The Dissertation Committee for Jinseok Choi
certifies that this is the approved version of the following dissertation:

Optimizing Communication Performance of Low-Resolution ADC Systems with Hybrid Beamforming

Committee:

Brian L. Evans, Supervisor

Jeffrey G. Andrews

Ross Baldick

Constantine Caramanis

Hazem Hajj

Optimizing Communication Performance of Low-Resolution ADC Systems with Hybrid Beamforming

by

Jinseok Choi

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2019

Dedicated to my parents.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my research supervisor at the University of Texas at Austin, Professor Brian L. Evans, for his invaluable advice and support for my research projects. His generosity and academic insights have remained the biggest motives throughout my 5-year journey in the Ph.D. program. I consider myself to be tremendously fortunate to have worked with him and as a member of his research group, Embedded Signal Processing Laboratory (ESPL).

I would also like to thank the rest of my dissertation committee members. Professor Jeffery Andrews' broad spectrum of knowledge on wireless communications helped me set the fundamental groundwork for the research area of interest. Optimization theories and skills from Professors Ross Baldick and Constantine Caramanis have provided me with better opportunities to explore and navigate through challenging issues. Professor Hazem Hajj, especially his deep understandings in data mining, has broadened my perspectives to step further to crossing the boundaries between academic fields.

Special thanks to Dr. Alan Gatherer for giving me many inspirations.

Many thanks to my colleagues from ESPL—Junmo Sung, Yunseong Cho, Debarati Kundu, Ghadi Sebaali, Faris Mismar, Hugo Andrade, Scott Johnston—and, more broadly, Wireless Networking Communications Group

(WNCG): Sungwoo Park, Junil Choi, Namyoon Lee, Junse Lee, Gilwon Lee, Changsik Choi, Jeonghun Park, Taewan Kim, among many others. They have continuously inspired my academic envisions that undoubtedly contributed to my research. I extend my thanks to Kyung Woo Min, Wanki Cho, and friends in Korea who have been mentally supportive. It would be hard to imagine my graduate life without my colleagues and friends.

Finally, I would like to express my gratitude to my family: Jiuk Choi, my father, Kyunglim Cho, my mother, and Eun Young Choi, my younger sister. My parents continue to inspire me to be a better person and to share what I have with others. I am grateful for their unconditional love and support throughout my life. And, most importantly, to my beloved half, who has entered my life in the mid-stages of my graduate career. From the bottom of my heart, I am indebted to her for her wise support in maintaining balance and discovering true happiness in life.

Jinseok Choi

Optimizing Communication Performance of Low-Resolution ADC Systems with Hybrid Beamforming

Publication No. _____

Jinseok Choi, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Brian L. Evans

Low-resolution analog-to-digital converter (ADC) systems and hybrid analog-and-digital beamforming systems have drawn extensive attention as a promising receiver architecture for millimeter wave (mmWave) communications by reducing hardware cost and power consumption. In this dissertation, hybrid beamforming systems that employ low-resolution ADCs are considered to achieve a better trade-off between communication performance and power consumption. Due to non-negligible quantization errors, however, existing state-of-the-art hybrid beamforming techniques cannot be directly applied to such systems as they ignore the impact of the quantization error. In this regard, I propose new receiver architectures and algorithms for hybrid beamforming with low-resolution ADC systems to enhance spectral efficiency under coarse quantization in different layers of the network stack, and provide subsequent analyses.

First, problems of optimizing the number of ADC bits and designing analog combiners with fixed-resolution ADCs are tackled to design an energy-efficient receiver architecture with phase shifter-based hybrid beamforming. A hybrid receiver architecture with resolution-adaptive ADCs for mmWave communications is proposed to optimize the power distribution over ADCs. For the proposed architecture, a near-optimal bit-allocation solution is derived in closed form. In addition, the performance lower bound of the proposed receiver architecture is derived in ergodic rate. For a fixed-resolution ADC system, a new analog combining architecture is proposed for mmWave communications. The proposed analog combiner consists of two consecutive analog combiners that maximize channel gain and minimize effective quantization error. An approximated ergodic rate of the proposed receiver is also derived in closed form. Next, considering switch-based analog beamforming, antenna selection at a base station is investigated for low-resolution ADC systems. Unlike downlink transmit antenna selection problems, a quantization-aware antenna selection criterion is necessary and derived to incorporate quantization error for uplink receive antenna selection problems. Leveraging the criterion, a quantization-aware antenna selection algorithm is proposed and analyzed for uplink. Last, in a higher layer of the network stack, a user scheduling problem is investigated for hybrid beamforming systems with low-resolution ADCs. New user scheduling criteria are derived to maximize scheduling gain under coarse quantization and efficient scheduling algorithms are proposed accordingly. Subsequent analysis for the proposed algorithm provides closed-form ergodic rates.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ADC	analog-to-digital converter
AoA	angle of arrival
AQNM	additive quantization noise model
ARV	array response vector
AWGN	additive white Gaussian noise
BA	bit allocation
BS	base station
CQI	channel quality indicator
CSI	channel state information
CSS	channel structure-based scheduling
DL	downlink
DFT	discrete Fourier transform
FAS	fast antenna selection
FFT	fast Fourier Transform
GMI	generalized mutual information
IID	independent and identically distributed
KKT	Karush-Kuhn-Tucker
LOS	line of sight
MCMC	Markov chain Monte Carlo
MI	mutual information
MIMO	multiple input multiple output
MIS	metropolized independent sampler

MMSE	minimum mean squared error
MMSQE	minimum mean squared quantization error
mmWave	millimeter wave
MRC	maximum ratio combining
MS	mobile station
MSE	mean squared error
MSQE	mean squared quantization error
MU-MIMO	multiuser multiple input multiple output
NBS	norm-based selection
OFDM	orthogonal frequency division multiplexing
OMP	orthogonal matching pursuit
QFAS	quantization-aware antenna selection
QMCMC	quantization-aware Markov chain Monte Carlo
RBF	random beamforming
RF	radio frequency
SINR	signal-to-interference-plus-noise ratio
SNR	signal-to-noise ratio
SUS	semi-orthogonal user scheduling
SVD	singular value decomposition
TSAC	two-stage analog combining
UE	user equipment
UL	uplink
ULA	uniform linear array
ZF	zero-forcing

Chapter 1

Introduction

This introductory chapter briefly overviews the background and motivation in this dissertation, followed by the brief summary of expected contributions. Section 1.1 presents the background regarding the systems with a large number of antennas and power-efficient system designs such as low-resolution analog-to-digital converter (ADC) and hybrid analog-and-digital beamforming architectures. Section 1.2 provides the motivation of the proposed research. Section 1.3 summarizes the contributions of the proposed research. The notations and abbreviations are summarized in Section 1.4.

1.1 Background

1.1.1 Wireless Communication Systems

Cellular networks are composed of a large number of users who use cellular devices such as mobile phones and tablets and a large number of base stations (BSs) that are fixed and arranged to provide coverage to the users. The physical area that a BS covers is called a cell. Mobile users in each cell are connected with an associated BS. Since a BS cannot in general serve all of the users in the cell, user scheduling is necessary to select users to serve

by maximizing the cell throughput while maintaining fairness among users. The wireless link from a BS to mobile users is called downlink, and the BS transmits data to the users on the downlink. On the other hand, the wireless link from the mobile users to the BS is called uplink, and the users transmit data to the BS on the uplink.

Unlike wireline communications, fading and interference are the two key impairments of wireless communications, which makes the problem even more challenging. Fading is the time variation of channel strengths that is induced by the small-scale effect of multi-path fading and the large-scale effects known as path loss and shadowing. Being different from thermal noise, interference is generated by other signals. The different delays on the multiple paths from the transmitter to the receiver cause interference at the receiver for subsequent transmissions, which is known as inter-symbol-interference. When multiple users communicate with the BS in the same time and frequency resource (co-channel), there is significant interference between them, which is called inter-user-interference. In the multi-cell environment, the incoming signals from other cells are interfering with the co-channel signals of the associated cell, and it is called inter-cell-interference. How to deal with such interference is one of the most important issues in the design of wireless communications.

When the channel is in deep fade, i.e., the channel strength is very low, it is almost impossible to achieve reliable communications. Many diversity techniques have been developed to overcome such problem. There are many ways to obtain diversity. Via coding and interleaving, diversity can be obtained

over time since the coded symbols are transmitted over time so that different parts of the codeword experience different fading channels. If a channel is frequency selective, similar diversity can be obtained over frequency. When multiple antennas are spaced sufficiently at the transmitter and/or receivers, diversity can also be achieved over space.

In addition to the diversity techniques, many interference mitigation techniques have been developed to deal with several kinds of interference. Linear equalizers such as maximum ratio combining, zero-forcing combining, and minimum mean squared error combining and nonlinear equalizers are widely used, and they can be applied over time, frequency, and space. Multiple access techniques such as code-division multiple access and orthogonal frequency-division multiple access were also developed to serve multiple users without interfering with each other. Cell sectorization is used to reduce interference among co-channel cell. Sectorization divides each cell spatially by employing directional antennas at the BS and provides substantial reduction of interference without requiring the acquisition of new BS sites.

As discussed above, using multiple antennas offers diversity gain. In addition to the diversity gain, it also provides power gain when a receiver is equipped with multiple antennas or a transmitter equipped with multiple antennas knows the channel state information. Having both multiple transmit and receive antennas, which is known as a multiple-input multiple-output (MIMO) system, gives a new way to use multiple antennas. The MIMO systems provide an additional spatial dimension and yields a degree-of-freedom

gain which can be exploited by spatially multiplexing several data streams onto the MIMO channel [1]. This leads to an increase in the channel capacity that is proportional to the degree-of-freedom. Thus, MIMO techniques have been the primary tool in the wireless communications to increase both the capacity and reliability.

Recently, using a large number of antennas at the BS has been widely investigated. The extra antennas can dramatically increase the capacity and improve the radiated energy efficiency by taking aggressive multiplexing and focusing energy into ever smaller regions of space [2]. It is also known that the multiple access layer can be simplified with the use of a large number of antennas [3]. Millimeter wave (mmWave) communication that operates at very high frequencies is likely to employ a large number of antennas to overcome its large path loss by accomplishing large beamforming gain. Since huge bandwidth available at mmWave frequencies can realize the rates of multiple gigabits per second per uses, mmWave communications have been considered as a potential future wireless communication technology to meet ever increasing demand for data rate.

1.1.2 Millimeter Wave Communications

Moving to a millimeter wave (mmWave) spectrum in range of 30–300 GHz enables the utilization of multi-gigahertz bandwidth and offers an order of magnitude increase in achievable rate [4–6]. Consequently, mmWave communication has drawn extensive attention as a promising technology for

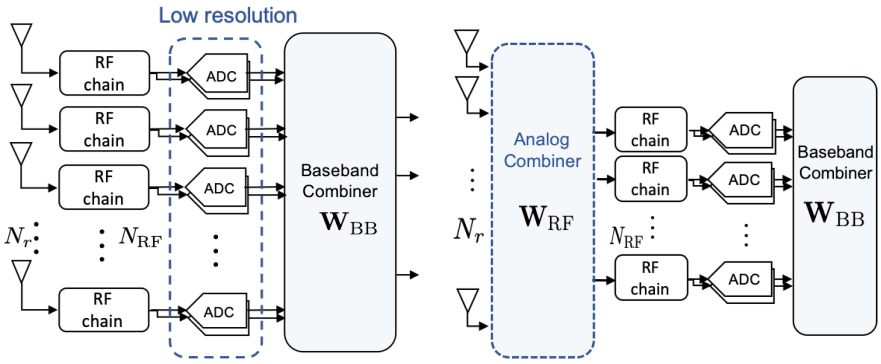


Figure 1.1: (a) A low-resolution ADC receiver and (b) hybrid beamforming receiver.

next-generation cellular systems [7–9], and evinced its feasibility [10]. Unlike the traditional MIMO communication that operates sub-3 GHz with a small number of antennas, the small wavelength of the mmWave spectrum allows a large number of antennas to be packed into transceivers with very small antenna spacing. Leveraging the large antenna arrays, mmWave systems can manipulate directional beamforming to produce high beamforming gain, which helps overcome large free-space pathloss of mmWave signals and maintains a reasonable level of received signal-to-noise ratio (SNR).

Problems with hardware cost and power consumption, however, arise from deploying large antenna arrays. Due to the large number of radio frequency (RF) chains and power-demanding high-resolution ADCs coupled with high sampling rates, the significant power consumption at the receivers becomes one of the primary challenges to resolve. To overcome these challenges, receivers that employ low-resolution ADCs [11] to dramatically reduce the

power consumption at the ADCs and hybrid analog-and-digital beamforming architectures [12] that attempt to reduce the burden of fully digital beamforming have attracted the most interest in recent years as shown in Fig. 1.1.

1.1.3 Low-Resolution ADC Systems

The power consumption of ADCs, P_{ADC}, scales exponentially in the number of quantization bits b, i.e., $P_{\mathrm{ADC}} \propto 2^b$ [13], leading high-speed and high-resolution ADCs to be the primary power consumers in the receiver with large antenna arrays. Although deploying low-resolution ADCs in mmWave communication systems with large antenna arrays greatly reduces power consumption at receivers, non-negligible quantization error due to coarse quantization degrades the performance of such system. Furthermore, the increased quantization error prevents the existing state-of-the-art multiple-input multiple-output (MIMO) techniques from achieving desirable performance.

As an effort to realize low-resolution ADC systems, essential wireless communication techniques such as channel estimation and detection have been developed in low-resolution ADC systems [14–19]. For the 1-bit ADC system which is the extreme case of low-resolution ADCs, compressive sensing [14], maximum-likelihood [15], and Bussgang decomposition-based techniques [16] were employed for channel estimation. Compressive sensing-based channel estimators were also developed for the systems with low-resolution ADCs [17], and achieved comparable estimation accuracy to that of infinite-bit ADC systems at low and medium signal-to-noise ratio (SNR). Unified frameworks for

channel estimation and symbol detection were developed for 1-bit ADC systems [15] and low-resolution ADC systems [17]. Achieving higher detection accuracy than a minimum mean squared error (MMSE) estimator, message passing de-quantization-based detectors were proposed in 1-bit ADC [18] and low-resolution ADC systems [19]. For mmWave channels, the main consideration in this dissertation, a generalized approximate message-passing (GAMP) algorithm with 1-bit ADCs showed a similar channel estimation performance as maximum-likelihood (ML) estimator with full-resolution ADCs in the low and medium SNR regimes [14] by exploiting the sparsity of mmWave channels in the angular domain. It was further proved in [20] that accurate estimation is also possible in wideband mmWave channel estimation by combining GAMP with the expectation-maximization algorithm.

1.1.4 Hybrid Beamforming Systems

In another line of research, hybrid beamforming architectures employ an analog beamformer to reduce the number of RF chains less than the number of antennas to reduce power consumption and system complexity [21, 22]. phase shifter-based analog beamforming and switch-based analog beamforming are often considered for analog beamformer networks [23] by offering different benefits and limitations. When the system uses the set of phase shifters for analog beamforming [24, 25], the design of an analog precoder and combiner is limited by its constant amplitude [21], which leads to separate analog and digital beamformer design.

State-of-the-art hybrid beamforming methods have been proposed with the goal of achieving spectral efficiency close to that of the system with fully digital beamformers [21, 26–31]. In [31], it was shown that the number of RF chains are required to be at least twice the number of data streams to realize the performance of fully digital beamforming. An analog beamformer is often designed by selecting array response vectors corresponding to the dominant channel eigenmodes [21, 26–30]. Indeed, it was shown that the optimal RF precoder and combiner converge to array response vectors in dominant eigenmodes [26]. Motivated by this, orthogonal matching pursuit (OMP) was used to develop beamformer design algorithms [21, 27, 28], which composes RF beamformer with the array response vectors by estimating the dominant eigenmodes. In addition, low-complexity hybrid precoding algorithms in multi-user MIMO downlink systems were proposed by considering zero-forcing precoding [29] and limited feedback [30]. When the system uses a switch network for analog beamforming, an analog beamforming problem becomes equivalent to an antenna selection problem. Although adopting the switch network keeps the system from using highly effective beamforming techniques, it requires much less hardware cost and complexity compared to the analog processing with phase shifters [23].

1.2 Motivation

1.2.1 Hybrid Beamforming with Low-Resolution ADCs

Previous studies consider two major architectures: hybrid analog-and-digital beamforming and low-resolution ADC systems. The former employs analog beamforming to decrease the number of RF chains to be less than that of antennas, thereby mitigating the burden on digital beamforming, and the latter adopts a small number of quantization bits to reduce ADC power consumption. In other words, the hybrid beamforming receivers consider a small number of RF chains with full-resolution ADCs, while the low-resolution ADC receivers assume no reduction on the number of RF chains. There is prior work [32] that studied a general version of these two extreme points: hybrid architecture with low-resolution ADCs as shown in Fig. 1.2. In [32], the spectral efficiency was analyzed under a constant channel assumption, and it was shown that hybrid architecture with low-resolution ADCs achieves high energy efficiency. In this dissertation, I consider the hybrid architecture with low-resolution ADCs to achieve the best trade-off between the performance and power consumption [32]. Then, I develop advanced receiver designs and algorithms that enhance spectral efficiency under coarse quantization for the considered system in different layers of the network stack, and further provide subsequent analyses. In the following three chapters, I focus on developing novel receiver architectures to incorporate the effect of coarse quantization in the receiver design. In the last two chapters, I investigate user scheduling problems to provide new scheduling criteria under coarse quantization, and

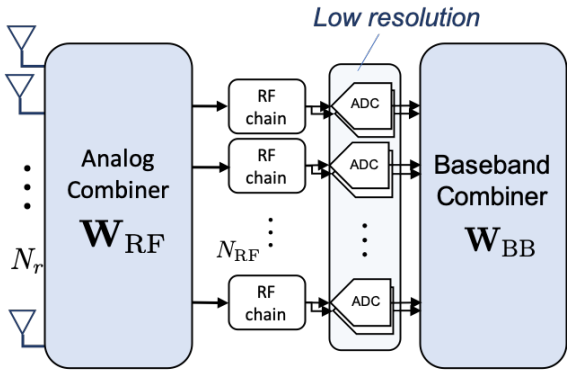


Figure 1.2: A hybrid beamforming receiver with low-resolution ADCs.

then I provide the summary of the dissertation and possible future research directions.

Adopting uplink hybrid analog-and-digital beamforming architecture, I propose a new receiver design with resolution-adaptive ADCs for mmWave communications. Although previous works consider hybrid beamforming architectures with low-resolution ADCs [32] or mixed-ADC architectures [33,34], they either assume predetermined ADC resolutions regardless of channel gain on each RF chain or force antennas to select between 1-bit ADC and ∞-bit ADC, which is far from an energy-efficient architecture. In this regard, employing resolution-adaptive ADCs at the hybrid receiver can provide significant flexibility in distributing energy over the ADCs with different channel gains, which leads to highly energy-efficient receiver architectures. To this end, I solve ADC bit-allocation problems to find an optimal bit distribution that minimizes total quantization error subject to limited power consumption, and further show its relevance to a generalized mutual information. To pro-

vide a performance lower bound of the proposed architecture, I also derive an approximated ergodic rate in closed form.

Moving the focus onto the analog beamformer design rather than the ADC design, I investigate an advanced hybrid combining technique for large-scale MIMO receivers with low-resolution ADCs. Conventional hybrid combiners were limited to high-resolution ADC cases and thus, a new hybrid combining architecture is required to achieve optimality in communication performance for hybrid MIMO systems with low-resolution ADCs. Although the analysis in [32, 35, 36] provided useful insights for the hybrid architecture with low-resolution ADCs such as the achievable rate and power trade-off, the quantization error was not explicitly taken into account in the hybrid beamformer design. Thus, I propose a new analog combining architecture and develop a two-stage analog combining algorithm which effectively reduces quantization error while maintaining large channel gains in the reduced signal dimension.

The resolution-adaptive ADC and new analog combining architectures consider phase shifter-based analog beamforming architectures. Avoiding the burden of implementing large phase shifter arrays for hybrid MIMO systems, employing switch-based analog beamforming is another power-efficient solution. In this regard, I also investigate antenna selection problems for systems with low-resolution ADCs, thereby providing more flexibility in resolution and number of ADCs without the necessity for implementing large phase-shifter arrays. Indeed, for channels measured at 2.6 GHz, a great number of RF chains could be turned off by using antenna selection without a substantial

performance loss [37]. Previously proposed antenna selection methods [38–43], however, focused on MIMO systems without any quantization errors. Consequently, for low-resolution ADC receivers, a new antenna selection method that incorporates coarse quantization effect needs to be developed. Therefore, I develop a quantization-aware antenna selection algorithm for uplink communications and also study the antenna selection problem for downlink communications.

Focusing on the higher layer of the network stack than the receiver design, user scheduling problems are investigated in a single cell environment and user scheduling algorithms are developed for hybrid receivers with low-resolution ADCs. Many user scheduling methods were developed under the no quantization system in which the number of quantization bits is considered to be infinite [44–49]. The quantization error from employing low-resolution ADCs, however, is a function of user channels and non-negligible. Existing user scheduling criteria mostly focus on channel orthogonality and channel amplitude, which does not incorporate the increase of the quantization error when scheduling users. Accordingly, the proposed user scheduling algorithm in this dissertation exploits new findings that are effective under coarse quantization.

1.3 Dissertation Summary

To summarize, I have contributed to advanced receiver designs and algorithm development for hybrid beamforming systems with low-resolution ADCs to improve their communication performance.

1.3.1 Thesis Statement

In this dissertation, I defend the following statement:

Advanced mixed-domain signal processing techniques can unlock gamechanging system-level tradeoffs in communication performance vs. power consumption in millimeter wave cellular base station designs.

1.3.2 Overview of Contributions

The main contributions of this dissertation are summarized as follows:

1. **Bit Allocation for Hybrid Receivers with Resolution-Adaptive ADCs:** A new hybrid receiver architecture with resolution-adaptive ADCs for mmWave communications is proposed to achieve power-efficient communications. A near-optimal bit-allocation solution that minimizes the total mean squared quantization error is derived in closed form. Exploiting the solution, a bit-allocation algorithm is developed for a total ADC power constraint case, outperforming conventional low-resolution ADC receivers in both in spectral and energy efficiencies. Finally, a closed-form performance lower bound of the proposed receiver architecture is derived in ergodic rate when the receiver employs maximum ratio combining (MRC).

2. **Two-Stage Analog Combining for Low-Resolution ADC Systems:** A new hybrid receiver architecture with low-resolution ADC is proposed for mmWave communications by splitting the analog combiner into two consecutive analog combiners. The main function of the first analog combiner is

to collect most channel gains into the lower dimension. The second analog combiner focuses on reducing quantization errors from the low-resolution ADCs by evenly spreading the collected signal over all available RF chains. It is shown that the proposed two-stage analog combiner achieves an optimal scaling law of the channel capacity with respect to the number of RF chains under the presence of quantization error and maximizes the capacity for channels with homogeneous singular values. An approximated ergodic rate with MRC is derived in closed form, showing that it also achieves the optimal scaling law.

3. **Base Station Antenna Selection for Low-Resolution ADC Systems:** Antenna selection at a base station with large antenna arrays and low-resolution ADCs is investigated for both downlink and uplink. For downlink transmit antenna selection, it is shown that although a selection criterion that maximizes sum rate with ZF precoding is equivalent to that of a perfect quantization system, sum rate loss decreases to zero as total transmit power increases unlike the perfect quantization system. For uplink receive antenna selection, a greedy antenna selection criterion is generalized to capture trade-offs between channel gain and quantization error. Leveraging the criterion, a quantization-aware fast antenna selection algorithm is developed and analyzed.

4. **Uplink User Scheduling for Hybrid Receivers with Low-Resolution ADCs:** Channel structure-based user scheduling criteria are derived to

maximize scheduling gain for low-resolution ADC systems. Using the derived criteria, user scheduling algorithms that maximizes the uplink sum rate in the low-resolution ADC system for full and partial channel state information are developed, showing improvement in communication performance compared to the conventional scheduling methods. Subsequent analysis for the proposed algorithm provides closed-form ergodic rates for different channel scenarios.

The main takeaway messages of the dissertation are summarized as:

- Employing low-resolution ADCs is one of the solutions to reduce power consumption of receivers with a large number of antennas.
- Non-negligible quantization error requires conventional wireless techniques to be modified to improve communication performance.
- Hybrid beamforming techniques that further decrease power consumption by reducing the number of RF chains need to consider the quantization error in their design.
- User scheduling also requires additional scheduling criteria to incorporate the effect of the coarse quantization when scheduling users.
- Adopting variable-resolution ADCs can be the other form of the low-resolution ADC receivers that achieves higher energy-efficient systems.

I believe that the contributions and key findings will pave the way for future wireless communication systems.

1.4 Notation and Abbreviations

This dissertation uses the following notation: $\mathbf{A}$ is a matrix and $\mathbf{a}$ is a column vector. $\mathbf{A}^H$ and $\mathbf{A}^T$ denote conjugate transpose and transpose. $[\mathbf{A}]_{i,:}$ and $\mathbf{a}_i$ indicate the ith row and column vector of $\mathbf{A}$. We denote $a_{i,j}$ or $[\mathbf{A}]_{i,j}$ as the $\{i,j\}$th element of $\mathbf{A}$ and a_i as the ith element of $\mathbf{a}$. $\lambda_i\{\mathbf{A}\}$ denotes the i-th largest singular value of $\mathbf{A}$. $\mathcal{CN}(\mu, \sigma^2)$ is the complex Gaussian distribution with mean μ and variance σ^2. $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ represent an expectation and variance operators, respectively. The correlation matrix is denoted as $\mathbf{R}_{\mathbf{xy}} = \mathbb{E}[\mathbf{x}\mathbf{y}^H]$. The diagonal matrix diag$\{\mathbf{A}\}$ has $\{a_{i,i}\}$ at its ith diagonal entry, and diag$\{\mathbf{a}\}$ or diag$\{\mathbf{a}^T\}$ has $\{a_i\}$ at its ith diagonal entry. BlkDiag$\{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$ is a block diagonal matrix with block diagonal entries $\mathbf{A}_1, \cdots, \mathbf{A}_N$. BlkCirc$\{\mathbf{A}_0, \mathbf{A}_1, \cdots, \mathbf{A}_N\}$ is a block circulant matrix with $[\mathbf{A}_0, \mathbf{A}_1, \cdots, \mathbf{A}_N]$ at its first block row.$\mathbf{I}$ denotes the identity matrix with a proper dimension and we indicate the dimension N by $\mathbf{I}_N$ if necessary. $\mathbf{0}$ denotes a matrix that has all zeros in its elements with a proper dimension. $\|\mathbf{A}\|$ represents L_2 norm. $|\cdot|$ indicates an absolute value, cardinality, and determinant for a scalar value a, a set $\mathcal{A}$, and a matrix $\mathbf{A}$, respectively. Tr$\{\cdot\}$ is a trace operator and $x(N) \sim y(N)$ means $\lim_{N\to\infty} \frac{x}{y} = 1$.

Chapter 2

Resolution-Adaptive Hybrid MIMO Architectures for Millimeter Wave Communications

In this chapter[1], a hybrid analog-digital beamforming architecture with resolution-adaptive ADCs is proposed for millimeter wave (mmWave) receivers with large antenna arrays. Array response vectors are adopted for the analog combiners and derive ADC bit-allocation (BA) solutions in closed form. The BA solutions reveal that the optimal number of ADC bits is logarithmically proportional to the RF chain's signal-to-noise ratio raised to the 1/3 power. Using the solutions, two proposed BA algorithms minimize the mean square quantization error of received analog signals under a total ADC power constraint. Contributions include 1) ADC bit-allocation algorithms to improve communication performance of a hybrid MIMO receiver, 2) approximation

[1]This chapter is based on the work published in the journal paper: J. Choi, B. L. Evans, and A. Gatherer, "Resolution-Adaptive Hybrid MIMO Architectures for Millimeter Wave Communications," in *IEEE Transactions on Signal Processing*, vol. 65, no. 23, pp. 6201-6216, Dec. 2017. Part of the work was also published in the conference paper: J. Choi, B. L. Evans, and A. Gatherer, "ADC Bit Allocation under a Power Constraint for MmWave Massive MIMO Communication Receivers," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 5-9, 2017, New Orleans, LA, USA. This work was supervised by Prof. Brian L. Evans. The useful feedback from Dr. Alan Gatherer improved the quality of the work.

of the capacity with the BA algorithm as a function of channels, and 3) a worst-case analysis of the ergodic rate of the proposed MIMO receiver that quantifies system tradeoffs and serves as the lower bound. Simulation results demonstrate that the BA algorithms outperform a fixed-ADC approach in both spectral and energy efficiency, and validate the capacity and ergodic rate formula. For a power constraint equivalent to that of fixed 4-bit ADCs, the revised BA algorithm makes the quantization error negligible while achieving 22% better energy efficiency. Having negligible quantization error allows existing state-of-the-art digital beamformers to be readily applied to the proposed system.

2.1 Introduction

In this chapter, I first investigate an advanced receiver design for phase shifter-based hybrid beamforming with low-resolution ADCs. Hybrid architectures employ fewer RF chains than the number of antennas to reduce power consumption and system complexity. An analog beamformer is the pivotal component that enables the hybrid structure to reduce the number of RF chains [21, 22]. An analog beamformer is often designed by selecting array response vectors corresponding to the dominant channel eigenmodes [21, 26–30]. Indeed, it was shown that the optimal RF precoder and combiner converge to array response vectors in dominant eigenmodes [26]. Motivated by this, orthogonal matching pursuit (OMP) was used to develop beamformer design algorithms [21, 27, 28]. Although the hybrid beamforming approaches in

[21,22,24–30] delivered remarkable achievements in the development of the low-power and low-complexity architecture with large antenna arrays, the hybrid architectures still assume high-resolution ADCs that consume a high power at receivers.

Since power consumption of ADCs scales exponentially in terms of the number of quantization bits [50], employing low-resolution ADCs can be indispensable to reduce hardware cost and power consumption in the large antenna array regime. Consequently, low-resolution ADC architectures have been investigated [14,17,18,20,51–58]. It was revealed that least-squares channel estimation and maximum-ratio combining (MRC) with 1-bit ADCs are sufficient to support multi-user operation with quadrature-phase-shift-keying [52], which is known to be optimal for 1-bit ADC systems [11, 51]. Deploying large antenna arrays provided an opportunity to use message-passing and expectation-maximization algorithms for symbol detection and channel estimation with low complexity [14, 17, 18, 20]. To examine the effect of quantization in achievable rate, the Bussgang decomposition [54, 55] was utilized for linear expressions of quantization operation. The analysis in [54] revealed that noise correlation can reduce the capacity loss to less than $\frac{2}{\pi}$ at low signal-to-noise ratio (SNR). A lower bound for the achievable rate of the 1-bit ADC massive MIMO system was derived [55], using MRC detection with a linear minimum mean square error (MMSE) channel estimator. Offering an analytical tractability, the additive quantization noise model (AQNM) [56–59] were adopted to derive the achievable rate of massive MIMO systems with low-resolution ADCs using

MRC in Rayleigh [57] and Rician fading channels [58].

The considered architectures in the previous studies, however, present two extreme points: (1) fewer number of RF chains with high-resolution ADCs and (2) low-resolution ADCs with full number of RF chains. One prior study with less extremity [60] focused on a generalized system consisting of fewer number of RF chains with low-resolution ADCs. In [60], the spectral efficiency was analyzed under a constant channel assumption. It is also assumed that each ADC's resolution is predetermined regardless of channel gain on each RF chain. In another line of research, mixed-ADC architectures were proposed [33, 34, 59]. In [59], performance analysis of mixed-ADC systems where receivers use a combination of low-resolution and high-resolution ADCs showed that the architecture can achieve a better energy-rate tradeoff compared to systems either with infinite-resolution ADCs or low-resolution ADCs. In [33, 34] each antenna uses different ADC resolution depending on its channel gain. This system has explicit benefits compared to fixed low-resolution ADC systems such as increase of channel estimation accuracy and spectral efficiency. In [33, 34], however, they force antennas to select between 1-bit ADC and ∞-bit ADC, which is far from an energy-efficient architecture mainly because the total ADC power consumption can be dominated by only a few high-resolution ADCs. Moreover, it assumes full number of RF chains, which leads to dissipation of energy. For these reasons, an adaptive ADC design for a hybrid beamforming architecture is still questionable.

2.1.1 Contributions

The main contribution of this chapter is the proposition of a hybrid beamforming MIMO architecture with resolution-adaptive ADCs to offer a potential energy-efficient mmWave receiver architecture. Under this architecture proposition, I investigate the architecture as follows: (i) two bit-allocation (BA) algorithms are first developed to exploit the flexible ADC architecture and derive a capacity expression for a given channel realization. (ii) Due to the intractable ergodic rate analysis with BA, I then perform the analysis without BA, offering the baseline performance of the proposed receiver architecture. The proposed architecture is distinguishable from many other systems because it not only consists of a lower number RF chains and low-resolution ADCs [60] but also adapts the ADCs resolutions [33,34]. In the context of mmWave communications, I design the analog combiner to be a set of array response vectors to aggregate channel gains in the angular domain. Such design approach is beneficial as the sparse nature of mmWave channels in the angular domain allows the number of RF chains to be less than the number of antennas. Leaving the design issue of digital combiners, this chapter primarily focuses on the quantization problem for the proposed system.

Given the different channel gains on RF chains, the system performance can be improved by leveraging the flexible ADC architecture. To this end, as an extension of the work [61], I derive a closed-form BA solution for a minimum mean square quantization error (MMSQE) problem subject to a constraint on the total ADC power. Using the solution, a BA algorithm is

developed, and it determines ADC resolutions depending on angular domain channel gains. The derived solution provides an explicit relationship between the number of quantization bits and channel environment. One major finding from the solution is that the optimal number of ADC bits is logarithmically proportional to the corresponding RF chain's SNR raised to the 1/3 power. This result quantifies the conclusion made in [33] that allocating more bits to the RF chain with stronger channel gain is beneficial. I also derive a solution for a revised MMSQE problem to modify the proposed BA method to be robust to noise. The revised MMSQE problem is equivalent to maximizing generalized mutual information (GMI) in the low SNR regime. Applying the solution to a capacity, I approximate the capacity with the revised MMSQE-BA algorithm as a function of channels. Simulation results disclose that the BA algorithms achieve a higher capacity and sum rate than the conventional fixed-ADC system where all ADCs have same resolution. In particular, the revised BA algorithm provides the sum rate close to the infinite-resolution ADC system while achieving higher energy efficiency than using fixed ADCs in the low-resolution regime.

Regarding the implementation issue of the BA algorithms, the best scenario is to operate the resolution switching at the time-scale of the channel coherence time. This is because the proposed BA algorithms allocate different quantization bits to each ADC depending on the channel gain on each RF chain. Accordingly, if the switch is able to operate at the channel coherence time, the proposed architecture is able to adapt to channel fluctuations. Such

coherence time switching in mmWave channels, however, may not be feasible due to the very short coherence time of mmWave channels [62]. Consequently, the switching period may need to be the multiples of the coherence time. In this case, switching at the time-scale of slowly changing channel characteristics such as large-scale fading and angle of arrival (AoA) marginally degrades the performance of the BA algorithms. Then, the worst-case scenario is not to exploit the flexibility of ADC resolutions, which is equivalent to have an infinitely long switching period, and indeed converges to fixed-ADC architectures.

To provide deeper insight for the proposed system, I further perform an ergodic rate analysis. As mentioned, due to the intractability of the analysis with the BA algorithms, we derive an approximation of the ergodic rate for the considered architecture without applying BA—the worst-case analysis—for analytical tractability. Although the analysis focuses on the worst-case scenario, the importance of the derived rate can be given as follows:

- The obtained achievable rate can serve as the lower bound of the proposed architecture. Hence, it is expected that the proposed system can achieve a higher ergodic rate than the derived rate by leveraging the flexible ADCs.
- As a function of system parameters, the tractable rate provides a broad insight for the considered system. It can be shown that the achievable rates for the BA algorithms and for the fixed ADC show similar trends. In this regard, the derived rate provides general tradeoffs of the proposed architecture in terms of system parameters including quantization effect.

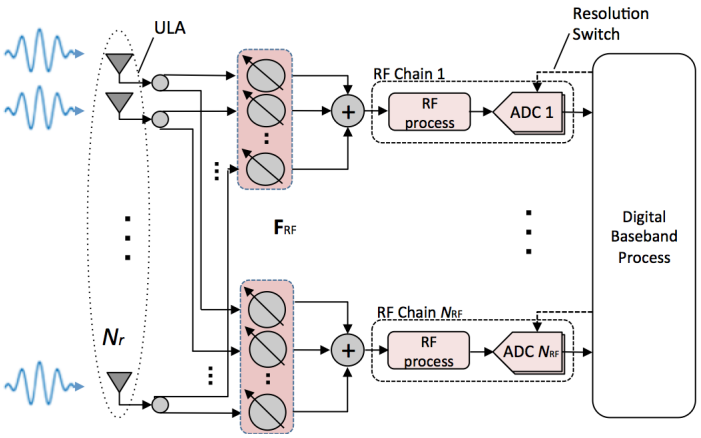


Figure 2.1: A hybrid beamforming receiver with resolution-adaptive ADCs.

- The analysis in [60] considered a quasi-static channel. This setting, however, ignores the transmission of a coded packet over different fading realizations so that rate adaptation cannot be applied over multiple fading realization. Especially, the quasi-static setting is not adequate in mmWave channels with the short coherence time [62]. Arguably, the ergodic rate analysis in this chapter offers more realistic evaluation in contemporary wireless systems that transmit a coded packet over multiple fading realizations [63].

2.2 System and Channel Model

2.2.1 Network and Signal Models

Single-cell MIMO uplink network is considered and N_u users with a single transmit antenna are served by a base station (BS) with N_r antennas. It is assumed that the BS is equipped with large antenna arrays ($N_r \gg N_u$). The hybrid architecture with low-resolution ADCs is employed at the BS. I

focus on uniform linear array (ULA) and assume that there are N_{RF} RF chains connected to N_{RF} pairs of ADCs. Employing adaptive ADCs such as flash ADCs, the proposed system is considered to be able to switch quantization resolution. Indeed, many power and resolution adaptive flash ADCs have been fabricated [64–66], and flash ADCs are the most suitable ADCs for applications requiring very large bandwidth with moderate resolution [67].

Assuming a narrowband channel, the received baseband analog signal $\mathbf{r} \in \mathbb{C}^{N_r}$ at the BS is expressed as

$$\mathbf{r} = \sqrt{p_u}\mathbf{H}\mathbf{s} + \tilde{\mathbf{n}} \tag{2.1}$$

where p_u is the average transmit power of users, $\mathbf{H}$ represents the $N_r \times N_u$ channel matrix between the BS and users, $\mathbf{s}$ indicates the $N_u \times 1$ vector of symbols transmitted by N_u users and $\tilde{\mathbf{n}} \in \mathbb{C}^{N_r}$ is the additive white Gaussian noise which follows complex Gaussian distribution $\tilde{\mathbf{n}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r})$. I further consider that the transmitted signal vector $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r})$ is Gaussian distributed with a zero mean and unit variance. It is also assumed that the channel $\mathbf{H}$ is perfectly known at the BS.

An analog beamformer $\mathbf{W}_{\mathrm{RF}} \in \mathbb{C}^{N_r \times N_{\mathrm{RF}}}$ is applied to $\mathbf{r}$ and constrained to satisfy $[\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{RF}}^H]_{i,i} = 1/N_r$, i.e., all element of $\mathbf{W}_{\mathrm{RF}}$ have equal norm of $1/\sqrt{N_r}$.

$$\mathbf{y} = \mathbf{W}_{\mathrm{RF}}^H \mathbf{r} = \sqrt{p_u}\mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{s} + \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{n}}. \tag{2.2}$$

I consider that the number of RF chains is less than the number of antennas ($N_{\mathrm{RF}} < N_r$), alleviating the power consumption and complexity at the BS.

Each beamforming output y_i is connected to an ADC pair as shown in Fig. 2.1. At each ADC, either a real or imaginary component of the complex signal y_i is quantized.

2.2.2 Channel Model

In this chapter, we consider mmWave channels. Since mmWave channels are expected to have limited scattering [21, 68, 69], each user channel is the sum of contributions of L scatterings and $L \ll N_r$. Adopting a geometric channel model, the kth user channel with L_k scatterers that contribute to L_k propagation paths is expressed as

$$\mathbf{h}_k = \sqrt{\gamma_k} \sum_{\ell=1}^{L_k} g_\ell^k \mathbf{a}(\theta_\ell^k) \in \mathbb{C}^{N_r} \tag{2.3}$$

where γ_k denotes the large-scale fading gain that includes geometric attenuation, shadow fading and noise power between the BS and kth user, g_ℓ^k is the complex gain of the ℓth path for the kth user and $\mathbf{a}(\theta_\ell^k)$ is the BS antenna array response vector corresponding to the azimuth AoA of the ℓth path for the kth user $\theta_\ell^k \in [-\pi/2, \pi/2]$. Each complex path gain $g_\ell^k \sim \mathcal{CN}(0, 1)$ is assumed to be an independent and identically distributed (IID) complex Gaussian random variable. It is also assumed that the number of propagation paths L_k is distributed as $L_k \sim \max\left\{Poisson(\lambda_{\mathrm{p}}), 1\right\}$ [70] for $k = 1, \cdots, N_u$. I call $\lambda_{\mathrm{p}} \in \mathbb{R}$ as the near average number of propagation paths.

Under the ULA assumption, the array response vector is expressed as

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N_r}} \left[1, e^{-j2\pi\vartheta}, e^{-j4\pi\vartheta}, \ldots, e^{-j2(N_r-1)\pi\vartheta}\right]^{\mathsf{T}}$$

where ϑ is the normalized spatial angle that is $\vartheta = \frac{d}{\lambda}\sin(\theta)$, λ is a signal wave length, and d is the distance between antenna elements. Considering the uniformly-spaced spatial angle, i.e., $\vartheta_i = \frac{d}{\lambda}\sin(\theta_i) = (i-1)/N_r$, the matrix of the array response vectors $\mathbf{A} = \left[\mathbf{a}(\theta_1), \cdots, \mathbf{a}(\theta_{N_r})\right]$ becomes a unitary discrete Fourier transform matrix; $\mathbf{A}^H\mathbf{A} = \mathbf{A}\mathbf{A}^H = \mathbf{I}$. Then, adopting the virtual channel representation [23, 69, 71], the channel vector $\mathbf{h}_k$ in (2.3) can be modeled as

$$\mathbf{h}_k = \mathbf{A}\tilde{\mathbf{h}}_{\mathrm{b},k} = \sum_{i=1}^{N_r} \tilde{h}_{\mathrm{b},(i,k)}\, \mathbf{a}(\theta_i)$$

where $\tilde{\mathbf{h}}_{\mathrm{b},k} \in \mathbb{C}^{N_r}$ is the beamspace channel of the kth user, i.e., $\tilde{\mathbf{h}}_{\mathrm{b},k}$ has L_k nonzero elements that contain the complex gains $\sim \mathcal{CN}(0,1)$ and the large-scale fading gain $\sqrt{\gamma_k}$. The beamspace channel matrix is denoted as $\tilde{\mathbf{H}}_{\mathrm{b}} = [\tilde{\mathbf{h}}_{\mathrm{b},1}, \cdots, \tilde{\mathbf{h}}_{\mathrm{b},N_u}]$ and it can be decomposed into $\tilde{\mathbf{H}}_{\mathrm{b}} = \tilde{\mathbf{G}}\mathbf{D}_\gamma^{1/2}$ where $\tilde{\mathbf{G}} \in \mathbb{C}^{N_r \times N_u}$ is the sparse matrix of complex path gains and $\mathbf{D}_\gamma = \mathrm{diag}(\gamma_1, \cdots, \gamma_{N_u})$. Accordingly, the beamspace channel of the kth user is expressed as $\tilde{\mathbf{h}}_{\mathrm{b},k} = \sqrt{\gamma_k}\tilde{\mathbf{g}}_k$. Finally, the channel matrix $\mathbf{H}$ is expressed as

$$\mathbf{H} = \mathbf{A}\tilde{\mathbf{H}}_{\mathrm{b}} = \mathbf{A}\tilde{\mathbf{G}}\mathbf{D}_\gamma^{1/2}. \tag{2.4}$$

It is assumed that the analog beamformer is composed of the array response vectors corresponding to the N_{RF} largest channel eigenmodes [26], i.e., $\mathbf{W}_{\mathrm{RF}} = \mathbf{A}_{\mathrm{RF}}$ where $\mathbf{A}_{\mathrm{RF}}$ is a $N_r \times N_{\mathrm{RF}}$ sub-matrix of $\mathbf{A}$. It is further assumed that the array response vectors in $\mathbf{A}_{\mathrm{RF}}$ capture all channel propagation paths from N_u users [72]. Then, the received signal after the analog beamforming in

Table 2.1: The Values of β for Different Quantization Bits b

b	1	2	3	4	5
β	0.3634	0.1175	0.03454	0.009497	0.002499

(2.2) reduces to

$$\mathbf{y} = \sqrt{p_u}\mathbf{A}_{\mathrm{RF}}^H\mathbf{H}\mathbf{s} + \mathbf{A}_{\mathrm{RF}}^H\tilde{\mathbf{n}} = \sqrt{p_u}\mathbf{H}_{\mathrm{b}}\mathbf{s} + \mathbf{n} \tag{2.5}$$

where $\mathbf{n} = \mathbf{A}_{\mathrm{RF}}^H\tilde{\mathbf{n}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_{\mathrm{RF}}})$ as $\mathbf{A}$ is unitary. Note that $\mathbf{H}_{\mathrm{b}}$ is the $N_{\mathrm{RF}} \times N_u$ sub-matrix of the beamspace channel matrix $\tilde{\mathbf{H}}_{\mathrm{b}}$ and contains $\sum_{k=1}^{N_u} L_k$ propagation path gains:

$$\mathbf{H}_{\mathrm{b}} = \mathbf{G}\mathbf{D}_{\gamma}^{1/2} \tag{2.6}$$

where $\mathbf{G}$ is the $N_{\mathrm{RF}} \times N_u$ sub-matrix of the complex gain matrix $\tilde{\mathbf{G}}$, corresponding to $\mathbf{A}_{\mathrm{RF}}$.

2.2.3 Quantization Model

I consider that each of the ith ADC pair has b_i quantization bits and adopt the AQNM [56, 73] as the quantization model to obtain a linearized quantization expression. The AQNM is accurate enough in low and medium SNR ranges [56]. After quantizing $\mathbf{y}$, we have the quantized signal vector

$$\begin{aligned}\mathbf{y}_{\mathrm{q}} &= \mathcal{Q}(\mathbf{y}) = \mathbf{D}_{\alpha}\,\mathbf{y} + \mathbf{n}_{\mathrm{q}} \\ &= \sqrt{p_u}\mathbf{D}_{\alpha}\mathbf{H}_{\mathrm{b}}\mathbf{s} + \mathbf{D}_{\alpha}\mathbf{n} + \mathbf{n}_{\mathrm{q}}\end{aligned} \tag{2.7}$$

where $\mathcal{Q}(\cdot)$ is an element-wise quantizer function separately applied to the real and imaginary parts and $\mathbf{D}_{\alpha}$ is a diagonal matrix with quantization

gains $\mathbf{D}_\alpha = \mathrm{diag}(\alpha_1, \cdots, \alpha_{N_{\mathrm{RF}}})$. The quantization gain α_i is a function of the number of quantization bits and defined as $\alpha_i = 1 - \beta_i$ where β_i is a normalized quantization error. Assuming the non-linear scalar MMSE quantizer and Gaussian transmit symbols, it can be approximated for $b_i > 5$ as $\beta_i = \frac{\mathbb{E}[|y_i - y_{\mathrm{q}i}|^2]}{\mathbb{E}[|y_i|^2]} \approx \frac{\pi\sqrt{3}}{2} 2^{-2b_i}$. The values of β_i are listed in Table 2.1 for $b_i \leq 5$. Note that b_i is the number of quantization bits for each real and imaginary part of y_i. The quantization noise $\mathbf{n}_{\mathrm{q}}$ is an additive noise which is uncorrelated with $\mathbf{y}$ and follows the complex Gaussian distribution with zero mean. For a fixed channel realization $\mathbf{H}_{\mathrm{b}}$, the covariance matrix of $\mathbf{n}_{\mathrm{q}}$ is

$$\mathbf{R}_{\mathbf{n}_{\mathrm{q}}\mathbf{n}_{\mathrm{q}}} = \mathbf{D}_\alpha \mathbf{D}_\beta \, \mathrm{diag}(p_u \mathbf{H}_{\mathrm{b}} \mathbf{H}_{\mathrm{b}}^H + \mathbf{I}_{N_{\mathrm{RF}}})$$

where $\mathbf{D}_\beta = \mathrm{diag}(\beta_1, \cdots, \beta_{N_{\mathrm{RF}}})$.

Assuming sampling at the Nyquist rate, the ADC power consumption is modeled as [56]

$$P_{\mathrm{ADC}}(b) = c \, f_s \, 2^b \tag{2.8}$$

where c is the energy consumption per conversion step (conv-step), called Walden's figure-of-merit, f_s is the sampling rate and b is the number of quantization bits. This model illustrates that the ADC power consumption scales exponentially in the number of quantization bits b.

2.3 ADC Bit Allocation Algorithms

In this section, BA algorithms are developed to improve the performance of the proposed system by leveraging the flexibility of ADC resolu-

tions. It is assumed that perfect knowledge of the channel state information (CSI) is available at the BS. The rationale behind this is that efficient algorithms have been proposed for mmWave channel estimation [14, 20, 22, 74, 75] by exploiting the sparse nature of mmWave channels. In the hybrid receiver structure with $N_{\mathrm{RF}} < N_r$, state-of-the-art mmWave channel estimators such as bisectional approach [22], modified OMP [74], and distributed grid message passing [75] validated the estimation performance. Assuming the use of high-resolution ADCs for a channel estimation phase, such estimation algorithms can be adopted in the considered system.

2.3.1 MMSQE Bit Allocation

I adopt the MSQE $\mathcal{E}(b) = \mathbb{E}\big[|y_- y_{\mathrm{q}}|^2\big]$ for $\mathbf{y}$ in (2.5) as a distortion measure. Assuming the MMSE quantizer and Gaussian transmit symbols, the MSQE of y_i with b_i quantization bits for $b_i > 5$ is modeled as [56]

$$\mathcal{E}_{y_i}(b_i) = \frac{\pi\sqrt{3}}{2}\sigma_{y_i}^2\, 2^{-2b_i} \tag{2.9}$$

where $\sigma_{y_i}^2 = p_u\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2 + 1$. Using (2.9) for any quantization bits,[2] I formulate the MMSQE problem through some relaxations. Then, the solution of the MMSQE problem minimizes the total quantization error by adapting quantization bits under constrained total ADC power consumption.

To avoid integer programming, the integer variables $\mathbf{b} \in \mathbb{Z}_+^{N_{\mathrm{RF}}}$ are re-

[2] Although (2.9) holds for $b_i > 5$, it can be validated by the performance of the proposed algorithms that (2.9) can provide a good approximation when formulating optimization problem even for a small number of quantization bits.

laxed to the real numbers $\mathbf{b} \in \mathbb{R}^{N_{\mathrm{RF}}}$ to find a closed-form solution. I also consider (2.9) to hold for $b_i \in \mathbb{R}$. Despite the fact that the ADC power consumption with b bits $P_{\mathrm{ADC}}(b) = 0$ for $b \leq 0$, I assume $P_{\mathrm{ADC}}(b) = cf_s 2^b$ in (2.8) to hold for $b \in \mathbb{R}$. Under the constraint of the total ADC power of the conventional fixed-ADC system in which all N_{RF} ADCs are equipped with $\bar{b}$ bits, the relaxed MMSQE problem is formulated as

$$\hat{\mathbf{b}} = \underset{\mathbf{b}=[b_1,\cdots,b_{N_{\mathrm{RF}}}]^\intercal}{\operatorname{argmin}} \sum_{i=1}^{N_{\mathrm{RF}}} \mathcal{E}_{y_i}(b_i) \tag{2.10}$$
$$\text{s.t.} \quad \sum_{i=1}^{N_{\mathrm{RF}}} P_{\mathrm{ADC}}(b_i) \leq N_{\mathrm{RF}} P_{\mathrm{ADC}}(\bar{b}),\ \mathbf{b} \in \mathbb{R}^{N_{\mathrm{RF}}}.$$

Here, $\bar{b}$ is the number of ADC bits for a fixed-ADC system, which is used to give a reference total ADC power in the constraint for the above MMSQE optimization problem. Proposition 1 provides the MMSQE-BA solution in a closed form by solving the Karush-Kuhn-Tucker (KKT) conditions for (2.10), which is different from the previously proposed greedy BA approach under a bit constraint in [76].

Proposition 1. *For the relaxed MMSQE problem in* (2.10)*, the optimal number of quantization bits which minimizes the total MSQE is derived as*

$$\hat{b}_i = \bar{b} + \log_2 \left(\frac{N_{\mathrm{RF}} \left(1 + \mathrm{SNR}_i^{\mathrm{rf}}\right)^{\frac{1}{3}}}{\sum_{j=1}^{N_{\mathrm{RF}}} \left(1 + \mathrm{SNR}_j^{\mathrm{rf}}\right)^{\frac{1}{3}}} \right),\ i = 1, \cdots, N_{\mathrm{RF}} \tag{2.11}$$

where $\mathrm{SNR}_i^{\mathrm{rf}} = p_u \| [\mathbf{H}_{\mathrm{b}}]_{i,:} \|^2$.

Proof. See Section 2.7. ■

In Proposition 1, $\mathrm{SNR}_i^{\mathrm{rf}}$ indicates the SNR of the ith received signal after analog beamforming y_i, which illustrates that the MMSQE-BA (2.11) depends on the channel gain of $\mathbf{y}$. The MMSQE-BA has the power of 1/3 which comes from the relationship between the MSQE $\mathcal{E}_{y_i}(b_i)$ and the ADC power $P_{\mathrm{ADC}}(b_i)$ in terms of b_i. Proposition 1 indicates that the optimal number of the ith ADC bits $\hat{b}_i$ increases logarithmically with $\left(1+\mathrm{SNR}_i^{\mathrm{RF}}\right)^{1/3}$ and decreases logarithmically with the sum of $\left(1+\mathrm{SNR}_j^{\mathrm{RF}}\right)^{1/3}$ for $j=1,\cdots,N_{\mathrm{RF}}$. Accordingly, the ADC pair with the relatively larger aggregated channel gain $\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2$ needs to have more quantization bits to minimize the total quantization distortion. Note that since the slowly changing channel characteristics such as large-scale fading and AoA mostly determines the channel gains and sparsity, they are the dominant factors for the BA solution in Proposition 1.

Since $\hat{b}_i$ in (2.11) is a real number solution, it is necessary to map it back to non-negative integers. Although the nearest integer mapping can be applied to the solution, it ignores the tradeoff between power consumption and quantization error and can violate the power constraint after the mapping. As an alternative, I propose a greedy-based tradeoff mapping method that is power efficient. First, the negative quantization bits ($\hat{b}_i < 0$) are mapped to zero, i.e., the ADC pairs with $\hat{b}_i \leq 0$ are deactivated. Note that this mapping does not violate the actual power constraint as $P_{\mathrm{ADC}}(b)=0$ for $b \leq 0$. Next, I map positive non-integer quantization bits ($\hat{b}_i > 0, \hat{b}_i \notin \mathbb{Z}$) to $\lceil \hat{b}_i \rceil$. If the power constraint is violated, i.e., $\sum_{i\in\mathbb{S}_+} P_{\mathrm{ADC}}(\lceil \hat{b}_i \rceil) > N_{\mathrm{RF}} P_{\mathrm{ADC}}(\bar{b})$ where $\mathbb{S}_+ = \{i \,|\, \hat{b}_i > 0\}$, it is necessary to map the subset of the positive non-

Algorithm 1: MMSQE-BA Algorithm

Set power constraint $P_{\max} = N_{\mathrm{RF}} P_{\mathrm{ADC}}(\bar{b})$ using (2.8) Set $\mathbb{S} = \{1 \dots N_{\mathrm{RF}}\}$ and $P_{\mathrm{total}} = 0$

for $i = 1 \dots N_{\mathrm{RF}}$ **do**

- (a) Compute $\hat{b}_i$ using (2.11) and $b_i = \max(0, \lceil \hat{b}_i \rceil)$
- (b) **if** $(b_i = 0)$, $\mathbb{S} = \mathbb{S} - \{i\}$
- (c) **else** $p_i = P_{\mathrm{ADC}}(b_i)$ and $P_{\mathrm{total}} = P_{\mathrm{total}} + p_i$
  - **if** $(\hat{b}_i \in \mathbb{Z})$, $\mathbb{S} = \mathbb{S} - \{i\}$

if $P_{\mathrm{total}} \leq P_{\max}$ **then**

- **return b**

for $i \in \mathbb{S}$ **do**

- compute $T_i = T(i)$ using (2.12)

while $P_{\mathrm{total}} > P_{\max}$ **do**

- (a) $i^* = \mathrm{argmin}_{i \in \mathbb{S}} T_i$
- (b) $b_{i^*} = b_{i^*} - 1$ and $\mathbb{S} = \mathbb{S} - \{i^*\}$
- (c) $P_{\mathrm{total}} = P_{\mathrm{total}} - p_{i^*} + P_{\mathrm{ADC}}(b_{i^*})$

return b

integer quantization bits to $\lfloor \hat{b}_i \rfloor$ instead of $\lceil \hat{b}_i \rceil$. Notice that the $\lfloor \hat{b}_i \rfloor$-mapping reduces the power consumption while increasing the MSQE. In this regard, it is necessary to find the best subset to perform power-efficient $\lfloor \hat{b}_i \rfloor$-mapping.

To determine the best subset of the positive non-integer quantization bits for $\lceil \hat{b}_i \rceil$, I propose a tradeoff function

$$T(i) = \left| \frac{\mathcal{E}_i(\hat{b}_i) - \mathcal{E}_i(\lfloor \hat{b}_i \rfloor)}{P_{\mathrm{ADC}}(\hat{b}_i) - P_{\mathrm{ADC}}(\lfloor \hat{b}_i \rfloor)} \right| \to \frac{2^{-2\lfloor \hat{b}_i \rfloor} - 2^{-2\hat{b}_i}}{2^{\hat{b}_i} - 2^{\lfloor \hat{b}_i \rfloor}} \sigma_{y_i}^2. \tag{2.12}$$

The proposed function in (2.12) represents the MSQE increase per unit power savings after mapping $\hat{b}_i$ to $\lfloor \hat{b}_i \rfloor$. For the $\lfloor \hat{b}_i \rfloor$-mapping, $\hat{b}_i$ with the smallest $T(i)$ is re-mapped to $\lfloor \hat{b}_i \rfloor$ from $\lceil \hat{b}_i \rceil$ to achieve the best tradeoff of quantization error vs. power consumption. This repeats for $\hat{b}_i$ with the next smallest $T(i)$ until the power constraint is satisfied. Algorithm 1 shows the proposed MMSQE-BA algorithm. The while-loop at line 7 will always end as this mapping algorithm can always satisfy the power constraint from the following reasons: (i) for $\hat{b}_i < 0$, the 0-bit mapping does not increase power, and (ii) for $\hat{b}_i > 0$, the total ADC power consumption always becomes $\sum_{i \in \mathbb{S}_+} P_{\mathrm{ADC}}(\lfloor \hat{b}_i \rfloor) \leq \sum_{i \in \mathbb{S}_+} P_{\mathrm{ADC}}(\hat{b}_i)$.

Note that the constant term in $1 + \mathrm{SNR}_i^{\mathrm{rf}}$ of (2.11) comes from the additive noise $\mathbf{n}$ in (2.5). Due to this noise term, the MMSQE-BA $\hat{b}_i$ would be almost the same for all ADCs when the transmit power p_u is small. In other words, in the low SNR regime, the noise term in $\hat{b}_i$ becomes dominant ($1 \gg \mathrm{SNR}_i^{\mathrm{rf}},\ i = 1, \cdots, N_{\mathrm{RF}}$). This leads to $\hat{b}_i \approx \bar{b}$ for $i = 1, \cdots, N_{\mathrm{RF}}$. The intuition behind this is that since we minimize the total MSQE of $\mathbf{y}$, which always includes the noise, the MMSQE-BA $\hat{b}_i$ minimizes mostly the quantization error of the noise in the low SNR regime, not the desired signal. Consequently, uniform bit allocation ($\hat{b}_i = \bar{b}$) across all the ADCs is likely to appear in the low SNR regime. In this perspective, the MMSQE-BA becomes more effective as the SNR increases while providing similar performance as fixed-ADCs in the low SNR regime. In Section 2.3.2, the MMSQE-BA is revised to overcome such noise-dependency.

