Centralized Cell-Free Massive MIMO with Low-Resolution Fronthaul

Dick Maryopi

This thesis is submitted for the degree of
Doctor of Philosophy

Communication Technologies Research Group
Electronics Engineering

January 2020

Supervisor:
Prof. Alister G. Burr

Internal examiner:
Dr. Kanapathippillai Cumanan

External examiner:
Dr. Deniz Gunduz

To my parents, my wife *Zara* and my son *Yusuf*.

Abstract

The increasingly new data-hungry applications in our digital society now might no longer be handled efficiently by the current cellular networks. Cell-free massive MIMO network comes to resolve the traditional way of deploying wireless networks by blurring the cell boundaries. The network comprises a large number of access points (APs) which connect the users to a central processing unit (CPU) via fronthauls for coherent transmission and reception. It is expected that this network can provide a uniformly high data rate per user and per unit area. In this thesis, we study a centralized approach to cell-free massive MIMO that can further exploit its potential with considering a practical issue of limited-capacity fronthauls. We develop different schemes as well as strategies that make the centralized approach feasible. Thereby, we propose the use of low-resolution fronthauls and analyse its performance by making use of Bussgang theorem.

The first part of this thesis considers a cell-free network with single-antenna APs, where a coarse scalar uniform quantizer is devised as an interface to the fronthauls. In the second part of this thesis, we extend the network to the case of multi-antenna APs, where two different processing schemes at the APs are studied: individual processing and joint processing. For each part, two strategies for acquiring the channel state information (CSI) under low-resolution fronthaul constraint are developed: estimate-and-quantize (EQ) and quantize-and-estimate (QE). We analyse the performance of both strategies and take them into account for deriving the achievable rate of the systems. Moreover, the scalability of the centralized approach is also discussed in terms of fronthaul load and AP processing. In the last part, we propose the use of a lattice vector quantizer at multi-antenna APs for the high-mobility and high-density scenario, in which two procedures for constructing the lattice codebook are developed.

Table of contents

List of tables

List of figures

If I have seen further it is by standing on the shoulders of giants

-Isaac Newton

Acknowledgements

All praise be to Allah for His blessing and merciful, without which I would not have been able to complete this thesis.

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Alister G. Burr, who has given me the opportunity and guidance in pursuing a doctoral degree at the University of York. I have benefited a lot from his wide knowledge and his deep insight into the research problems. His constant support for my academic development and many stimulating meetings during the last four years have helped me very much to get to this point. Special thanks go to my thesis advisor Dr. Kanapathippillai Cumanan who has given me his valuable suggestions and invited me to fruitful discussions.

I would like to express the deepest appreciation to Indonesia Endowment Fund for Education (LPDP) for their financial support and their commitment to bring the nation forward through education. I am particularly thankful for the assistance given by Camilla and Helen in everything to do with the PhD administration. Moreover, I would like to thank all my friends at the University of York. Especially, I would like to thank Manijeh, Qinhui, Cheng, Tong, Abi, Nils, Ibrahim, Aliyu and Sunghyun for the helpful discussions and sharing the room with a friendly working atmosphere. I would also like to express my gratitude to all my friends in KIBAR-UK, especially to Bern, Hendry, Agung, Ismu, Abram, Arif, Iwan, Syaiful, Hanief, Taufik, Taufiq, Ataka and Syihan for their kind supports and for sharing their experiences about living in the UK.

Finally and most importantly, I am indebted to my families from whom I have received generous supports. I owe my deepest gratitude to my mother for all her unconditional love, care and sacrifices for the best of her sons. I am deeply grateful to my father, who has thought me many things in life. His patience, endurance, determination and sacrifice for his family have been my source of inspiration. I am particularly grateful for the support, encouragement and trust given by my father-in-law and my mother-in-law.

Special thanks also go to my brothers Yudi, Refa, Iwan, Irfan, my sister-in-law Pipit and my brother-in-law Aulia for their kind supports. My deepest gratitude goes to my wife Zara and my son Yusuf for accompanying me during my PhD journey, with its many ups and downs. Their endless love and boundless support have been my source of energy.

Dick Maryopi
January 2020

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Dick Maryopi
January 2020

Chapter 1

Introduction

Mobile wireless communication is probably the fastest technology developed in the last few decades. The fact that it has changed significantly the way we are living now, however, might not be realized by most people. The next wireless generation, 5G and beyond are also predicted to shape our society even more dramatically. The success of their deployment promises a long list of new applications ranging from seamless video telephony and 3D video streaming to real-time mobile gaming, telemedicine, self-driving car, and smart city. The latter showcases just a few applications we can think of as a product of expanding the network from merely connecting people to additionally connecting things. Under the umbrella Internet of Things (IoT), the network will embrace all possible real-world physical objects, such as sensors, traffic objects, machines, etc.

To achieve that ambition, the next wireless generation has to face a huge challenge. According to the vision of the 5G association 5GPPP, the network should facilitate very dense wireless communication links with 1000 times higher area capacity to connect over 7 trillion wireless devices serving over 7 billion people [1]. At the same time, it should save 90% of energy per service provided, and create a secure, reliable Internet. On the other hand, the radio spectrum as the main resource for wireless communication has been very crowded. As the number of connected devices dramatically increases, whereas the available spectrum is increasingly scarce, interference is more inevitable. Thus, carrying out a communication system with this specification becomes then a delicate engineering task. Despite the launching of the 5G network will be done soon in several countries by 2020, there are still many requirements that can not be fulfilled by the 5G.

In response to this challenge, researchers are now questioning whether the currently deployed network, which is based on a cellular system, is still relevant for providing such diverse applications with such stringent requirements. Historically, the motivation for deploying cellular networks is to provide wireless service to a large area. Because the

radio signal attenuates proportionally with the distance, the service area is divided into multi cells where a certain base station is dedicated to cover a given non-overlapping cell of area. In order not to interfere between cells, adjacent cells can be operated in a different spectrum. However, with the increasing number of subscribers and high-bandwidth consuming applications, the spectrum becomes scarce such that the cells are pushed to reuse the same frequency to maintain the spectral efficiency. As a consequence, this creates an inter-cell interference which can severely degrade the user performance especially for the user at the cell edges. Hence, it requires a deliberately designed technique to suppress inter-cell interference.

Apart from the success story of the cellular system, the deployment of cells as we have today seems to constrain the design flexibility. Therefore, the time to rethink the concept of the cellular system might have come as many tools have been available and many technologies have been developed. One among others is massive MIMO (Multiple Input Multiple Output). Massive MIMO offers a solution to the spectrum crunch by making use of a large number of antennas at the base station while serving many terminals in the same time-frequency resource [2, 3]. Recently, an unconventional notion has emerged to deploy massive MIMO without the restriction of cells, which is called as *cell-free* massive MIMO [4, 5]. A strong contrast to the cellular system can be seen in how the service area is defined. In cell-free massive MIMO, a user is not served only by a base station, which is dedicated to a given area, but by all base stations simultaneously in a large coverage area. The signal from other base stations, which is conventionally treated as interference, is now becoming a useful signal. Moreover, there is a massive number of distributed base stations, or also called access points (AP), who serve a smaller number of users. By deploying access points in a distributed manner, many advantages can be obtained such as better coverage and the availability of macro-diversity such that the throughput per user per area can be increased. To enable this, the access points are connected to a central processing unit (CPU) via fronthaul links.

1.1 Motivation

Due to its potential to meet high data rate demand for a large number of users with uniform coverage, cell-free massive MIMO has initiated a new research direction and has attracted much attention as indicated by the increasing number of published research papers in the past few years. The initial work on cell-free massive MIMO mostly focused on the performance analysis when the signal processing, which comprises detection in the uplink and precoding in the downlink, is performed in a distributed manner at the

access points, making use of the locally-obtained channel state information (CSI) [6]. The purpose of doing this is usually to deal with the fronthaul load and scalability issues that exist inherently in a large distributed antenna system (DAS) such as cell-free massive MIMO. To perform coherent processing at the CPU, the transmission over the fronthaul links between APs and CPU is performed in the baseband. This means that the advantages of cell-free massive MIMO must be paid by stringent fronthaul requirements. Due to the large number of APs, there is then a concern that the CSI signalling over the fronthaul links becomes unscalable. That is, the fronthaul load will increase excessively with the number of served users. Thus, when the signal processing is performed at the APs, it can presumably avoid the CSI signalling over the fronthaul. While this may be true, the approach is indeed not scalable in terms of data signalling, since the processing at the AP increases the number of data signals to be transmitted over the fonthaul links proportionally to the number of users.

In contrast, we study in this thesis a centralized approach to cell-free massive MIMO. Here, we refer to a scheme as centralized when the CSI is available at the CPU and joint processing is performed at the CPU. The benefits of the centralized approach over the distributed approach are twofold. The first and perhaps the most important thing is that the centralized approach can deliver significantly higher spectral efficiency. This is shown in [7–10] for various types of joint processing at the CPU. A rather comprehensive comparison has been provided independently in [11] where the authors sort cell-free massive MIMO schemes based on the degree of cooperation among APs. From the perspective of the application, the ability to support a large number of users and at the same time to deliver a high spectral efficiency can be crucial for the new IoT scenario. For instance, in [8, 12] it is mentioned that the connected vehicle in autonomous driving requires very high data rates in the uplink to transmit the surrounding information generated by the many sensors on it. The second benefit of the centralized approach is that it is more scalable in terms of data signalling. Indeed, at a particular operating point, the overall fronthaul load is much lower than the distributed approach as investigated in [11]. Moreover, a low-complexity strategy can also be applied to reduce the load of the CSI signalling as shown in [10].

Another issue that is often overlooked in the study of the performance of cell-free massive MIMO is the fact that the fronthaul links are in practice not perfect. Instead, the fronthaul links must be subject to a limited capacity. It is well known from the information theory literature that error-free transmission can be achieved at the channel output when the rate of the channel input is less than the channel capacity. To that end, in the uplink scenario, we should represent the received signal at the APs at maximum

with the rate of the fronthaul capacity. In this case, we represent it as a bitstream and compress at a rate below than the fronthaul capacity. Hence, an analog-digital converter (ADC) must be utilized at the APs as an interface to the fronthaul. Regarding the scalability issue, the unscalable scheme may then lead to a high data rate that requires high precision ADCs. Furthermore, cell-free massive MIMO is also attractive to support wireless transmission at millimeter-wave due to its ability to exploit macro-diversity. This might be useful to reduce the outage probability due to blockages and shadowing to which the millimeter-wave transmission is sensitive. Obviously, operating at millimeter-wave can provide us a huge channel bandwidth which allows us to transmit a large amount of data, but in turn, can exhaust the utilization of ADC. As a consequence, the AP becomes power-hungry. This is due to the high-resolution quantizer inside the ADC whose power consumption increases exponentially with the number of quantization bits [13]. In this situation, utilizing low-resolution quantization might be a feasible option. This is in line with the prospect of 6G [14], where low-resolution technology is expected to be an enabler to support the processing of high data rates with low power consumption.

