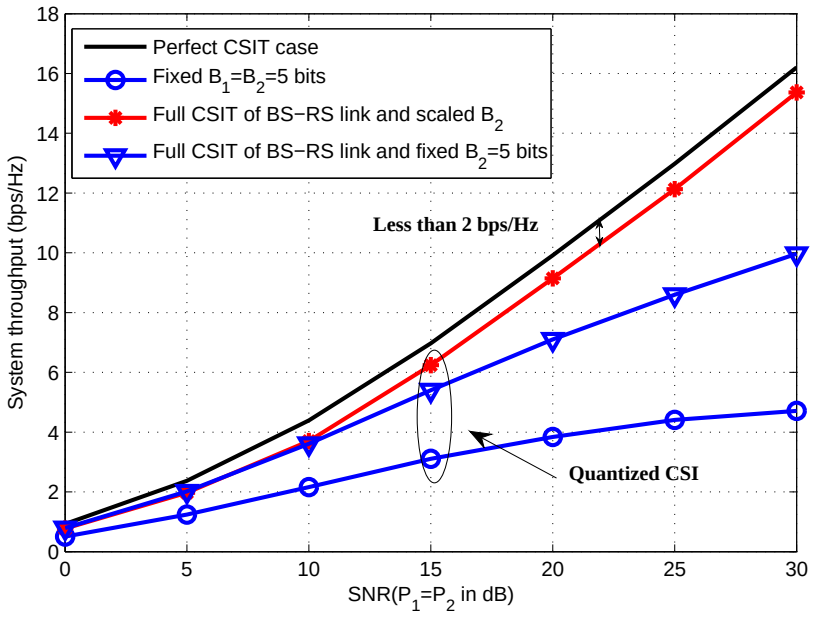

Figure 2.5: Multi-antenna relay-assisted broadcast channel with perfect BS-RS link CSI known at the BS, $M = 6, N_t = 4, N_r = 2$, and $K = \frac{N_t}{N_r} = 2$.

2.4 Summary

In this chapter, we have studied the capacity performance of multi-antenna relay-assisted broadcast channels with limited feedback CSI from both multi-antenna users and the RS. We have derived an upper bound for the system rate loss due to quantization errors incurred by limited feedback. Then a strategy was proposed to scale feedback bits of both links with respect to the transmit power at the BS and RS to maintain a bounded rate loss. The proposed scaling law provides insights into the influence of transmit power on the feedback sizes and gives useful guidelines for practical feedback design in wireless relay systems. Numerical results have verified the accuracy and effectiveness of the derived upper bound and the feedback quality control strategy.

Chapter 3

Low-Complexity Hybrid Precoding for Massive Multiuser MIMO

In the previous chapter, we proposed a limited feedback-based precoding strategy for a relay-assisted multiuser communication system. This scheme has the potential of significantly increasing cell coverage areas and boosting the capacity performance of cell-boundary users in particular. Now back to the very original point, why would we need such a relay station getting involved between the base station and a multitude of users? A typical answer would be that the long distance between the two ends results in a low received signal-to-noise ratio (SNR) at the user side. This will subsequently lead to an unreliable communication link from the transmitter to the receivers, which seems to necessitate the deployment of some intermediate nodes to help amplify the signal and manage interference in the meanwhile. However, recent studies have developed from the very fundamentals of multiple-input multiple-output (MIMO) communication theory and established the fact that more antennas would always help in improving link capacity and reliability [28]. They have brought up an upsurge of research efforts under the terminology of massive MIMO. Unlike currently implemented MIMO systems, where a modest number of transmit antennas are mounted at the base station (e.g., at most 8 antennas in Long Term Evolution-Advanced (LTE-A) systems), the massive MIMO technology proposes to employ a much larger number of antennas (e.g., 256 antennas or even more) to serve multiple users. Most notably, massive MIMO is shown to achieve substantial improvement in both spectral and energy efficiency with simplified transmit precoding/receive combining design [29]-[32]. Simple linear precoding schemes, such as zero-forcing (ZF), are

virtually optimal and comparable to nonlinear precoding like the capacity-achieving dirty paper coding (DPC) in massive MIMO systems [30], [33].

However, to realize the merits of massive MIMO, we are still faced with several practical challenges, one of which is the low-complexity precoding design in the downlink [34]. In most established MIMO systems, to exploit multiple antennas, the prevalent practice is to modify the amplitudes and phases of the complex symbols at the baseband and then upcovert the processed signal to around the carrier frequency after passing through digital-to-analog (D/A) converters, mixers, and power amplifiers (often referred to as the radio frequency (RF) chain). Outputs of the RF chain are then coupled with the antenna elements. In other words, this pure digital implementation of multiuser precoding requires that each antenna element needs to be supported by a dedicated RF chain. This is in fact too expensive to be implemented in massive MIMO systems due to the very large size of the antenna array considered in the scenario.

On the other hand, cost-effective variable phase shifters are readily available with current circuitry technology, making it possible to apply high dimensional phase-only RF or analog processing [35]-[38]. This would potentially enable splitting the full baseband precoding in traditional systems into two cascaded steps, one still in the baseband and the other in the RF domain. Phase-only precoding is considered in [35], [36] to achieve full diversity order and near-optimal beamforming performance through iterative algorithms. The limited baseband processing power can further be exploited to perform multi-stream signal processing as in [37], where both diversity and multiplexing transmissions of MIMO communications are addressed with less RF chains than antennas. [38] then takes into account more practical constraints such as only quantized phase control and finite-precision analog-to-digital (A/D) conversion. Works in [35]-[38], however, do not consider the multiuser scenario and are not aimed to maximize the capacity performance in the large array regime.

In this chapter, we consider the practical constraints of RF chains in massive MIMO systems and propose to design the RF precoder by extracting the phases of the conjugate transpose of the aggregate downlink channel to harvest the large array gain in massive MIMO systems, inspired by [37]. Low-dimensional baseband ZF precoding is then performed based on the equivalent channel obtained from the product of the RF precoder and the actual channel matrix. This hybrid precoding scheme, termed PZF, is shown to approach the performance of the virtually optimal yet practically infeasible full-complexity ZF precoding in a massive multiuser MIMO scenario. We

also note that hybrid baseband and RF precoding has been considered for millimeter wave (mmWave) communications in [39]-[41]. They share the common idea of capturing "dominant" paths present in mmWave channels using RF phase control and the RF processing is constrained, more or less, to choose from array response vectors. In the last part of this chapter, we will briefly show in the simulation the desirable performance of our proposed PZF scheme in a mmWave channel whereas a detailed study on practical precoding design for mmWave communications is deferred to the next chapter.

The rest of this chapter is organized as follows. The system model is introduced in Section 3.1. Section 3.2 presents the details of the proposed low-complexity hybrid precoding scheme for massive MIMO communications including both scheme clarification and spectral efficiency analysis. Section 3.3 provides simulation results of the proposed hybrid precoding scheme in both independent and identically distributed (i.i.d.) Rayleigh fading and sparse mmWave channels. Section 3.4 gives concluding remarks for this chapter.

3.1 System Model

It is universally accepted that in a fully scattered propagation environment, the more antennas at the base station (BS), the more capacity gains downlink users can harness in a broadcast channel. This observation facilitates the research of massive MIMO, where multiuser precoding is performed to boost intended signal while suppress interuser interference. However, as stated above, pure digital implementation of precoding comes at an extremely high cost. In particular, the number of RF chains has to scale with the number of transmit antennas although a much lesser number of users are involved in a typical transmission system. Thus, we might need to exploit the rapid development of cost-effective phase shifters and split a high dimensional precoding into two cascaded stages, namely, the baseband precoding and the RF processing, respectively.

A block diagram for the proposed hybrid baseband and RF precoding scheme of a massive multiuser MIMO system is illustrated in Fig. 3.1, where the BS is equipped with N_t transmit antennas, but driven by a far smaller number of RF chains, namely, K. This practical chain limitation restricts the maximum number of transmitted streams to be K and we assume scheduling exactly K single-antenna users, each supporting single-stream transmission. As discussed, the downlink precoding is di-

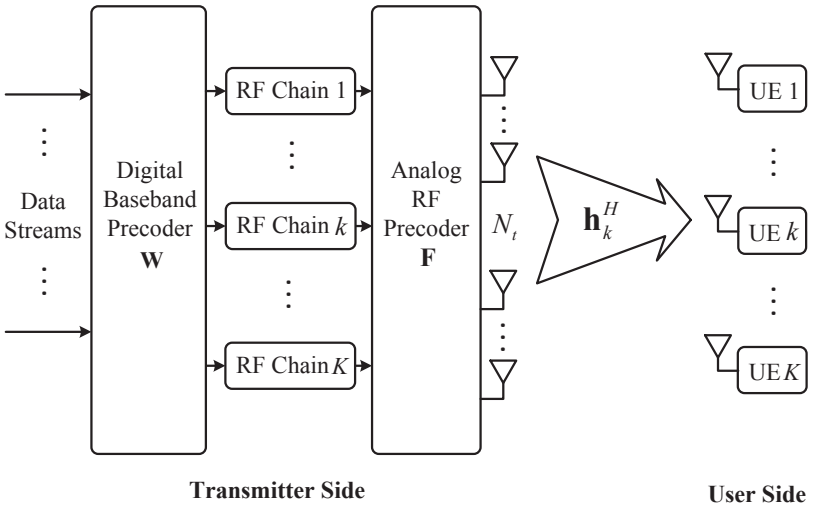

Figure 3.1: System model of a large multiuser MIMO system with hybrid precoding.

vided among digital baseband and analog RF processing, denoted by $\mathbf{W}$ of dimension $K \times K$ and $\mathbf{F}$ of dimension $N_t \times K$, respectively. Notably, both amplitude and phase modifications are feasible for the baseband precoder $\mathbf{W}$, but only phase changes can be made for the RF precoder $\mathbf{F}$ with variable phase shifters and combiners [37]. Thus each entry of $\mathbf{F}$ differs only in phases and the magnitudes are normalized to satisfy $|\mathbf{F}_{i,j}| = \frac{1}{\sqrt{N_t}}$ where $|\mathbf{F}_{i,j}|$ denotes the magnitude of the (i,j)th element of $\mathbf{F}$.

We adopt a narrowband flat fading channel and obtain the sampled baseband signal received at the kth user

$$y_k = \mathbf{h}_k^H \mathbf{F}\mathbf{W}\mathbf{s} + n_k \tag{3.1}$$

where $\mathbf{h}_k^H$ is the downlink channel from the BS to the kth user, and $\mathbf{s} \in \mathbb{C}^{K\times 1}$ denotes the complex signal vector intended for a total of K users, satisfying $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \frac{P}{K}\mathbf{I}_K$ where P is the transmit power at the BS and $\mathbb{E}[\cdot]$ is the expectation operator. To meet the total transmit power constraint, we further normalize $\mathbf{W}$ to satisfy $\|\mathbf{F}\mathbf{W}\|_F^2 = K$. n_k denotes the additive noise, assumed to be circular symmetric Gaussian with unit variance, i.e., $n_k \sim \mathcal{CN}(0,1)$. Then the received signal-to-interference-plus-noise-ratio (SINR) at the kth user is given by

$$\mathrm{SINR}_k = \frac{\frac{P}{K}|\mathbf{h}_k^H \mathbf{F}\mathbf{w}_k|^2}{1 + \sum_{j\neq k} \frac{P}{K}|\mathbf{h}_k^H \mathbf{F}\mathbf{w}_j|^2} \tag{3.2}$$

where $\mathbf{w}_j$ denotes the jth column of $\mathbf{W}$. According to Shannon's third theorem [42], if Gaussian inputs are used, the system can achieve a long-term average (over the fading distribution) spectral efficiency

$$R = \sum_{k=1}^{K} \mathbb{E}\left[\log_2(1 + \mathrm{SINR}_k)\right]. \tag{3.3}$$

3.2 Hybrid Precoding in Massive Multiuser MIMO Systems

In massive MIMO systems, ZF precoding is known as a prominent linear precoding scheme to achieve virtually optimal capacity performance due to the asymptotic orthogonality of user channels in richly scattering environment [30]. It is typically realized through baseband processing, requiring N_t RF chains performing RF-baseband frequency translation and A/D conversion. This tremendous hardware requirement, however, restricts the array size from scaling large.

To alleviate the hardware constraints while realize full potentials provided by excessive antennas in massive multiuser MIMO systems, pure phase changes can be made to the K RF chain outputs using cost-effective RF phase shifters before coupling with N_t transmit antennas. Low-dimensional multi-stream signal processing is then performed at the baseband to enable multiuser communciations. In this section, we propose a low-complexity hybrid precoding scheme, termed phased-ZF (PZF), to approach the performance of the full-complexity ZF precoding, which is, as stated above, practically infeasible due to the requirement of supporting each antenna with a dedicated RF chain. The spectral efficiency achieved by the proposed PZF scheme in i.i.d. Rayleigh fading channels is then analyzed and presented in a closed-form for practical system design references.