2.3.2 Revised MMSQE Bit Allocation

The MMSQE-BA (2.11) is dependent to the additive noise as it minimizes the quantization error of y_i, not solely the desired signal. Accordingly, the MMSQE-BA is less effective in the low SNR regime. To address this problem, I modify the previous MMSQE problem (2.10) by considering to minimize the quantization error of only the desired signal. I ignore the additive noise $\mathbf{n}$ in $\mathbf{y}$ and consider the quantization of the desired signal $\mathbf{x} = \sqrt{p_u}\mathbf{H}_{\mathrm{b}}\mathbf{s}$ at the ADCs. According to the AQNM, the quantization of $\mathbf{x}$ can be modeled as

$$\mathbf{x}_{\mathrm{q}} = \sqrt{p_u}\mathbf{D}_{\alpha}\mathbf{H}_{\mathrm{b}}\mathbf{s} + \hat{\mathbf{n}}_{\mathrm{q}}$$

where $\hat{\mathbf{n}}_{\mathrm{q}}$ is the additive quantization noise uncorrelated with $\mathbf{x}_{\mathrm{q}}$. The corresponding MSQE for the ith signal x_i becomes

$$\mathcal{E}_{x_i}(b_i) = \mathbb{E}\Big[|x_i - x_{\mathrm{q}i}|^2\Big] = \frac{\pi\sqrt{3}}{2}\sigma_{x_i}^2\, 2^{-2b_i} \tag{2.13}$$

where $\sigma_{x_i}^2 = p_u\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2$. Using (2.13), the revised MMSQE problem is formulated as

$$\hat{\mathbf{b}}^{rev} = \underset{\mathbf{b}=[b_1,\cdots,b_{N_{\mathrm{RF}}}]^{\intercal}}{\operatorname{argmin}} \sum_{i=1}^{N_{\mathrm{RF}}} \mathcal{E}_{x_i}(b_i) \tag{2.14}$$

$$\text{s.t.} \quad \sum_{i=1}^{N_{\mathrm{RF}}} P_{\mathrm{ADC}}(b_i) \leq N_{\mathrm{RF}} P_{\mathrm{ADC}}(\bar{b}),\ \mathbf{b} \in \mathbb{R}^{N_{\mathrm{RF}}}.$$

Note that while the MMSQE-BA algorithm in Section 2.3.1 is developed with the proper AQNM quantization modeling (2.7), the revised MMSQE-BA (revMMSQE-BA) algorithm will be developed based on the quantization

modeling only for the desired signal term in (2.5). Consequently, this modeling approach may be inaccurate since the actual quantization process involves noise. Adopting the GMI which serves a lower bound on the channel capacity [77, 78], however, I show that (2.14) is equivalent to maximizing the GMI in the low SNR regime. Under the assumptions of IID Gaussian signaling $s_i \sim \mathcal{CN}(0,1)$ and applying a linear combiner $\mathbf{W}$ to the quantized signal $\mathbf{y}_{\mathrm{q}}$ with nearest-neighbor decoding, the GMI of user n [33] is expressed as

$$I_n^{\mathrm{GMI}}(\mathbf{w}_n, \mathbf{b}) = \log_2\left(1 + \frac{\kappa(\mathbf{w}_n,\ \mathbf{b})}{1 - \kappa(\mathbf{w}_n,\ \mathbf{b})}\right) \tag{2.15}$$

where

$$\kappa(\mathbf{w}_n,\ \mathbf{b}) = \frac{\left|\mathbb{E}[\mathbf{w}_n^H \mathbf{y}_{\mathrm{q}} \sqrt{p_u} s_n]\right|^2}{p_u \mathbb{E}[|\mathbf{w}_n^H \mathbf{y}_{\mathrm{q}}|^2]} = \frac{\mathbf{w}_n^H \mathbf{R}_{\mathbf{y}_{\mathrm{q}} s_n} \mathbf{R}_{\mathbf{y}_{\mathrm{q}} s_n}^H \mathbf{w}_n}{\mathbf{w}_n^H \mathbf{R}_{\mathbf{y}_{\mathrm{q}} \mathbf{y}_{\mathrm{q}}} \mathbf{w}_n} \tag{2.16}$$

Proposition 2. *Using the IID Gaussian signaling and linear combiner* $\mathbf{W}$ *to the quantized signal* $\mathbf{y}_{\mathrm{q}}$ *with nearest-neighbor decoding, the revised MMSQE problem* (2.14) *is equivalent to* (2.17) *in the low SNR regime.*

$$\hat{\mathbf{b}}^{GMI} = \underset{\mathbf{w}_n,\ \mathbf{b}}{\operatorname{argmax}} \sum_{n=1}^{N_u} I_n^{\mathrm{GMI}}(\mathbf{w}_n, \mathbf{b}) \tag{2.17}$$

$$s.t. \quad \sum_{i=1}^{N_{\mathrm{RF}}} P_{\mathrm{ADC}}(b_i) \leq N_{\mathrm{RF}} P_{\mathrm{ADC}}(\bar{b}),\ \mathbf{b} \in \mathbb{R}^{N_{\mathrm{RF}}}.$$

Proof. See Section 2.8. ■

Now, I solve (2.14) and derive the revMMSQE-BA solution $\hat{\mathbf{b}}^{rev}$ in the following proposition.

Proposition 3. *Assuming* $\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\| \neq 0$ *for* $i = 1, \cdots, N_{\mathrm{RF}}$*, the optimal number of quantization bits which minimizes the total MSQE of desired signals* $\mathbf{x}$ *for the revised MMSQE problem* (2.14) *is*

$$\hat{b}_i^{rev} = \bar{b} + \log_2 \left(\frac{N_{\mathrm{RF}} \|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^{\frac{2}{3}}}{\sum_{j=1}^{N_{\mathrm{RF}}} \|[\mathbf{H}_{\mathrm{b}}]_{j,:}\|^{\frac{2}{3}}} \right), \ i = 1, \cdots, N_{\mathrm{RF}}. \tag{2.18}$$

Proof. Replacing $\sigma_{y_i}^2$ with $\sigma_{x_i}^2$ ($c_i = \sigma_{x_i}^2$) in (2.37) and following the same steps in the proof of Proposition 1 in Section 2.7, we obtain (2.43). Then, (2.18) is obtained by putting $z_i = 2^{-2b_i}$, $\bar{z} = 2^{-2\bar{b}}$ and $c_i = \sigma_{x_i}^2$ into (2.43). ■

Corollary 1. *The revMMSQE-BA solution* $\hat{\mathbf{b}}^{rev}$ *maximizes the GMI in the low SNR regime and minimizes the quantization error of the beam-domain received signal* $\mathbf{y}$ *in the high SNR.*

Proof. When the SNR is low, Proposition 2 holds. For the high SNR, the MMSQE-BA solution reduces to the revMMSQE-BA solution, $\hat{\mathbf{b}} \to \hat{\mathbf{b}}^{rev}$, as $\mathrm{SNR}_i^{rf} \gg 1$. ■

Accordingly, even in the low SNR regime, ADC bits can be selectively assigned to maximize GMI, which can be considered as maximizing achievable rate. In this regard, the revMMSQE-BA provides noise-robust BA performance. Similar non-negative integer mapping can be performed by replacing $\sigma_{y_i}^2$ in (2.12) with $\sigma_{x_i}^2$.

2.3.3 Capacity Analysis with Bit Allocation

In this subsection, the capacity of the proposed system is analyzed for given $(\mathbf{b}, \mathbf{H}_{\mathrm{b}})$ when the SNR is low. Let $\eta = \mathbf{D}_{\alpha}\mathbf{n} + \mathbf{n}_{\mathrm{q}}$, then the capacity is expressed as

$$C\big(\mathbf{b}, \mathbf{H}_{\mathrm{b}}\big) = \log_2 \left| \mathbf{I}_{N_{\mathrm{RF}}} + p_u \mathbf{R}_{\eta\eta}^{-1} \mathbf{D}_{\alpha} \mathbf{H}_{\mathrm{b}} \mathbf{H}_{\mathrm{b}}^{H} \mathbf{D}_{\alpha}^{H} \right| \tag{2.19}$$

where $\mathbf{R}_{\eta\eta} = \mathbf{D}_{\alpha}\mathbf{D}_{\alpha}^{H} + \mathbf{R}_{\mathbf{n}_{\mathrm{q}}\mathbf{n}_{\mathrm{q}}}$.

Lemma 1. *For a given ADC bit allocation* $\mathbf{b}$*, the capacity of* (2.7) *in the low SNR regime is approximated as*

$$C_{low}\big(\mathbf{b}, \mathbf{H}_{\mathrm{b}}\big) = \log_2 \left(1 + \sum_{i=1}^{N_{\mathrm{RF}}} \frac{p_u \alpha_i \|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2}{1 + p_u(1-\alpha_i)\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2} \right). \tag{2.20}$$

Proof. In the low SNR regime, the capacity (2.19) can be approximated as

$$\begin{aligned} C\big(\mathbf{b}, \mathbf{H}_{\mathrm{b}}\big) &\approx \log_2 \left(1 + p_u \mathrm{tr}\left(\mathbf{R}_{\eta\eta}^{-1} \mathbf{D}_{\alpha} \mathbf{H}_{\mathrm{b}} \mathbf{H}_{\mathrm{b}}^{H} \mathbf{D}_{\alpha}^{H} \right) \right) \\ &= \log_2 \left(1 + p_u \mathrm{tr}\left(\left[\mathbf{I}_{N_{\mathrm{RF}}} + p_u \mathbf{D}_{\beta} \mathrm{diag}\big(\mathbf{H}_{\mathrm{b}} \mathbf{H}_{\mathrm{b}}^{H}\big) \right]^{-1} \mathbf{D}_{\alpha} \mathbf{H}_{\mathrm{b}} \mathbf{H}_{\mathrm{b}}^{H} \right) \right) \\ &= \log_2 \left(1 + p_u \sum_{i=1}^{N_{\mathrm{RF}}} \left(1 + p_u \beta_i \|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2 \right)^{-1} \alpha_i \|\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2 \right). \end{aligned}$$

This completes the proof for Lemma 1. ■

Lemma 1 gives the same intuition as the BA solutions (2.11), (2.18) that to maximize the capacity, it is necessary to assign more bits to the RF chain with larger channel gains in the low SNR regime. I further derive an approximation of the capacity with the proposed BA algorithms by applying a BA

solution to (2.20). In particular, I consider the case in which the revMMSQE-BA algorithm is applied to the resolution-adaptive ADC architecture since it is more effective in the low SNR regime.

Proposition 4. *For the low SNR, the capacity under the proposed resolution-adaptive ADC architecture with the revMMSQE-BA algorithm,* $C_{low}^{RBA}\big(\mathbf{b}, \mathbf{H}_{\mathrm{b}}\big)$*, can be approximated as*

$$\tilde{C}_{low}^{RBA}\big(\mathbf{H}_{\mathrm{b}}\big) = \log_2\left(1 + \Phi(\bar{b})\right) \tag{2.21}$$

where

$$\Phi = \sum_{i=1}^{N_{\mathrm{RF}}} \frac{p_u\left(1-\pi\sqrt{3}\,2^{-(2\bar{b}+1)}\left(N_{\mathrm{RF}}^{-2}\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^{-\frac{4}{3}}\left\{\sum_{j=1}^{N_{\mathrm{RF}}}\|[\mathbf{H}_{\mathrm{b}}]_{j,:}\|^{\frac{2}{3}}\right\}^2\right)_{2^{2\bar{b}}}^{-}\right)\big\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\big\|^2}{1+p_u\pi\sqrt{3}\,2^{-(2\bar{b}+1)}\left(N_{\mathrm{RF}}^{-2}\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^{-\frac{4}{3}}\left\{\sum_{j=1}^{N_{\mathrm{RF}}}\|[\mathbf{H}_{\mathrm{b}}]_{j,:}\|^{\frac{2}{3}}\right\}^2\right)_{2^{2\bar{b}}}^{-}\big\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\big\|^2}.$$

Proof. Forcing non-negativity to the revMMSQE-BA solution (2.18) as $b_i = (\hat{b}_i^{rev})^+$ where $(a)^+ = \max(a, 0)$, we apply $b_i = (\hat{b}_i^{rev})^+$ to (2.20). Then, the capacity $C_{low}^{RBA}\big(\mathbf{b}, \mathbf{H}_{\mathrm{b}}\big)$ can be approximated as

$$\begin{aligned} C_{low}^{RBA}\big(\mathbf{b}, \mathbf{H}_{\mathrm{b}}\big) &\approx C_{low}\big((\hat{\mathbf{b}}^{rev})^+, \mathbf{H}_{\mathrm{b}}\big) \\ &\overset{(a)}{\approx} \log_2\left(1 + \sum_{i=1}^{N_{\mathrm{RF}}} \frac{p_u\left(1 - \frac{\pi\sqrt{3}}{2}2^{-2(\hat{b}_i^{rev})^+}\right)^+\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2}{1 + p_u\frac{\pi\sqrt{3}}{2}2^{-2(\hat{b}_i^{rev})^+}\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2}\right) \\ &\overset{(b)}{\approx} \log_2\left(1 + \sum_{i=1}^{N_{\mathrm{RF}}} \frac{p_u\left(1 - \frac{\pi\sqrt{3}}{2}2^{-2(\hat{b}_i^{rev})^+}\right)\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2}{1 + p_u\frac{\pi\sqrt{3}}{2}2^{-2(\hat{b}_i^{rev})^+}\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2}\right) \end{aligned} \tag{2.22}$$

where (a) is from the approximation of α_i and (b) comes from removing the non-negativity condition of α_i. Since p_u and $\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|$, which corresponds to $\alpha_i < 0$ are small, the error from the approximation (b) can be negligible. Rearranging (2.22), we derive (2.21). ■

Since the revMMSQE-BA solution $\hat{\mathbf{b}}^{rev}$ is the function of $\mathbf{H}_{\mathrm{b}}$, $\tilde{C}_{low}^{RBA}$ in (2.21) is only a function of channels and captures the capacity that the proposed flexible ADC architecture can achieve adaptively for a given channel by using the revMMSQE-BA algorithm.

Now, regarding the implementation issue of the algorithm, I remark the following ADC resolution switching scenarios.

Remark 1. *Resolution switching at every channel coherence time allows the proposed architecture to adapt to different channel fading realizations, implying that it is the best switching scenario. Such coherence time switching in mmWave channels, however, may not be feasible due to the very short coherence time of mmWave channels [62]. Consequently, the switching period needs to be the multiples of the coherence time. In this case, switching at the time-scale of slowly changing channel characteristics marginally degrades the performance of the algorithms. Then, the worst-case scenario is not to exploit the flexibility of ADC resolutions, which is equivalent to have a infinitely long switching period, and converges to the fixed-ADC system with analog beamforming.*

In the next section, using a practical receiver, e.g., MRC, I analyze the worst-case scenario in terms of an *ergodic achievable rate* due to the intractability of the analysis with the BA solutions. The derived ergodic rate of the proposed system for the worst-case scenario offers the insight of the system performance as a function of the system parameters.

2.4 Worst-Case Analysis

I derive the ergodic achievable rate of user n for the hybrid beamforming architecture with fixed-ADCs over mmWave channels. The number of quantization bits in (2.7) is considered to be the same, i.e., $b_i = b$, and thus, $\alpha_i = \alpha$ for $i = 1, \cdots, N_{\mathrm{RF}}$. Using MRC, the quantized signal vector is

$$\mathbf{y}_{\mathrm{q}}^{\mathrm{mrc}} = \mathbf{H}_{\mathrm{b}}^H \mathbf{y}_{\mathrm{q}} = \alpha\sqrt{p_u}\mathbf{H}_{\mathrm{b}}^H\mathbf{H}_{\mathrm{b}}\mathbf{s} + \alpha\mathbf{H}_{\mathrm{b}}^H\mathbf{n} + \mathbf{H}_{\mathrm{b}}^H\mathbf{n}_{\mathrm{q}}$$

and the nth element of $\mathbf{y}_{\mathrm{q}}^{\mathrm{mrc}}$ for the user n is expressed as

$$y_{\mathrm{q},n}^{\mathrm{mrc}} = \alpha\sqrt{p_u}\mathbf{h}_{\mathrm{b},n}^H\mathbf{h}_{\mathrm{b},n}s_n + \alpha\sqrt{p_u}\sum_{\substack{k=1\\k\neq n}}^{N_u}\mathbf{h}_{\mathrm{b},n}^H\mathbf{h}_{\mathrm{b},k}s_k + \alpha\mathbf{h}_{\mathrm{b},n}^H\mathbf{n} + \mathbf{h}_{\mathrm{b},n}^H\mathbf{n}_{\mathrm{q}}. \tag{2.23}$$

Since $\mathbf{h}_{\mathrm{b},n} = \sqrt{\gamma_n}\mathbf{g}_n$ from (2.6), the desired signal power in (2.23) becomes $p_u\alpha^2\gamma_n^2\|\mathbf{g}_n\|^4$ and the noise-plus-interference power is given by

$$\Psi_{\mathbf{G}} = p_u\alpha^2\gamma_n\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k|\mathbf{g}_n^H\mathbf{g}_k|^2 + \alpha^2\gamma_n\|\mathbf{g}_n\|^2 + \alpha\beta\gamma_n\mathbf{g}_n^H\mathrm{diag}(p_u\mathbf{G}\mathbf{D}_\gamma\mathbf{G}^H + \mathbf{I}_{N_{\mathrm{RF}}})\mathbf{g}_n.$$

Simplifying the ratio of the two power terms, the achievable rate of the nth user is expressed as

$$R_n = \mathbb{E}\left[\log_2\left(1 + \frac{p_u\alpha\gamma_n\|\mathbf{g}_n\|^4}{\tilde{\Psi}_{\mathbf{G}}}\right)\right] \tag{2.24}$$

where

$$\tilde{\Psi}_{\mathbf{G}} = p_u\alpha\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k\left|\mathbf{g}_n^H\mathbf{g}_k\right|^2 + \alpha\|\mathbf{g}_n\|^2 + (1-\alpha)\mathbf{g}_n^H\mathrm{diag}\left(p_u\mathbf{G}\mathbf{D}_\gamma\mathbf{G}^H + \mathbf{I}_{N_{\mathrm{RF}}}\right)\mathbf{g}_n.$$

Considering large antenna arrays at the receiver, I use Lemma 2 to characterize the achievable rate (2.24).

Lemma 2. *Considering large antenna arrays at the BS, the uplink ergodic achievable rate* (2.24) *for the user* n *can be approximated as*

$$\tilde{R}_n = \log_2\left(1 + \frac{p_u \alpha \gamma_n \mathbb{E}\big[\|\mathbf{g}_n\|^4\big]}{\mathbb{E}\big[\tilde{\Psi}_{\mathbf{G}}\big]}\right) \tag{2.25}$$

where

$$\mathbb{E}\big[\tilde{\Psi}_{\mathbf{G}}\big] = \mathbb{E}\left[p_u \alpha \sum_{\substack{k=1 \\ k\neq n}}^{N_u} \gamma_k \big|\mathbf{g}_n^H \mathbf{g}_k\big|^2 + \alpha \|\mathbf{g}_n\|^2 + \beta \mathbf{g}_n^H \mathrm{diag}\big(p_u \mathbf{G}\mathbf{D}_\gamma \mathbf{G}^H + \mathbf{I}_{N_{\mathrm{RF}}}\big)\mathbf{g}_n\right]. \tag{2.26}$$

Proof. Lemma 1 in [79] is used for (2.24). ■

According to Lemma 2 in [79], the approximation in (2.25) becomes more accurate as the number of the BS antennas increases. Thus, this approximation will be particularly accurate in systems with the large number of antennas. Using Lemma 2, we derive the closed-form approximation of (2.24) as a function of system parameters: the transmit power, the number of BS antennas, RF chains, users and quantization bits, and the near average number of propagation paths.

Theorem 1. *The uplink ergodic achievable rate of the user* n *in the considered system with fixed ADCs is derived in a closed-form approximation as*

$$\tilde{R}_n = \log_2\left(1 + \frac{p_u \gamma_n \alpha \big(\lambda_{\mathrm{p}}^2 + 2\lambda_{\mathrm{p}} + 2e^{-\lambda_{\mathrm{p}}}\big)}{\eta}\right) \tag{2.27}$$

where

$$\eta = \Big(\lambda_{\mathrm{p}} + e^{-\lambda_{\mathrm{p}}}\Big)\left(1 + 2p_u \gamma_n (1-\alpha) + \big(\lambda_{\mathrm{p}} + e^{-\lambda_{\mathrm{p}}}\big)\frac{p_u}{N_{\mathrm{RF}}} \sum_{\substack{k=1 \\ k\neq n}}^{N_u} \gamma_k\right).$$

Proof. See Section 2.9 ■

Note that since the obtained ergodic rate in Theorem 1 is from the worst-case scenario, it can serve as the lower bound of the proposed architecture. This further implies that the proposed system can achieve higher ergodic rate than the derived rate by leveraging the flexibility of ADC resolutions. In addition, the derived ergodic rate explains general tradeoffs of the proposed system thanks to its tractability as a function of the system parameters. In contrast to the prior work [60] which assumes the quasi-static setting, the achievable rate in Theorem 1 considers mmWave fading channels in the ergodic sense. Accordingly, the derived ergodic rate measures the achievable rates by adopting the rate to the different fading realizations and thus offers more realistic evaluation than the quasi-static analysis in contemporary wireless systems.

Corollary 2 is derived for simplifying the ergodic rate in (2.27) when the near average number of propagation paths λ_p is moderate or large, and further provide remarks on the derived rate in behalf of profound understanding.

Corollary 2. *When the near average number of propagation paths* λ_{p} *is moderate or large,* (2.27) *can be approximated as*

$$\tilde{R}_n^{\dagger} = \log_2\left(1 + \frac{p_u \gamma_n \alpha(\lambda_{\mathrm{p}} + 2)}{1 + p_u\Big(2\gamma_n(1-\alpha) + \frac{\lambda_{\mathrm{p}}}{N_{\mathrm{RF}}}\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k\Big)}\right). \tag{2.28}$$

Proof. When λ_{p} is moderate or large enough, we can approximate $\lambda_{\mathrm{p}} + e^{-\lambda_{\mathrm{p}}} \approx$

$\lambda_{\rm p}$. Hence, we have the approximation (2.28) by replacing $\lambda_{\rm p} + e^{-\lambda_{\rm p}}$ with $\lambda_{\rm p}$ in (2.27). ■

Remark 2. *For fixed λ_p, (2.27) with $b \to \infty$ reduces to*

$$\tilde{R}_n \to \log_2 \left(1 + \frac{p_u \gamma_n (\lambda_{\rm p}^2 + 2\lambda_{\rm p} + 2e^{-\lambda_{\rm p}})}{1 + (\lambda_{\rm p} + e^{-\lambda_p}) \frac{p_u}{N_{\rm RF}} \sum_{\substack{k=1 \\ k \neq n}}^{N_u} \gamma_k} \right). \tag{2.29}$$

It is clear from (2.29) that the uplink rate can be improved by using more RF chains (larger $N_{\rm RF}$), which reduces the user interference. Let $N_{\rm RF} = \tau N_r$ with $0 < \tau < 1$, then for the fixed λ_p, the full-resolution rate (2.29) increases to

$$\tilde{R}_n \to \log_2 \Big(1 + p_u \gamma_n \big(\lambda_{\rm p}^2 + 2\lambda_{\rm p} + 2e^{-\lambda_{\rm p}} \big) \Big), \ \textit{as } N_r \to \infty.$$

Remark 3. *When using MRC, the uplink user rate transfers to the interference-limited regime from the noise-limited regime as p_u increases. Consequently, for fixed λ_p, (2.27) with the infinite transmit power ($p_u \to \infty$), converges to $\tilde{R}_n \to$*

$$\log_2 \left(1 + \frac{\gamma_n \alpha (\lambda_{\rm p}^2 + 2\lambda_{\rm p} + 2e^{-\lambda_{\rm p}})}{\big(\lambda_{\rm p} + e^{-\lambda_{\rm p}} \big) \left(2\gamma_n (1-\alpha) + \frac{(\lambda_{\rm p} + e^{-\lambda_{\rm p}})}{N_{\rm RF}} \sum_{\substack{k=1 \\ k \neq n}}^{N_u} \gamma_k \right)} \right). \tag{2.30}$$

The interference power can be eliminated by using an infinite number of antennas with $N_{\rm RF} = \tau N_r$ where $0 < \tau < 1$. Then, (2.30) approaches to

$$\tilde{R}_n \to \log_2 \left(1 + \frac{\alpha (\lambda_{\rm p}^2 + 2\lambda_{\rm p} + 2e^{-\lambda_{\rm p}})}{2(1-\alpha) \big(\lambda_{\rm p} + e^{-\lambda_{\rm p}} \big)} \right), \ \textit{as } N_r \to \infty. \tag{2.31}$$

The result (2.31) shows that even the infinite transmit power ($p_u \to \infty$) and the infinite number of BS antennas ($N_r \to \infty$) cannot fully compensate for the degradation caused by the quantization distortion when mmWave channels have a fixed number of propagation paths independent to N_r.

Now it is considered that λ_{p} is an increasing function of N_r [80] since larger antenna arrays with a fixed antenna spacing capture more physical paths due to larger array aperture. Then, Corollary 2 holds in the large antenna array regime.

Remark 4. *Without loss of generality, we assume $\lambda_{\mathrm{p}} = \epsilon N_r$ where $0 < \epsilon < 1$. Considering large antenna arrays with $N_{\mathrm{RF}} = \tau N_r$,* (2.28) *becomes*

$$\tilde{R}_n^{\dagger} = \log_2 \left(1 + \frac{p_u \gamma_n \alpha (\epsilon N_r + 2)}{1 + p_u \Big(2\gamma_n (1-\alpha) + \epsilon/\tau \sum_{\substack{k=1 \\ k \neq n}}^{N_u} \gamma_k \Big)} \right). \tag{2.32}$$

The achievable rate (2.32) *increases to infinity as $N_r \to \infty$ for any quantization bits b, which is not the case for the fixed λ_{p} as previously shown in Remark 2 and 3. With finite N_r, however,* (2.32) *cannot increase to infinity but converges to*

$$\tilde{R}_n^{\dagger} \to \log_2 \left(1 + \frac{\gamma_n \alpha (\epsilon N_r + 2)}{2\gamma_n (1-\alpha) + \frac{\epsilon}{\tau} \sum_{\substack{k=1 \\ k \neq n}}^{N_u} \gamma_k} \right), \ \ as\ p_u \to \infty. \tag{2.33}$$

It is observed that the convergence in (2.33) *is from the limited number of propagation paths.*

Remark 5. *Assuming that the transmit power inversely scales with the number of RF chains that is proportional to the number of BS antennas, i.e., $p_u = E_s/N_{\mathrm{RF}} = E_s/(\tau N_r)$, the rate in* (2.28) *with fixed E_s and $\lambda_{\mathrm{p}} = \epsilon N_r$ reduces to*

$$\tilde{R}_n^{\dagger} \to \log_2 \big(1 + E_s \gamma_n \alpha \epsilon / \tau \big), \quad as\ N_r \to \infty. \tag{2.34}$$

Thus, (2.34) *shows that we can scale down the user transmit power* p_u *proportionally to* $1/N_r$ *maintaining a desirable rate. In addition,* (2.34) *can be improved by using more quantization bits (larger* α*). This result is similar to that of the uplink rate of low-resolution massive MIMO systems with Rayleigh channels [57] but different in that* (2.34) *includes the factor of* ϵ/τ *due to the analog beamforming and the sparse nature of mmWave channels.*

In the following section, the performance of the proposed BA algorithms is evaluated through simulations. I also validate Theorem 1 and Corollary 2, and confirm the observations made in this section.

2.5 Simulation Results

A single cell with a radius of 200 m is considered and $N_u = 8$ users are distributed randomly over the cell. The minimum distance between the BS and users is 30 m, i.e., $30 \leq d_n \leq 200$ for $n = 1, \cdots, N_u$ where d_n $[m]$ is the distance between the BS and user n. Considering that the system operates at a 28 GHz carrier frequency, I adopt the mmWave pathloss model in [70] given as $PL(d_n)$ [dB] $= \alpha_{\text{pl}} + \beta_{\text{pl}} 10 \log_{10} d_n + \chi$ where $\chi \sim \mathcal{N}(0, \sigma_s^2)$ is the lognormal shadowing with $\sigma_s^2 = 8.7$ dB. The least square fits are $\alpha_{\text{pl}} = 72$ dB and $\beta_{\text{pl}} = 2.92$ dB [70]. Noise power is calculated as P_{noise} [dBm] $= -174 + 10 \log_{10} W + n_f$ where W and n_f are the transmission bandwidth and noise figure at the BS, respectively. I assume $W = 1$ GHz so as $f_s = 1$ GHz in (2.8), and $n_f = 5$ dB. Since I assume the normalized noise variance in the system model (2.1), the large-scale fading gain incorporating the normalization

is $\gamma_{n,\mathrm{dB}}$ [dB] $= -(PL(d_n) + P_{\mathrm{noise}})$. I consider the near average number of propagation paths $\lambda_{\mathrm{p}} = \epsilon N_r$ and the number of RF chains $N_{\mathrm{RF}} = \tau N_r$ with $\epsilon = 0.1$ and $\tau = 0.5$. It is assumed that the slowly changing characteristics of mmWave channels are consistent over $100 \times$ *the channel coherence time*, i.e., large-scale fading gains γ_n and the sparse structure of $\mathbf{G}$ in (2.6) are fixed over 100 channel realizations but the complex gains in $\mathbf{G}$ change at every channel realization. This simulation environment holds for the rest of this chapter unless mentioned otherwise.

The proposed algorithms are evaluated in terms of the capacity (2.19), uplink sum rate with MRC, and energy efficiency. The uplink sum rate is defined as $R = \sum_{n=1}^{N_u} R_n$. The ergodic rate of the nth user R_n is computed as follows. Applying MRC $\mathbf{D}_\alpha \mathbf{H}_{\mathrm{b}}$ to the quantized signal vector $\mathbf{y}_{\mathrm{q}}$ in (2.7), the ergodic rate of user n with ADC bit allocation $\mathbf{b}$ is given as

$$R_n(\mathbf{b}) = \mathbb{E}\left[\log_2\left(1 + \frac{p_u \gamma_n |\boldsymbol{\alpha}^H \mathbf{v}_n|^2}{\Psi_{\mathbf{G}}^{\mathrm{BA}}}\right)\right] \tag{2.35}$$

where

$$\Psi_{\mathbf{G}}^{\mathrm{BA}} = p_u \sum_{\substack{m=1 \\ m \neq n}}^{N_u} \gamma_m |\mathbf{g}_n^H \mathbf{D}_\alpha^2 \mathbf{g}_m|^2 + \mathbf{g}_n^H (\mathbf{D}_\alpha^4 + \mathbf{D}_\alpha^H \mathbf{R}_{\mathbf{n}_{\mathrm{q}}\mathbf{n}_{\mathrm{q}}} \mathbf{W}_\alpha) \mathbf{g}_n$$

with $\boldsymbol{\alpha} = [\alpha_1^2, \cdots, \alpha_{N_{\mathrm{RF}}}^2]^\intercal$ and $\mathbf{v}_n = [|g_{1,n}|^2, \cdots, |g_{N_{\mathrm{RF}},n}|^2]^\intercal$. Note that when quantization bits are same across ADCs, $b_i = b_j$, $\forall i, j$, (2.35) reduces to (2.24).

2.5.1 Average Capacity

We compare the proposed BA algorithms with the fixed-ADC case and include the infinite-resolution ADC case to indicate an upper bound. In Fig.

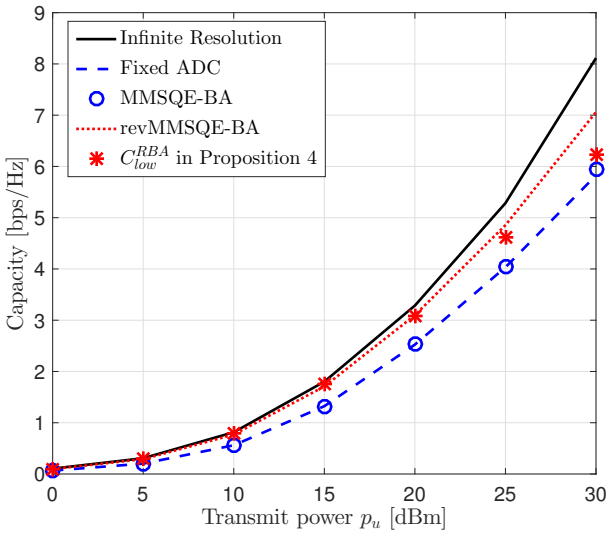
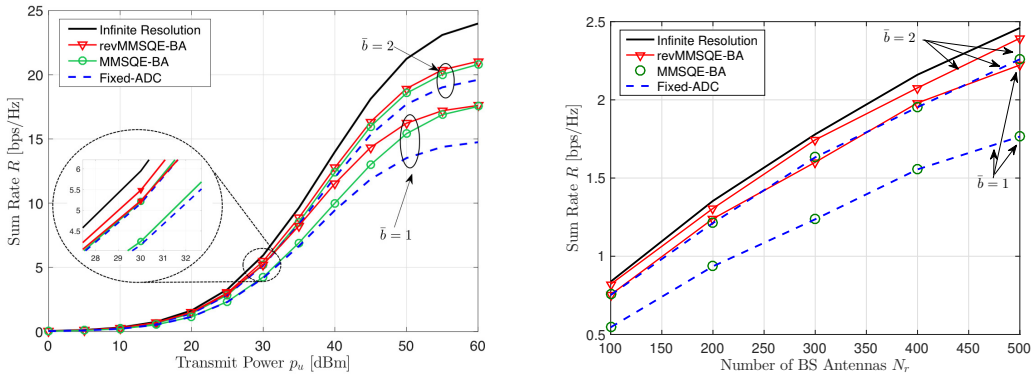


Figure 2.2: Simulation results of the average capacity for $N_u = 8$ users and $N_r = 256$ BS antennas with $\bar{b} = 1$ constraint bit.

2.2, the BA algorithms are applied with $\bar{b} = 1$. Recall that $\bar{b}$ is the number of ADC bits for a fixed-ADC system, which is used to give a reference total ADC power in the constraint for the MMSQE problem. This indicates that the total ADC power consumption with the algorithms is equal or less than that of N_{RF} 1-bit ADCs. In Fig. 2.2, the revMMSQE-BA improves the average capacity compared to the fixed ADCs. Moreover, it nearly achieves the capacity similar to the one with infinite-resolution ADCs in the low SNR regime, offering large energy saving from ADCs. The MMSQE-BA, however, does not show capacity improvement because the large pathloss makes the noise dominant over the range of p_u in Fig. 2.2. Consequently, the performance gap between the algorithms demonstrates the noise-robustness of the revMMSQE-BA. Although



(a) (b)

Figure 2.3: Simulation results of uplink sum rate for $\bar{b} \in \{1, 2\}$ constraint bits and $N_u = 8$ users (a) with $N_r = 256$ BS antennas and (b) with $p_u = 20$ dBm.

the gap between the capacity with the revMMSQE-BA and its approximation $\tilde{C}_{low}^{RBA}$ in (2.21) increases as p_u increases, $\tilde{C}_{low}^{RBA}$ provides a good approximation of the capacity with the revMMSQE-BA algorithm in the low SNR regime.

2.5.2 Average Uplink Sum Rate

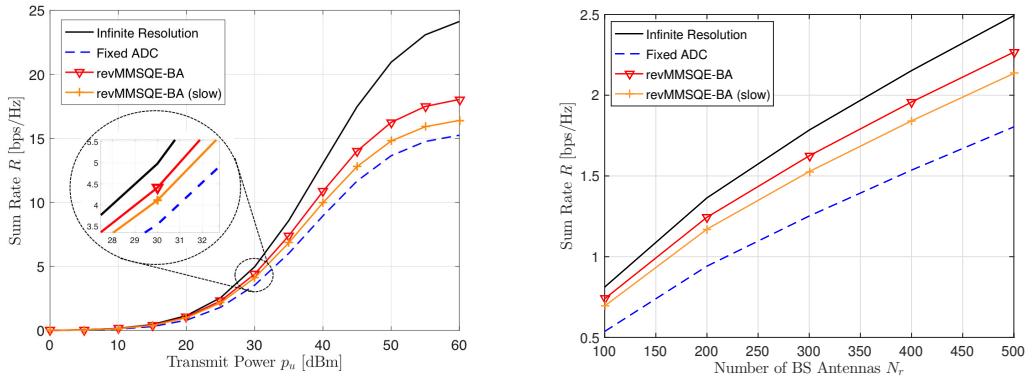
Fig. 2.3 shows the uplink sum rate of the MMSQE-BA, revMMSQE-BA and fixed-ADC systems (a) over different transmit power p_u with $N_r = 256$ antennas and $N_u = 8$ users and (b) over the different number of BS antennas N_r with $p_u = 20$ dBm transmit power and $N_u = 8$ users. In Fig. 2.3(a), the MMSQE-BA and revMMSQE-BA achieve the higher sum rate than the fixed-ADC system for both cases of $\bar{b} = 1$ and $\bar{b} = 2$. In particular, the revMMSQE-BA provides the best sum rate over the entire p_u while the MMSQE-BA shows a similar sum rate to the fixed-ADC case in the low SNR regime due to additive noise. This demonstrates that the revMMSQE-BA is

Table 2.2: Average Ratio of ADCs after Bit Allocation (%)

Constraint Bits	ADC Resolutions (bits)						
	0	1	2	3	4	5	6
$\bar{b} = 1$	40.78	28.20	26.46	4.46	0.10	0	0
$\bar{b} = 2$	32.10	16.32	25.54	19.36	6.54	0.14	0
$\bar{b} = 3$	19.40	7.46	18.42	28.54	22.58	3.48	0.12

robust to the noise. Notably, the rate of the MMSQE-BA becomes close to that of the revMMSQE-BA in the high SNR regime, which corresponds to the intuition that the revMMSQE-BA performs similarly to the MMSQE-BA in the high SNR regime.

In Fig. 2.3(b), the revMMSQE-BA also offers the best sum rate for all cases over the entire N_r. Notice that the sum rate of the revMMSQE-BA with $\bar{b} = 1$ shows similar rate to the fixed-ADC system with $\bar{b} = 2$, thus implying that the revMMSQE-BA achieves about the 1-bit better sum rate than the fixed-ADC system for the considered system. In contrast to the revMMSQE-BA, the MMSQE-BA shows no improvement for $p_u = 20$ dBm because the noise power is dominant when allocating quantization bits due to the large pathloss of mmWave channels. This, again, validates the noise-robustness of the revMMSQE-BA. Table 2.2 shows the average ratio of ADCs for different resolutions after applying the revMMSQE-BA algorithm for $\bar{b} = 1, 2$ and 3 with $p_u = 20$ dBm, $N_u = 8$, $N_r = 256$, and $N_{\mathrm{RF}} = 128$. Intuitively, the number of ADCs with higher resolution increases while that with lower resolution decreases as the constraint bits $\bar{b}$ increases. For example, the average number of 1-bit ADCs decreases from 36.10 (0.282×128) to 9.55 while that of 3-bit



(a) (b)

Figure 2.4: Simulation results of uplink sum rate for $\bar{b} = 1$ constraint bit and $N_u = 8$ users (a) with $N_r = 256$ BS antennas and (b) with $p_u = 20$ dBm, including the case of switching at slowly changing channel characteristics.

ADCs increases from 5.89 to 36.53 as $\bar{b}$ increases from 1 to 3.

In Fig. 2.4, to consider more realistic implementation of the proposed BA algorithms, we evaluate the revMMSQE-BA with two different switching periods: the channel coherence time and the time-scale of slowly changing channel characteristics (slow switching). It is observed that the slow switching results in small decrease of the sum rate from the coherence-time switching, while still achieving higher sum rate than the fixed-ADCs. Accordingly, the simulation results imply that the proposed hybrid architecture with slow switching can achieve the sum rate in between the revMMSQE-BA with the coherence-time switching and fixed-ADC systems. In addition, the general trends of the ergodic rate for the proposed BA algorithm and the fixed-ADC (worst-case scenario) are similar with the performance gap.

Regarding the total power consumption of the receiver, there will be an additional benefit of using the BA algorithms. The power saving from turning off the RF process associated with 0-bit ADCs (deactivated ADCs) as a consequence of BA can be accomplished. In Section 2.5.3, I provide energy efficiency for different ADC configurations to incorporate the additional advantage of the BA algorithms in performance evaluation.

2.5.3 Energy Efficiency

In this subsection, the revMMSQE-BA is evaluated in terms of energy efficiency. Energy efficiency can be defined as [32]

$$\eta_{\text{eff}} = \frac{RW}{P_{\text{tot}}} \text{ bits/Joule}$$

where P_{tot} is the receiver power consumption. Recall that R is the sum rate over a single cell, W is the transmission bandwidth. Let P_{LNA}, P_{PS}, P_{RFchain}, and P_{BB} represent power consumption in the low-noise amplifier, phase shifter, RF chain, and baseband processor, respectively. Applying an additional power consumption term due to the ADC resolution switching $P_{\text{SW}}(b_i)$, the receiver power consumption of the considered system in Fig. 2.1 is given as

$$P_{\text{tot}} = N_r P_{\text{LNA}} + N_{\text{act}}(N_r P_{\text{PS}} + P_{\text{RFchain}}) + 2\sum_{i=1}^{N_{\text{RF}}} \Big(P_{\text{ADC}}(b_i) + P_{\text{SW}}(b_i) \Big) + P_{\text{BB}}$$

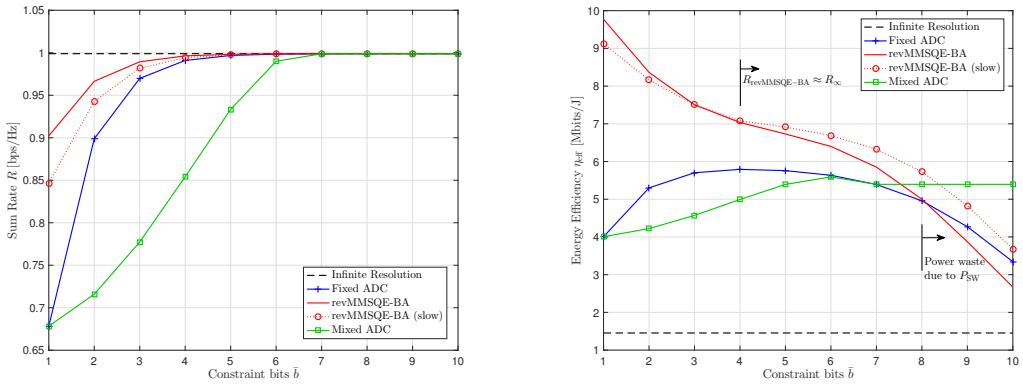
where N_{act} is the number of activated ADC pairs ($b_i \neq 0$). I assume $P_{\text{LNA}} = 20$ mW, $P_{\text{PS}} = 10$ mW, $P_{\text{RFchain}} = 40$ mW, and $P_{\text{BB}} = 200$ mW [23,32]. I consider $c = 494$ fJ/conv-step [56, 81] for $P_{\text{ADC}}(b_i)$ in (2.8). According to the measures

in [65], the switching power consumption $P_{\mathrm{SW}}(b_i)$ when switching from b_i^{p} bits to b_i bits can be modeled as

$$P_{\mathrm{SW}}(b_i) = c_{\mathrm{sw}}\left|2^{b_i} - 2^{b_i^{\mathrm{p}}}\right|, \quad i = 1, \cdots, N_{\mathrm{RF}} \tag{2.36}$$

where $c_{\mathrm{sw}} = 3.47$ (or 0.94) mW/conv-step if the resolution increases, $b_i > b_i^{\mathrm{p}}$ (or decreases, $b_i < b_i^{\mathrm{p}}$). Notice that (2.36) becomes zero when there is no change in resolution ($b_i = b_i^{\mathrm{p}}$).

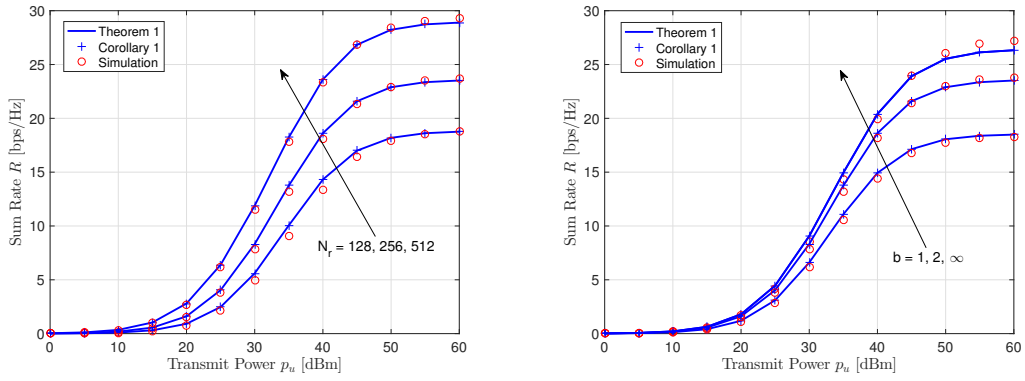
In the simulation, the following cases are compared: 1) fixed-ADC, 2) revMMSQE-BA with coherence-time switching, 3) reMMSQE-BA with slow switching, and 4) mixed-ADC systems [33]. I also simulate the infinite-resolution ADC case for benchmarking, assuming $b_\infty = 12$ quantization bits for the case. For the mixed-ADC system, we employ 1-bit and 7-bit ADCs, and assigns 7-bit ADCs to the RF chains with the strongest channel gains by satisfying the total ADC power constraint $N_{\mathrm{RF}}P_{\mathrm{ADC}}(\bar{b})$. Consequently, the number of 1-bit and 7-bit ADCs varies depending on the power constraint. Note that, except for the revMMSQE-BA, the number of activated ADC pairs is equal to that of RF chains $N_{\mathrm{act}} = N_{\mathrm{RF}}$. In addition, I impose two harsh simulation constraints on the algorithm. First, I apply the switching power consumption $P_{\mathrm{SW}}(b_i)$ only to the revMMSQE-BA despite the fact that the mixed-ADC system also consumes ADC switching power. Second, it is assumed that channel coherence time is equal to symbol duration, implying that if the switching operates at the channel coherence time, it occurs at every transmission.



(a) (b)

Figure 2.5: Uplink (a) sum rate and (b) energy efficiency simulation results with $N_r = 256$ BS antennas, $N_u = 8$ users and $p_u = 20$ dBm transmit power.

In Fig. 2.5, the sum rate and energy efficiency are simulated over different constraint bits $\bar{b}$. Note that the fixed-ADC, revMMSQE-BA, revMMSQE-BA (slow) and mixed-ADC system consume the similar total ADC power while the total power consumptions P_{tot} of the revMMSQE-BA and revMMSQE-BA (slow) are not equal to the other cases due to the deactivated (0-bit) ADCs and the switching power $P_{\text{SW}}(i)$. In Fig. 2.5(a), the revMMSQE-BA shows the higher sum rate than the fixed-ADC and mixed-ADC cases in the low-resolution regime ($\bar{b} \leq 4$), and it converges to the sum rate of the infinite-resolution case faster than the other two cases. Since the slow switching cannot capture the channel fluctuations caused by small-scale fading, the revMMSQE-BA (slow) shows a lower sum rate than the revMMSQE-BA in the low-resolution regime. The revMMSQE-BA (slow), however, achieves the higher sum rate than the fixed-ADC and mixed-ADC cases in the low-



(a) (b)

Figure 2.6: Uplink sum rate of the analytical approximations and the simulation results for $N_u = 8$ users with (a) $b = 2$ quantization bits and $N_r \in \{128, 256, 512\}$ BS antennas, and (b) $N_r = 256$ and $b \in \{1, 2, \infty\}$.

resolution regime ($\bar{b} \leq 4$). Given the same power constraint, the mixed-ADC system discloses the lowest sum rate due to the dominant ADC power consumption from the high-resolution ADCs.

In Fig. 2.5(b), the revMMSQE-BA provides the highest energy efficiency in the low-resolution regime, achieving the highest rate. In the high-resolution regime ($\bar{b} \geq 8$), the energy efficiency of the revMMSQE-BA is lower than that of the fixed-ADC and mixed-ADC systems due to the dissipation of power consumption in resolution switching. Note that although the revMMSQE-BA (slow) shows a lower energy efficiency than the revMMSQE-BA when $\bar{b} < 4$, it achieves a higher energy efficiency as $\bar{b}$ increases. This is because the slow switching accomplishes a better tradeoff between the rate and the switching power consumption than the coherence-time switching as $\bar{b}$ increases. Regarding the sum rate and energy efficiency, it is not worth-

while to consider the number of constraint bits above $\bar{b} = 6$ because the sum rate of the revMMSQE-BA is already comparable with the infinite-resolution system around $\bar{b} = 4$ with 22% better energy efficiency than the fixed-ADC case. Therefore, the simulation results demonstrate that the revMMSQE-BA with coherence-time switching provides the best performance, and that the slow switching approach offers performance improvement concerning the implementation. Fig. 2.5 indeed, implies that the proposed BA algorithm eliminates most of the quantization distortion requiring the minimum power consumption. Accordingly, I can employ existing digital beamformers to the power-constrained system when using the proposed BA algorithm in the low-resolution regime.

2.5.4 Worst-Case Analysis Validation

In this subsection, I validate Theorem 1 and Corollary 2, and confirm the observations in Section 2.4. For simulation, user locations are fixed once they are dropped in the cell, which corresponds to the setting of the analytical derivations in Section 2.4. Fig. 2.6 illustrates the sum rate for (a) $N_r \in \{128, 256, 512\}$ BS antennas with $b = 2$ quantization bits and for (b) $N_r = 256$ with $b \in \{1, 2, \infty\}$ over different transmit power p_u. The analytical results show accurate alignments with the simulation results in Fig. 2.6(a) and Fig. 2.6(b), which validates Theorem 1 and Corollary 2. The sum rates show the transition from the noise-limited regime to the interference-limited regime as p_u increases. This observation corresponds to the convergence of the achievable

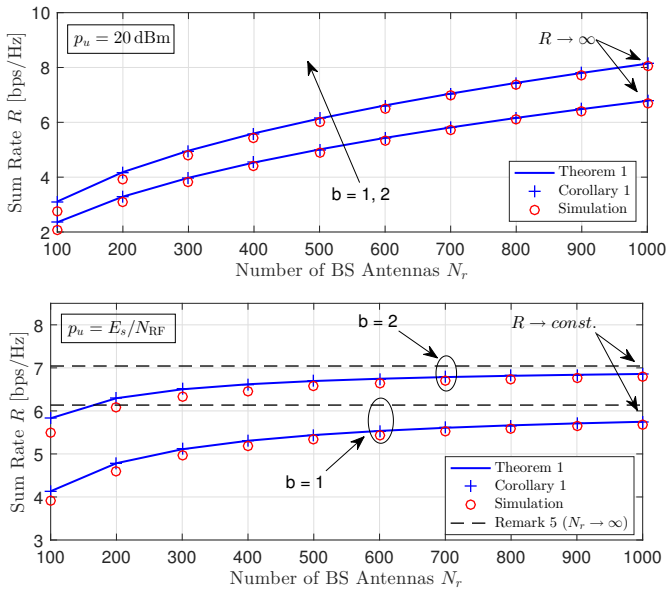


Figure 2.7: Uplink sum rate of the analytical and simulation results for $b \in \{1, 2\}$ quantization bits with $N_u = 8$ users. Two different cases of the transmit power are considered: i) $p_u = 20$ dBm and ii) $p_u = E_s/N_{\mathrm{RF}}$ with $E_s = 45$ dBm.

rate with increasing transmit power in Remark 4. Notably, the sum rate with $b = \infty$ also tends to converge in Fig. 2.6(b) due to the interference in (2.33).

I also evaluate the analytical results over the different number of BS antennas. The fixed transmit power of 20 dBm ($p_u = 20$ dBm) and the power-scaling law ($p_u = E_s/N_{\mathrm{RF}}$) with $E_s = 45$ dBm are considered in Fig. 2.7. It is observed that Fig. 2.7 validates the derived approximations of the achievable rate for the different power assumptions and offers intuitions discussed in Remark 4 and 5: the uplink sum rate with $p_u = 20$ dBm keeps increasing as N_r increases, and we can maintain the sum rate by decreasing the transmit

power p_u proportionally to $1/N_r$ ($N_{\mathrm{RF}} = \tau N_r$). In addition, the sum rate with the power-scaling law converges to (2.34) and can be improved by increasing the number of quantization bits (larger α) as illustrated in Fig. 2.7. In Section 2.5.2, the fixed-ADC approach serves the lower bound of the sum rate in the proposed architecture showing the similar trend to the BA strategies with performance gap. Therefore, the derived ergodic rate in Theorem 1 explains general tradeoffs of the proposed system serving the lower bound of the sum rate.

2.6 Conclusion

This chapter proposes the resolution-adaptive ADC network for the hybrid MIMO receiver for mmWave communications. Employing array response vectors for analog beamforming, I investigate the ADC bit-allocation problem to minimize the quantization distortion of received signals by leveraging the flexibility of ADC resolutions. One key finding is that the optimal number of ADC bits increases logarithmically proportional to the RF chain's SNR raised to the 1/3 power. The proposed algorithms outperform the conventional fixed ADCs in the proposed architecture in the low-resolution regime. In particular, the revised algorithm shows a higher capacity, sum rate and energy efficiency in any communication environment. Furthermore, the revised algorithm makes the quantization error of desired signals negligible while achieving higher energy efficiency than the fixed-ADC system. Having negligible quantization distortion allows existing state-of-the-art digital beamforming techniques to

be readily applied to the proposed system. The approximated capacity expression captures the capacity that the proposed flexible ADC architecture can achieve adaptively for a given channel by using the revised algorithm. The derived ergodic rate from the worst-case analysis explains the tradeoffs of the proposed system in terms of system parameters, serving as the lower performance bound of the proposed system. In the next chapter, I will also propose another advanced receiver architecture by focusing on optimization of analog combining rather than that of ADC resolutions so that quantization errors can be mitigated in the analog preprocessing.

2.7 Proof of Proposition 1

By defining $z_i = 2^{-2b_i}$, $\bar{z} = 2^{-2\bar{b}}$ and $c_i = \sigma_{y_i}^2$ where $\sigma_{y_i}^2 = p_u \|[\mathbf{H}]_{i,:}\|^2 + 1$, we can convert (2.10) into a simpler form given as

$$\hat{\mathbf{z}} = \underset{\mathbf{z} > \mathbf{0}_{N_{\mathrm{RF}}}}{\operatorname{argmin}} \mathbf{c}^{\intercal}\mathbf{z} \quad \text{s.t.} \quad \sum_{i=1}^{N_{\mathrm{RF}}} z_i^{-\frac{1}{2}} \leq N_{\mathrm{RF}} \bar{z}^{-\frac{1}{2}} \tag{2.37}$$

where $\mathbf{0}_{N_{\mathrm{RF}}}$ is a $N_{\mathrm{RF}} \times 1$ zero vector. Note that (2.37) is the equivalent problem to (2.10) and is a convex optimization problem. The global optimal solution of (2.10) can be achieved by the KKT conditions for (2.37).

By relaxing $\mathbf{z} > \mathbf{0}_{N_{\mathrm{RF}}}$ to $\mathbf{z} \geq \mathbf{0}_{N_{\mathrm{RF}}}$ and defining $\mathbf{v}$ as

$$\mathbf{v} = \begin{bmatrix} \sum_{i=1}^{N_{\mathrm{RF}}} z_i^{-\frac{1}{2}} - N_{\mathrm{RF}} \bar{z}^{-\frac{1}{2}} \\ -\mathbf{z} \end{bmatrix}, \tag{2.38}$$

KKT conditions become

$$\mathbf{c} + J_{\mathbf{v}}(\mathbf{z})^{\intercal}\boldsymbol{\mu} = \mathbf{0}_{N_{\mathrm{RF}}} \tag{2.39}$$

$$\mu_i \, v_i = 0, \quad \forall \, i \tag{2.40}$$

$$\mathbf{v} \leq \mathbf{0}_{(N_{\mathrm{RF}}+1)} \tag{2.41}$$

$$\boldsymbol{\mu} \geq \mathbf{0}_{(N_{\mathrm{RF}}+1)} \tag{2.42}$$

where the Jacobian matrix of $\mathbf{v}$ is defined as $J_{\mathbf{v}}(\mathbf{z}) = \begin{bmatrix}\mathbf{p} & -\mathbf{I}_{N_{\mathrm{RF}}}\end{bmatrix}^{\intercal}$ with $\mathbf{p} = \left[-\frac{1}{2}z_1^{-\frac{3}{2}}, \cdots, -\frac{1}{2}z_{N_{\mathrm{RF}}}^{-\frac{3}{2}}\right]^{\intercal}$, and $\boldsymbol{\mu} \in \mathbb{R}^{(N_{\mathrm{RF}}+1)}$ is the vector of the Lagrangian multipliers. Since $z_i \neq 0$, $i = 1, \cdots, N_{\mathrm{RF}}$, the Lagrangian multipliers become $\mu_j = 0$, $j = 2, \cdots, N_{\mathrm{RF}} + 1$, from (2.40). Hence, (2.39) guarantees $\mu_1 \neq 0$ as $\mathbf{c} \neq \mathbf{0}_{N_{\mathrm{RF}}}$, and (2.40) gives $v_1 = 0$ meaning that the equality holds for the power constraint. From (2.39) and (2.40), I have $c_i = \frac{1}{2}z_i^{-\frac{3}{2}}\mu_1$ and $\sum_{i=1}^{N_{\mathrm{RF}}} z_i^{-\frac{1}{2}} = N_{\mathrm{RF}}\bar{z}^{-\frac{1}{2}}$, which gives $\mu_1 = \left\{\frac{\bar{z}^{\frac{1}{2}}}{N_{\mathrm{RF}}}\sum_{j=1}^{N_{\mathrm{RF}}}(2c_j)^{\frac{1}{3}}\right\}^3 > 0$. Putting the Lagrangian multipliers $\mu_1 = \left\{\frac{\bar{z}^{\frac{1}{2}}}{N_{\mathrm{RF}}}\sum_{j=1}^{N_{\mathrm{RF}}}(2c_j)^{\frac{1}{3}}\right\}^3$ into $c_i = \frac{1}{2}z_i^{-\frac{3}{2}}\mu_1$, I have

$$\hat{z}_i = \bar{z}\left\{1/N_{\mathrm{RF}} \cdot \sum_{j=1}^{N_{\mathrm{RF}}}(c_j/c_i)^{\frac{1}{3}}\right\}^2. \tag{2.43}$$

Since $\hat{z}_i > 0$, the solution $\hat{\mathbf{z}}$ meets the KKT conditions. Using the definitions of $z_i, \bar{z}$ and c_i, (2.11) is obtained from (2.43). ■

2.8 Proof of Proposition 2

With the optimal combiner $\mathbf{w}_n^{\mathrm{opt}} = \mathbf{R}_{\mathbf{y}_{\mathrm{q}}\mathbf{y}_{\mathrm{q}}}^{-1}\mathbf{R}_{\mathbf{y}_{\mathrm{q}}s_n}$ [33], (2.16) becomes

$$\kappa(\mathbf{w}_n^{\mathrm{opt}}\ \mathbf{b}) = \mathbf{R}_{\mathbf{y}_{\mathrm{q}}s_n}^{H}\mathbf{R}_{\mathbf{y}_{\mathrm{q}}\mathbf{y}_{\mathrm{q}}}^{-1}\mathbf{R}_{\mathbf{y}_{\mathrm{q}}s_n}. \tag{2.44}$$

In the low SNR regime, $\mathbf{R}^H_{\mathbf{y}_\mathrm{q}\mathbf{y}_\mathrm{q}}$ is computed as

$$\begin{aligned}\lim_{p_u\to 0}\mathbf{R}_{\mathbf{y}_\mathrm{q}\mathbf{y}_\mathrm{q}} &= \lim_{p_u\to 0}\Big(p_u\mathbf{D}_\alpha\mathbf{H}_\mathrm{b}\mathbf{H}_\mathrm{b}^H\mathbf{D}_\alpha^H+\mathbf{D}_\alpha^2+\mathbf{R}_{\mathbf{n}_\mathrm{q}\mathbf{n}_\mathrm{q}}\Big)\\ &= \lim_{p_u\to 0}\Big(\mathbf{D}_\alpha+\mathbf{D}_\alpha\mathbf{D}_\beta\,\mathrm{diag}\big(p_u\mathbf{H}_\mathrm{b}\mathbf{H}_\mathrm{b}^H\big)\Big)\\ &= \mathbf{D}_\alpha. \end{aligned} \tag{2.45}$$

The correlation vector $\mathbf{R}_{\mathbf{y}_\mathrm{q}s_n}$ is computed as

$$\mathbf{R}_{\mathbf{y}_\mathrm{q}s_n} = \mathbb{E}[\mathbf{y}_\mathrm{q}s_n] = \sqrt{p_u}\mathbf{D}_\alpha\mathbf{h}_{\mathrm{b},n}. \tag{2.46}$$

Using (2.45) and (2.46), $\kappa(\mathbf{w}_n^{\mathrm{opt}}\ \mathbf{b})$ (2.44) becomes

$$\kappa(\mathbf{w}_n^{\mathrm{opt}},\ \mathbf{b}) = p_u\sum_{i=1}^{N_\mathrm{RF}}\Big(1-\frac{\pi\sqrt{3}}{2}2^{-2b_i}\Big)|h_{\mathrm{b},n,i}|^2 \tag{2.47}$$

where $h_{\mathrm{b},n,i}$ is the ith element of $\mathbf{h}_{\mathrm{b},n}$. Since we have $\log(1+x/(1-x)) = x+o(x)$ as $x\to 0$, the GMI becomes $I_n^{\mathrm{GMI}}(\mathbf{w}_n^{\mathrm{opt}},\mathbf{b}) = \kappa(\mathbf{w}_n^{\mathrm{opt}},\mathbf{b}) + o(\kappa(\mathbf{w}_n^{\mathrm{opt}},\mathbf{b}))$ in the low SNR regime, where $o(\cdot)$ is little-o. Thus, the objective function in the GMI maximization problem (2.17) with the low SNR approximation becomes

$$\begin{aligned}\hat{\mathbf{b}}^{GMI} &\simeq \operatorname*{argmax}_{\mathbf{b}}\sum_{n=1}^{N_u}\kappa(\mathbf{w}_n^{\mathrm{opt}},\mathbf{b})\\ &= \operatorname*{argmin}_{\mathbf{b}}\sum_{n=1}^{N_u}\sum_{i=1}^{N_\mathrm{RF}}p_u\frac{\pi\sqrt{3}}{2}2^{-2b_i}|h_{\mathrm{b},n,i}|^2. \end{aligned} \tag{2.48}$$

Note that (2.48) is equal to the objective function in (2.14). This completes the proof. ■

2.9 Proof of Theorem 1

Since the beamspace channels $\mathbf{g}$ are sparse, I use an indicator function to characterize the sparsity. The indicator function $\mathbb{1}_{\{i\in\mathcal{A}\}}$ is defined by

$$\mathbb{1}_{\{i\in\mathcal{A}\}} = \begin{cases} 1 & \text{if } i \in \mathcal{A} \\ 0 & \text{else.} \end{cases}$$

Utilizing the function $\mathbb{1}_{\{\cdot\}}$, I first model the ith complex path gain of the nth user $g_{i,n}$ as

$$g_{i,n} = \mathbb{1}_{\{i\in\mathcal{P}_n\}}\xi_{i,n}, \quad n = 1, \cdots, N_u \tag{2.49}$$

where $\mathcal{P}_n = \left\{ i \,\middle|\, g_{i,n} \neq 0, i = 1, \cdots, N_{\mathrm{RF}} \right\}$ and $\xi_{i,n}$ is an IID complex Gaussian random variable which follows $\mathcal{CN}(0,1)$. I compute the expectation of the number of propagation paths $\mathbb{E}[L]$. Since I assume $L \sim \max\{Q, 1\}$ with $Q \sim Poisson(\lambda_{\mathrm{p}})$, the expectation $\mathbb{E}[L]$ is derived as

$$\mathbb{E}\big[L\big] = e^{-\lambda_{\mathrm{p}}} + \sum_{\ell=1}^{\infty} \ell \, \frac{\lambda_{\mathrm{p}}^{\ell} e^{-\lambda_{\mathrm{p}}}}{\ell!} = e^{-\lambda_{\mathrm{p}}} + \lambda_{\mathrm{p}}. \tag{2.50}$$

Similarly, $\mathbb{E}\big[L^2\big]$ can be given as

$$\mathbb{E}\big[L^2\big] = e^{-\lambda_{\mathrm{p}}} + \sum_{\ell=1}^{\infty} \ell^2 \, \frac{\lambda_{\mathrm{p}}^{\ell} e^{-\lambda_{\mathrm{p}}}}{\ell!} \overset{(a)}{=} e^{-\lambda_{\mathrm{p}}} + \lambda_{\mathrm{p}} + \lambda_{\mathrm{p}}^2 \tag{2.51}$$

where (a) comes from $\mathbb{E}\big[Q^2\big] = \sum_{\ell=1}^{\infty} \ell^2 \, \frac{\lambda_{\mathrm{p}}^{\ell} e^{-\lambda_{\mathrm{p}}}}{\ell!}$ and $\mathbb{E}\big[Q^2\big] = \mathrm{Var}[Q] + \big\{\mathbb{E}[Q]\big\}^2$.