In that respect, it seems important to study a scheme for cell-free massive MIMO that is scalable and more friendly to the fronthaul. Although much work has been done in this area [15–25], cell-free massive MIMO is still in its infancy, where further research needs to be carried out, particularly when the assumption of limited fronthaul capacity is applied. This thesis aims to fill the remaining gap by proposing a centralized approach to cell-free massive MIMO with respect to low-resolution fronthaul.

1.2 Contributions

In this thesis, we study the concept of centralized cell-free massive MIMO to further extent. In this case, we develop some schemes as well as strategies that make the centralized approach feasible under the practical constraint of limited-capacity fronthaul links. The fact that a large number of APs is deployed in cell-free-massive MIMO, we propose the use of low-resolution fronthauls and analyse its performance by making use of Bussgang theorem. Specifically, the contributions are:

- For single-antenna APs, we devise a low-resolution scalar uniform quantizer as an interface to the fronthaul, where we model the quantizer using Bussgang decomposition and characterize its optimum step size.

- Using the low-resolution scalar uniform quantizer, we develop two strategies for acquiring the CSI at the CPU. The first is the estimate-and-quantize (EQ) strategy, and the second is the more scalable quantize-and-estimate (QE) strategy.
- We analyse the performance of both strategies in terms of the mean squared error (MSE). Through numerical simulation we verify our analysis and asses the performance of our developed strategies.
- Taking into account the quantization distortion that affects the CSI acquisition and the data transmission, we derive the achievable rate per-user for the case of zero-forcing (ZF) detection performed at the CPU. We provide also a simpler SINR expression for ZF detection.
- We examine also the scalability of the centralized cell-free massive MIMO with single-antenna APs.
- In the case of multi-antenna APs, we study two different processing schemes at the APs for low-resolution fronthauls. We investigate in the first scheme the case, where the received signals across the multiple antennas are treated individually, and a scalar uniform quantizer with low resolution is used at each antenna. We develop the EQ and QE strategies for acquiring the CSI for this scheme and derive the achievable rate.
- Considering that the resulting fronthaul load from the multi-antenna AP increases, we propose in the second scheme joint processing of the signals received across the multi antennas. With low bit-rate per dimension, we utilize vector quantizer, which is modelled by Bussgang decomposition. We develop the EQ and QE strategies for acquiring the CSI for this scheme and derive the achievable rate.
- Similarly, we examine the scalability of the centralized cell-free massive MIMO with multi-antenna APs.
- We investigate further the joint processing scheme at multi-antenna APs for the high-mobility and high-density scenario, for which we suggest the use of a lattice vector quantizer. For this purpose, we study the lattice quantizer design problem with a fast processing requirement.
- We address the lattice codebook design problem. To deal with this, we adopt the geometrical shaping approach. In this case, we propose two procedures for constructing a lattice codebook devoted to uncorrelated and correlated channel

scenario. We do a modification to Conway and Sloane algorithm and introduce a scale factor that brings the codebook to a near-optimal performance.

- To the codebook with a correlated input signal, we propose the use of ellipsoidal Voronoi shaping and the use of Karhunen–Loève transform in the quantization process.

Parts of this thesis have been published in a journal article and in conference and workshop proceedings as listed below.

Journal Papers

- D. Maryopi, M. Bashar, and A. Burr, "On the Uplink Throughput of Zero Forcing in Cell-Free Massive MIMO With Coarse Quantization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7220–7224, Jul. 2019, ISSN: 0018-9545. DOI: 10.1109/TVT.2019.2920070

Conference Papers

- A. Burr, M. Bashar, and D. Maryopi, "Cooperative Access Networks: Optimum Fronthaul Quantization in Distributed Massive MIMO and Cloud RAN," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, Jun. 2018, pp. 1–5. DOI: 10.1109/VTCSpring.2018.8417560

- M. Bashar, H. Q. Ngo, A. G. Burr, D. Maryopi, K. Cumanan, and E. G. Larsson, "On the Performance of Backhaul Constrained Cell-Free Massive MIMO with Linear Receivers," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct. 2018, pp. 624–628. DOI: 10.1109/ACSSC.2018.8645433

- A. Burr, M. Bashar, and D. Maryopi, "Ultra-Dense Radio Access Networks for Smart Cities: Cloud-RAN, Fog-RAN and "Cell-Free" massive MIMO," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Workshop CorNer*, Bologna, Italy, Sep. 2018

- D. Maryopi and A. G. Burr, "Few-Bit CSI Acquisition for Centralized Cell-Free Massive MIMO with Spatial Correlation," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Marrakech, Morocco, Apr. 2019

- M. Bashar, A. G. Burr, D. Maryopi, K. Haneda, and K. Cumanan, "Robust Geometry-Based User Scheduling for Large MIMO Systems Under Realistic Channel Conditions," in *European Wireless 2018; 24th European Wireless Conference*, May 2018

- A. Burr and D. Maryopi, "On the Modelling of Coarse Vector Quantization in Distributed Massive MIMO," *IEEE Statistical Signal Processing Workshop*, 2021, Submitted

1.3 Outline of the Thesis

The thesis is organized as follows. **Chapter** 2 provides background to the subsequent chapters. In the first part of this chapter, we will first introduce the basics of massive MIMO, where we will consider two types of deployment, namely multicell and cell-free massive MIMO. The implication of having a large number of antennas at the base station is discussed and the performance of both deployments is given. In the second part of this chapter, we will present the basic concept of quantization, where we will give an overview of scalar quantization and vector quantization. Further, we will also describe two tools that can be used to model a quantization process based on the assumption of the required resolution.

In **Chapter** 3, a framework for designing a centralized cell-free massive MIMO with capacity-limited fronthaul is introduced in the case where only a single antenna is available at the APs. We begin the discussion in this chapter firstly by defining centralized cell-free massive MIMO and explaining its general concept. We will then give the scope of the discussion by describing the considered system model. To deal with a limited-capacity fronthaul, we wish to design an optimum quantizer for the fronthaul. This will be presented subsequently for a scalar uniform quantizer with low resolution. We will then look at an important part of centralized cell-free massive MIMO which is the CSI acquisition. This is then followed by the scheme for data transmission considering the limited-capacity fronthaul. Further, the achievable rates of this approach are given where we will also derive a simpler SINR expression. Then, a scalability issue of cell-free massive MIMO will also be discussed. We will see that the centralized approach can resolve some aspects of this issue. Numerical results for validating our analysis and evaluating the performance of our proposed scheme are then given afterwards.

Chapter 4 will extend the previous chapter to multi-antenna APs, and consider two schemes of AP processing. In the first scheme, we process the received signal across the multi-antenna AP separately, where we use independent scalar quantization at each antenna similar to the work in [9]. Since the resulting fronthaul load from the multi-antenna AP is larger than for a single-antenna AP, we propose in the second scheme to jointly process the received signal at the APs. By exploiting the fact that the channel across the AP's antenna is correlated, we employ Vector Quantization (VQ) with a small

number of bits per dimension. After describing the quantization model, we will also describe how the CSI is acquired in the joint processing scheme and determine the data rate achievable with this scheme. Before we evaluate its system performance, we will discuss the scalability and the possibility to extend the centralized cell-free massive MIMO with multi-antenna APs into a larger network. Then, we will show using simulations that the joint processing can deliver higher data rate than the individual processing counterpart at the same fronthaul resolution. In the opposite situation, we can set the joint processing at the same throughput performance as the individual processing, which compensates with a lower requirement of the fronthaul load.

In **Chapter** 5 we will look at lattice vector quantization, which is intended to be applied at the multi-antenna APs to jointly quantize the received signals following the scheme in the previous chapter. Therefore, the focus of this chapter is to devise a codebook for the above-mentioned application. In this case, we consider constructing a codebook from lattice points. Although the constructed codebook is suboptimal, as we will see later, by using a lattice we aim to find a good trade-off between performance and complexity. After giving a brief introduction to lattices, we will describe the codebook design problem using a lattice. Then, we propose two procedures for constructing a lattice codebook which is fast and near-optimal. These are intended for uncorrelated and correlated channel applications, respectively. Finally, we give some numerical results which evaluate the performance of our proposed procedures.

Finally, in **Chapter** 6, we summarize the main results of this thesis and provide some potential research in the future.

1.4 Notation

In this section we introduce some essential notation we use in this thesis. Any specific notation will be described when it is appeared for the first time. Roman letter, lower-case boldface letter and upper-case boldface letter are used respectively to denote a scalar, a column vector and a matrix. The set of all complex and real $M \times N$ matrix are represented by $\mathbb{R}^{M\times N}$ and $\mathbb{C}^{M\times N}$ respectively. By $\langle \cdot, \cdot \rangle$ we denote the inner product with $\| \cdot \|$ as its corresponding vector norm or Frobenius norm. The expectation of a random variable is represented by $\mathbb{E}\{\cdot\}$. We denote circularly complex Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$ by $\mathcal{CN}(\mathbf{m}, \mathbf{\Sigma})$. We use $\mathbf{I}_N$ for the $N \times N$ identity matrix and $\mathbf{1}_N$ for the all-one vector of dimension N. We denote the complex conjugate by the superscript $(\cdot)^*$ and the transpose conjugate by $(\cdot)^H$. For a vector $\mathbf{a}$, $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix with the diagonal elements created from vector $\mathbf{a}$.

Chapter 2

Preliminaries

This chapter is intended as a preview and background to the subsequent chapters. In the first part of this chapter, we will first introduce the basics of massive MIMO, where we will consider two types of deployment namely multicell and cell-free massive MIMO. The implication of having a large number of antennas at the base station is discussed and the performance of both deployments is given. In the second part of this chapter, we will present the basic concept of quantization, where we will give an overview of scalar quantization and vector quantization. Further, we will also describe two tools that can be used to model a quantization process based on the assumption of the required resolution.

2.1 Massive MIMO

The availability of many antennas in Multiple Input Multiple Output (MIMO) systems can be utilized in different ways. By sending multiple data streams through many independent paths we can exploit the so called multiplexing gain such that the data rate is increased. On the other hand, by sending the same data through different paths we exploit the diversity to make the transmission more reliable. In the multi-user scenario such as a cellular system, the many users can be treated as a collection of antennas that form a MIMO system together with the antenna array at the base station. It allows either many users to be served simultaneously or a good propagation channel to be allocated to each user.

Massive MIMO is a form of multi-user MIMO system. The most distinguishing thing between massive MIMO and general multi user MIMO is the number of antennas which serves the user. In massive MIMO the number of antennas involved is very large and much greater than the number of user antennas. As the number of antennas increases the propagation channel begins to enjoy the *channel hardening* property which means that

the singular value distribution of the propagation matrix becomes more deterministic [3]. As explained in [3] it is desired to have all singular values equal to possibly obtain parallel independent links such that the upper bound of the capacity can be reached. Furthermore, scaling up the number of antennas at the base station for a fixed number of users makes the propagation more *favorable* in the sense that the channels tend to be nearly orthogonal. Those properties lead then to the practical benefit of massive MIMO such as increasing the spectral efficiency, optimal use of linear processing etc.