3.2.1 Hybrid Precoding Vector Design

The structure shown in Fig. 3.1 is exploited to perform the proposed hybrid baseband and RF joint processing, where the baseband precoder $\mathbf{W}$ modifies both the amplitudes and phases of incoming complex symbols and the RF precoder $\mathbf{F}$ controls phases of the upconverted RF signal. We propose to perform phase-only control at the RF domain by extracting phases of the conjugate transpose of the aggregate

downlink channel from the BS to multiple users. This is to align the phases of channel elements and can thus harvest the large array gain provided by the massive multiuser MIMO systems. To clarify, denote $\mathbf{F}_{i,j}$ as the (i,j)th element of $\mathbf{F}$ and we perform the RF precoding according to

$$\mathbf{F}_{i,j} = \frac{1}{\sqrt{N_t}} e^{j\varphi_{i,j}} \tag{3.4}$$

where $\varphi_{i,j}$ is the phase of the (i,j)th element of the conjugate transpose of the composite downlink channel, i.e., $[\mathbf{h}_1, \cdots, \mathbf{h}_K]$. Here we implicitly assume perfect channel knowledge at the BS which can potentially be obtained, e.g., through uplink channel estimation combined with channel reciprocity in time division duplex (TDD) systems [29]. We note that efficient channel estimation techniques leveraging hybrid structures and rigorous treatment of frequency selectivity are an ongoing research topic of great practical interest.

Then at the baseband, we observe an equivalent channel $\mathbf{H}_{eq} = \mathbf{HF}$ of a low dimension $K \times K$ where $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_K]^H$ is the composite downlink channel. Hence multi-stream baseband precoding can be applied to $\mathbf{H}_{eq}$, where simple low-dimensional ZF precoding is performed as

$$\mathbf{W} = \mathbf{H}_{eq}^H (\mathbf{H}_{eq}\mathbf{H}_{eq}^H)^{-1}\mathbf{\Lambda} \tag{3.5}$$

where $\mathbf{\Lambda}$ is a diagonal matrix, introduced for column power normalization. With this PZF scheme, to support simultaneous transmission of K streams, *hardware complexity is substantially reduced, where only K RF chains are needed, as compared to N_t required by the full-complexity ZF precoding.*

Quantized RF Phase Control: According to (3.4), each entry of the RF precoder $\mathbf{F}$ is designed to differ only in phases and assumes a continuous value. However, in practical implementation, the phase of each entry tends to be heavily quantized due to practical constraints of variable phase shifters. Therefore, we investigate the performance of our proposed hybrid precoding scheme in this realistic scenario, i.e., phases of the KN_t entries of $\mathbf{F}$ are quantized up to B bits of precision, each quantized to its nearest neighbor based on closest Euclidean distance. The phase of each entry of $\mathbf{F}$ can thus be written as $\hat{\varphi} = (2\pi\hat{n}) \,/\, \left(2^B\right)$ where $\hat{n}$ is chosen according to

$$\hat{n} = \arg \min_{n \in \{0, \cdots, 2^B - 1\}} \left| \varphi - \frac{2\pi n}{2^B} \right| \tag{3.6}$$

where φ is the unquantized phase obtained from (3.4).

In the simulation results shown in Figs. 3.2–3.4, for the quantized case, each entry's phase of $\mathbf{F}$ is quantized by choosing from $\{0, \pi\}$ ($B = 1$) and $\{0, \pm\frac{\pi}{2}, \pi\}$ ($B = 2$) based on (3.6) to formulate the quantized $\hat{\mathbf{F}}$. Then the baseband precoder $\mathbf{W}$ is computed by (3.5) with the RF precoder $\hat{\mathbf{F}}$.

3.2.2 Spectral Efficiency Analysis

In this part, we analyze the spectral efficiency achieved by our proposed PZF and the full-complexity ZF precoding in i.i.d. Rayleigh fading channels in the limit of large transmit antenna size N_t. Closed-form expressions are derived herein, revealing the roles different parameters play in affecting system capacity.

Denoting the kth column of $\mathbf{F}$ by $\mathbf{f}_k$, we obtain the received signal at the kth user as

$$y_k = \left[\mathbf{h}_k^H \mathbf{f}_1, \cdots, \mathbf{h}_k^H \mathbf{f}_k, \cdots, \mathbf{h}_k^H \mathbf{f}_K\right] \mathbf{W}\mathbf{s} + n_k \tag{3.7}$$

based on (3.1). As described in Section 3.2.1, $\mathbf{f}_k$ is designed by extracting the phases of $\mathbf{h}_k$, we thus have the diagonal term

$$\mathbf{h}_k^H \mathbf{f}_k = \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} |h_{i,k}| \tag{3.8}$$

where $h_{i,k}$ denotes the ith element of the vector $\mathbf{h}_k$. Under the assumption that each element of $\mathbf{h}_k$ is i.i.d. complex Gaussian random variable with unit variance and zero mean, i.e., $h \sim \mathcal{CN}(0, 1)$, we conclude that $|h|$ follows Rayleigh distribution with mean $\frac{\sqrt{\pi}}{2}$ and variance $1 - \frac{\pi}{4}$. When N_t tends to infinity, the central limit theorem indicates

$$\mathbf{h}_k^H \mathbf{f}_k \sim \mathcal{N}(\frac{\sqrt{\pi N_t}}{2}, 1 - \frac{\pi}{4}). \tag{3.9}$$

Regarding the off-diagonal term, i.e., $j \neq k$, we have

$$\mathbf{h}_k^H \mathbf{f}_j = \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} h_{i,k}^* e^{j\varphi_{i,j}} \tag{3.10}$$

where $(\cdot)^*$ stands for the complex conjugate operation. Its distribution is characterized

in the following lemma.

Lemma 4. *Under i.i.d. Rayleigh fading assumptions, the off-diagonal term* $\mathbf{h}_k^H \mathbf{f}_j$ *in the proposed PZF hybrid precoding scheme is distributed according to* $\mathbf{h}_k^H \mathbf{f}_j \sim \mathcal{CN}(0,1)$.

Proof. We write $h_{i,k}^* = a_{i,k} + jb_{i,k}$ and obtain

$$\begin{aligned}&\mathbf{h}_k^H \mathbf{f}_j \\ =&\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} \left(a_{i,k} \cos \varphi_{i,j} - b_{i,k} \sin \varphi_{i,j}\right) + j \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} \left(a_{i,k} \sin \varphi_{i,j} + b_{i,k} \cos \varphi_{i,j}\right). \end{aligned} \tag{3.11}$$

Under the i.i.d. Rayleigh fading assumption, here comes two useful properties: (1) $a_{i,k}$ and $b_{i,k}$ are independently distributed according to $\mathcal{N}(0, \frac{1}{2})$ and i.i.d. across different i,k's; (2) $\varphi_{i,j}$ is uniformly distributed over $[0, 2\pi]$ and i.i.d. across different i,j's.

For each summation term in the real part of (3.11), we have its mean value

$$\mu_{\text{real}} = \mathbb{E}[a_{i,k}]\mathbb{E}[\cos \varphi_{i,j}] - \mathbb{E}[b_{i,k}]\mathbb{E}\left[\sin \varphi_{i,j}\right] = 0 \tag{3.12}$$

and variance

$$\begin{aligned}\sigma_{\text{real}}^2 =&\mathbb{E}\left[\left(a_{i,k} \cos \varphi_{i,j} - b_{i,k} \sin \varphi_{i,j}\right)^2\right] \\ =&\mathbb{E}\left[a_{i,k}^2\right] \mathbb{E}\left[\cos^2 \varphi_{i,j}\right] + \mathbb{E}\left[b_{i,k}^2\right] \mathbb{E}\left[\sin^2 \varphi_{i,j}\right] - 2\mathbb{E}\left[a_{i,k} b_{i,k} \sin \varphi_{i,j} \cos \varphi_{i,j}\right] \\ =&\frac{1}{2}. \end{aligned} \tag{3.13}$$

Then the independence of these summation terms facilitates using the Central Limit Theorem, yielding

$$\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} a_{i,k} \cos(\varphi_{i,j}) \sim \mathcal{N}(0, \frac{1}{2}). \tag{3.14}$$

The same reasoning applies to the imaginary part of (3.11) which gives

$$\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} b_{i,k} \sin(\varphi_{i,j}) \sim \mathcal{N}(0, \frac{1}{2}). \tag{3.15}$$

Finally, to prove the independence of the real and imaginary parts of (3.11), we write

their covariance as

$$cov = \frac{1}{N_t}\mathbb{E}\left[\sum_{i=1}^{N_t}\left(a_{i,k}\cos\varphi_{i,j} - b_{i,k}\sin\varphi_{i,j}\right)\sum_{i=1}^{N_t}\left(a_{i,k}\sin\varphi_{i,j} + b_{i,k}\cos\varphi_{i,j}\right)\right]. \quad (3.16)$$

The expectation of the cross terms can be easily shown to be zero due to the i.i.d. $\mathcal{N}(0, \frac{1}{2})$ distribution of $a_{i,k}$ and $b_{i,k}$ as well as the independence between $a_{i,k}, b_{i,k}$ and $\varphi_{i,j}$. Then *cov* reduces to

$$\begin{aligned}
&\frac{1}{N_t}\sum_{i=1}^{N_t}\Big\{\mathbb{E}\left[a_{i,k}^2\sin\varphi_{i,j}\cos\varphi_{i,j}\right] - \mathbb{E}\left[b_{i,k}^2\sin\varphi_{i,j}\cos\varphi_{i,j}\right] \\
&\qquad + \mathbb{E}\left[a_{i,k}b_{i,k}\cos^2\varphi_{i,j}\right] - \mathbb{E}\left[a_{i,k}b_{i,k}\sin^2\varphi_{i,j}\right]\Big\} \\
=&\frac{1}{N_t}\sum_{i=1}^{N_t}\left(\mathbb{E}\left[a_{i,k}^2\right] - \mathbb{E}\left[b_{i,k}^2\right]\right)\mathbb{E}\left[\sin\varphi_{i,j}\cos\varphi_{i,j}\right] \\
=&0. \qquad (3.17)
\end{aligned}$$

Since both the real and imaginary parts of (3.11) have been shown to be Gaussian distributed, their covariance being zero then indicates that they are independent. Thus we can conclude $\mathbf{h}_k^H\mathbf{f}_j \sim \mathcal{CN}(0,1)$, which completes the proof. □

Based on Lemma 4, we derive that the magnitude of the off-diagonal term, i.e., $|\mathbf{h}_k^H\mathbf{f}_j|$ follows a Rayleigh distribution with mean $\frac{\sqrt{\pi}}{2}$ and variance $1 - \frac{\pi}{4}$. Compared with the diagonal term $\mathbf{h}_k^H\mathbf{f}_k$ given by (3.9), it is safe at this point to say that *the off-diagonal term is negligible when the transmit antenna number N_t is fairly large.* This implies that the inter-user interference is negligible even without any baseband processing at large N_t! However in our proposed PZF scheme, considering when N_t assumes some medium high value, the residual interference may still deteriorate the system performance, we incorporate ZF processing at the baseband to suppress it as shown in (3.5).

We reason that even with ZF processing at the baseband, the spectral efficiency achieved is still less than it would be if the off-diagonal terms $\mathbf{h}_k^H\mathbf{f}_j$ were precisely zero after RF precoding. In other words, the spectral efficiency achieved by PZF is upper bounded by $K\mathbb{E}\left[\log_2\left(1 + \frac{P}{K}|\mathbf{h}_k^H\mathbf{f}_k|^2\right)\right]$ where the off-diagonal terms are assumed to be zero. The spectral efficiency achieved by PZF is then characterized in the following theorem using the limit equivalence type of argument [43].

Theorem 3. *The spectral efficiency achieved by the proposed low-complexity PZF precoding scheme is tightly upper bounded by* $R_{PZF} \leq K\mathcal{R}$*, where*

$$\lim_{N_t \to \infty} \frac{\mathcal{R}}{\log_2\left(1 + \frac{\pi}{4}\frac{PN_t}{K}\right)} = 1. \tag{3.18}$$

Proof. The spectral efficiency upper bound can be derived as

$$\begin{aligned}
\mathcal{R} &= \mathbb{E}\left[\log_2\left(1 + \frac{P}{K}|\mathbf{h}_k^H \mathbf{f}_k|^2\right)\right] \\
&= \mathbb{E}\left[\log_2\left(1 + \frac{P}{K}\left(y + \frac{\sqrt{\pi N_t}}{2}\right)^2\right)\right] \\
&= \log_2\left(1 + \frac{\pi}{4}\frac{PN_t}{K}\right) + \underbrace{\mathbb{E}\left[\log_2 \frac{1 + \frac{P}{K}\left(y + \frac{\sqrt{\pi N_t}}{2}\right)^2}{1 + \frac{\pi N_t}{4}\frac{P}{K}}\right]}_{\Delta}
\end{aligned} \tag{3.19}$$

where $y \sim \mathcal{N}(0, 1 - \frac{\pi}{4})$. Next, we will prove that Δ is infinitesimally small with increasing N_t, i.e., $\lim_{N_t \to \infty} \Delta = 0$. For notational brevity, we define $\rho \triangleq \frac{P}{K}$ and $a \triangleq \frac{\sqrt{\pi N_t}}{2}$ and equivalently write the limit of Δ as

$$\begin{aligned}
\lim_{N_t \to \infty} \Delta &= \lim_{a \to \infty} \mathbb{E}\left[\log_2 \frac{1 + \rho(y + a)^2}{1 + \rho a^2}\right] \\
&\leq \log_2 \frac{1 + \rho\mathbb{E}\left[(y + a)^2\right]}{1 + \rho a^2} \qquad (3.20) \\
&= \lim_{a \to \infty} \log_2 \frac{1 + \rho(1 - \frac{\pi}{4} + a^2)}{1 + \rho a^2} \\
&= 0 \qquad (3.21)
\end{aligned}$$

where (3.20) comes from applying the Jensen's Inequality.