Now, I solve the expectations in Lemma 2. I have $|g_{i,n}|^2 = \mathbb{1}_{\{i\in\mathcal{P}_n\}}|\xi_{i,n}|^2$ and $|\xi_{i,n}|^2$ is distributed as exponential random variable with mean of the value 1, i.e., $|\xi_{i,n}|^2 \sim \exp(1)$. Despite the fact that the dimension of $\mathbf{g}_{i,n}$ is N_{RF}, $\|\mathbf{g}_{i,n}\|^2$ follows the chi-square distribution of $2L_n$ degrees of freedom $\|\mathbf{g}_n\|^2 \sim$

$\chi^2_{2L_n}$ due to the channel sparsity, where L_n is the number of propagation paths for the nth user. Then, I derive the expectation of $\|\mathbf{g}_n\|^2$ for the AWGN noise power in (2.26) as

$$\mathbb{E}\big[\|\mathbf{g}_n\|^2\big] = \mathbb{E}\Big[\mathbb{E}\big[\|\mathbf{g}_n\|^2\big|L_n\big]\Big] \overset{(a)}{=} e^{-\lambda_{\mathrm{p}}} + \lambda_{\mathrm{p}} \tag{2.52}$$

where (a) comes from $\|\mathbf{g}_n\|^2 \sim \chi^2_{2L_n}$ and (2.50). Similarly, the expectation of the desired signal power in (2.25) is

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{g}_n\|^4\Big] &= \mathbb{E}\Big[\mathbb{E}\big[\|\mathbf{g}_n\|^4 \,\big|\, L_n\big]\Big] \\ &= \mathbb{E}\Big[\mathrm{Var}\big[\|\mathbf{g}_n\|^2\big|\, L_n\big] + \Big\{\mathbb{E}\big[\|\mathbf{g}_n\|^2\big|\, L_n\big]\Big\}^2\Big] \\ &\overset{(a)}{=} \lambda_{\mathrm{p}}^2 + 2\lambda_{\mathrm{p}} + 2e^{-\lambda_{\mathrm{p}}}\end{aligned} \tag{2.53}$$

where (a) comes from $\|\mathbf{g}_n\|^2 \sim \chi^2_{2L_n}$, (2.50) and (2.51).

To further derive $\mathbb{E}\big[\tilde{\Psi}_{\mathbf{G}}\big]$ in (2.26), I solve the inter-user interference power $\mathbb{E}\big[|\mathbf{g}_n^H \mathbf{g}_k|^2\big]$ for $k \neq n$, which is given as

$$\begin{aligned}\mathbb{E}\Big[\big|\mathbf{g}_n^H \mathbf{g}_k\big|^2\Big] &= \mathbb{E}\left[\left(\sum_{i=1}^{N_{\mathrm{RF}}} g_{i,n}^* g_{i,k}\right)\left(\sum_{j=1}^{N_{\mathrm{RF}}} g_{j,n} g_{j,k}^*\right)\right] \\ &= \sum_{i=1}^{N_{\mathrm{RF}}} \mathbb{E}\Big[|g_{i,n}|^2 |g_{i,k}|^2\Big] \\ &\overset{(a)}{=} \sum_{i=1}^{N_{\mathrm{RF}}} \mathbb{E}\Big[\mathbb{1}_{\{i\in\mathcal{P}_n, i\in\mathcal{P}_k\}}\Big].\end{aligned} \tag{2.54}$$

Note that (a) comes from $g_{i,n} = \mathbb{1}_{\{i\in\mathcal{P}_n\}}\xi_{i,n}$ defined in (2.49) and the independence between $\xi_{i,n}$ and $\xi_{i,k}$ when $k \neq n$. Furthermore, $\mathbb{E}\Big[\mathbb{1}_{\{i\in\mathcal{P}_n, i\in\mathcal{P}_k\}}\Big]$ in (2.54)

can be computed as

$$\mathbb{E}\Big[\mathbb{1}_{\{i\in\mathcal{P}_n, i\in\mathcal{P}_k\}}\Big] \overset{(a)}{=} \left\{\mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\{i\in\mathcal{P}_n\}}\big|L_n\big]\Big]\right\}^2 = \left(\frac{\mathbb{E}\big[L_n\big]}{N_{\mathrm{RF}}}\right)^2 \overset{(b)}{=} \left(\frac{\lambda_{\mathrm{p}}+e^{-\lambda_{\mathrm{p}}}}{N_{\mathrm{RF}}}\right)^2 \tag{2.55}$$

where (a) is from the IID of L_n and the independence between the two events: $\{i \in \mathcal{P}_n\}$ and $\{i \in \mathcal{P}_k\}$, and (b) comes from (2.50). Putting (2.55) into (2.54), $\mathbb{E}\big[|\mathbf{g}_n^H\mathbf{g}_k|^2\big]$ finally becomes

$$\mathbb{E}\Big[|\mathbf{g}_n^H\mathbf{g}_k|^2\Big] = \frac{(\lambda_{\mathrm{p}}+e^{-\lambda_{\mathrm{p}}})^2}{N_{\mathrm{RF}}}. \tag{2.56}$$

Lastly, I compute the quantization noise power in (2.26) as

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{g}_n^H \mathrm{diag}\big(p_u\mathbf{G}\mathbf{D}_\gamma\mathbf{G}^H + \mathbf{I}_{N_{\mathrm{RF}}}\big)\mathbf{g}_n\Big] \\
&= \mathbb{E}\left[\sum_{i=1}^{N_{\mathrm{RF}}}|g_{i,n}|^2\left(p_u\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k|g_{i,k}|^2 + p_u\gamma_n|g_{i,n}|^2 + 1\right)\right] \\
&= \sum_{i=1}^{N_{\mathrm{RF}}}\left(p_u\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k\mathbb{E}\Big[|g_{i,k}|^2|g_{i,n}|^2\Big] + \mathbb{E}\Big[p_u\gamma_n|g_{i,n}|^4 + |g_{i,n}|^2\Big]\right) \\
&\overset{(a)}{=} \sum_{i=1}^{N_{\mathrm{RF}}}\left(p_u\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k\mathbb{E}\Big[\mathbb{1}_{\{i\in\mathcal{P}_k, i\in\mathcal{P}_n\}}\Big] + \big(2p_u\gamma_n+1\big)\mathbb{E}\Big[\mathbb{1}_{\{i\in\mathcal{P}_n\}}\Big]\right) \\
&\overset{(b)}{=} p_u\frac{(\lambda_{\mathrm{p}}+e^{-\lambda_{\mathrm{p}}})^2}{N_{\mathrm{RF}}}\sum_{\substack{k=1\\k\neq n}}^{N_u}\gamma_k + (\lambda_{\mathrm{p}}+e^{-\lambda_{\mathrm{p}}})(2p_u\gamma_n+1)
\end{aligned} \tag{2.57}$$

where (a) and (b) are from (2.49) and (2.55), respectively. Substituting (2.52), (2.53), (2.56) and (2.57) into (2.25) and simplifying the equations, we derive the final result (2.27). ■

Chapter 3

Two-Stage Analog Combining in Hybrid Beamforming Systems with Low-Resolution ADCs

In this chapter[1], I investigate hybrid analog/digital beamforming for multiple-input multiple-output (MIMO) systems with low-resolution analog-to-digital converters (ADCs) for millimeter wave (mmWave) communications. In the receiver, I propose to split the analog combining subsystem into a channel gain aggregation stage followed by a spreading stage. Both stages use phase shifters. The goal is to design the two-stage analog combiner to optimize mutual information (MI) between the transmitted and quantized signals by effectively managing quantization error. To this end, I formulate an unconstrained MI maximization problem without a constant modulus constraint on analog combiners, and derive a two-stage analog combining solution. The solution achieves the optimal scaling law with respect to the number of radio

[1]This chapter is based on the work published in the journal paper: J. Choi, G. Lee, and B. L. Evans, "Two-Stage Analog Combining in Hybrid Beamforming Systems with Low-Resolution ADCs," in *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2410-2425, May 1, 2019. Part of the work was also published in the conference paper: J. Choi, G. Lee, and B. L. Evans, "A Hybrid Beamforming Receiver with Two-Stage Analog Combiner and Low-Resolution ADCs," in *Proceedings of IEEE International Conference on Communications (ICC)*, May 2019. This work was supervised by Prof. Brian L. Evans and Dr. Gilwon Lee provided valuable feedback and contributions that improved the work.

frequency chains and maximizes the MI for homogeneous singular values of a MIMO channel. I further develop a two-stage analog combining algorithm to implement the derived solution for mmWave channels. By decoupling channel gain aggregation and spreading functions from the derived solution, the proposed algorithm implements the two functions by using array response vectors and a discrete Fourier transform matrix under the constant modulus constraint on each matrix element. Therefore, the proposed algorithm provides a near optimal solution for the unconstrained problem, whereas conventional hybrid approaches offer a near optimal solution only for a constrained problem. The closed-form approximation of the ergodic rate is derived for the algorithm, showing that a practical digital combiner with two-stage analog combining also achieves the optimal scaling law. Simulation results validate the algorithm performance and the derived ergodic rate.

3.1 Introduction

Unlike the previous chapter, a new hybrid beamforming architecture is proposed for homogeneous resolution ADCs to incorporate the impact of quantization error in the analog preprocessing. Hybrid beamforming architectures have been widely investigated to reduce the number of RF chains with minimum communication performance degradation. Singular value decomposition (SVD)-based analog combining designs were proposed [24, 82, 83] as the SVD transceiver maximizes the channel capacity. In [24], hybrid precoder and combiner design methods were developed by extracting the phases of the el-

ements of the singular vectors. Considering correlated channels, the SVD of the MIMO channel covariance matrix was used for analog combiner design to maximize mutual information in [82]. The performance of hybrid precoding systems was analyzed for MIMO downlink communications [84, 85]. It was shown that hybrid beamforming systems with a small number of RF chains can achieve the performance comparable to fully digital beamforming systems. For MIMO uplink communications, the Gram-schmidt based analog combiner design algorithm was developed in [86] to orthogonalize multiuser signals.

For mmWave channels, hybrid beamforming techniques were proposed by exploiting the limited scattering of the channels [21–23,29,30,87–89]. Adopting array response vectors (ARVs) for analog beamformer design, orthogonal matching pursuit (OMP)-based algorithms were developed in [21, 22, 87–89]. The proposed OMP-based algorithm in [21] approximates the minimum mean squared error (MMSE) combiner with a fewer number of RF chains than the number of antennas by using ARV-based analog combiners. The OMP-based algorithm in [21] was further improved by combining OMP and local search to reduce the computational complexity [88] and by iteratively updating the phases of the phase shifters [89]. A channel estimation technique was also proposed by using hierarchical multi-resolution codebook-based ARVs with low training overhead in [22]. By leveraging the sparse nature of mmWave channels, the proposed algorithms with ARV-based analog beamformers achieved the comparable performance with greatly reduced cost and power consumption compared to fully digital systems.

While the previous studies [21–24, 29, 30, 82–89] considered infinite-resolution ADCs in hybrid MIMO systems, hybrid beamforming systems with low-resolution ADCs were investigated in [25, 32, 35, 36, 90, 91] to take advantage of both the hybrid beamforming and low-resolution ADC architectures. The proposed algorithm in [25] attempted to design an analog combiner by minimizing the MSE including the quantization error. The analog combiner, however, is not constrained with a constant modulus, and the entire combining matrix needs to be designed for each transmitted symbol separately. Without considering the coarse quantization effect in combiner design, bit allocation techniques [90] and user scheduling methods [91] were developed for a given ARV-based analog combiner. In [32, 35], an alternating projection method was adopted to implement SVD-based analog combiners. The performance analysis of hybrid MIMO systems with low-resolution ADCs in [32] showed the superior tradeoff between performance and power consumption compared to fully digital systems and hybrid systems with infinite-resolution ADCs. In [36], a subarray antenna structure was considered, and an ARV-based combining algorithm was used to select the ARV that maximizes the aggregated channel gain. Although the analysis in [32, 35, 36] provided useful insights for the hybrid architecture with low-resolution ADCs such as the achievable rate and power tradeoff, the quantization error was not explicitly taken into account in the hybrid beamformer design. Consequently, considering the coarse quantization effect in the analog combiner design is still an open question.

3.1.1 Contributions

In this chapter, I derive a near optimal analog combining solution for an unconstrained MI maximization problem in hybrid MIMO systems with low-resolution ADCs. I, then, propose a two-stage analog combining architecture to properly implement the derived solution under a constant modulus constraint on each phase shifter. Splitting the solution into a channel gain aggregation stage by using ARVs and a gain spreading stage by using a discrete Fourier transform (DFT) matrix, the two-stage analog combining structure realizes the derived near optimal combining solution with phase shifter-based analog combiners for mmWave communications. The contributions of this paper can be summarized as follows:

- Without imposing a constant modulus constraint on an analog combiner, I formulate an unconstrained MI maximization problem for a hybrid MIMO system with low-resolution ADCs. For a general channel, I derive a near optimal analog combining solution which consists of (1) any semi-unitary matrix that includes the singular vectors of the signal space in the channel matrix and (2) any unitary matrix with constant modulus. The first and second parts in the derived solution can be considered as a channel gain aggregation function that collects the entire channel gains into the lower dimension and a spreading function that reduces quantization error by spreading the aggregated gains over RF chains, respectively. I show that the derived solution achieves the optimal scaling law with respect to the

number of RF chains and maximizes the MI when the singular values of a MIMO channel are the same.

- An ARV-based two-stage analog combining algorithm is further developed to implement the derived solution for mmWave channels under the constant modulus constraint on each phase shifter. Decoupling the channel gain aggregation and spreading functions from the solution, the algorithm implements the aggregation and spreading functions by using ARVs and a DFT matrix without losing the optimality of the solution in the large antenna array regime. Therefore, the two-stage analog combiner obtained from the proposed algorithm under the constant modulus constraint also provides a near optimal solution for the unconstrained MI maximization problem, whereas conventional hybrid approaches offer a near optimal solution only for a constrained problem. Since the DFT matrix is independent of channels, only passive phase shifters need to be appended to a conventional hybrid MIMO architecture with marginal complexity and cost increase, while achieving a large MI gain.

- A closed-form approximation of the ergodic rate with a maximum ratio combining (MRC) digital combiner is derived for the proposed algorithm. The derived rate characterizes the ergodic rate performance of the proposed two-stage analog combining architecture in terms of the system parameters including quantization resolution. The derived rate reveals that the ergodic rate of the MRC combiner achieves the same optimal scaling law with the

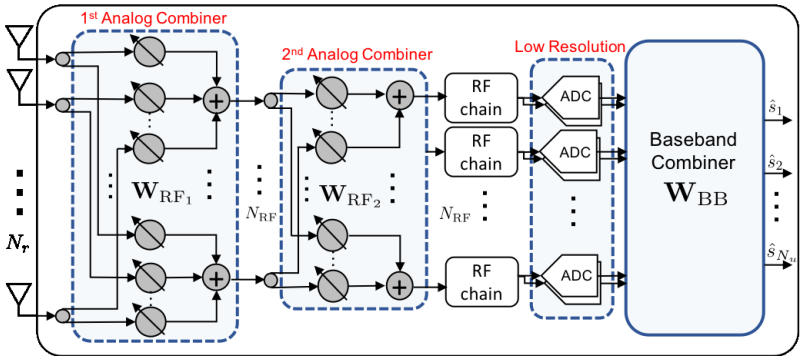


Figure 3.1: A receiver architecture with two-stage analog combining, low-resolution ADCs and digital combining.

proposed two-stage analog combiner by reducing the quantization error as the number of RF chains increases.

Simulation results demonstrate that the proposed two-stage analog combining algorithm outperforms conventional algorithms and validate the derived ergodic rate.

3.2 System Model

A single-cell uplink wireless network is considered in which the BS is equipped with N_r receive antennas and N_{RF} RF chains with $N_{\mathrm{RF}} < N_r$. As shown in Fig. 3.1, the antennas are uniform linear arrays (ULA), and each RF chain is followed by a pair of low-resolution ADCs. It is assumed that the BS serves N_u users each with a single transmit antenna with $N_u \leq N_{\mathrm{RF}}$.

3.2.1 Channel Model

The channel $\mathbf{h}_{\gamma,k}$ of user k is assumed to be the sum of the contributions of scatterers that contribute L_k propagation paths to the channel $\mathbf{h}_{\gamma,k}$ [92]. For mmWave channels, the number of channel paths L_k is expected to be small due to the limited scattering [10]. Here, I assume a narrowband channel where the components of each user signal propagating through L_k propagation paths are arriving within the sampling time. The discrete-time narrowband channel of user k can be modeled as

$$\mathbf{h}_{\gamma,k} = \frac{1}{\sqrt{\gamma_k}}\mathbf{h}_k = \sqrt{\frac{N_r}{\gamma_k L_k}} \sum_{\ell=1}^{L_k} g_{\ell,k}\mathbf{a}(\phi_{\ell,k}) \tag{3.1}$$

where γ_k denotes the pathloss of user k, $g_{\ell,k}$ is the complex gain of the ℓth propagation path of user k, and $\mathbf{a}(\phi_{\ell,k})$ is the ARV of the receive antennas corresponding to the azimuth AoA of the ℓth path of the kth user $\phi_{\ell,k} \in [-\pi/2, \pi/2]$. The complex channel gain $g_{\ell,k}$ follows an independent and identically distributed (i.i.d.) complex Gaussian distribution, $g_{\ell,k} \overset{i.i.d}{\sim} \mathcal{CN}(0,1)$. The ARV $\mathbf{a}(\theta)$ for the ULA antennas of the BS is given as

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N_r}}\Big[1, e^{-j\pi\vartheta}, e^{-j2\pi\vartheta}, \ldots, e^{-j(N_r-1)\pi\vartheta}\Big]^T$$

where the spatial angle $\vartheta = \frac{2d}{\lambda}\sin(\theta)$ is related to the physical AoA θ, d is the distance between antennas, and λ is the signal wave length. I use ϕ and θ to denote the physical AoAs of a user channel and physical angles of analog combiners, respectively. I also use φ and ϑ to denote the spatial angles for ϕ and θ, respectively, where $\varphi, \vartheta \in [-1, 1]$.

3.2.2 Signal and Quantization Model

For simplicity, a homogeneous long-term received SNR network[2] is considered, where a conventional uplink power control compensates for the pathloss and shadowing effect to achieve the same long-term received SNR target for all users in the cell [93, 94]. Let $\mathbf{x} = \mathbf{Ps}$ be the transmitted user signals where $\mathbf{P} = \text{diag}\{\sqrt{\rho\,\gamma_1}, \ldots, \sqrt{\rho\,\gamma_{N_u}}\}$ is the transmit power matrix and $\mathbf{s}$ is the $N_u \times 1$ transmitted symbol vector from N_u users. Further, let $\mathbf{H}_\gamma = \mathbf{HB}$ represent the $N_r \times N_u$ channel matrix where $\mathbf{H}_\gamma = [\mathbf{h}_{\gamma,1}, \ldots, \mathbf{h}_{\gamma,N_u}]$ is the channel matrix, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{N_u}]$ is the channel matrix after the uplink power control, and $\mathbf{B} = \text{diag}\{\sqrt{1/\gamma_1}, \ldots, \sqrt{1/\gamma_{N_u}}\}$. The analog baseband received signal vector is given as

$$\mathbf{r} = \mathbf{H}_\gamma \mathbf{x} + \mathbf{n} = \mathbf{HBPs} + \mathbf{n} = \sqrt{\rho}\mathbf{Hs} + \mathbf{n}$$

where $\mathbf{n}$ indicates the $N_r \times 1$ additive white noise vector. I assume zero mean and unit variance for the user symbols $\mathbf{s}$ and noise $\mathbf{n}$. The noise follows the complex Gaussian distribution $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r})$ and thus, ρ is considered to be the SNR. Also, the channel state information of $\mathbf{H}$ is assumed to be available at the BS. Note that the uplink power control offers homogeneous long-term average SNR, which can relieve the problem of user signals with small power being buried under the quantization error due to the limited dynamic range of low-resolution ADCs. More investigation to resolve with problem, however, is needed as a future work.

[2]We remark that the derived analysis in this chapter can also be applicable to a heterogeneous long-term received SNR network with minor modification.

After the BS receives the signals from users, the signals are combined via two analog combiners as shown in Fig. 3.1. Then, the received baseband analog signal vector becomes

$$\begin{aligned}\mathbf{y} &= \sqrt{\rho}\mathbf{W}_{\mathrm{RF}_2}^H\mathbf{W}_{\mathrm{RF}_1}^H\mathbf{H}\mathbf{s} + \mathbf{W}_{\mathrm{RF}_2}^H\mathbf{W}_{\mathrm{RF}_1}^H\mathbf{n} \\ &= \sqrt{\rho}\mathbf{W}_{\mathrm{RF}}^H\mathbf{H}\mathbf{s} + \mathbf{W}_{\mathrm{RF}}^H\mathbf{n} \end{aligned} \tag{3.2}$$

where $\mathbf{W}_{\mathrm{RF}} = \mathbf{W}_{\mathrm{RF}_1}\mathbf{W}_{\mathrm{RF}_2}$ denotes the two-stage analog combiner, $\mathbf{W}_{\mathrm{RF}_1} \in \mathbb{C}^{N_r \times N_{\mathrm{RF}}}$ is the first analog combiner, and $\mathbf{W}_{\mathrm{RF}_2} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{RF}}}$ is the second analog combiner. Each real and imaginary part of the combined signal (3.2) are quantized at ADCs with b quantization bits. Assuming a MMSE scalar quantizer and Gaussian signaling $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_u})$, I adopt an additive quantization noise model (AQNM) [73] which shows reasonable accuracy in the low to medium SNR ranges [56]. The AQNM approximates the quantization process in linear form, which is equivalent to the approximation with Bussgang decomposition for low-resolution ADCs [54]. The quantized signal vector is expressed as [54, 73]

$$\mathbf{y}_{\mathrm{q}} = \mathcal{Q}(\mathbf{y}) = \alpha_b\sqrt{\rho}\mathbf{W}_{\mathrm{RF}}^H\mathbf{H}\mathbf{s} + \alpha_b\mathbf{W}_{\mathrm{RF}}^H\mathbf{n} + \mathbf{q} \tag{3.3}$$

where $\mathcal{Q}(\cdot)$ is the element-wise quantizer, the scalar quantization gain is $\alpha_b = 1 - \beta_b$ where $\beta_b = \mathbb{E}[|y - y_{\mathrm{q}}|^2]/\mathbb{E}[|y|^2]$, and $\mathbf{q}$ denotes the quantization noise vector. For $b > 5$ quantization bits, β_b is approximated as $\beta_b \approx \frac{\pi\sqrt{3}}{2}2^{-2b}$. For $b \leq 5$, the values of β_b are listed in Table 1 in [57]. The quantization noise vector $\mathbf{q}$ is uncorrelated to the quantization input $\mathbf{y}$ and follows the complex

Gaussian distribution $\mathbf{q} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R_{qq}})$, where the covariance matrix is given as [73]

$$\mathbf{R_{qq}} = \alpha_b \beta_b \mathrm{diag}\big\{\rho \mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{H}^H \mathbf{W}_{\mathrm{RF}} + \mathbf{W}_{\mathrm{RF}}^H \mathbf{W}_{\mathrm{RF}}\big\}. \tag{3.4}$$

Then, a digital combiner $\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{RF}}}$ is applied to the quantized signal in (3.3) as

$$\mathbf{z} = \alpha_b \sqrt{\rho} \mathbf{W}_{\mathrm{BB}}^H \mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{s} + \alpha_b \mathbf{W}_{\mathrm{BB}}^H \mathbf{W}_{\mathrm{RF}}^H \mathbf{n} + \mathbf{W}_{\mathrm{BB}}^H \mathbf{q}. \tag{3.5}$$

3.3 Optimality of Two-Stage Analog Combining

In this section, I provide a near optimal structure for the first and second analog combiners $\mathbf{W}_{\mathrm{RF}_1}, \mathbf{W}_{\mathrm{RF}_2}$ in low-resolution ADC systems for a general channel. To this end, I first formulate an unconstrained MI maximization problem without a constant modulus condition on the analog combiner $\mathbf{W}_{\mathrm{RF}}$. Then, I derive a near optimal solution for the unconstrained problem, which can be split into two different functions corresponding to the two-stage analog combiner.

Let $\mathcal{C}(\mathbf{W}_{\mathrm{RF}}) \triangleq \mathrm{I}(\mathbf{s}; \mathbf{y}_{\mathrm{q}})$. I consider the MI between the transmit symbols $\mathbf{s}$ and quantized signals $\mathbf{y}_{\mathrm{q}}$ under the AQNM model as a measure to maximize. The MI is given as

$$\mathcal{C}(\mathbf{W}_{\mathrm{RF}}) = \log_2 \Big| \mathbf{I}_{N_{\mathrm{RF}}} + \rho \alpha_b^2 \big(\alpha_b^2 \mathbf{W}_{\mathrm{RF}}^H \mathbf{W}_{\mathrm{RF}} + \mathbf{R_{qq}}\big)^{-1} \mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{H}^H \mathbf{W}_{\mathrm{RF}} \Big|. \tag{3.6}$$

Using (3.6), I formulate the maximum MI problem by only assuming a semi-unitary constraint on the analog combiner $\mathbf{W}_{\mathrm{RF}}^H \mathbf{W}_{\mathrm{RF}} = \mathbf{I}_{N_{\mathrm{RF}}}$ as in [32] to

keep the effective noise being white Gaussian noise. Note that the MI in (3.6) incorporates the effect of inter-user interference and noise. Accordingly, the relaxed MI maximization problem is formulated as [3]

$$\mathcal{P}1: \ \mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}} = \underset{\mathbf{W}_{\mathrm{RF}}}{\mathrm{argmax}} \ \mathcal{C}(\mathbf{W}_{\mathrm{RF}}), \ \text{s.t.} \ \mathbf{W}_{\mathrm{RF}}^{H}\mathbf{W}_{\mathrm{RF}} = \mathbf{I}. \tag{3.7}$$

From the data processing inequality given below, the MI between transmitted signals $\mathbf{s}$ and quantized signals $\mathbf{y}_{\mathrm{q}}$ is larger or equal to the MI between transmitted signals $\mathbf{s}$ and digitally combined signal $\mathbf{z}$, i.e., $\mathrm{I}(\mathbf{s};\mathbf{y}_{\mathrm{q}}) \geq \mathrm{I}(\mathbf{s};\mathbf{z})$ [97]. In this work, I maximize $\mathrm{I}(\mathbf{s};\mathbf{y}_{\mathrm{q}})$ so that a derived solution can maximize the upper bound of $\mathrm{I}(\mathbf{s};\mathbf{z})$.

Under the perfect quantization system where the number of quantization bits is assumed to be infinite, the unconstrained optimal analog combiner for the problem $\mathcal{P}1$ is given as the matrix $\mathbf{U}_{1:N_{\mathrm{RF}}}$ that consists of the first N_{RF} left singular vectors of $\mathbf{H}$. The optimal solution $\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}}$ of the problem $\mathcal{P}1$ with a finite number of quantization bits, however, is still not known. I first derive an optimal scaling law with respect to the number of RF chains N_{RF}, and provide a solution that achieves the scaling law.

Theorem 2 (Optimal scaling law)**.** *For fixed $N_{\mathrm{RF}}/N_r = \kappa$ with $\kappa \in (0,1)$, the MI with the optimal combiner $\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}}$ for the problem $\mathcal{P}_1$ scales with N_{RF} as*

$$\mathcal{C}(\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}}) \sim N_u \log_2 N_{\mathrm{RF}} \tag{3.8}$$

[3] To take into account the fairness among users, solving the problem of maximizing the minimum signal-to-interference-plus-noise ratio would be necessary [95, 96]. Since it is beyond the scope of this work, I only consider the MI maximization, leaving the fairness problem as future work.

and the optimal scaling law is achieved by using $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ *such that:*

(*i*) $\mathbf{W}_{\mathrm{RF}_1}^{\star} = [\mathbf{U}_{1:N_u}\ \mathbf{U}_{\perp}]$, *and*

(*ii*) $\mathbf{W}_{\mathrm{RF}_2}^{\star}$ *is any* $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ *unitary matrix that satisfies the constant modulus condition on its elements,*

where $\mathbf{U}_{1:N_u}$ *is the matrix of the left-singular vectors corresponding to the first* N_u *largest singular values of* $\mathbf{H}$ *and* $\mathbf{U}_{\perp}$ *denotes the matrix of any orthonormal vectors whose column space is orthogonal to that of* $\mathbf{U}_{1:N_u}$.

Proof. Since the optimal solution for $\mathcal{P}1$ is not known, I first derive an upper bound of $\mathcal{C}(\mathbf{W}_{\mathrm{RF}})$ and its scaling law with respect to N_{RF}. I, then, show that adopting $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$, which satisfies the conditions (*i*) and (*ii*) in Theorem 2, achieves the same scaling law of the upper bound.

An arbitrary semi-unitary combiner $\mathbf{W}_{\mathrm{RF}}$ can be decomposed into

$$\mathbf{W}_{\mathrm{RF}} = [\mathbf{U}_{||}\ \mathbf{U}_{\perp}]\bar{\mathbf{W}}_{\mathrm{RF}}, \tag{3.9}$$

where $\mathbf{U}_{||}$ is an $N_r \times m$ matrix composed of m orthonormal basis vectors whose column space is in the subspace of $\mathrm{Span}(\mathbf{u}_1, \cdots, \mathbf{u}_{N_u})$ with $1 \leq m \leq N_u$, $\mathbf{U}_{\perp}$ is an $N_r \times (N_{\mathrm{RF}} - m)$ matrix composed of $(N_{\mathrm{RF}} - m)$ orthonormal basis vectors whose column space is in the subspace of $\mathrm{Span}^{\perp}(\mathbf{u}_1, \cdots, \mathbf{u}_{N_u})$, and $\bar{\mathbf{W}}_{\mathrm{RF}}$ is an $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ unitary matrix. Here, $\mathbf{u}_i$ is the i-th left-singular vector of $\mathbf{H}$.

Using (3.9), the term $\mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{H}^H \mathbf{W}_{\mathrm{RF}}$ in (3.6) can be re-written as

$$\begin{aligned}\mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{H}^H \mathbf{W}_{\mathrm{RF}} &= \bar{\mathbf{W}}_{\mathrm{RF}}^H [\mathbf{U}_{||}\ \mathbf{U}_{\perp}]^H \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H [\mathbf{U}_{||}\ \mathbf{U}_{\perp}] \bar{\mathbf{W}}_{\mathrm{RF}} \\ &= \bar{\mathbf{W}}_{\mathrm{RF}}^H \underbrace{\begin{bmatrix} \mathbf{U}_{||}^H \mathbf{U}_{1:N_u} \mathbf{\Lambda}_{N_u} \mathbf{U}_{1:N_u}^H \mathbf{U}_{||} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\triangleq \mathbf{Q}} \bar{\mathbf{W}}_{\mathrm{RF}} \end{aligned} \tag{3.10}$$

where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \cdots, \lambda_{N_u}, 0, \cdots, 0\} \in \mathbb{C}^{N_r \times N_r}$, $\mathbf{\Lambda}_{N_u} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_{N_u}\}$, λ_i is the ith largest singular value of $\mathbf{H}\mathbf{H}^H$, and $\mathbf{U}_{1:N_r} = [\mathbf{u}_1, \cdots, \mathbf{u}_{N_r}]$. The matrix $\mathbf{Q}$ has rank m and can be decomposed into $\mathbf{Q} = \mathbf{U}_{\mathbf{Q}} \bar{\mathbf{\Lambda}} \mathbf{U}_{\mathbf{Q}}^H$, where $\mathbf{U}_{\mathbf{Q}}$ is the $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ matrix consisting of N_{RF} singular vectors of $\mathbf{Q}$; and $\bar{\mathbf{\Lambda}} = \mathrm{diag}\{\bar{\lambda}_1, \cdots, \bar{\lambda}_m, 0, \cdots, 0\} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{RF}}}$. Here, $\bar{\lambda}_i$ is the ith largest singular value of $\mathbf{Q}$. Since $\mathbf{U}_{\mathbf{Q}}$ is unitary, $\bar{\mathbf{W}}_{\mathrm{RF}}$ can be re-expressed as

$$\bar{\mathbf{W}}_{\mathrm{RF}} = \mathbf{U}_{\mathbf{Q}} \overline{\mathbf{W}}_{\mathrm{RF}}. \tag{3.11}$$

and $\overline{\mathbf{W}}_{\mathrm{RF}}$ is still unitary. Substituting (3.11) into (3.10), we have $\mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{H}^H \mathbf{W}_{\mathrm{RF}} = \overline{\mathbf{W}}_{\mathrm{RF}}^H \bar{\mathbf{\Lambda}} \overline{\mathbf{W}}_{\mathrm{RF}}$ and the MI in (3.6) becomes

$$\mathcal{C}(\mathbf{W}_{\mathrm{RF}}) = \log_2 \left| \mathbf{I} + \frac{\alpha_b}{\beta_b} \mathrm{diag}^{-1} \left\{ \overline{\mathbf{W}}_{\mathrm{RF}}^H \bar{\mathbf{\Lambda}} \overline{\mathbf{W}}_{\mathrm{RF}} + \frac{1}{\beta_b \rho} \mathbf{I} \right\} \overline{\mathbf{W}}_{\mathrm{RF}}^H \bar{\mathbf{\Lambda}} \overline{\mathbf{W}}_{\mathrm{RF}} \right|. \tag{3.12}$$

Let $\mathbf{G} = \overline{\mathbf{W}}_{\mathrm{RF}}^H \bar{\mathbf{\Lambda}}^{1/2} = [\mathbf{G}_{\mathrm{sub}}\ \mathbf{0}]$, where $\mathbf{G}_{\mathrm{sub}}$ is the $N_{\mathrm{RF}} \times m$ submatrix of $\mathbf{G}$.

Then, the MI can be upper bounded as

$$\begin{aligned}
\mathcal{C}(\mathbf{W}_{\mathrm{RF}}) &= \log_2 \left| \mathbf{I}_{N_{\mathrm{RF}}} + \frac{\alpha_b}{\beta_b} \mathbf{G}^H \mathrm{diag}^{-1} \left\{ \|[\mathbf{G}]_{i,:}\|^2 + \frac{1}{\beta_b \rho} \right\} \mathbf{G} \right| \\
&= \log_2 \left| \mathbf{I}_m + \frac{\alpha_b}{\beta_b} \mathbf{G}_{\mathrm{sub}}^H \mathrm{diag}^{-1} \left\{ \|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2 + \frac{1}{\beta_b \rho} \right\} \mathbf{G}_{\mathrm{sub}} \right| \\
&\overset{(a)}{=} \log_2 \left| \mathbf{I}_m + \frac{\alpha_b}{\beta_b} \tilde{\mathbf{G}}_{\mathrm{sub}}^H \tilde{\mathbf{G}}_{\mathrm{sub}} \right| \\
&= \sum_{i=1}^{m} \log_2 \left(1 + \frac{\alpha_b}{\beta_b} \lambda_i \{ \tilde{\mathbf{G}}_{\mathrm{sub}}^H \tilde{\mathbf{G}}_{\mathrm{sub}} \} \right) \\
&\overset{(b)}{\leq} m \log_2 \left(1 + \frac{\alpha_b}{\beta_b m} \sum_{i=1}^{m} \lambda_i \{ \tilde{\mathbf{G}}_{\mathrm{sub}}^H \tilde{\mathbf{G}}_{\mathrm{sub}} \} \right) \\
&\overset{(c)}{=} m \log_2 \left(1 + \frac{\alpha_b}{\beta_b m} \sum_{i=1}^{N_{\mathrm{RF}}} \frac{\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2}{\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2 + \frac{1}{\beta_b \rho}} \right)
\end{aligned} \tag{3.13}$$

where (a) follows by letting $\tilde{\mathbf{G}}_{\mathrm{sub}}$ be the matrix whose each row i is given as i-th row of $\mathbf{G}_{\mathrm{sub}}$ normalized by $\left(\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2 + \frac{1}{\beta_b \rho}\right)^{1/2}$; (b) comes from Jensen's inequality and the concavity of $\log_2(1+x)$ for $x > 0$; and (c) is from

$$\sum_{i=1}^{m} \lambda_i \{ \tilde{\mathbf{G}}_{\mathrm{sub}}^H \tilde{\mathbf{G}}_{\mathrm{sub}} \} = \mathrm{Tr} \{ \tilde{\mathbf{G}}_{\mathrm{sub}}^H \tilde{\mathbf{G}}_{\mathrm{sub}} \} = \sum_{i=1}^{N_{\mathrm{RF}}} \frac{\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2}{\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2 + \frac{1}{\beta_b \rho}}.$$

The upper bound of $\mathcal{C}(\mathbf{W}_{\mathrm{RF}})$ in (3.13) can further be upper bounded by $m \log_2(1 + \frac{\alpha_b N_{\mathrm{RF}}}{\beta_b m})$ because $\frac{\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2}{\|[\mathbf{G}_{\mathrm{sub}}]_{i,:}\|^2 + \frac{1}{\beta_b \rho}} < 1$. Since the derivative of this bound with respect to m is positive for $m > 0$ with any given $\alpha_b, N_{\mathrm{RF}} > 0$, it is maximized when $m = N_u$, and thus, it scales as $N_u \log_2 N_{\mathrm{RF}}$, as $N_{\mathrm{RF}} \to \infty$.

Now, I prove that the scaling law can be achieved by the two-stage analog combiner $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star} \mathbf{W}_{\mathrm{RF}_2}^{\star}$ in Theorem 2. Let $\mathbf{C} \triangleq \mathbf{W}_{\mathrm{RF}_2}^{\star H} \mathbf{\Lambda}_{N_{\mathrm{RF}}} \mathbf{W}_{\mathrm{RF}_2}^{\star}$. From the relationship $\mathbf{W}_{\mathrm{RF}}^{\star H} \mathbf{H} \mathbf{H}^H \mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_2}^{\star H} \mathbf{\Lambda}_{N_{\mathrm{RF}}} \mathbf{W}_{\mathrm{RF}_2}^{\star} = \mathbf{C}$, where $\mathbf{\Lambda}_{N_{\mathrm{RF}}} =$

$\text{diag}\{\lambda_1, \cdots, \lambda_{N_u}, 0, \cdots, 0\} \in \mathbb{C}^{N_{\text{RF}} \times N_{\text{RF}}}$ and (3.12), I have

$$\mathcal{C}(\mathbf{W}_{\text{RF}}^{\star}) = \log_2 \left| \mathbf{I}_{N_{\text{RF}}} + \frac{\alpha_b}{\beta_b} \text{diag}^{-1} \left\{ \mathbf{C} + \tfrac{1}{\beta_b \rho} \mathbf{I}_{N_{\text{RF}}} \right\} \mathbf{C} \right| \tag{3.14}$$

$$\overset{(a)}{=} \log_2 \left| \mathbf{I} + \frac{\alpha_b}{\beta_b} \left(\frac{\sum_{i=1}^{N_u} \lambda_i}{N_{\text{RF}}} + \frac{1}{\beta_b \rho} \right)^{-1} \mathbf{W}_{\text{RF}_2}^{\star H} \mathbf{\Lambda}_{N_{\text{RF}}} \mathbf{W}_{\text{RF}_2}^{\star} \right| \tag{3.15}$$

$$= \sum_{k=1}^{N_u} \log_2 \left(1 + \frac{\alpha_b \rho N_{\text{RF}} \lambda_k}{N_{\text{RF}} + (1-\alpha_b)\rho \sum_{i=1}^{N_u} \lambda_i} \right)$$

$$= \sum_{k=1}^{N_u} \log_2 \left(1 + \frac{\alpha_b \rho N_{\text{RF}} \lambda_k / N_r}{\kappa + (1-\alpha_b)\rho \sum_{i=1}^{N_u} \lambda_i / N_r} \right) \tag{3.16}$$

$$\overset{(b)}{\sim} N_u \log_2 N_{\text{RF}}, \text{ as } N_{\text{RF}} \to \infty.$$

Here, (a) is from that all diagonal entries of $\mathbf{W}_{\text{RF}_2}^{\star H} \mathbf{\Lambda}_{N_{\text{RF}}} \mathbf{W}_{\text{RF}_2}^{\star}$ are the same as $d_j = \frac{\sum_{i=1}^{N_u} \lambda_i}{N_{\text{RF}}}$, for $j = 1, \cdots, N_{\text{RF}}$ because of the constant modulus property of $\mathbf{W}_{\text{RF}_2}^{\star}$; (b) follows from the fact that as $N_{\text{RF}} \to \infty$, i.e., as $N_r \to \infty$, I have $\frac{1}{N_r} \mathbf{H}^H \mathbf{H} \to \text{diag}\{ \frac{1}{L_1} \sum_{\ell=1}^{L_1} |g_{\ell,1}|^2, \cdots, \frac{1}{L_{N_u}} \sum_{\ell=1}^{L_{N_u}} |g_{\ell,N_u}|^2 \}$ [98] by the channel model in (3.1) without the pathloss component and the law of large numbers, which implies

$$\frac{\lambda_i}{N_r} \to \frac{1}{L_i} \sum_{\ell=1}^{L_i} |g_{\ell,i}|^2 < \infty, \text{ for } i = 1, \cdots, N_u.$$

This completes the proof of Theorem 2. ■

I note from (3.14) that $\mathbf{W}_{\text{RF}_1}^{\star}$ of the two-stage analog combining solution $\mathbf{W}_{\text{RF}}^{\star}$ aggregates all channel gains into the smaller dimension and provides $(N_{\text{RF}} - N_u)$ extra dimensions. Then, as observed in (3.15), $\mathbf{W}_{\text{RF}_2}^{\star}$ spreads the aggregated channels gains over all N_{RF} dimensions, which reduces the quantization error by exploiting the extra dimensions. Accordingly, as the number of

RF chains N_{RF} increases, the proposed solution $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ achieves the optimal scaling law (3.8) by reducing the quantization error.

Corollary 3. *The conventional optimal solution* $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}} = [\mathbf{U}_{1:N_u}\ \mathbf{U}_{\perp}]$ *for perfect quantization systems cannot achieve the optimal scaling law* (3.8) *in coarse quantization systems, and it is upper bounded by*

$$\mathcal{C}\big(\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}\big) < \mathcal{C}_{\mathrm{svd}}^{\mathrm{ub}} = N_u \log_2\left(1 + \frac{\alpha_b}{1-\alpha_b}\right). \tag{3.17}$$

Proof. From (3.14), we have the following MI by setting $\mathbf{W}_{\mathrm{RF}_2} = \mathbf{I}$:

$$\begin{aligned}
\mathcal{C}\big(\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}\big) &= \log_2\left|\mathbf{I} + \frac{\alpha_b}{\beta_b}\mathrm{diag}^{-1}\left\{\mathbf{\Lambda}_{N_{\mathrm{RF}}} + \tfrac{1}{\beta_b\rho}\mathbf{I}\right\}\mathbf{\Lambda}_{N_{\mathrm{RF}}}\right| \\
&= \sum_{i=1}^{N_u}\log_2\left(1 + \frac{\alpha_b\lambda_i}{\beta_b\lambda_i + 1/\rho}\right) \\
&\overset{(a)}{<} N_u \log_2\left(1 + \frac{\alpha_b}{\beta_b}\right).
\end{aligned}$$

where (a) comes from $\rho > 0$. ■

Corollary 3 shows that the conventional unconstrained optimal analog combiner $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$ can capture all channel gains but the MI does not scale as that of $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$. Since all channel gains after processed through $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$ are concentrated on only N_u RF chains out of N_{RF} RF chains, using $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$ results in severe quantization errors at each of the N_u RF chains. Although the channel gains $\{\lambda_i\}$ increase as N_r increases, the quantization errors also increase in proportion to the channel gains for $\mathcal{C}\big(\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}\big)$, yielding only the bounded MI in (3.17).

Again, unlike the conventional solution, the additional second stage analog combiner $\mathbf{W}_{\mathrm{RF}_2}^{\star}$ proposed in Theorem 2 spreads the channel gains captured by the first stage combiner $\mathbf{W}_{\mathrm{RF}_1}^{\star}$ to all N_{RF} RF chains evenly, leading to achieving the optimal scaling law by greatly alleviating quantization errors. Intuitively, adopting the second combiner $\mathbf{W}_{\mathrm{RF}_2}^{\star}$ results in distributing the burden of ADCs confined in few RF chains over all available ADCs of the total RF chains. Later, I show that such performance gain from adopting the two-stage analog combining structure can be significant even with a reasonable number of RF chains.

Theorem 3. *For the case of homogeneous singular values of* $\mathbf{H}^H\mathbf{H}$ *where all singular values* $\{\lambda_i\}$ *are equal, the two-stage analog combining solution* $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ *in Theorem 2 maximizes the MI in* (3.7) *with finite* N_{RF}*, i.e.,*

$$\mathbf{W}_{\mathrm{RF}}^{\star} = \arg\max_{\mathbf{W}_{\mathrm{RF}}} \mathcal{C}(\mathbf{W}_{\mathrm{RF}})$$
$$s.t.\ \mathbf{W}_{\mathrm{RF}}^H\mathbf{W}_{\mathrm{RF}} = \mathbf{I}_{N_{\mathrm{RF}}}\ \textit{and}\ \lambda_1 = \cdots = \lambda_{N_u} = \lambda,$$

and the corresponding optimal MI is given as

$$\mathcal{C}_{\mathrm{opt}} \triangleq \mathcal{C}(\mathbf{W}_{\mathrm{RF}}^{\star}) = N_u \log_2\left(1 + \frac{\alpha_b \lambda N_{\mathrm{RF}}}{\lambda N_u(1-\alpha_b) + N_{\mathrm{RF}}/\rho}\right). \tag{3.18}$$

Proof. Recall $\mathbf{G} = \overline{\mathbf{W}}_{\mathrm{RF}}^H \bar{\mathbf{\Lambda}}^{1/2} = [\mathbf{G}_{\mathrm{sub}}\ \mathbf{0}]$ in the proof of Theorem 2, where $\mathbf{G}_{\mathrm{sub}}$ is the $N_{\mathrm{RF}} \times m$ submatrix of $\mathbf{G}$ and $\bar{\mathbf{\Lambda}} = \mathrm{diag}\{\bar{\lambda}_1, \cdots, \bar{\lambda}_m, 0, \cdots, 0\}$ is the diagonal matrix composed of the singular values of $\mathbf{Q}$, defined in (3.10).

From the assumption of $\lambda_1 = \cdots = \lambda_{N_u} = \lambda$, we have

$$\begin{aligned}\max_{\mathbf{x}\in\mathbb{C}^{N_{\mathrm{RF}}}:\|\mathbf{x}\|=1} \mathbf{x}^H\mathbf{Q}\mathbf{x} &= \max_{\mathbf{y}\in\mathbb{C}^{m}:\|\mathbf{y}\|=1} \lambda\|\mathbf{U}_{1:N_u}^H\mathbf{U}_{||}\mathbf{y}\|^2 \\ &\overset{(a)}{\leq} \max_{\mathbf{y}\in\mathbb{C}^{m}:\|\mathbf{y}\|=1} \lambda\|\mathbf{U}_{1:N_u}^H\|^2\|\mathbf{U}_{||}\|^2\|\mathbf{y}\|^2 \\ &= \lambda,\end{aligned}$$

where (a) comes from the sub-multiplicativity of the norm, and the last equality holds by $\|\mathbf{U}_{1:N_u}^H\| = 1$ and $\|\mathbf{U}_{||}\| = 1$. This implies the singular values of $\mathbf{Q}$ are bounded as $\bar{\lambda}_i \leq \lambda$ for $i = 1, \cdots, m$. Hence, $\|[\mathbf{G}_{\mathrm{sub}}]_{j,:}\|^2$ is maximized for any given $\overline{\mathbf{W}}_{\mathrm{RF}}$ when $\bar{\lambda}_i$ achieves λ for all $i = 1, \cdots, m$.

I consider the upper bound of $\mathcal{C}(\mathbf{W}_{\mathrm{RF}})$ in (3.13) and define

$$\mathbf{G}_{\mathrm{sub}}^{\star} = \overline{\mathbf{W}}_{\mathrm{RF}}^H \begin{bmatrix} \sqrt{\lambda}\mathbf{I}_m \\ \mathbf{0} \end{bmatrix}.$$

Then, (3.13) is further upper bounded as

$$\begin{aligned}\mathcal{C}(\mathbf{W}_{\mathrm{RF}}) &\leq m\log_2\left(1 + \frac{\alpha_b}{\beta_b m}\sum_{i=1}^{N_{\mathrm{RF}}}\frac{\|[\mathbf{G}_{\mathrm{sub}}^{\star}]_{i,:}\|^2}{\|[\mathbf{G}_{\mathrm{sub}}^{\star}]_{i,:}\|^2 + \frac{1}{\beta_b\rho}}\right) \\ &\overset{(a)}{\leq} m\log_2\left(1 + \frac{\alpha_b N_{\mathrm{RF}}}{\beta_b m}\frac{\sum_{i=1}^{N_{\mathrm{RF}}}\|[\mathbf{G}_{\mathrm{sub}}^{\star}]_{i,:}\|^2}{\sum_{i=1}^{N_{\mathrm{RF}}}\|[\mathbf{G}_{\mathrm{sub}}^{\star}]_{i,:}\|^2 + \frac{N_{\mathrm{RF}}}{\beta_b\rho}}\right) \\ &\overset{(b)}{=} m\log_2\left(1 + \frac{\alpha_b\lambda N_{\mathrm{RF}}}{\lambda m\beta_b + N_{\mathrm{RF}}/\rho}\right), \qquad (3.19)\end{aligned}$$

where (a) holds by Jensen's inequality and the concavity of $\frac{x}{x+1}$ for $x > 0$; and (b) comes from $\sum_{i=1}^{N_{\mathrm{RF}}}\|[\mathbf{G}_{\mathrm{sub}}^{\star}]_{i,:}\|^2 = \|\mathbf{G}_{\mathrm{sub}}^{\star}\|_F^2 = \lambda m$. Note that (3.19) is maximized when $m = N_u$ since the derivative of (3.19) with respect to m is positive for $m > 0$ for any given $\alpha_b, \lambda, \rho, N_{\mathrm{RF}} > 0$. By substituting $\lambda_1 = \cdots =$

$\lambda_{N_u} = \lambda$ into (3.16), it can be shown that the upper bound of $\mathcal{C}(\mathbf{W}_{\mathrm{RF}})$ in (3.19) with $m = N_u$ can be achieved by adopting $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$. This completes the proof of Theorem 3. ■

Theorem 3 shows the optimality of the proposed two-stage analog combining solution $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ in maximizing the MI for any number of RF chains $N_{\mathrm{RF}} \geq N_u$ with homogeneous singular values. Note that such optimality of $\mathbf{W}_{\mathrm{RF}}^{\star}$ can be nearly achieved for a fixed number of users in large-scale MIMO systems as shown in Remark 6.

Remark 6. *From Theorem 3, the two-stage analog combining solution* $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ *in Theorem 2 maximizes the MI for* $\mathcal{P}1$ *as well as achieves the optimal scaling law* (3.8) *in homogeneous massive MIMO networks with a large number of antennas* N_r*, where each channel elemen* $h_{ij} \overset{i.i.d.}{\sim} \mathcal{CN}(0,1)$*. This is because as the number of receive antennas* N_r *increases,* $\frac{1}{N_r}\mathbf{H}^H\mathbf{H} \to \mathbf{I}_{N_u}$*, i.e,* $\frac{1}{N_r}\lambda_i \to 1$*,* $\forall i$ *[99].*

Fig. 3.2 shows the simulation results of the MI of the proposed two-stage analog combiner $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ in Theorem 2 and the conventional analog combiner $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$ in Corollary 3 which is optimal for infinite-resolution ADC systems. Here, I use $\mathbf{W}_{\mathrm{RF}_1}^{\star} = \mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}} = \mathbf{U}_{1:N_{\mathrm{RF}}}$ and $\mathbf{W}_{\mathrm{RF}_2}^{\star} = \mathbf{W}_{\mathrm{DFT}}$, where $\mathbf{W}_{\mathrm{DFT}}$ is an $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ normalized DFT matrix, and consider Rayleigh MIMO channels described in Remark 6. As shown in Fig. 3.2(a), the MI of the proposed two-stage analog combiner almost achieves the optimal MI $\mathcal{C}_{\mathrm{opt}}$ (3.18) in Theorem 3 with $\lambda/N_r = 1$ even in the regime of a finite N_r. I

further note that compared with the conventional one-stage combiner $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$ converging to the upper limit $\mathcal{C}_{\mathrm{svd}}^{ub}$, the MI of the two-stage analog combiner logarithmically increases without a limit as N_r increases with $\kappa \approx 1/3$. This follows the optimal scaling law in Theorem 2.

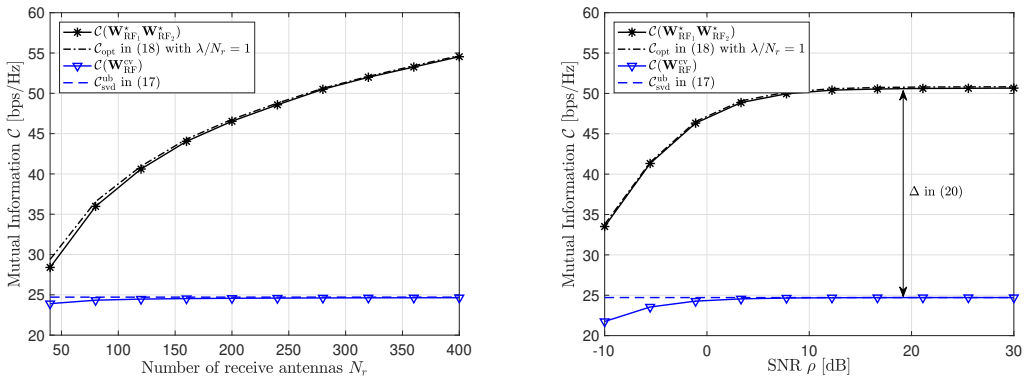
Fig. 3.2(b) shows the MI simulation results with respect to the SNR ρ. The two-stage combiner $\mathbf{W}_{\mathrm{RF}}^{\star} = \mathbf{W}_{\mathrm{RF}_1}^{\star}\mathbf{W}_{\mathrm{RF}_2}^{\star}$ yields superior MI performance to that of $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$, and the MI of $\mathbf{W}_{\mathrm{RF}}^{\star}$ converges to $N_u \log_2\left(1+\frac{\alpha_b N_{\mathrm{RF}}}{(1-\alpha_b)N_u}\right)$, which is obtained from $\mathcal{C}_{\mathrm{opt}}$ (3.18) with $\rho \to \infty$. Therefore, the MI gap between the upper limits of the two combiners $(\mathbf{W}_{\mathrm{RF}}^{\star}, \mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}})$ is

$$\Delta = N_u\left(\log_2\left(1+\frac{\alpha_b N_{\mathrm{RF}}}{(1-\alpha_b)N_u}\right) - \log_2\left(1+\frac{\alpha_b}{1-\alpha_b}\right)\right). \tag{3.20}$$

Since $N_{\mathrm{RF}} \geq N_u$ is considered in this chapter, the proposed two-stage combiner $\mathbf{W}_{\mathrm{RF}}^{\star}$ always yields the higher upper limit of the MI than the SVD-based one-stage combiner $\mathbf{W}_{\mathrm{RF}}^{\mathrm{cv}}$.

3.4 Two-Stage Analog Combining Algorithm

In the previous section, I derived the analog combining solution for the unconstrained problem $\mathcal{P}1$. However, the constant modulus constraint on each matrix element should be taken into account in designing analog combiners since it is implemented using phase shifters. I further consider a pre-defined set of phases with a finite cardinality for phase shifters. Considering channels known at the receiver, I propose a codebook-based two-stage analog combining algorithm for mmWave communications.



(a) $\mathcal{C}$ vs. N_r (b) $\mathcal{C}$ vs. ρ

Figure 3.2: The simulation results of the MI with the proposed two-stage analog combining solution $\mathbf{W}^{\star}_{\mathrm{RF}_1}\mathbf{W}^{\star}_{\mathrm{RF}_2}$ and the conventional unconstrained optimal analog combiner $\mathbf{W}^{\mathrm{cv}}_{\mathrm{RF}}$ in the Rayleigh MIMO channels: (a) for $(\rho, N_{\mathrm{RF}}, N_u, b) = (5\ \mathrm{dB}, \lceil \frac{N_r}{3} \rceil, 8, 2)$ as N_r increases, and (b) for $(N_r, N_{\mathrm{RF}}, N_u, b) = (256, \lceil \frac{N_r}{3} \rceil, 8, 2)$ as ρ increases where b denotes the number of quantization bits.