To gain insight to how massive MIMO works, let us consider a multi user MIMO system consisting of a base station equipped with M antennas and having K users which are connected by the propagation channel $\mathbf{G} \in \mathbb{C}^{M \times K}$. For reliable communication the knowledge of channel state information (CSI) is important to compensate the effect of the channel. If we let the number of antennas M and the number of users K grow very large, then the matrix $\mathbf{G}$ that should be estimated will also grow proportionally. As long as we can provide the required CSI it seems that we can serve an unlimited number of users by increasing the number of antennas at the base station. However, CSI acquisition in large dimension is not trivial and is one of the main limiting factors.

The standard way of acquiring CSI is by sending training pilots, and then the channel will be estimated from the received pilots. To get a good quality of CSI a certain number of pilots is required, depending on the propagation environment. In Frequency Division Duplex (FDD) the uplink and downlink transmission is separated by frequency, and therefore the channel is different between uplink and downlink. The strategy usually used in this protocol uses a feedback link. In the case of the downlink, where CSI is needed for precoding, the base station sends pilots at each antenna at intervals related to the coherence interval, which is the time-frequency block with approximately static fading. The users estimate the channel and send the estimated channel back after the uplink training pilots. When the number of antennas grows very large, this strategy becomes inefficient because the number of pilots scales with the number of antennas and thus many resources are used only for pilots in the feedback channel.

To overcome that problem, it is more efficient to operate massive MIMO in Time Division Duplex (TDD), where uplink and downlink take place in succession. Because they operate in the same frequency resource, they have the same frequency response over the available coherence interval. In this situation we can use the estimated channel in the uplink also for the downlink due to the channel reciprocity. The base station can use the conjugate transpose of matrix $\mathbf{G}$ for doing precoding. Thus, the costly channel feedback can be saved. On the other hand, there have been some attempts to realize massive MIMO in FDD such as in [31, 32]. Most of the motivation for it is the availability

of many frequency bands, which so far have been dedicated for FDD by the previous technologies. The FDD protocol is also preferred in the situation where the calibration of the antenna front end between the transmitter and receiver is difficult to perform in order to maintain the channel reciprocity [33]. However, TDD is still more efficient than FDD and generally better in terms of the performance for utilizing a large number of antennas at the base stations [34, 35].

The main features of the massive MIMO scheme in the initial work of Marzetta [2] are the TDD protocol and the use of large numbers of antennas at non-cooperating base stations. In the following sections we keep those features and outline the scheme in more detail based on its system architecture. A comprehensive discussion of massive MIMO can be found in [33] and in [36] for the context of 5G .

2.1.1 Multicell Massive MIMO

We discuss first in this section a multicell massive MIMO system with co-located antenna architecture. Suppose that we have a system of L cells each of which has one base station with M co-located antennas serving K single-antenna users. The propagation channel between the k-th user in the l-th cell and the m-th base station antenna in the j-th cell can be modelled by

$$\begin{aligned} g_{jlmk} &= h_{jlmk}\beta_{jlk}^{1/2}, \text{ where} \\ j &= 1, ..., L, \quad l = 1, ..., L, \\ m &= 1, ..., M, \quad k = 1, ..., K. \end{aligned} \tag{2.1}$$

The coefficient h_{jlmk} represents the small scale fading between the k-th user and the m-th base station antenna of the corresponding l-th and j-th cell. We assume this coefficient to be uncorrelated and complex Gaussian distributed with zero mean and unit variance. The large scale fading is denoted by β_{jlk}, which includes the path loss attenuation and shadowing. Due to the co-located configuration, every single antenna at the base station of the j-th cell perceives the same large scale fading from the k-th user in the l-th cell. Thus, we can conveniently write the channel for all K users in particular cell l to the base station of cell j as a matrix given by

$$\mathbf{G}_{jl} = \mathbf{H}_{jl}\mathbf{D}_{jl}^{1/2}, \tag{2.2}$$

where the coefficient g_{jlmk} and h_{jlmk} respectively compose the m-th row and k-th column of matrix $\mathbf{G}_{jl}, \mathbf{H}_{jl} \in \mathbb{C}^{M\times K}$. Moreover, we denote the k-th column of matrix $\mathbf{G}_{jl}$ by

$\mathbf{g}_{jlk}$. Here, the matrix $\mathbf{D}_{jl} = \text{diag}(\beta_{jlk}) \in \mathbb{C}^{K\times K}$ is a diagonal matrix where it has all zeros in the off diagonal positions and has large scale fading coefficient β_{jlk} as its diagonal elements. Further, we would like our channel to posses the channel hardening and favorable propagation properties. They are respectively fulfilled when the following conditions

$$\frac{\|\mathbf{g}_{jjk}\|^2}{\mathbb{E}\{\|\mathbf{g}_{jjk}\|^2\}} \to 1 \quad \text{and} \quad \frac{\mathbf{g}_{jlm}^H \mathbf{g}_{jjk}}{\sqrt{\mathbb{E}\{\|\mathbf{g}_{jlm}\|^2\}\mathbb{E}\{\|\mathbf{g}_{jjk}\|^2\}}} \to 0 \tag{2.3}$$

hold almost surely as $M \to \infty$ [33]. To achieve those properties with high probability we set therefore $M \gg K$. This setting will also imply

$$\begin{aligned} \frac{\mathbf{G}_{jl}^H \mathbf{G}_{jl}}{M} &= \mathbf{D}_{jl}^{1/2} \left(\frac{\mathbf{H}_{jl}^H \mathbf{H}_{jl}}{M} \right) \mathbf{D}_{jl}^{1/2} \\ &\approx \mathbf{D}_{jl} \end{aligned} \tag{2.4}$$

Pilot Transmission

Prior to receiving or transmitting data every user sends a pilot sequence intended for the base station to estimate the channel. We assign each pilot sequence according to the time frequency resource, which has a consequence that the number of orthogonal pilot sequences is limited by the available coherence time τ_c and the coherence bandwidth ω_c of the channel. We assume the frequency reuse factor one is applied and the transmission in each cell is synchronously performed such that the same time-frequency resource is shared among all K users in L cells. It follows that the same set of pilot sequences should be reused in each cell.

Suppose that the j-th cell is the cell under observation. The base station in the j-th cell wants to estimate the channel only from the K users in the j-th cell denoted by $\mathbf{G}_{jj}$. Each user in the j-th cell sends its complex valued pilot sequence $\boldsymbol{\varphi}_k \in \mathbb{C}^{\tau_p \times 1}, k = 1, ..., K$, where it has the length $\tau_p < \tau_c$ samples and unit energy $\|\boldsymbol{\varphi}_k\|^2 = 1$. This user-specific pilot is taken from a set of orthogonal sequence $\mathcal{P}_\varphi = \{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_K\}$. At the same time, the other K users in the $L-1$ cells send their pilots, and hence the received pilot signal at the j-th base station is

$$\mathbf{Y}_{p,j} = \sqrt{\rho_p} \sum_{l=1}^{L} \mathbf{G}_{jl} \boldsymbol{\Theta}_l + \mathbf{W}_j, \tag{2.5}$$

where $\mathbf{Y}_{p,j}$ has the dimension $M \times \tau_c$ and the matrix $\boldsymbol{\Theta}_l = [\boldsymbol{\varphi}_{1l}, \ldots, \boldsymbol{\varphi}_{Kl}]^T \in \mathbb{C}^{K\times \tau_c}$ is composed of pilot sequences from the l-th cell. The notation ρ_p denotes the normalized

transmit power of the pilot which is obtained from dividing the pilot transmit power by the noise power at the receiver. The matrix $\mathbf{W}_j \in \mathbb{C}^{M \times \tau_c}$ denotes the additive noise matrix whose entries are independent complex normal distributed as $\sim \mathcal{CN}(0,1)$. At the base station j the received pilot signal is projected onto its own orthogonal pilot sequences $\mathbf{\Theta}_j$ resulting

$$\mathbf{Y}_{p,j}\mathbf{\Theta}_j^* = \sqrt{\rho_p}\sum_{l=1}^{L} G_{jl}\mathbf{\Theta}_l\mathbf{\Theta}_j^* + \mathbf{W}_j\mathbf{\Theta}_j^* \tag{2.6}$$

The channel matrix from the user in the cell j is then estimated by least squares as

$$\hat{\mathbf{G}}_{jj} = \sqrt{\rho_p}\mathbf{G}_{jj} + \sqrt{\rho_p}\sum_{\substack{l=1 \\ l\neq j}}^{L} \mathbf{G}_{jl}\mathbf{\Theta}_l\mathbf{\Theta}_j^* + \mathbf{W}_j\mathbf{\Theta}_j^* \tag{2.7}$$

It can be seen that there exists interference during the pilot transmission, called pilot contamination, expressed by the second term in equation (2.7). In worst case, where the pilot sequences from all L cells are taken from the same set $\mathcal{P}_\varphi$ and possibly non-orthogonal, we obtain

$$\hat{\mathbf{G}}_{jj} = \sqrt{\rho_p}\sum_{l=1}^{L} \mathbf{G}_{jl} + \mathbf{W}_j\mathbf{\Theta}_j^*. \tag{2.8}$$

Because the distance of the K users in the j-th cell is distributed nearer to their base station, the portion of $\mathbf{G}_{jj}$ in equation (2.8) is larger than $\mathbf{G}_{jl}\forall l \neq j$. Nevertheless, it degrades the estimated channel significantly which in turn limits the performance of massive MIMO more than any other kind of impairment. Due to the greater number of users allowed in massive MIMO, pilot contamination also has a larger impact in massive MIMO than in conventional multiuser MIMO. Some solutions to deal with this problem have been studied. One of the most established way is by implementing pilot reuse and power control in the pilot transmission [37].