On the other hand, the limit of Δ can be lower bounded as follows. For brevity, we write the standard deviation of y as $\sigma = \sqrt{1 - \pi/4}$.

$$\begin{aligned}
\lim_{N_t \to \infty} \Delta &\geq \lim_{a \to \infty} \mathbb{E}\left[\log_2 \frac{(y + a)^2}{\frac{1}{\rho} + a^2}\right] \\
&= \lim_{a \to \infty} \mathbb{E}\left[\log_2\left(1 + \frac{y}{a}\right)^2\right] + \lim_{a \to \infty} \log_2 \frac{a^2}{\frac{1}{\rho} + a^2}
\end{aligned} \tag{3.22}$$

$$= \lim_{a\to\infty} \frac{a\log_2 e}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \ln x^2 e^{-\frac{a^2(x-1)^2}{2\sigma^2}} dx \tag{3.23}$$
$$= \lim_{a\to\infty} \frac{2a\log_2 e}{\sqrt{2\pi}\sigma} \int_{0}^{+\infty} (\ln x) e^{-\frac{a^2(x-1)^2}{2\sigma^2}} \left(1 + e^{-\frac{2a^2 x}{\sigma^2}}\right) dx$$

where (3.23) comes from the variable substitution $x = 1 + \frac{y}{a}$ and that the limit of the second term in (3.22) is zero. We further split the integral range and note that the integral over the range $(1, +\infty)$ is strictly positive, yielding

$$\lim_{N_t\to\infty} \Delta \geq \lim_{a\to\infty} \frac{2a\log_2 e}{\sqrt{2\pi}\sigma} \left(1 + e^{-\frac{2a^2\xi}{\sigma^2}}\right) \int_0^1 e^{-\frac{a^2(x-1)^2}{2\sigma^2}} \ln x dx \tag{3.24}$$
$$= \frac{2\log_2 e}{\sqrt{2\pi}\sigma} \int_0^1 \ln x \left[\lim_{a\to\infty} a\left(1 + e^{-\frac{2a^2\xi}{\sigma^2}}\right) e^{-\frac{a^2(1-x)}{2\sigma^2}}\right.$$
$$\left. + \lim_{a\to\infty} a\left(1 + e^{-\frac{2a^2\xi}{\sigma^2}}\right)\left(e^{-\frac{a^2(1-x)^2}{2\sigma^2}} - e^{-\frac{a^2(1-x)}{2\sigma^2}}\right)\right] dx \tag{3.25}$$
$$= \lim_{a\to\infty} \frac{2ae^{-\frac{a^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma\ln 2} \left(1 + e^{-\frac{2a^2\xi}{\sigma^2}}\right) \int_0^1 (\ln x) e^{\frac{a^2 x}{2\sigma^2}} dx \tag{3.26}$$

where (3.24) holds by applying the Mean Value Theorem for Integral with $\xi \in (0, 1)$. (3.26) is valid since the limit of the second term in (3.25) is zero.

Before proceeding, we prove the inequality

$$\int_0^1 (\ln x) e^{mx} dx \geq \frac{-e^m + 1}{m} \tag{3.27}$$

for arbitrarily positive m by showing

$$\int_0^1 (\ln x) e^{mx} dx - \frac{-e^m + 1}{m}$$
$$= \int_0^1 e^{mx}(\ln x + 1) dx$$
$$= \int_0^{1/e} e^{mx}(\ln x + 1) dx + \int_{1/e}^1 e^{mx}(\ln x + 1) dx$$
$$\geq \int_0^{1/e} e^{\frac{m}{e}}(\ln x + 1) dx + \int_{1/e}^1 e^{\frac{m}{e}}(\ln x + 1) dx$$
$$= e^{\frac{m}{e}} \int_0^1 (\ln x + 1) dx$$
$$= 0. \tag{3.28}$$

Then from the proved relation, we substitute $a^2/2\sigma^2$ for m in (3.26) and obtain

$$\lim_{N_t\to\infty}\Delta \geq \lim_{a\to\infty}\frac{2ae^{-\frac{a^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma\ln 2}\left(1+e^{-\frac{2a^2\xi}{\sigma^2}}\right)\frac{2\sigma^2\left(1-e^{\frac{a^2}{2\sigma^2}}\right)}{a^2} \tag{3.29}$$

$$= \frac{4\sigma}{\sqrt{2\pi}\ln 2}\lim_{a\to\infty}\left(1+e^{-\frac{2a^2\xi}{\sigma^2}}\right)\frac{e^{-\frac{a^2}{2\sigma^2}}-1}{a}$$

$$= 0. \tag{3.30}$$

Based on both (3.21) and (3.30), we conclude that $\lim_{N_t\to\infty}\Delta = 0$. Thus as N_t goes large and according to (3.19), we are justified to write

$$\lim_{N_t\to\infty}\frac{\mathcal{R}}{\log_2\left(1+\frac{\pi}{4}\frac{PN_t}{K}\right)} = 1 \tag{3.31}$$

which completes the proof. □

Remark 3. *Considering that the off-diagonal terms $\mathbf{h}_k^H\mathbf{f}_j$'s are essentially negligible when N_t is large, we expect the derived closed-form upper bound to be very tight in the large antenna regime. This is verified in the simulation results as shown in Figs. 3.2–3.4. We observe that the closed-form expression is fairly tight at high N_t, which can serve as a good approximation of the actual spectral efficiency achieved by the proposed PZF hybrid precoding scheme.*

The full-complexity ZF precoding vector (with unit norm) for the kth data stream is found by projecting $\mathbf{h}_k$ onto the nullspace of $\tilde{\mathbf{H}}_k = [\mathbf{h_1}, \cdots, \mathbf{h}_{k-1}, \mathbf{h}_{k+1}, \cdots, \mathbf{h}_K]^H$. In the spectral efficiency analysis, we employ the fact users' channels are asymptotically orthogonal in massive multiuser MIMO systems [29]. Thus we obtain that the full-complexity ZF precoding converges to the conjugate beamforming *with inter-user interference forced to zero*, achieving an SINR

$$\text{SINR}_k \longrightarrow \frac{P}{K}|\mathbf{h}_k|^2, \quad \text{as } N_t\to\infty. \tag{3.32}$$

Then according to (3.3), we obtain the spectral efficiency of full-complexity ZF precoding in the limit of large N_t as [44, Eq. (78)]

$$R_{\text{FC-ZF}} \to K\mathbb{E}\left[\log_2\left(1+\frac{P}{K}|\mathbf{h}_k|^2\right)\right]$$

$$= K e^{\frac{K}{P}} \log_2 e \sum_{n=1}^{N_t} E_n\left(\frac{K}{P}\right) \tag{3.33}$$

by acknowledging that $|\mathbf{h}_k|^2$ follows chi-squared distribution with $2N_t$ degrees of freedom and $E_n(x)$ is the exponential integral of order n defined as

$$E_n(x) = \int_1^\infty e^{-xt} t^{-n} dt, \qquad n = 0, 1, \cdots, \Re(x) \geq 0. \tag{3.34}$$

3.3 Simulation Results

In this section, we provide simulation results to demonstrate the superior performance of our proposed low-complexity PZF precoding scheme in large multiuser MIMO systems. Notably, we respectively apply the proposed PZF scheme in i.i.d. Rayleigh fading channels and the mmWave propagation environment with limited scattering, both achieving highly desirable spectral efficiency performance. Please note that practical precoding design of mmWave communications will be discussed in much greater detail in the next chapter and here we just briefly present the simulation results of applying the proposed PZF precoding scheme to mmWave scenarios.

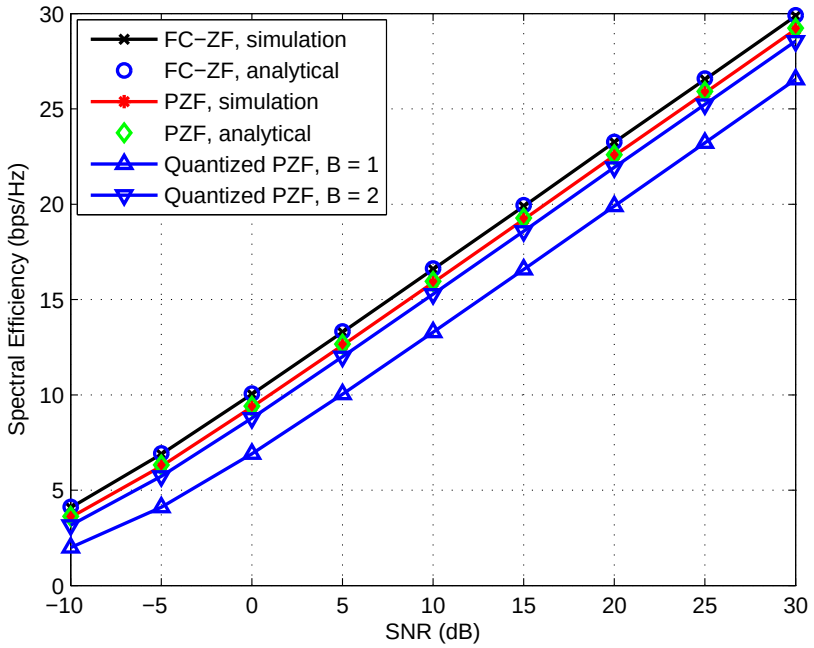

Figure 3.2: Spectral efficiency achieved by different precoding schemes in large multiuser MIMO systems with i.i.d. Rayleigh fading channels where $N_t = 64$ and $K = 2$.

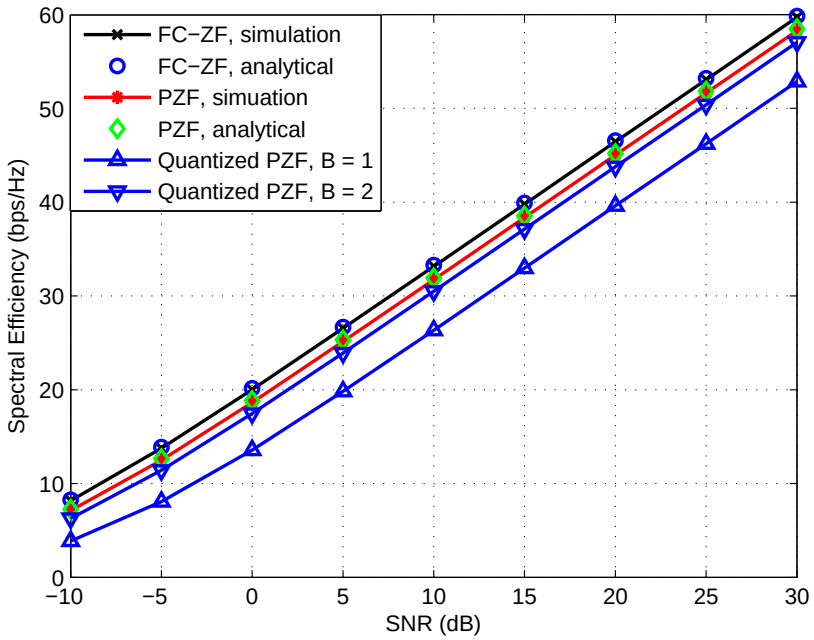

Figure 3.3: Spectral efficiency achieved by different precoding schemes in large multiuser MIMO systems with i.i.d. Rayleigh fading channels where $N_t = 128$ and $K = 4$.

3.3.1 Large Rayleigh Fading Multiuser Channels

We numerically compare the performance of the proposed PZF precoding scheme in Figs. 3.2–3.4 along with its quantized version against the full-complexity ZF scheme, which is deemed virtually optimal in the large array regime but practically infeasible due to the requirement of N_t costly RF chains. It is observed that the proposed PZF precoding performs measurably close to the ideal yet infeasible full-complexity ZF precoding, with less than 1 dB loss but substantially reduced complexity. Compared with the full-complexity ZF precoding, the proposed PZF requires only K RF chains instead of N_t. This complexity reduction is of great practical interest as the number of transmit antennas N_t will normally assume a very high value in typical massive MIMO settings while the user number K remains in a small to medium range. As for the more practical scenario, where the RF phase control is heavily quantized, we find that even with $B = 2$ bits of precision, i.e., rough phase control from $\{0, \pm\frac{\pi}{2}, \pi\}$, the proposed scheme suffers negligible degradation, say less than 1 dB.