3.4.1 Proposed Two-Stage Analog Combining Algorithm

Theorem 2 provides a practical analog combiner structure that is implementable with a two-stage analog combiner $\mathbf{W}_{\mathrm{RF}} = \mathbf{W}_{\mathrm{RF}_1}\mathbf{W}_{\mathrm{RF}_2}$: the first analog combiner and the second analog combiner can be considered as a channel gain aggregation matrix and spreading matrix, respectively. Leveraging such insight and the finding in the following Corollary 4, I propose an ARV-based two-stage analog combining (ARV-TSAC) algorithm for mmWave channels.

Corollary 4. *When the sum of all channel paths from each user* $\sum_{k=1}^{N_u} L_k$ *is a finite value and the AoA of each path is different than that of other paths, the optimal scaling in* (3.8) *can be achieved by using* $\tilde{\mathbf{W}}^{\star}_{\mathrm{RF}} = \mathbf{W}_{\mathrm{AoA}}\mathbf{W}^{\star}_{\mathrm{RF}_2}$ *as* $N_r \to \infty$ *for fixed* $\kappa \in (0, 1)$*, where* $\mathbf{W}_{\mathrm{AoA}} = [\mathbf{A}_{\mathrm{AoA}}, \mathbf{A}^{\perp}_{\mathrm{AoA}}]$*,* $\mathbf{A}_{\mathrm{AoA}} =$

$[\mathbf{a}(\phi_{1,1}), \mathbf{a}(\phi_{2,1}), \cdots, \mathbf{a}(\phi_{L_{N_u},N_u})]$, *and* $\mathbf{A}_{\mathrm{AoA}}^{\perp}$ *is an* $N_r \times (N_{\mathrm{RF}} - \sum_{k=1}^{N_u} L_k)$ *matrix composed of orthonormal basis vectors whose column space is in* $\mathrm{Span}^{\perp}(\mathbf{A}_{\mathrm{AoA}})$.

Proof. See Section 3.7. ■

Note that a conventional optimal solution can be obtained using $\tilde{\mathbf{W}}_{\mathrm{RF}}^{\mathrm{cv}\star} = \mathbf{W}_{\mathrm{AoA}}$ under perfect quantization [100, 101] as $N_r \to \infty$, which does not require the second spreading combiner $\mathbf{W}_{\mathrm{RF}_2}^{\star}$ but the sum rate will be bounded as shown in Corollary 3. According to Corollary 4, using ARVs provides a fair tradeoff between practicality in implementation and performance. To design the first analog combiner $\mathbf{W}_{\mathrm{RF}_1}$, I adopt an ARV-codebook based maximum channel gain aggregation approach to collect most channel gains into the lower signal dimension by exploiting the sparse nature of mmWave channels. I set the codebook of the evenly spaced spatial angles $\mathcal{V} = \{\vartheta_1, \ldots, \vartheta_{|\mathcal{V}|}\}$. Since selecting N_{RF} ARVs out of the total $|\mathcal{V}|$ ARVs in the codebook requires $\binom{|\mathcal{V}|}{N_{\mathrm{RF}}}$ search complexity for the exhaustive method, I propose a greedy-based algorithm to find the best N_{RF} ARVs with greatly reduced complexity[4].

Algorithm 2 describes the proposed ARV-TSAC method. In Step (a), the ARV with the spatial angle $\vartheta^{\star}$ which captures the largest channel gain in the remaining channel dimensions $\mathbf{H}_{\mathrm{rm}}$ is selected and it composes a column of the first analog combiner in Step (b). In Step (c), the channel matrix on

[4]Selecting N_{RF} angles from the codebook $\mathcal{V}$ that are closest to the AoAs of channels can be an alternative approach for implementing the first analog combiner with low complexity in the ARV-TSAC algorithm when all AoAs of channels are available at the BS.

the remaining dimensions $\mathbf{H}_{\mathrm{rm}}$ is projected onto the subspace of $\mathrm{Span}^{\perp}(\mathbf{a}(\vartheta^{\star}))$ to remove the channel gain on the space of the selected ARV. Algorithm 2 repeats these steps until N_{RF} ARVs are selected from the codebook $\mathcal{V}$. Note that Algorithm 2 nearly finds $\mathbf{U}_{1:N_u}$ in the first N_u iterations and $\mathbf{U}_{\perp}$ in the remaining $(N_{\mathrm{RF}} - N_u)$ iterations for the first analog combiner. This is because the algorithm sequentially searches for the array response vectors that have the principal components of $\mathbf{H}$ with rank N_u.

Remark 7. *The second-stage analog combiner that satisfies the condition (ii) of Theorem 2 can be implemented by adopting a normalized $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ DFT matrix, i.e., $\mathbf{W}^{\star}_{\mathrm{RF}_2} = \mathbf{W}_{\mathrm{DFT}}$.*

Employing the DFT matrix for the second analog combiner $\mathbf{W}_{\mathrm{RF}_2} = \mathbf{W}_{\mathrm{DFT}}$ (or any unitary matrix with constant modulus) offers benefits in reducing implementation complexity and power consumption since $\mathbf{W}_{\mathrm{DFT}}$ does not depend on the channel $\mathbf{H}$ and can be constructed by using passive (or fixed) analog phase shifters. Accordingly, although the additional N_{RF}^2 fully-connected passive phase shifters for the second analog combiner add to the complexity of the proposed architecture in physical area and power consumption, it can be implemented with very low complexity and power consumption in the practical system. Furthermore, if N_{RF} is a power of two, the fast Fourier transform version of the DFT calculation can be implemented, which reduces the number of additional passive phase shifters to $N_{\mathrm{RF}} \log_2 N_{\mathrm{RF}}$.

Passive phase shifters consume negligible power compared to active phase shifters, and advanced passive phase shifters were designed to increase

the accuracy and minimize the power attenuation [102–104]. To implement the DFT matrix for the second analog combiner $\mathbf{W}_{\mathrm{RF}_2}$, a Butler matrix can be used, which can further reduce the cost and complexity [105]. When the number of RF chains N_{RF} is a power of two, a Hadamard matrix that is composed of 1s and -1s can be adopted for the second combiner, which only requires $(N_{\mathrm{RF}}^2 - N_{\mathrm{RF}})/2$ passive phase shifters with 180° phase shift. Accordingly, deploying the Hadamard matrix for the second analog combiner $\mathbf{W}_{\mathrm{RF}_2}$ would require lower implementation cost and complexity than using the DFT matrix, and it can also be implemented with passive phase shifters only.

Although the first stage analog combiner designed by the proposed ARV-TSAC algorithm is similar to existing one-stage analog combiner designs, the primary contributions of this work are threefold. The first is to propose a new two-stage analog combining architecture for hybrid MIMO receivers with a reduced number of RF chains having low-resolution ADCs. The second is to derive a near optimal unconstrained two-stage analog combining solution for the proposed architecture and show the theoretical performance gap between the proposed two-stage architecture and the conventional SVD-based combining architecture in low-resolution ADC systems. Finally, I further provide the theoretical performance analysis of the proposed two-stage analog combining architecture with the developed ARV-TSAC algorithm in the next subsection.

Algorithm 2: ARV-based TSAC

Initialization: set $\mathbf{W}_{\mathrm{RF}_1}$ = empty matrix, $\mathbf{H}_{\mathrm{rm}} = \mathbf{H}$, and $\mathcal{V} = \{\vartheta_1, \ldots, \vartheta_{|\mathcal{V}|}\}$ where $\vartheta_n = \frac{2n}{|\mathcal{V}|} - 1$

for $i = 1 : N_{\mathrm{RF}}$ **do**

Maximum channel gain aggregation

(a) $\mathbf{a}(\vartheta^\star) = \mathrm{argmax}_{\vartheta \in \mathcal{V}} \|\mathbf{a}(\vartheta)^H \mathbf{H}_{\mathrm{rm}}\|^2$

(b) $\mathbf{W}_{\mathrm{RF}_1} = \left[\mathbf{W}_{\mathrm{RF}_1} \mid \mathbf{a}(\vartheta^\star) \right]$

(c) $\mathbf{H}_{\mathrm{rm}} = \mathcal{P}^{\perp}_{\mathbf{a}(\vartheta^\star)} \mathbf{H}_{\mathrm{rm}}$, where $\mathcal{P}^{\perp}_{\mathbf{a}(\vartheta)} = \mathbf{I} - \mathbf{a}(\vartheta)\mathbf{a}(\vartheta)^H$

(d) $\mathcal{V} = \mathcal{V} \setminus \{\vartheta^\star\}$

Set $\mathbf{W}_{\mathrm{RF}_2} = \mathbf{W}_{\mathrm{DFT}}$ where $\mathbf{W}_{\mathrm{DFT}}$ is a normalized $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ DFT matrix.

return $\mathbf{W}_{\mathrm{RF}_1}$ and $\mathbf{W}_{\mathrm{RF}_2}$;

3.4.2 Performance Analysis

The ergodic sum rate of the ARV-TSAC algorithm with an MRC baseband combiner is analyzed. Once I derive the closed-form ergodic rate, I compare the rate with the one without the second analog combiner $\mathbf{W}_{\mathrm{RF}_2}$ to quantify the ergodic rate gain from employing $\mathbf{W}_{\mathrm{RF}_2}$. To this end, a virtual channel representation [71] is adopted for analytic tractability which captures the sparse property of mmWave channels [23, 69]. Under the virtual channel representation, the channel vector $\mathbf{h}_k$ in (3.1) can be modeled as

$$\mathbf{h}_k = \sqrt{\frac{N_r}{L_k}} \mathbf{A} \tilde{\mathbf{g}}_k = \mathbf{A} \tilde{\mathbf{h}}_{\mathrm{b},k}$$

where $\tilde{\mathbf{h}}_{\mathrm{b},k} = \sqrt{\frac{N_r}{L_k}} \tilde{\mathbf{g}}_k$ is the L_k-sparse beamspace channel of user k, i.e., $\tilde{\mathbf{g}}_k$ has L_k nonzero entries $\overset{i.i.d.}{\sim} \mathcal{CN}(0, 1)$, and $\mathbf{A} = [\mathbf{a}(\varphi_1), \ldots, \mathbf{a}(\varphi_{N_r})]$ with uniformly

spaced spatial angles φ_i.

Under this representation, I consider the case where the codebook size of Algorithm 2 is equal to the number of antennas $|\mathcal{V}| = N_r$. Accordingly, the first analog combiner is the $N_r \times N_{\mathrm{RF}}$ submatrix of $\mathbf{A}$ which captures the most channel gain, $\mathbf{W}_{\mathrm{RF}_1} = \mathbf{A}_{\mathrm{sub}}$. It is assumed that $\mathbf{W}_{\mathrm{RF}_1}$ captures all channel propagation paths from N_u users [72, 90], i.e., L_k channels paths for each user fall within N_{RF} RF chains. For simplicity, I further assume $L_k = L$, $\forall k$, in the analysis[5]. Thus, after combining with $\mathbf{W}_{\mathrm{RF}_1} = \mathbf{A}_{\mathrm{sub}}$, the channel becomes $\mathbf{H}_{\mathrm{b}} = \mathbf{W}_{\mathrm{RF}_1}^H \mathbf{H}$, and the channel vector of user k with the reduced dimension $\mathbf{h}_{\mathrm{b},k} \in \mathbb{C}^{N_{\mathrm{RF}}}$ is

$$\mathbf{h}_{\mathrm{b},k} = \sqrt{\frac{N_r}{L}} \mathbf{g}_k. \tag{3.21}$$

I consider L nonzero channel gains to be uniformly distributed within each user channel $\mathbf{h}_{\mathrm{b},k}$ and use an indicator function $\mathbb{1}_{\{i \in \mathcal{A}\}}$ to characterize the channel sparsity where $\mathbb{1}_{\{i \in \mathcal{A}\}} = 1$ if $i \in \mathcal{A}$, and $\mathbb{1}_{\{i \in \mathcal{A}\}} = 0$ otherwise. Utilizing $\mathbb{1}_{\{.\}}$, the ℓth complex path gain of user k is modeled as

$$g_{\ell,k} = \xi_{\ell,k} \mathbb{1}_{\{\ell \in \mathcal{P}_k\}}, \quad \ell = 1, \cdots, N_{\mathrm{RF}}, \ k = 1, \cdots, N_u$$

where $\xi_{\ell,k} \overset{i.i.d.}{\sim} \mathcal{CN}(0,1)$, $\forall \ell, k$ and $\mathcal{P}_k = \left\{ i \,\middle|\, g_{i,k} \neq 0, i = 1, \cdots, N_{\mathrm{RF}} \right\}$ is the nonzero index set.

I consider the MRC combiner $\mathbf{W}_{\mathrm{BB}} = \bar{\mathbf{H}}_{\mathrm{b}}$ where $\bar{\mathbf{H}}_{\mathrm{b}} = \mathbf{W}_{\mathrm{RF}_2}^H \mathbf{W}_{\mathrm{RF}_1}^H \mathbf{H}$,

[5]The similar results can be derived with minor changes for general L_k.

and the received signal k in (3.5) becomes

$$z_k = \alpha_b\sqrt{\rho}\bar{\mathbf{h}}_{\mathrm{b},k}^H\bar{\mathbf{h}}_{\mathrm{b},k}s_k + \alpha_b\sqrt{\rho}\sum_{i\neq k}^{N_u}\bar{\mathbf{h}}_{\mathrm{b},k}^H\bar{\mathbf{h}}_{\mathrm{b},i}s_i + \alpha_b\bar{\mathbf{h}}_{\mathrm{b},k}^H\mathbf{W}_{\mathrm{RF}}^H\mathbf{n} + \bar{\mathbf{h}}_{\mathrm{b},k}^H\mathbf{q}. \tag{3.22}$$

From (3.22), the achievable rate of the proposed system for the MRC combiner is given as [106, 107]

$$r_k^{\mathrm{mrc}} = \log_2\left(1 + \frac{\rho\alpha_b\|\bar{\mathbf{h}}_{\mathrm{b},k}\|^4}{\rho\alpha_b\sum_{i\neq k}^{N_u}|\bar{\mathbf{h}}_{\mathrm{b},k}^H\bar{\mathbf{h}}_{\mathrm{b},i}|^2 + \|\bar{\mathbf{h}}_{\mathrm{b},k}\|^2 + \rho\beta_b\Psi_k}\right) \tag{3.23}$$

where $\Psi_k = \bar{\mathbf{h}}_{\mathrm{b},k}^H\mathrm{diag}\{\bar{\mathbf{H}}_{\mathrm{b}}\bar{\mathbf{H}}_{\mathrm{b}}^H\}\bar{\mathbf{h}}_{\mathrm{b},k}$, and the ergodic rate is

$$\begin{aligned}\bar{r}_k^{\mathrm{mrc}} &= \mathbb{E}\left[r_k^{\mathrm{mrc}}\right] \\ &= \mathbb{E}\left[\log_2\left(1 + \frac{\rho\alpha_b\|\bar{\mathbf{h}}_{\mathrm{b},k}\|^4}{\rho\alpha_b\sum_{i\neq k}^{N_u}|\bar{\mathbf{h}}_{\mathrm{b},k}^H\bar{\mathbf{h}}_{\mathrm{b},i}|^2 + \|\bar{\mathbf{h}}_{\mathrm{b},k}\|^2 + \rho\beta_b\Psi_k}\right)\right].\end{aligned} \tag{3.24}$$

Since $\mathbf{W}_{\mathrm{RF}_2} = \mathbf{W}_{\mathrm{DFT}}$ is unitary, I have $\|\bar{\mathbf{h}}_{\mathrm{b},i}^H\bar{\mathbf{h}}_{\mathrm{b},j}\| = \|\mathbf{h}_{\mathrm{b},i}^H\mathbf{h}_{\mathrm{b},j}\|$, $\forall i, j$. The ergodic rate (3.24) is approximated as

$$\begin{aligned}\bar{r}_k^{\mathrm{mrc}} &= \mathbb{E}\left[\log_2\left(1 + \frac{\rho\alpha_b\|\mathbf{h}_{\mathrm{b},k}\|^4}{\rho\alpha_b\sum_{i\neq k}^{N_u}|\mathbf{h}_{\mathrm{b},k}^H\mathbf{h}_{\mathrm{b},i}|^2 + \|\mathbf{h}_{\mathrm{b},k}\|^2 + \rho\beta_b\Psi_k}\right)\right] \\ &\overset{(a)}{\approx} \log_2\left(1 + \frac{\rho\alpha_b\mathbb{E}\left[\|\mathbf{h}_{\mathrm{b},k}\|^4\right]}{\rho\alpha_b\sum_{i\neq k}^{N_u}\mathbb{E}\left[|\mathbf{h}_{\mathrm{b},k}^H\mathbf{h}_{\mathrm{b},i}|^2\right] + \mathbb{E}\left[\|\mathbf{h}_{\mathrm{b},k}\|^2\right] + \rho\beta_b\mathbb{E}\left[\Psi_k\right]}\right)\end{aligned} \tag{3.25}$$

where (a) follows from Lemma 1 in [79]. Note that the approximation becomes more accurate as the number of receive antennas N_r increases for the non-sparse channel environment [79]. However, this also holds for mmWave channels. As the number of receive antennas N_r increases, the resolution of beamformer also increases, thereby increasing the number of major channel elements. Consequently, although the rate of increase of the number of the

effective channel elements may be slower than $O(N_r)$, I can consider the number of effective channel elements increases as the number of receive antennas increases.

I first analyze the average quantization error with two-stage analog combining and MRC $\mathbb{E}[\Psi_k]$ in (3.25). Noting that

$$\Psi_k = \mathbf{h}_{\mathrm{b},k}^H \mathbf{W}_{\mathrm{DFT}} \mathrm{diag}\big\{\mathbf{W}_{\mathrm{DFT}}^H \mathbf{H}_{\mathrm{b}} \mathbf{H}_{\mathrm{b}}^H \mathbf{W}_{\mathrm{DFT}}\big\} \mathbf{W}_{\mathrm{DFT}}^H \mathbf{h}_{\mathrm{b},k},$$

I decompose $\mathbb{E}[\Psi_k]$ as $\mathbb{E}[\Psi_k] = \mathbb{E}[\Psi_k^{\mathrm{auto}}] + \mathbb{E}[\Psi_k^{\mathrm{cross}}]$, and define the auto quantization noise and cross quantization noise variances as

$$\mathbb{E}\big[\Psi_k^{\mathrm{auto}}\big] = \mathbb{E}\Big[\mathbf{h}_{\mathrm{b},k}^H \mathbf{W}_{\mathrm{DFT}} \mathrm{diag}\big\{\mathbf{W}_{\mathrm{DFT}}^H \mathbf{h}_{\mathrm{b},k} \mathbf{h}_{\mathrm{b},k}^H \mathbf{W}_{\mathrm{DFT}}\big\} \mathbf{W}_{\mathrm{DFT}}^H \mathbf{h}_{\mathrm{b},k}\Big], \tag{3.26}$$

$$\mathbb{E}\big[\Psi_k^{\mathrm{cross}}\big] = \mathbb{E}\Big[\mathbf{h}_{\mathrm{b},k}^H \mathbf{W}_{\mathrm{DFT}} \mathrm{diag}\big\{\mathbf{W}_{\mathrm{DFT}}^H \mathbf{H}_{\mathrm{b}\backslash k} \mathbf{H}_{\mathrm{b}\backslash k}^H \mathbf{W}_{\mathrm{DFT}}\big\} \mathbf{W}_{\mathrm{DFT}}^H \mathbf{h}_{\mathrm{b},k}\Big] \tag{3.27}$$

where $\mathbf{H}_{\mathrm{b}\backslash k}$ denotes the channel matrix $\mathbf{H}_{\mathrm{b}}$ without its kth column. Then, (3.26) and (3.27) represent the average quantization errors for the associated user caused by the associated user itself and other users, respectively.

Lemma 3. *For the considered mmWave channel, the auto quantization noise variance for the two-stage analog combining of the ARV-TSAC algorithm with MRC* (3.26) *is derived as*

$$\mathbb{E}\big[\Psi_k^{\mathrm{auto}}\big] = \frac{2N_r^2}{N_{\mathrm{RF}}}. \tag{3.28}$$

Proof. See Section 3.8. ■

Note that the quantization noise variance decreases as the number of RF chains N_{RF} increases, which corresponds to the following intuition: the

second DFT analog combiner spreads the aggregated signal power at each RF chain over the N_{RF} chains and thus decreases the quantization error more as N_{RF} increases.

Lemma 4. *For the considered mmWave channel, the cross quantization noise variance for the two-stage analog combining of the ARV-TSAC algorithm with MRC* (3.27) *is derived as*

$$\mathbb{E}\left[\Psi_k^{\mathrm{cross}}\right] = \frac{N_r^2(N_u-1)}{N_{\mathrm{RF}}}. \tag{3.29}$$

Proof. See Section 3.9. ■

Since both $\mathbb{E}\left[\Psi_k^{\mathrm{auto}}\right]$ and $\mathbb{E}\left[\Psi_k^{\mathrm{cross}}\right]$ decrease with N_{RF}, the quantization error with the proposed two-stage analog combining and MRC combining is expected to decrease as N_{RF} increases, leading the ergodic rate to the same scaling law as in (3.8). I derive the approximated ergodic sum rate of (3.23) in closed form and validate the insight.

Theorem 4. *For the considered mmWave channel with low-resolution ADCs, the ergodic sum rate of the ARV-based TSAC method with MRC is approximated as*

$$\bar{\mathcal{R}}^{\mathrm{mrc}} \approx N_u \log_2\left(1+\frac{\rho\alpha_b N_r N_{\mathrm{RF}}(1+1/L)}{N_{\mathrm{RF}}+\rho N_r(N_u-1)+2\rho(1-\alpha_b)N_r}\right). \tag{3.30}$$

Proof. See Section 3.10. ■

Note that the derived ergodic rate in (3.30) is a function of system parameters and provides insights how the ergodic rate is improved with the proposed two-stage analog combining.

Remark 8. *Let* $\kappa = N_{\mathrm{RF}}/N_r$ *where* $\kappa \in (0,1)$ *is a constant value. Then,* (3.30) *can reduce to*

$$\bar{\mathcal{R}}^{\mathrm{mrc}} \approx N_u \log_2\left(1 + \frac{\rho\alpha_b N_{\mathrm{RF}}(1+1/L)}{\kappa + \rho(N_u-1) + 2\rho(1-\alpha_b)}\right). \tag{3.31}$$

The ergodic sum rate in (3.31) *achieves the optimal scaling law* $\sim N_u \log N_{\mathrm{RF}}$ *with respect to* N_{RF} *as in* (3.8).

Remark 8 shows that the optimal scaling law can be achieved by the proposed two-stage analog combining algorithm even with the practical baseband combiner. This result verifies that the two-stage analog combining architecture is effective to enhance the achievable rate in mmWave hybrid MIMO systems with low-resolution ADCs. To specify the effect of employing the second analog combiner $\mathbf{W}_{\mathrm{RF}_2}$, I also derive the ergodic rate (3.24) without using $\mathbf{W}_{\mathrm{RF}_2}$.

Corollary 5. *For the considered mmWave channel with low-resolution ADCs, the MRC ergodic rate of the ARV-TSAC without the second analog combiner is approximated as*

$$\bar{\mathcal{R}}^{\mathrm{mrc}}_{\mathrm{one}} \approx N_u \log_2\left(1 + \frac{\rho\alpha_b N_r N_{\mathrm{RF}}(1+1/L)}{N_{\mathrm{RF}} + \rho N_r(N_u-1) + 2\rho(1-\alpha_b)N_r N_{\mathrm{RF}}/L}\right). \tag{3.32}$$

Proof. See Section 3.11. ■

Unlike the quantization noise term $2\rho(1-\alpha_b)N_r$ in (3.30), that $2\rho(1-\alpha_b)N_rN_{\mathrm{RF}}/L$ in (3.32) includes N_{RF}/L, which prevents the optimal scaling of the ergodic sum rate as in (3.8) with respect to N_{RF} for fixed L.

Remark 9. *Let $\kappa = N_{\mathrm{RF}}/N_r$ where $\kappa \in (0,1)$ is a constant value. Then,* (3.32) *can reduce to*

$$\bar{\mathcal{R}}_{\mathrm{one}}^{\mathrm{mrc}} \approx N_u \log_2\left(1+\frac{\rho\alpha_b N_{\mathrm{RF}}(1+1/L)}{\kappa+\rho(N_u-1)+2\rho(1-\alpha_b)N_{\mathrm{RF}}/L}\right). \tag{3.33}$$

Note that unlike the ergodic rate of the two-stage analog combining $\bar{\mathcal{R}}^{\mathrm{mrc}}$ in (3.31)*, that of the one-stage analog combining $\bar{\mathcal{R}}_{\mathrm{one}}^{\mathrm{mrc}}$ in* (3.33) *cannot achieve the optimal scaling law with respect to the number of RF chains N_{RF}.*

As we show throughout this chapter, using an additional analog combiner provides noticeable improvement in the mutual information, and we could also use more analog combiners such as three stages or four stages. Adding more stages, however, would require additional implementation cost and complexity. The two-stage solution provides good results in both theory and simulation. Therefore, the two-stage analog combiner is considered to be the best when considering such penalty in increasing cost and complexity.

3.5 Simulation Results

In this section, the performance of the proposed two-stage analog combing algorithm is evaluated in the MI and ergodic sum rate. In the simulations, the codebook size is set to be $|\mathcal{V}| = N_r$, which guarantees $\mathbf{W}_{\mathrm{RF}}^H\mathbf{W}_{\mathrm{RF}} = \mathbf{I}_{N_{\mathrm{RF}}}$.

Consequently, analog combiners used in the simulations are semi-unitary. To provide a reference performance of a conventional one-stage analog combining approach, I simulate a greedy-based MI maximization method which solves the following problem for the given ARV codebook in a greedy way:

$$\mathcal{P}2: \qquad \mathbf{W}_{\mathrm{RF}}^{\mathrm{opt,c}} = \underset{\mathbf{W}_{\mathrm{RF}}}{\operatorname{argmax}} \ \mathcal{C}(\mathbf{W}_{\mathrm{RF}})$$
$$\text{s.t.} \quad \mathbf{W}_{\mathrm{RF}}^{H}\mathbf{W}_{\mathrm{RF}} = \mathbf{I}, \ \ |[\mathbf{W}_{\mathrm{RF}}]_{i,j}| = \frac{1}{\sqrt{N_r}}, \forall i, j.$$

At each iteration, the greedy method searches for a single ARV from the codebook $\mathcal{V}$ which maximizes the MI with the previously selected ARVs and thus can nearly provide the optimal MI performance of the one-stage analog combining for the given codebook.

In the simulations, the following cases are evaluated:

1. ARV-TSAC: proposed two-stage analog combining.
2. ARV: one-stage analog combining with $\mathbf{W}_{\mathrm{RF}} = \mathbf{W}_{\mathrm{RF}_1}$ selected from the ARV-TSAC.
3. SVD+DFT: two-stage analog combining with $\mathbf{W}_{\mathrm{RF}_1} = \mathbf{U}_{1:N_{\mathrm{RF}}}$ and $\mathbf{W}_{\mathrm{RF}_2} = \mathbf{W}_{\mathrm{DFT}}$ based on Theorem 2.
4. SVD: one-stage analog combining $\mathbf{W}_{\mathrm{RF}} = \mathbf{U}_{1:N_{\mathrm{RF}}}$,
5. Greedy-MI: one-stage analog combining with greedy-based MI maximization.

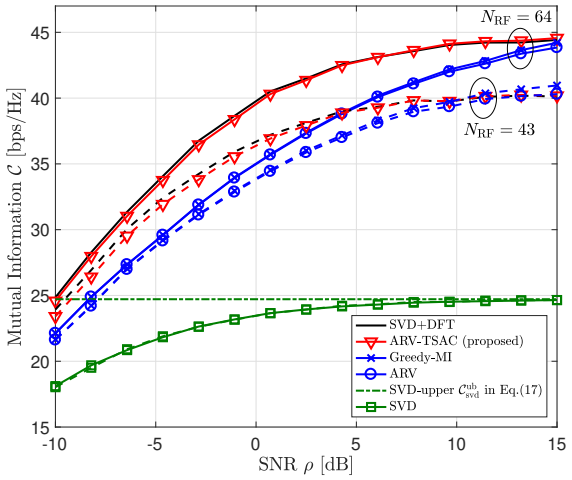
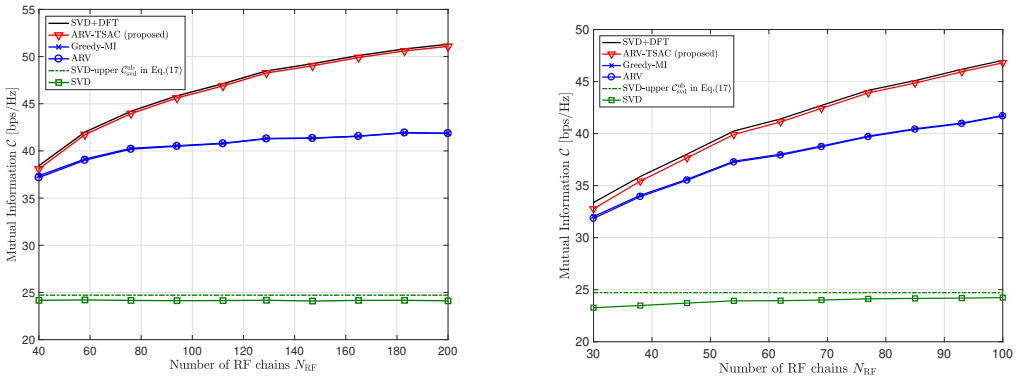


Figure 3.3: The MI simulation results for $N_r = 128$ receive antennas, $N_u = 8$ users, $\lambda_L = 3$ average channel paths, $b = 2$ quantization bits, and $N_{\mathrm{RF}} \in \{43, 64\}$ RF chains that are $\lceil N_r/3 \rceil$ and $\lceil N_r/2 \rceil$, respectively.

The SVD+DFT and SVD cases are infeasible in practice due to violating the constant modulus constraint, and SVD+DFT provides a tight upper bound on MI for a homogeneous singular value case from Theorem 3. Here, $L_k = \max\{1, \mathrm{Poisson}(\lambda_L)\}$ [70] is adopted unless mentioned otherwise, where λ_L is considered as the average number of channel paths.

3.5.1 Mutual Information

Fig. 3.3 shows the MI simulation results for $N_r = 128$, $N_{\mathrm{RF}} \in \{43, 64\}$, $N_u = 8$, $\lambda_L = 3$, and $b = 2$ with respect to the SNR ρ. The proposed ARV-TSAC algorithm achieves a similar MI as does the SVD+DFT case, and they show the best MI over the most SNR values. The Greedy-MI and ARV cases provide similar MI to each other but show the MI gap from the ARV-TSAC.

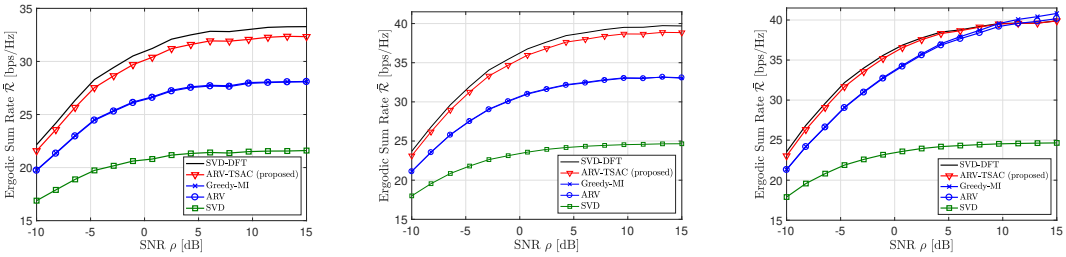


(a) $N_r = 256$ (b) $\kappa = 1/3$

Figure 3.4: The MI simulation results with $N_u = 8$ users, $\lambda_L = 4$ average channel paths, $b = 2$ quantization bits, and $\rho = 0$ dB SNR for (a) $N_r = 256$ receive antennas and (b) $\kappa = N_{\rm RF}/N_r = 1/3$.

The gap decreases as ρ increases in the high SNR regime, and the Greedy-MI and ARV cases with $N_{\rm RF} = 43$ show the higher MI than SVD+DFT and ARV-TSAC in the very high SNR regime. Such phenomenon occurs as the channel environment does not guarantee the optimality condition for the two-stage analog combining solution in Theorem 3. As more RF chains are used, however, the MI gap between ARV-TSAC/SVD+DFT and Greedy-MI/ARV becomes larger and the performance reversal would happen in even the higher SNR regime. This is because the proposed two-stage analog combining can exploit more RF chains to further reduce quantization errors. The SVD case results in the worst MI performance and it converges to the theoretic upper bound $\mathcal{C}_{\rm svd}^{\rm ub}$ due to the quantization error.

Fig. 3.4 shows the MI simulation results with $N_u = 8$, $\lambda_L = 4$, $b = 2$, and $\rho = 0$ dB in terms of $N_{\rm RF}$. In Fig. 3.4(a), N_r is fixed to be $N_r = 256$.



(a) MRC (b) ZF (c) MMSE

Figure 3.5: Perfect CSI simulation results of the ergodic sum rate with $N_r = 128$ receive antennas, $N_{\mathrm{RF}} = 43$ RF chains, $N_u = 8$ users, $\lambda_L = 3$ average channel paths, and $b = 2$ quantization bits for (a) maximum ratio combining (MRC), (b) zero-forcing (ZF), and (c) minimum mean squared error (MMSE) digital combiners.

The two-stage combining cases, i.e., SVD+DFT and ARV-TSAC, show that the MI increases logarithmically with N_{RF}, and this corresponds to the scaling law derived in Theorem 2. The one-stage combining cases such as the Greedy-MI, ARV, and SVD cases, however, show a marginal increase of the MI as N_{RF} increases. In Fig. 3.4(b), the ratio between N_r and N_{RF} is fixed to be $\kappa = 1/3$. Here, the Greedy-MI and ARV cases also increase more slowly compared to the SVD+DFT and ARV-TSAC cases. This is because more channel gains can be collected as N_r increases for all cases, but the two-stage combining can reduce more quantization error as N_{RF} increases. Accordingly, the MI gap between the two-stage combining and one-stage combining cases increases as N_{RF} increases.

3.5.2 Ergodic Sum Rate

Now, I evaluate the ergodic rate for linear digital combiners $\mathbf{W}_{\mathrm{BB}}$ such as MRC, zero-forcing (ZF), and MMSE. Let $\mathbf{H}_{\mathrm{eq}} = \mathbf{W}_{\mathrm{RF}}^H \mathbf{H}$. The MRC, ZF, and MMSE combiners are given as: $\mathbf{W}_{\mathrm{BB,mrc}} = \mathbf{H}_{\mathrm{eq}}$, $\mathbf{W}_{\mathrm{BB,zf}} = \mathbf{H}_{\mathrm{eq}}(\mathbf{H}_{\mathrm{eq}}^H \mathbf{H}_{\mathrm{eq}})^{-1}$, and $\mathbf{W}_{\mathrm{BB,mmse}} = \mathbf{R}_{\mathbf{y_q y_q}}^{-1} \mathbf{R}_{\mathbf{y_q x}}$, where $\mathbf{R}_{\mathbf{y_q x}} = \alpha\rho\mathbf{H}_{\mathrm{eq}}$ and $\mathbf{R}_{\mathbf{y_q y_q}} = \alpha^2\rho\mathbf{H}_{\mathrm{eq}}\mathbf{H}_{\mathrm{eq}}^H + \alpha^2\mathbf{W}_{\mathrm{RF}}^H\mathbf{W}_{\mathrm{RF}} + \mathbf{R}_{\mathbf{qq}}$. For the given analog and digital combiners $(\mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}})$ with $\mathbf{W}_{\mathrm{RF}}^H\mathbf{W}_{\mathrm{RF}} = \mathbf{I}_{N_{\mathrm{RF}}}$, the ergodic rate of user k is expressed as

$$\bar{r}_k(\mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}}) = \mathbb{E}\Big[\log_2\left(1 + \alpha_b^2\rho|\mathbf{w}_{\mathrm{BB},k}^H\mathbf{h}_{\mathrm{eq},k}|^2/\eta_{\mathrm{BB},k}\right)\Big]$$

where $\eta_{\mathrm{BB},k} = \alpha_b^2\rho\sum_{u\neq k}^{N_u}|\mathbf{w}_{\mathrm{BB},k}^H\mathbf{h}_{\mathrm{eq},u}|^2 + \alpha_b^2\|\mathbf{w}_{\mathrm{BB},k}\|^2 + \mathbf{w}_{\mathrm{BB},k}^H\mathbf{R}_{\mathbf{qq}}\mathbf{w}_{\mathrm{BB},k}$.

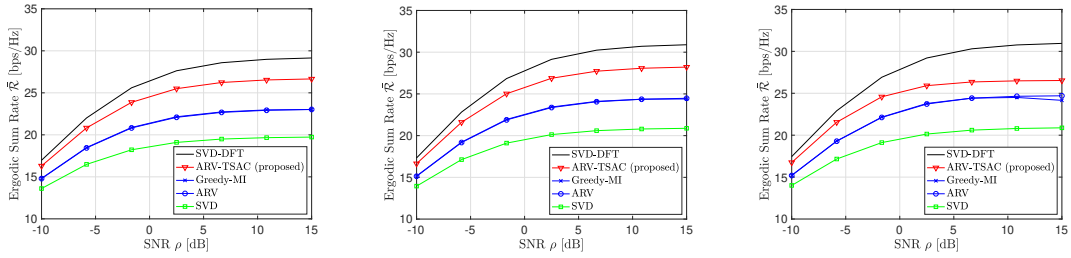
In addition to the perfect CSI case, I also consider the imperfect CSI case in which the estimated channel matrix at the receiver has channel estimation error at path coefficients and AoAs to provide a numerical study of the impact of the channel estimation error in the proposed system. I assume the estimated channel matrix as [108]

$$\tilde{\mathbf{h}}_k = \sqrt{\frac{N_r}{L_k}}\sum_{\ell=1}^{L_k}(g_{\ell,k} + g_{\ell,k}^e)\mathbf{a}(\phi_{\ell,k} + \phi_{\ell,k}^e)$$

where $g_{\ell,k} \overset{i.i.d}{\sim} \mathcal{CN}(0,1)$ and $g_{\ell,k}^e \overset{i.i.d}{\sim} \mathcal{CN}(0,\sigma_e^2)$ denote the channel path gain and estimated error term for each path ℓ of each user k respectively. $\phi_{\ell,k}$ and $\phi_{\ell,k}^e$ denote the channel AoA and estimated error term for each path ℓ of each user k with $\phi_{\ell,k}^e \overset{i.i.d}{\sim} \mathcal{U}[-e,e]$ where $e \in [0,\pi/2]$, respectively. Here $\mathcal{U}[-e,e]$ represents the uniform distribution.

Fig. 3.5 illustrates the ergodic sum rates with $N_r = 128$, $N_{\mathrm{RF}} = 43$, $N_u = 8$, $\lambda_L = 3$, and $b = 2$ versus the SNR ρ for different digital combiners: (a) MRC, (b) ZF, and (c) MMSE. Similarly to the MI results, ARV-TSAC shows the comparable ergodic rate to that of SVD+DFT and outperforms the one-stage combining such as the Greedy-MI and ARV cases in most cases. I note that the SVD case also shows the worst sum rate performance in the considered systems. The gaps between the two-stage combining cases and one-stage combining cases for the MRC and ZF combiners are much larger than the gap for the MMSE combiner. In addition, SVD+DFT and ARV-TSAC with the ZF combiner achieve the ergodic rates comparable to the MMSE combiner, while the Greedy-MI and ARV cases with the ZF combiner show much lower ergodic sum rates than that with the MMSE combiner. Since the MRC and ZF combiners ignore the AWGN and quantization noise whereas the MMSE combiner does not, using the MMSE combiner improves the ergodic rate of the one-stage analog combining cases. The two-stage analog combining cases, however, already reduced the quantization noise by using the second analog combiner, and thus, they provide the MMSE-like ergodic rate performance with the ZF combiner. Therefore, the proposed two-stage analog combining with the ARV-TSAC algorithm can achieve significant rate improvement with the MRC or ZF combiners compared to the one-stage analog combining approach.

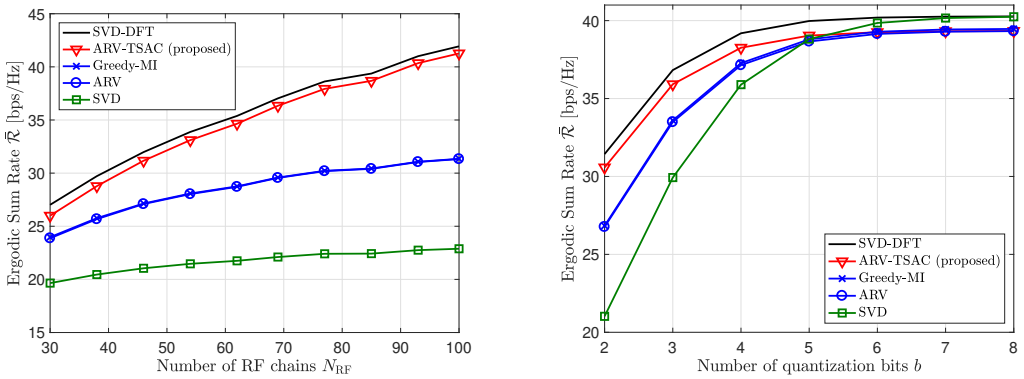
Fig. 3.6 shows the ergodic rate simulation results of the proposed algorithm with MRC, ZF, and MMSE combining for the imperfect CSI case.



(a) MRC (b) ZF (c) MMSE

Figure 3.6: Imperfect CSI simulation results of the ergodic sum rate with $N_r = 128$ receive antennas, $N_{\text{RF}} = 43$ RF chains, $N_u = 8$ users, $\lambda_L = 3$ average channel paths, $b = 2$ quantization bits, $\sigma_e^2 = 10^{-1}$, and $\phi_{\ell,k}^e \sim \mathcal{U}\big[\sin^{-1}(-\frac{1}{N_r}), \sin^{-1}(\frac{1}{N_r})\big]$ for (a) maximum ratio combining (MRC), (b) zero-forcing (ZF), and (c) minimum mean squared error (MMSE) digital combiners.

Compared to the perfect CSI case in Fig. 3.5, the results show degradation in ergodic rates while maintaining a similar trend. Although the proposed ARV-TSAC shows a larger gap to the SVD-DFT case for imperfect CSI vs. perfect CSI, the ARV-TSAC still achieves higher rates with large gap from the one-stage analog combining case such as the ARV, Greedy-MI, and SVD cases. Unlike the perfect CSI case, there is no performance reversal between ARV-TSAC and the other one-stage analog combining cases for MMSE combining, and ARV-TSAC with ZF achieves higher rate than with MMSE combining. This shows that in the considered hybrid beamforming systems with low-resolution ADCs, the MMSE combining is more vulnerable to channel estimation error. Since simulation results show that the proposed ARV-TSAC with ZF combining achieves similar performance to the MMSE combining, it is expected that the proposed two-stage analog combining with ZF digital combining can provide good performance in general and offer more robust



(a) $\bar{\mathcal{R}}^{\mathrm{mrc}}$ vs. N_{RF} (b) $\bar{\mathcal{R}}^{\mathrm{mrc}}$ vs. b

Figure 3.7: Simulation results of the ergodic sum rate of the MRC combiner with $N_u = 8$ users, $\lambda_L = 3$ average channel paths, and $\rho = 0$ dB SNR for (a) $b = 2$ quantization bits and $\kappa = N_{\mathrm{RF}}/N_r = 1/3$ and (b) $N_r = 128$ receive antennas and $N_{\mathrm{RF}} = 43$ RF chains.

performance to channel estimation error, thereby achieving higher ergodic rate than the other linear digital combiners.

Fig. 3.7 provides the simulation results of the ergodic rate with the MRC digital combiner for $N_u = 8$, $\lambda_L = 3$, and $\rho = 0$ dB in terms of the number of (a) RF chains N_{RF} and (b) quantization bits b. In Fig. 3.7(a), we consider $b = 2$ and $\kappa = N_{\mathrm{RF}}/N_r = 1/3$. The ergodic rates of SVD+DFT and ARV-TSAC are similar and both increase logarithmically with N_{RF}, whereas the ergodic rates of the Greedy-MI and ARV cases increase more slowly. Such scaling results correspond to Remark 8 and 9. As N_r increases with a fixed κ, SVD+DFT and ARV-TSAC effectively reduce the more quantization error while obtaining larger channel gains, but the Greedy-MI and ARV cases only obtain larger channel gains without mitigating the quantization error. In Fig.

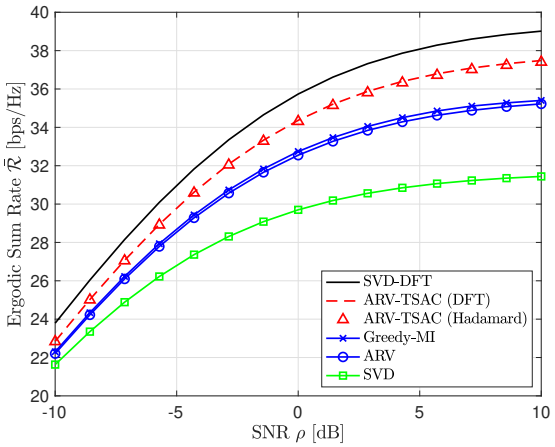


Figure 3.8: Simulation results of the ergodic sum rate with $N_r = 128$ receive antennas, $N_{\text{RF}} = 32$ RF chains, $N_u = 8$ users, $\lambda_L = 3$ average channel paths, and $b = 3$ quantization bits for maximum ratio combining (MRC).

3.7(b), I consider $N_r = 128$ and $N_{\text{RF}} = 43$. Note that in the low-resolution ADC regime, the ARV-TSAC algorithm achieves the ergodic rate comparable to that of SVD+DFT and shows a noticeable improvement compared to the Greedy-MI, ARV, and SVD cases. As b increases, the ergodic rates of the ARV-TSAC, Greedy-MI, and ARV algorithms converge to each other with a small gap from the SVD+DFT case. The ergodic rate of the SVD case, however, converges to that of SVD+DFT without any gap because the SVD combining is optimal in maximizing the MI of infinite-resolution ADC systems. The simulation results validate the effectiveness of the proposed two-stage combining in low-resolution ADC systems.

Fig. 3.8 shows the simulation results for the MRC digital combiner, including a DFT-based second analog combiner and a Hadamard-based sec-

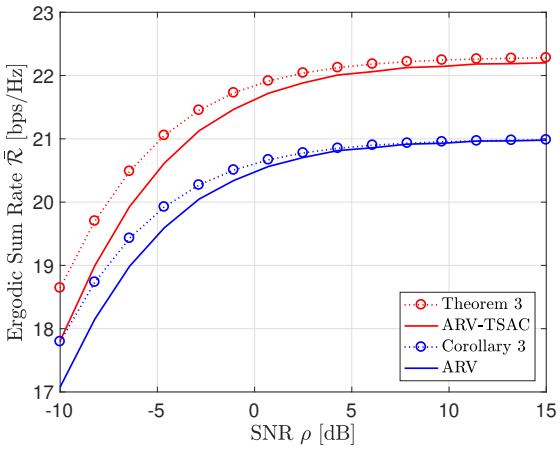


Figure 3.9: Comparison of the ergodic rate for the theoretical and simulation results with $N_r = 128$ receive antennas, $N_{\text{RF}} = 43$ RF chains, $N_u = 8$ users each with $L = 8$ channel paths for the virtual channels.

ond analog combiner. The simulation results demonstrated that using the Hadamard-based second analog combiner also achieves the sum rate that is the same as the DFT case since the Hadamard matrix also satisfies the condition (ii) in Theorem 2. Therefore, adding $\mathbf{W}_{\text{RF}_2}$ can still maintain similar power consumption to one-stage analog combining systems.

Finally, I validate the derived ergodic rates in Theorem 4 and Corollary 5. $N_r = 128$ receive antennas, $N_{\text{RF}} = 43$ RF chains, $N_u = 8$ users each with $L = 8$ channel paths for the virtual channels, and $b = 2$ quantization bits are considered. In Fig. 3.9, the theoretical ergodic rates tightly align with the simulation results in the medium to high SNR regime, and show similar trend as the simulation results do. Thus, the derived ergodic rates can characterize the ergodic rate performance of the proposed algorithm for

the two-stage analog combining system in terms of the system parameters including quantization resolution.

Overall, the two-stage analog combining structure with the ARV-TSAC algorithm almost achieves the performance of SVD+DFT that is a near optimal solution for the unconstrained problem $\mathcal{P}1$, while the greedy-MI and ARV algorithms provide a near optimal solution only for the constrained problem $\mathcal{P}2$. Since $\mathcal{P}1$ has a larger feasible set than $\mathcal{P}2$ to find an optimal solution for the same objective function, this leads to $\mathcal{C}(\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt}}) \geq \mathcal{C}(\mathbf{W}_{\mathrm{RF}}^{\mathrm{opt,c}})$. In this regard, the ARV-TSAC algorithm achieves the higher performance than that of the Greedy-MI and ARV algorithms in most cases. This shows that the proposed two-stage analog combining architecture with the ARV-TSAC is a practical solution suitable for the mmWave hybrid MIMO systems with low-resolution ADCs.

3.6 Conclusion

In this chapter, I derived a near optimal analog combining solution for an unconstrained MI maximization problem in hybrid MIMO systems with low-resolution ADCs. I showed optimalities of the solution in the scaling law and in maximizing the mutual information for a homogeneous channel singular value case. To implement the derived solution, I proposed a two-stage analog combining architecture that decouples the channel gain aggregation and spreading functions in the solution into two cascaded analog combiners. Accordingly, the proposed two-stage analog combining also provides a near

optimal solution for the unconstrained problem whereas conventional hybrid algorithms offer a near optimal solution only for the constrained problem. In addition, I derived a closed-form approximation to the ergodic rate, which reveals that the two-stage analog combiner achieves the optimal scaling law with a practical digital combiner. Simulation results validated the key insights obtained in this chapter and the derived ergodic rate, and also demonstrated that the proposed two-stage analog combining algorithm outperforms conventional algorithms. In the next chapter, switch-based analog beamforming will be considered as different power-efficient solution to avoid the burden of implementing large phase shifter arrays.

3.7 Proof of Corollary 4

Let $\mathbf{H}$ be decomposed into $\mathbf{H} = \mathbf{A}_{\mathrm{AoA}}\mathbf{H}_{\mathrm{V}}$, where the beamdomain channel is $\mathbf{H}_{\mathrm{V}} = \mathrm{blkdiag}\{\tilde{\mathbf{g}}_1, \cdots, \tilde{\mathbf{g}}_{N_u}\}$ and $\tilde{\mathbf{g}}_k = \sqrt{\frac{N_r}{L_k}}[g_{1,k}, \cdots, g_{L_k,k}]^T$. Then, it can be shown [98] that as $N_r \to \infty$,

$$\mathbf{W}_{\mathrm{AoA}}^H \mathbf{W}_{\mathrm{AoA}} \to \mathbf{I}_{N_{\mathrm{RF}}}, \quad \frac{1}{\sqrt{N_r}} \mathbf{W}_{\mathrm{AoA}}^H \mathbf{H} \to \frac{1}{\sqrt{N_r}} \begin{bmatrix} \mathbf{H}_{\mathrm{V}} \\ \mathbf{0} \end{bmatrix}. \tag{3.34}$$

Let $\tilde{\mathbf{H}}_{\mathrm{V}} = [\mathbf{H}_{\mathrm{V}}^T, \mathbf{0}^T]^T$ and $\mathbf{C}_{\mathrm{AoA}} = \mathbf{W}_{\mathrm{RF}_2}^{\star H} \tilde{\mathbf{H}}_{\mathrm{V}} \tilde{\mathbf{H}}_{\mathrm{V}}^H \mathbf{W}_{\mathrm{RF}_2}^{\star}$. Using (3.34), we show $\mathcal{C}(\mathbf{W}_{\mathrm{RF}})$ in (3.12) with $\mathbf{W}_{\mathrm{RF}} = \tilde{\mathbf{W}}_{\mathrm{RF}}^{\star}$ converges as $N_r \to \infty$ to

$$\left(\mathcal{C}(\tilde{\mathbf{W}}_{\mathrm{RF}}^{\star}) - \log_2 \left| \mathbf{I} + \frac{\alpha_b}{\beta_b} \mathrm{diag}^{-1} \left\{ \mathbf{C}_{\mathrm{AoA}} + \tfrac{1}{\beta_b \rho} \mathbf{I} \right\} \mathbf{C}_{\mathrm{AoA}} \right| \right) \to 0. \tag{3.35}$$

Note that each diagonal of $\mathbf{W}_{\mathrm{RF}_2}^{\star H} \tilde{\mathbf{H}}_{\mathrm{V}} \tilde{\mathbf{H}}_{\mathrm{V}}^H \mathbf{W}_{\mathrm{RF}_2}^{\star}$ is $\frac{1}{\kappa} \sum_{k=1}^{N_u} \frac{1}{L_k} (\sum_{\ell=1}^{L_k} |g_{\ell,k}|)^2 = c_1 < \infty$. Let $\mathcal{C}_{\infty}(\tilde{\mathbf{W}}_{\mathrm{RF}}^{\star})$ denote the second term in (3.35). Then, $\mathcal{C}_{\infty}(\tilde{\mathbf{W}}_{\mathrm{RF}}^{\star})$

can be lower bounded as

$$\begin{aligned}\mathcal{C}_{\infty}(\tilde{\mathbf{W}}_{\mathrm{RF}}^{\star}) &> \log_2 \left| \mathbf{I}_{N_{\mathrm{RF}}} + \frac{\alpha_b \rho}{c_1 \beta_b \rho + 1} \mathbf{W}_{\mathrm{RF}_2}^{\star H} \tilde{\mathbf{H}}_{\mathrm{V}} \tilde{\mathbf{H}}_{\mathrm{V}}^{H} \mathbf{W}_{\mathrm{RF}_2}^{\star} \right| \\ &\overset{(a)}{\sim} N_u \log_2 N_{\mathrm{RF}}, \text{ as } N_{\mathrm{RF}} \to \infty, \end{aligned} \tag{3.36}$$

where (a) follows from the same reason of (b) below (3.16). This implies that $\mathcal{C}(\tilde{\mathbf{W}}_{\mathrm{RF}}^{\star})$ follows the optimal scaling law. ■

3.8 Proof of Lemma 3

The auto quantization noise variance term in (3.26) is expressed as

$$\begin{aligned}\mathbb{E}\left[\Psi_k^{\mathrm{auto}}\right] &= \mathbb{E}\left[\sum_{i=1}^{N_{\mathrm{RF}}} \left|\mathbf{h}_{\mathrm{b},k}^{H} \mathbf{w}_i\right|^4\right] \\ &= \left(\frac{N_r}{L}\right)^2 \sum_{i=1}^{N_{\mathrm{RF}}} \mathbb{E}\left[\left|\mathbf{g}_k^{H} \mathbf{w}_i\right|^4\right] \\ &= \left(\frac{N_r}{L}\right)^2 \sum_{i=1}^{N_{\mathrm{RF}}} \left(\mathbb{V}\left[\left|\mathbf{g}_k^{H} \mathbf{w}_i\right|^2\right] + \left(\mathbb{E}\left[\left|\mathbf{g}_k^{H} \mathbf{w}_i\right|^2\right]\right)^2\right) \end{aligned} \tag{3.37}$$

where $\mathbf{w}_i$ is the ith column of $\mathbf{W}_{\mathrm{DFT}}$. The expectation term $\mathbb{E}[|\mathbf{g}_k^{H} \mathbf{w}_i|^2]$ in (3.37) is computed as

$$\mathbb{E}\left[\left|\mathbf{g}_k^{H} \mathbf{w}_i\right|^2\right] = \frac{1}{N_{\mathrm{RF}}} \mathbb{E}\left[\sum_{\ell=1}^{N_{\mathrm{RF}}} |g_{\ell,k}|^2\right] = \frac{L}{N_{\mathrm{RF}}}. \tag{3.38}$$

Now, let $\hat{\mathbf{w}}_i = \sqrt{N_{\mathrm{RF}}}\mathbf{w}_i$. Then, the variance term $\mathbb{V}[|\mathbf{g}_k^H \mathbf{w}_i|^2]$ in (3.37) can be computed as

$$\begin{aligned}
\mathbb{V}\Big[\big|\mathbf{g}_k^H \mathbf{w}_i\big|^2\Big] &= \frac{1}{N_{\mathrm{RF}}^2}\mathbb{V}\left[\sum_{\ell=1}^{N_{\mathrm{RF}}} |g_{\ell,k}|^2 + \sum_{\ell_1\neq\ell_2}^{N_{\mathrm{RF}}} g_{\ell_1,k}^* g_{\ell_2,k} \hat{w}_{\ell_1,i}^* \hat{w}_{\ell_2,i}\right] \\
&\overset{(a)}{=} \frac{1}{N_{\mathrm{RF}}^2}\left(\mathbb{V}\left[\sum_{\ell=1}^{N_{\mathrm{RF}}} |g_{\ell,k}|^2\right] + \mathbb{V}\left[\sum_{\ell_1\neq\ell_2}^{N_{\mathrm{RF}}} g_{\ell_1,k}^* g_{\ell_2,k} \hat{w}_{\ell_1,i}^* \hat{w}_{\ell_2,i}\right]\right) \\
&\overset{(b)}{=} \frac{1}{N_{\mathrm{RF}}^2}\left(\mathbb{V}\Big[\|\mathbf{g}_k\|^2\Big] + \sum_{\ell_1\neq\ell_2}^{N_{\mathrm{RF}}} \mathbb{V}\Big[g_{\ell_1,k}^* g_{\ell_2,k}\Big]\right) \qquad (3.39)
\end{aligned}$$

where (a) and (b) hold as the associated terms are uncorrelated, which can be shown from straight forward mathematics, and $|\hat{w}_{\ell,i}| = 1$, $\forall \ell, i$. Since $\|\mathbf{g}_k\|^2 \sim \chi_{2L}^2$, which is a chi-square distribution with $2L$ degrees of freedom, I have $\mathbb{V}[\|\mathbf{g}_k\|^2] = L$, and $\mathbb{V}[g_{\ell_1,k}^* g_{\ell_2,k}]$ is computed as

$$\begin{aligned}
\mathbb{V}\Big[g_{\ell_1,k}^* g_{\ell_2,k}\Big] &= \mathbb{V}\Big[\xi_{\ell_1,k}^* \xi_{\ell_2,k} \mathbb{1}_{\{\ell_1\in\mathcal{P}_k\}} \mathbb{1}_{\{\ell_2\in\mathcal{P}_k\}}\Big] \\
&\overset{(a)}{=} \mathbb{E}\big[|\xi_{\ell_1,k}^* \xi_{\ell_2,k}|^2\big]\mathbb{E}\left[\mathbb{1}_{\{\ell_1,\ell_2\in\mathcal{P}_k\}}\right] - \Big(\mathbb{E}\left[\xi_{\ell_1,k}^* \xi_{\ell_2,k}\right]\Big)^2 \Big(\mathbb{E}\big[\mathbb{1}_{\{\ell_1,\ell_2\in\mathcal{P}_k\}}\big]\Big)^2 \\
&= \frac{L(L-1)}{N_{\mathrm{RF}}(N_{\mathrm{RF}}-1)},
\end{aligned}$$

where (a) holds by $\mathbb{V}[XY] = \mathbb{E}[X^2]\mathbb{E}[Y^2] - (\mathbb{E}[X])^2(\mathbb{E}[Y])^2$ for independent X and Y. Therefore, (3.39) is derived as

$$\mathbb{V}\Big[\big|\mathbf{g}_k^H \mathbf{w}_i\big|^2\Big] = \frac{1}{N_{\mathrm{RF}}^2}\left(L + \sum_{\ell_1\neq\ell_2}^{N_{\mathrm{RF}}} \frac{L(L-1)}{N_{\mathrm{RF}}(N_{\mathrm{RF}}-1)}\right) = \left(\frac{L}{N_{\mathrm{RF}}}\right)^2. \qquad (3.40)$$

Putting (3.38) and (3.40) into (3.37), the auto quantizaiton noise variance $\mathbb{E}\big[\Psi_k^{\mathrm{auto}}\big]$ becomes (3.28). ■

3.9 Proof of Lemma 4

The cross quantization noise variance in (3.27) is derived as

$$\begin{aligned}\mathbb{E}\big[\Psi_k^{\text{cross}}\big] &= \mathbb{E}\bigg[\sum_{i=1}^{N_{\text{RF}}}\sum_{u\neq 1}^{N_u}\mathbf{h}_{\text{b},k}^H\mathbf{w}_i\mathbf{w}_i^H\mathbf{h}_{\text{b},u}\mathbf{h}_{\text{b},u}^H\mathbf{w}_i\mathbf{w}_i^H\mathbf{h}_{\text{b},k}\bigg] \\ &= \left(\frac{N_r}{L}\right)^2\mathbb{E}_{\mathbf{g}_k}\bigg[\sum_{i=1}^{N_{\text{RF}}}\sum_{u\neq 1}^{N_u}\mathbf{g}_k^H\mathbf{w}_i\mathbf{w}_i^H\mathbb{E}_{\mathbf{g}_u}\Big[\mathbf{g}_u\mathbf{g}_u^H\Big]\mathbf{w}_i\mathbf{w}_i^H\mathbf{g}_k\bigg] \\ &= \frac{N_r^2(N_u-1)}{LN_{\text{RF}}}\sum_{i=1}^{N_{\text{RF}}}\mathbb{E}_{\mathbf{g}_k}\Big[\mathbf{g}_k^H\mathbf{w}_i\mathbf{w}_i^H\mathbf{g}_k\Big] \\ &\overset{(a)}{=} \frac{N_r^2(N_u-1)}{N_{\text{RF}}}\end{aligned}$$

where (a) follows from $\mathbb{E}\big[|\mathbf{g}_k^H\mathbf{w}_i|^2\big] = \frac{L}{N_{\text{RF}}}$ in (3.38). ■

3.10 Proof of Theorem 4

To compute (3.25), I first derive $\mathbb{E}[\|\mathbf{h}_{\text{b},k}\|^2]$ as

$$\mathbb{E}\Big[\|\mathbf{h}_{\text{b},k}\|^2\Big] = \frac{N_r}{L}\mathbb{E}\big[\|\mathbf{g}_k\|^2\big] \overset{(a)}{=} N_r \tag{3.41}$$

where (a) follows from $\|\mathbf{g}_k\|^2 \sim \chi^2_{2L}$. Next, $\mathbb{E}[\|\mathbf{h}_{\text{b},k}\|^4]$ is computed as

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{h}_{\text{b},k}\|^4\Big] &= \mathbb{V}\big[\|\mathbf{h}_{\text{b},k}\|^2\big] + \big(\mathbb{E}\big[\|\mathbf{h}_{\text{b},k}\|^2\big]\big)^2 \\ &= \left(\frac{N_r}{L}\right)^2\bigg(\mathbb{V}\big[\|\mathbf{g}_k\|^2\big] + \Big(\mathbb{E}\big[\|\mathbf{g}_k\|^2\big]\Big)^2\bigg) \\ &= \frac{N_r^2(1+L)}{L}.\end{aligned} \tag{3.42}$$

The inter-user interference term $\mathbb{E}[|\mathbf{h}_{\mathrm{b},k}^H \mathbf{h}_{\mathrm{b},i}|^2]$ is computed as

$$\begin{aligned}
\mathbb{E}\Big[|\mathbf{h}_{\mathrm{b},k}^H \mathbf{h}_{\mathrm{b},i}|^2\Big] &= \left(\frac{N_r}{L}\right)^2 \mathbb{E}\big[|\mathbf{g}_k^H \mathbf{g}_i|^2\big] \\
&= \left(\frac{N_r}{L}\right)^2 \sum_{\ell=1}^{N_{\mathrm{RF}}} \mathbb{E}\big[|g_{\ell,k}^* g_{\ell,i}|^2\big] \\
&= \left(\frac{N_r}{L}\right)^2 \sum_{\ell=1}^{N_{\mathrm{RF}}} \mathbb{E}\Big[|\xi_{\ell,k}^* \mathbb{1}_{\{\ell \in \mathcal{P}_k\}} \xi_{\ell,i} \mathbb{1}_{\{\ell \in \mathcal{P}_i\}}|^2\Big] \\
&= \frac{N_r^2}{N_{\mathrm{RF}}}. \qquad (3.43)
\end{aligned}$$

Finally, we compute the quantization variance term $\mathbb{E}[\Psi_k]$ as

$$\begin{aligned}
\mathbb{E}\big[\Psi_k\big] &= \mathbb{E}\big[\Psi_k^{\mathrm{auto}}\big] + \mathbb{E}\big[\Psi_k^{\mathrm{cross}}\big] \\
&\overset{(a)}{=} \frac{2N_r^2}{N_{\mathrm{RF}}} + \frac{N_r^2(N_u - 1)}{N_{\mathrm{RF}}}, \qquad (3.44)
\end{aligned}$$

where $\mathbb{E}\big[\Psi_k^{\mathrm{auto}}\big]$ and $\mathbb{E}\big[\Psi_k^{\mathrm{cross}}\big]$ are in (3.26) and (3.27), respectively, and (a) follows from Lemma 3 and Lemma 4.