Uplink Data Transmission

Multicarrier modulation as well as single carrier modulation can be employed in massive MIMO. We consider here the single carrier modulation, but the extension to the multicarrier case should be straightforward without many changes. In the uplink phase, for each channel use, the k-th user in the l-th cell is modeled to send an independent identically distributed (i.i.d.) random message $q_{kl} \in \mathbb{N}$ which is obtained from some finite alphabet $\mathcal{M}$. The message is mapped through modulation onto the complex plane generating the

transmit symbol $x_{kl} \sim \mathcal{CN}(0,1)$ which is assumed to be complex Gaussian distributed to attain the maximum entropy. At the base station of the j-th cell the data-bearing signal $\mathbf{y}_j$ is received as a superposition of x_{kl} described by

$$\mathbf{y}_j = \sqrt{\rho_u} \sum_{l=1}^{L} \mathbf{G}_{jl} \mathbf{x}_l + \mathbf{v}_j, \tag{2.9}$$

where $\mathbf{v_j} \sim \mathcal{CN}(0, \mathbb{I}_K)$ is the received noise vector. We also assume that the transmit symbols satisfy $\mathbb{E}\{|x_{kl}|^2\} = 1$ over the codebook, such that the notation ρ_u represents the normalized transmit power for the uplink data. Since we normalize the transmit power by the noise power at the receiver, the quantity ρ_u can also be seen as the normalized transmit SNR. Further, the received symbol at each antenna and the transmit symbol of each user can be arranged in column vectors respectively given by

$$\mathbf{y_j} = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{Mj} \end{pmatrix}, \quad \mathbf{x_l} = \begin{pmatrix} x_{1l} \\ \vdots \\ x_{Kl} \end{pmatrix}.$$

Due to favorable propagation a linear detector can be used by the base station j to recover each symbol x_{kj} from the received signal $\mathbf{y}_j$ with relatively good performance. It is done by multiplying the received signal $\mathbf{y}_j$ by the detector matrix $\mathbf{A}^H$

$$\begin{aligned} \mathbf{r_j} &= \mathbf{A}^H \mathbf{y}_j && (2.10) \\ &= \sqrt{\rho_u} \sum_{l=1}^{L} \mathbf{A}^H \mathbf{G}_{jl} \mathbf{x}_l + \mathbf{A}^H \mathbf{v_j} && (2.11) \\ &= \sqrt{\rho_u} \mathbf{A}^H \mathbf{G}_{jj} \mathbf{x}_j + \sqrt{\rho_u} \sum_{\substack{l=1 \\ l \neq j}}^{L} \mathbf{A}^H \mathbf{G}_{jl} \mathbf{x}_l + \mathbf{A}^H \mathbf{v}_j, \end{aligned}$$

where $\mathbf{A} \in \mathbb{C}^{M \times K}$ is chosen according to the optimized system parameter. The received symbol r_{kj} for particular user k, which is the k-th element of $\mathbf{r_j}$, can be derived further as

$$r_{kj} = \sqrt{\rho_u} \mathbf{a}_k^H \mathbf{g}_{jjk} x_{kj} + \sqrt{\rho_u} \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbf{a}_k^H \mathbf{g}_{jji} x_{ij} + \sqrt{\rho_u} \sum_{\substack{l=1 \\ l \neq j}}^{L} \sum_{i=1}^{K} \mathbf{a}_k^H \mathbf{g}_{jli} x_{ij} + \mathbf{a}_k^H \mathbf{v_j}, \tag{2.12}$$

where $\mathbf{a}_k$ and $\mathbf{g}_{jjk}$ are the k-th column of matrix $\mathbf{A}$ and $\mathbf{G}_{jj}$ respectively. In equation (2.12), it can be seen that the first term expresses the desired symbol, whereas the second

and the third term express respectively the intracell and intercell interference. It follows that the Signal to Interference Noise Ratio (SINR) of user k can be expressed by

$$\mathrm{SINR}_{u,k} = \frac{\rho_u |\mathbf{a}_k^H \mathbf{g}_{jjk}|^2}{\rho_u \sum_{l=1}^{L} \sum_{i=1}^{K} |\mathbf{a}_k^H \mathbf{g}_{jli}|^2 - \rho_u |\mathbf{a}_k^H \mathbf{g}_{jjk}|^2 + \|\mathbf{a}_k\|^2}. \tag{2.13}$$

Then, the achievable rate for the k-th user in uplink can be expressed in term of SINR_k as [2]

$$R_{u,k} = \gamma^{UL}(1 - \frac{\tau_p}{\tau_c})\log_2(1 + \mathrm{SINR}_k), \tag{2.14}$$

where γ^{UL} is a coefficient specifying the portion of uplink data.

The achievable rate given in (2.14) implicitly depends on the detector scheme or the choice of detector matrix A. Among the most studied linear detectors are Maximum Ratio Combining (MRC), Zero Forcing (ZF) and Minimum Mean squared Error (MMSE), where each of them is motivated by a different approach. In MRC the aim is to maximize the SNR of each user, so that the column vector of $\mathbf{A}$ is chosen as

$$\mathbf{a}_{\mathrm{MRC},k} = \underset{\mathbf{a}_k \in \mathbb{C}^{M\times 1}}{\arg\max} \quad \frac{\rho_u |\mathbf{a}_k^H \mathbf{g}_{jjk}|^2}{\|\mathbf{a}_k\|^2}, \tag{2.15}$$

where the maximum is reached when $\mathbf{a}_k = \mathbf{g}_{jjk}$. On the other hand, the design goal of the ZF detector is suppressing the undesired signal from other users. Specifically, we look for a detector $\mathbf{a}_{\mathrm{ZF},k}$ which satisfies the following equation

$$\mathbf{g}_{jjk}^H \mathbf{a}_{\mathrm{ZF},k} = 1 \quad \text{and} \quad \mathbf{g}_{jjk'}^H \mathbf{a}_{\mathrm{ZF},k} = 0 \quad \forall k' \neq k, \tag{2.16}$$

or equally satisfies

$$\mathbf{G}_{jj}^H \mathbf{A}_{\mathrm{ZF},k} = \mathbf{I}_K, \tag{2.17}$$

where $\mathbf{I}_k$ is an identity matrix of size K. The condition is well known to be fulfilled by pseudo inverse matrix of $\mathbf{G}^H$.

In contrast to MRC and ZF, MMSE makes an effort to take care of SNR and interference at the same time by minimizing the mean squared error between the received and the transmitted symbol written as

$$\mathbf{a}_{\mathrm{MMSE},k} = \underset{\mathbf{a}_k \in \mathbb{C}^{M\times 1}}{\arg\min}\, \mathbb{E}\left\{ |\mathbf{a}_k^H \mathbf{y}_j - x_{kj}|^2 \right\}. \tag{2.18}$$

It is also well known that the solution of MMSE is a linear *Wiener filter*. In brief, assuming a known CSI at receiver the linear detector matrices are generally given by [38]

$$\mathbf{A} = \begin{cases} \mathbf{G} & \text{MRC} \\ \mathbf{G}(\mathbf{G}^H\mathbf{G})^{-1} & \text{ZF} \\ \mathbf{G}(\mathbf{G}^H\mathbf{G} + \frac{1}{\rho_u}\mathbf{I}_k)^{-1} & \text{MMSE} \end{cases} \tag{2.19}$$

Downlink Data Transmission

In the downlink we consider the opposite direction of transmission where the base station in the j-th cell sends specific data to all K users in the j-th cell. As in the uplink we consider the same model of single carrier signal with $x_{kj} \in \mathbb{C}$ as the transmit symbol intended for the user k in the j-th cell. The data-bearing signal received by the k-th user in the j-th cell can then be described as

$$\mathbf{y_j} = \sqrt{\rho_d} \sum_{l=1}^{L} G_{jl}^T \mathbf{A}\mathbf{x_l} + \mathbf{v_j}, \tag{2.20}$$

where $\mathbf{x_l}$ is the transmit vector containing the transmit symbol of all K users in the l-th cell and $\mathbf{A}$ is the precoding matrix. We hold the assumption that the base station knows the channel from the estimate of the uplink training pilot. Further, we make use of channel reciprocity, where due to the channel hardening the statistics of the actual channel in the downlink is equal to the statistics of the estimated channel in uplink. As a consequence, we can apply the same matrix in (2.19) for precoding and the performance is close to the case of known CSI at users. As in the uplink, we can derive from (2.20) the SINR in the downlink as

$$\text{SINR}_{d,k} = \frac{\rho_d |\mathbf{a}_k^H \mathbf{g}_{jjk}|^2}{\rho_d \sum\limits_{l=1}^{L} \sum\limits_{i=1}^{K} |\mathbf{a}_k^H \mathbf{g}_{jli}|^2 - \rho_d |\mathbf{a}_k^H \mathbf{g}_{jjk}|^2 + 1} \tag{2.21}$$

and the achievable rate as

$$R_{d,k} = \gamma^{DL} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2(1 + \text{SINR}_k), \tag{2.22}$$

with $\gamma^{DL} = 1 - \gamma^{UL}$ denoting the remaining portion of the payload in downlink.

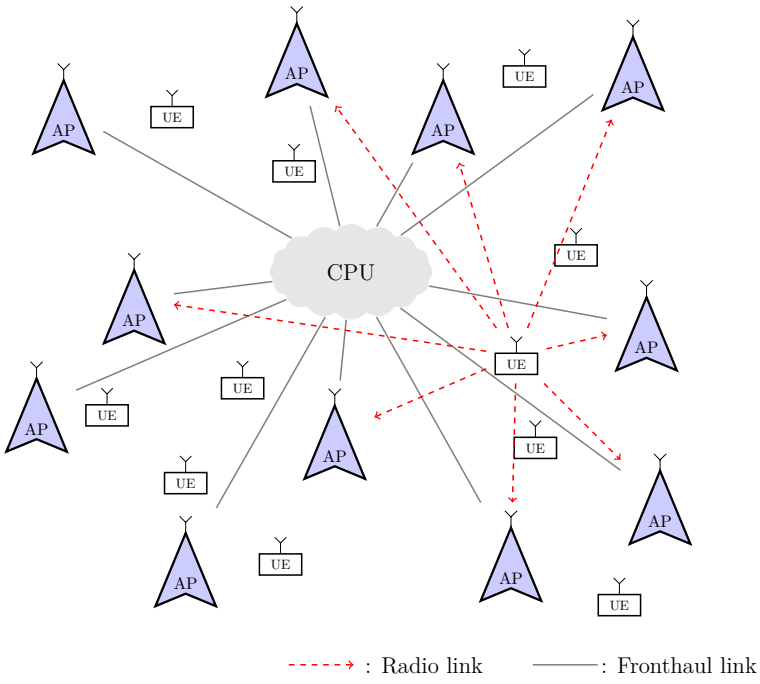


Fig. 2.1 A Cell-Free Massive MIMO networks with M access points and K users.

2.1.2 Cell-Free Massive MIMO

We observe now a deployment where the massive antennas at each base station from the multicell networks are spread over a wide area. In this case, we have a distributed antenna system (DAS) with a very large number of service antennas or access points (APs) serving a smaller number of users. However, in contrast to multicell massive MIMO the service area in the networks is not divided into cells. Since there are no cell boundaries and a large number of APs is used, this network is called cell-free massive MIMO. In this network, a user is not limited to be served by a dedicated AP in a given area, but a user can be served jointly by many distributed APs which are connected to central processing unit (CPU) via fronthaul links. This is illustrated in Fig. 2.1.

In this section we introduce first the simple model of cell free-massive MIMO according to [6], where the APs and the UEs are each equipped with a single antenna and conjugate beamforming is used for the detection and precoding at APs. As in the multicell setup, the channel between the k-th user and the m-th AP is specified by

$$g_{mk} = h_{mk}\beta_{mk}^{1/2}, \tag{2.23}$$

where the coefficient h_{mk} models the small scale fading between the k-th user and the m-th AP with the assumption to be i.i.d. $\sim \mathcal{CN}(0,1)$. However, the i.i.d. assumption

of the small scale fading here is stronger than in multicell massive MIMO, because the distributed configuration in a large area makes the distances among APs probably to be far apart so that they are scattered differently. Furthermore, due to the distance between APs, the average channel gain from the APs to the users, or the large scale fading coefficient denoted by β_{mk}, is likely to be uncorrelated for each user k and each AP m. This so called macro diversity makes the transmission more robust to blockages and shadowing effects which allows all users to be served with adequate signal strength. As opposed to multicell networks all users are closer to the APs such that a uniformly good coverage can be assured with high probability.