The derived analytical spectral efficiency expressions (3.18) and (3.33) are also shown in Figs. 3.2–3.4. We observe that the derived closed-form expressions are quite accurate in characterizing spectral efficiencies achieved by the proposed PZF precoding and full-complexity ZF precoding schemes throughout the whole signal-to-

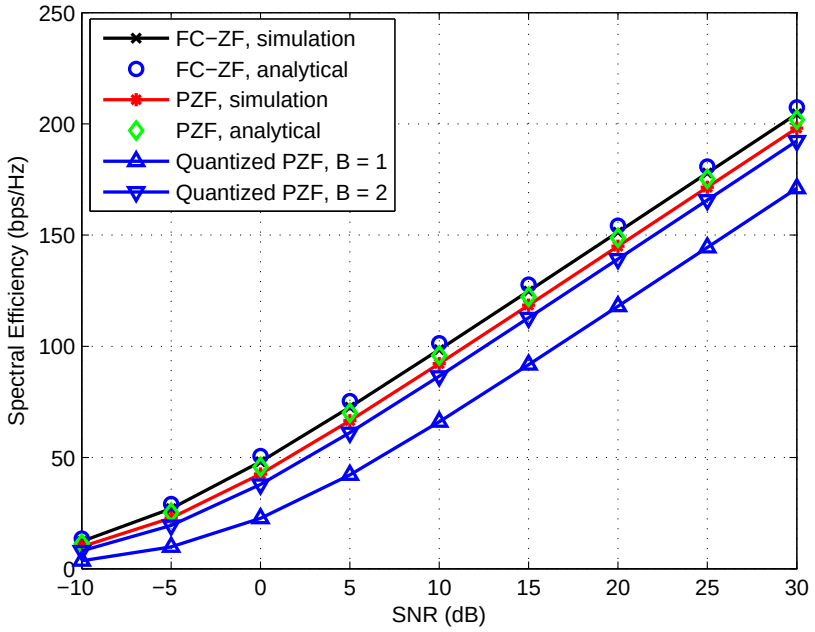

Figure 3.4: Spectral efficiency achieved by different precoding schemes in large multiuser MIMO systems with i.i.d. Rayleigh fading channels where $N_t = 128$ and $K = 16$.

noise (SNR) [1] range, thus providing useful guidelines in practical system designs.

3.3.2 Large mmWave Multiuser Channels

Apart from ideal i.i.d. Rayleigh fading channels, our proposed PZF scheme can also be applied to the mmWave communication which is known to have very limited multipath components. To capture this poor scattering nature, in the simulation, we adopt a geometric channel model [39]-[41]

$$\mathbf{h}_k^H = \sqrt{\frac{N_t}{N_p}} \sum_{l=1}^{N_p} \alpha_l^k \mathbf{a}^H \left(\phi_l^k, \theta_l^k\right) \tag{3.35}$$

where each user is assumed to observe the same number of propagation paths, denoted by N_p, the strength associated with the lth path seen by the kth user is represented by α_l^k, and $\phi_l^k(\theta_l^k)$ is the random azimuth (elevation) angle of departure drawn independently from some continuous distributions, e.g., uniform distributions. The complex gain α_l^k is assumed to be circular symmetric complex Gaussian with unit

[1] Here SNR $= P$ is the common SNR received at each antenna with noise variance normalized to unity.

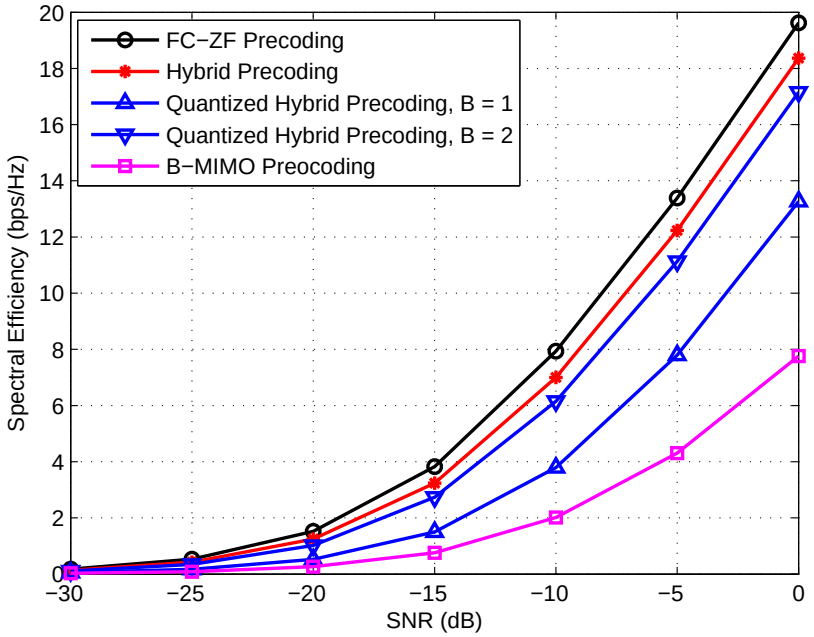

Figure 3.5: Spectral efficiency achieved by different precoding schemes in large mmWave multiuser systems with $N_t = 128, K = 4, d = \frac{1}{2}$ and $N_p = 10$.

variance, i.e., $\alpha_l^k \sim \mathcal{CN}(0,1)$. $\mathbf{a}(\phi_l^k, \theta_l^k)$ is the array response vector depending only on array structures. Here we consider a uniform linear array (ULA) whose array response vector admits a simple expression, given by [40]

$$\mathbf{a}(\phi) = \frac{1}{\sqrt{N_t}} \left[1, e^{j2\pi d \sin(\phi)}, \cdots, e^{j2\pi(N_t-1)d\sin(\phi)}\right]^T \tag{3.36}$$

where d is the normalized antenna separation, normalized to the unit of the carrier wavelength.

As mmWave systems exploit orders of magnitude higher frequency, the free space path loss is then much greater than its low frequency counterparts. We thus investigate the performance of our proposed PZF scheme in the low SNR regime for large multiuser mmWave MIMO communications in Fig. 3.5 by simulations. We compare our proposed PZF scheme against the beamspace MIMO (B-MIMO) scheme proposed in [41], which essentially steers data streams onto the approximate strongest paths (using DFT matrix columns) at the RF domain and performs low-dimensional baseband ZF precoding based on the equivalent channel. For fair comparison, the BS is also assumed to have a total of K chains. The B-MIMO scheme achieves desirable performance in line-of-sight (LoS) channel but fails to capture the energy of sparse

multipath components in non-LoS channels.

3.4 Summary

In this paper, we studied a large multiuser MIMO system under practical RF hardware constraints. We proposed a low-complexity hybrid PZF scheme to approach the desirable yet infeasible full-complexity ZF precoding. The RF processing was designed to harvest the large power gain provided by the excess of antennas with reasonable complexity, and the baseband precoder was then introduced to enable multi-stream processing. Its performance has been characterized in a closed form and further demonstrated in both i.i.d. Rayleigh fading channels and poorly scattered mmWave propagation environments through computer simulations.

Chapter 4

Sparse Precoding for Chain Limited mmWave Multiuser Systems

In forthcoming years, the wireless communication system is envisioned to face critical challenges of explosive increase of demands for high data rates due to continuous roll-out of various bandwidth-intensive mobile devices. Large amounts of efforts have been invested in nearly all aspects of increasing spectral efficiency, including frequency reuse, further splitting cells, coordinated multi-point transmission (CoMP), relaying, etc. But none of them alone is seen as a viable solution to meet the ever increasing traffic demands forseen in 2020 and beyond [39]. In recent years, the availability of abundant underutilized spectrum at millimeter wave (mmWave) frequencies (normally in the range of 30-300 GHz) has spurred an interest to enable multi-Gbps high speed communication through using large chunk of spectrum [45], [46]. Apart from much touted applications such as wireless local area network (WLAN), wireless personal area network (WPAN), high definition (HD) video streaming, etc., the outdoor cellular broadband system has also been nominated as another potential application scenario [47]–[50]. One promising characteristic of mmWave communications is the dramatic decrease of carrier wavelength, which makes it possible to pack a large number of antennas into a reasonable physical size. This excess of transmit antennas provides considerable beamforming gain to combat unfavorable propagation impairments experienced by orders-of-magnitude higher mmWave carrier frequency, enabled by the massive multiple-input multiple-output (MIMO) technology.

As discussed in Chapter 3, to reap full gains provided by the massive amount of antennas, we are facing a number of challenges, with one being the reduction of precoding complexity given the hardware constraints. This is true not only at lower frequencies which is the primary focus in Chapter 3, but also at high mmWave bands. On top of common characteristics like having excessive number of antennas and being unable to support each transmit antenna with a dedicated radio frequency (RF) chain, the mmWave band has a differentiating characteristic compared to its low frequency counterparts, i.e., very sparse propagation paths. This is both a friend and foe. The sparsity in multi-path components of mmWave propagation makes the channel between transmitter and receiver more likely to be rank deficient, i.e., being unable to harvest as much multiplexing gain as in a rich scattering environment. On the other hand, as long as we are not transmitting "too many" data streams simultaneously (less than the channel matrix rank), we will be able to design very low-complexity precoding schemes by sending data streams specifically in those sparse channel directions with even simplified channel estimation. Along this line, we develop sparse precoding in this chapter for RF chain-limited mmWave MIMO systems and theoretically evaluate the performance of the proposed scheme compared to the virtually optimal yet practically infeasible full-complexity zero-forcing (ZF) precoding.

Note that the capacity optimality of beam steering in sparse mmWave MIMO systems have been studied in [49]. A hybrid design of jointly combining RF beamforming and baseband precoding is proposed in [50] and its associated prototyping is introduced in great detail in [39], where the RF beamforming and baseband precoding vectors are chosen among predetermined RF beam sets and baseband codebooks, respectively. [40] also proposes a hybrid precoding design for mmWave single user MIMO communications, which essentially approaches the optimal singular value decomposition (SVD)-based precoding and receiver processing through combining RF beamforming and baseband precoding using the sparse reconstruction method mostly known to be studied in the compressed sensing (CS) theory. The hybrid channel estimation method is then correspondingly presented in [52]. Different from these hybrid approaches, the low-resolution ADC design is considered to address the high complexity and power consumption issues associated with mmWave communications in [51], [53]. However, none of them is specialized to the multiuser communications at mmWave frequencies and no rigorous mathematical evaluation of the performance achieved by the proposed schemes is presented therein. This chapter of the thesis is written to specifically address these issues. We also note that another popular way to

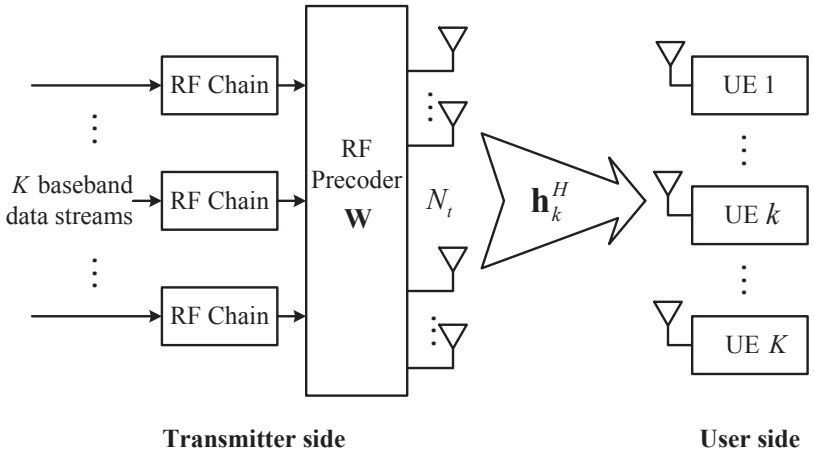

Figure 4.1: System model of the large mmWave MIMO broadcast channel with only RF processing using variable phase shifters.

reduce RF chain requirement is to use antenna selection [54]-[57] where even simple phase shifters are made unnecessary. Selection schemes, however, provide very limited array gain and are shown to achieve poor performance in correlated channels such those experienced in mmWave communications [57].

The rest of the chapter is organized as follows. The system model is presented in Section 4.1, and the proposed MUBS scheme along with its analytical evaluations is discussed in Section 4.2. Two enhancements of the MUBS scheme is developed in Section 4.3, followed by simulation results presented in Section 4.4. Concluding remarks are finally made in Section 4.5.