Putting (3.41), (3.42), (3.43), and (3.44) into (3.25), I derive the approximated ergodic rate of (3.25) in closed form. The ergodic rate is equivalent to N_u users, which leads to the ergodic sum rate in (3.30). This completes the proof of Theorem 4. ■

3.11 Proof of Corollary 5

Without the second analog combiner $\mathbf{W}_{\mathrm{RF}}$, the approximated ergodic rate of user k can be computed as (3.25) by substituting the average quantization noise variance for the two-stage analog combining $\mathbb{E}[\Psi_k]$ with the following

average quantization noise variance:

$$\begin{aligned}\mathbb{E}\big[\hat{\Psi}_k\big] &= \mathbb{E}\Big[\mathbf{h}_{\mathrm{b},k}^H \mathrm{diag}\big\{\mathbf{H}_\mathrm{b}\mathbf{H}_\mathrm{b}^H\big\}\mathbf{h}_{\mathrm{b},k}\Big] \\ &= \mathbb{E}\left[\left(\frac{N_r}{L}\right)^2 \sum_{\ell=1}^{N_\mathrm{RF}} |g_{\ell,k}|^2 \sum_{u=1}^{N_u} |g_{\ell,u}|^2\right] \\ &= \left(\frac{N_r}{L}\right)^2 \left(\sum_{\ell=1}^{N_\mathrm{RF}} \mathbb{E}\left[|g_{\ell,k}|^4\right] + \sum_{\ell=1}^{N_\mathrm{RF}} \sum_{u\neq k}^{N_u} \mathbb{E}\Big[|g_{\ell,k}|^2 |g_{\ell,u}|^2\Big]\right). \end{aligned} \tag{3.45}$$

Here, $\mathbb{E}[|g_{\ell,k}|^4]$ in (3.45) is computed as

$$\begin{aligned}\mathbb{E}\Big[|g_{\ell,k}|^4\Big] &= \mathbb{E}\Big[\mathbb{1}_{\{\ell\in\mathcal{P}_k\}}\Big]\mathbb{E}\Big[\big|\xi_{\ell,k}\big|^4\Big] \\ &= \frac{L}{N_\mathrm{RF}}\left(\mathbb{V}\Big[\big|\xi_{\ell,k}\big|^2\Big] + \Big(\mathbb{E}\Big[\big|\xi_{\ell,k}\big|^2\Big]\Big)^2\right) \\ &= \frac{2L}{N_\mathrm{RF}}, \end{aligned} \tag{3.46}$$

and the second expectation term $\mathbb{E}[|g_{\ell,k}|^2|g_{\ell,u}|^2]$ is derived as

$$\begin{aligned}\mathbb{E}\Big[|g_{\ell,k}|^2|g_{\ell,u}|^2\Big] &= \mathbb{E}\Big[\mathbb{1}_{\{\ell\in\mathcal{P}_k\}}\mathbb{1}_{\{\ell\in\mathcal{P}_u\}}\Big]\mathbb{E}\Big[|\xi_{\ell,k}|^2|\xi_{\ell,u}|^2\Big] \\ &= \left(\frac{L}{N_\mathrm{RF}}\right)^2. \end{aligned} \tag{3.47}$$

Putting (3.46) and (3.47) into (3.45), I derive the average quantization noise variance for the one-stage analog combining as

$$\mathbb{E}\big[\hat{\Psi}_k\big] = N_r^2\left(\frac{2}{L} + \frac{N_u - 1}{N_\mathrm{RF}}\right).$$

This completes the proof of Corollary 5. ■

Chapter 4

Base Station Antenna Selection for Low-Resolution ADC Systems

In this chapter[1], I investigate antenna selection at a base station with large antenna arrays and low-resolution analog-to-digital converters. For downlink transmit antenna selection for narrowband channels, I show (1) a selection criterion that maximizes sum rate with zero-forcing precoding equivalent to that of a perfect quantization system; (2) maximum sum rate increases with number of selected antennas; (3) derivation of the sum rate loss function from using a subset of antennas; and (4) unlike high-resolution converter systems, sum rate loss reaches a maximum at a point of total transmit power and decreases beyond that point to converge to zero. For wideband orthogonal-frequency-division-multiplexing (OFDM) systems, the results hold when entire subcarriers share a common subset of antennas. For uplink receive antenna

[1]This chapter is based on the work: J. Choi, J. Sung, N. Prasad, X. Qi, B. L. Evans, and A. Gatherer, "Base Station Antenna Selection for Low-Resolution ADC Systems", submitted to *IEEE Transactions on Communications*, 2019. Part of the work was also published in the conference paper: J. Choi, J. Sung, B. L. Evans, and A. Gatherer, "Antenna Selection for Large-Scale MIMO Systems with Low-Resolution ADCs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 15-20, 2018. This work was supervised by Prof. Brian L. Evans, and valuable feedback from Junmo Sung, Narayan Prasad, Xiao-Feng Qi, and Alan Gatherer improved the quality of this work.

selection for narrowband channels, I (1) generalize a greedy antenna selection criterion to capture tradeoffs between channel gain and quantization error; (2) propose a quantization-aware fast antenna selection algorithm using the criterion; and (3) derive a lower bound on sum rate achieved by the proposed algorithm based on submodular functions. For wideband OFDM systems, I extend the proposed algorithm and derive a lower bound on its sum rate. Simulation results validate theoretical analyses and show increases in sum rate over conventional algorithms.

4.1 Introduction

Although phase shifter-based analog beamforming that was considered the previous chapters can offer high flexibility in analog processing, the implementation of large phase shifter arrays requires additional cost and complexity. Greatly reducing such implementation burden, switch-based analog beamforming that is equivalent to antenna selection is a different energy-efficient architecture to reduce the number of RF chains and ADCs. Antenna selection problems have been widely studied without quantization error for high-resolution ADC systems. For the transmit antenna selection, it was shown that single antenna selection achieves full diversity gain which the transmitter without antenna selection (the transmitter uses all antennas) achieves [109], and it is optimal in the low signal-to-noise ratio (SNR) [110]. To find the best transmit antenna subset, convex optimization techniques were adopted by relaxing a binary integer problem to a real number problem [111,112]. Transmit

antenna selection was also jointly studied with other problems [113, 114]. An outage probability was derived for single user selection and antenna selection in [113], and a precoder was designed jointly with antenna selection [114]. Energy and spectral efficiency tradeoff was maximized in [115] by solving a multi-objective antenna selection problem. For special systems such as spatial modulation systems, a Euclidean distance-based antenna selection method was developed [116].

Receive antenna selection methods were also developed for last decade [38–43]. In [38], a greedy antenna selection method was developed by minimizing capacity loss. It was shown in [38] that the diversity order of the receive antenna selection system is same as the full diversity order. In [39], a correlation-based method and mutual information-based method were developed, showing that selecting receive antennas more than the number of transmit antennas can nearly achieve the performance of full receive antenna systems. Convex optimization approach was also taken in receive antenna selection [40]. To provide a lower bound of greedy selection methods, modularity and submodularity concepts were used in [41]. In [117] a sampling-based selection method was proposed by employing cross entropy optimization technique.

Antenna selection problems have been studied for various channels. For correlated channels, selection algorithms were proposed by exploiting partial channel state information (CSI) such as a channel covariance matrix [118]. Antenna selection problems were also solved for millimeter wave channels jointly with precoder design [119, 120]. In orthogonal frequency division multiplexing

(OFDM) systems, both transmit antenna selection [121, 122] and receive antenna selection algorithms [42, 43] were developed. An adaptive Markov chain Monte Carlo (MCMC) method was adopted for antenna selection [42], and optimal power allocation between training and data symbols with antenna selection was derived to minimize performance loss due to channel estimation error [43]. An outage probability was analyzed for per-subcarrier antenna selection in [121], and an adaptive antenna selection method that balances between per-subcarrier and bulk selection was proposed in [122].

Most prior work on antenna selection, however, focused on MIMO systems without any quantization errors. Accordingly, antenna selection for low-resolution ADC systems that incorporates coarse quantization effect needs to be investigated. In [123], a cross entropy maximization approach in [117] was extended for low-resolution ADC systems by jointly solving the user scheduling problem. Transmit antenna selection was analyzed for single antenna selection by utilizing Weibul distribution in low-resolution ADC systems [124]. In [124], it was shown that although the TAS gain is limited when compared to the gain for perfect quantization, the TAS gain can still provide a large increase of ergodic rate. Although the proposed receive antenna selection algorithm in [123] demonstrated its high performance, it can require high complexity when the number of candidate antennas are large due to its parameters such as the number of iterations and sampling. In addition, the transmit antenna selection in [124] considers single antenna selection and thus, it is difficult to be generalized to multiple antenna selection.

4.1.1 Contributions

In this chapter, I investigate antenna selection at a BS with a large number of antenna arrays in low-resolution ADC systems where both the BS and mobile stations (MSs) are equipped with low-resolution ADCs. I investigate DL transmit antenna selection and UL receive antenna selection. The contributions are summarized as follows:

- For narrowband channels, I show that the DL transmit antenna selection problem with zero-forcing (ZF) precoding in low-resolution ADC systems is equivalent to that in high-resolution ADC systems when antennas are selected to maximize the DL sum rate. Observing the quantization effect in the SNR, I further analyze the DL sum rate with antenna selection by incorporating quantization effects. I show that selecting more transmit antennas provides larger maximum sum rate for low-resolution ADC systems as well as high-resolution ADC systems. Unlike the rate loss in high-resolution ADC systems, I prove that the rate loss decreases beyond a certain point of transmit power and converges to zero in low-resolution ADC systems.

- For an UL receive antenna selection problem in the narrowband, an existing criterion for a greedy capacity-maximization antenna selection method is generalized to incorporate quantization effects. The derived objective function offers an opportunity to select an antenna with the best tradeoff between the additional channel gain and increase in quantization error. A lower bound of the sum rate achieved by the proposed greedy algorithm is

also derived by using a concept of submodularity. In addition, I modify the adaptive MCMC antenna selection [42] for the low-resolution ADC systems to provide a numerical upper bound of the sum rate.

- The antenna selection problem is extended to the wideband OFDM systems. The wideband OFDM systems under coarse quantization for both DL and UL communications is first derived. Then, I show that the derived results in the DL narrowband communications also hold for the DL OFDM communication when subcarriers share a common antenna subset. For the UL OFDM communications, I modify the proposed received antenna selection algorithms and derive the lower bound of the capacity with the greedy algorithm.

- Simulation results validate the theoretical results and demonstrate that the proposed algorithm outperforms conventional algorithms in achievable rate. The proposed receive antenna selection algorithm provides near optimal sum rate performance in the large antenna array regime.

4.2 System Model

A single-cell multiuser network is considered, in which a BS serves N_{MS} MSs. As shown in Fig. 4.1, The BS is equipped with N_{BS} antennas and low-resolution ADCs. Each MS is equipped with a single antenna and low-resolution ADCs. I assume that the number of the BS antennas is much larger than the number of MSs, $N_{\mathrm{BS}} \gg N_{\mathrm{MS}}$. The CSI is assumed to be known at

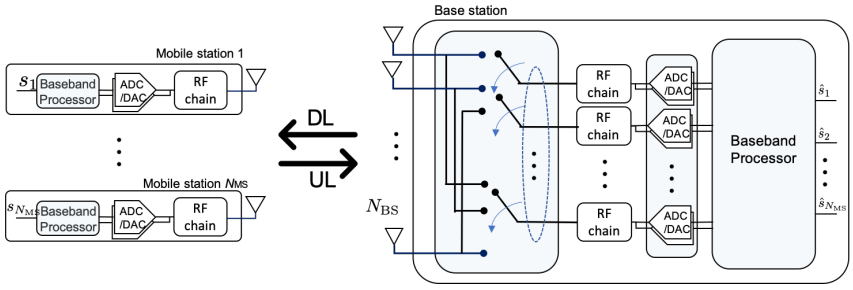


Figure 4.1: A multiuser communication system in which a base station (BS) serves N_{MS} mobile stations (MSs). The BS is equipped with N_{BS} antennas and low-resolution ADCs. Each MS is equipped with a single antenna and low-resolution ADCs.

the BS.

4.2.1 Downlink Narrowband System

The BS selects N_t transmit antennas and employs a ZF precoding to null multiuser interference signals by using the CSI. The vector of the precoded transmit signals $\mathbf{x}^{\mathrm{dl}} \in \mathbb{C}^{N_t}$ is given as

$$\mathbf{x}^{\mathrm{dl}} = \mathbf{W}_{\mathrm{BB}}(\mathcal{T})\mathbf{P}^{1/2}\mathbf{s}^{\mathrm{dl}}$$

where $\mathbf{W}_{\mathrm{BB}}(\mathcal{T}) \in \mathbb{C}^{N_t \times N_{\mathrm{MS}}}$ is the precoder with the selected antennas in the subset of antenna indices $\mathcal{T}$, $\mathbf{P} = \mathrm{diag}\{p_1, \ldots, p_{N_{\mathrm{MS}}}\}$ is the matrix of transmit power for $\mathbf{s}^{\mathrm{dl}}$, and $\mathbf{s}^{\mathrm{dl}} \in \mathbb{C}^{N_{\mathrm{MS}}}$ is the user symbol vector. The transmit power is constrained by the total power constraint P as

$$\mathrm{tr}(\mathbb{E}[\mathbf{x}^{\mathrm{dl}}\mathbf{x}^{\mathrm{dl}\,H}]) = \mathrm{tr}(\mathbf{W}_{\mathrm{BB}}(\mathcal{T})\mathbf{P}\mathbf{W}_{\mathrm{BB}}^{H}(\mathcal{T})) \leq P. \tag{4.1}$$

With ZF precoding, the precoder $\mathbf{W}_{\mathrm{BB}}(\mathcal{T})$ becomes $\mathbf{W}_{\mathrm{BB}}(\mathcal{T}) = \mathbf{H}_{\mathcal{T}}^{\mathrm{dl}\,H}(\mathbf{H}_{\mathcal{T}}^{\mathrm{dl}}\mathbf{H}_{\mathcal{T}}^{\mathrm{dl}\,H})^{-1}$. Accordingly, the received analog baseband signals at the MSs is given as

$$\mathbf{r}^{\mathrm{dl}} = \mathbf{H}_{\mathcal{T}}^{\mathrm{dl}}\mathbf{x}^{\mathrm{dl}} + \mathbf{n}^{\mathrm{dl}} = \mathbf{P}^{1/2}\mathbf{s}^{\mathrm{dl}} + \mathbf{n}^{\mathrm{dl}} \tag{4.2}$$

where $\mathbf{H}_{\mathcal{T}}^{\mathrm{dl}} \in \mathbb{C}^{N_{\mathrm{MS}} \times N_t}$ is the DL narrowband channel matrix, which consists of N_t selected columns of the DL channel $\mathbf{H}^{\mathrm{dl}} \in \mathbb{C}^{N_{\mathrm{MS}} \times N_{\mathrm{BS}}}$, and $\mathbf{n}^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_{\mathrm{MS}}})$ is the additive white circularly complex Gaussian noise (AWGN) vector.

Using the additive quantization noise model (AQNM) [73], which provides a reasonable accuracy for low to medium SNR [56], the quantized DL received signal vector is expressed as

$$\begin{aligned}\mathbf{y}^{\mathrm{dl}} &= \mathcal{Q}\big(\mathrm{Re}\{\mathbf{r}^{\mathrm{dl}}\}\big) + j\mathcal{Q}\big(\mathrm{Im}\{\mathbf{r}^{\mathrm{dl}}\}\big) \\ &= \alpha_b\mathbf{P}^{1/2}\mathbf{s}^{\mathrm{dl}} + \alpha_b\mathbf{n}^{\mathrm{dl}} + \mathbf{q}^{\mathrm{dl}}\end{aligned} \tag{4.3}$$

where $\mathcal{Q}(\cdot)$ is the element-wise quantizer function. Here, α_b is defined as $\alpha_b = 1 - \beta_b$ and considered to be the quantization gain ($\alpha_b < 1$), and β_b is the normalized mean squared quantization error $\beta_b = \frac{\mathbb{E}[|r_i - y_i|^2]}{\mathbb{E}[|r_i|^2]}$. Assuming a scalar minimum mean squared error (MMSE) quantizer and Gaussian signaling $\mathbf{s}^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_{\mathrm{MS}}})$, β_b is approximated as $\beta_b \approx \frac{\pi\sqrt{3}}{2}2^{-2b}$ for $b > 5$ [125], where b is the number of quantization bits for each real and imaginary part. The values of β_b for $b \leq 5$ are shown in Table 1 in [57]. The vector $\mathbf{q}^{\mathrm{dl}} \in \mathbb{C}^{N_{\mathrm{MS}}}$ represents the additive quantization noise that is uncorrelated with the quantization input $\mathbf{r}^{\mathrm{dl}}$ [73]. It is assumed that the quantization noise follows the complex Gaussian distribution with a zero mean $\mathbf{q}^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\mathbf{q}^{\mathrm{dl}}\mathbf{q}^{\mathrm{dl}}})$ [57]. The covariance matrix

of $\mathbf{q}^{\mathrm{dl}}$ is derived as [57]

$$\mathbf{R}_{\mathbf{q}^{\mathrm{dl}}\mathbf{q}^{\mathrm{dl}}} = \alpha_b(1-\alpha_b)\mathrm{diag}\big\{\mathbb{E}\big[\mathbf{r}^{\mathrm{dl}}\mathbf{r}^{\mathrm{dl}\,H}\big]\big\} = \alpha_b(1-\alpha_b)(\mathbf{P}+\mathbf{I}_{N_{\mathrm{MS}}}). \tag{4.4}$$

4.2.2 Uplink Narrowband System

The BS selects N_r receive antennas and receives signals from N_{MS} MSs. The selected antennas are connected to RF chains followed by low-resolution ADCs. The UL narrowband channel matrix between the BS and MSs is denoted as $\mathbf{H}^{\mathrm{ul}} \in \mathbb{C}^{N_{\mathrm{BS}}\times N_{\mathrm{MS}}}$. The received baseband analog signals at the N_r selected antennas $\mathbf{r}^{\mathrm{ul}} \in \mathbb{C}^{N_r}$ can be expressed as

$$\mathbf{r}^{\mathrm{ul}} = \sqrt{\rho}\mathbf{H}^{\mathrm{ul}}_{\mathcal{K}}\mathbf{s}^{\mathrm{ul}} + \mathbf{n}^{\mathrm{ul}} \tag{4.5}$$

where ρ, $\mathbf{H}^{\mathrm{ul}}_{\mathcal{K}} \in \mathbb{C}^{N_r\times N_{\mathrm{MS}}}$, $\mathbf{s}^{\mathrm{ul}} \in \mathbb{C}^{N_{\mathrm{MS}}}$, and $\mathbf{n}^{\mathrm{ul}} \in \mathbb{C}^{N_r}$ denotes the transmit power, the channel matrix for the selected antennas in the subset of antenna indices $\mathcal{K}$, the user symbol vector, and the AWGN vector, respectively. I assume $\mathbf{s}^{\mathrm{ul}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_{\mathrm{MS}}})$ and $\mathbf{n}^{\mathrm{ul}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r})$.

After the antenna selection, each real and imaginary component of the complex output r^{ul}_i, where r^{ul}_i denotes the ith element of $\mathbf{r}^{\mathrm{ul}}$ in (4.5), is quantized at the pair of ADCs. Adopting the AQNM [73], the quantized UL received baseband signals becomes

$$\begin{aligned}\mathbf{y}^{\mathrm{ul}} &= \mathcal{Q}\big(\mathrm{Re}\{\mathbf{r}^{\mathrm{ul}}\}\big) + j\mathcal{Q}\big(\mathrm{Im}\{\mathbf{r}^{\mathrm{ul}}\}\big) \\ &= \alpha_b\sqrt{\rho}\mathbf{H}^{\mathrm{ul}}_{\mathcal{K}}\mathbf{s}^{\mathrm{ul}} + \alpha_b\mathbf{n}^{\mathrm{ul}} + \mathbf{q}^{\mathrm{ul}}\end{aligned} \tag{4.6}$$

where $\mathbf{q}^{\text{ul}}$ represents the additive quantization noise that is uncorrelated with $\mathbf{r}^{\text{ul}}$. I assume $\mathbf{q}^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\mathbf{q}^{\text{ul}}\mathbf{q}^{\text{ul}}})$ [57]. The covariance matrix of $\mathbf{q}^{\text{ul}}$ is

$$\mathbf{R}_{\mathbf{q}^{\text{ul}}\mathbf{q}^{\text{ul}}} = \alpha_b(1-\alpha_b)\,\text{diag}(\rho \mathbf{H}_{\mathcal{K}}^{\text{ul}} \mathbf{H}_{\mathcal{K}}^{\text{ul}\,H} + \mathbf{I}_{N_r}). \tag{4.7}$$

In the following sections, antenna selection is explored for the considered DL and UL systems.

4.3 Downlink Transmit Antenna Selection

In this section, I first show that a transmit antenna selection problem with ZF precoding for narrowband channels in low-resolution ADC systems is equivalent to that in high-resolution ADC systems. The resulting achievable rate, however, involves the quantization error and thus, the sum rate in low-resolution ADC systems is also analyzed.

4.3.1 Sum Rate Maximization Problem

From the quantized signals $\mathbf{y}^{\text{dl}}$ in (4.3) and quantization covariance matrix $\mathbf{R}_{\mathbf{q}^{\text{dl}}\mathbf{q}^{\text{dl}}}$ in (4.4), the DL achievable rate for user i with selected transmit antennas in $\mathcal{T}$ becomes

$$\gamma_i^{\text{dl}}(\mathcal{T}) = \log_2\left(1 + \frac{\alpha_b^2 p_i}{\alpha_b^2 + \alpha_b(1-\alpha_b)(1+p_i)}\right). \tag{4.8}$$

Assuming equal power distribution, $p_i = p_{\mathcal{T}}$, $\forall i$, and ZF precoding with maximum transmit power from (4.1), we have

$$p_{\mathcal{T}} = \frac{P}{\text{tr}\big(\mathbf{W}_{\text{BB}}^H(\mathcal{T})\mathbf{W}_{\text{BB}}(\mathcal{T})\big)} = \frac{P}{\text{tr}\big((\mathbf{H}_{\mathcal{T}}\mathbf{H}_{\mathcal{T}}^H)^{-1}\big)}. \tag{4.9}$$

Using (4.8) and (4.9), the DL achievable sum rate reduces to

$$\mathcal{R}^{\mathrm{dl}}(\mathcal{T}) = N_{\mathrm{MS}} \log_2 \left(1 + \frac{\alpha_b p_{\mathcal{T}}}{1 + (1 - \alpha_b) p_{\mathcal{T}}} \right). \tag{4.10}$$

The transmit antennas selection problem is formulated by adopting the achievable sum rate in (4.10) as an objective function. Let $\mathcal{S} = \{1, 2, \ldots, N_{\mathrm{BS}}\}$ be the index set of the BS antennas. Then, the transmit antenna selection problem for maximum sum rate is formulated as

$$\mathcal{P}1: \qquad \mathcal{T}^{\star} = \underset{\mathcal{T} \subseteq \mathcal{S}: N_{\mathrm{MS}} \leq |\mathcal{T}| \leq N_t}{\operatorname{argmax}} \mathcal{R}^{\mathrm{dl}}(\mathcal{T}).$$

where N_t is the given maximal number of transmit antennas that can be selected.

Remark 10. *The transmit antenna selection problem* $\mathcal{P}1$ *with ZF precoding and equal power allocation for narrowband channels is equivalent to that in high-resolution ADC systems.*

Accordingly, I show that any state-of-the-art transmit antenna selection methods for multiuser communications with the ZF precoding [112, 126] can be applicable in low-resolution ADC systems. The achievable rate $\mathcal{R}^{\mathrm{dl}}(\mathcal{T})$, however, includes the quantization effect as a noise that is proportional to the transmit power, which differs from perfect quantization systems. In this regards, I provide theoretical analysis for the transmit antenna selection problem to characterize the sum rate and draw intuitions for the low-resolution ADC regime in the following subsection.

4.3.2 Sum Rate Analysis of Transmit Antenna Selection

Here, a property of the sum rate in the considered low-resolution ADC system is first derived with respect to the number of selected antennas. To this end, I introduce Lemma 5.

Lemma 5. *For any matrix* $\mathbf{H} \in \mathbb{C}^{m\times n}$ *with* $\mathrm{rank}(\mathbf{H}) = m$*, the following inequality holds:*

$$\mathrm{tr}\left(\mathbf{Q}\tilde{\mathbf{H}}(\mathbf{I}_\ell - \tilde{\mathbf{H}}^H\mathbf{Q}\tilde{\mathbf{H}})^{-1}\tilde{\mathbf{H}}^H\mathbf{Q}\right) > 0$$

where $\mathbf{Q} = (\tau\mathbf{I}_m + \mathbf{H}\mathbf{H}^H)^{-1}$ *with* $\tau \geq 0$ *and* $\tilde{\mathbf{H}}$ *is a* $m \times \ell$ *sub-matrix of* $\mathbf{H}$ *which consists of the columns of* $\mathbf{H}$ *for* $1 \leq \ell \leq (n-m)$*.*

Proof. See Lemma 2 in [126]. ■

Theorem 5. *The maximum sum rate of MSs with low-resolution ADCs in* (4.10) *is monotonically increasing with the number of selected transmit antennas in ZF precoding DL systems* (4.2)*:*

$$\mathcal{R}^{\mathrm{dl}}(\mathcal{T}_{\mathrm{opt1}}) < \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_{\mathrm{opt2}})$$

where $\mathcal{T}_{\mathrm{opt1}}$ *and* $\mathcal{T}_{\mathrm{opt2}}$ *are the optimal antenna subsets with* $|\mathcal{T}_{\mathrm{opt1}}| < |\mathcal{T}_{\mathrm{opt2}}|$*.*

Proof. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be antenna subsets with $\mathcal{T}_1 \subset \mathcal{T}_2 \subseteq \mathcal{S}$, and $\bar{\mathcal{T}}$ be $\bar{\mathcal{T}} = \mathcal{T}_2 - \mathcal{T}_1$. The average sum rate difference between the sum rates with the two antenna subsets, $\mathcal{T}_1$ and $\mathcal{T}_2$, is

$$\begin{aligned}\frac{\mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}})}{N_{\mathrm{MS}}} &= \frac{\mathcal{R}^{\mathrm{dl}}(\mathcal{T}_2) - \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_1)}{N_{\mathrm{MS}}} \\ &= \log_2\left(1 + \frac{\alpha_b p_{\mathcal{T}_2}}{1 + (1-\alpha_b)p_{\mathcal{T}_2}}\right) - \log_2\left(1 + \frac{\alpha_b p_{\mathcal{T}_1}}{1 + (1-\alpha_b)p_{\mathcal{T}_1}}\right). \quad (4.11)\end{aligned}$$

Using $p_{\mathcal{T}_i} = P/\mathrm{tr}((\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_i}\mathbf{H}^{\mathrm{dl}\,H}_{\mathcal{T}_i})^{-1})$ for $i = 1, 2$, (4.11) is rewritten as

$$\frac{\mathcal{R}^{\mathrm{dl}}_D(\bar{\mathcal{T}})}{N_{\mathrm{MS}}} = \log_2\left(\frac{(\mathrm{tr}((\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_2}\mathbf{H}^{\mathrm{dl}\,H}_{\mathcal{T}_2})^{-1}) + P)(\mathrm{tr}((\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_1}\mathbf{H}^{\mathrm{dl}\,H}_{\mathcal{T}_1})^{-1}) + (1-\alpha_b)P)}{(\mathrm{tr}((\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_2}\mathbf{H}^{\mathrm{dl}\,H}_{\mathcal{T}_2})^{-1}) + (1-\alpha_b)P)(\mathrm{tr}((\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_1}\mathbf{H}^{\mathrm{dl}\,H}_{\mathcal{T}_1})^{-1}) + P)}\right).$$

Let $\mathbf{Q} = (\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_2}\mathbf{H}^{\mathrm{dl}\,H}_{\mathcal{T}_2})^{-1}$ and $\mathbf{\Psi}_{\bar{\mathcal{T}}} = \mathbf{Q}\mathbf{H}^{\mathrm{dl}}_{\bar{\mathcal{T}}}(\mathbf{I}_{|\bar{\mathcal{T}}|} - \mathbf{H}^{\mathrm{dl}\,H}_{\bar{\mathcal{T}}}\mathbf{Q}\mathbf{H}^{\mathrm{dl}}_{\bar{\mathcal{T}}})^{-1}\mathbf{H}^{\mathrm{dl}\,H}_{\bar{\mathcal{T}}}\mathbf{Q}$. Then, leveraging the matrix inversion lemma, the rate difference $\mathcal{R}^{\mathrm{dl}}_D(\bar{\mathcal{T}})$, which I also call as the rate loss, becomes

$$\begin{aligned}\mathcal{R}^{\mathrm{dl}}_D(\bar{\mathcal{T}}) &= N_{\mathrm{MS}}\log_2\left(\frac{(\mathrm{tr}(\mathbf{Q})+P)(\mathrm{tr}(\mathbf{Q})+\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})+(1-\alpha_b)P)}{(\mathrm{tr}(\mathbf{Q})+(1-\alpha_b)P)(\mathrm{tr}(\mathbf{Q})+\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})+P)}\right) \\ &= N_{\mathrm{MS}}\log_2\left(1+\frac{\alpha_b\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})P}{\mathrm{tr}(\mathbf{Q})^2+\big(\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})+P\big)\mathrm{tr}(\mathbf{Q})+(1-\alpha_b)\big(P^2+P(\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})+\mathrm{tr}(\mathbf{Q}))\big)}\right) \qquad (4.12)\end{aligned}$$

$$\overset{(a)}{>} 0 \qquad (4.13)$$

where (a) holds from the following reasons: I have $\mathrm{tr}(\mathbf{Q}) > 0$, and from Lemma 5 with $\tau = 0$, I have $\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}}) > 0$ for any channel matrix $\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_2}$ with $\mathrm{rank}(\mathbf{H}^{\mathrm{dl}}_{\mathcal{T}_2}) = N_{\mathrm{MS}}$ and its $N_{\mathrm{MS}} \times |\bar{\mathcal{T}}|$ sub-matrix $\mathbf{H}^{\mathrm{dl}}_{\bar{\mathcal{T}}}$ with $1 \le |\bar{\mathcal{T}}| \le (|\mathcal{T}_2| - N_{\mathrm{MS}})$. In addition, α_b is always less than one ($\alpha_b < 1$) since it is the quantization gain defined as $\alpha_b = 1 - \mathbb{E}[|r_i - y_i|^2]/\mathbb{E}[|r_i|^2]$.

Now, let $\mathcal{T}_2$ be the antenna subset that satisfies $\mathcal{T}_{\mathrm{opt1}} \subset \mathcal{T}_2$ and $|\mathcal{T}_{\mathrm{opt1}}| < |\mathcal{T}_2| = |\mathcal{T}_{\mathrm{opt2}}|$. Then, I obtain the following inequalities:

$$\mathcal{R}^{\mathrm{dl}}(\mathcal{T}_{\mathrm{opt1}}) < \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_2) \le \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_{\mathrm{opt2}})$$

where $\mathcal{R}^{\mathrm{dl}}(\mathcal{T}_{\mathrm{opt1}}) < \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_2)$ follows from leveraging $\mathcal{R}^{\mathrm{dl}}_D(\bar{\mathcal{T}}) > 0$ in (4.12) and $\mathcal{R}^{\mathrm{dl}}(\mathcal{T}_2) \le \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_{\mathrm{opt2}})$ comes from the optimality definition of $\mathcal{T}_{\mathrm{opt2}}$. This completes the proof. ■

Although adding more transmit antennas is not guaranteed to increase the sum rate [41] in general because of a transmit power constraint, Theorem 5 shows that the maximum sum rate increases with the number of selected transmit antennas N_t even with the coarse quantization at the user mobile. This result was also shown to be true for high-resolution ADC systems [126]. Now I will show that the sum rate loss $\mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}})$ has a different property compared to the high-resolution ADC systems where the loss monotonically increases with P and converges to an upper bound [126]. Having $\mathcal{T}_2 = \mathcal{S}$, $\mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}})$ can be considered as the sum rate loss due to antennas selection and minimized to zero by increasing the transmit power constraint P.

Corollary 6. *Let $\mathcal{T}_1 \subset \mathcal{T}_2 \subseteq \mathcal{S}$, then the achievable sum rate loss $\mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}}) = \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_2) - \mathcal{R}^{\mathrm{dl}}(\mathcal{T}_1)$ goes to zero under coarse quantization as the transmit power constraint P increases*

$$\mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}}) \to 0 \quad \textit{as } P \to \infty.$$

In addition, the achievable rate converges to $\mathcal{R}^{\mathrm{dl}}(\mathcal{T}) \to N_{\mathrm{MS}} \log_2 \left(1 + \frac{\alpha_b}{1-\alpha_b}\right)$ as $P \to \infty$.

Proof. If $P \to \infty$, the achievable sum rate loss in (4.12) goes to zero and the sum rate in (4.10) converges to $N_{\mathrm{MS}} \log_2 \left(1 + \frac{\alpha_b}{1-\alpha_b}\right)$. ■

Unlike the high-resolution ADC system, this result suggests that antenna selection can have the marginal rate loss from the system using the entire antennas by increasing P.

Corollary 7. *Let $\mathcal{T}_1 \subset \mathcal{T}_2 \subseteq \mathcal{S}$. The transmit power constraint that leads to the maximum sum rate loss from not using antennas in $\bar{\mathcal{T}} = \mathcal{T}_2 - \mathcal{T}_1$ is*

$$P_D^{\max} = \sqrt{\frac{\mathrm{tr}(\mathbf{Q})\mathrm{tr}(\mathbf{K})}{1-\alpha_b}} \tag{4.14}$$

where $\mathbf{Q} = (\mathbf{H}_{\mathcal{T}_2}^{\mathrm{dl}}\mathbf{H}_{\mathcal{T}_2}^{\mathrm{dl}\,H})^{-1}$ and $\mathbf{K} = (\mathbf{H}_{\mathcal{T}_1}^{\mathrm{dl}}\mathbf{H}_{\mathcal{T}_1}^{\mathrm{dl}\,H})^{-1}$, and the maximum sum rate loss is

$$\mathcal{R}_D^{\mathrm{dl,max}}(\bar{\mathcal{T}}) = N_{\mathrm{MS}} \log_2 \left(1 + \frac{\alpha_b\big(\mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{Q})\big)}{\mathrm{tr}(\mathbf{Q}) + (1-\alpha_b)\mathrm{tr}(\mathbf{K}) + 2\sqrt{(1-\alpha_b)\mathrm{tr}(\mathbf{Q})\mathrm{tr}(\mathbf{K})}} \right). \tag{4.15}$$

Proof. Let $\mathbf{Q} = (\mathbf{H}_{\mathcal{T}_2}^{\mathrm{dl}}\mathbf{H}_{\mathcal{T}_2}^{\mathrm{dl}\,H})^{-1}$ and $\mathbf{\Psi}_{\bar{\mathcal{T}}} = \mathbf{Q}\mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}}(\mathbf{I}_{|\bar{\mathcal{T}}|} - \mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}\,H}\mathbf{Q}\mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}})^{-1}\mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}\,H}\mathbf{Q}$. The derivative of (4.12) with respect to the transmit power constraint is derived as

$$\frac{d\mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}})}{dP} = \frac{\alpha_b N_{\mathrm{MS}} \mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})\big(\mathrm{tr}(\mathbf{Q})^2 + \mathrm{tr}(\mathbf{Q})\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}}) + (\alpha_b - 1)P^2\big)}{\Gamma_{\bar{\mathcal{T}}}} \tag{4.16}$$

where $\Gamma_{\bar{\mathcal{T}}} = \ln 2(\mathrm{tr}(\mathbf{Q}) + P)(\mathrm{tr}(\mathbf{Q}) + \mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}}) + P)(\mathrm{tr}(\mathbf{Q}) + (1-\alpha_b)P)(\mathrm{tr}(\mathbf{Q}) + \mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}}) + (1-\alpha_b)P)$. Since $0 < \alpha_b < 1$ and $\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}}) > 0$, by setting (4.16) to be zero, $P_D^{\max}$ is derived as

$$P_D^{\max} = \sqrt{\frac{\mathrm{tr}(\mathbf{Q})^2 + \mathrm{tr}(\mathbf{Q})\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})}{1-\alpha_b}}. \tag{4.17}$$

Using $\mathrm{tr}\big((\mathbf{H}_{\mathcal{T}_1}^{\mathrm{dl}}\mathbf{H}_{\mathcal{T}_1}^{\mathrm{dl}\,H})^{-1}\big) = \mathrm{tr}(\mathbf{Q}) + \mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})$, the maximizer $P_D^{\max}$ (4.17) is rewritten as (4.14). With respect to the transmit power constraint P, the maximum sum rate loss for $\mathcal{T}_1$ and $\mathcal{T}_2$ can be determined by putting $P = P_D^{\max}$ into (4.13), which leads to (4.15). This completes the proof. ■

According to Corollary 7, the transmit antenna selection in low-resolution ADC systems always achieves the sum rate with the rate loss less than $\mathcal{R}_D^{\mathrm{dl,max}}(\bar{\mathcal{T}})$ in (4.15) for a selected antenna subset. Note that if there is no quantization error, i.e., $\alpha_b = 1$, $P_D^{\max}$ goes to infinity. Then, the sum rate loss cannot decrease with P in the perfect quantization system, which corresponds to the upper bound of the sum rate loss in [126]. Since $\Gamma_{\bar{\mathcal{T}}}$ and $\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}})$ are positive, $\partial \mathcal{R}_D^{\mathrm{dl}}(\bar{\mathcal{T}})/\partial P$ in (4.16) becomes positive when $P < P_D^{\max}$ and negative when $P > P_D^{\max}$, i.e., for $P < P_D^{\max}$, the sum rate loss increases as P increases, and for $P > P_D^{\max}$, the loss decreases to zero as P increases. Therefore, (4.14) can be considered as the reference power constraint that is required to reduce the sum rate loss while achieving a reasonable sum rate.

Corollary 8. *The maximum rate loss in low-resolution ADC systems is less than that in high-resolution ADC systems, i.e.,* $\mathcal{R}_D^{\mathrm{dl,max}}(\bar{\mathcal{T}}; b) \leq \mathcal{R}_D^{\mathrm{dl,max}}(\bar{\mathcal{T}}; \infty)$.

Proof. Since $\mathrm{tr}(\mathbf{\Psi}_{\bar{\mathcal{T}}}) = \mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{Q}) > 0$ from Lemma 5, where $\mathbf{Q} = (\mathbf{H}_{\mathcal{T}_2}^{\mathrm{dl}} \mathbf{H}_{\mathcal{T}_2}^{\mathrm{dl}\,H})^{-1}$, $\mathbf{K} = (\mathbf{H}_{\mathcal{T}_1}^{\mathrm{dl}} \mathbf{H}_{\mathcal{T}_1}^{\mathrm{dl}\,H})^{-1}$, and $\mathbf{\Psi}_{\bar{\mathcal{T}}} = \mathbf{Q}\mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}} (\mathbf{I}_{|\bar{\mathcal{T}}|} - \mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}\,H} \mathbf{Q} \mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}})^{-1} \mathbf{H}_{\bar{\mathcal{T}}}^{\mathrm{dl}\,H} \mathbf{Q}$, the maximum rate loss in (4.15) is a monotonically increasing function with respect to α_b with $0 < \alpha_b < 1$. When $\alpha_b \to 1$, the considered system becomes equivalent to the high-resolution ADC system. ■

Based on Corollary 8, the transmit antenna selection can be more effective in low-resolution ADC systems as the rate loss is smaller than that in high-resolution ADC systems.

4.4 Uplink Receive Antenna Selection

In this section, I examine the key difference of the receive antenna selection problem at the BS with low-resolution ADCs from the conventional problem and propose a quantization-aware receive antenna selection method.

4.4.1 Capacity Maximization Problem

For the considered UL narrowband system in (4.6), the capacity can be expressed as

$$\mathcal{R}^{\mathrm{ul}}(\mathcal{K}) = \log_2 \left| \mathbf{I}_{N_r} + \rho \alpha_b^2 \big(\alpha_b^2 \mathbf{I}_{N_r} + \mathbf{R}_{\mathbf{q}^{\mathrm{ul}}\mathbf{q}^{\mathrm{ul}}} \big)^{-1} \mathbf{H}_{\mathcal{K}}^{\mathrm{ul}} \mathbf{H}_{\mathcal{K}}^{\mathrm{ul}\,H} \right| \tag{4.18}$$

where $\mathbf{R}_{\mathbf{q}^{\mathrm{ul}}\mathbf{q}^{\mathrm{ul}}}$ is given in (4.7). Note from (4.18) that in the low-resolution ADC system, the capacity involves the quantization noise covariance matrix $\mathbf{R}_{\mathbf{q}^{\mathrm{ul}}\mathbf{q}^{\mathrm{ul}}}$ as a penalty term for each antenna. I use $\mathbf{f}_i^H$ to indicate the ith row of $\mathbf{H}^{\mathrm{ul}}$ and $\mathcal{K}(i)$ to denote the ith selected antenna.

Remark 11. *Since each diagonal entry of* $\mathbf{R}_{\mathbf{q}^{\mathrm{ul}}\mathbf{q}^{\mathrm{ul}}}$ *contains an aggregated channel gains at each selected antenna* $\|\mathbf{f}_{\mathcal{K}(i)}\|^2$*, the tradeoff between the channel gain from adding antennas and its influence on quantization error needs to be considered in antenna selection.*

Using the capacity in (4.18), we formulate the UL receive antenna selection problem as follows:

$$\mathcal{P}2: \qquad \mathcal{K}^\star = \underset{\mathcal{K}\subseteq\mathcal{S}:|\mathcal{K}|=N_r\geq N_{\mathrm{MS}}}{\operatorname{argmax}} \mathcal{R}^{\mathrm{ul}}\big(\mathcal{K}\big), \tag{4.19}$$

where $\mathcal{S} = \{1, \ldots, N_{\mathrm{BS}}\}$. Notice that the large number of BS antennas N_{BS} makes it almost infeasible to perform an exhaustive search. Accordingly, to avoid searching over all possible antenna subsets $\mathcal{K}$, I propose two algorithms: a quantization-aware antenna selection algorithm based on the greedy approach and a Markov chain Monte Carlo (MCMC)-based algorithm.

4.4.2 Greedy Approach

Now, let $\mathbf{D}_{\mathcal{K}} = \mathrm{diag}\{1+\rho(1-\alpha_b)\|\mathbf{f}_{\mathcal{K}(i)}\|^2\}$ be the diagonal matrix with $(1+\rho(1-\alpha_b)\|\mathbf{f}_{\mathcal{K}(i)}\|^2)$ for $i = 1, \ldots, N_r$ at its diagonal entries. Then, the capacity in (4.18) can be rewritten as

$$\mathcal{R}^{\mathrm{ul}}(\mathcal{K}) = \log_2 \left| \mathbf{I}_{N_r} + \rho\alpha_b \mathbf{D}_{\mathcal{K}}^{-1} \mathbf{H}_{\mathcal{K}}^{\mathrm{ul}} \mathbf{H}_{\mathcal{K}}^{\mathrm{ul}\,H} \right|. \tag{4.20}$$

Let $\mathcal{K}_t$ be the set of selected antennas during the first t greedy selections and $\mathbf{H}_{\mathcal{K}_t \cup \{j\}}$ be the channel matrix of t selected antennas during the first t greedy selections and a candidate antenna $j \in \mathcal{S} \setminus \mathcal{K}_t$ at the next selection stage. Then, I formulate the greedy selection problem as

$$J = \underset{j \in \mathcal{S} \setminus \mathcal{K}_t}{\mathrm{argmax}}\, \mathcal{R}^{\mathrm{ul}}(\mathcal{K}_t \cup \{j\}). \tag{4.21}$$

To reduce the complexity of solving the problem in (4.21), I decompose the capacity formula (4.20). At the $(t+1)$th selection stage with a candidate antenna j, I have

$$\begin{aligned} \mathcal{R}^{\mathrm{ul}}(\mathcal{K}_t \cup \{j\}) &= \log_2 \left| \mathbf{I}_{N_r} + \rho\alpha_b \mathbf{D}_{\mathcal{K}_t \cup \{j\}}^{-1} \mathbf{H}_{\mathcal{K}_t \cup \{j\}}^{\mathrm{ul}} \mathbf{H}_{\mathcal{K}_t \cup \{j\}}^{\mathrm{ul}\,H} \right| \\ &= \log_2 \left| \mathbf{I}_{N_{\mathrm{MS}}} + \rho\alpha_b \Big(\mathbf{H}_{\mathcal{K}_t}^{\mathrm{ul}\,H} \mathbf{D}_{\mathcal{K}_t}^{-1} \mathbf{H}_{\mathcal{K}_t}^{\mathrm{ul}} + \frac{1}{d_j} \mathbf{f}_j \mathbf{f}_j^H \Big) \right|. \end{aligned} \tag{4.22}$$

Recall that $\mathbf{f}_j^H$ denotes the jth row of $\mathbf{H}^{\text{ul}}$ and d_j is the corresponding diagonal entry of $\mathbf{D}_{\mathcal{K}_t \cup \{j\}}$.

Using the matrix determinant lemma $|\mathbf{A} + \mathbf{u}\mathbf{v}^H| = |\mathbf{A}|(1 + \mathbf{v}^H \mathbf{A}^{-1}\mathbf{u})$, we rewrite (4.22) as

$$\mathcal{R}^{\text{ul}}(\mathcal{K}_t \cup \{j\}) = \mathcal{R}^{\text{ul}}(\mathcal{K}_t) + \log_2\left(1 + \frac{\rho\alpha_b}{d_j} c_t(j)\right) \tag{4.23}$$

where

$$c_t(j) = \mathbf{f}_j^H \left(\mathbf{I}_{N_{\text{MS}}} + \rho\alpha_b \mathbf{H}_{\mathcal{K}_t}^{\text{ul}\,H} \mathbf{D}_{\mathcal{K}_t}^{-1} \mathbf{H}_{\mathcal{K}_t}^{\text{ul}}\right)^{-1} \mathbf{f}_j. \tag{4.24}$$

To maximize $\mathcal{R}^{\text{ul}}(\mathbf{H}_{\mathcal{K}_t \cup \{j\}})$ given the t selected antennas, the next antenna j which maximizes $c_t(j)/d_j$ needs to be selected at the $(t+1)$th stage as

$$J = \underset{j \in \mathcal{S} \setminus \mathcal{K}_t}{\text{argmax}} \frac{c_t(j)}{d_j}. \tag{4.25}$$

Unlike the criterion with no quantization error in [127], the derived criterion $c_t(j)/d_j$ incorporates (i) the effect of the existing quantization error from the previously selected t antennas to the next antenna j in $c_t(j)$, and (ii) the additional quantization error from the antenna j as a penalty for selecting the antenna j in the form of $1/d_j$. In this regard, solving the problem (4.25) gives the antenna J which offers the best tradeoff between the channel gain from selecting an antenna and its influence on the increase of the quantization error. Note that (4.25) is the generalized antenna selection criterion of the one in [127]; as the number of quantization bits b increases, the quantization gain α_b increases as $\alpha_b \to 1$, which leads to $d_j \to 1$ and $\mathbf{D}_{\mathcal{K}_t} \to \mathbf{I}_t$.

Algorithm 3: Quantization-aware Fast Antenna Selection

Initialization: $\mathcal{S} = \{1, \ldots, N_{\mathrm{BS}}\}$, $\mathcal{K} = \emptyset$ and $\mathbf{Q} = \mathbf{I}_{N_{\mathrm{MS}}}$.
Compute initial antenna gain and compute penalty:
$c(j) = \|\mathbf{f}_j\|^2$ and $d_j = 1 + \rho(1 - \alpha_b)\|\mathbf{f}_j\|^2$ for $j \in \mathcal{S}$.
for $t = 1 : N_r$ **do**
  Select antenna J using (4.25): $J = \mathrm{argmax}_{j \in \mathcal{S}}\, c(j)/d_j$.
  Update sets: $\mathcal{S} = \mathcal{S} \setminus \{J\}$ and $\mathcal{K} = \mathcal{K} \cup \{J\}$
  Compute: $\mathbf{a} = \left(c(J) + \frac{d_J}{\rho\alpha_b}\right)^{-\frac{1}{2}} \mathbf{Q}\mathbf{f}_J$ and $\mathbf{Q} = \mathbf{Q} - \mathbf{a}\mathbf{a}^H$.
  Update $c(j) = c(j) - |\mathbf{f}_j^H \mathbf{a}|^2$ for $j \in \mathcal{S}$.
return $\mathcal{K}$;

I now propose a quantization-aware fast antenna selection (QFAS) algorithm by using the derived criterion in (4.25) and modifying the selection algorithm in [127] without increasing the overall complexity. Unlike the perfect quantization case, the quantization error term d_j needs to be computed prior to selection. At each selection stage, the proposed algorithm adopts (4.25). To compute $c_t(j)$ in (4.24), we define $\mathbf{Q}_t = \left(\mathbf{I}_{N_{\mathrm{MS}}} + \rho\alpha_b \mathbf{H}_{\mathcal{K}_t}^H \mathbf{D}_{\mathcal{K}_t}^{-1} \mathbf{H}_{\mathcal{K}_t}\right)^{-1}$. Then, $c_t(j)$ is updated as

$$c_{t+1}(j) = \mathbf{f}_j^H \mathbf{Q}_{t+1} \mathbf{f}_j \overset{(a)}{=} c_t(j) - |\mathbf{f}_j^H \mathbf{a}|^2.$$

where (a) follows from that $\mathbf{Q}_t$ can be efficiently updated by using the matrix inversion lemma as $\mathbf{Q}_{t+1} = \mathbf{Q}_t - \mathbf{a}\mathbf{a}^H$ with $\mathbf{a} = \left(c_t(J) + \frac{d_J}{\rho\alpha_b}\right)^{-1/2} \mathbf{Q}_t \mathbf{f}_J$. The proposed QFAS algorithm is described in Algorithm 3. Note that the complexity for step 5 and 6 are $\mathcal{O}(N_r N_{\mathrm{MS}}^2)$ and $\mathcal{O}(N_r N_{\mathrm{MS}} N_{\mathrm{BS}})$, respectively. The overall complexity becomes $\mathcal{O}(N_r N_{\mathrm{MS}} N_{\mathrm{BS}})$ because of $(N_{\mathrm{BS}} \gg N_{\mathrm{MS}})$. Thus, the proposed algorithm does not increase the overall complexity from the conventional algorithm [127], which provides the opportunity to be practically

implemented.

Now, the performance of the proposed QFAS method is analyzed by using submodularity.

Definition 1 (Submodularity)**.** *If $\mathcal{V}$ is a finite set, a submodular function is a set function $f : 2^{\mathcal{V}} \to \mathbb{R}$ which meets the following condition: for every $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$ with $\mathcal{A} \subseteq \mathcal{B}$ and every element $v \in \mathcal{V} \setminus \mathcal{B}$, f satisfies that $f(\mathcal{A} \cup \{v\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{v\}) - f(\mathcal{B})$.*

Definition 2 (Monotone)**.** *A set function $f : 2^{\mathcal{V}} \to \mathbb{R}$ is monotone if for every $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$, we have that $f(\mathcal{A}) \leq f(\mathcal{B})$. f is said to be normalized if $f(\phi) = 0$, where ϕ denotes the empty set.*

From the definition of a submodular set function, it exhibits a diminishing return property. The following theorem provides a performance lower bound of greedy methods for optimizing submodular objective functions.

Theorem 6 ([128])**.** *For a normalized nonnegative and monotone submodular function $f : 2^{\mathcal{V}} \to \mathbb{R}_+$, let $\mathcal{A}_{\mathrm{G}} \subseteq \mathcal{V}$ be a set with $|\mathcal{A}_{\mathrm{G}}| = k$ obtained by selecting elements one at a time and choosing an element that provides the largest marginal increase in the function value at each time. Let $\mathcal{A}^\star$ be the optimal set that maximizes the value of f with $|\mathcal{A}^\star| = k$. Then, $f(\mathcal{A}_{\mathrm{G}}) \geq (1 - \frac{1}{e}) f(\mathcal{A}^\star)$.*

Based on Theorem 6, it was shown in [41] that the achievable rate of a point-to-point MIMO system is a submodular function, and hence, the greedy

antenna selection algorithm for high-resolution ADC systems provides at least $\left(1-\frac{1}{e}\right)\mathcal{R}^{\text{opt}}$, where $\mathcal{R}^{\text{opt}}$ the achievable rate with the optimal antenna subset for high-resolution ADC systems. I extend this result to the capacity with the quantization error in (4.18).

Corollary 9. *The capacity achieved by the proposed QFAS method is lower bounded by*

$$\mathcal{R}^{\text{ul}}(\mathcal{K}_{\text{qfas}}) \geq \left(1-\frac{1}{e}\right)\mathcal{R}^{\text{ul}}(\mathcal{K}^{\star}). \tag{4.26}$$

Proof. I first need to show that the achievable rate with the quantization error $\mathcal{R}^{\text{ul}}(\mathcal{K})$ in (4.18) is submodular. I define a function $\mathbf{\Gamma}_{\mathcal{K}}$ as

$$\mathbf{\Gamma}_{\mathcal{K}} = \mathbf{I}_{N_r} + \rho\alpha_b^2\left(\alpha_b^2\mathbf{I}_{N_r} + \mathbf{R}_{\mathbf{q}^{\text{ul}}\mathbf{q}^{\text{ul}}}\right)^{-1/2}\mathbf{H}_{\mathcal{K}}^{\text{ul}}\mathbf{H}_{\mathcal{K}}^{\text{ul}\,H}\left(\alpha_b^2\mathbf{I}_{N_r} + \mathbf{R}_{\mathbf{q}^{\text{ul}}\mathbf{q}^{\text{ul}}}\right)^{-1/2}. \tag{4.27}$$

Let $\mathbf{x}_{\mathcal{K}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Gamma}_{\mathcal{K}})$. Since $\mathbf{\Gamma}_{\mathcal{K}}$ is nonsingular, the entropy of $\mathbf{x}_{\mathcal{K}}$ is given as

$$h(\mathbf{x}_{\mathcal{K}}) = \ln|\pi e\mathbf{\Gamma}_{\mathcal{K}}| = N_r\ln(\pi e) + \frac{1}{\log_2 e}\mathcal{R}^{\text{ul}}(\mathcal{K}).$$

Exploiting the form of $\mathbf{R}_{\mathbf{q}^{\text{ul}},\mathbf{q}^{\text{ul}}}$ in (4.7), for any sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{S}$ and element such that $\{s\} \notin \mathcal{B}$ and $\{s\} \in \mathcal{S}$, I have $h(\mathbf{x}_{\{s\}}|\mathbf{x}_{\mathcal{A}}) \geq h(\mathbf{x}_{\{s\}}|\mathbf{x}_{\mathcal{B}})$, i.e., $h(\mathbf{x}_{\mathcal{A}\cup\{s\}}) - h(\mathbf{x}_{\mathcal{A}}) \geq h(\mathbf{x}_{\mathcal{B}\cup\{s\}}) - h(\mathbf{x}_{\mathcal{B}})$. The entropy is submodular and $\mathcal{R}^{\text{ul}}(\mathcal{K})$ in (4.18) is also submodular. In addition, $\mathcal{R}^{\text{ul}}(\mathcal{K})$ is normalized and monotone. Since $\mathcal{R}^{\text{ul}}(\mathcal{K})$ (4.18) is submodular, monotone, and nonnegative, the capacity with the greedy maximization in (4.21) is lower bounded by (4.26) from Theorem 6. Thus, the capacity with the proposed QFAS is also lower bounded by (4.26). ■

4.4.3 Markov Chain Monte Carlo Approach

To find a numerical upper bound of the capacity for the antenna selection without exhaustive search, I provide an algorithm that finds an approximated optimal solution for the problem $\mathcal{P}2$ in (4.19). The adaptive MCMC-based selection method [42] is modified by adopting (4.18) for formulating an original probability density function (PDF). To develop the MCMC-based algorithm for low-resolution ADC systems, I define a binary vector $\boldsymbol{\omega} \in \{0,1\}^{N_{\mathrm{BS}}}$ with $\|\boldsymbol{\omega}\|_0 = N_r$ where 1 indicates that the corresponding receive antenna is selected and vice versa. Here, $\boldsymbol{\omega}$ can be considered as a codeword of the codebook $\mathcal{V}$ that contains all possible combinations of antenna subsets of size N_r, i.e., $|\mathcal{V}| = \binom{N_{\mathrm{BS}}}{N_r}$. Now, let the original PDF be

$$\pi(\boldsymbol{\omega}) \triangleq \exp\left(\frac{1}{\tau}\mathcal{R}^{\mathrm{ul}}(\boldsymbol{\omega})\right)/\Gamma \tag{4.28}$$

where τ is a rate constant and Γ is a normalizing factor. I reformulate $\mathcal{P}2$ in (4.19) as

$$\boldsymbol{\omega}^{\star} = \underset{\boldsymbol{\omega}\in\mathcal{V}}{\operatorname{argmax}}\, \pi(\boldsymbol{\omega}). \tag{4.29}$$

To solve (4.29), the proposed algorithm uses a Metropolized independence sampler (MIS) [129] for the MCMC sampling, which is performed as follows: for a given current sample $\boldsymbol{\omega}(i)$, a new sample $\boldsymbol{\omega}^{\mathrm{new}}$ is selected according to a proposal distribution $q(\boldsymbol{\omega})$. Based on a accepting probability $p_{\mathrm{accept}}(\pi, q) = \min\{1, \frac{\pi(\boldsymbol{\omega}^{\mathrm{new}})}{\pi(\boldsymbol{\omega}(i))}\frac{q(\boldsymbol{\omega}(i))}{q(\boldsymbol{\omega}^{\mathrm{new}})}\}$, a next sample is obtained as $\boldsymbol{\omega}(i+1) = \boldsymbol{\omega}^{\mathrm{new}}$ if accepted, or I have $\boldsymbol{\omega}(i+1) = \boldsymbol{\omega}(i)$, otherwise. After N_{MCMC} iterations,

I have a set of $(1 + N_{\text{MCMC}})$ samples including an initial sample $\boldsymbol{\omega}(0)$, i.e., $\{\boldsymbol{\omega}(0), \boldsymbol{\omega}(1), \ldots, \boldsymbol{\omega}(N_{\text{MCMC}})\}$.