In consequence of the distributed configuration the overall channel between UEs and APs can only then be written in matrix notation as

$$\mathbf{G} = \mathbf{H} \odot \mathbf{D}^{1/2}, \tag{2.24}$$

where $\odot$ denotes the Hadamard product or element-wise product. The components of $\mathbf{H}, \mathbf{D} \in \mathbb{C}^{M \times K}$ are respectively the small scale and large scale fading coefficient for each m and k. If we have more than one antenna, suppose N at each AP, then every N rows of the matrix D have the same components. Nevertheless, we consider in this section the case of single-antenna APs and UEs. The discussion with multi-antenna APs is deferred until we come to Chapter 4.

Pilot Transmission

As in multicell massive MIMO we follow the same procedure in the cell-free configuration, where each user sends a specific pilot sequence to acquire the channel before transmitting data. We also utilize pilot sequences taken from an orthonormal set

$$\mathcal{P}_\varphi = \{\boldsymbol{\varphi}_k \in \mathbb{C}^{\tau_p \times 1} : \langle \boldsymbol{\varphi}_k, \boldsymbol{\varphi}_l \rangle = \delta_{kl}, \|\boldsymbol{\varphi}_k\|^2 = 1, k = 1, ..., K\}, \tag{2.25}$$

where they should have a unit energy and have an inner product equal to one if and only if the user index $k = l$ otherwise zero. The set $\mathcal{P}_\varphi$ for instance can be obtained from a unitary matrix computed by the singular value decomposition of a complex random symmetric matrix. In this case, the pilot length τ_p specifies the number of user K that can use an orthogonal pilot sequence. This can be a problem for a cell-free system due to the large number of users that should be served simultaneously in a large area. When $K > \tau_p$, it can happen that different users send the same pilot and causes a sort of pilot contamination. To deal with it, a greedy pilot assignment method has been proposed in [6], which seems to work well.

At the m-th AP the received pilot $\mathbf{y}_m$ from all K users is observed as

$$\mathbf{y}_m = \sqrt{\rho_p} \sum_{k=1}^{K} g_{mk} \boldsymbol{\varphi}_k + \mathbf{w}, \tag{2.26}$$

where ρ_p is the normalized transmit power of the pilot and $\mathbf{w}$ is the noise added at receiver $\sim \mathcal{CN}(0, \mathbf{1})$. To obtain the channel coefficient for user k the received pilot is then projected onto the pilot sequence $\boldsymbol{\varphi}_k^H$ resulting in

$$\boldsymbol{\varphi}_k^H \mathbf{y}_m = \sqrt{\rho_p} \sum_{k=1}^{K} g_{mk} \boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_k + \boldsymbol{\varphi}_k^H \mathbf{w}, \tag{2.27}$$

and estimated by least squares as

$$\hat{g}_{mk} = \sqrt{\rho_p} g_{mk} + \sum_{k \neq k'}^{K} g_{mk'} \boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_k' + \boldsymbol{\varphi}_k^H \mathbf{w}. \tag{2.28}$$

Under the assumption that there is a sufficient number of orthogonal pilot sequences available for K users, the second term will vanish.

To get another perspective we can express all received pilots without additive noise from M APs as

$$\mathbf{Y} = \mathbf{G} \boldsymbol{\Theta}^H, \tag{2.29}$$

where $\boldsymbol{\Theta} \in \mathbb{C}^{K \times K}$ is a pilot matrix with $\boldsymbol{\varphi}_k$ as its k-th column and $\mathbf{Y} \in \mathbb{C}^{M \times K}$ the received pilot matrix with $\mathbf{y}_m^T$ as its m-th row. We can then estimate $\mathbf{G}$ by performing right multiplication of $\mathbf{Y}$ with $\boldsymbol{\Theta}$.

Uplink Data Transmission

The data-bearing signal $x_{u,k} \in \mathbb{C}$ with $\mathbb{E}\{|x_{u,k}|^2\} = 1$ is sent in the uplink by the k-th user. All users send their data simultaneously and the m-th AP receives them as

$$y_{u,m} = \sqrt{\rho_u} \sum_{k=1}^{K} g_{mk} x_{u,k} + w_{u,m}. \tag{2.30}$$

The user detection is done at each AP by multiplying the received signal $y_{u,m}$ with the detector coefficient a_{mk}, which is equal to g_{mk}^* for conjugate beamforming. We assume that the channel is not known at the CPU and should be locally obtained from the estimation at the AP. Thus, the signal $\hat{g}_{mk}^* y_{u,m}$ is sent via the m-th fronthaul link, and

then the CPU receives from all M APs the signal of the k-th user as

$$\begin{aligned} r_{u,k} &= \sum_{m=1}^{M} \hat{g}_{mk}^{*} y_{u,m} \\ &= \sqrt{\rho_u} \sum_{k'=1}^{K} \sum_{m=1}^{M} \hat{g}_{mk}^{*} g_{mk'} x_{u,k'} + \sum_{m=1}^{M} \hat{g}_{mk}^{*} w_{u,m}. \end{aligned} \tag{2.31}$$

By plugging equation (2.28) in the equation (2.31) we can obtain the SINR in the uplink given as [6]

$$\mathrm{SINR}_{u,k}^{\mathrm{MRC}} = \frac{\rho_u \left(\sum\limits_{m=1}^{M} \gamma_{mk} \right)^2}{\rho_u \sum\limits_{k' \neq k}^{K} \left(\sum\limits_{m=1}^{M} \gamma_{mk} \frac{\beta_{mk'}}{\beta_{mk}} \right)^2 |\boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_{k'}|^2 + \rho_u \sum\limits_{k'=1}^{K} \sum\limits_{m=1}^{M} \gamma_{mk} \beta_{mk'} + \sum\limits_{m=1}^{M} \gamma_{mk}}, \tag{2.32}$$

where $\gamma_{mk} \triangleq \mathbb{E}\{|\hat{g}_{mk}|^2\}$.

Downlink Data Transmission

In the downlink the CPU has the data symbol $x_{d,k} \in \mathbb{C}$ for user k with $\mathbb{E}\{|x_{d,k}|^2\} = 1$. In the same fashion as in the uplink we can do precoding at the APs using the estimated channel $\hat{g}_{mk}$. The m-th AP sends the data-bearing signal for all K users as

$$y_{d,m} = \sqrt{\rho_d} \sum_{k=1}^{K} \hat{g}_{mk}^{*} x_{d,k}. \tag{2.33}$$

The k-th user receives the signal intended for it as

$$\begin{aligned} r_{d,k} &= \sum_{m=1}^{M} g_{mk} y_{d,m} + w_{d,k} \\ &= \sqrt{\rho_d} \sum_{m=1}^{M} \sum_{k'=1}^{K} g_{mk} \hat{g}_{mk'}^{*} x_{d,k'} + w_{d,k}. \end{aligned} \tag{2.34}$$

By treating the estimate $\hat{g}_{mk'}$ in (2.34) as the true channel, the effective channel term $g_{mk}\hat{g}_{mk'}^{*}$ will approximately turn in to a constant $\|g_{mk}\|^2$ for $k' = k$. Further, we can blindly decode the transmitted symbol $x_{d,k}$ since we are allowed to cancel out the effective channel term relying on the channel hardening (2.3). Here, the users do not need to know the channel realization g_{mk} but only the channel statistics. The SINR in downlink

for the k-th user is then given by [6]

$$\mathrm{SINR}_{d,k}^{\mathrm{MRC}} = \frac{\rho_u \left(\sum_{m=1}^{M} \gamma_{mk} \right)^2}{\rho_d \sum_{k' \neq k}^{K} \left(\sum_{m=1}^{M} \gamma_{mk'} \frac{\beta_{mk}}{\beta_{mk'}} \right)^2 |\boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k|^2 + \rho_d \sum_{k'=1}^{K} \sum_{m=1}^{M} \gamma_{mk'} \beta_{mk} + 1}. \tag{2.35}$$

2.2 Quantization

One of the main problems in the cell-free massive MIMO scheme described in the previous section is the high data rate of the I/Q baseband signal to be carried via the fronthaul. The load is even higher when the number of the user served is large. On the other hand, the capacity of the fronthaul must be limited to some extent. For this reason, a sort of signal compression should be performed at APs as well as at the CPU for efficient transmission. In this section we give an overview of signal compression, particularly quantization which is an important stage in the compression of a continuous valued signal.

The subject of signal compression is commonly studied under the terminology of *source coding*. It goes back to Shannon who first developed a theory underlying the quantification of how much a source signal can be compressed under some fidelity criteria. He gave a formal description when a source signal can be compressed without loss of information and when it can be compressed with possibly minimum distortion. It brought us then to two types of source coding namely *lossless* source coding and *lossy* source coding.

The aim of source coding is essentially to represent an information source in an efficient form which is appropriate for transmission or storage. For analysis, the information source is usually modeled as a random process, where information as a realization of random variables is produced by a source in each time unit according to some distribution. In general, the property desired by the coded source should have a data rate as low as possible while being able to recover the original information without error. If the source is in discrete form, the original information can exactly be recovered from its compressed version by lossless source coding. But if the source is in continuous form such as real signal, it is often that the source can be recovered only as an approximation of the original source. This is because a continuous source has an alphabet with an uncountably infinite number of elements, such that a part of them should be ignored for efficient computation. In such a case we deal with lossy source coding where an irreversible loss of information must be accepted.

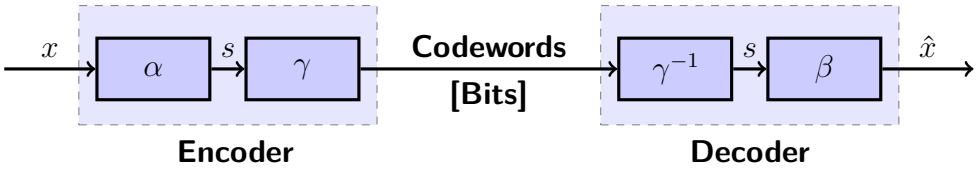


Fig. 2.2 The diagram of basics source coding scheme consists of lossy and lossless coding

A common approach for compressing a continuous source is by combining lossless and lossy source coding as depicted in Fig. 2.2. We have an encoder-decoder pair each of which consists of a lossy and a lossless part. The lossy encoder-decoder pair are denoted by α and β, whereas the lossless encoder-decoder pair are denoted by γ and γ^{-1} respectively. The encoded source, which is intended to be transmitted or stored in a medium, is represented as codewords in unit of bits. A comprehensive study of these topics is given in [39, 40], to which we consult for most of our following discussion.

In a lossy source coding system we discard intentionally some amount of information to make it possible to be further processed with finite precision. As shown in Fig. 2.2, the lossy part is presented by encoder α and β. The source output x is considered to be the realization of a continuous random source. To represent x with finite precision, the encoder α decides in which subset x should be contained. Each subset is associated with an index s, based on which the encoder γ selects to which codeword the source symbol can be assigned. Because we have γ and γ^{-1} as a lossless source coding system, we can obtain the index s accurately. According to the index s, the decoder β reconstructs then the source symbol as $\hat{x}$. This encoder-decoder pair, α and β together, is called *quantizer* $\beta(\alpha(x))$, and the process of mapping x to $\hat{x}$ is called *quantization*.