4.1 System Model

We consider a large mmWave multiuser MIMO system as shown in Fig. 4.1, where an antenna array with N_t elements is mounted on the base station (BS). Due to the high cost of mixed signal processing, only $K(\ll N_t)$ RF chains are available to convert digital baseband signals to analog RF waveforms. We assume scheduling the same number of single-antenna users as that of RF chains, i.e., K and the data for user k is weighted by a vector $\mathbf{w}_k$ at the RF domain using pure phase shifters before being sent by N_t transmit antennas. Thus each entry of the precoding vector $\mathbf{w}_k$ has constant modulus constraint and normalized to satisfy $|\mathbf{w}_i| = \frac{1}{\sqrt{N_t}}$ where $|\mathbf{w}_i|$ denotes the magnitude of the ith element of $\mathbf{w}$. The sampled baseband signal received by the

user k is then given by

$$y_k = \sum_{j=1}^{K} \mathbf{h}_k^H \mathbf{w}_j s_j + n_k \tag{4.1}$$

where s_j denotes the scalar symbol intended for the jth user and n_k is the complex Gaussian noise with zero mean and unit variance, i.e., $n_k \sim \mathcal{CN}(0,1)$. The downlink channel from the BS to the kth user is represented by $\mathbf{h}_k^H$. Since mmWave channels are expected to have very limited multipath components, we adopt a geometric channel model to capture this poor scattering nature [40], [41], [58]-[60], characterized by

$$\mathbf{h}_k^H = \sqrt{\frac{N_t}{N_p}} \sum_{l=1}^{N_p} \alpha_l^k \mathbf{a}^H \left(\phi_l^k, \theta_l^k\right) \tag{4.2}$$

where N_p is the number of paths seen by each user, α_l^k is the complex gain of the lth path seen by the kth user, $\phi_l^k(\theta_l^k)$ is the random azimuth (elevation) angle of departure drawn independently from such distributions as truncated Laplacian distribution or uniform distribution over $[0, 2\pi]$ [60]. The complex gain α_l^k is assumed to be circular symmetric complex Gaussian with unit variance, i.e., $\alpha_l^k \sim \mathcal{CN}(0,1)$. Directional antennas can be incorporated by letting $\alpha_l^k \sim \mathcal{CN}\left(0, \Lambda(\phi_l^k, \theta_l^k)\right)$ where $\Lambda(\cdot)$ denotes the directional element gain [49]. $\mathbf{a}(\phi_l^k, \theta_l^k)$ is the unit spatial signature in the corresponding angle depending on array structures. Here we consider a uniform linear array (ULA), where the antennas are evenly spaced on a straight line. The unit spatial signature, in this case, follows a simple expression, given by

$$\mathbf{a}(\phi) = \frac{1}{\sqrt{N_t}} \left[1, e^{j2\pi d \sin(\phi)}, \cdots, e^{j(N_t-1)2\pi d \sin(\phi)}\right]^T \tag{4.3}$$

where d is the normalized antenna separation, normalized to the unit of the carrier wavelength. Note that the effects of elevation angles θ_l^k can be accounted for by employing the uniform planar array (UPA) structure [49] and conclusions made in this thesis about ULA are readily extended to the UPA structure.

With the system model introduced above, we can easily obtain the signal-to-interference-plus-noise-ratio (SINR) of the kth user as

$$\mathrm{SINR}_k = \frac{\frac{P}{K}|\mathbf{h}_k^H \mathbf{w}_k|^2}{1 + \sum_{j \neq k} \frac{P}{K}|\mathbf{h}_k^H \mathbf{w}_j|^2} \tag{4.4}$$

under the assumption that each of the symbols has power P/K where P is the total transmit power at the BS. If Gaussian inputs are used, the system can achieve a long-term average (over the fading distribution) per-user rate as given by

$$R = \mathbb{E}\left[\log_2(1 + \mathrm{SINR}_k)\right]. \tag{4.5}$$

In the following, we will analyze the rate performance of the proposed low-complexity precoding scheme based on this formula.

4.2 Multiuser Beam Steering in Large mmWave Channels

As understood from Chapter 3, ZF precoding is known to be virtually optimal in massive multiuser MIMO systems due to its capability to concentrate intended signal energy and meanwhile cause no interuser interference in asymptotically orthogonal multiuser channel settings [30]. We also learn that de-facto implementation of MIMO technologies through pure baseband precoding (referred to as full-complexity ZF precoding thereafter) is facing critical challenges when the number of transmit antennas scales large as considered in Chapter 3. This is the case at lower frequencies, e.g., less than 3 GHz in most deployed cellular systems. It is even more so at mmWave bands because the high costs and power consumption of mmWave mixed-signal processing hardware make the issue even worse. That is, it's more of practical challenges in mmWave communication systems to support each antenna element with a dedicated RF chain performing digital-to-analog conversion, frequency translation, power amplifying, etc.

However, the mmWave band boasts a distinct feature than its lower frequency counterparts that might be taken good advantage of when designing precoding schemes, i.e., very sparse channels. The channel sparsity means there exist only a few dominant paths connecting the transmitter to the receiver and we are able to simplify the precoding design by sending data streams only along those sparse paths. In this section, we propose a sparse multiuser beam steering scheme (MUBS) in an effort to realize full potentials of the massive antennas mounted at mmWave devices with reasonable hardware complexity. We further study the large system performance of both the virtually optimal full-complexity ZF precoding and the proposed MUBS

schemes in mmWave channels, and then characterize the asymptotic performance gap between them. The proposed MUBS scheme is shown to be able to achieve most of full-complexity ZF rates with, however, considerably reduced hardware complexity. Considering the near-optimality of the full-complexity ZF precoding, we are justified to claim that the proposed MUBS scheme embodies a good tradeoff between hardware complexity and capacity performance in mmWave MIMO systems.

4.2.1 Multiuser Beam Steering Vector Design

Before proceeding, we first introduce the following lemma from [49] characterizing the asymptotic orthogonality property of large array mmWave multiuser systems, which will be exploited to introduce our proposed scheme.

Lemma 5. *[49] For a ULA system with azimuth angles of departure drawn independently from a continuous distribution, the unit array spatial signatures satisfy* $\mathbf{a}(\phi_{l_0}^{k_0}) \perp span\{\mathbf{a}(\phi_l^k)|\forall k \neq k_0 \text{ or } l \neq l_0\}$ *as the antenna number* N_t *goes infinite and the total number of paths is* $KN_p = o(N_t)$.

According to Lemma 5, each path signature contributing to the channel $\mathbf{h}_{k_1}$ is asymptotically orthogonal to that of $\mathbf{h}_{k_2}, \forall k_1 \neq k_2$, hence their linear combinations are also orthogonal in asymptotic sense, i.e., $\mathbf{h}_{k_1} \perp \mathbf{h}_{k_2}, \forall k_1 \neq k_2$. Meanwhile, we know that the ZF precoding vector for the kth data stream is found by projecting $\mathbf{h}_k$ onto the nullspace of the concatenation $\tilde{\mathbf{H}}_k = [\mathbf{h_1}, \cdots, \mathbf{h}_{k-1}, \mathbf{h}_{k+1}, \cdots, \mathbf{h}_K]^H$. We thus come to the conclusion: the ZF precoding converges, as the number of transmit antennas N_t tends to infinity, to conjugate beamforming with interuser interference forced to zero[1], where the beamforming vector for each user is simply the conjugate transpose of the corresponding downlink channel vector between the BS and that user. In mathematical representations, it would be

$$\mathbf{w}_k^{ZF} \longrightarrow \tilde{\mathbf{h}}_k, \quad \text{as } N_t \to +\infty. \tag{4.6}$$

where $\tilde{\mathbf{h}}_k = \frac{\mathbf{h}_k}{||\mathbf{h}_k||}$ is the normalized channel state information.

Therefore, to approach ZF precoding given RF chain limitations in large mmWave multiuser systems, a good solution would be trying to steer the beams for the kth user towards the channel $\mathbf{h}_k$ itself in the RF domain. This is generally difficult to realize as

[1]It is worth pointing out that the ZF precoding considered in our system setup actually tends to the interference free benchmark system as discussed in [30] due to that K is much less than N_t.

wireless channels typically comprise an aggregation of multiple paths and it is hard to capture them with pure RF phase changes. Fortunately, mmWave propagation features very limited multipath components. In other words, we can capture the channel power by using just a few RF beams to track each of the paths. When the RF chains are limited, it is then justified to track only the strongest path in the RF domain for each user and try to capture as much channel power as possible. Hence we summarize our scheme, termed MUBS, as follows:

- **Step 1**: Apply compressed sensing-based signaling to estimate all departure angles ϕ_l^k and path strength α_l^k at each user [61].
- **Step 2**: Each user feeds back to the BS the angle ϕ_{max}^k of the strongest path (i.e., largest $|\alpha_l^k|$) among all N_p paths seen by that user.
- **Step 3**: The BS performs analog beamforming according to $\mathbf{a}(\phi_{max}^k)$ at the RF domain before sending the symbol intended for the kth user to the antennas.

The proposed MUBS scheme exploits the sparse scattering nature of mmWave propagations and is applicable even in frequency division duplexing (FDD) systems by using compressed sensing-based downlink channel estimation techniques [61][2]. The channel feedback load is rather light as only one scalar symbol feedback is needed for each user during one coherent period. Most notably, the MUBS scheme can be implemented with considerably low complexity as merely K RF chains are necessitated instead of N_t as demanded by baseband ZF precoding. Baseband preprocessing is even made unnecessary and all MUBS requires is a phase-shifting network at the RF domain.

In the following, we will present the spectral efficiency analysis of both full-complexity ZF precoding and the proposed MUBS schemes in the large mmWave multiuser channels to show the desirable performance of our proposed scheme.

4.2.2 Spectral Efficiency Achieved by ZF precoding

As discussed in Section 4.2.1, the full-complexity ZF precoding for large mmWave channels converges to conjugate beamforming, i.e., $\mathbf{w}_k = \mathbf{h}_k$, with *interuser interference forced to zero*. This amounts to considering only the channel direction itself for each user without paying attention to any interuser interference when analyzing

[2]Please refer to [62]–[65] for further reading of the compressed sensing theory.

its received SINR. Then according to (4.4), we know that as the transmit antenna number N_t goes large, the SINR achieved by ZF precoding can be written in the following form

$$\begin{aligned}\mathrm{SINR}_k &= \frac{P}{K}\left|\left[\sqrt{\frac{N_t}{N_p}}\sum_{l=1}^{N_p}\alpha_l^k\mathbf{a}^H\left(\phi_l^k\right)\right]\left[\frac{\sum_{l=1}^{N_p}\left(\alpha_l^k\right)^H\mathbf{a}(\phi_l^k)}{\sqrt{\sum_{l=1}^{N_p}|\alpha_l^k|^2}}\right]\right|^2 \\ &\overset{a.s.}{=} \frac{PN_t}{KN_p}\sum_{l=1}^{N_p}\left|\alpha_l^k\right|^2 \end{aligned} \tag{4.7}$$

where $a.s.$ stands for almost sure convergence and results from the orthogonality $\mathbf{a}(\phi_l^k) \perp \mathbf{a}(\phi_m^k), \forall m \neq l$ when N_t is fairly large as described in Lemma 5. After assuming $\alpha_l^k \sim \mathcal{CN}(0,1)$, we conclude that the sum $\sum_{l=1}^{N_p}\left|\alpha_l^k\right|^2$ is a chi-squared random variable with $2N_p$ degrees of freedom, i.e., distributed according to $\chi^2_{2N_p}$.

With the received SINR given by (4.7) and according to the rate formula (4.5), we compute the per user rate for the full-complexity ZF precoding scheme, which is summarized in the following lemma.