For the proposal distribution, the product of Bernoulli distributions is used, which is given as

$$q(\boldsymbol{\omega}; \mathbf{p}) = \frac{1}{\Gamma'} \prod_{j=1}^{N_{\text{BS}}} p_j^{[\boldsymbol{\omega}_v]_j} (1 - p_j)^{1 - [\boldsymbol{\omega}_v]_j} \tag{4.30}$$

where p_j represents the probability of receive antenna j to be selected and $[\boldsymbol{\omega}]_j$ denotes the jth element of $\boldsymbol{\omega}$. Since Γ' is unnecessary for computing the accepting probability p_{accept}, $q(\boldsymbol{\omega}; \mathbf{p})$ is used without Γ'. Similarly, $\pi(\boldsymbol{\omega})$ is also used without the normalizing factor Γ for p_{accept}.

The selection probabilities $\mathbf{p}$ will be adaptively updated at each iteration in the algorithm to increase the similarity between $\pi(\boldsymbol{\omega})$ and $q(\boldsymbol{\omega}; \mathbf{p})$. We update the probability entries p_j to update the proposal distribution $q(\boldsymbol{\omega}; \mathbf{p})$ by minimizing the Kullback-Leibler divergence between $\pi(\boldsymbol{\omega})$ and $q(\boldsymbol{\omega}; \mathbf{p})$ [42]. Then, the update at $(t + 1)$th iteration becomes

$$p_j^{(t+1)} = p_j^{(t)} + r^{(t+1)} \left(\frac{1}{N_{\text{MCMC}}} \sum_{i=1}^{N_{\text{MCMC}}} [\boldsymbol{\omega}(i)]_j - p_j^{(t)} \right) \tag{4.31}$$

where $r^{(t)}$ is a sequence of decreasing step sizes that satisfies $\sum_{t=0}^{\infty} r^{(t)} = \infty$ and $\sum_{t=0}^{\infty} (r^{(t)})^2 < \infty$ [130]. Finally, Algorithm 4 describes the quantization-aware MCMC-based antenna selection (QMCMC-AS) algorithm. Algorithm 4 stops once it reaches a stopping criterion, which we set as the number of maximum iterations τ_{stop}. The computational complexity of the QMCMC-AS method is $\mathcal{O}(N_r N_{\text{MS}}^2 N_{\text{MCMC}} \tau_{\text{stop}})$ [42]. Note that unlike the QFAS method, the

Algorithm 4: Quantization-aware MCMC-Antenna Selection

Initialization: Set original distribution $\pi(\boldsymbol{\omega})$ as (4.28) and proposal distribution $q(\boldsymbol{\omega}; \mathbf{p})$ as (4.30) without normalizing factors. Set $\boldsymbol{\omega}(0)$ as selected antennas from Algorithm 3, and $\hat{\boldsymbol{\omega}}_C^* = \boldsymbol{\omega}(0)$. Set $p_j^{(0)} = 1/2$, $\forall j$.
for $t = 1 : \tau_{\text{stop}}$ **do**
  Run the MIS to draw samples $\{\boldsymbol{\omega}(i)\}_{i=1}^{N_{\text{MCMC}}}$ with $p_{\text{accept}}(\pi, q)$
  If $|\boldsymbol{\omega}(i)| > N_r$, keep only first N_r entries with largest $p_j^{(k)}$. If $|\boldsymbol{\omega}(i)| < N_r$, randomly select $(N_r - |\boldsymbol{\omega}(i)|)$ more antennas.
  Update $p_j^{(t)}$ according to (4.31).
  If $\pi(\boldsymbol{\omega}(i)) > \pi(\hat{\boldsymbol{\omega}}_C^*)$, for $i = 1, \ldots, N_{\text{MCMC}}$, set $\pi(\hat{\boldsymbol{\omega}}_C^*) = \pi(\boldsymbol{\omega}(i))$.
return $\hat{\boldsymbol{\omega}}_C^*$;

complexity of the QMCMC-AS method involves additional parameters such as the sample size N_{MCMC} and the number of iterations τ_{stop}. When $\binom{N_{\text{BS}}}{N_r}$ is large, the QMCMC-AS method is required to have large N_{MCMC} and τ_{stop} to find a good subset of antennas [117]. Accordingly, the complexity of the QMCMC-AS can be unnecessarily high. Thus, I use the QMCMC-AS method only to provide an approximated optimal performance as a benchmark.

4.5 Extension to Wideband Channels

In this section, I derive the multiuser OFDM system models with quantization error and extend the DL and UL antenna selection problems to the wideband OFDM system.

4.5.1 Downlink OFDM Communications

Let N_{sc} be the number of subcarriers for the OFDM system and $\mathbf{u}_n \in \mathbb{C}^{N_{\text{MS}}}$ be the frequency domain symbol vector of N_{MS} MSs at the nth subcarrier after ZF precoding for the selected antennas in $\mathcal{T}$. I consider bulk selection where all subcarriers share a same antenna subset. Then, $\mathbf{u}_n \in \mathbb{C}^{N_{\text{MS}}}$ is

$$\mathbf{u}_n = \mathbf{W}_{\text{BB},n}(\mathcal{T})\mathbf{P}_n^{1/2}\mathbf{s}_n^{\text{dl}}$$

where $\mathbf{W}_{\text{BB},n}(\mathcal{T}) \in \mathbb{C}^{N_t \times N_{\text{MS}}}$ is the ZF precoder, $\mathbf{P}_n = \text{diag}\{p_{n,1}, \ldots, p_{n,N_{\text{MS}}}\}$ is the power allocation matrix, and $\mathbf{s}_n = [s_{n,1}, s_{n,2}, \ldots, s_{n,N_{\text{MS}}}]^T$ is the frequency symbol vector for the nth subcarrier. Let $\mathbf{x}_n^{\text{dl}}$ be the DL OFDM symbol vectors at time n. Assuming equal transmit power allocation $p_{n,u} = p_{\mathcal{T}}$, $\forall n, u$, I stack $\mathbf{x}_n^{\text{dl}}$ for N_{sc} time duration $\underline{\mathbf{x}} = [\mathbf{x}_1^{\text{dl}\,T}, \mathbf{x}_2^{\text{dl}\,T}, \ldots, \mathbf{x}_{N_{\text{sc}}}^{\text{dl}\,T}]^T \in \mathbb{C}^{N_{\text{sc}}N_t}$, which is

$$\begin{aligned}
\underline{\mathbf{x}}^{\text{dl}} &= (\mathbf{W}_{\text{DFT}}^H \otimes \mathbf{I}_{N_t})\underline{\mathbf{u}} \\
&= \sqrt{p_{\mathcal{T}}}(\mathbf{W}_{\text{DFT}}^H \otimes \mathbf{I}_{N_t})\text{BlkDiag}\{\mathbf{W}_{\text{BB},1}(\mathcal{T}), \mathbf{W}_{\text{BB},2}(\mathcal{T}), \ldots, \mathbf{W}_{\text{BB},N_{\text{sc}}}(\mathcal{T})\}\underline{\mathbf{s}}^{\text{dl}} \\
&= \sqrt{p_{\mathcal{T}}}(\mathbf{W}_{\text{DFT}}^H \otimes \mathbf{I}_{N_t})\underline{\mathbf{W}}_{\text{BB}}\underline{\mathbf{s}}^{\text{dl}}
\end{aligned}$$

where $\mathbf{W}_{\text{DFT}}$ is the normalized N_{sc}-point DFT matrix, $\underline{\mathbf{u}} = [\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_{N_{\text{sc}}}^T]^T \in \mathbb{C}^{N_{\text{sc}}N_t}$, $\underline{\mathbf{s}}^{\text{dl}} = [\mathbf{s}_1^{\text{dl}\,T}, \mathbf{s}_2^{\text{dl}\,T}, \ldots, \mathbf{s}_{N_{\text{sc}}}^{\text{dl}\,T}]^T \in \mathbb{C}^{N_{\text{sc}}N_{\text{MS}}}$, and the block diagonal matrix $\underline{\mathbf{W}}_{\text{BB}} = \text{BlkDiag}\{\mathbf{W}_{\text{BB},1}(\mathcal{T}), \ldots, \mathbf{W}_{\text{BB},N_{\text{sc}}}(\mathcal{T})\}$.

Let the analog received signals of N_{MS} MSs after CP removal at time n be $\mathbf{r}_n^{\text{dl}} \in \mathbb{C}^{N_{\text{MS}}}$. The vectors of received signals $\mathbf{r}_n^{\text{dl}}$ for N_{sc} time duration are

stacked as

$$\begin{aligned}\underline{\mathbf{r}}^{\mathrm{dl}} &= \underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{x}}^{\mathrm{dl}} + \underline{\mathbf{n}}^{\mathrm{dl}} \\ &= \sqrt{p_{\mathcal{T}}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}(\mathbf{W}_{\mathrm{DFT}}^{H} \otimes \mathbf{I}_{N_t})\underline{\mathbf{W}}_{\mathrm{BB}}\underline{\mathbf{s}}^{\mathrm{dl}} + \underline{\mathbf{n}}^{\mathrm{dl}} \end{aligned} \tag{4.32}$$

where $\underline{\mathbf{r}}^{\mathrm{dl}} = [\mathbf{r}_1^{\mathrm{dl}\,T}, \mathbf{r}_2^{\mathrm{dl}\,T}, \ldots, \mathbf{r}_{N_{\mathrm{sc}}}^{\mathrm{dl}\,T}]^T \in \mathbb{C}^{N_{\mathrm{sc}}N_{\mathrm{MS}}}$, and the DL channel matrix for N_t selected transmit antennas $\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}} \in \mathbb{C}^{N_{\mathrm{sc}}N_{\mathrm{MS}} \times N_{\mathrm{sc}}N_t}$ is given as

$$\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}} = \mathrm{BlkCirc}\big\{\mathbf{H}_{\mathcal{T},0}^{\mathrm{dl}}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{H}_{\mathcal{T},L-1}^{\mathrm{dl}}, \cdots, \mathbf{H}_{\mathcal{T},1}^{\mathrm{dl}}\big\} \tag{4.33}$$

where $\mathbf{H}_{\mathcal{T},\ell}^{\mathrm{dl}} \in \mathbb{C}^{N_{\mathrm{MS}} \times N_t}$ is the channel matrix of the selected antennas in $\mathcal{T}$ for the $(\ell+1)$th channel tap, L is the number of channel taps, and $\underline{\mathbf{n}}^{\mathrm{dl}} = [\mathbf{n}_1^{\mathrm{dl}\,T}, \mathbf{n}_2^{\mathrm{dl}\,T}, \ldots, \mathbf{n}_{N_{\mathrm{sc}}}^{\mathrm{dl}\,T}]^T \in \mathbb{C}^{N_{\mathrm{sc}}N_{\mathrm{MS}}}$ denotes the vector of the AWGN noise vectors stacked for N_{sc} time duration.

The received OFDM signals $\underline{\mathbf{r}}^{\mathrm{dl}}$ are quantized at the ADCs. The quantized signal are expressed with the AQNM as [73]

$$\underline{\mathbf{y}}^{\mathrm{dl}} = \alpha_b\sqrt{p_{\mathcal{T}}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}(\mathbf{W}_{\mathrm{DFT}}^{H} \otimes \mathbf{I}_{N_t})\underline{\mathbf{W}}_{\mathrm{BB}}\underline{\mathbf{s}}^{\mathrm{dl}} + \alpha_b\underline{\mathbf{n}}^{\mathrm{dl}} + \underline{\mathbf{q}}^{\mathrm{dl}}$$

where $\underline{\mathbf{q}}^{\mathrm{dl}} = [\mathbf{q}_1^{\mathrm{dl}\,T}, \mathbf{q}_2^{\mathrm{dl}\,T}, \ldots, \mathbf{q}_{N_{\mathrm{sc}}}^{\mathrm{dl}\,T}]^T \in \mathbb{C}^{N_{\mathrm{sc}}N_{\mathrm{MS}}}$ is the additive quantization noise vector and $\underline{\mathbf{q}}^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}}})$. Finally, the quantized signal is combined through a DFT matrix as

$$\begin{aligned}\underline{\mathbf{z}}^{\mathrm{dl}} &= (\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_{\mathrm{MS}}})\underline{\mathbf{y}}^{\mathrm{dl}} \\ &= \alpha_b\sqrt{p_{\mathcal{T}}}(\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_{\mathrm{MS}}})\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}(\mathbf{W}_{\mathrm{DFT}}^{H} \otimes \mathbf{I}_{N_t})\underline{\mathbf{W}}_{\mathrm{BB}}\underline{\mathbf{s}}^{\mathrm{dl}} + (\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_{\mathrm{MS}}})(\alpha_b\underline{\mathbf{n}}^{\mathrm{dl}} + \underline{\mathbf{q}}^{\mathrm{dl}}) \\ &= \alpha_b\sqrt{p_{\mathcal{T}}}\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{W}}_{\mathrm{BB}}\underline{\mathbf{s}}^{\mathrm{dl}} + \underline{\mathbf{v}}^{\mathrm{dl}} \\ &\overset{(a)}{=} \alpha_b\sqrt{p_{\mathcal{T}}}\underline{\mathbf{s}}^{\mathrm{dl}} + \underline{\mathbf{v}}^{\mathrm{dl}}.\end{aligned}$$

Here, $\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}} = (\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_{\mathrm{MS}}})\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}(\mathbf{W}_{\mathrm{DFT}}^{H} \otimes \mathbf{I}_{N_t}) = \mathrm{BlkDiag}\{\mathbf{G}_{\mathcal{T},1}^{\mathrm{dl}}, \cdots, \mathbf{G}_{\mathcal{T},N_{\mathrm{sc}}}^{\mathrm{dl}}\}$ where $\mathbf{G}_{\mathcal{T},n}^{\mathrm{dl}} = \sum_{\ell=0}^{L-1} \mathbf{H}_{\mathcal{T},\ell}^{\mathrm{dl}}\, e^{-\frac{j2\pi(n-1)\ell}{N_{\mathrm{sc}}}}$ is the frequency domain DL channel matrix for subcarrier n, and $\underline{\mathbf{v}}^{\mathrm{dl}} = (\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_{\mathrm{MS}}})(\alpha_b \underline{\mathbf{n}}^{\mathrm{dl}} + \underline{\mathbf{q}}^{\mathrm{dl}}) = [\mathbf{v}_1^{\mathrm{dl}\,T}, \cdots, \mathbf{v}_{N_{\mathrm{sc}}}^{\mathrm{dl}\,T}]^T$. The equality (a) follows from $\underline{\mathbf{W}}_{\mathrm{BB}} = \mathrm{BlkDiag}\{\mathbf{W}_{\mathrm{BB},1}(\mathcal{T}), \cdots, \mathbf{W}_{\mathrm{BB},N_{\mathrm{sc}}}(\mathcal{T})\} = \underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}\,H}(\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}\,H})^{-1}$, i.e.,

$$\mathbf{W}_{\mathrm{BB},n}(\mathcal{T}) = \mathbf{G}_{\mathcal{T},n}^{\mathrm{dl}\,H}(\mathbf{G}_{\mathcal{T},n}^{\mathrm{dl}}\mathbf{G}_{\mathcal{T},n}^{\mathrm{dl}\,H})^{-1}.$$

Under coarse quantization, the received digital signal after DFT for subcarrier n becomes

$$\mathbf{z}_n^{\mathrm{dl}} = \alpha_b\sqrt{p_{\mathcal{T}}}\mathbf{s}_n^{\mathrm{dl}} + \mathbf{v}_n^{\mathrm{dl}}.$$

I compute the covariance matrix of $\mathbf{v}_n^{\mathrm{dl}}$. Let $\mathbf{W}_{\mathrm{MS}} = (\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_{\mathrm{MS}}})$ and $\mathbf{W}_{\mathrm{BS}} = (\mathbf{W}_{\mathrm{DFT}} \otimes \mathbf{I}_{N_t})$. Then, the covariance matrix of $\mathbf{v}_n^{\mathrm{dl}}$ is expressed as

$$\begin{aligned}\mathbf{R}_{\mathbf{v}_n^{\mathrm{dl}}\mathbf{v}_n^{\mathrm{dl}}} &= \alpha_b^2\mathbf{W}_{\mathrm{MS},n}\mathbb{E}\big[\underline{\mathbf{n}}^{\mathrm{dl}}\underline{\mathbf{n}}^{\mathrm{dl}\,H}\big]\mathbf{W}_{\mathrm{MS},n}^{H} + \mathbf{W}_{\mathrm{MS},n}\mathbb{E}\big[\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}\,H}\big]\mathbf{W}_{\mathrm{MS},n}^{H} \\ &= \alpha_b^2\mathbf{I}_{N_{\mathrm{MS}}} + \mathbf{W}_{\mathrm{MS},n}\mathbf{R}_{\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}}}\mathbf{W}_{\mathrm{MS},n}^{H}\end{aligned}$$

where $\mathbf{W}_{\mathrm{MS},n} = ([\mathbf{W}_{\mathrm{DFT}}]_{n,:} \otimes \mathbf{I}_{N_{\mathrm{MS}}})$, and $\mathbf{R}_{\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}}} = \mathbb{E}\big[\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}\,H}\big]$ is the covariance matrix of $\underline{\mathbf{q}}^{\mathrm{dl}}$. To derive $\mathbf{R}_{\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}}}$, I first simplify the precoding matrix $\underline{\mathbf{W}}_{\mathrm{BB}}$ as follows:

$$\begin{aligned}\underline{\mathbf{W}}_{\mathrm{BB}} &= \underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}\,H}(\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}\,H})^{-1} \\ &\overset{(a)}{=} \mathbf{W}_{\mathrm{BS}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H}\mathbf{W}_{\mathrm{MS}}^{H}\left(\mathbf{W}_{\mathrm{MS}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\mathbf{W}_{\mathrm{BS}}^{H}\mathbf{W}_{\mathrm{BS}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H}\mathbf{W}_{\mathrm{MS}}^{H}\right)^{-1} \\ &\overset{(b)}{=} \mathbf{W}_{\mathrm{BS}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H}\left(\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H}\right)^{-1}\mathbf{W}_{\mathrm{MS}}^{-1} \qquad (4.34)\end{aligned}$$

where (a) comes from the definition of $\underline{\mathbf{G}}_{\mathcal{T}}^{\mathrm{dl}} = \mathbf{W}_{\mathrm{MS}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\mathbf{W}_{\mathrm{BS}}^{H}$ and (b) follows from the fact that $\mathbf{W}_{\mathrm{MS}}$, $\mathbf{W}_{\mathrm{BS}}$, and $\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H}$ are invertible. Then, the covariance matrix of $\underline{\mathbf{q}}^{\mathrm{dl}}$ becomes [57, 73]

$$\begin{aligned}\mathbf{R}_{\underline{\mathbf{q}}^{\mathrm{dl}}\underline{\mathbf{q}}^{\mathrm{dl}}} &= \alpha_b(1-\alpha_b)\mathrm{diag}\left\{\mathbb{E}[\underline{\mathbf{r}}^{\mathrm{dl}}\underline{\mathbf{r}}^{\mathrm{dl}\,H}]\right\} \\ &= \alpha_b(1-\alpha_b)\mathrm{diag}\big\{p_{\mathcal{T}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\mathbf{W}_{\mathrm{BS}}^{H}\underline{\mathbf{W}}_{\mathrm{BB}}\underline{\mathbf{W}}_{\mathrm{BB}}^{H}\mathbf{W}_{\mathrm{BS}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H} + \mathbf{I}_{N_{\mathrm{sc}}N_{\mathrm{MS}}}\big\} \\ &\overset{(a)}{=} \alpha_b(1-\alpha_b)(p_{\mathcal{T}}+1)\mathbf{I}_{N_{\mathrm{sc}}N_{\mathrm{MS}}} \qquad (4.35)\end{aligned}$$

where (a) follows from (4.34). Finally, using (4.35), the covariance matrix $\mathbf{R}_{\mathbf{v}_n^{\mathrm{dl}}\mathbf{v}_n^{\mathrm{dl}}}$ becomes $\mathbf{R}_{\mathbf{v}_n^{\mathrm{dl}}\mathbf{v}_n^{\mathrm{dl}}} = (\alpha_b + \alpha_b(1-\alpha_b)p_{\mathcal{T}})\mathbf{I}_{N_{\mathrm{MS}}}$. Accordingly, the SINR of user u for nth subcarrier is given as

$$\mathrm{SINR}_{u,n}(\mathcal{T}) = \frac{\alpha_b p_{\mathcal{T}}}{1+(1-\alpha_b)p_{\mathcal{T}}}. \qquad (4.36)$$

Using (4.36), the transmit antenna selection problem for the OFDM system is formulated as

$$\mathcal{P}3: \qquad \mathcal{T}_{\mathrm{ofdm}}^{\star} = \underset{\mathcal{T}\subseteq\mathcal{S}:|\mathcal{T}|=N_t\geq N_{\mathrm{MS}}}{\mathrm{argmax}} \; \mathcal{R}^{\mathrm{dl,ofdm}}(\mathcal{T})$$

where $\mathcal{R}^{\mathrm{dl,ofdm}}(\mathcal{T}) = \frac{1}{N_{\mathrm{sc}}}\sum_{n=1}^{N_{\mathrm{sc}}}\sum_{u=1}^{N_{\mathrm{MS}}}\log_2\left(1+\mathrm{SINR}_{u,n}(\mathcal{T})\right)$ is the average sum rate. From (4.36), it can be shown that the achievable rate is equal for all u and n. Consequently, maximizing the sum rate is equivalent to maximizing the SINR in (4.36), and it is necessary need to select transmit antennas that maximize the transmit power $p_{\mathcal{T}}$. It is considered that the total transmit power is constrained by P as $\mathrm{tr}\{\mathbb{E}[\underline{\mathbf{x}}^{\mathrm{dl}}\underline{\mathbf{x}}^{\mathrm{dl}\,H}]\} \leq P$. Assuming equal power allocation for each user and subcarrier, I have $\mathrm{tr}\big\{\mathbb{E}[\underline{\mathbf{x}}^{\mathrm{dl}}\underline{\mathbf{x}}^{\mathrm{dl}\,H}]\big\} =$

$p_{\mathcal{T}}\mathrm{tr}\big\{\mathbf{W}_{\mathrm{BS}}^H\underline{\mathbf{W}}_{\mathrm{BB}}\underline{\mathbf{W}}_{\mathrm{BB}}^H\mathbf{W}_{\mathrm{BS}}\big\} = p_{\mathcal{T}}\mathrm{tr}\big\{(\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H})^{-1}\big\}$ and thus, the power allocation $p_{\mathcal{T}}$ with maximum transmit power is given as

$$p_{\mathcal{T}} = \frac{P}{\mathrm{tr}\big\{(\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}\,H})^{-1}\big\}}. \tag{4.37}$$

Remark 12. *The transmit power in* (4.37) *shows that the transmit antenna selection problem for DL OFDM communications in low-resolution ADC systems with ZF precoding and equal power allocation is equivalent to that in high-resolution ADC systems.*

Accordingly, any state-of-the-art transmit antenna selection methods for high-resolution ADC OFDM systems with ZF-precoding can be employed for low-resolution ADC OFDM systems, which was also true for narrowband communications as shown in Section 4.3. In addition, the analysis derived in Section 4.3.2 also holds for the DL OFDM systems.

Corollary 10. *For the multiuser DL OFDM system with ZF precoding and equal power distribution in* (4.32)*, the maximum achievable sum rate of MSs with low-resolution ADCs is monotonically increasing with the number of selected transmit antennas:*

$$\mathcal{R}^{\mathrm{dl,ofdm}}(\mathcal{T}_{\mathrm{opt1}}) < \mathcal{R}^{\mathrm{dl,ofdm}}(\mathcal{T}_{\mathrm{opt2}})$$

where $\mathcal{T}_{\mathrm{opt1}}$ *and* $\mathcal{T}_{\mathrm{opt2}}$ *are the optimal antenna subsets with* $|\mathcal{T}_{\mathrm{opt1}}| < |\mathcal{T}_{\mathrm{opt2}}|$.

Proof. Replace $\mathbf{H}_{\mathcal{T}}^{\mathrm{dl}}$ in the proof of Theorem 5 with $\underline{\mathbf{H}}_{\mathcal{T}}^{\mathrm{dl}}$ and follow the same proof. ■

According to Corollary 10, we need to use as many antennas at the BS for DL OFDM systems with ZF-precoding to maximize the achievable rate even with quantization error at the MSs.

4.5.2 Uplink ODFM Communications

Similarly to the DL OFDM system model with low-resolution ADCs derived in the previous section, the UL ODFM system with low-resolution ADCs can be modeled as follows [131]. Let $\mathbf{x}_n^{\mathrm{ul}} \in \mathbb{C}^{N_{\mathrm{MS}}}$ be a vector of the OFDM symbols of N_{MS} MSs at time n. Let $\underline{\mathbf{x}}^{\mathrm{ul}} = [\mathbf{x}_1^{\mathrm{ul}\,T}, \mathbf{x}_2^{\mathrm{ul}\,T}, \ldots, \mathbf{x}_{N_{\mathrm{sc}}}^{\mathrm{ul}\,T}]^T \in \mathbb{C}^{N_{\mathrm{sc}}N_{\mathrm{MS}}}$, which is given as

$$\underline{\mathbf{x}}^{\mathrm{ul}} = \sqrt{\rho}(\mathbf{W}_{\mathrm{DFT}}^H \otimes \mathbf{I}_{N_{\mathrm{MS}}})\underline{\mathbf{s}}^{\mathrm{ul}}$$

where $\underline{\mathbf{s}}^{\mathrm{ul}} = [\mathbf{s}_1^{\mathrm{ul}\,T}, \mathbf{s}_2^{\mathrm{ul}\,T}, \ldots, \mathbf{s}_{N_{\mathrm{sc}}}^{\mathrm{ul}\,T}]^T \in \mathbb{C}^{N_{\mathrm{sc}}N_{\mathrm{MS}}}$ and $\mathbf{s}_n^{\mathrm{ul}} = [s_{n,1}^{\mathrm{ul}}, s_{n,2}^{\mathrm{ul}}, \ldots, s_{n,N_{\mathrm{MS}}}^{\mathrm{ul}}]^T$.

Let the analog received signals at the BS with N_r selected antennas in $\mathcal{K}$ after CP removal at time n be $\mathbf{r}_n^{\mathrm{ul}} \in \mathbb{C}^{N_r}$. The vectors of received signals $\mathbf{r}_n^{\mathrm{ul}}$ for N_{sc} time duration are stacked as

$$\begin{aligned}\underline{\mathbf{r}}^{\mathrm{ul}} &= \underline{\mathbf{H}}_{\mathcal{K}}^{\mathrm{ul}}\underline{\mathbf{x}}^{\mathrm{ul}} + \underline{\mathbf{n}}^{\mathrm{ul}} \\ &= \sqrt{\rho}\underline{\mathbf{H}}_{\mathcal{K}}^{\mathrm{ul}}(\mathbf{W}_{\mathrm{DFT}}^H \otimes \mathbf{I}_{N_{\mathrm{MS}}})\underline{\mathbf{s}}^{\mathrm{ul}} + \underline{\mathbf{n}}^{\mathrm{ul}}\end{aligned}$$

where $\underline{\mathbf{r}}^{\mathrm{ul}} = [\mathbf{r}_1^{\mathrm{ul}\,T}, \mathbf{r}_2^{\mathrm{ul}\,T}, \ldots, \mathbf{r}_{N_{\mathrm{sc}}}^{\mathrm{ul}\,T}]^T \in \mathbb{C}^{N_{\mathrm{sc}}N_r}$, and the UL channel matrix in the time domain for N_r selected antennas $\underline{\mathbf{H}}_{\mathcal{K}}^{\mathrm{ul}} \in \mathbb{C}^{N_{\mathrm{sc}}N_r \times N_{\mathrm{sc}}N_{\mathrm{MS}}}$ is given as

$$\underline{\mathbf{H}}_{\mathcal{K}}^{\mathrm{ul}} = \mathrm{BlkCirc}\big\{\mathbf{H}_{\mathcal{K},0}^{\mathrm{ul}}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{H}_{\mathcal{K},L-1}^{\mathrm{ul}}, \cdots, \mathbf{H}_{\mathcal{K},1}^{\mathrm{ul}}\big\}$$

where $\mathbf{H}^{\text{ul}}_{\mathcal{K},\ell}$ is the UL channel matrix of the selected antennas for the $(\ell+1)$th channel tap, L is the number of channel taps, and $\underline{\mathbf{n}}^{\text{ul}} = [\mathbf{n}_1^{\text{ul}\,T}, \mathbf{n}_2^{\text{ul}\,T}, \ldots, \mathbf{n}_{N_{\text{sc}}}^{\text{ul}\,T}]^T \in \mathbb{C}^{N_{\text{sc}}N_r}$ denotes the vector of the AWGN noise vectors.

After quantization, the quantized OFDM signals are expressed by adopting the AQNM as [73]

$$\underline{\mathbf{y}}^{\text{ul}} = \alpha_b\sqrt{\rho}\underline{\mathbf{H}}^{\text{ul}}_{\mathcal{K}}(\mathbf{W}^H_{\text{DFT}} \otimes \mathbf{I}_{N_{\text{MS}}})\underline{\mathbf{s}}^{\text{ul}} + \alpha_b\underline{\mathbf{n}}^{\text{ul}} + \underline{\mathbf{q}}^{\text{ul}}$$

where $\underline{\mathbf{q}}^{\text{ul}} = [\mathbf{q}_1^{\text{ul}\,T}, \mathbf{q}_2^{\text{ul}\,T}, \ldots, \mathbf{q}_{N_{\text{sc}}}^{\text{ul}\,T}]^T \in \mathbb{C}^{N_{\text{sc}}N_r}$ is the additive quantization noise vector and $\mathbf{q}_n^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}})$. The covariance matrix $\mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}}$ is [73]

$$\begin{aligned}\mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}} &= \alpha_b(1-\alpha_b)\text{diag}\{\mathbb{E}[\mathbf{r}_n^{\text{ul}}\mathbf{r}_n^{\text{ul}\,H}]\} \\ &= \alpha_b(1-\alpha_b)\text{diag}\big\{\rho\mathbf{B}_{\mathcal{K}}\mathbf{B}_{\mathcal{K}}^H + \mathbf{I}_{N_r}\big\} \end{aligned} \tag{4.38}$$

where $\mathbf{B}_{\mathcal{K}} = [\mathbf{H}^{\text{ul}}_{\mathcal{K},0}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{H}^{\text{ul}}_{\mathcal{K},L-1}, \cdots, \mathbf{H}^{\text{ul}}_{\mathcal{K},1}]$. We note that $\mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}} = \mathbf{R}_{\mathbf{q}_m^{\text{ul}}\mathbf{q}_m^{\text{ul}}}$, $\forall n \neq m$, i.e., $\mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}}$ is independent to subcarriers. Finally, $\underline{\mathbf{y}}^{\text{ul}}$ is combined through a DFT matrix as

$$\begin{aligned}\underline{\mathbf{z}}^{\text{ul}} &= (\mathbf{W}_{\text{DFT}} \otimes \mathbf{I}_{N_r})\underline{\mathbf{y}}^{\text{ul}} \\ &= \alpha_b\sqrt{\rho}(\mathbf{W}_{\text{DFT}} \otimes \mathbf{I}_{N_r})\underline{\mathbf{H}}^{\text{ul}}_{\mathcal{K}}(\mathbf{W}^H_{\text{DFT}} \otimes \mathbf{I}_{N_{\text{MS}}})\underline{\mathbf{s}}^{\text{ul}} + (\mathbf{W}_{\text{DFT}} \otimes \mathbf{I}_{N_r})(\alpha_b\underline{\mathbf{n}}^{\text{ul}} + \underline{\mathbf{q}}^{\text{ul}}) \\ &= \alpha_b\sqrt{\rho}\underline{\mathbf{G}}^{\text{ul}}_{\mathcal{K}}\underline{\mathbf{s}}^{\text{ul}} + \underline{\mathbf{v}}^{\text{ul}}\end{aligned}$$

where $\underline{\mathbf{G}}^{\text{ul}}_{\mathcal{K}} = (\mathbf{W}_{\text{DFT}} \otimes \mathbf{I}_{N_r})\underline{\mathbf{H}}^{\text{ul}}_{\mathcal{K}}(\mathbf{W}^H_{\text{DFT}} \otimes \mathbf{I}_{N_{\text{MS}}}) = \text{BlkDiag}\{\mathbf{G}^{\text{ul}}_{\mathcal{K},1}, \cdots, \mathbf{G}^{\text{ul}}_{\mathcal{K},N_{\text{sc}}}\}$, $\mathbf{G}^{\text{ul}}_{\mathcal{K},n} = \sum_{\ell=0}^{L-1}\mathbf{H}^{\text{ul}}_{\mathcal{K},\ell}e^{-\frac{j2\pi(n-1)\ell}{N_{\text{sc}}}}$, and the noise $\underline{\mathbf{v}}^{\text{ul}} = (\mathbf{W}_{\text{DFT}} \otimes \mathbf{I}_{N_r})(\alpha_b\underline{\mathbf{n}}^{\text{ul}} + \underline{\mathbf{q}}^{\text{ul}}) = [\mathbf{v}_1^{\text{ul}\,T}, \cdots, \mathbf{v}_{N_{\text{sc}}}^{\text{ul}\,T}]^T$. Accordingly, under coarse quantization, the received digital signal after DFT for subcarrier n becomes

$$\mathbf{z}_n^{\text{ul}} = \alpha_b\sqrt{\rho}\mathbf{G}^{\text{ul}}_{\mathcal{K},n}\mathbf{s}_n^{\text{ul}} + \mathbf{v}_n^{\text{ul}}. \tag{4.39}$$

The covariance matrix of $\mathbf{v}_n^{\text{ul}}$ is derived as $\mathbf{R}_{\mathbf{v}_n^{\text{ul}}\mathbf{v}_n^{\text{ul}}} = \alpha_b^2\mathbf{I}_{N_r} + \mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}}$ where $\mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}}$ is defined in (4.38). Using (4.39), the UL capacity for subcarrier n is derived as

$$\mathcal{R}_n^{\text{ul}}(\mathcal{K}) = \log_2 \left| \mathbf{I}_{N_r} + \rho\alpha_b^2(\alpha_b^2\mathbf{I}_{N_r} + \mathbf{R}_{\mathbf{q}_n^{\text{ul}}\mathbf{q}_n^{\text{ul}}})^{-1}\mathbf{G}_{\mathcal{K},n}^{\text{ul}}\mathbf{G}_{\mathcal{K},n}^{\text{ul}\,H} \right|. \tag{4.40}$$

Note that the capacity of the wideband OFDM system for each subcarrier in (4.40) shows similar structure as that of the narrowband system in (4.18).

Since all subcarriers share a same subset of antennas, i.e., $\mathcal{K}$ is same for all subcarriers, the maximization cannot be applied to each subcarrier separately. Accordingly, it is necessary to find the best subset of antennas $\mathcal{K}$ for the entire subcarriers, and the receive antenna selection problem for the wideband UL OFDM system is formulated as

$$\mathcal{P}4: \qquad \mathcal{K}_{\text{ofdm}}^{\star} = \underset{\mathcal{K}\subseteq\mathcal{S}:|\mathcal{K}|=N_r\geq N_{\text{MS}}}{\text{argmax}} \sum_{n=1}^{N_{\text{sc}}} \mathcal{R}_n^{\text{ul}}(\mathcal{K}). \tag{4.41}$$

To solve (4.41), I extend the greedy approach for the narrowband communications in Section 4.4. It is also shown that the MCMC approach can be naturally adopted with modification.

Similarly to (4.21), let $\mathbf{G}_{\mathcal{K}_t\cup\{j\},n}^{\text{ul}}$ be the channel matrix of t selected antennas during the first t greedy selections and a candidate antenna $j \in \mathcal{S}\backslash\mathcal{K}_t$ at the next selection. The greedy maximization problem is formulated as

$$J = \underset{j\in\mathcal{S}\backslash\mathcal{K}_t}{\text{argmax}} \sum_{n=1}^{N_{\text{sc}}} \mathcal{R}_n^{\text{ul}}\left(\mathcal{K}_t \cup \{j\}\right). \tag{4.42}$$

Now, I decompose (4.40). Let $\bar{\mathbf{D}}_{\mathcal{K}_t\cup\{j\}} = \mathbf{I}_{t+1} + \rho(1-\alpha_b)\mathrm{diag}\{\mathbf{B}_{\mathcal{K}_t\cup\{j\}}\mathbf{B}^H_{\mathcal{K}_t\cup\{j\}}\}$. At the $(t+1)$th selection stage, I have

$$\begin{aligned}\mathcal{R}^{\mathrm{ul}}_n(\mathcal{K}_t\cup\{j\}) &= \log_2\left|\mathbf{I}_{N_{\mathrm{MS}}} + \rho\alpha_b \mathbf{G}^{\mathrm{ul}\,H}_{\mathcal{K}_t\cup\{j\},n}\bar{\mathbf{D}}^{-1}_{\mathcal{K}_t\cup\{j\}}\mathbf{G}^{\mathrm{ul}}_{\mathcal{K}_t\cup\{j\},n}\right| \\ &= \log_2\left|\mathbf{I}_{N_{\mathrm{MS}}} + \rho\alpha_b\left(\mathbf{G}^{\mathrm{ul}\,H}_{\mathcal{K}_t,n}\bar{\mathbf{D}}^{-1}_{\mathcal{K}_t}\mathbf{G}^{\mathrm{ul}}_{\mathcal{K}_t,n} + \frac{1}{\bar{d}_j}\mathbf{f}_{n,j}\mathbf{f}^H_{n,j}\right)\right| \\ &= \mathcal{R}^{\mathrm{ul}}_n(\mathcal{K}_t) + \log_2\left(1+\frac{\rho\alpha_b}{d_j}c_{n,t}(j)\right) \end{aligned} \tag{4.43}$$

where $\mathbf{f}^H_{n,j}$ is jth row of $\mathbf{G}^{\mathrm{ul}}_n$, $\bar{d}_j$ is the corresponding diagonal entry of $\bar{\mathbf{D}}_{\mathcal{K}_t\cup\{j\}}$, and $c_{n,t}(j)$ is

$$c_{n,t}(j) = \mathbf{f}^H_{n,j}\left(\mathbf{I}_{N_{\mathrm{MS}}} + \rho\alpha_b\mathbf{G}^{\mathrm{ul}\,H}_{\mathcal{K}_t,n}\bar{\mathbf{D}}^{-1}_{\mathcal{K}_t}\mathbf{G}^{\mathrm{ul}}_{\mathcal{K}_t,n}\right)^{-1}\mathbf{f}_{n,j}. \tag{4.44}$$

With (4.43), the greedy maximization problem in (4.42) reduces to

$$J = \underset{j\in\mathcal{S}\setminus\mathcal{K}_t}{\mathrm{argmax}}\sum_{n=1}^{N_{\mathrm{sc}}}\log_2\left(1+\frac{\rho\alpha_b}{d_j}c_{n,t}(j)\right). \tag{4.45}$$

Therefore, a greedy algorithm that is similar to Algorithm 3 can be used for (4.45). In addition, let $\mathbf{Q}_{n,t} = (\mathbf{I}_{N_{\mathrm{MS}}} + \rho\alpha_b\mathbf{G}^{\mathrm{ul}\,H}_{\mathcal{K}_t,n}\bar{\mathbf{D}}^{-1}_{\mathcal{K}_t}\mathbf{G}^{\mathrm{ul}}_{\mathcal{K}_t,n})^{-1}$. Then, $c_{n,t}(j)$ in (4.44) can also be updated without matrix inversion for each subcarrier as shown in Algorithm 3. Accordingly, the complexity of the proposed QFAS algorithm for the UL OFDM system becomes $\mathcal{O}(N_{\mathrm{sc}}N_rN_{\mathrm{MS}}N_{\mathrm{BS}})$.

Corollary 11. *The capacity of the QFAS method for the UL OFDM system is lower bounded by*

$$\sum_{n=1}^{N_{\mathrm{sc}}}\mathcal{R}^{\mathrm{ul}}_n(\mathcal{K}_{\mathrm{qfas}}) \geq \left(1-\frac{1}{e}\right)\sum_{n=1}^{N_{\mathrm{sc}}}\mathcal{R}^{\mathrm{ul}}_n(\mathcal{K}^\star_{\mathrm{ofdm}}) \tag{4.46}$$

where $\mathcal{K}^\star_{\mathrm{ofdm}}$ is the optimal subset of receive antennas defined in (4.41).

Proof. The class of submodular functions is closed under nonnegative linear combinations, and I showed that the capacity with the quantization error is submodular in the proof of Corollary 9. Consequently, the sum capacity for all carrier frequencies in (4.41) is also submodular. Since the proposed QFAS for the wideband OFDM system solves (4.45), which is equivalent to the greedy maximization in (4.42), from Theorem 6, I derive (4.46). ■

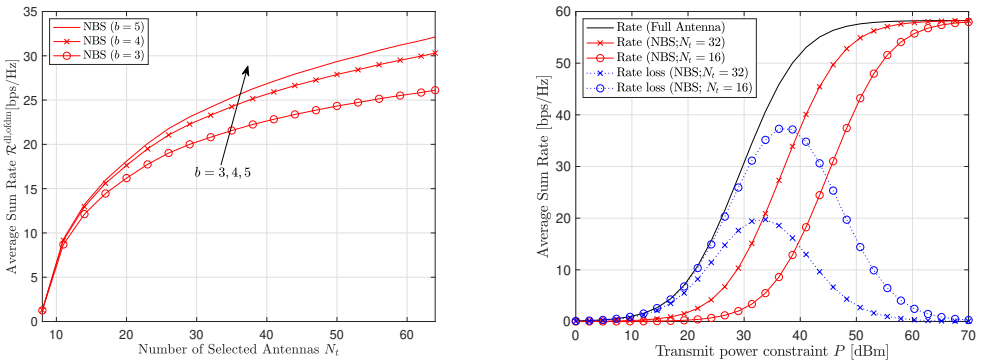
To find an approximated optimal solution, the adaptive MCMC approach described in Section 4.4.3 can be also used. To this end, the original PDF $\pi(\boldsymbol{\omega})$ needs to be modified as

$$\pi(\boldsymbol{\omega}) \triangleq \exp\left(\frac{1}{\tau}\sum_{n=1}^{N_{\mathrm{sc}}} \mathcal{R}_n^{\mathrm{ul}}(\boldsymbol{\omega})\right) / \Gamma_{\mathrm{ofdm}} \tag{4.47}$$

where τ is a rate constant and Γ_{ofdm} is a normalizing factor for the PDF. Then, the adaptive MCMC-based antenna selection method for the OFDM system can be performed similarly to the QMCMC-AS method in Section 4.4.3. The complexity of the QMCMC-AS method for the OFDM system is computed as $\mathcal{O}(N_{\mathrm{sc}} N_r N_{\mathrm{MS}}^2 N_{\mathrm{MCMC}} \tau_{\mathrm{stop}})$.

4.6 Simulation Results

In this section, the theoretical results and proposed methods are validated through simulations. Rayleigh channels are assumed with a zero mean and unit variance for small scale fading. The log-distance pathloss model [132] is adopted for a large scale fading. I consider randomly distributed MSs over a single cell with radius of $1km$ and the minimum distance between the BS



(a) (b)

Figure 4.2: Average sum rate $\mathcal{R}^{\mathrm{dl,ofdm}}$ (a) with respect to the number of selected antennas N_t for $N_{\mathrm{BS}} = 64$ BS antennas, $N_{\mathrm{MS}} = 8$ MSs, $P = 30$ dBm total power constraint, and $b \in \{3, 4, 5\}$ ADC bits, and (b) with respect to the total transmit power constraint P for $N_{\mathrm{BS}} = 128$ BS antennas, $N_{\mathrm{MS}} = 12$ MSs, $N_t = 16$ selected antennas, and $b = 3$ ADC bits.

and MSs to be $100m$. Considering a 2.4 GHz carrier frequency with 10 MHz bandwidth, I use 8.7 dB lognormal shadowing variance and 12 dB noise figure at receivers.

4.6.1 Downlink Transmit Antenna Selection

I consider the DL ODFM system with $N_{\mathrm{sc}} = 64$ subcarriers for channels with $L = 4$ taps. To validate the analysis, the norm-based selection (NBS) method is used in simulations, which selects antennas in the order of channel norm that corresponds to each antenna [42, 117]. Note that the NBS method always provides $\mathcal{T}_1 \subseteq \mathcal{T}_2$ when $|\mathcal{T}_1| \leq |\mathcal{T}_2|$ for the same channel. In Fig. 4.2(a), the average sum rate increases with the number of selected antennas, which validates the derived Theorem 5 and Corollary 10. Fig. 4.2(b) shows the aver-

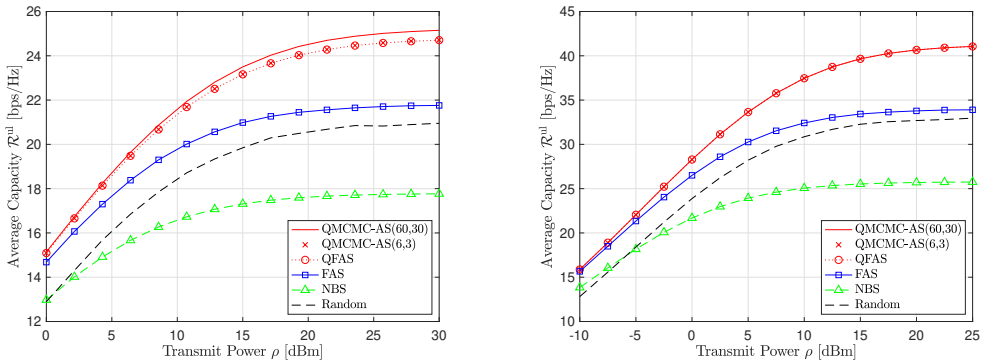


Figure 4.3: Average capacity $\mathcal{R}^{\text{ul}}$ with respect to transmit power ρ for (a) $N_{\text{BS}} = 32$ BS antennas, $N_{\text{MS}} = 8$ MSs, $N_r = 8$ selected antennas, and $b = 3$ quantization bits, and for (b) $N_{\text{BS}} = 128$ BS antennas, $N_{\text{MS}} = 12$ MSs, $N_r = 16$ selected antennas, and $b = 3$ ADC bits.

age sum rate versus the total power constraint P. Unlike the high-resolution ADC systems, there exists a point $P_D^{\max}$ for the maximum rate loss from not using all antennas, and the rate loss decreases after the point $P_D^{\max}$ in (4.14) for the OFDM channel $\underline{\mathbf{H}}^{\text{dl}}$. Theoretical $P_D^{\max}$ for the NBS method with $N_t = 32$ and $N_t = 16$ are 33.1351 dBm and 37.2850, respectively. In addition, the theoretical maximum rate loss in (4.15) for the OFDM channel $\underline{\mathbf{H}}^{\text{dl}}$ with $N_t = 32$ and $N_t = 16$ are 19.8034 bps/Hz and 37.5282 bps/Hz, respectively, which also corresponds to the simulation results.

4.6.2 Uplink Receive Antenna Selection

The proposed algorithms for the UL antenna selection—QFAS and QMCMC-AS methods are evaluated. I also simulate the NBS method [42,117] and the fast antenna selection (FAS) algorithm in [127], which shows a com-

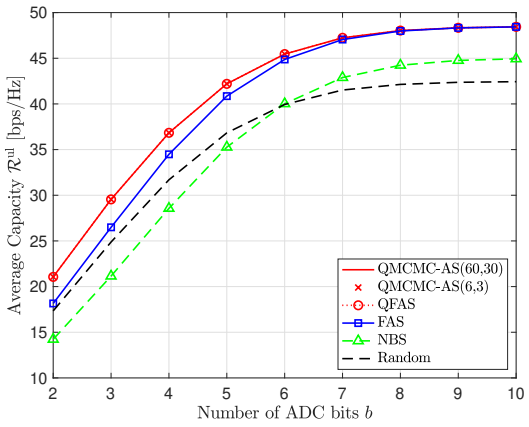
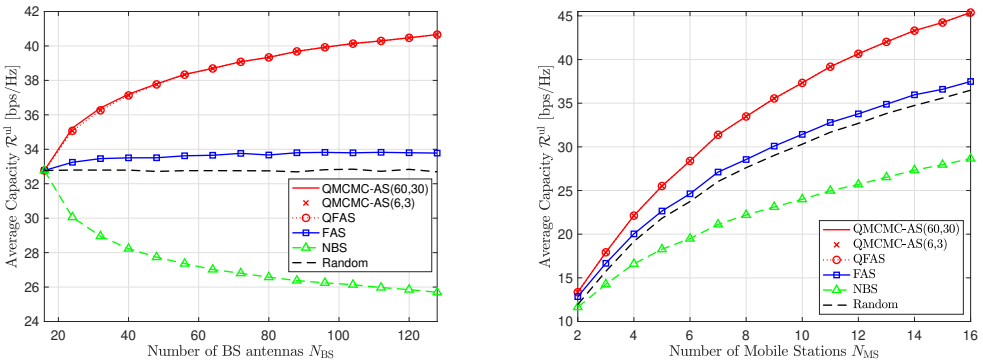


Figure 4.4: Average capacity $\mathcal{R}^{\text{ul}}$ with respect to the number of ADC bits b for $N_{\text{BS}} = 128$ BS antennas, $N_{\text{MS}} = 8$ MSs, $N_r = 16$ selected antennas, and $\rho = 10$ dBm transmit power.

parable performance to the optimal selection under perfect quantization. Although the NBS method presents low performance improvement, because of its low complexity $\mathcal{O}(N_{\text{MS}}N_r)$, it is considered as a reasonable antenna selection method for high-resolution ADC systems [117]. A random selection is simulated to offer a reference performance.

4.6.2.1 Narrowband Communications

In Fig. 4.3(a) the QFAS shows higher capacity than FAS, NBS, and random selection cases. Noting that the initial point of the QMCMC-AS method is the antenna subset from the QFAS, the QMCMC-AS with ($N_{\text{MCMC}} = 6, \tau_{\text{stop}} = 3$) provides no capacity increase from the QFAS method. Although the QMCMC-AS with $(60, 30)$ shows capacity increase from the QFAS method, it is marginal. Accordingly, the QFAS method achieves a near optimal per-



(a) (b)

Figure 4.5: Average capacity $\mathcal{R}^{\text{ul}}$ (a) with respect to the number of BS antennas N_{BS} for $N_{\text{MS}} = 12$ MSs, $N_r = 16$ selected antennas, $\rho = 20$ dBm transmit power, and $b = 3$ ADC bits, and (b) with respect to the number of MSs N_{MS} for $N_{\text{BS}} = 128$ BS antennas, $N_r = 16$ selected antennas, $\rho = 20$ dBm transmit power, and $b = 3$ ADC bits.

formance in terms of capacity with low complexity. The FAS method offers marginal improvement from the random selection case as it ignores quantization error associated with selected antennas. The NBS method shows the worst performance in low-resolution ADC systems, which means that selecting the subset of antennas that gives the largest channel gains not only increases the inter-user interference but also increases quantization error.

With the increased number of receive antennas, selected antennas, and MSs, the trend of the curves in Fig. 4.3(b) is similar to Fig. 4.3(a). The QMCMC-AS with $(60, 30)$, however, shows no improvement from the QFAS. This shows that the QMCMC-AS is not scalable with the number of BS antennas and selected antennas. In both Fig. 4.3(a) and (b), the capacity gap between the QFAS algorithm and the conventional algorithms increases with

the transmit power ρ because the quantization error becomes more dominant than the AWGN as the transmit power increases. In addition, the results in Fig. 4.3 demonstrate that the conventional UL antenna selection approaches are not applicable to the low-resolution ADC receivers.

In Fig. 4.4, in the low-resolution ADC regime, the capacity of the QFAS method is higher than the FAS, NBS, and random selection. This corresponds to the intuition for the proposed method such that considering the quantization error is critical when selecting antennas in low-resolution ADC systems. The capacity of the QFAS and FAS methods converges as the number of ADC bits b increases, thereby showing that the proposed QFAS method is generalized version of the FAS in terms of quantization precision. The NBS method performs better than the random selection in high-resolution ADC regime while it still performs worse in the low-resolution ADC regime. Again, this validates the intuition that the antenna selection approaches for high-resolution ADC systems cannot directly be applied to the low-resolution ADC receivers.

In Fig. 4.5(a), it is observed that there is large improvement from the random selection for the QFAS method as N_{BS} increases whereas the FAS and NBS cannot provide such improvement. Accordingly, the proposed QFAS method can be effective in the large antenna array systems with low-resolution ADCs by efficiently reducing the number of RF chains. The capacity with the NBS method even decreases with the number of BS antennas since the increased candidate antenna size worsens the resulting subset of antennas by significantly increasing quantization error and interference. In Fig. 4.5(b), the

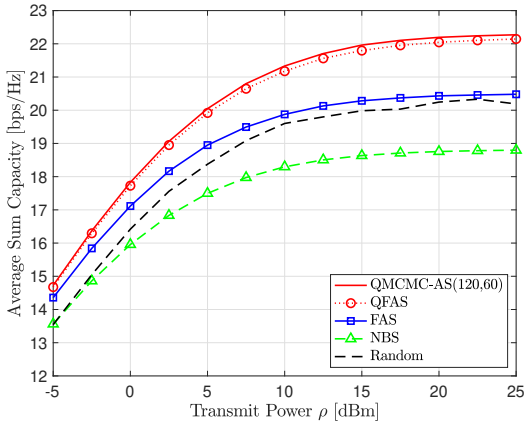


Figure 4.6: Average sum capacity $\frac{1}{N_{\text{sc}}}\sum_n \mathcal{R}_n^{\text{ul}}$ with respect to transmit power ρ for $N_{\text{BS}} = 32$ BS antennas, $N_{\text{MS}} = 8$ MSs, $N_r = 8$ selected antennas, $b = 3$ quantization bits, and $N_{\text{sc}} = 64$ subcarriers with $L = 4$-tap channels.

capacity gap between the QFAS and FAS methods increases with N_{MS}, which is desirable in term of maximizing the sum rate. Overall, the performance improvement with the proposed QFAS becomes larger as more users are served and more antennas are deployed for the fixed number of selected antennas (equivalently RF chains), which is desirable for future communication systems that are likely to serve more users with more antennas.

4.6.2.2 Wideband OFDM Communications

I consider UL wideband ODFM communications with $N_{\text{sc}} = 64$ subcarriers for channels with $L = 4$ taps. Similarly to the simulation results for the narrowband system, the proposed QFAS method in Fig. 4.6 shows higher capacity than the FAS, NBS, and random selection. In addition, the QFAS method almost achieves the capacity of the QMCMC-AS with the increased

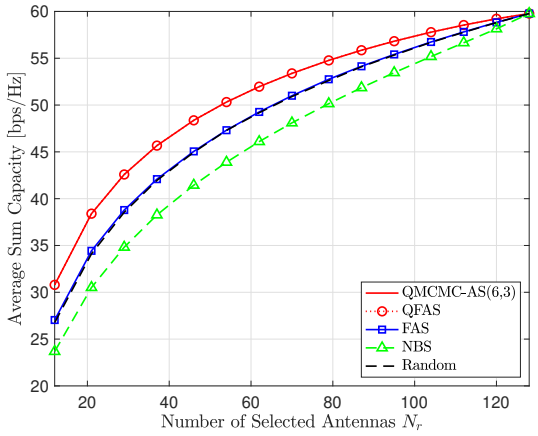


Figure 4.7: Average sum capacity $\frac{1}{N_{\text{sc}}}\sum_n \mathcal{R}_n^{\text{ul}}$ with respect to the number of selected antennas N_r for $N_{\text{BS}} = 128$ BS antennas, $N_{\text{MS}} = 12$ MSs, $b = 3$ ADC bits, $N_{\text{sc}} = 64$ subcarriers with $L = 4$-tab channels, and $\rho = 20$ dBm.

number of sampling and iterations ($N_{\text{MCMC}} = 120, \tau_{\text{stop}} = 60$). Therefore, the QFAS can also achieve near optimal selection performance in wideband OFDM systems while the FAS method shows marginal improvement from the random selection and the NBS method shows the worst performance in low-resolution ADC systems.

In Fig. 4.7, the proposed QFAS performs better than the FAS, NBS, and random selection for any size of antenna subset N_r. The QFAS provides saving of about 10 RF chains on average compared to the FAS and random selection, Such saving can be considered as large for receivers with the relatively small number RF chains compared to the number of antennas. Overall, the simulation results demonstrate that the conventional receive antenna selection is not adequate under non-negligible quantization error and that the proposed

QFAS can effectively incorporate the quantization error in antenna selection.

4.7 Conclusion

In this chapter, I investigated antenna selection at a BS in low-resolution ADC systems to achieve power-efficient wireless communication systems. For downlink narrowband and wideband OFDM systems, I showed that the existing state-of-the-art transmit antenna selection techniques can be applicable to the low-resolution ADC systems when the BS employs the ZF precoding with equal power distribution. In addition, I proved that it is beneficial to use more antennas in terms of maximizing the sum rate. Unlike the high-resolution ADC systems, I validated that the transmit antenna selection can achieve a comparable sum rate to the system that uses all antennas by increasing the total transmit power constraint, which allows to reduce the number of RF chains with marginal sum rate loss. For an uplink narrowband and wideband OFDM systems, I showed that the conventional receive antenna selection criteria are insufficient for the low-resolution ADC systems. The generalized greedy selection criterion provided that capturing the balance between the channel gain and increase in quantization error is critical when there is non-negligible quantization error at the receiver. The propose greedy selection algorithm showed that it guarantees $(1 - \frac{1}{e})$ of the capacity with an optimal antenna subset. I have proposed advanced receiver designs to mitigate quantization error in the last three chapters. In the next chapter, however, I will focus on developing a technique that is used in the higher network layer.

Chapter 5

User Scheduling for Millimeter Wave Hybrid Beamforming Systems with Low-Resolution ADCs

In this chapter[1], I investigate uplink user scheduling for millimeter wave (mmWave) hybrid analog/digital beamforming systems with low-resolution analog-to-digital converters (ADCs). Deriving new scheduling criteria for the mmWave systems, I show that the channel structure in the beamspace, in addition to the channel magnitude and orthogonality, plays a key role in maximizing the achievable rates of scheduled users due to quantization error. The criteria show that to maximize the achievable rate for a given channel gain, the channels of the scheduled users need to have (1) as many propagation paths as possible with unique angle-of-arrivals (AoAs) and (2) even power distribution in the beamspace. Leveraging the derived criteria, an efficient scheduling

[1]This chapter is based on the work published in the journal paper: J. Choi, G. Lee, and B. L. Evans, "User Scheduling for Millimeter Wave Hybrid Beamforming Systems with Low-Resolution ADCs," in *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2401-2414, Apr. 2019. Part of the work was also published in the conference paper: J. Choi, and B. L. Evans, "User Scheduling for Millimeter Wave MIMO Communications with Low-Resolution ADCs," in *Proceedings of IEEE International. Conference. on Communications (ICC)*, May 20-24, 2018, Kansas City, MO, USA. This work was supervised by Prof. Brian L. Evans and valuable feedback and contributions from Dr. Gilwon Lee improved the quality of this work.

algorithm is proposed for mmWave zero-forcing receivers with low-resolution ADCs. I further propose a chordal distance-based scheduling algorithm that exploits only the AoA knowledge and analyze the performance by deriving ergodic rates in closed form. Based on the derived rates, I show that the beamspace channel leakage resulting from phase offsets between AoAs and quantized angles of analog combiners can lead to sum rate gain by reducing quantization error compared to the channel without leakage. Simulation results validate the sum rate performance of the proposed algorithms and derived ergodic rate expressions.