2.2.1 Scalar Quantization

For the case where we have a single sample x at a unit time as the input to the quantizer, we deal with a scalar quantization. It performs a mapping from a scalar input to a scalar *reconstruction value* in a countable set. An example of simple scalar mapping is the rounding operation to the nearest integer. The set from which the reconstruction value is taken is called a *codebook*. In the former example the codebook is an integer set which is of infinite size but countable. To be convenient for practical implementation the codebook size is chosen to be finite.

In general, a scalar quantization mapping with *codebook size* or *quantization level* S can be expressed as

$$Q : \mathbb{R} \to \hat{\mathcal{X}} = \{\hat{x}_0, \ldots, \hat{x}_{S-1}\} \subset \mathbb{R}. \tag{2.36}$$

The subset of $\mathbb{R}$, whose elements are mapped onto the same reconstruction value $\hat{x}_s$, is called as the quantization *interval*. To avoid ambiguity, we partition the real line into intervals given by

$$\mathcal{I}_s = \{x \in \mathbb{R} : Q(x) = \hat{x}_s\} = [u_s, u_{s+1}), \tag{2.37}$$

where u_s and u_{s+1} are the *decision thresholds*. The partition requires that the union of the interval $\mathcal{I}_s$ should cover the whole real line and there is no intersection between them. Further, we arrange the thresholds in increasing order $u_{s-1} < u_s < u_{s+1}$ such that we can calculate the interval size or *step size* as $\Delta_s = u_{s+1} - u_s$. Since we should consider the whole real line as the possible input value, we set the decision threshold $u_0 = -\infty$ and $u_S = \infty$. We call the unbounded intervals (u_0, u_1) and $[u_{S-1}, u_S)$ as the *overload region*, whereas the region with the bounded intervals $[u_1, u_{S-1})$ is called the *granular region*. By choosing the remaining $S-1$ thresholds and S reconstruction values as system parameters we can then specify the quantization mapping as a cascade of threshold operators given by

$$Q(\mathbf{x}) = \sum_{s=0}^{S-1} \hat{x}_s T_s(\mathbf{x}), \qquad \text{where } T_s(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{I}_s \\ 0 & \text{otherwise} \end{cases} \tag{2.38}$$

An example of a scalar quantization function $Q(x)$ is shown in Fig. 2.3 where we have a typical staircase characteristic that arises from the cascade of threshold operators in (2.38).

For efficient storage, the codebook size $|\hat{\mathcal{X}}| = S$ should be chosen with regard to the expected rate or *resolution*. While the term rate is universally used to describe the amount of information per channel use, the term resolution refers more to the measure of precision in describing an analog signal in digital form. We use both of them in this thesis interchangeably. More precisely, the average rate of a scalar quantizer can be given by

$$R = \mathbb{E}\{|\gamma(Q(X))|\} = \sum_{s=0}^{S-1} p(\hat{x}_s)|\gamma(\hat{x}_s)|, \tag{2.39}$$

where it depends on $|\gamma(\hat{x}_s)|$, which denotes the length of codeword assigned to $\hat{x}_s$, and depends on the probability of the quantizer output $\hat{x}_s$ given by

$$p(\hat{x}_s) = \Pr(X \in \mathcal{I}_s) = \int_{u_s}^{u_{s+1}} f(x)dx \tag{2.40}$$

with a probability density function of the input $f(x)$.

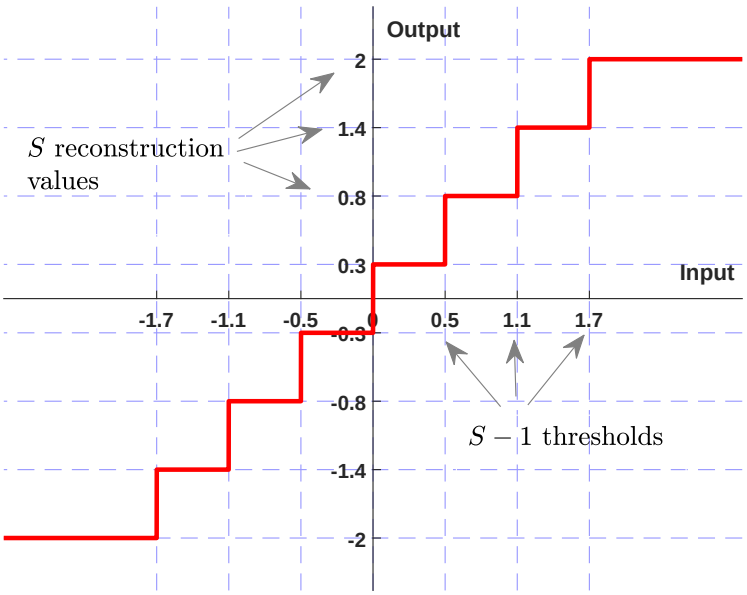


Fig. 2.3 Scalar quantization function of *mid-riser* quantizer (The quantizer is symmetric respected to the origin as decision threshold).

As given by equation (2.39), because the average rate R is dictated by γ, we should make the right assumption for γ in designing the quantizer Q. By applying a fixed-length code at γ, where a binary codeword with the same length is assigned to every reconstruction value x_s, the quantization level S should be equal to 2^R. If we know the probability of the quantizer output $p(x_s)$ for every index s, then we can apply variable length coding such as a Huffman code at the encoder γ. That is, the quantizer output with high probability will be assigned to a short codeword, whereas those with low probability will be assigned to the longer codeword. This allows us then to reduce the average rate even more.

As mentioned before, the consequence of representing a signal with finite precision is that we have to accept some errors. However, as long as these errors do not exceed our fidelity criteria, we can still expect the system to work well. A variety of distortion measures can be used as the fidelity criteria. We use here the most common one which is the squared error $d(x, \hat{x}) = (x - \hat{x}_s)^2$. Accordingly, we can measure the overall distortion based on its statistical average. That is, the average distortion of the scalar quantization is given by

$$D = \mathbb{E}\{d(X, Q(X))\} = \sum_{s=0}^{S-1} \int_{u_s}^{u_{s+1}} (x - \hat{x}_s)^2 f_X(x) \mathrm{d}x, \tag{2.41}$$

The expression in equation (2.39) and equation (2.41) are usually used to assess the performance of a particular quantizer by deriving the *operational rate-distortion function* $R(D)$. By doing so we can evaluate how much rate reduction can be achieved under some given distortion constraint.

2.2.1.1 Pulse Code Modulation (PCM)

The most straightforward scalar quantization is *Pulse Code Modulation* (PCM). For this type of quantizer the step size of each interval is chosen to have the same size:

$$\Delta = \frac{A}{S} = A \cdot 2^{-R}, \tag{2.42}$$

where the maximum value x_{max} and the minimum value x_{min} from the input are assumed to be known, so the range $A = x_{max} - x_{min}$ can be determined. In practice, maintaining the input signal in the finite range A can be done for example using Automatic Gain Control (AGC). The quantizer Q can be expressed then in closed form as

$$Q(x) = \left\lfloor \frac{x - x_{min}}{\Delta} + 0.5 \right\rfloor \cdot \Delta + x_{min}. \tag{2.43}$$

If we assume the input value x to be uniformly distributed as $f_X(x) = 1/A$ for $-A/2 \leq x \leq A/2$, so we can obtain the average distortion from equation (2.41) and (2.43) as

$$\begin{aligned} D &= \frac{\Delta^2}{12} \\ &= \frac{A^2}{12} \cdot 2^{-2R} \\ &= \sigma^2 \cdot 2^{-2R} = D(R) \end{aligned} \tag{2.44}$$

2.2.1.2 Optimized Scalar Quantizer

The PCM quantizer is not necessarily optimal particularly if the distribution of the input signal is not uniform. In designing the optimal quantizer the task is to find the codebook $\hat{\mathcal{X}}$ and the intervals $\mathcal{I}_s$ such that the objective function is minimized or maximized under some given constraints. The standard objective function used in practice is the average distortion which is given by equation (2.41) for the scalar quantizer. Given a fixed-rate scalar quantizer with size S, we should determine the $S-1$ thresholds u_s with $1 \leq s \leq S-1$ and the S reconstruction values $\hat{x}_s$ with $0 \leq s \leq S-1$ that minimize the average distortion D. The problem can be decomposed into these two following conditions

(i) Given the interval $\mathcal{I}_s = [u_s, u_{s+1})$, find the corresponding optimal codebook as

$$\begin{aligned} \{\hat{x}_s^*\} &= \arg\min_{\{\hat{x}_s\}} \quad \mathbb{E}\{d(X, Q(X))\} \\ &= \arg\min_{\{\hat{x}_s\}} \quad \sum_{s=0}^{S-1} D_s \\ &= \arg\min_{\{\hat{x}_s\}} \quad \sum_{s=0}^{S-1} \int_{u_s}^{u_{s+1}} d(x, \hat{x}_s) f_X(x) \mathrm{d}x \end{aligned} \tag{2.45}$$

The minimization can be observed separately for each $\mathcal{I}_s$ because they are assumed to be independent, where D is the sum of D_s depending only on each $\hat{x}_s$. Making use of the Bayes' rule

$$f_X(x) = f_{X|\hat{x}_s}(x|\hat{x}_s) \cdot p(\hat{x}_s), \tag{2.46}$$

we can express D_s as

$$\begin{aligned} D_s(\hat{x}_s) &= p(\hat{x}_s) \int_{u_s}^{u_{s+1}} d(x, \hat{x}_s) f_{X|\hat{x}_s}(x|\hat{x}_s) \mathrm{d}x \\ &= p(\hat{x}_s) \cdot \mathbb{E}\{d(X, \hat{x}_s) | X \in \mathcal{I}_s\}. \end{aligned} \tag{2.47}$$

Further, because $p(\hat{x}_s)$, given by equation (2.64), does not depend on $\hat{x}_s$, we have for each intervals the optimal reconstruction value as

$$\hat{x}_s^* = \arg\min_{\{\hat{x}_s\}} \quad \mathbb{E}\{d(X, \hat{x}_s) | X \in \mathcal{I}_s\}. \tag{2.48}$$

This condition is called the *centroid condition* and for the squared error distortion measure it has the solution

$$\begin{aligned} \hat{x}_s^* &= \mathbb{E}\{X | X \in \mathcal{I}_s\} \\ &= \frac{\int_{u_s}^{u_{s+1}} x f_X(x) \mathrm{d}x}{\int_{u_s}^{u_{s+1}} f_X(x) \mathrm{d}x} \end{aligned} \tag{2.49}$$

(ii) Given the codebook $\hat{\mathcal{X}} = \{\hat{x}_s\}$, find the optimal intervals $\mathcal{I}_s^* = [u_s^*, u_{s+1}^*)$, which satisfy

$$\mathcal{I}_s^* = \arg\min_{\mathcal{I}_s} \quad \sum_{s=0}^{S-1} \int_{u_s}^{u_{s+1}} d(x, \hat{x}_s) f_X(x) \mathrm{d}x. \tag{2.50}$$