Lemma 6. *When the transmit antenna number N_t is large and the number of users K is substantially less than N_t in the considered multiuser mmWave system, the full-complexity ZF precoding scheme achieves a per user rate*

$$R_{ZF} \overset{a.s.}{=} e^{\frac{1}{\rho}}\log_2 e\sum_{k=1}^{N_p}E_k\left(\frac{1}{\rho}\right) \tag{4.8}$$

where $E_k(x)$ is the exponential integral of order k defined as

$$E_k(x) = \int_1^\infty e^{-xt}t^{-k}dt, \quad k = 1,\cdots,\Re(x) > 0. \tag{4.9}$$

Proof. For notational brevity, we define $\rho = PN_t/(KN_p)$. The per-user rate achieved by baseband ZF precoding can then be computed according to (4.5) as

$$R_{ZF} \overset{a.s.}{=} \mathbb{E}\left[\log_2\left(1+\rho\sum_{l=1}^{N_p}\left|\alpha_l^k\right|^2\right)\right]$$

$$= \frac{\log_2 e}{\Gamma(N_p)\rho^{N_p}} \int_0^\infty \ln(1+t) t^{N_p-1} e^{-\frac{t}{\rho}} dt \tag{4.10}$$

by plugging in the probability density function of the chi-squared random variable. The above integral can be evaluated in a closed form as given in [44, Eq. (78)]

$$\begin{aligned} &\int_0^\infty \ln(1+t) t^{n-1} e^{-\mu t} dt \\ =&(n-1)! e^{\mu} \sum_{k=1}^{n} \frac{\Gamma(-n+k, \mu)}{\mu^k} \end{aligned} \tag{4.11}$$

where $\Gamma(\cdot,\cdot)$ is the complementary incomplete gamma function defined by $\Gamma(\alpha, x) = \int_x^\infty e^{-t} t^{\alpha-1} dt$. Hence we have

$$\begin{aligned} R_{ZF} &\overset{a.s.}{=} e^{\frac{1}{\rho}} \log_2 e \sum_{k=1}^{N_p} \rho^{k-N_p} \Gamma\left(k - N_p, \frac{1}{\rho}\right) \\ &= e^{\frac{1}{\rho}} \log_2 e \sum_{k=1}^{N_p} E_k\left(\frac{1}{\rho}\right) \end{aligned} \tag{4.12}$$

where the last equation follows from the relation $E_k(x) = x^{k-1}\Gamma(1-k, x)$. □

4.2.3 Rate Achieved by MUBS precoding

According to the MUBS precoding scheme described in Section 4.2.1, MUBS essentially steers the beam towards the strongest paths seen by each user as represented by $\mathbf{a}\left(\phi_{max}^k\right)$ by manipulating RF phase shifters. Assuming no distortion happens during the channel estimation (using compressed sensing-based techniques) and feedback phases, we obtain the SINR achieved by the MUBS precoding scheme at very large N_t based on (4.4) as

$$\begin{aligned} \mathrm{SINR}_k &= \frac{P}{K} \left| \left[\sqrt{\frac{N_t}{N_p}} \sum_{l=1}^{N_p} \alpha_l^k \mathbf{a}^H\left(\phi_l^k, \theta_l^k\right) \right] \mathbf{a}(\phi_{max}^k) \right|^2 \\ &\overset{a.s.}{=} \frac{P N_t}{K N_p} \left|\alpha_{max}^k\right|^2 \end{aligned} \tag{4.13}$$

where α_{max}^k stands for the complex gain α_l^k with the largest magnitude among all N_p paths seen by the kth user and $\mathbf{a}(\phi_{max}^k)$ stands for the spatial signature corresponding

to that path. The last almost sure convergence result comes from the orthogonality $\mathbf{a}(\alpha_l^k) \perp \mathbf{a}(\alpha_m^k), \forall m \neq l$ when N_t is fairly large as described in Lemma. 5. According to the rate formula (4.5), we may then analytically evaluate the per user rate achieved by our proposed MUBS precoding scheme under the assumption the complex gain $\alpha_l^k \sim \mathcal{CN}(0,1)$. The result is summarized in the lemma below.

Lemma 7. *When the transmit antenna number N_t is large and the number of users K is substantially less than N_t in the considered multiuser mmWave system, the proposed MUBS precoding scheme achieves a per user rate*

$$R_{MUBS} = \log_2 e \sum_{k=1}^{N_p} (-1)^{k-1} \binom{N_p}{k} e^{\frac{k}{\rho}} E_1\left(\frac{k}{\rho}\right) \tag{4.14}$$

where $\binom{n}{k}$ is the binomial coefficient indexed by n and k, given by $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ under suitable circumstances. $E_1(x)$ is defined in (4.9).

Proof. As we have $\alpha_l^k \sim \mathcal{CN}(0,1)$ and thus $\left|\alpha_l^k\right|^2$ is exponentially distributed with mean 1 whose probability density function is given by

$$f_{\alpha_l^k}(x) = e^{-x}. \tag{4.15}$$

Since $|\alpha_{max}^k|^2$ is the maximum amongst N_p complex gains $|\alpha_l^k|^2$, we can obtain the probability density function of $\left|\alpha_{max}^k\right|^2$ as given by

$$f(x) = \begin{cases} N_p \left(1 - e^{-x}\right)^{N_p - 1} e^{-x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The per-user rate achieved by the proposed MUBS precoding scheme is developed based on (4.5) as follows where we also define $\rho = PN_t/(KN_p)$ for notational brevity.

$$\begin{aligned} & R_{MUBS} \\ =& \mathbb{E}\left[1 + \frac{PN_t}{KN_p}\left|\alpha_{max}^k\right|^2\right] \\ =& N_p \int_0^\infty \log_2(1+\rho x)\left(1 - e^{-x}\right)^{N_p - 1} e^{-x} dx \\ \overset{(a)}{=}& N_p \log_2 e \sum_{k=1}^{N_p} (-1)^{k-1} \binom{N_p - 1}{k-1} \int_0^\infty \ln(1+\rho x) e^{-kx} dx \end{aligned}$$

$$\overset{(b)}{=} N_p \log_2 e \sum_{k=1}^{N_p} \frac{(-1)^k}{k} \binom{N_p - 1}{k-1} e^{-\frac{k}{\rho}} E_i\left(-\frac{k}{\rho}\right)$$
$$\overset{(c)}{=} \log_2 e \sum_{k=1}^{N_p} (-1)^{k-1} \binom{N_p}{k} e^{\frac{k}{\rho}} E_1\left(\frac{k}{\rho}\right) \tag{4.16}$$

where (a) results from the binomial expansion theorem and (b) follows from [66, Eq. (4.337.2)] with $E_i(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt$. Step (c) follows from the relation $E_i(-x) = -E_1(x)$ for $\Re(x) > 0$. □

4.2.4 Asymptotic Rate Loss Convergence

To this end, we have analytically evaluated the rates for the full-complexity ZF precoding and our proposed MUBS precoding schemes for large mmWave multiuser MIMO systems under the assumption $K \ll N_t$. We know that the ZF precoding is virtually optimal in the large antenna regime from massive MIMO theories [30]. Hence we are interested in the performance gap between our simplified MUBS precoding design and the ZF scheme. It will indicate how well the proposed low-complexity scheme approaches the optimal performance in large mmWave communications and suggest possible further improvements. In the following theorem, we analyze the asymptotic performance of the rate loss of the proposed MUBS as compared against the full-complexity ZF precoding scheme when N_t is taken to infinity.

Theorem 4. *The rate loss of MUBS relative to ZF precoding incurred by RF chain limitations, defined by $\Delta R = R_{ZF} - R_{MUBS}$, converges asymptotically to*

$$\Delta R \to \log_2 e \sum_{k=2}^{N_p} \left[\frac{1}{k-1} + (-1)^{k-1} \binom{N_p}{k} \ln k \right] \tag{4.17}$$

as the transmit antenna number N_t tends to infinity.

Proof. When N_t approaches infinity, we equivalently obtain $\rho = PN_t/(KN_p)$ goes infinite for a given system setting. Then by substituting the analytical rate expressions of ZF and MUBS, namely (4.8) and (4.14) into $\Delta R = R_{ZF} - R_{MUBS}$, we obtain the

rate loss

$$\begin{aligned}
&\lim_{\rho\to+\infty}\frac{\Delta R}{\log_2 e}\\
&=\lim_{\rho\to+\infty} e^{\frac{1}{\rho}}\sum_{k=1}^{N_p}E_k\left(\frac{1}{\rho}\right)-\sum_{k=1}^{N_p}(-1)^{k-1}\binom{N_p}{k}e^{\frac{k}{\rho}}E_1\left(\frac{k}{\rho}\right)\\
&=\lim_{\rho\to+\infty}\Bigg\{\underbrace{E_1\left(\frac{1}{\rho}\right)\left[e^{\frac{1}{\rho}}+\sum_{k=1}^{N_p}(-1)^{k}\binom{N_p}{k}e^{\frac{k}{\rho}}\right]}_{S_1}\\
&\quad+\underbrace{\sum_{k=1}^{N_p}(-1)^{k}\binom{N_p}{k}e^{\frac{k}{\rho}}\int_k^1 e^{-\frac{t}{\rho}}t^{-1}dt}_{S_2}+\underbrace{e^{\frac{1}{\rho}}\sum_{k=2}^{N_p}E_k\left(\frac{1}{\rho}\right)}_{S_3}\Bigg\}
\end{aligned}\tag{4.18}$$

where the last equation is based on the following argument

$$\begin{aligned}
E_1(kx)&=\int_1^\infty e^{-kxt}t^{-1}dt\\
&\overset{\alpha=kt}{=}\int_k^\infty e^{-x\alpha}\left(\frac{\alpha}{k}\right)^{-1}\frac{1}{k}d\alpha\\
&=E_1(x)+\int_k^1 e^{-x\alpha}\alpha^{-1}d\alpha.
\end{aligned}\tag{4.19}$$

In the following, we will show that the three summation terms S_1, S_2, and S_3 in (4.18) have finite limiting values, respectively and they will add up to the right side of (4.17).

For the first term S_1, we have

$$\begin{aligned}
&\lim_{\rho\to+\infty}S_1\\
&\overset{(d)}{=}\lim_{\rho\to+\infty}\left\{E_1\left(\frac{1}{\rho}\right)\left(1-e^{\frac{1}{\rho}}\right)\left[\left(1-e^{\frac{1}{\rho}}\right)^{N_p-1}-1\right]\right\}\\
&=\lim_{x\to0^+}\left\{\frac{E_1(x)}{\frac{1}{1-e^x}}\left[(1-e^x)^{N_p-1}-1\right]\right\}\\
&\overset{(e)}{=}\lim_{x\to0^+}\left\{\frac{-(1-e^{-x})^2}{x}\right\}\times\lim_{x\to0^+}\left\{\left[(1-e^x)^{N_p-1}-1\right]\right\}\\
&\overset{(f)}{=}0
\end{aligned}\tag{4.20}$$

where step (d) follows from using binomial expansion theorem and step (e) results from applying the L'Hopital's rule to the term $\frac{E_1(x)}{\frac{1}{1-e^x}}$ and the relation $E_1'(x) = -E_0(x) = -e^{-x}/x$. Step (f) is obtained by applying the L'Hopital's rule to the term $\frac{-\left(1-e^{-x}\right)^2}{x}$ and acknowledging that $\lim_{x\to 0^+}\left[(1-e^x)^{N_p-1}-1\right] = -1$.

For the second term S_2, we have

$$\begin{aligned}\lim_{\rho\to+\infty} S_2 &= \lim_{\rho\to+\infty}\left\{\sum_{k=1}^{N_p}(-1)^k\binom{N_p}{k}e^{\frac{k}{\rho}}\int_k^1 e^{-\frac{t}{\rho}}t^{-1}dt\right\}\\ &=\sum_{k=1}^{N_p}(-1)^k\binom{N_p}{k}\int_k^1 t^{-1}dt\\ &=\sum_{k=1}^{N_p}(-1)^{k-1}\binom{N_p}{k}\ln k. \end{aligned} \tag{4.21}$$

For the third term S_3, we use the formula [66, Eq. (8.352.5)] (for $n = 2, 3, \cdots$)

$$E_n(x) =\frac{(-1)^{n+1}x^{n-1}}{(n-1)!}\left[E_1(x)-e^{-x}\sum_{m=0}^{n-2}\frac{(-1)^m m!}{x^{m+1}}\right] \tag{4.22}$$

by virtue of $\Gamma(0,x) = E_1(x)$ for $\Re(x) > 0$ and obtain

$$\begin{aligned}&\lim_{x\to 0^+} E_n(x)\\ &= \lim_{x\to+0}\left\{\frac{(-1)^{n+1}}{(n-1)!}x^{n-1}E_1(x)\right\}\\ &\quad+\lim_{x\to+0}\frac{(-1)^n e^{-x}}{(n-1)!}\sum_{m=0}^{n-2}x^{n-m-2}(-1)^m m!\\ &\overset{(g)}{=}\frac{(-1)^{n+1}}{(n-1)!}\lim_{x\to 0^+}\left\{\frac{e^{-x}}{(1-n)x^{1-n}}\right\}+\frac{1}{n-1}\\ &\overset{(h)}{=}\frac{1}{n-1}\end{aligned} \tag{4.23}$$

where (g) results from applying the L'Hopital's rule to the first term and the relation $E_1'(x) = -E_0(x) = -e^{-x}/x$. (h) follows also by using the L'Hoptal's rule. Then we

get the limiting value of the third term in (4.18) as

$$\lim_{\rho\to+\infty} S_3 = \lim_{\rho\to+\infty} e^{\frac{1}{\rho}} \sum_{k=2}^{N_p} E_k \left(\frac{1}{\rho}\right)$$
$$= \sum_{k=2}^{N_p} \frac{1}{k-1}. \tag{4.24}$$

Adding the limiting values from (4.20), (4.21), and (4.24) gives the limiting rate loss of (4.17) in *Theorem* 4 after being multiplied with the term $\log_2 e$, which completes the proof. □

Remark 4. *Theorem 4 suggests that in the asymptotic sense, the rate loss of MUBS compared to ZF converges to a constant depending only on the number of paths N_p seen by each user. This N_p relevant constant is essentially small in a typical mmWave propagation settings as will be evidenced in the upcoming numerical results of Section 4.4. On the other hand, the derived rate expressions* (4.8) *and* (4.14) *both indicate that rates achieved by MUBS and ZF precoding schemes grow without limit as the number of transmit antenna N_t increases. Based on these observations, we are now justified to conclude that the proposed MUBS precoding is able to achieve most of ZF capacity in the large array regime, but with substantially reduced hardware complexity, i.e., only K RF chains are needed instead of N_t as required by the full-complexity ZF precoding.*

4.3 Rate Enhancement through A Hybrid Design

So far we have limited, in our proposed scheme, the system to employ RF phase shifters to perform analog processing only. We move a step further in this section by incorporating the limited baseband processing power of the mmWave system. That is, the system will now be capable of performing a hybrid precoding: pure digital processing at the baseband (represented by $\mathbf{F}$ of dimension $L \times L$ where L is the number of RF chains) followed by analog RF processing (represented by $\mathbf{W}$ of dimension $N_t \times L$). The new system block diagram is illustrated in Fig. 4.2. In the following subsections, we discuss two enhancements to the MUBS scheme proposed in Section 4.2 through employing a hybrid structure when more RF chain resources are available.