5.1 Introduction

Unlike the previous chapters, I focus on developing new user scheduling criteria for hybrid beamforming with low-resolution ADC systems to reduce quantization error without changing the receiver architecture. In recent years, low-resolution ADC systems with hybrid analog/digital beamforming have been investigated to take advantage of both the reduced number of ADC bits and radio frequency (RF) chains [32,90,133,134]. It was shown in [32] that the hybrid beamforming systems with low-resolution ADCs achieve comparable rate to that of infinite-bit ADC systems, providing better energy-rate tradeoff compared to conventional hybrid multiple-input multiple-output (MIMO) systems and low-resolution ADC systems. To further increase spectral and energy efficiency of mmWave receivers, deploying adaptive-resolution ADCs in hybrid MIMO systems was proposed with ADC bit-allocation algorithms

[90, 133]. Channel estimation techniques were also investigated for hybrid MIMO systems with low-resolution ADCs [134]. Understanding the superior spectral and energy efficiency of the architecture, this chapter focuses on the hybrid MIMO receiver with low-resolution ADCs to solve a user scheduling problem in mmWave communications.

Although user scheduling in multiuser MIMO systems has been extensively studied for more than a decade, it has not been investigated for low-resolution ADC systems. One representative method of user scheduling is the semi-orthogonal user selection (SUS) method [44]. This method selects users in a greedy manner such that the channel vectors of the selected users are nearly orthogonal and have large magnitudes based on the full channel state information (CSI) knowledge of all users at the base station (BS). Another representative approach is the random beamforming (RBF) method [45] that selects the user who has the maximum signal-to-interference-noise ratio (SINR) for each beam when a set of orthogonal beams are determined a priori at the BS before scheduling. Similarly, to capture the orthogonality between channels of scheduled users, user scheduling algorithms that adopt chordal distance as a selection measure were proposed in [135, 136].

Unlike the user scheduling methods that have been studied under the Rayleigh fading channel model by assuming rich scattering [44–46], different approaches have investigated user scheduling under the channels with poor scattering such as mmWave channels [47–49]. In [47], user scheduling algorithms were proposed for mmWave communications by leveraging the knowl-

edge of channel gain and angle of departure. In addition, the achievable sum rate was quantified for the BS which employs an iterative matrix decomposition based hybrid beamforming scheme proposed in [137]. The RBF method was analyzed in both the uniform random single path [48] and multi-path channel models [49]. By exploiting the sparse nature of mmWave channels, beam aggregation-based scheduling and fairness-aware scheduling algorithms were developed in [49]. Although the user scheduling algorithms were proposed for mmWave communications, they still focused on user scheduling without quantization error. Consequently, user scheduling in mmWave systems with low-resolution ADCs remains questionable.

5.1.1 Contributions

In this chapter, I investigate uplink user scheduling for mmWave hybrid MIMO zero-forcing receivers with low-resolution ADCs. Noting that non-negligible quantization error can be a primary bottleneck for attaining scheduling gain in the low-resolution ADC system, I provide following contributions:

- User scheduling criteria is derived to maximize the scheduling gain by finding the best tradeoff between channel gains and corresponding quantization noise. Adopting the virtual channel model [71], the criteria can be interpreted as follows: for a given channel gain, (i) unique AoAs of each scheduled user and (ii) equal power spread across the beamspace complex gains within each user maximize sum rate. Accordingly, the derived scheduling criteria reveal that the channel structure in the beamspace, in addition to the chan-

nel magnitude and orthogonality, plays a key role in maximizing sum rate under coarse quantization.

- Leveraging the derived criteria, an efficient scheduling algorithm is proposed for hybrid low-resolution ADC systems. The proposed algorithm combines semi-orthogonal user filtering [44] and non-overlap filtering of dominant beams [49] to enforce orthogonality among scheduled users and to reduce quantization error. Using an approximated SINR as a scheduling measure, the algorithm captures the trade-off between channel gain and corresponding quantization error, and reduces computational complexity by avoiding matrix inversion.

- Considering the difficulty of acquiring instantaneous full CSI, I further propose a chordal distance-based scheduling algorithm which only requires AoAs of mmWave channels, known as slowly-varying channel characteristics [138]. Unlike the previously developed chordal distance-based algorithms [135, 136] that use full CSI and adopt a simple greedy structure which requires prohibitively high complexity, the proposed algorithm exploits only the AoA information of mmWave channels and reduces the complexity by filtering a user candidate set.

- To analyze the performance of the chordal distance-based algorithm, closed-form sum rates are derived for two channel scenarios: (1) AoAs exactly align with quantized angles of analog combiners and (2) arbitrary AoAs produce phase offsets from the quantized angles, which results in channel leakage. For

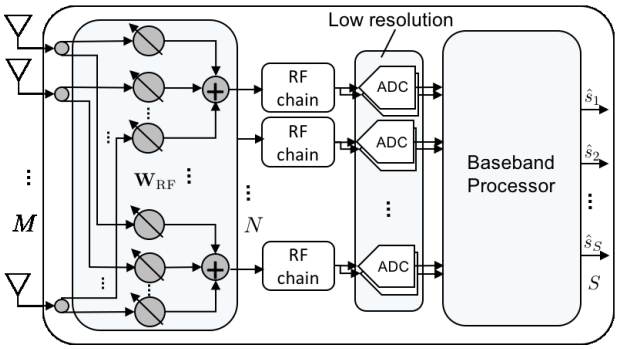


Figure 5.1: A receiver architecture with large antenna arrays and analog combiners $\mathbf{W}_{\mathrm{RF}}$, followed by low-resolution ADCs.

the first scenario, an ergodic rate is derived as the sum of the ergodic rate with no quantization and the rate loss due to quantization. Accordingly, the derived rate provides the expected ergodic rate loss due to quantization in closed form. For the second scenario, an approximated lower bound of the ergodic rate is derived in closed form. It is observed that the two channel scenarios result in different sum rates as a consequence of coarse quantization, and the channel leakage provides sum rate gain by reducing quantization error, which challenges the conventional negative understanding towards channel leakage.

Simulation results demonstrate the superior ergodic sum rate performance of the proposed algorithms and validate the analysis and intuition obtained in this chapter.

5.2 System Model

5.2.1 Signal and Channel Models

We consider a single-cell multiuser MIMO network for uplink communications. A BS employs a uniform linear array (ULA) of M receive antennas. Analog combiners are applied at the BS, followed by $N \leq M$ chains as shown in Fig. 5.1. We assume that K single-antenna users are distributed in the cell and the BS schedules $S \leq N$ users to serve among the K users in the cell. The ADCs are considered to be low-resolution ADCs to reduce the receiver power consumption.

Focusing on mmWave communications, the channel $\mathbf{h}_k$ for user k is assumed to be a sum of the contributions of limited scatterers that contribute L_k propagation paths to the channel $\mathbf{h}_k$ [139]. Therefore, the discrete-time narrowband channel of user k can be modeled as [71]

$$\mathbf{h}_{\gamma,k} = \sqrt{\frac{1}{\gamma_k}}\mathbf{h}_k = \sqrt{\frac{M}{\gamma_k L_k}} \sum_{\ell=1}^{L_k} g_{k,\ell}\mathbf{a}(\phi_{k,\ell}) \tag{5.1}$$

where γ_k denotes the pathloss of user k, $g_{k,\ell}$ is the complex gain of the ℓth propagation path of user k, and $\mathbf{a}(\phi_{k,\ell})$ is the array steering vector of the BS receive antennas corresponding to the azimuth AoA of the ℓth path of the kth user $\phi_{k,\ell} \in [-\pi/2, \pi/2]$. It is considered that $g_{k,\ell}$ is an independent and identically distributed (IID) complex Gaussian random variable as $g_{k,\ell} \overset{i.i.d}{\sim} \mathcal{CN}(0,1)$. The array steering vector $\mathbf{a}(\theta)$ for the ULA antennas of the BS is given as

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}}\left[1, e^{-j\pi\vartheta}, e^{-j2\pi\vartheta}, \ldots, e^{-j(M-1)\pi\vartheta}\right]^{\mathsf{T}} \tag{5.2}$$

where $\vartheta = \frac{2d}{\lambda}\sin(\theta)$ is the spatial angle that is related to the physical AoA θ, d denotes the distance between antenna elements, and λ represents the signal wave length. Throughout this chapter, θ and ϕ are used to denote the physical angles of analog combiners and physical AoAs of a user channel, respectively. I also use ϑ and φ to indicate the spatial angles for θ and ϕ, respectively. It is assumed that ϑ is a constant value in the range of $[-1, 1]$ and φ is a uniform random variable $\varphi \sim \text{Unif}[-1, 1]$.

For simplicity, I consider a homogeneous long-term received SNR network[2] where a conventional uplink power control compensates for the pathloss and shadowing effect to achieve the same long-term received SNR target for all users in the cell [93, 94]. Let $\mathbf{x} = \mathbf{Ps}$ be the transmitted user signals where $\mathbf{P} = \text{diag}\{\sqrt{\rho\,\gamma_1}, \ldots, \sqrt{\rho\,\gamma_S}\}$ is the transmit power matrix and $\mathbf{s}$ is the $S \times 1$ transmitted symbol vector from S users. Let $\mathbf{H}_\gamma = \mathbf{HB}$ represent the $M \times S$ channel matrix where $\mathbf{H}_\gamma = [\mathbf{h}_{\gamma,1}, \ldots, \mathbf{h}_{\gamma,S}]$ is the channel matrix, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_S]$ is the channel matrix after the uplink power control, and $\mathbf{B} = \text{diag}\{\sqrt{1/\gamma_1}, \ldots, \sqrt{1/\gamma_S}\}$ is the pathloss matrix. Then, the received baseband analog signal $\mathbf{r} \in \mathbb{C}^M$ is given as

$$\mathbf{r} = \mathbf{H}_\gamma\mathbf{x} + \mathbf{n} = \mathbf{HBPs} + \mathbf{n} = \sqrt{\rho}\mathbf{Hs} + \mathbf{n} \tag{5.3}$$

where I assume $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_S)$, and $\mathbf{n}$ indicates the additive white Gaussian noise (AWGN) vector $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. Thus, ρ can be regarded as the SNR.

[2]The proposed scheduling criteria and the proposed algorithms in this chapter can also be applicable to a heterogeneous long-term received SNR network.

The received analog signals in (5.3) are combined via an $M \times N$ analog combiner $\mathbf{W}_{\mathrm{RF}}$. The combiner $\mathbf{W}_{\mathrm{RF}}$ is implemented using analog phase shifters, and its elements are constrained to have the equal norm of $1/\sqrt{M}$. After analog combining, (5.3) becomes

$$\mathbf{y} = \mathbf{W}_{\mathrm{RF}}^H \mathbf{r} = \sqrt{\rho}\mathbf{W}_{\mathrm{RF}}^H \mathbf{H}\mathbf{s} + \mathbf{W}_{\mathrm{RF}}^H \mathbf{n}. \tag{5.4}$$

Assuming uniformly-spaced spatial angles, the matrix of array steering vectors $\mathbf{A} = \left[\mathbf{a}(\theta_1), \ldots, \mathbf{a}(\theta_M)\right]$ becomes a unitary discrete Fourier transform (DFT) matrix. Noting that the antenna space and beamspace are related through a spatial Fourier transform, a sub-matrix of the DFT matrix is adopted as the analog combiner $\mathbf{W}_{\mathrm{RF}} = \tilde{\mathbf{A}}$ [22, 90] to project the received signals onto the beamspace, where $\tilde{\mathbf{A}}$ consists of N columns of $\mathbf{A}$. Through the projection, the BS can exploit the sparsity of the mmWave channels to capture channel gains with the reduced number of RF chains [21]. Using $\mathbf{W}_{\mathrm{RF}} = \tilde{\mathbf{A}}$, (5.4) is rewritten as

$$\mathbf{y} = \sqrt{\rho}\tilde{\mathbf{A}}^H \mathbf{H}\mathbf{s} + \tilde{\mathbf{A}}^H \mathbf{n} = \sqrt{\rho}\mathbf{H}_{\mathrm{b}}\mathbf{s} + \mathbf{v}. \tag{5.5}$$

I denote $\mathbf{H}_{\mathrm{b}} = \tilde{\mathbf{A}}^H \mathbf{H}$, which is the projection of the channel matrix onto the beamspace. Since $\mathbf{A}$ is a unitary matrix, the projected noise vector $\mathbf{v} = \tilde{\mathbf{A}}^H \mathbf{n}$ is distributed as $\mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$.

5.2.2 Quantization Model

In this subsection, I introduce an additive quantization noise model [73] which approximates quantization process in a linear form for analytical

tractability. Such linear approximation of quantization provides reasonable accuracy in low and medium SNR ranges [56]. After processed through the RF chains, each complex sample y_i in (5.5) is quantized at the ith pair of ADCs, and each ADC quantizes either a real or imaginary component of y_i. The quantized signal $\mathbf{y}_{\mathrm{q}}$ is [73]

$$\mathbf{y}_{\mathrm{q}} = \mathcal{Q}\big(\mathrm{Re}\{\mathbf{y}\}\big) + j\mathcal{Q}\big(\mathrm{Im}\{\mathbf{y}\}\big) = \alpha\sqrt{\rho}\mathbf{H}_{\mathrm{b}}\mathbf{s} + \alpha\mathbf{v} + \mathbf{q} \tag{5.6}$$

where $\mathcal{Q}(\cdot)$ is the element-wise quantizer function. The quantization gain α is defined as $\alpha = 1 - \beta$, $\beta = \mathbb{E}[|y - y_{\mathrm{q}}|^2]/\mathbb{E}[|y|^2]$ is a normalized mean squared quantization error, and $\mathbf{q}$ is the additive quantization noise vector.

For a scalar MMSE quantizer of a Gaussian random variable, β can be approximated as $\beta \approx \frac{\pi\sqrt{3}}{2}2^{-2b}$ for $b > 5$ [125] where b denotes the number of quantization bits for each real and imaginary part of y. The values of β for $b \leq 5$ are shown in Table 1 in [90]. Although the quantization error is neither Gaussian nor is its covariance matrix diagonal in an exact nonlinear quantization model, approximations are provided based on [54, 56, 73] as follows: considering a lower bound of achievable rate, I assume $\mathbf{q} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{\mathbf{qq}}(\mathbf{H}_{\mathrm{b}}))$ [54]. Since $\mathbf{q}$ is uncorrelated with $\mathbf{y}$ [73], the covariance matrix of $\mathbf{q}$ with $\mathbf{H}_{\mathrm{b}}$ is given as [54, 73]

$$\mathbf{R}_{\mathbf{qq}}(\mathbf{H}_{\mathrm{b}}) = \alpha(1-\alpha)\,\mathrm{diag}(\rho\mathbf{H}_{\mathrm{b}}\mathbf{H}_{\mathrm{b}}^H + \mathbf{I}_N). \tag{5.7}$$

In the following section, I investigate a user scheduling problem based on the considered system model.

5.3 User Scheduling

In this section, I focus on ZF combining $\mathbf{W}_{\mathrm{zf}} = \mathbf{H}_{\mathrm{b}}(\mathbf{H}_{\mathrm{b}}^H\mathbf{H}_{\mathrm{b}})^{-1}$ at the BS and investigate user scheduling to derive scheduling criteria and propose an algorithm by exploiting the obtained criteria. To this end, I first consider the case where the effective CSI $\mathbf{H}_{\mathrm{b}}$ is known at the BS and then extend the problem to the case where only the partial CSI is available. For low-resolution ADC systems, state-of-the-art channel estimation techniques have been developed and have shown remarkable estimation accuracy with few-bit ADCs [17,88] or even with one-bit ADCs [14–16]. With the ZF combiner $\mathbf{W}_{\mathrm{zf}}$, the quantized signal in (5.6) is given as

$$\mathbf{y}_{\mathrm{q}}^{\mathrm{zf}} = \mathbf{W}_{\mathrm{zf}}^H\mathbf{y}_{\mathrm{q}} = \alpha\sqrt{\rho}\mathbf{W}_{\mathrm{zf}}^H\mathbf{H}_{\mathrm{b}}\mathbf{s} + \alpha\mathbf{W}_{\mathrm{zf}}^H\mathbf{v} + \mathbf{W}_{\mathrm{zf}}^H\mathbf{q}.$$

Nulling out the inter-user interference, the achievable rate of user k is derived as

$$r_k(\mathbf{H}_{\mathrm{b}}) = \log_2\left(1 + \frac{\alpha^2\rho}{\mathbf{w}_{\mathrm{zf},k}^H\mathbf{R}_{\mathbf{qq}}(\mathbf{H}_{\mathrm{b}})\mathbf{w}_{\mathrm{zf},k} + \alpha^2\|\mathbf{w}_{\mathrm{zf},k}\|^2}\right) \tag{5.8}$$

Using the achievable rate with quantization error (5.8), I formulate a user scheduling problem5:

$$\mathcal{P}1: \quad \mathcal{R}(\mathbf{H}_{\mathrm{b}}(\mathcal{S}^\star)) = \max_{\mathcal{S}\subset\{1,\ldots,K\}:|\mathcal{S}|\leq S}\sum_{k\in\mathcal{S}} r_k(\mathbf{H}_{\mathrm{b}}(\mathcal{S})) \tag{5.9}$$

where $\mathcal{S}$ represents the set of scheduled users, $\mathbf{H}_{\mathrm{b}}(\mathcal{S})$ is the beamspace channel matrix of the users in $\mathcal{S}$, and $\mathcal{R}(\mathbf{H}_{\mathrm{b}}(\mathcal{S}))$ is the sum rate of the scheduled users in $\mathcal{S}$. Unlike the user scheduling without quantization, which considers the

channel orthogonality and the large channel gains, the user scheduling with the coarse quantization needs to consider an additional condition.

Remark 13. *To maximize the achievable rate* (5.8)*, the aggregated beamspace channel gain at each RF chain* $\|[\mathbf{H}_{\mathrm{b}}]_{i,:}\|^2$ *needs to be minimized to reduce the quantization noise variance* $\mathbf{R}_{\mathbf{qq}}$ *in addition to forcing the channel orthogonality* ($\mathbf{h}_{\mathrm{b},k} \perp \mathbf{h}_{\mathrm{b},k'},\ k \neq k'$) *and maximizing the beamspace channel gain* $\|\mathbf{h}_{\mathrm{b},k}\|^2$*, which reduces* $\|\mathbf{w}_{\mathrm{zf},k}\|^2$*.*

5.3.1 Analysis of Scheduling Criteria

The scheduling criteria are derived for channels in the beamspace based on the finding in Remark 13 to propose an efficient scheduling algorithm that solves $\mathcal{P}1$ in (5.9). To focus on key scheduling ingredients besides the channel magnitude, I consider the case where the magnitude of each user channel is given in the analysis, i.e., $\|\mathbf{h}_{\mathrm{b},k}\| = \sqrt{\gamma_k},\ \forall k$ with $\gamma_k > 0$. Given $\|\mathbf{h}_{\mathrm{b},k}\| = \sqrt{\gamma_k},\ \forall k$, I reformulate $\mathcal{P}1$ to the problem of finding the optimal channel matrix that maximizes the uplink sum rate to characterize the channel matrix that fully extracts scheduling gains.

$$\mathcal{P}2: \quad \mathcal{R}(\mathbf{H}_{\mathrm{b}}^{\star}) = \max_{\mathbf{H}_{\mathrm{b}} \in \mathbb{C}^{N \times S}} \sum_{k=1}^{S} r_k(\mathbf{H}_{\mathrm{b}}), \quad \text{s.t. } \|\mathbf{h}_{\mathrm{b},k}\| = \sqrt{\gamma_k}\ \ \forall k. \tag{5.10}$$

To provide geometrical interpretation for the channel matrix analysis, I further adopt the virtual channel representation [71], where each beamspace channel $\mathbf{h}_{\mathrm{b},k}$ contains $(N - L_k)$ zeros and L_k complex gains of the L_k channel paths. I first consider the single user scheduling ($S = 1$) and derive the channel

characteristics required to maximize the achievable rate for $\mathcal{P}2$. Then, the result is utilized to derive the scheduling criteria for the multiuser scheduling case.

Lemma 6. *For a single user scheduling, scheduling a user who has the following channel characteristics maximizes the uplink achievable rate in $\mathcal{P}2$:*

(i) the largest number of channel propagation paths and

(ii) equal power spread across the beamspace complex gains.

Proof. The ZF combiner for a single user becomes $\mathbf{w}_{\text{zf}} = \mathbf{h}_{\text{b}}/\|\mathbf{h}_{\text{b}}\|^2$. Then, (5.8) is given as

$$\begin{aligned}\mathcal{R}(\mathbf{h}_{\text{b}}) &= \log_2\left(1+\frac{\alpha\rho}{(1-\alpha)\frac{\mathbf{h}_{\text{b}}^H}{\|\mathbf{h}_{\text{b}}\|^2}\text{diag}\left(\rho\mathbf{h}_{\text{b}}\mathbf{h}_{\text{b}}^H+\mathbf{I}_N\right)\frac{\mathbf{h}_{\text{b}}}{\|\mathbf{h}_{\text{b}}\|^2}+\frac{\alpha}{\|\mathbf{h}_{\text{b}}\|^2}}\right)\\ &= \log_2\left(1+\frac{\alpha\rho\|\mathbf{h}_{\text{b}}\|^4}{\rho(1-\alpha)\sum_{i\in\mathcal{L}}|h_{\text{b},i}|^4+\|\mathbf{h}_{\text{b}}\|^2}\right), \end{aligned} \tag{5.11}$$

where $\mathcal{L}$ is the set of indices of non-zero complex gains in $\mathbf{h}_{\text{b}}$ with $|\mathcal{L}| = L$. With the constraint of $\|\mathbf{h}_{\text{b}}\| = \sqrt{\gamma}$, the problem of maximizing $\mathcal{R}(\mathbf{h}_{\text{b}})$ in (5.11) reduces to

$$\min_{\mathbf{h}_{\text{b}}} \sum_{i\in\mathcal{L}} |h_{\text{b},i}|^4 \quad \text{s.t. } \|\mathbf{h}_{\text{b}}\|^2 = \gamma. \tag{5.12}$$

The Karush-Kuhn-Tucker condition is used to solve the reduced problem in (5.12). Let $x_i = |h_{\text{b},i}|^2$ for $i = 1, 2, \ldots, N$. The Lagrangian of the problem with a Lagrangian multiplier μ is given as

$$\mathfrak{L}(\mathbf{x}, \mu) = \|\mathbf{x}\|^2 + \mu\left(\sum_{i\in\mathcal{L}} x_i - \gamma\right).$$

By taking a derivative of $\mathfrak{L}(\mathbf{x}, \mu)$ with respect to x_i for $i \in \mathcal{L}$ and setting it to zero, I obtain $x_i = -\mu/2$. Putting it to $\sum_{i \in \mathcal{L}} x_i = \gamma$, I have $\mu = -2\gamma/L$. Finally, the solution becomes

$$x_i = \gamma/L, \quad i \in \mathcal{L}. \tag{5.13}$$

Under the virtual channel representation, x_i indicates the power of the beamspace complex gains and L is the number of propagation paths. Accordingly, the physical meaning of (5.13) is that the achievable rate for the single user case with the given channel power $\|\mathbf{h}_{\mathrm{b}}\|^2 = \gamma$ can be maximized when the channel power γ is evenly spread to the L beamspace complex gains.

By applying the solution $|h^{\star}_{\mathrm{b},i}|^2 = \gamma/L$ in (5.13) for $i \in \mathcal{L}$, the achievable rate in (5.11) becomes

$$\mathcal{R}(\mathbf{h}^{\star}_{\mathrm{b}}) = \log_2 \left(1 + \frac{\alpha\rho}{\rho(1-\alpha)/L + 1/\gamma} \right). \tag{5.14}$$

The quantization noise variance term in (5.14) decreases as L increases. Therefore, the achievable rate $\mathcal{R}(\mathbf{h}^{\star}_{\mathrm{b}})$ can be further maximized if the scheduled user channel $\mathbf{h}^{\star}_{\mathrm{b}}$ has the largest number of propagation paths with equal power distribution across the beamspace complex gains. ■

Unlike the conventional understanding that scheduling a user with the largest channel gain achieves the maximum achievable rate for the single user communication in the noise-limited environment, Lemma 6 shows that the achievable rate is related not only to the channel magnitude $\|\mathbf{h}_{\mathrm{b}}\|$ but also to the channel structure in the beamspace when received signals are coarsely

quantized. I further show that if the number of propagation paths L is limited, the maximum rate for the single user case converges to a finite value as the channel magnitude increases.

Corollary 12. *With the finite number of channel propagation paths L, the maximum achievable rate with single user scheduling converges to*

$$\mathcal{R}(\mathbf{h}_{\mathrm{b}}^{\star}) \to \log_2\left(1+\alpha L/(1-\alpha)\right), \quad \textit{as } \|\mathbf{h}_{\mathrm{b}}\| \to \infty. \tag{5.15}$$

Proof. The maximum achievable rate of the single user scheduling with the given L and $\|\mathbf{h}_{\mathrm{b}}\|^2 = \gamma$ is derived in (5.14). Then, (5.14) converges to (5.15) as increasing the channel gain ($\gamma \to \infty$). ■

Corollary 12 shows that the quantization error ($\alpha < 1$) limits the achievable rate to remain finite because the quantization noise variance also increases with the increase of the channel magnitude. This implies that the conventional scaling law $\log\log K$ [45] cannot be met in the low-resolution ADCs regime. Accordingly, as the SNR increases, mitigation of the quantization error becomes a more critical problem that needs to be considered in user scheduling.

Now, the multiuser scheduling is investigated for the channel environment where $\sum_{k=1}^{S} L_{\mathcal{S}(k)} \leq N$. Here, $\mathcal{S}(k)$ is the kth scheduled user. This condition is relevant to mmWave channels where the number of channel paths L_k is presumably very small [70]. The problem $\mathcal{P}2$ is solved to characterize the channel properties that maximize the scheduling gain. Theorem 7 shows

the structural scheduling criteria of channels to maximize the sum rate in $\mathcal{P}2$ for the considered case.

Theorem 7. *For $\sum_{k=1}^{S} L_{\mathcal{S}(k)} \leq N$, scheduling a set of users $\mathcal{S}$ that satisfies the following channel characteristics maximizes the uplink sum rate in $\mathcal{P}2$.*

(i) Unique AoAs at the receiver for the channel propagation paths of each scheduled user:

$$\mathcal{L}_{\mathcal{S}(k)} \cap \mathcal{L}_{\mathcal{S}(k')} = \emptyset \text{ if } k \neq k', \tag{5.16}$$

where $\mathcal{L}_{\mathcal{S}(k)}$ represents the set of indices of non-zero complex gains in $\mathbf{h}_{\mathrm{b},\mathcal{S}(k)}$.

(ii) Equal power spread across the beamspace complex gains within each user channel:

$$|h_{\mathrm{b},i,\mathcal{S}(k)}| = \sqrt{\gamma_{\mathcal{S}(k)}/L_{\mathcal{S}(k)}} \text{ for } i \in \mathcal{L}_{\mathcal{S}(k)}. \tag{5.17}$$

Proof. I take a two-stage maximization approach and show the sufficient conditions for maximizing the sum rate in $\mathcal{P}2$ with the constraint of $\sum_{k=1}^{S} L_{\mathcal{S}(k)} \leq N$. Using the diagonal structure of $\mathbf{R_{qq}}$ as shown in (5.7), (5.8) is rewritten in a simpler form as

$$r_k(\mathbf{H}_\mathrm{b}) = \log_2\left(1+\frac{\alpha\rho}{\rho(1-\alpha)\mathbf{w}_{\mathrm{zf},k}^H \mathrm{diag}\left(\mathbf{H}_\mathrm{b}\mathbf{H}_\mathrm{b}^H\right)\mathbf{w}_{\mathrm{zf},k} + \|\mathbf{w}_{\mathrm{zf},k}\|^2}\right). \tag{5.18}$$

In the first stage, I focus on minimizing $\|\mathbf{w}_{\mathrm{zf},k}\|^2$ in (5.18). When user channels are orthogonal, $\mathbf{h}_{\mathrm{b},k} \perp \mathbf{h}_{\mathrm{b},k'}$ for $k \neq k'$, we have $\mathbf{w}_{\mathrm{zf},k} = \mathbf{h}_{\mathrm{b},k}/\|\mathbf{h}_{\mathrm{b},k}\|^2$. Since

$\mathbf{w}_{\mathrm{zf},k}$ with minimum norm is known as $\mathbf{w}_{\mathrm{zf},k} = \mathbf{h}_{\mathrm{b},k}/\|\mathbf{h}_{\mathrm{b},k}\|^2$, $\|\mathbf{w}_{\mathrm{zf},k}\|^2$ can be minimized with the orthogonality condition.

In the second stage, I minimize the achievable rate of (5.18) by imposing the orthogonality condition from the first stage as follows:

$$r_k(\mathbf{H}_{\mathrm{b}}|\mathbf{h}_{\mathrm{b},k} \perp \mathbf{h}_{\mathrm{b},k'}) \overset{(a)}{=} \log_2\left(1+\frac{\alpha\rho\|\mathbf{h}_{\mathrm{b},k}\|^4}{\rho(1-\alpha)\mathbf{h}_{\mathrm{b},k}^H \mathrm{diag}\big(\mathbf{H}_{\mathrm{b}}\mathbf{H}_{\mathrm{b}}^H\big)\mathbf{h}_{\mathrm{b},k}+\|\mathbf{h}_{\mathrm{b},k}\|^2}\right) \tag{5.19}$$

$$= \log_2\left(1+\frac{\alpha\rho\gamma_k^2}{\rho(1-\alpha)\sum\limits_{i\in\mathcal{L}_k}|h_{\mathrm{b},i,k}|^2\left(|h_{\mathrm{b},i,k}|^2+\sum\limits_{u\neq k}|h_{\mathrm{b},i,u}|^2\right)+\gamma_k}\right)$$

$$\overset{(b)}{\leq} \log_2\left(1+\frac{\alpha\rho\gamma_k^2}{\rho(1-\alpha)\sum_{i\in\mathcal{L}_k}|h_{\mathrm{b},i,k}|^4+\gamma_k}\right) \tag{5.20}$$

$$\overset{(c)}{\leq} \log_2\left(1+\frac{\alpha\rho}{\rho(1-\alpha)/L_k+1/\gamma_k}\right). \tag{5.21}$$

The equality (a) is from $\mathbf{w}_{\mathrm{zf},k} = \mathbf{h}_{\mathrm{b},k}/\|\mathbf{h}_{\mathrm{b},k}\|^2$. The equality in (b) holds if and only if $|h_{\mathrm{b},i,u}| = 0$, $\forall\, u \neq k$ and $i \in \mathcal{L}_k$. This implies that each user needs to have channel paths with unique AoAs to maximize the achievable rate. Note that (5.20) is equivalent to the achievable rate of the single user scheduling in (5.11) due to the channel orthogonality and unique AoA conditions. Consequently, applying Lemma 6, I have the inequality (c) which comes from the fact that (5.20) is maximized when $|h_{\mathrm{b},i,k}| = \sqrt{\gamma_k/L_k}$ for $i \in \mathcal{L}_k$, i.e., channel power is spread evenly across the beamspace complex gains within each user channel. The upper bound in (5.21) is equivalent to the maximum achievable rate for the single user case in (5.14). Therefore, (5.21) is also the maximum achievable rate of each user for the problem $\mathcal{P}2$, which also maximizes the sum

rate in $\mathcal{P}2$.

Throughout the proof, it is shown that the derived conditions—the orthogonality, the unique AoA, and the equal power spread conditions—are sufficient to maximize the sum rate in $\mathcal{P}2$ for the case of $\sum_{k=1}^{S} L_{\mathcal{S}(k)} \leq N$. Since, the unique AoA condition implies the orthogonality, only the unique AoA and equal power spread conditions are required to be satisfied by the beamspace channel matrix $\mathbf{H}_{\mathrm{b}}$ for maximizing the uplink sum rate. This completes the proof. ■

Distinguished from conventional channels, there are channel orthogonality cases related to mmWave massive MIMO communications: (a) asymptotic orthogonality of array steering vectors across different angles [98], (b) orthogonality of beamspace channel sub-vectors having common AoAs, and (c) orthogonality of array steering vectors in (5.2) with angle offsets of multiples of $2/M$ [49]. Note that the first condition in (5.16) particularly emphasizes the third case (c) which forces the beamspace channel orthogonality and further minimizes the aggregated channel gain at each RF chain by avoiding overlap between channel gains in the same AoA, which reduces the quantization noise variance as discussed in Remark 13. The second condition in (5.17) also minimizes the aggregated channel gain by evenly spreading the channel power across the beamspace gains, and thus, reduces the quantization error. Consequently, Theorem 7 emphasizes the importance of the channel structure in maximizing the sum rate under coarse quantization, while conventional user scheduling approaches ignore such criteria.

Algorithm 5: Channel Structure-based Scheduling (CSS)

Initialization: $\mathcal{K}_1 = \{1, \ldots, K\}$, $\mathcal{S} = \phi$, and $i = 1$.
for *k = 1:K* **do**
  BS stores $N_b \geq L_k$ indices of dominant spatial angles of $\mathbf{h}_{\mathrm{b},k}$ in $\mathcal{B}_k$.
Iteration: **while** $i \leq S$ *and* $\mathcal{K}_i \neq \emptyset$ **do**
  for $k \in \mathcal{K}_i$ **do**
    BS computes approximated SINR of user k, $\mathrm{SINR}_k\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})\big)$ in (5.25).
  BS schedules user who has the largest SINR as

$$\mathcal{S}(i) = \underset{k \in \mathcal{K}_i}{\operatorname{argmax}} \, \mathrm{SINR}_k\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})\big) \tag{5.22}$$

  and updates scheduled user set $\mathcal{S} = \mathcal{S} \cup \{\mathcal{S}(i)\}$.
  Then, BS computes orthogonal component $\mathbf{f}_{\mathcal{S}(i)}$ for filtering as in (5.23).
  Using $\mathbf{f}_{\mathcal{S}(i)}$ and $\mathcal{B}_{\mathcal{S}(i)}$, BS filters candidate set $\mathcal{K}_i$ as in (5.24) and sets $i = i + 1$;
return Scheduled user set $\mathcal{S}$;

Therefore, I propose a quantization-aware scheduling algorithm based on the criteria in Theorem 7. Although the scheduling criteria in Theorem 7 is derived under the condition of $\sum_{k=1}^{S} L_{\mathcal{S}(k)} \leq N$, I show that the proposed algorithm which exploits the criteria still achieves higher performance compared to conventional algorithms for $\sum_{k=1}^{S} L_{\mathcal{S}(k)} > N$ in Section 5.5.

5.3.2 Proposed Algorithm

In this subsection, a user scheduling algorithm with low complexity is proposed by using the criteria in Theorem 7. Adopting a greedy manner, the proposed algorithms make it possible to schedule users without examining all

combinations of users. At each iteration, the proposed algorithm schedules a user and reduces the size of a user candidate set $\mathcal{K}$ through filtering. To extract user diversity, the algorithm filter the user set $\mathcal{K}$ by enforcing semi-orthogonality between scheduled user channels, not perfect orthogonality. In addition to the scheduling criteria in Theorem 7, the orthogonality condition in (5.19) is also applied for the filtering to provide higher precision in the semi-orthogonality.

Algorithm 5 describes the proposed scheduling method, called channel structure-based scheduling (CSS). After each user selection, the proposed algorithm filters the user candidate set $\mathcal{K}$ by leveraging the orthogonality condition in (5.19) as in [44] by utilizing (5.23)

$$\begin{aligned}\mathbf{f}_{\mathcal{S}(i)} &= \mathbf{h}_{\mathrm{b},\mathcal{S}(i)} - \sum_{j=1}^{i-1} \frac{\mathbf{f}_{\mathcal{S}(j)}^H \mathbf{h}_{\mathrm{b},\mathcal{S}(i)}}{\|\mathbf{f}_{\mathcal{S}(j)}\|^2} \mathbf{f}_{\mathcal{S}(j)} \\ &= \left(\mathbf{I} - \sum_{j=1}^{i-1} \frac{\mathbf{f}_{\mathcal{S}(j)} \mathbf{f}_{\mathcal{S}(j)}^H}{\|\mathbf{f}_{\mathcal{S}(j)}\|^2}\right) \mathbf{h}_{\mathrm{b},\mathcal{S}(i)} \end{aligned} \tag{5.23}$$

where $\mathbf{f}_{\mathcal{S}(i)}$ is the component of $\mathbf{h}_{\mathrm{b},\mathcal{S}(i)}$ that is orthogonal to the subspace $\mathrm{span}\{\mathbf{f}_{\mathcal{S}(1)}, \ldots, \mathbf{f}_{\mathcal{S}(i-1)}\}$. Unlike the algorithm in [44] which computes the orthogonal component $\mathbf{f}_k$ for the entire users in the candidate set, the proposed CSS algorithm calculates $\mathbf{f}_{\mathcal{S}(i)}$ only for the currently scheduled user $\mathcal{S}(i)$. The algorithm also enforces additional spatial orthogonality in the beamspace to the filtered set as in [49] by modifying the unique AoA condition in (5.16). Since there can exist phase offsets that lead to more than L_k dominant channel gains in $\mathbf{h}_{\mathrm{b},k}$ due to the quantized angles of the analog combiner, the

algorithm stores $N_b \geq L_k$ indices of dominant spatial angles in $\mathcal{B}_k$ and filters the user set $\mathcal{K}$ by removing users whose angle indices in $\mathcal{B}_k$ show more than N_{OL} overlaps with those of the scheduled user in $\mathcal{B}_{\mathcal{S}(i)}$. The semi-orthogonality filtering becomes

$$\mathcal{K}_{i+1} = \left\{ k \in \mathcal{K}_i \backslash \{\mathcal{S}(i)\} \middle| \frac{|\mathbf{f}_{\mathcal{S}(i)}^H \mathbf{h}_{\mathrm{b},k}|}{\|\mathbf{f}_{\mathcal{S}(i)}\| \|\mathbf{h}_{\mathrm{b},k}\|} < \epsilon, |\mathcal{B}_{\mathcal{S}(i)} \cap \mathcal{B}_k| \leq N_{\mathrm{OL}} \right\}. \tag{5.24}$$

These filtering operations not only reduce the size of the user set $\mathcal{K}$, but also offer semi-orthogonality between the scheduled users in $\mathcal{S}$ and the candidate users in $\mathcal{K}$. As a result, the filtering leads the ZF combiner to be approximated as $\mathbf{w}_{\mathrm{zf},k} \approx \mathbf{h}_{\mathrm{b},k}/\|\mathbf{h}_{\mathrm{b},k}\|^2$ for a user $k \in \mathcal{K}$, and the SINR of user $k \in \mathcal{K}$ with previously scheduled users in $\mathcal{S}$ is approximated as

$$\mathrm{SINR}_k\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})\big) \approx \frac{\alpha\rho\|\mathbf{h}_{\mathrm{b},k}\|^4}{(1-\alpha)\mathbf{h}_{\mathrm{b},k}^H \mathbf{D}\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})\big)\mathbf{h}_{\mathrm{b},k}} \tag{5.25}$$

where $\mathbf{D}\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})\big) = \mathrm{diag}\big(\rho\,\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})\mathbf{H}_{\mathrm{b}}(\mathcal{S} \cup \{k\})^H + \frac{1}{1-\alpha}\mathbf{I}_N\big)$. For a scheduling measure, the proposed algorithm adopts the approximated SINR (5.25) to incorporate the scheduling criteria in Theorem 7 with the channel magnitude and orthogonality[3]. At each iteration, the algorithm schedules the user who has the largest SINR among the users in $\mathcal{K}$ as shown in (5.22). Using the approximated SINR (5.25) for the selection measure greatly reduces the computational complexity by avoiding the matrix inversion for computing the ZF combiner $\mathbf{W}_{\mathrm{zf}}$.

[3]By treating the approximate SINR as the true SINR and following the technique used in [44] and [49], the proposed method can be incorporated with the proportional fairness (PF) policy [140] for fairness-aware scheduling in a heterogeneous system.

Algorithm 6: Greedy Max-Sum Rate Scheduling

Initialization: $\mathcal{K}_{\mathrm{G},1} = \{1, \ldots, K\}$, $\mathcal{S}_{\mathrm{G}} = \emptyset$, and $i = 1$.
Iteration: **while** $i \leq S_{\mathrm{G}}$ **do**
  for $k \in \mathcal{K}_{\mathrm{G},i}$ **do**
    Compute sum rate using r_j in (5.8) for scheduled users and each user $k \in \mathcal{K}_{\mathrm{G},i}$ as

$$\mathcal{R}_k = \sum_{j \in \mathcal{S}_{\mathrm{G}} \cup \{k\}} r_j\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S}_{\mathrm{G}} \cup \{k\})\big) \tag{5.26}$$

  BS schedules user who maximizes sum rate as $\mathcal{S}_{\mathrm{G}}(i) = \mathrm{argmax}_{k \in \mathcal{K}_{\mathrm{G},i}} \mathcal{R}_k$ and
  updates $\mathcal{K}_{\mathrm{G},i+1} = \mathcal{K}_{\mathrm{G},i} \setminus \{\mathcal{S}_{\mathrm{G}}(i)\}$, $\mathcal{S}_{\mathrm{G}} = \mathcal{S}_{\mathrm{G}} \cup \{\mathcal{S}_{\mathrm{G}}(i)\}$, and $i = i + 1$;
return Scheduled user set $\mathcal{S}_{\mathrm{G}}$;

To provide a reference in sum rate performance, I also propose a high-complexity and high-performance greedy algorithm which schedules the user who achieves the highest sum rate at each iteration as shown in Algorithm 6. At each iteration, the greedy algorithm computes sum rate in (5.8), i.e., the algorithm computes the exact SINR for scheduled users in $\mathcal{S}_{\mathrm{G}}$ and a candidate user k, $\forall k \in \mathcal{K}_{\mathrm{G},i}$. Thus, the algorithm carries the huge burden of computing a matrix inversion $|\mathcal{K}_{\mathrm{G},i}|$ times at each selection. At the ith stage, the greedy algorithm computes the achievable rate in (5.8) $|\mathcal{K}_{\mathrm{G},i}| \times i$ times and compares the derived $|\mathcal{K}_{\mathrm{G},i}|$ sum rates, whereas the CSS algorithm only computes the approximated SINR in (5.25) $|\mathcal{K}_i|$ times and compares $|\mathcal{K}_i|$ SINRs. Moreover, unlike the greedy algorithm, the CSS algorithm reduces the size of the user set $\mathcal{K}_i$ by filtering in (5.24) at each iteration. This leads to $|\mathcal{K}_i| \ll |\mathcal{K}_{\mathrm{G},i}|$, and

the gap $|\mathcal{K}_{\mathrm{G},i}| - |\mathcal{K}_i|$ will increase with iteration; the CSS algorithm becomes more efficient with larger K and /or S.

Remark 14. *The proposed algorithm can be applied to an orthogonal frequency division multiplexing (OFDM) system for a wideband channel case. Since the system with a given analog combiner is considered, the proposed algorithm can be performed independently for each subcarrier index i. However, the structure of the quantization noise $\mathbf{q}[i]$ in the wideband OFDM system becomes different from that of the narrowband system so that the spatial filtering in the proposed user scheduling algorithm may not be desirable. Nonetheless, the approximated SINR can still be applicable with the semi-orthogonality filtering by computing the quantization noise variance for each subcarrier i of the OFDM system $\mathbf{R}_{\mathbf{qq}}[i]$. Thus, the BS can perform the proposed algorithms to schedule users to be served on each subcarrier by relaxing the spatial filtering.*

The proposed method schedules users with minimum overlap among quantized AoAs of user channels to satisfy the derived scheduling criterion (i) in Theorem 7. Accordingly, by using the proposed scheduling method, the beamforming-based Doppler effect reduction techniques such as a per-beam synchronization approach in [141] can be performed at the BS since the BS can see each beam with a single dedicated user signal with large channel gains and possibly with other user signals with negligible channel gains. Therefore, the proposed scheduling method can provide potential benefit in reducing Doppler effect when jointly used with Doppler effect mitigation techniques at the BS.

5.3.3 Beam Training-Based Channel Acquisition

Assuming time-division duplex communications, I briefly provide an example of extension of the proposed algorithm to a practical system where the BS uses beam training and receives channel quality indicators (CQIs) from users. A procedure of beam training and CQI feedback can be as follows:

1. The BS constructs a set of $N_s \geq N$ beam vectors $\{\mathbf{a}(\bar{\vartheta}_1), \ldots, \mathbf{a}(\bar{\vartheta}_{N_s})\}$ with the angles within the angles of the analog combiner $\tilde{\mathbf{A}}$, i.e., there exists i such that $\bar{\vartheta}_n \in [\vartheta_i - 1/M, \vartheta_i + 1/M]$, $\forall n$, where ϑ_i is the spatial angle of the ith analog beamformer. Then, the BS transmits each beam of the set in time to all users in the cell during a training phase.

2. Each user k can estimate the channel gain corresponding to each beam and have the estimate of $\mathbf{h}_k^H \tilde{\mathbf{A}} = \mathbf{h}_{\mathrm{b},k}^H$ at the end of the beam training. From the sparsity of the mmWave channel, few elements of $\mathbf{h}_{\mathrm{b},k}$ have non-negligible beam gains and we can implement an efficient feedback method that exploits the sparsity of the effective channel $\mathbf{h}_{\mathrm{b},k}$ as described in [142]. For instance, each user can feed back the beam indices of the non-negligible beam gains and their corresponding channel coefficients in a quantized form to the BS.

3. After the feedback from all users is over, the BS can create an estimate of $\mathbf{H}_b$ with the feedback information by simply padding zeros in the unreported beam indices. Then, the BS can directly apply the proposed scheduling algorithm by using the estimated channel.

5.4 User Scheduling with Partial Channel Information

In this section, a user scheduling algorithm is proposed when only partial CSI is known at the BS since it can be challenging to obtain reliable instantaneous CSI estimates for entire users as the number of antennas or users becomes large. A reasonable alternative is to use slowly-varying channel characteristics, in particular, AoAs of mmWave channels [138]; AoAs persist over longer than the coherence time of mmWave channels, and mmWave channels have a limited number of AoAs. In this regard, by using the AoA knowledge, the proposed algorithm can greatly reduce the burden of estimating instantaneous full CSI at each channel coherence time. After scheduling, it is assumed that the BS acquires the effective CSI of the scheduled users for decoding.

5.4.1 Proposed Algorithm

According to (5.1), the channel $\mathbf{h}_k$ lies in the subspace spanned by its array response vectors, i.e., $\mathbf{h}_k \in \text{span}\{\mathbf{a}(\phi_{k,1}), \ldots, \mathbf{a}(\phi_{k,L_k})\}$. To measure the separation between the subspaces, I adopt chordal distance which measures the angle between the subspaces. In the initialization phase, the algorithm removes users whose AoAs are not in the range of angles of RF chains (reduced range of angles)[4] from the initial candidate user set $\mathcal{K}_{\text{cd},1}$. In the scheduling phase, a first user is scheduled by randomly selecting a user among the set of users with the most AoAs in the reduced range of angles. To schedule a next

[4]The range of angles of RF chains indicates the set of angles corresponding to $\bigcup_i \{\vartheta : |\vartheta - \vartheta_i| < \frac{1}{M}\}$, i.e., the AoAs in the reduced range of angles are $\varphi_{k,\ell} \in \bigcup_i \{\vartheta : |\vartheta - \vartheta_i| < \frac{1}{M}\}$.

user, the algorithm updates the candidate user set $\mathcal{K}_{\mathrm{cd},i}$ by filtering users whose chordal distance is shorter than the threshold d_{th} to impose semi-orthogonality among scheduled users. Due to the filtering, the remaining users in $\mathcal{K}_{\mathrm{cd},i+1}$ are guaranteed to have a certain level of orthogonality with the scheduled users $\mathcal{S}(j)$ for $j = 1, 2, \ldots, i-1$. Then, the algorithm schedules the user with the longest chordal distance among the remaining users with the most AoAs in the reduced range of angles.

To this end, I generate the matrix of array response vectors for each user by exploiting the AoA knowledge as $\mathbf{A}_k = [\mathbf{a}(\phi_{k,\mathcal{V}_k(1)}), \ldots, \mathbf{a}(\phi_{k,\mathcal{V}_k(V_k)})]$ where $\mathcal{V}_k$ is the set of AoAs indices within the reduced range of angles for user k and $V_k = |\mathcal{V}_k|$. Let $\mathcal{A}_k = \mathrm{span}\{\mathbf{A}_k\}$ is the subspace for user k. The chordal distance between the two subspaces $(\mathcal{A}_k, \mathcal{A}_{k'})$ is defined as $d_{\mathrm{cd}}(k, k') = \sqrt{\sum_{\ell=1}^{L_{min}} \sin^2 \theta_\ell}$ where $L_{min} = \min\{L_k, L_{k'}\}$ and $\theta_\ell \leq \pi/2$ is the principal angle between $\mathcal{A}_k$ and $\mathcal{A}_{k'}$. Let $\mathbf{Q}_k$ be the unitary matrix whose columns are orthonormal basis vectors of $\mathcal{A}_k$. According to [143], $d_{\mathrm{cd}}(k, k')$ is rewritten as $d_{\mathrm{cd}}(k, k') = \sqrt{L_{min} - \mathrm{tr}\left(\mathbf{Q}_k^H \mathbf{Q}_{k'} \mathbf{Q}_{k'}^H \mathbf{Q}_k\right)}$. The proposed chordal distance-based user scheduling method is described in Algorithm 7.

Let $\tilde{\mathbf{h}}_k = \sqrt{\frac{M}{L_k}} \sum_{i \in \mathcal{V}_k} g_{k,i} \mathbf{a}(\phi_{k,i})$. Then, the algorithm provides an opportunity to schedule users with nearly $\tilde{\mathbf{h}}_k \perp \tilde{\mathbf{h}}_{k'}$ while the effective channel that the BS sees is the beamspace channel $\mathbf{h}_{\mathrm{b},k} = \mathbf{W}_{\mathrm{RF}}^H \mathbf{h}_k$. Since the AoAs $\phi_{k,i}$, $i \in \mathcal{V}_k$ are in the range of angles of RF chains, $\tilde{\mathbf{h}}_k$ can be regarded to be

Algorithm 7: Chordal Distance-based User Scheduling

Initialization: $\mathcal{K}_{\text{cd},1} = \{1, \ldots, K\}$, $\mathcal{S}_{\text{cd}} = \phi$, and $i = 1$
for *k = 1:K* **do**
  Let $\mathcal{V}_k$ be set of AoA indices in range of angles of steering vectors for user k. If $\mathcal{V}_k = \emptyset$, do $\mathcal{K}_{\text{cd},1} = \mathcal{K}_{\text{cd},1} \setminus \{k\}$, otherwise, set $\mathbf{A}_k = [\mathbf{a}(\phi_{k,\mathcal{V}_k(1)}), \ldots, \mathbf{a}(\phi_{k,\mathcal{V}_k(V_k)})]$. Generate unitary matrix $\mathbf{Q}_k$ = column basis of $\mathbf{A}_k$.
Iteration: **while** $i \leq S_{\text{cd}}$ *and* $\mathcal{K}_{\text{cd},i} \neq \emptyset$ **do**
  if *i = 1* **then**
    Randomly schedule first user $\mathcal{S}_{\text{cd}}(1) \in \mathcal{K}_{\text{cd},1}$ among users with largest $|\mathcal{V}_k|$. Update candidate user set $\mathcal{K}_{\text{cd},2} = \mathcal{K}_{\text{cd},1} \setminus \mathcal{S}_{\text{cd}}(1)$ and $\mathcal{S}_{\text{cd}} = \mathcal{S}_{\text{cd}} \cup \{\mathcal{S}_{\text{cd}}(1)\}$.
  else
    for $k \in \mathcal{K}_{\text{cd},i}$ **do**
      Let $L_{min} = \min\{L_{\mathcal{S}_{\text{cd}}(i-1)}, L_k\}$, and compute

$$d_{\text{cd}}\left(\mathcal{S}_{\text{cd}}(i-1), k\right) = \sqrt{L_{min} - \text{tr}\left(\mathbf{Q}^H_{\mathcal{S}_{\text{cd}}(i-1)} \mathbf{Q}_k \mathbf{Q}_k^H \mathbf{Q}_{\mathcal{S}_{\text{cd}}(i-1)}\right)}. \tag{5.27}$$

    Filter candidate user set based on (5.27)

$$\mathcal{K}_{\text{cd},i+1} = \left\{k \in \mathcal{K}_{\text{cd},i} \middle| d_{\text{cd}}\left(\mathcal{S}_{\text{cd}}(i-1), k\right)/\sqrt{L_{min}} > d_{\text{th}}\right\}. \tag{5.28}$$

    Let $\mathcal{U}$ be set of users with largest $|\mathcal{V}_k|$, $\forall k \in \mathcal{K}_{\text{cd},i+1}$. Schedule user in $\mathcal{U}$ as

$$\mathcal{S}_{\text{cd}}(i) = \underset{k \in \mathcal{U}}{\text{argmax}}\ d_{\text{cd}}\left(\mathcal{S}_{\text{cd}}(i-1), k\right). \tag{5.29}$$

    Update $\mathcal{K}_{\text{cd},i+1} = \mathcal{K}_{\text{cd},i+1} \setminus \{\mathcal{S}_{\text{cd}}(i)\}$, $\mathcal{S}_{\text{cd}} = \mathcal{S}_{\text{cd}} \cup \{\mathcal{S}_{\text{cd}}(i)\}$.
  Set $i = i + 1$;
return Scheduled user set $\mathcal{S}_{\text{cd}}$;

in the subspace of $\mathbf{W}_{\mathrm{RF}}$, i.e., almost $\tilde{\mathbf{h}}_k \in \mathrm{span}\{\mathbf{W}_{\mathrm{RF}}\}$[5]. Accordingly, using $\mathbf{W}_{\mathrm{RF}}^H \mathbf{W}_{\mathrm{RF}} = \mathbf{I}_N$ which comes from the definition i.e., a sub-matrix of the DFT matrix $\mathbf{W}_{\mathrm{RF}} = \tilde{\mathbf{A}}$, $\tilde{\mathbf{h}}_k$ can be rewritten as

$$\tilde{\mathbf{h}}_k \approx \mathbf{W}_{\mathrm{RF}}(\mathbf{W}_{\mathrm{RF}}^H \mathbf{W}_{\mathrm{RF}})^{-1} \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{h}}_k = \mathbf{W}_{\mathrm{RF}} \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{h}}_k \tag{5.30}$$

In addition, I have $\mathbf{h}_{\mathrm{b},k} = \mathbf{W}_{\mathrm{RF}}^H \mathbf{h}_k \approx \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{h}}_k$ as the impact of $\mathbf{a}(\phi_{k,j})$, $\forall j \notin \mathcal{V}_k$ on the beam domain channel $\mathbf{h}_{b,k}$ is relatively small compared to that of $\mathbf{a}(\phi_{k,i})$, $\forall i \in \mathcal{V}_k$ after analog combining. In this regard, as the algorithm gives $\tilde{\mathbf{h}}_k \perp \tilde{\mathbf{h}}_{k'}$, I can nearly have $\mathbf{h}_{\mathrm{b},k} \perp \mathbf{h}_{\mathrm{b},k'}$ by

$$\begin{aligned}
\epsilon &= \tilde{\mathbf{h}}_k^H \tilde{\mathbf{h}}_{k'} \\
&\overset{(a)}{\approx} \tilde{\mathbf{h}}_k^H \mathbf{W}_{\mathrm{RF}} \mathbf{W}_{\mathrm{RF}}^H \mathbf{W}_{\mathrm{RF}} \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{h}}_{k'} \\
&= \tilde{\mathbf{h}}_k^H \mathbf{W}_{\mathrm{RF}} \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{h}}_{k'} \\
&\overset{(b)}{\approx} \mathbf{h}_{\mathrm{b},k}^H \mathbf{h}_{\mathrm{b},k'}
\end{aligned}$$

where (a) is from (5.30) and (b) is from $\mathbf{h}_{\mathrm{b},k} \approx \mathbf{W}_{\mathrm{RF}}^H \tilde{\mathbf{h}}_k$. Thus, the proposed algorithm guarantees a certain level of orthogonality between the beamspace channels of the scheduled users.

As discussed in Section 5.3.3, the beam indices for non-negligible channel gains can be obtained by using CQI feedback, i.e., AoAs can be estimated for each user. When the capacity of amount of feedback is limited and small, such beam index-only feedback which requires only few integer numbers can be applied to faciliate the proposed chordal distance-based algorithm.

[5]If the AoAs of $\tilde{\mathbf{h}}_k$ exactly align with the quantized angles of the analog combiner, $\tilde{\mathbf{h}}_k$ perfectly lies in the subspace of $\mathbf{W}_{\mathrm{RF}}$.

5.4.2 Ergodic Rate Analysis

Now, the performance of the chordal distance-based algorithm is analyzed in ergodic rate. I focus on the case where each channel has a single propagation path, which corresponds to the sparse nature of mmWave channels [48], and the number of RF chains are equal to the number of antennas $N = M$ in the analysis.

Remark 15. *When there is a single path for each user channel, the filtering in* (5.28) *reduces to* $\mathcal{K}_{\mathrm{cd},i+1} = \left\{k \in \mathcal{K}_{\mathrm{cd},i} \;\middle|\; |\mathbf{a}^H(\phi_{\mathcal{S}(i-1)})\mathbf{a}(\phi_k)| < \epsilon_{\mathrm{th}}\right\}$ *where* $\epsilon_{\mathrm{th}} \ll 1$*, and the scheduling problem in* (5.29) *becomes*

$$\mathcal{S}_{\mathrm{cd}}(i) = \underset{k\in\mathcal{K}_{\mathrm{cd},i+1}}{\operatorname{argmin}} |\mathbf{a}^H(\phi_{\mathcal{S}(i-1)})\mathbf{a}(\phi_k)|.$$

Based on Remark 15, I derive closed-form expressions of the ergodic sum rate for two different cases: (1) AoAs of channels exactly align with the quantized angles of the analog combiner, and (2) channels have arbitrary AoAs regardless of the quantized angles of the analog combiner. For the first case, there is no channel leakage in the beamspace and thus, it is often considered as a more favorable channel condition since it improves communication performance such as channel estimation accuracy [134] and achievable rate [21, 144].

Proposition 5. *When AoAs of channels exactly align with the quantized angles of the analog combiner with a single propagation path, the ergodic sum rate for* $|\mathcal{S}_{\mathrm{cd}}| = S$ *scheduled users with the proposed chordal distance-based*

scheduling algorithm is derived as

$$\bar{\mathcal{R}}_1 = \frac{S}{\ln 2}\left(e^{\frac{1}{\rho M}}\,\Gamma\left(0, \frac{1}{\rho M}\right) - e^{\frac{1}{\rho(1-\alpha)M}}\,\Gamma\left(0, \frac{1}{\rho(1-\alpha)M}\right)\right) \tag{5.31}$$

where $\Gamma(a,z)$ *is an incomplete gamma function defined as* $\Gamma(a,z) = \int_z^\infty t^{a-1}e^{-t}dt$.

Proof. See Section 5.7. ■

Corollary 13. *The derived ergodic rate* (5.31) *can be expressed as the sum of the ergodic rate without quantization error* $\bar{\mathcal{R}}_{\text{inf}}$ *and the ergodic rate loss due to quantization error* $\bar{\mathcal{R}}_{\text{loss}}(\alpha)$

$$\bar{\mathcal{R}}_1 = \bar{\mathcal{R}}_{\text{inf}} + \bar{\mathcal{R}}_{\text{loss}}(\alpha)$$

where $\bar{\mathcal{R}}_{\text{inf}} = \frac{S}{\ln 2}e^{\frac{1}{\rho M}}\Gamma(0, \frac{1}{\rho M})$ *and* $\bar{\mathcal{R}}_{\text{loss}}(\alpha) = -\frac{S}{\ln 2}e^{\frac{1}{\rho(1-\alpha)M}}\Gamma(0, \frac{1}{\rho(1-\alpha)M})$.

Proof. The quantization error term in (5.34) can be removed by having $\alpha = 1$. Then, I have

$$\mathbb{E}\left[\log_2\left(1+\rho\|\mathbf{h}_{\text{b},k}\|^2\right)\right] = \frac{1}{\ln 2}e^{\frac{1}{\rho M}}\Gamma\left(0, \frac{1}{\rho M}\right)$$

as $\frac{1}{M}\|\mathbf{h}_{\text{b},k}\|^2 = |g_k|^2 \sim \text{Exp}(1)$, and the ergodic sum rate becomes $\bar{\mathcal{R}}_{\text{inf}} = \frac{S}{\ln 2}e^{\frac{1}{\rho M}}\Gamma(0, \frac{1}{\rho M})$. ■

Note that as the number of quantization bits decreases to zero, $\bar{\mathcal{R}}_{\text{loss}}(\alpha)$ increases to $\bar{\mathcal{R}}_{\text{inf}}$, which leads $\bar{\mathcal{R}}_1 \to 0$. On the other hand, as the number of quantization bits increases to infinity, $\bar{\mathcal{R}}_{\text{loss}}(\alpha)$ decreases to zero, which leads $\bar{\mathcal{R}}_1 \to \bar{\mathcal{R}}_{\text{inf}}$. This complies with intuition.