Because $\hat{x}_s$ is fixed, it means if $x \in \mathcal{I}_s$ then

$$d(x, \hat{x}_s) \leq d(x, \hat{x}_j), \forall j \neq s \tag{2.51}$$

or the condition

$$Q(x) = \arg\min_{\hat{x}_s} \quad d(x, \hat{x}_s) \tag{2.52}$$

should be satisfied. This condition is called the *nearest neighbour condition*. Because the choice of u_s affects only the distortion of neighbouring intervals, the solution is u_s^* which satisfies

$$d(u_s^*, \hat{x}_s) = d(u_s^*, \hat{x}_{s+1}) \text{ for } 1 \leq s \leq S-1. \tag{2.53}$$

For the squared error distortion measure the optimal decision threshold is

$$u_s^* = \frac{1}{2}(\hat{x}_s + \hat{x}_{s+1}) \tag{2.54}$$

Algorithm 1: Lloyd Algorithm

input : realization $\{x\}$, quantizer size S
output : optimum recontruction values $\{\hat{x}_s\}$ and thresholds $\{u_s\}$

1 Choose initial reconstruction values $\{\hat{x}_s\}$;
2 Nearest neighbour condition: associate all samples of the training set $\{x\}$ with quantization interval $\mathcal{I}_s$ according to $\alpha(x) = \arg\min_{\forall s} \quad d(x, \hat{x}_s)$;
3 Update accordingly the decision threshold u_s;
4 Centroid condition : update $\{\hat{x}_s\}$ according to

$$\hat{x}_s = \arg\min_{\hat{x}_s \in \mathbb{R}} \quad \mathbb{E}\{d(X, \hat{x}_s) | \alpha(X) = s\}$$

with the expectation taken from the training set;
5 Repeat steps 2-4 until convergence;

To obtain both the optimal codebook and intervals, Lloyd gave a procedure known as the *Lloyd algorithm*, which iterates between the centroid and nearest neighbours conditions. The required pdf for the centroid condition given in equation (2.49) is commonly not known in practice. However, we can replace it with a sufficiently large training set in the algorithm. For a quantizer of size S with encoder mapping α and squared error distortion $d(x, \hat{x}_s)$ the Lloyd algorithm is given in Algorithm 1. We note that the convergence of Lloyd Algorithm is guaranteed and therefore we may take arbitrary initial reconstruction values in the first step. However, implausible initialization will result in long convergence time and convergence to a local optimum [41].

2.2.2 Vector Quantization

Vector quantization can be seen as a generalization of scalar quantization into higher dimensions. Suppose that we have a continuous random source $\{X\}$ with the realization x, we can arrange N samples of x as N an dimensional vector $\mathbf{x}$. A vector quantization maps the N dimensional space $\mathbb{R}^N$ into a countable set of reconstruction values or vectors in $\mathbb{R}^N$. More precisely, it is given by

$$Q : \mathbb{R}^N \to \hat{\mathcal{X}}, \text{ where } \hat{\mathcal{X}} = \{\hat{\mathbf{x}}_s\}_{s=0}^{S-1} \subset \mathbb{R}^N \tag{2.55}$$

is the *codebook* of finite size $|\hat{\mathcal{X}}| = S$ with $\hat{\mathbf{x}}_s \in \mathbb{R}^N$ for each $s \in \mathcal{J} \triangleq \{1, 2, \ldots, S\}$. The subset of $\mathbb{R}^N$, whose elements are mapped onto the same reconstruction value $\hat{\mathbf{x}}_s$, is called as the quantization *cell* given by

$$\mathcal{C}_s = \{\mathbf{x} \in \mathbb{R}^N : Q(\mathbf{x}) = \hat{\mathbf{x}}_s\}. \tag{2.56}$$

In similar fashion to intervals in scalar quantization, by doing vector quantization we implicitly divide the space $\mathbb{R}^N$ into cell partitions expressed by

$$\mathbb{R}^N = \bigcup_{s=0}^{S-1} \mathcal{C}_s, \text{ with } \forall s \neq j : \mathcal{C}_s \cap \mathcal{C}_j = \emptyset \tag{2.57}$$

Having defined the cells and their corresponding reconstruction values we can specify the quantization mapping as a cascade of threshold operators given by

$$Q(\mathbf{x}) = \sum_{s=0}^{S-1} \hat{\mathbf{x}}_s T_s(\mathbf{x}), \text{ where } T_s(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{C}_s \\ 0 & \text{otherwise} \end{cases} \tag{2.58}$$

It is obvious that the reconstruction value $\hat{\mathbf{x}}_s$ tends to be different from the element of input value $\mathbf{x}$. Representing all values in cell $\mathcal{C}_s$ by the reconstruction value $\hat{\mathbf{x}}_s$ means that we introduce a distortion to $\mathbf{x}$ to some degree. For a quantizer Q with input source $\{\mathbf{X}_n\}$ we define the average distortion as

$$D = \mathbb{E}\{d_N(\mathbf{X}_n, Q(\mathbf{X}_n))\} = \sum_{s=0}^{S-1} \int_{\mathcal{C}_s} d_N(\mathbf{x}, Q(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{2.59}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the joint probability density function (pdf) of the random vector $\mathbf{x}$ and $d_N(\mathbf{x}, \hat{\mathbf{x}}) \geq 0$ is the distortion measure between $\mathbf{x}$ and $\hat{\mathbf{x}}$. We use here the mean squared

error (MSE) as the distortion measure, where we have

$$d_N(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N}\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \tag{2.60}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}(x - \hat{x}_n)^2 . \tag{2.61}$$

Plugging this to (2.59) we obtain the average distortion as

$$D = \frac{1}{N}\sum_{s=0}^{S-1}\int_{\mathcal{C}_s}\|\mathbf{x} - \hat{\mathbf{x}}\|^2 f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{2.62}$$

Algorithm 2: Linde-Buzo-Gray (LBG) Algorithm

input : vector realization $\{\mathbf{x}_n\}$, codebook size S
output : optimum recontruction values $\{\hat{\mathbf{x}}_s\}$

1 Choose initial reconstruction values $\{\hat{\mathbf{x}}_s\}$;
2 Nearest neighbour condition: associate all samples of the training set $\{\mathbf{x}_n\}$ with quantization cell $\mathcal{C}_s$ according to $\alpha(\mathbf{x}_n) = \arg\min_{\forall s} \quad d(\mathbf{x}_n, \hat{\mathbf{x}}_s)$;
3 Update accordingly the decision threshold u_s;
4 Centroid condition : update $\{\hat{\mathbf{x}}_s\}$ according to

$$\hat{\mathbf{x}}_s = \arg\min_{\hat{\mathbf{x}}_s \in \mathbb{R}} \quad \mathbb{E}\{d(\mathbf{X}, \hat{\mathbf{x}}_s)|\alpha(\mathbf{X}) = s\}$$

with the expectation taken from the training set;
5 Repeat steps 2-4 until convergence;

To see the relation between the lossy and lossless coding part we define the average rate of the source coding system which includes the quantizer Q and lossless coding γ as

$$R = \frac{1}{N}\mathbb{E}\{|\gamma(Q(\mathbf{X}_n))|\} = \frac{1}{N}\sum_{s=0}^{S-1} p(\hat{\mathbf{x}}_s)|\gamma(\hat{\mathbf{x}}_s)|, \tag{2.63}$$

where $|\gamma(\hat{\mathbf{x}}_s)|$ is the codeword length of the encoded reconstruction value $\hat{\mathbf{x}}_s$ and $p(\hat{\mathbf{x}}_s)$ is the pmf of $\hat{\mathbf{x}}_s$ given by

$$p(\hat{\mathbf{x}}_s) = \int_{\mathcal{C}_s} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{2.64}$$

As stated in equation (2.64), the probability of reconstruction value $\hat{\mathbf{x}}_s$ depends on the distribution of the input value $\mathbf{x}$ which falls into the cell $\mathcal{C}_s$. In this context, quantization can be seen as discretization of the pdf.

To obtain an optimal codebook for vector quantization, the nearest neighbour and centroid conditions for scalar quantization can be extended to higher dimensions. Using a similar principle the *Linde-Buzo-Gray* (LBG) algorithm iteratively alternates between centroid and nearest neighbour conditions until a convergence is found. The procedure is presented in detail in Algorithm. 2. In case of two dimensional vector quantization, the resulting codebooks from the LBG algorithm are illustrated in Fig. 2.4 for some iterations. We fed the quantizer with independent random Gaussian signals with unit variance and chose the codebook of size $S = 16$. As can be seen after 50 iterations, the arrangement of the reconstruction values conforms the input distribution and most of the cells roughly have an hexagonal shape. It is well known that the hexagonal shape is optimum for packing and covering problem in two dimension. This demonstrates that the codebook has converged and approximately has reached its optimum.

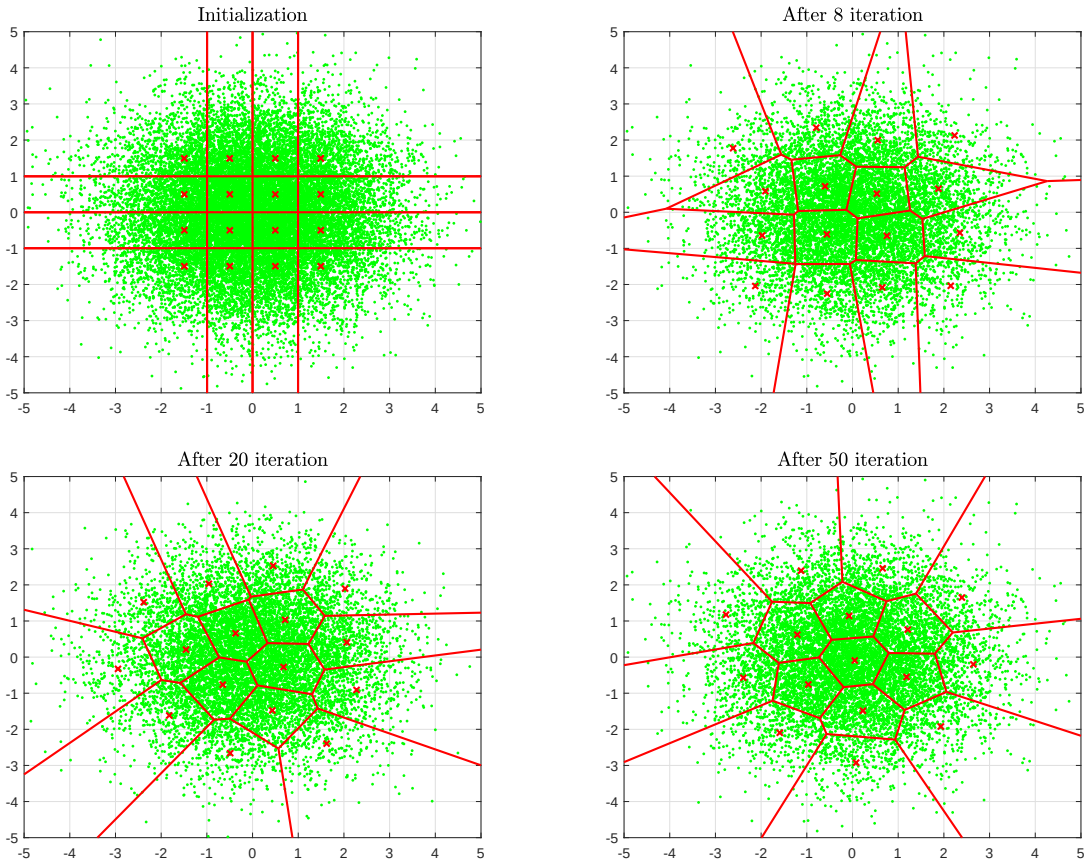


Fig. 2.4 The convergence of the LBG algorithm, in which the optimal codebook is obtained after 50 iterations.