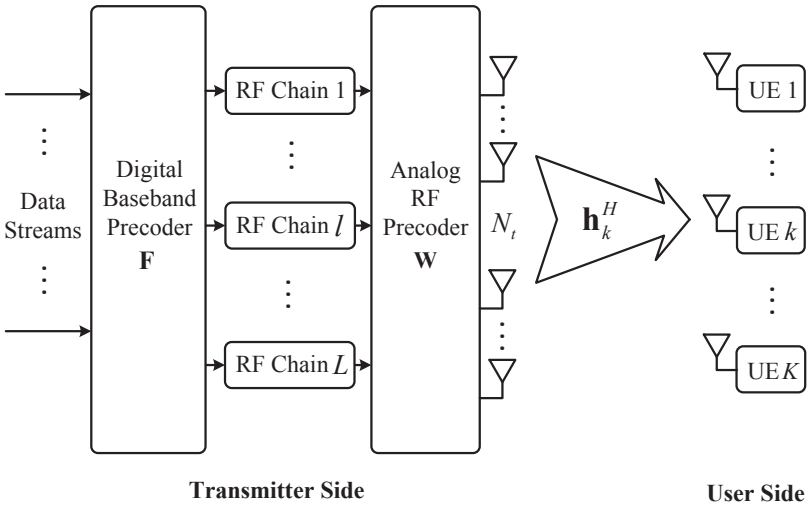

Figure 4.2: Block diagram for a hybrid precoding mmWave MIMO communication system with joint baseband and RF processing.

4.3.1 Multiple Chains-Enabled MUBS

Suppose now we have more RF chains available than users in the system setup, e.g., $L = 2K$ chains. As described in Section 4.2.1, when only limited number of RF chains are available, we want to use them to steer beams for each user towards its channel direction. The strongest path is tracked when exactly K RF chains are mounted on the BS. But in case when more chains are available, we might have the freedom to allocate more chains for each user to better track its channel direction and capture the power. For simplicity and initial attempt, we evenly distribute the RF chains among all users, i.e., 2 chains per user if a total of $2K$ chains are available.

To be more specific, the RF precoder $\mathbf{W}$ is designed according to

$$\mathbf{W}_{M-MUBS}\left[\mathbf{a}\left(\phi_{max}^{1}\right),\mathbf{a}\left(\phi_{max,2}^{1}\right),\cdots,\mathbf{a}\left(\phi_{max}^{K}\right),\mathbf{a}\left(\phi_{max,2}^{K}\right)\right] \tag{4.25}$$

where $\mathbf{a}\left(\phi_{max,2}^{k}\right)$ stands for the unit response vector for the kth user with the second largest magnitude.

Apart from the RF processing, the baseband precoding is also needed to weight the unit response vectors according to the channel breakdown so as to better approximate the channel direction of each user. Along this line, the baseband processing matrix

$\mathbf{F}$ is given by

$$\mathbf{F}_{M-MUBS}\begin{bmatrix} \alpha_{max}^{1} & & 0 \\ \alpha_{max,2}^{1} & & 0 \\ \vdots & \vdots & \vdots \\ 0 & & \alpha_{max}^{K} \\ 0 & & \alpha_{max,2}^{K} \end{bmatrix}\Lambda$$

where α_{max}^{k} denotes the complex gain of the strongest path of the kth user while $\alpha_{max,2}^{k}$ represents the complex gain of the second strongest path. Λ is a diagonal matrix introduced for column power normalization purposes. We name this scheme multiple chains-enabled MUBS (M-MUBS) for simplicity. As shown in Figs. 4.5–4.6, the performance of the hybrid M-MUBS precoding scheme is much closer to the virtually optimal full-complexity ZF precoding.

4.3.2 Multiple Chains-Enabled MUBS with ZF Processing

With the RF precoding matrix $\mathbf{W}$ designed according to (4.25) when $2K$ RF chains are available, more channel power can be captured through RF processing. All the baseband processing matrix $\mathbf{F}$ does is combine the unit response vectors according to their original weights in channel breakdown and normalize the transmit power. We further deduce that if the antenna number N_t is not large enough, e.g., on the order of tens of antennas, we might have the issue of power saturation as the signal-to-noise ratio (SNR) goes. This can be observed from the increasing gap between two chains-enabled MUBS scheme and the ZF precoding in Figs. 4.5–4.6 with growing SNR. Further analysis reveals that this power saturation is due to the residual interuser interference when the orthogonality of users' channels does not accurately hold in the medium-large array regime. One way to completely nullify interuser interference is to introduce ZF precoding at the low-dimensional baseband processing $\mathbf{F}$, which will then ensure desirable performance with our proposed reduced-complexity scheme, hence the name ZF-MUBS. In mathematical representations, the baseband precoding matrix $\mathbf{F}$ is designed according to

$$\mathbf{F}_{ZF-MUBS} = \mathbf{W}_{M-MUBS}^{H}(\mathbf{W}_{M-MUBS}\mathbf{W}_{M-MUBS}^{H})^{-1}\Gamma \tag{4.26}$$

where $\mathbf{W}_{M-MUBS}$ follows from (4.25) and Γ is a diagonal matrix introduced for column power normalization purposes. In Figs. 4.5–4.6, it is shown that the proposed ZF-MUBS scheme performs measurably close to the baseband ZF precoding. But the hardware complexity, especially the number of RF chains, is substantially reduced (from N_t for full-complexity ZF to $2K$ in the proposed ZF-MUBS where $K \ll N_t$).

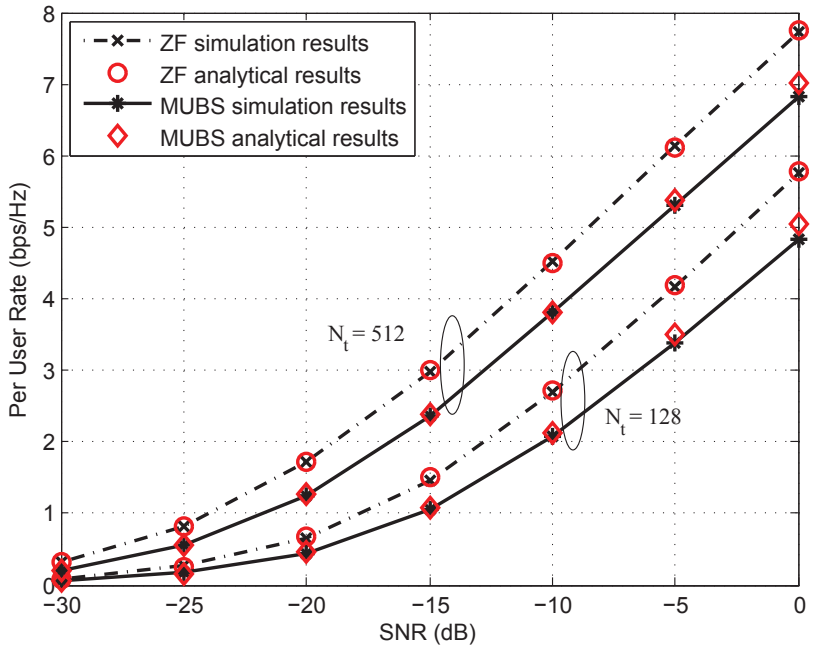

Figure 4.3: Simulated and analytical per-user rates achieved by full-complexity ZF and the proposed MUBS precoding with $K = 2$ and $N_p = 3$.

4.4 Numerical Results

We provide computer simulation results in this section to demonstrate the performance of the proposed MUBS scheme and its enhancements. Throughout the simulations, the channel is generated according to (4.2) where the antenna separation is taken to be half-wavelength and the number of paths per user N_p is assumed to take on 3 to reflect the sparse scattering nature of mmWave propagation [49]. The ULA structure is exploited in the simulation with departure angles ϕ's drawn independently from uniform distribution over $[0, 2\pi]$. All reported simulation results are averaged over 1000 channel realizations.

In Fig. 4.3[3], we plot both analytical and simulation results of the per-user rates

[3]One might notice that we study the performance during extremely low SNR range $(-30, 0)$ dB.

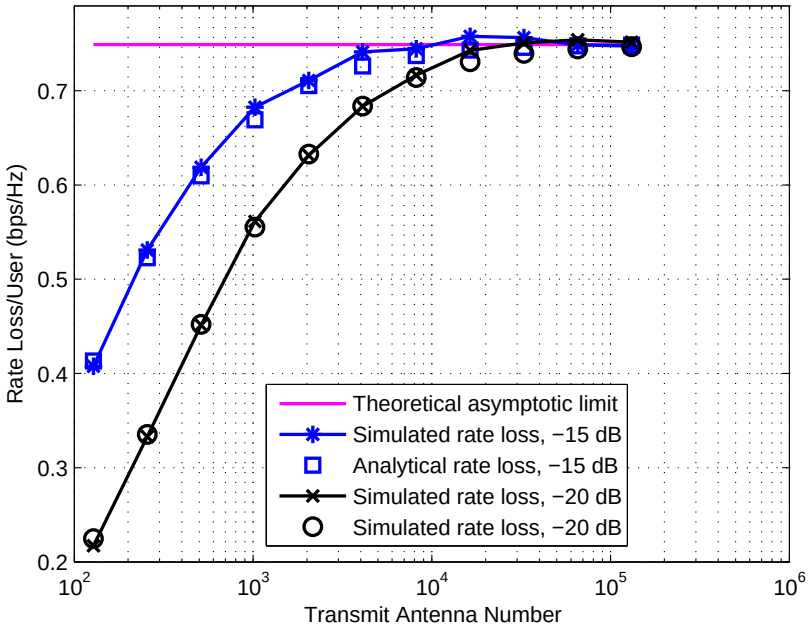

Figure 4.4: Per-user rate loss of MUBS compared against full-complexity ZF precoding from both simulation and analytical results with $K = 2$ and $N_p = 3$.

achieved by the proposed MUBS and full-complexity ZF precoding when the number of users K is 2 and the transmit antenna number N_t equals 128 and 512, respectively. It is observed from Fig. 4.3 that the proposed MUBS scheme achieves most capacity of the desirable yet infeasible ZF precoding with substantially reduced complexity (K chains for MUBS versus N_t chains for ZF) in large mmWave channels. It is also found that the analytical expressions (4.8) and (4.14) characterize the rates of corresponding schemes more accurately for larger N_t.

Fig. 4.4 depicts the performance gap between the proposed MUBS scheme and the full-complexity ZF precoding. The rate loss convergence phenomenon as indicated in (4.17) is demonstrated for various SNR cases both analytically and numerically. We observe from Fig. 4.4 that both the analytical and numerical rate losses converge to a finite limit when N_t goes large and in the case of $N_p = 3$, the asymptotic limit is approximately 0.75 bps/Hz, which depends only on N_p as evidenced from (4.17). This gap convergence guarantees that the proposed MUBS scheme can approach ZF

Here SNR $= P$ is the common average SNR received at each antenna with noise variance normalized to unity. As mmWave systems exploit orders of magnitude higher frequency, the free space path loss is much greater than its low GHz counterparts. Hence, the received power at each antenna in this case is orders of magnitude less than in Rayleigh fading. We thus investigate the performance of our proposed MUBS scheme in the low SNR regime for large multiuser mmWave MIMO communications.

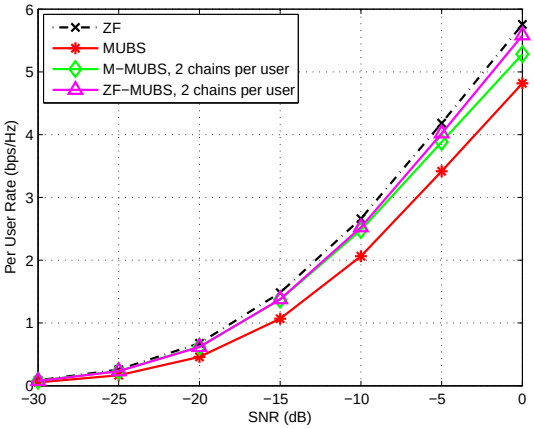

Figure 4.5: Simulated per-user rates for comparing different precoding schemes with $N_t = 128$, $K = 2$ and $N_p = 3$.