Now, I focus on the second case where channels have arbitrary AoAs, which leads to the channel leakage effect in the beam domain due to phase offsets. The derived ergodic rate for the second case is shown in Proposition 6.

Proposition 6. *When channels have a single path and arbitrary AoAs regardless of the quantized angles of the analog combiner, a lower bound of the ergodic sum rate for $|\mathcal{S}_{\mathrm{cd}}| = S$ scheduled users with the proposed chordal distance-based scheduling algorithm is approximated as*

$$\bar{\mathcal{R}}_2^{lb} = \frac{S}{\ln 2}\Bigg(e^{\frac{1+\rho(1-\alpha)(S-1)M^2\mathcal{F}_2(M)}{\rho\alpha M+\rho(1-\alpha)M^2\mathcal{F}_1(M)}}\Gamma\Bigg(0,\frac{1+\rho(1-\alpha)(S-1)M^2\mathcal{F}_2(M)}{\rho\alpha M+\rho(1-\alpha)M^2\mathcal{F}_1(M)}\Bigg) - e^{\frac{1+\rho(1-\alpha)(S-1)M^2\mathcal{F}_2(M)}{\rho(1-\alpha)M^2\mathcal{F}_1(M)}}\Gamma\Bigg(0,\frac{1+\rho(1-\alpha)(S-1)M^2\mathcal{F}_2(M)}{\rho(1-\alpha)M^2\mathcal{F}_1(M)}\Bigg)\Bigg) \quad (5.32)$$

where $\mathcal{F}_1(M) = \int_0^1 F^4(\delta, M)\, d\delta$, $\mathcal{F}_2(M) = \left(\int_0^1 F^2(\delta, M)\, d\delta\right)^2$, and $F(\delta, M)$ is the Fejér kernel.

Proof. See Section 5.8. ■

Remark 16. *The derived ergodic rate expressions in* (5.31) *and* (5.32) *both converge to $\bar{\mathcal{R}}_{\mathrm{inf}}$ as the number of quantization bits increases:*

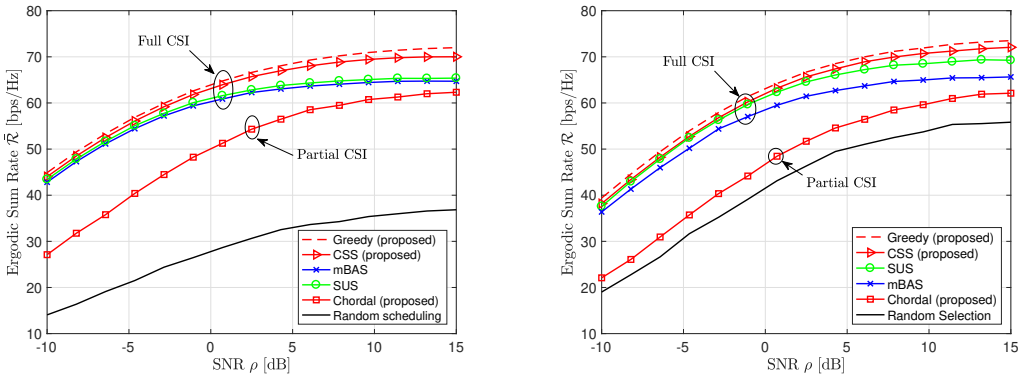
$$\bar{\mathcal{R}}_1,\ \bar{\mathcal{R}}_2^{lb} \to \frac{S}{\ln 2}e^{\frac{1}{\rho M}}\,\Gamma\left(0,\frac{1}{\rho M}\right), \quad \text{as } \alpha \to 1.$$

As the quantization precision increases, the lower bound in (5.38) becomes an exact expression, and (5.32) becomes an approximation of the ergodic rate itself rather than its lower bound. Accordingly, it can be inferred from Remark 16 that the two channel scenarios lead to different ergodic rates as a

consequence of quantization. In this regard, although a single path channel is considered, Propositions 5 and 6 still convey meaningful information as they not only provide closed-form ergodic rates but also specify the channel leakage effect in terms of ergodic rate for low-resolution ADCs. In addition, the single-path channel model is relevant to the case of unmanned aerial vehicle systems [145], which is of interest in upcoming 5G wireless communication systems. In Section 5.5, based on the intuition from Propositions 5 and 6, it can be shown that the channel leakage, indeed, positively affects the ergodic rate in the low-resolution ADC regime, and thus, makes the difference in the ergodic rates of the two channel scenarios.

5.5 Simulation Results

In this section, the proposed algorithms are evaluated, the derived ergodic rates are validated, and intuitions in this chapter are confirmed through simulations. In simulations, the number of channel paths L_k is distributed as $L_k \sim \max\{\text{Poission}(\lambda_L), 1\}$ [70] where λ_L represents the near average number of channel paths. I consider $M = 128$ BS antennas and $K = 200$ candidate users, and the BS schedules $S = 12$ users to serve at each transmission [146, 147]. Without imposing the constraint of $\|\mathbf{h}_{\mathrm{b},k}\| = \sqrt{\gamma_k}$, the following cases are evaluated through simulation: (1) CSS algorithm, (2) greedy algorithm, (3) chordal distance-based algorithm, (4) mmWave beam aggregation-based scheduling (mBAS) algorithm [49], and (5) SUS algorithm [44]. To provide a reference for a performance lower bound, a random scheduling case



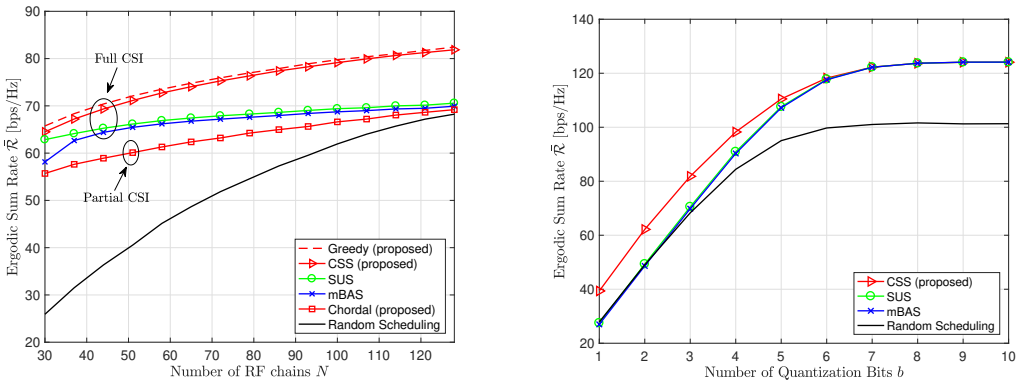
(a) $\lambda_L = 3$ (b) $\lambda_L = 8$

Figure 5.2: Uplink sum rate simulation results for $M = 128$ BS antennas, $N = 40$ RF chain, $K = 200$ candidate users, $S = 12$ scheduled users, and $b = 3$ quantization bits with (a) $\lambda_L = 3$ and (b) $\lambda_L = 8$ average channel paths,.

is also included. For the CSS and the mBAS algorithms, the BS stores $N_b = L_k$ indices of dominant elements in the effective channel $\mathbf{h}_{\mathrm{b},k}$. Parameters such as ϵ_{th}, N_{OL}, and d_{th} are optimally chosen unless mentioned otherwise.

5.5.1 Performance Validation

I first focus on performance validation of the proposed algorithms in sum rate. In Fig. 5.2, I consider $N = 40$ RF chains which is about 30% of the number of antennas $M = 128$ and $b = 3$ quantization bits. Fig. 5.2(a) shows the uplink sum rate with respect to the SNR ρ for $\lambda_L = 3$. The proposed CSS algorithm achieves the higher sum rate compared to the SUS and mBAS algorithms. In addition, the CSS algorithm attains the sum rate that is comparable to that of the proposed greedy algorithm which achieves the sub-optimal rate by requiring much higher complexity. The sum rate gap between

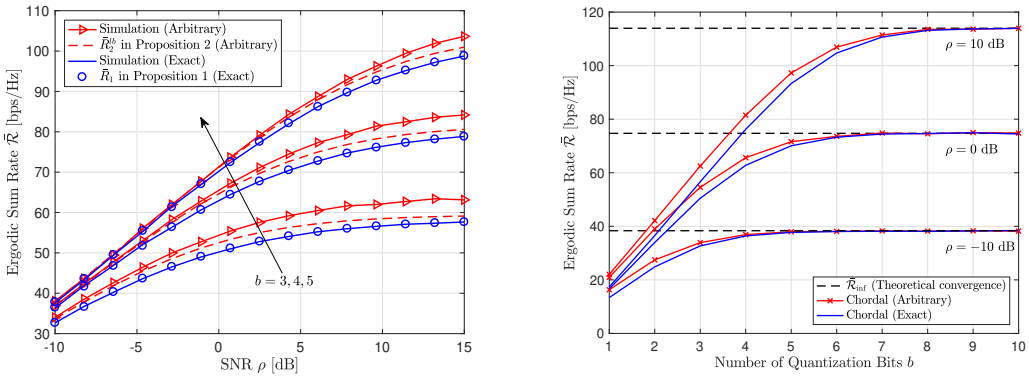


(a) (b)

Figure 5.3: Uplink sum rate for $M = 128$ antennas, $K = 200$ candidate users, $S = 12$ scheduled users, $\lambda_L = 3$ average channel paths, and $\rho = 6$ dB SNR with respect to the number of (a) RF chains N with $b = 3$ and (b) quantization bits b with $N = 128$.

the CSS and the prior algorithms—the SUS and mBAS algorithms—increases as ρ increases because the quantization noise becomes dominant compared to the AWGN in the high SNR regime.

Fig. 5.2(b) plots simulation results with $\lambda_L = 8$ average channel paths for $\sum_{k=1}^{S} L_S(k) > N$ where the condition in Theorem 1 does not hold. The proposed CSS algorithm achieves a higher sum rate than conventional scheduling methods, which shows that although the derived scheduling criteria may not be optimal in a practical system, they can still be effective for mmWave user scheduling as they capture a relationship between the sparse property of mmWave channels and quantization error. In Fig. 5.2(a) and (b), the chordal distance-based algorithm which only exploits the AoA knowledge improves the sum rate compared to random scheduling, closing the gap between the SUS



(a) (b)

Figure 5.4: (a) The analytical and simulation results for the uplink sum rate of the system with chordal distance-based scheduling, and (b) simulation results for the uplink sum rate of the system with chordal distance-based scheduling for $M = 128$ BS antennas, $N = 128$ RF chains, $K = 200$ candidate users, $S = 12$ scheduled users, and $L_k = 1$ channel path $\forall k$,

and mBAS algorithms. Therefore, the simulation results validate the sum rate performance of the proposed algorithms.

In Fig. 5.3(a), the sum rate results with respect to the number of RF chains N are presented for $\rho = 6$ dB. The CCS algorithm shows its sum rate that tightly aligns with that of the greedy algorithm, achieving the higher rate than the SUS and mBAS. In addition, the chordal distance-based algorithm shows a large improvement compared to the random scheduling for the low to medium N. As N increases, the effective channels $\mathbf{h}_{\text{b},k}$ are more likely to be orthogonal to each other for the fixed number of scheduled users, which enhances the performance of random scheduling. In this regard, the sum rates of the SUS and mBAS algorithms show the marginal sum rate increase compared to the random scheduling as N increases, whereas the CSS algorithm

still provides the noticeable improvement by mitigating quantization error.

Fig. 5.3(b) shows the uplink sum rate with respect to the number of quantization bits b. The CSS algorithm also attains the sum rate of the greedy algorithm with lower complexity and outperforms the SUS and mBAS algorithms. Note that the sum rate of the SUS and mBAS algorithms converges to that of the CSS and greedy algorithms as the number of quantization bits b increases; i.e., quantization error becomes negligible. This convergence corresponds to the fact that the derived criteria is effective under coarse quantization. Thus, in the low-resolution ADC regime, the CSS algorithm provides the noticeable sum rate improvement compared to the other algorithms that ignore quantization error.

5.5.2 Analysis Validation

The performance analysis and intuitions obtained from the analyses are validated in this subsection. In Fig. 5.4, $N = 128$ and $L_k = 1$, $\forall k$ are considered. As shown in Fig. 5.4(a), the derived ergodic rate (5.31) in Proposition 5 exactly matches the ergodic rate from the simulation. In addition, the lower bound approximation of ergodic rate (5.32) in Proposition 6 shows a small gap from the ergodic rate of the simulation, validating its analytical accuracy. In this regard, the derived ergodic rates can provide a performance guideline for the hybrid MIMO systems with the proposed chordal distance-based algorithm. From Fig. 5.4(a), the two different channel scenarios—exact AoA alignment and arbitrary AoAs—show difference in sum rate for the same

system configuration, as discussed in Remark 16. In the following simulation results, this phenomenon is numerically examined based on intuitions obtained in this chapter.

The sum rate of the chordal distance-based scheduling algorithm is evaluated with respect to the number of quantization bits b to find the behavior of the sum rate gap between the two channel scenarios: exact AoA alignment and arbitrary AoAs. In Fig. 5.4(b), it is shown that the uplink sum rates converges to $\bar{\mathcal{R}}_{\text{inf}} = \frac{S}{\ln 2} e^{\frac{1}{\rho M}} \Gamma\left(0, \frac{1}{\rho M}\right)$ as b increases. As discussed in Remark 16, such convergence of the sum rates implies that the two channel scenarios lead to different effects on quantization error. Note that the convergence rates are different for different ρ. When the SNR is low, the quantization noise is less dominant compared to the AWGN, which results in faster convergence in terms of the number of b, and vice versa. Therefore, it is concluded that coarse quantization causes the different sum rates from the channel scenarios.

In Fig. 5.5, I simulate the sum rates for the two channel scenarios with $N = 40$, $\lambda_L = 3$, and $b = 3$. Note that the sum rate for the arbitrary AoA channel is higher than that for the exact AoA alignment channel in the medium and high SNR regime in which the quantization noise is dominant over the AWGN. The quantization noise variance at the ith ADC is computed as $\mathbb{E}[|y_i - y_{\text{q},i}|^2] = \frac{\pi\sqrt{3}}{2}\sigma_i^2 2^{-2b}$ [56], where $\sigma_i^2 = \mathbb{E}[|y_i|^2] = p_u \|[\mathbf{H}_{\text{b}}]_{i,:}\|^2 + 1$. Therefore, without the phase offset, most σ_i^2 would be large whereas most σ_i^2 would be moderate with the phase offsets as the phase offsets spread the channel path gain at certain angles over the entire angles of RF chains. Consequently,

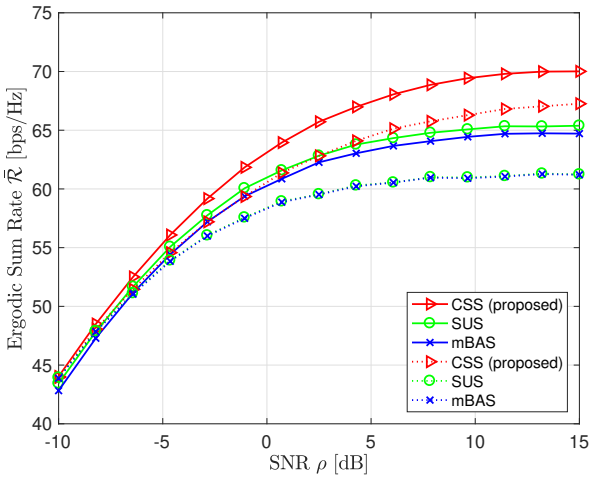


Figure 5.5: Uplink sum rate simulation results for $M = 128$ BS antennas, $N = 40$ RF chains, $K = 200$ candidate users, $S = 12$ scheduled users, $\lambda_L = 3$ average channel paths, and $b = 3$ quantization bits.

the phase offset reduces the overall quantization noise variance and this leads to the performance gain. This corresponds to the results in Theorem 1-(ii), i.e., it is more beneficial to have more spread beamspace gains than to have concentrated beamspace gains.

5.6 Conclusion

This chapter investigated user scheduling for mmWave hybrid beamforming systems with low-resolution ADCs. I derived new user scheduling criteria that are effective under coarse quantization. Leveraging the criteria, I developed the user scheduling algorithm which achieves the sub-optimal sum rate with low complexity, outperforming the conventional scheduling algorithms. I further proposed the chordal distance-based scheduling algorithm which only

exploits the AoA knowledge of channels. The chordal distance-based scheduling algorithm improved the sum rate compared to the random scheduling case, closing the gap between the full CSI-based conventional scheduling methods as the SNR increases. I also provided the performance analysis for the algorithm in ergodic rate, and the derived rates are the functions of system parameters including quantization bits. I obtained an intuition from the derived rates that channel leakage due to the phase offsets between the arbitrary AoAs and quantized angles of analog combiners offers the sum rate gain by reducing the quantization error compared to the channel without leakage. Therefore, for mmWave communications, this chapter provides not only new user scheduling algorithms for low-resolution ADC systems, but also new scheduling criteria and intuition for mmWave channels under coarse quantization. Concluding this dissertation, I will provide the summary of the contributions in the previous chapters and discuss potential future research directions in Chapter 6.

5.7 Proof of Proposition 1

Let the ZF combiner $\mathbf{W}_{\text{zf}} = \mathbf{H}_{\text{b}}(\mathcal{S}_{\text{cd}})(\mathbf{H}_{\text{b}}(\mathcal{S}_{\text{cd}})^H\mathbf{H}_{\text{b}}(\mathcal{S}_{\text{cd}}))^{-1}$. Using the achievable rate (5.8), the ergodic rate of user $k \in \mathcal{S}_{\text{cd}}$ is defined as

$$\begin{aligned}\bar{r}_k &= \mathbb{E}\Big[r_k\big(\mathbf{H}_{\text{b}}(\mathcal{S}_{\text{cd}})\big)\Big] \\ &= \mathbb{E}\left[\log_2\left(1+\frac{\alpha^2\rho}{\mathbf{w}_{\text{zf},k}^H\mathbf{R}_{\mathbf{qq}}(\mathbf{H}_{\text{b}}(\mathcal{S}_{\text{cd}}))\mathbf{w}_{\text{zf},k}+\alpha^2\|\mathbf{w}_{\text{zf},k}\|^2}\right)\right].\end{aligned} \tag{5.33}$$

Based on Remark 15, the algorithm schedules a user $j \in \mathcal{K}_{\text{cd}}$ who provides the smallest value of $|\mathbf{a}^H(\phi_k)\mathbf{a}(\phi_j)|$. Under the assumption of the ex-

act AoA alignment, $|\mathbf{a}^H(\phi_k)\mathbf{a}(\phi_j)|$ is equivalent to zero when $\mathcal{L}_k \cap \mathcal{L}_j = \emptyset$ for $k \neq j$, i.e., user channels are spatially orthogonal to each other. For the exact AoA alignment scenario with $L = 1$, there is only one non-zero element in $\mathbf{h}_{\mathrm{b},k}$. Accordingly, any scheduled users have to satisfy $\mathcal{L}_k \cap \mathcal{L}_j = \emptyset$ to avoid rank deficiency of a channel matrix, which can be guaranteed by setting $|\mathbf{a}^H(\phi_k)\mathbf{a}(\phi_{k'})| < \epsilon_{th} \ll 1$ in the filtering. Hence, the ZF combiner for user $k \in \mathcal{S}_{\mathrm{cd}}$ becomes $\mathbf{w}_{\mathrm{zf},k} = \mathbf{h}_{\mathrm{b},k}/\|\mathbf{h}_{\mathrm{b},k}\|^2$, and (5.33) is solved as

$$\bar{r}_k = \mathbb{E}\left[\log_2\left(1 + \frac{\alpha\rho\|\mathbf{h}_{\mathrm{b},k}\|^4}{\rho(1-\alpha)\mathbf{h}_{\mathrm{b},k}^H\mathbf{D}_{\mathbf{H}_{\mathrm{b}}}(\mathcal{S}_{\mathrm{cd}})\mathbf{h}_{\mathrm{b},k} + \|\mathbf{h}_{\mathrm{b},k}\|^2}\right)\right] \tag{5.34}$$

$$\overset{(a)}{=} \mathbb{E}\left[\log_2\left(1 + \frac{\alpha\rho}{(1-\alpha)\rho + 1/(M|g_k|^2)}\right)\right]$$

$$\overset{(b)}{=} \frac{1}{\ln 2}\left(e^{\frac{1}{\rho M}}\,\Gamma\left(0, \frac{1}{\rho M}\right) - e^{\frac{1}{\rho(1-\alpha)M}}\,\Gamma\left(0, \frac{1}{\rho(1-\alpha)M}\right)\right) \tag{5.35}$$

where $\mathbf{D}_{\mathbf{H}_{\mathrm{b}}}(\mathcal{S}_{\mathrm{cd}}) = \mathrm{diag}\Big(\mathbf{H}_{\mathrm{b}}(\mathcal{S}_{\mathrm{cd}})\mathbf{H}_{\mathrm{b}}(\mathcal{S}_{\mathrm{cd}})^H\Big)$, g_k is the complex gain of the propagation path of user k. Here, (a) is from $L = 1$ with $\mathcal{L}_k \cap \mathcal{L}_{k'} = \emptyset$ for $k, k' \in \mathcal{S}_{\mathrm{cd}}$, and (b) comes from the fact that $|g_k|^2$ is an exponential random variable with the rate parameter $\lambda = 1$, $|g_k|^2 \sim \mathrm{Exp}(1)$. Due to the randomness of g_k, the ergodic rate of each user is equal, which leads to (5.31). This completes the proof. ■

5.8 Proof of Proposition 2

To find a lower bound of the ergodic sum rate achieved by the proposed algorithm, I consider the random scheduling method and find its ergodic sum rate for the lower bound. Since I focus on a large antenna array system at the

BS, the array response vectors of the scheduled users are almost orthogonal with large M [98], and thus I adopt $\mathbf{w}_{\mathrm{zf},k} \approx \frac{\mathbf{A}^H \mathbf{h}_k}{\|\mathbf{h}_k\|^2}$. Then, the ergodic rate of the scheduled user k can be approximated as

$$\begin{aligned}\bar{r}_k =& \mathbb{E}\left[\log_2\left(1+\frac{\alpha^2\rho}{\mathbf{w}_{\mathrm{zf},k}^H \mathbf{R}_{\mathbf{qq}}\big(\mathbf{H}_{\mathrm{b}}(\mathcal{S}_{\mathrm{cd}})\big)\mathbf{w}_{\mathrm{zf},k}+\alpha^2\|\mathbf{w}_{\mathrm{zf},k}\|^2}\right)\right] \\ \overset{(a)}{\approx}& \mathbb{E}\left[\log_2\left(1+\frac{\alpha\rho\|\mathbf{h}_k\|^4}{\rho(1-\alpha)(\mathbf{A}^H\mathbf{h}_k)^H\mathbf{D}_{\mathbf{A}^H\mathbf{H}}(\mathcal{S}_{\mathrm{cd}})\mathbf{A}^H\mathbf{h}_k+\|\mathbf{h}_k\|^2}\right)\right] \end{aligned} \tag{5.36}$$

where $\mathbf{D}_{\mathbf{A}^H\mathbf{H}}(\mathcal{S}_{\mathrm{cd}}) = \mathrm{diag}\big(\mathbf{A}^H\mathbf{H}(\mathcal{S}_{\mathrm{cd}})\mathbf{H}(\mathcal{S}_{\mathrm{cd}})^H\mathbf{A}\big)$, (a) comes from $\mathbf{w}_{\mathrm{zf},k} \approx \frac{\mathbf{A}^H\mathbf{h}_k}{\|\mathbf{h}_k\|^2}$. Without loss of generality, let $\mathcal{S}_{\mathrm{cd}} = \{1, 2, \ldots, S\}$. The channel matrix of scheduled users can be represented as $\mathbf{H}(\mathcal{S}_{\mathrm{cd}}) = \sqrt{M}\mathbf{A}_u\mathbf{G}$ where $\mathbf{A}_u = [\mathbf{a}(\varphi_1), \ldots, \mathbf{a}(\varphi_S)]$ and $\mathbf{G} = \mathrm{diag}(g_1, \ldots, g_S)$, and (5.36) becomes

$$\begin{aligned}&\mathbb{E}\left[\log_2\left(1+\frac{M^2\alpha\rho|g_k|^4}{M^2\rho(1-\alpha)|g_k|^2\mathbf{a}^H(\varphi_k)\mathbf{A}\bar{\mathbf{D}}\mathbf{A}^H\mathbf{a}(\varphi_k)+M|g_k|^2}\right)\right] \\ &= \mathbb{E}\left[\log_2\left(1+\frac{M\alpha\rho|g_k|^2}{\Psi_k+1}\right)\right] \\ &= \mathbb{E}_{g_k}\left[\mathbb{E}\left[\log_2\left(1+\frac{M\alpha\rho|g_k|^2}{\Psi_k+1}\right)\middle| g_k\right]\right]. \end{aligned} \tag{5.37}$$

where $\bar{\mathbf{D}} = \mathrm{diag}\big(\mathbf{A}^H\mathbf{A}_u\mathbf{G}\,\mathbf{G}^H\mathbf{A}_u^H\mathbf{A}\big)$ and

$$\Psi_k = M\rho(1-\alpha)\sum_{m,s=1}^{M,S}|g_s|^2|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)|^2|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_s)|^2.$$

To compute the inner expectation in (5.37), I can use Lemma 1 in [148] as g_k is considered to be a constant given the condition, which makes the signal power and the interference-plus-noise power independent to each other. Then

the inner expectation in (5.37) becomes

$$\mathbb{E}\left[\log_2\left(1+\frac{M\alpha\rho|g_k|^2}{\Psi_k+1}\right)\middle|g_k\right] \overset{(a)}{=} \frac{1}{\ln 2}\int_0^\infty \frac{e^{-z}}{z}\left(1-e^{-zM\alpha\rho|g_k|^2}\right)\mathbb{E}\left[e^{-z\Psi_k}\middle|g_k\right]dz$$
$$\overset{(b)}{\geq} \frac{1}{\ln 2}\int_0^\infty \frac{e^{-z}}{z}\left(1-e^{-zM\alpha\rho|g_k|^2}\right)e^{-z\mathbb{E}[\Psi_k|g_k]}dz \tag{5.38}$$

where (a) follows from Lemma 1 in [148] and (b) comes from Jensen's inequality. To compute the expectation in (5.38), I rewrite it as

$$\mathbb{E}\big[\Psi_k|g_k\big] = M\rho(1-\alpha)\Bigg(\mathbb{E}\left[\sum_{m=1}^{M}|g_k|^2\left|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)\right|^4\middle|g_k\right]$$
$$+\,\mathbb{E}\left[\sum_{m=1}^{M}\sum_{s\neq k}^{S}|g_s|^2\left|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)\right|^2\left|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_s)\right|^2\right]\Bigg). \tag{5.39}$$

The first expectation term in (5.39) can be computed as

$$\mathbb{E}\left[\sum_{m=1}^{M}|g_k|^2\left|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)\right|^4\middle|g_k\right] = |g_k|^2\sum_{m=1}^{M}\mathbb{E}\left[\left|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)\right|^4\right]$$
$$\overset{(a)}{=} |g_k|^2 M\int_0^1 F^4\left(\delta;M\right)\,d\delta \tag{5.40}$$

where (a) comes from the fact that $\delta_{m,k} := \vartheta_m - \varphi_k$ can be regarded as $\delta_{m,k} \overset{i.i.d.}{\sim} \text{Unif}\big[-1,1\big]$ due to the symmetry of the Fejér kernel of order M, $F(\vartheta;M)$ [149].

Then, with $\mathbb{E}[|g_s|^2] = 1$, the second expectation term can be expressed as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{m=1}^{M}\sum_{s\neq k}^{S}|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)|^2|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_s)|^2\right] \\
&=\sum_{m=1}^{M}\sum_{s\neq k}^{S}\mathbb{E}\Big[|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_k)|^2\Big]\mathbb{E}\Big[|\mathbf{a}^H(\vartheta_m)\mathbf{a}(\varphi_s)|^2\Big] \\
&=\sum_{m=1}^{M}\sum_{s\neq k}^{S}\mathbb{E}\left[F^2\left(\delta_{m,k};M\right)\right]\mathbb{E}\left[F^2\left(\delta_{m,s};M\right)\right] \\
&=(S-1)M\left(\int_0^1 F^2\left(\delta;M\right)\,d\delta\right)^2. \qquad (5.41)
\end{aligned}$$

Let $c_1 = M\alpha\rho$, $c_2 = M^2\rho(1-\alpha)\int_0^1 F^4\left(\delta;M\right)\,d\delta$, and $c_3 = M^2\rho(1-\alpha)(S-1)\left(\int_0^1 F^2\left(\delta;M\right)\,d\delta\right)^2$. From (5.37), (5.38),(5.40), and (5.41), the ergodic rate $\bar{r}_k$ is approximately lower bounded by

$$\begin{aligned}
\bar{r}_k &\approx \mathbb{E}_{g_k}\left[\mathbb{E}\left[\log_2\left(1+\frac{c_1|g_k|^2}{\Psi_k+1}\right)\middle|g_k\right]\right] \\
&\geq \frac{1}{\ln 2}\mathbb{E}_{g_k}\left[\int_0^\infty \frac{e^{-z}}{z}\left(1-e^{-zc_1|g_k|^2}\right)e^{-z\mathbb{E}[\Psi_k|g_k]}dz\right] \\
&= \frac{1}{\ln 2}\int_0^\infty \frac{e^{-(1+c_3)z}}{z}\left(\mathbb{E}_{g_k}\left[e^{-c_2z|g_k|^2}\right]-\mathbb{E}_{g_k}\left[e^{-(c_1+c_2)z|g_k|^2}\right]\right)dz \\
&\overset{(a)}{=} \frac{1}{\ln 2}\int_0^\infty \frac{e^{-(1+c_3)z}}{z}\left(\frac{1}{1+c_2z}-\frac{1}{1+(c_1+c_2)z}\right)dz \\
&= \frac{1}{\ln 2}\left(e^{\frac{1+c_3}{c_1+c_2}}\,\Gamma\left(0,\frac{1+c_3}{c_1+c_2}\right)-e^{\frac{1+c_3}{c_2}}\,\Gamma\left(0,\frac{1+c_3}{c_2}\right)\right) \qquad (5.42)
\end{aligned}$$

where (a) comes from the Laplace transform of the exponential distribution $|g_k|^2 \sim \exp(1)$. Without the fading information of channels, the ergodic rate for each user after the user scheduling is equivalent to each other, which results in (5.32). This completes the proof. ■

Chapter 6

Concluding Remarks

This chapter concludes the dissertation with a summary of contributions in Section 6.1 and potential future research directions in Section 6.2.

6.1 Summary

In this dissertation, I developed advanced receiver designs and derived user scheduling criteria for hybrid analog-and-digital beamforming systems with low-resolution ADCs. Due to the non-negligible quantization error, existing hybrid beamforming techniques cannot be directly applied to the considered systems as they ignore the change of the quantization error. Accordingly, it is essential to consider advanced low-resolution ADC systems that can adopt existing hybrid beamforming techniques without significant performance loss and that can mitigate quantization error in the analog preprocessing while maintaining large channel gains. In addition to the advanced receiver design, it is also critical to develop techniques that are used in the higher stack of network such as user scheduling for hybrid beamforming systems with low-resolution ADCs by incorporating the effect of quantization error.

In the first part of this dissertation, I focused on optimizing the reso-

lutions of ADCs under a power constraint and proposed resolution-adaptive ADC networks for hybrid beamforming receivers for phase shifter-based hybrid beamforming systems. To find the optimal ADC bit distribution for a given power constraint, I derived a near-optimal bit allocation solution that minimizes the total mean squared quantization error. Since the solution is derived in closed form, the ADC bit distribution can be determined with low-complexity. In addition, existing hybrid beamforming techniques can be readily applied to the proposed system as the solution minimizes quantization error for the limited power consumption.

In the second part of this dissertation, I focused on optimizing an analog combining architecture to mitigate quantization error for fixed-resolution ADC receivers. By solving a mutual information (MI) maximization problem without a constant modulus constraint on analog combiners, I derived an optimal two-stage combiner: a channel gain aggregation stage followed by a spreading stage to maximize the MI by effectively managing quantization error. I showed that the derived two-stage combiner achieves the optimal scaling law with respect to the number of RF chains and maximizes the MI for homogeneous singular values of a MIMO channel. Then, I developed a two-stage analog combining algorithm to implement the derived solution under a constant modulus constraint for mmWave channels.

Considering switch-based analog beamforming instead of phase shifter-based beamforming for reducing implementation cost and complexity, I studied antenna selection problems for low-resolution ADC systems in the third part of

this dissertation. For the downlink transmit antenna selection with ZF precoding case, I showed that the problem is same for both high- and low-resolution ADC receivers. For the uplink receive antenna selection case, however, the quantization error makes the problem different from that of high-resolution ADC systems. In this regard, I derived a quantization-aware selection criterion and developed a quantization-aware greedy antenna selection algorithm with subsequent analysis.

In the fourth part of this dissertation, I derived user scheduling criteria for hybrid beamforming receivers with low-resolution ADCs. Since existing criteria ignore the impact of quantization error when scheduling users, the derived scheduling criteria provides the two key ideas to reduce the quantization error: *i*) unique AoAs for the channel paths of each scheduled user and *ii*) equal power spread across the complex path gains within each user channel. Leveraging the derived criteria, I developed user scheduling algorithms for coarse quantization systems for perfect and partial CSI cases. Simulation results validated the performance of the proposed algorithms and analyses in this dissertation.

6.2 Future work

In this dissertation, I addressed some of the main critical issues to adopt hybrid analog-and-digital beamforming with low-resolution ADCs in large antenna array systems. There are still issues left that need to be resolved to successfully realize mmWave communication systems. Therefore, I present

promising future research directions related to the topics in this dissertation.

- **Channel estimation in the two-stage analog combining system:** The two-stage analog combining structure was proposed in Chapter 3. Assuming the CSI at the receiver, the proposed two-stage analog combining achieved optimality in the scaling law and maximizing the mutual information. Then, the next question would be how to estimate the channel with the two-stage analog combining structure. Based on the linear approximation model of the quantization process, existing channel estimation techniques for hybrid beamforming systems can be applied to the two-stage analog combining with low-resolution ADC systems after multiplying the matrix inversion of the second analog combiner since it is nonsingular. As the second analog combiner leads to relatively even distribution of quantization errors over ADCs while maintaining the total quantization error, it is expected that the estimation of the quantization variance would be easier than one-stage analog combining case. Due to the approximation error, however, it is possible to obtain CSI with undesirable amount of distortion if the existing techniques are applied without modification. In addition, possible phase error of the cascaded phase shifter networks can further make the channel estimation more challenging. Therefore, it is necessary to thoroughly investigate channel estimation for the two-stage combining systems with the exact quantization model under the potential error from phase shifter networks.

- **Extension of the receiver design work into wideband communications:** The antenna selection problems studied in Chapter 4 showed that similar intuitions and solutions hold for both narrowband and wideband OFDM communications. The other system designs—resolution-adaptive ADCs in Chapter 2 and two-stage analog combining in Chapter 3—considered narrowband communications only. It is also possible for the systems to have similar results in both narrowband and wideband channels as the antenna selection system. The resolution-adaptive ADC system, however, results in different quantization resolutions for each received signal, and thus, the different quantization distortion level can significantly degrade the system performance in the wideband OFDM system. In this respect, more rigorous and precise study would be desirable for the wideband OFDM communications. For the two-stage analog combining, the proposed combining solution may not be near optimal in the OFDM system since the entire subcarriers share the same analog combining. Consequently, it is necessary to find the optimal analog combining structure that works for all subcarriers.

- **Cooperation of multiple base stations under limited total power consumption:** The bit allocation solution for the resolution-adaptive ADC system was derived by considering the power constraint within a single BS. To achieve higher energy-efficiency with higher sum spectral efficiency over multiple cells, the optimization of the bit allocation needs to be solved by considering the total power constraint for multiple BSs.

In this case, BSs with different channel conditions and user distributions can be allowed to use more/less power so that the distribution of the energy for the BSs can be more flexible and achieve higher efficiency. To this end, the optimization can be performed in two steps for low complexity. The amount of energy distribution can be first decided assuming perfect quantization at each BSs. Then, the closed-form bit allocation solution derived in Chapter 2 can be applied to each BSs with minor modifications. However, this approach will only provide suboptimal solutions which may be far from the optimal solution. In addition, a fairness issue needs to be considered in the multi-cell optimization problem. Accordingly, the more elaborate study is necessary to accomplish highly spectrum- and energy-efficient future wireless systems.

Bibliography

[1] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

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

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

[4] Z. Pi and F. Khan, "A millimeter-wave massive MIMO system for next generation mobile broadband," in *Proc. Asilomar Conf. Signals, Systems and Comp.*, Nov. 2012, pp. 693–698.

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

[6] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, 2015.

[7] 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.

[8] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal Sel. Areas in Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[9] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[10] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[11] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Jul. 2015.

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

[13] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Sel. Areas in Commun.*, vol. 17, no. 4, pp. 539–550, Apr.

1999.

[14] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath, "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *Proc. Asilomar Conf. Signals, Systems and Comp.*, Nov. 2014, pp. 957–961.

[15] J. Choi, J. Mo, and R. W. Heath, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2005–2018, Mar. 2016.

[16] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. on Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug 2017.

[17] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, 2016.

[18] S. Wang, Y. Li, and J. Wang, "Multiuser detection for uplink large-scale MIMO under one-bit quantization," in *IEEE Int. Conf. Commun.*, 2014, pp. 4460–4465.

[19] ——, "Multiuser Detection in Massive Spatial Modulation MIMO With Low-Resolution ADCs," *IEEE Trans. on Wireless Commun.*, vol. 14,

no. 4, pp. 2156–2168, April 2015.

[20] J. Mo, P. Schniter, and R. W. Heath, "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," *IEEE Trans. on Signal Process.*, vol. 66, no. 5, pp. 1141–1154, 2017.

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

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

[23] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, Jan. 2016.

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

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

[26] O. El Ayach, R. W. Heath, S. Abu-Surra, S. Rajagopal, and Z. Pi, "The capacity optimality of beam steering in large millimeter wave MIMO systems," in *IEEE Int. Work. Signal Process. Advances in Wireless Commun.*, 2012, pp. 100–104.

[27] ——, "Low complexity precoding for large millimeter wave MIMO systems," in *IEEE Int. Conf. Commun.*, 2012, pp. 3724–3729.

[28] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *IEEE Inform. Theory and App. Work.*, 2013, pp. 1–5.

[29] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, 2014.

[30] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, 2015.

[31] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale MIMO systems," in *IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2015, pp. 2929–2933.

[32] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, "Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate

tradeoffs," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 4, pp. 2274–2287, 2017.

[33] N. Liang and W. Zhang, "Mixed-ADC massive MIMO," *IEEE Journal Sel. Areas in Commun.*, vol. 34, no. 4, pp. 983–997, Mar. 2016.

[34] T.-C. Zhang, C.-K. Wen, S. Jin, and T. Jiang, "Mixed-ADC massive MIMO detectors: Performance analysis and design optimization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7738–7752, 2016.

[35] W. B. Abbas, F. Gomez-Cuba, and M. Zorzi, "Millimeter wave receiver efficiency: A comprehensive comparison of beamforming schemes with low resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8131–8146, 2017.

[36] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek, "A Comparison of Hybrid Beamforming and Digital Beamforming with Low-Resolution ADCs for Multiple Users and Imperfect CSI," *IEEE J. Sel. Topics Signal Process.*, pp. 484–498, 2018.

[37] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in real propagation environments: Do all antennas contribute equally?" *IEEE Trans. on Commun.*, vol. 63, no. 11, pp. 3917–3928, Nov. 2015.

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

[39] A. F. Molisch, M. Z. Win, , and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Trans. on Wireless Commun.*, vol. 4, no. 4, pp. 1759–1772, July 2005.

[40] A. Dua, K. Medepalli, and A. J. Paulraj, "Receive antenna selection in MIMO systems using convex optimization," *IEEE Trans. on Wireless Commun.*, vol. 5, no. 9, pp. 2353–2357, Sep. 2006.

[41] R. Vaze and H. Ganapathy, "Sub-Modularity and Antenna Selection in MIMO Systems," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1446–1449, Sep. 2012.

[42] Y. Liu, Y. Zhang, C. Ji, W. Q. Malik, and D. J. Edwards, "A low-complexity receive-antenna-selection algorithm for MIMO–OFDM wireless systems," *IEEE Trans. on Veh. Technol.*, vol. 58, no. 6, pp. 2793–2802, Dec. 2009.

[43] A. B. Narasimhamurthy and C. Tepedelenlioglu, "Antenna Selection for MIMO-OFDM Systems With Channel Estimation Error," *IEEE Trans. on Veh. Technol.*, vol. 58, no. 5, pp. 2269–2278, Jun 2009.

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

[45] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. on Inform. Theory*, vol. 51,

no. 2, pp. 506–522, 2005.

[46] E. Liu and K. K. Leung, "Expected throughput of the proportional fair scheduling over Rayleigh fading channels," *IEEE Commun. Lett.*, vol. 14, no. 6, 2010.

[47] R. Rajashekar and L. Hanzo, "User Selection Algorithms for Block Diagonalization Aided Multiuser Downlink mm-Wave Communication," *IEEE Access*, vol. 5, pp. 5760–5772, 2017.

[48] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in millimeter-wave multiuser MISO downlink," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, 2016.

[49] G. Lee, Y. Sung, and M. Kountouris, "On the performance of random beamforming in sparse millimeter wave channels," *IEEE Journal Sel. Topics in Signal Process.*, vol. 10, no. 3, pp. 560–575, 2016.

[50] H.-S. Lee and C. G. Sodini, "Analog-to-digital converters: Digitizing the analog world," *Proc. of the IEEE*, vol. 96, no. 2, pp. 323–334, 2008.

[51] A. Mezghani and J. A. Nossek, "On ultra-wideband MIMO systems with 1-bit quantized outputs: Performance analysis and input optimization," in *IEEE Int. Symposium Inform. Theory*, 2007, pp. 1286–1289.

[52] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," *arXiv preprint arXiv:1404.7736*, Apr. 2014.

[53] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: Channel estimation and high-order modulations," in *IEEE Int. Conf. Commun. Work.*

[54] A. Mezghani and J. A. Nossek, "Capacity lower bound of MIMO channels with output quantization and correlated noise," in *IEEE Int. Symposium Inform. Theory*, 2012.

[55] Y. Li, C. Tao, L. Liu, G. Seco-Granados, and A. L. Swindlehurst, "Channel estimation and uplink achievable rates in one-bit massive MIMO systems," in *IEEE Sensor Array and Multichannel Signal Process. Work.*, 2016, pp. 1–5.

[56] O. Orhan, E. Erkip, and S. Rangan, "Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance," in *IEEE Inform. Theory and App. Work.*, Feb. 2015, pp. 191–198.

[57] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Oct. 2015.

[58] J. Zhang, L. Dai, S. Sun, and Z. Wang, "On the spectral efficiency of massive MIMO systems with low-resolution ADCs," *IEEE Commun. Lett..*, vol. 20, no. 5, pp. 842–845, Feb. 2016.

[59] J. Zhang, L. Dai, Z. He, S. Jin, and X. Li, "Performance Analysis of Mixed-ADC Massive MIMO Systems over Rician Fading Channels," *IEEE Journal Sel. Areas in Commun.*, vol. 35, no. 6, pp. 1327–1338, Jun. 2017.

[60] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, "Achievable rates of hybrid architectures with few-bit ADC receivers," in *VDE Int. ITG Work. Smart Antennas*, 2016, pp. 1–8.

[61] J. Choi, B. L. Evans, and A. Gatherer, "ADC Bit Allocation under a Power Constraint for MmWave Massive MIMO Communication Receivers," in *IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2017.

[62] F. Khan, Z. Pi, and S. Rajagopal, "Millimeter-wave mobile broadband with large scale spatial processing for 5G mobile communication," in *IEEE Annual Allerton Conf. Commun., Control, and Computing*, 2012, pp. 1517–1523.

[63] A. Lozano and N. Jindal, "Are yesterday-s information-theoretic fading models and performance metrics adequate for the analysis of today's wireless systems?" *IEEE Commun. Mag.*, vol. 50, no. 11, 2012.

[64] J. Yoo, D. Lee, K. Choi, and J. Kim, "A power and resolution adaptive flash analog-to-digital converter," in *ACM Int. Symposium on Low Power Electronics and Design*, 2002, pp. 233–236.

[65] S. Nahata, K. Choi, and J. Yoo, "A high-speed power and resolution adaptive flash analog-to-digital converter," in *IEEE Int. System-on-Chip Conf.*, 2004, pp. 33–36.

[66] G. Rajashekar and M. Bhat, "Design of Resolution Adaptive TIQ Flash ADC using AMS 0.35μm technology," in *IEEE Int. Conf. Electronic Design*, 2008, pp. 1–6.

[67] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 69–77, 2005.

[68] A. M. Sayeed and V. Raghavan, "Maximizing MIMO capacity in sparse multipath with reconfigurable antenna arrays," *IEEE Journal Sel. Topics in Signal Process.*, vol. 1, no. 1, pp. 156–166, 2007.

[69] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE Journal Sel. Topics in Signal Process.*, vol. 10, no. 3, pp. 436–453, Feb. 2016.

[70] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal Sel. Areas in Commun.*, vol. 32, no. 6, pp. 1164–1179, 2014.

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

[72] T. Kim and D. J. Love, "Virtual AoA and AoD estimation for sparse millimeter wave MIMO channels," in *IEEE Int. Work. Signal Process. Advances in Wireless Commun., 2015*, 2015, pp. 146–150.

[73] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust predictive quantization: Analysis and design via convex optimization," *IEEE Journal Sel. Topics in Signal Process.*, vol. 1, no. 4, pp. 618–632, 2007.

[74] J. Lee, G.-T. Gil, and Y. H. Lee, "Exploiting spatial sparsity for estimating channels of hybrid MIMO systems in millimeter wave communications," in *IEEE Global Commun. Conf.*, 2014, pp. 3326–3331.

[75] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, 2016.

[76] J. Choi, B. L. Evans, and A. Gatherer, "Space-time fronthaul compression of complex baseband uplink LTE signals," in *in Proc. IEEE Int. Conf. Commun.*, July. 2016, pp. 1–6.

[77] W. Zhang, "A general framework for transmission with transceiver distortion and some applications," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 384–399, 2012.

[78] ——, "A remark on channels with transceiver distortion," in *IEEE Inform. Theory and App. Work.,*, 2016, pp. 1–4.

[79] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE Journal Sel. Topics in Signal Process.*, vol. 8, no. 5, pp. 966–981, 2014.

[80] V. Raghavan and A. M. Sayeed, "Sublinear capacity scaling laws for sparse MIMO channels," *IEEE Trans. Inform. Theory*, vol. 57, no. 1, pp. 345–364, 2011.

[81] H. Chung, A. Rylyakov, Z. T. Deniz, J. Bulzacchelli, G.-Y. Wei, and D. Friedman, "A 7.5-GS/s 3.8-ENOB 52-mW flash ADC with clock duty cycle control in 65nm CMOS," in *Symposium VLSI Circuits*, 2009, pp. 268–269.

[82] P. Sudarshan, N. B. Mehta, A. F. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, 2006.

[83] F. Gholam, J. Vía, and I. Santamaría, "Beamforming design for simplified analog antenna combining architectures," *IEEE Trans. Veh. Technol.*, vol. 60, no. 5, pp. 2373–2378, 2011.

[84] D. Ying, F. W. Vook, T. A. Thomas, and D. J. Love, "Hybrid structure in massive MIMO: Achieving large sum rate with fewer RF chains," in *IEEE Int. Conf. Commun.*, 2015, pp. 2344–2349.

[85] F. Sohrabi and W. Yu, "Hybrid Digital and Analog Beamforming Design for Large-Scale Antenna Arrays," *IEEE J. Sel. Topics in Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.

[86] J. Li, L. Xiao, X. Xu, and S. Zhou, "Robust and low complexity hybrid beamforming for uplink multiuser mmWave MIMO systems," *IEEE Commu. Lett.*, vol. 20, no. 6, pp. 1140–1143, 2016.

[87] T. E. Bogale and L. B. Le, "Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital," *IEEE Global Commun. Conf.*, pp. 4066–4071, 2014.

[88] C. Rusu, R. Méndez-Rial, N. González-Prelcicy, and R. W. Heath, "Low complexity hybrid sparse precoding and combining in millimeter wave MIMO systems," in *IEEE Int. Conf. Commun.*, 2015, pp. 1340–1345.

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

[90] J. Choi, B. L. Evans, and A. Gatherer, "Resolution-adaptive hybrid MIMO architectures for millimeter wave communications," *IEEE Trans. Signal Process.*, vol. 65, no. 23, pp. 6201–6216, 2017.

[91] J. Choi, G. Lee, and B. L. Evans, "User scheduling for millimeter wave hybrid beamforming systems with low-resolution ADCs," *IEEE Trans. on Wireless Commun.*, vol. 18, no. 4, pp. 2401–2414, 2019.

[92] R. B. Ertel, P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Personal Commun.*, vol. 5, no. 1, pp. 10–22, 1998.

[93] A. Simonsson and A. Furuskar, "Uplink power control in LTE-overview and performance, subtitle: principles and benefits of utilizing rather than compensating for SINR variations," in *IEEE Veh. Technol. Conf.*, 2008, pp. 1–5.

[94] E. Tejaswi and B. Suresh, "Survey of power control schemes for LTE uplink," *Int. J. Computer Science and Inform. Technol.*, vol. 10, pp. 369–373, 2013.

[95] S. He, Y. Huang, S. Jin, F. Yu, and L. Yang, "Max-Min Energy Efficient Beamforming for Multicell Multiuser Joint Transmission Systems." *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1956–1959, 2013.

[96] Y.-F. Liu, M. Hong, and Y.-H. Dai, "Max-Min Fairness Linear Transceiver Design Problem for a Multi-User SIMO Interference Channel is Polynomial Time Solvable." *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 27–30, 2013.

[97] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.

[98] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive MIMO," in *European Signal Process. Conf.*,

2014, pp. 76–80.

[99] 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.

[100] O. E. Ayach, R. Heath, S. Abu-surra, S. Rajagopal, and Z. Pi, "The capacity optimality of beam steering in large millimeter wave MIMO systems," in *IEEE SPAWC*, 2012, pp. 100–104.

[101] Y. Chen, D. Chen, and T. Jiang, "Non-uniform quantization codebook based hybrid precoding to reduce feedback overhead in millimeter wave MIMO systems," *to appear in IEEE Trans. Commun.*, 2018.

[102] Y. Yu, P. Baltus, A. van Roermund, D. Jeurissen, A. de Graauw, E. van der Heijden, and R. Pijper, "A 60GHz digitally controlled phase shifter in CMOS," in *European Solid-State Circuits Conf.* IEEE, 2008, pp. 250–253.

[103] F. Ellinger, H. Jackel, and W. Bachtold, "Varactor-loaded transmission-line phase shifter at C-band using lumped elements," *IEEE Trans. on Microwave Theory and Tech.*, vol. 51, no. 4, pp. 1135–1140, 2003.

[104] T.-W. Li and H. Wang, "A Millimeter-Wave Fully Integrated Passive Reflection-Type Phase Shifter With Transformer-Based Multi-Resonance Loads for 360 Phase Shifting," *IEEE Trans. on Circuits and Systems I*, vol. 65, no. 4, pp. 1406–1419, 2018.

[105] F. E. Fakoukakis, T. N. Kaifas, E. E. Vafiadis, and G. A. Kyriacou, "Design and implementation of Butler matrix-based beam-forming networks for low sidelobe level electronically scanned arrays," *Int. J. of Microwave and Wireless Technol.*, vol. 7, no. 1, pp. 69–79, 2015.

[106] P. Liu, S. Jin, T. Jiang, Q. Zhang, and M. Matthaiou, "Pilot power allocation through user grouping in multi-cell massive MIMO systems," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1561–1574, 2017.

[107] E. Bjornson, M. Matthaiou, and M. Debbah, "Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4353–4368, 2015.

[108] Y. Yapıcı and I. Güvenç, "Low-complexity adaptive beam and channel tracking for mobile mmWave communications," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 572–576.

[109] Z. Chen, J. Yuan, and B. Vucetic, "Analysis of transmit antenna selection/maximal-ratio combining in Rayleigh fading channels," *IEEE Trans. on Veh. Technol.*, vol. 54, no. 4, pp. 1312–1321, Jul. 2005.

[110] S. Sanayei and A. Nosratinia, "Capacity of MIMO Channels With Antenna Selection," *IEEE Trans. on Inform. Theory*, vol. 53, no. 11, pp. 4356–4362, Nov. 2007.

[111] X. Gao, O. Edfors, J. Liu, and F. Tufvesson, "Antenna selection in measured massive MIMO channels using convex optimization," in *IEEE*

Global Commun. Conf. Workshops, Dec. 2013, pp. 129–134.

[112] S. Khademi, E. DeCorte, G. Leus, and A. van der Veen, "Convex optimization for joint zero-forcing and antenna selection in multiuser MISO systems," in *IEEE Int. Workshop on Signal Process. Adv. in Wireless Commun.*, Jun. 2014, pp. 30–34.

[113] X. Zhang, Z. Lv, and W. Wang, "Performance analysis of multiuser diversity in MIMO systems with antenna selection," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 1, pp. 15–21, Jan. 2008.

[114] P. V. Amadori and C. Masouros, "Large Scale Antenna Selection and Precoding for Interference Exploitation," *IEEE Trans. on Commun.*, vol. 65, no. 10, pp. 4529–4542, Oct. 2017.

[115] Z. Liu, W. Du, and D. Sun, "Energy and Spectral Efficiency Tradeoff for Massive MIMO Systems With Transmit Antenna Selection," *IEEE Trans. on Veh. Technol.*, vol. 66, no. 5, pp. 4453–4457, May 2017.

[116] P. Yang, Y. Xiao, Y. L. Guan, S. Li, and L. Hanzo, "Transmit Antenna Selection for Multiple-Input Multiple-Output Spatial Modulation Systems," *IEEE Trans. on Commun.*, vol. 64, no. 5, pp. 2035–2048, May 2016.

[117] Y. Zhang, C. Ji, W. Q. Malik, D. C. O'Brien, and D. J. Edwards, "Receive antenna selection for MIMO systems over correlated fading chan-

nels," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 9, pp. 4393–4399, Sep. 2009.

[118] L. Dai, S. Sfar, and K. B. Letaief, "Optimal antenna selection based on capacity maximization for MIMO systems in correlated channels," *IEEE Trans. on Commun.*, vol. 54, no. 3, pp. 563–573, Mar. 2006.

[119] P. V. Amadori and C. Masouros, "Low RF-complexity millimeter-wave beamspace-MIMO systems by beam selection," *IEEE Trans. on Commun.*, vol. 63, no. 6, pp. 2212–2223, May 2015.

[120] H. Li, Q. Liu, Z. Wang, and M. Li, "Joint Antenna Selection and Analog Precoder Design With Low-Resolution Phase Shifters," *IEEE Trans. on Veh. Technol.*, vol. 68, no. 1, pp. 967–971, Jan 2019.

[121] M. Torabi, "Antenna selection for MIMO-OFDM systems," *Elsevier Signal Process.*, vol. 88, no. 10, pp. 2431–2441, 2008.

[122] N. P. Le, F. Safaei, and L. C. Tran, "Antenna Selection Strategies for MIMO-OFDM Wireless Systems: An Energy Efficiency Perspective," *IEEE Trans. on Veh. Technol.*, vol. 65, no. 4, pp. 2048–2062, April 2016.

[123] J.-C. Chen, "Joint Antenna Selection and User Scheduling for Massive Multiuser MIMO Systems With Low-Resolution ADCs," *IEEE Trans. on Veh. Technol.*, vol. 68, no. 1, pp. 1019–1024, Nov. 2019.

[124] J. Choi and B. L. Evans, "Analysis of Ergodic Rate for Transmit Antenna Selection in Low-Resolution ADC Systems," *IEEE Trans. on Veh. Technol.*, vol. 68, no. 1, pp. 952–956, Oct. 2019.

[125] A. Gersho and R. M. Gray, *Vector quantization and signal compression.* Springer 2012 (originally published 1992).

[126] P.-H. Lin and S.-H. Tsai, "Performance analysis and algorithm designs for transmit antenna selection in linearly precoded multiuser MIMO systems," *IEEE Trans. on Veh. Technol.*, vol. 61, no. 4, pp. 1698–1708, Mar. 2012.

[127] M. Gharavi-Alkhansari and A. B. Gershman, "Fast antenna subset selection in MIMO systems," *IEEE Trans. on Signal Process.*, vol. 52, no. 2, pp. 339–347, Feb. 2004.

[128] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions–I," *Math. Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[129] J. S. Liu, *Monte Carlo strategies in scientific computing.* Springer Science & Business Media, 2008.

[130] J. Harold, G. Kushner, and Y. George, "Stochastic Approximation Algorithms and Applications," 1997.

[131] N. Prasad, X.-F. Qi, and A. Gatherer, "Optimizing beams and bits: A novel approach for massive MIMO base station design," in *IEEE Int.*

Conf. on Computing, Networking and Commun., Apr. 2019, pp. 970–976.

[132] V. Erceg, L. J. Greenstein, S. Y. Tjandra, S. R. Parkoff, A. Gupta, B. Kulic, A. A. Julius, and R. Bianchi, "An empirically based path loss model for wireless channels in suburban environments," *IEEE Journal on Sel. Areas in Commun.*, vol. 17, no. 7, pp. 1205–1211, Jul. 1999.

[133] J. Choi, B. L. Evans, and A. Gatherer, "ADC bit allocation under a power constraint for mmWave massive MIMO communication receivers," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process.*, 2017, pp. 3494–3498.

[134] J. Sung, J. Choi, and B. L. Evans, "Narrowband Channel Estimation for Hybrid Beamforming Millimeter Wave Communication Systems with One-Bit Quantization," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, 2018.

[135] B. Zhou, B. Bai, Y. Li, D. Gu, and Y. Luo, "Chordal distance-based user selection algorithm for the multiuser MIMO downlink with perfect or partial CSIT," in *IEEE Int. Conf. Advanced Inform. Networking and App.*, 2011, pp. 77–82.

[136] K. Ko and J. Lee, "Multiuser MIMO user selection based on chordal distance," *IEEE Trans. Commun.*, vol. 60, no. 3, pp. 649–654, 2012.

[137] R. Rajashekar and L. Hanzo, "Iterative matrix decomposition aided block diagonalization for mm-wave multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1372–1384, 2017.

[138] S. Park and R. W. Heath, "Spatial Channel Covariance Estimation for the Hybrid MIMO Architecture: A Compressive Sensing-Based Approach," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 12, pp. 8047–8062, 2018.

[139] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. on Antennas and Propagation*, vol. 61, no. 7, pp. 3814–3827, 2013.

[140] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.

[141] L. You, X. Gao, G. Y. Li, X.-G. Xia, and N. Ma, "BDMA for millimeter-wave/terahertz massive MIMO transmission with per-beam synchronization," *IEEE Journal on Sel. Areas in Commun.*, vol. 35, no. 7, pp. 1550–1563, 2017.

[142] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, "Codebook-based hybrid precoding for millimeter wave multiuser systems," *IEEE Trans. Signal Process*, vol. 65, no. 20, pp. 5289–5304, 2017.

[143] J. H. Conway, R. H. Hardin, and N. J. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian spaces," *Exper. Math.*, vol. 5, no. 2, pp. 139–159, 1996.

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

[145] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 120–127, 2018.

[146] S. Malkowsky, J. Vieira, L. Liu, P. Harris, K. Nieman, N. Kundargi, I. C. Wong, F. Tufvesson, V. Öwall, and O. Edfors, "The World's First Real-Time Testbed for Massive MIMO: Design, Implementation, and Validation," *IEEE Access*, vol. 5, pp. 9073–9088, 2017.

[147] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors, and F. Tufvesson, "A flexible 100-antenna testbed for massive MIMO," in *IEEE Globecom Workshops*, 2014, pp. 287–293.

[148] K. A. Hamdi, "A useful lemma for capacity analysis of fading interference channels," *IEEE Trans. Commun.*, vol. 58, no. 2, 2010.

[149] R. S. Strichartz, *The way of analysis*. Jones & Bartlett Learning, 2000.

Vita

Jinseok Choi received his B.S. in the Department of Electrical and Electronic Engineering at The Yonsei University, Seoul, Korea in 2014 and his M.S. in Electrical and Computer Engineering at The University of Texas at Austin, TX, USA, in 2016. He is currently pursuing Ph.D. degree in Electrical and Computer Engineering at The University of Texas at Austin. He is joining Wireless Networking and Communications Group as a student member and working at Embedded Signal Processing Laboratory under supervision of Professor Brian L. Evans. His primal research interest is to develop and analyze future wireless communication systems.

Permanent address: jinseokchoi89@utexas.edu

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.