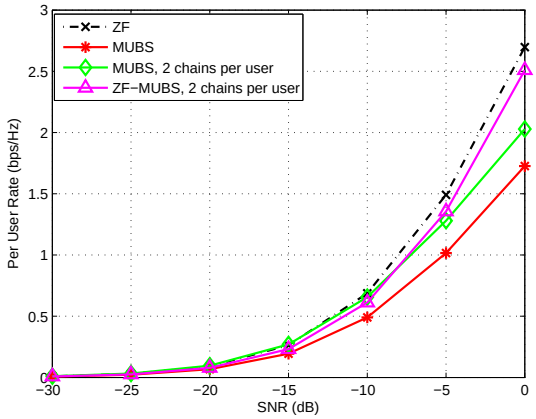

Figure 4.6: Simulated per-user rates for comparing different precoding schemes with $N_t = 128$, $K = 16$ and $N_p = 3$.

precoding in the large array regime as rates achieved by both schemes grow without limit as N_t increases.

In Figs. 4.5–4.6, we present the simulation results of the MUBS scheme and its two enhancements proposed in Section 4.3. It is observed that with 2 chains available for each user, the incorporation of baseband processing in addition to the RF precoding facilitates a massive increase in the rate performance thanks to the sparse scattering of mmWave propagations. In particular, the ZF-enhanced variant of the MUBS scheme, i.e., ZF-MUBS, has the most desirable performance among others as it completely cancels interuser interference at the baseband while capturing more channel power through RF processing. More importantly, all of the three proposed schemes achieve desirable performance close to that of the full-complexity ZF precoding, but with substantially reduced hardware complexity.

4.5 Summary

In this chapter, we studied a large multiuser mmWave system comprising an excess of transmit antennas at the BS but driven by a far smaller number of RF chains. By considering the limited scattering nature of mmWave channels and the asymptotic orthogonality associated with increasing transmit antenna numbers, we proposed to steer the beam for each user towards its strongest path at the RF domain to approach the full-complexity ZF precoding, which is highly desirable but practically infeasible

due to the requirement of supporting each antenna element with a dedicated RF chain. We theoretically showed that the proposed MUBS scheme can approach the performance upper bound imposed by ZF precoding based on asymptotic rate loss analysis. Two enhancements were further proposed based on MUBS, namely, M-MUBS and ZF-MUBS, which would exploit a hybrid structure to help improve rate performance. The findings were finally validated through computer simulations.

Chapter 5

Conclusions and Future Work

This thesis focuses on the precoding design for multiuser MIMO communication systems to achieve high capacity performance under practical constraints. Three currently prevalent multiuser MIMO scenarios are considered. First, given the quantized channel information obtained through limited feedback, a relay-assisted multiuser MIMO system suffers significant rate loss when the transmitter performs linear precoding. An upper bound is derived for the system throughput loss due to channel quantization error, and a feedback quality control strategy is then proposed to maintain a bounded rate loss relative to the perfect channel knowledge scenario. Second, in a massive multiuser MIMO channel, the de-facto pure baseband implementation of virtually optimal zero-forcing precoding is made infeasible due to the large array size. To address this issue, a low-complexity hybrid precoding scheme is proposed where a low-dimensional baseband precoding is preceded by phase-only processing at the radio frequency domain. The desirable performance of the proposed scheme is demonstrated through both theoretical evaluation and computer simulations. Third, in a mmWave multiuser MIMO scenario, where excessive amounts of antennas can be incorporated in a compact form, the hardware complexity issue is even more challenging. However, its distinct characteristic, i.e. channel sparsity, allows a simplified precoding scheme, termed multiuser beam steering, to steer the beam for each user towards its dominant propagation paths at the radio frequency domain only. The proposed scheme is theoretically shown to approach the capacity upper bound imposed by full-complexity zero-forcing precoding based on asymptotic rate loss analysis.

In the multiuser relay system considered in the thesis, it is worthwhile to further consider channel estimation errors in addition to quantization inaccuracy associated with limited feedback in the future work. Besides, for the hybrid precoding schemes

proposed in the thesis for massive multiuser MIMO systems, a joint design of channel estimation and transmit precoding is well worth future efforts as the channel estimation technique for such hybrid systems is far from being well investigated.

Appendix A

Publication List

- L. Liang, W. Xu, and X. Dong, "Limited feedback-based multi-antenna relay broadcast channels with block diagonalization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4092–4101, Aug. 2013.
- L. Liang, W. Xu and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Letters*, vol. 3, no. 6, pp. 653–656, Oct. 2014.
- L. Liang, Y. Dai, W. Xu, and X. Dong, "How to approach zero-forcing under RF chain limitations in large mmWave multiuser systems?" in *Proc. IEEE/CIC Int. Conf. Commun. in China*, Oct. 2014, pp. 518–522.

Bibliography

[1] J. G. Andrews, S. Buzzi, *et al.* "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] W. Yu, *Competition and cooperation in multi-user communication environments*, Ph.D. thesis, Stanford University, Jun. 2002.

[3] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a mutliantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[4] Q. Spencer, L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[5] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna system," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.

[6] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.

[7] N. Ravindran and N. Jindal, "Limited feedback-based block diagonalization for the MIMO broadcast channel," *IEEE J. Select. Areas Commun.*, vol. 26, no. 8, pp. 1473–1482, Oct. 2008.

[8] R. Pabst, *et al*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Commun. Mag.*, vol. 42, no. 9, pp. 80–89, Sep. 2004.

[9] B. Wang, J. Zhang, and A. Host-Madsen, "On the capacity of MIMO relay channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 29–43, Jan. 2005.

[10] C. B. Chae, T. Tang, R. W. Heath, Jr., and S. Cho, "MIMO relaying with linear processing for multiuser transmission in fixed relay networks," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 727–738, Feb. 2008.

[11] Z. Wang, W. Chen, F. Gao, and J. Li, "Capacity performance of relay beamformings for MIMO multirelay networks with imperfect R-D CSI at relays," *IEEE Trans. Veh. Technol.*, vol. 60, no. 6, pp. 2608–2619, Jul. 2011.

[12] B. K. Chalise, L. Vandendorpe, Y. D. Zhang, and M. G. Amin, "Local CSI based selection beamforming for amplify-and-forward MIMO relay networks," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2433–2446, May 2012.

[13] Z. Wang, W. Chen, and J. Li, "Efficient beamforming for MIMO relaying broadcast channel with imperfect channel estimation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 1, pp. 419–426, Jan. 2012.

[14] H. Wan and W. Chen, "Joint source and relay design for multiuser MIMO nonregenerative relay networks with directed links," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2871–2876, Jul. 2012.

[15] W. Xu, X. Dong, and W.-S. Lu, "Joint precoding optimization for multiuser multi-antenna relaying downlinks using quadratic programming," *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1228–1235, May 2011.

[16] Y. Huang, L. Yang, M. Bengtsson, and B. Ottersten, "A limited feedback joint precoding for amplify-and-forward relaying," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1347–1357, Mar. 2010.

[17] Y. Liu and W. Chen, "Limited-feedback-based adaptive power allocation and subcarrier pairing for OFDM DF relay networks with diversity," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2559–2571, Jul. 2012.

[18] B. Khoshnevis, W. Yu, and R. Adve, "Grassmannian beamforming for MIMO amplify-and-forward relaying," *IEEE J. Select. Areas Commun.*, vol. 26, no. 8, pp. 1397–1407, Oct. 2008.

[19] B. Zhang, Z. He, K. Niu, and L. Zhang, "Robust linear beamforming for MIMO relay broadcast channel with limited feedback," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 209–212, Feb. 2010.

[20] W. Xu and X. Dong, "Optimized one-way relaying strategy with outdated CSI quantization for spatial multiplexing," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4458–4464, Aug. 2012.

[21] W. Xu, X. Dong, and W.-S. Lu, "MIMO relaying broadcast channels with linear precoding and quantized channel state information feedback," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5233–5245, Oct. 2010.

[22] T. Yoo, A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[23] W. Dai, Y. Liu, and B. Rider, "Quantization bounds on Grassmann manifolds and applications to MIMO communications," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1108–1123, Mar. 2008.

[24] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etx.: Packings in Grassmannian space," *Exper. Math.*, vol. 5, pp. 139–159, 1996.

[25] N. Jindal, "Antenna combining for the MIMO downlink channel," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3834–3844, Oct. 2008.

[26] R. A. Horn and C. R. Johnson, *Matrix analysis*. New York: Cambridge Univ. Press, 1985.

[27] A. Gupta and D. Nagar, *Matrix variate distributions.*, Chapman & Hall/CRC, 2000.

[28] T. L. Marzetta, "How much training is required for multiuser MIMO?", *Fortieth Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2006.

[29] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

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

[31] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, pp. 186–195, vol. 52, no. 2, Feb. 2014.

[32] H. Q. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2012.

[33] H. Q. Ngo, M. Matthaiou, T. Q. Duong, and E. G. Larsson, "Uplink performance analysis of mulituser MU-SIMO aystems with ZF receivers," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4471–4483, Nov. 2013.

[34] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[35] D. J. Love and R. W. Heath, Jr. "Equal gain transmission in multiple-input multiple-output wireless systems," *IEEE Trans. Commun.*, vol. 51, no. 7, pp. 1102–1110, Jul. 2003.

[36] X. Zheng, Y. Xie, J. Li, and P. Stoica, "MIMO transmit beamforming under uniform elemental power constraint," *IEEE Trans. Sig. Process.*, vol. 55, no. 11, pp. 5395–5406, Nov. 2007.

[37] X. Zhang, A. F. Molisch, and S.Y. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Trans. Sig. Process.*, vol. 53, no. 11, pp. 4091–4103, Nov. 2005.

[38] V. Venkateswaran and A. J. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Trans. Sig. Process.*, vol. 58, no. 8, pp. 4131–4143, Aug. 2010.

[39] W. Roh *et al.*, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[40] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr. "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[41] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE Global Telecommun. Conf. (Globecom)*, Dec. 2013, pp. 3679–3684.

[42] S. Haykin, *Communication Systems*, 4th Ed. New York, NY: John Wiley & Sons, 2001.

[43] J. Choi, V. Raghavan, and D. J. Love, "Limited feedback design for the spatially correlated multi-antenna broadcast channel," in *Proc. IEEE Global Telecommun. Conf. (Globecom)*, Dec. 2013, pp. 3481–3486.

[44] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. Veh. Technol.*, vol. 48, no. 4, pp. 1165–1181, Jul. 1999.

[45] S. K. Yong and C.-C. Chong, "An overview of multigigabit wireless through millimeter wave technology: Potentials and technical challenges," *EURASIP J. Wireless Commun. and Netw.*, vol. 2007, no. 1, pp. 1–10, Jan. 2007.

[46] R. C. Daniels and R. W. Heath, Jr., "60 GHz wireless communications: Emerging requirements and design recommendations," *IEEE Veh. Technol. Mag.*, vol. 2, no. 3, pp. 41–50, Sep. 2007.

[47] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[48] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[49] O. E. Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "The capacity optimality of beam steering in large millimeter wave MIMO systems," in *Proc. IEEE 13th Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Jun. 2012, pp. 100–104.

[50] T. Kim *et al.*, 'Tens of Gbps support with mmWave beamforming systems for next generation communications," in *Proc. IEEE Global Telecommun. Conf. (Globecom)*, Dec. 2013, pp. 3790–3795.

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

[52] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[53] J. Mo and R. Heath, "High SNR capacity of millimeter wave MIMO systems with one-bit quantization," *Proc. Inf. Theory and Applications Workshop (ITA)*, Feb. 2014, pp. 1–5.

[54] A. Gorokhov, D. A. Gore, and A. J. Paulraj, "Receive antenna selection for MIMO spatial multiplexing: theory and algorithms," *IEEE Trans. Signal Process.*, vol. 51, no. 11, pp. 2796–2807, Nov. 2003.

[55] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[56] A. F. Molisch, M. Z. Win, Y.-S. Choi, and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Wireless Commun.*, vol. 4, no. 4, pp. 1759–1772, Jul. 2005

[57] Z. Xu, S. Sfar, and R. S. Blum, "Analysis of MIMO systems with receive antenna selection in spatially correlated Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 251–262, Jan. 2009.

[58] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[59] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60-GHz inddor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, Apr. 2002.

[60] A. Forenza, D. Love, and R. W. Heath, Jr., "Simplified spatial correlation models for clustered MIMO channels with different array configurations," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1924–1934, Jul. 2009.

[61] D. Ramasamy, S. Venkateswaran, and U. Madhow, "Compressive adaption of large steerable arrays," in *Information Theory and Applications Workshop (ITA)*, Feb. 2012, pp. 234–239.

[62] E. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[63] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[64] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?", *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[65] D. Dohono, "Compressed sensing", *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[66] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th Ed. San Diego, CA: Academic Press, 2007.