2.2.3 High Resolution Approximation

In this and following subsections, we will discuss two options of modeling quantization based on assumptions about the resolution. A very common assumption made in practice is that the input signal is quantized with quite high resolution in order to achieve nearly the same quality as the original signal. This is done by choosing a large quantization level (in scalar quantization) or number of cells S (in vector quantization). If we further assume that that there is no input in the overload region and the pdf of the input signal is smooth, often a convenient approximation and a reasonable implication can be made.

In scalar quantization, making the quantization level large and ensuring the input lies in a finite range lead to the condition where the step size Δ becomes small and the overload probability of the input becomes low. If the step size is small enough and the input pdf is smooth, then the input pdf can be approximated as

$$f_X(x) \approx f_s; x \in \mathcal{I}_s. \tag{2.65}$$

That is, in each interval the input distribution is roughly constant. The probability of the quantizer output, which is equivalently to the probability of input falling in the interval $\mathcal{I}_s$, can also be given by

$$p(\hat{x}_s) = \Pr(X \in \mathcal{I}_s) = \int_{u_s}^{u_{s+1}} f(x)dx \approx (u_s - u_{s+1})f_s \tag{2.66}$$

A similar behaviour is also valid for the quantization error. It is Bennett [40] who investigated this for the first time and showed that the quantization error is uniformly distributed under the high-resolution assumption. More precisely, Bennett showed as follows [42]. Suppose that $\epsilon = Q(X) - X$ is the quantization error of the input with cumulative density function (cdf)

$$F_\epsilon(\alpha) = \Pr(\epsilon \leq \alpha) \text{ and pdf} \tag{2.67}$$

$$f_\epsilon(\alpha) = \frac{dF_\epsilon(\alpha)}{d\alpha}; \alpha \in (-\Delta/2, \Delta/2). \tag{2.68}$$

Then, it implies

$$\Pr(\epsilon \leq \alpha) = \sum_{s=0}^{S-1} \Pr(\epsilon \leq \alpha \text{ and } X \in \mathcal{I}_s) \tag{2.69}$$

$$= \sum_{s=0}^{S-1} \int_{-S/2+s\Delta}^{S/2+s\Delta+\alpha} f_X(\beta)d\beta \tag{2.70}$$

$$\approx \sum_{s=0}^{S-1} f_X(\hat{x}_s)\,\alpha \tag{2.71}$$

$$= \frac{\alpha}{\Delta} \sum_{s=0}^{S-1} f_X(\hat{x}_s)\,\Delta \tag{2.72}$$

$$\approx \frac{\alpha}{\Delta}, \tag{2.73}$$

where equation (2.70) follows from the assumption of smooth pdf and the mean value theorem of calculus. The last equation is due to the approximation of the Riemann integral with a summation. We obtain then the pdf of the quantization error as

$$f_\epsilon(\alpha) \approx \frac{1}{\Delta} \text{ for } \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right). \tag{2.74}$$

Additionally, Bennett showed that the power spectral density of ϵ is also flat. Those two conditions of the noise process ϵ, namely uniformly distributed and white, tend to be used as arguments for assuming that the signal x and ϵ are independent. Since for a given quantization error, one could expect that this error has been caused possibly by many different input signals. If they are fulfilled, then the quantization process can conveniently be represented using the *additive noise model* given by

$$Q(X) = X + \epsilon. \tag{2.75}$$

This model reminds us of the additive white Gaussian noise (AWGN) channel model in which the output is expressed as simple addition of the original signal with white noise. In a similar fashion, it is expected that using this model we can provide a tractable analysis for quantization.

Further, the above assumptions lead to an approximation of the average distortion in (2.41), which was formulated initially by Bennett [40] given as

$$D \approx \frac{1}{12}\frac{1}{S^2}\int_{x_1}^{x_{S-1}} f_X(\hat{x})\lambda(\hat{x})^{-2}d\hat{x}. \tag{2.76}$$

It expresses the average distortion as a function of the point density of reconstruction value $\lambda(x)$, the input pdf $f_X(x)$ and the quantization level S. Thus, the approximation accuracy is increasing with the increasing of quantization level S.

2.2.4 Low Resolution Approximation

In spite of its simple formulation, the additive noise model for quantization is valid only based on a weak assumption that the quantization error is uncorrelated to the input signal. In fact, quantization is a nonlinear operation and the quantization error is a deterministic function of the input. Hence, the input signal and the quantization error are essentially correlated, and their correlation is even stronger at low resolution. Modeling quantization at low resolution using the additive noise model can therefore result in rather poor performance.

As an alternative, we may make use of the result from Bussgang's theorem to model the quantization at low resolution. The theorem describes the statistical property of a nonlinear system when the input is Gaussian. It is stated in its original form as follows.

Theorem 2.2.1 (Bussgang's Theorem [43])**.**
The crosscorrelation function of two Gaussian signals taken after one of them has undergone nonlinear amplitude distortion is identical, except for a factor of proportionality, to the crosscorrelation function taken before the distortion.

If we consider only one Gaussian signal that has undergone a nonliear system, then the above theorem holds for the relationship between the input-output crosscorrelation and the autocorrelation of the input signal. Let $x(t)$ be a real Gaussian input signal that undergoes a nonlinear function $Q(x)$ resulting in a signal $y(t)$ at the output, then it holds that:

$$R_{xy}(\tau) = \alpha R_{xx}(\tau), \tag{2.77}$$

where $\forall \tau \in \mathbb{R}$

$$R_{xy}(\tau) = \mathbb{E}\{x(t)y(t+\tau)\} \text{ and} \tag{2.78}$$

$$R_{xx}(\tau) = \mathbb{E}\{x(t)x(t+\tau)\} \tag{2.79}$$

are the crosscorrelation and autocorrelation function of the stationary random process. The proportionality factor α depends on the characteristic of the nonlinear system Q, which is given as

$$\alpha = \frac{1}{\mathbb{E}\{|x|^2\}} \int_X xQ(x) f_X(x) dx. \tag{2.80}$$

As investigated by Bussgang in [43], the proportional factor α in (2.80) can be seen as cross-covariance function between the output and the input of the system, which should give some constant irrespective of τ. The implication of this Bussgang theorem is that

we can approximate the nonlinear system Q with the following linear model

$$y = Q(x) = \alpha x + d, \tag{2.81}$$

where the distortion d is now uncorrelated to the input x. To verify this, we can compute

$$\begin{aligned}
\mathbb{E}\{d(t)x(t+\tau)\} &= \mathbb{E}\{(y(t) - \alpha x(t))x(t+\tau)\} && (2.82)\\
&= \mathbb{E}\{y(t)x(t+\tau)\} - \alpha\mathbb{E}\{x(t)x(t+\tau)\} && (2.83)\\
&= R_{xy}(\tau) - \alpha R_{xx}(\tau) && (2.84)\\
&= 0, && (2.85)
\end{aligned}$$

where the last equation is obtained by plugging in the expression α from equation (2.77). Further, it is also of particular interest for the analysis to specify the proportionality factor between the power of the output and the power of the input from a nonlinear system. For this purpose, a factor λ is defined as

$$\begin{aligned}
\lambda &\triangleq \frac{\mathbb{E}\{|y|^2\}}{\mathbb{E}\{|x|^2\}}\\
&= \frac{1}{\mathbb{E}\{|x|^2\}} \int_X |Q(x)|^2 f_X(x)dx. && (2.86)
\end{aligned}$$

Although the Bussgang decomposition model in (2.81) is not as popular as the additive noise model, this model has been increasingly used for modeling quantization as reported among others in [44–47] especially when the resolution is low.

Chapter 3

Centralized Cell-Free massive MIMO with Single-Antenna Access Points

The demand for higher data rate communication through a wireless medium has been unstoppable for a long time. Rather, it increases further in the presence of newly envisioned applications such as autonomous vehicles, where a uniformly high data rate is expected by all users to be provided simultaneously across a wide area. While the cell-free massive MIMO discussed in the previous chapter has attracted much attention due to its ability to increase the capacity per user per unit area, its potential has not been entirely exploited. In this chapter, we aim to carry out cell-free massive MIMO to its greatest advantage when only a single antenna is available at access points. We show that the data throughput per user can be significantly improved if we apply the centralized approach to cell-free massive MIMO.

The use of a single antenna at APs is particularly worthwhile when the low-cost implementation of the AP is given a high priority. In this case, the number of required components such as the channel estimation module, RF-chain and ADC may not be more than one unit per AP. However, if we choose to operate the network using the distributed approach, the feasible choice of processing becomes restricted to conjugate beamforming. This is because the distributed approach prevents any AP obtaining CSI from other APs. Thus, the only way to process the received data-bearing signal is based on the locally obtained CSI, where for single antenna AP it is most advantageous to multiply the received signal by a single coefficient of the corresponding conjugate channel estimate. Unlike the distributed approach, we allow the APs in a centralized approach to transfer the CSI to the CPU via the fronthaul. This enables then another form of processing such as zero-forcing to be applied at the CPU.

There have been some works that analysed the performance of centralized cell-free massive MIMO with single antenna APs. For instance, the performance of the centralized approach using zero-forcing processing in the downlink is studied in [48], whereas MMSE processing in the uplink is compared in [7] with the standard distributed approach. However, none of the works mentioned above considered the limited capacity of the fronthaul links. At the beginning of this thesis project, very few works, such as [26, 49] had studied cell-free massive MIMO with limited-capacity fronthaul. Moreover, they did not consider how to realize centralized processing in practice particularly in terms of obtaining the global CSI.

In this chapter, a framework for designing a centralized cell-free massive MIMO with capacity-limited fronthauls is introduced in the case where only a single antenna is available at the APs. We begin the discussion in this chapter firstly by defining centralized cell-free massive MIMO and explaining its general concept. We will then give the scope of the discussion by describing the considered system model. To deal with a limited-capacity fronthaul, we wish to design an optimum quantizer for the fronthaul. This will be presented in the subsequent section for a scalar uniform quantizer with low resolution. We will then look at the important part of centralized cell-free massive MIMO which is the CSI acquisition. This is then followed by the scheme of data transmission considering the limited-capacity fronthaul. Further, the achievable rates of this approach are given where we will also derive a simpler SINR expression. Then, the scalability issue in cell-free massive MIMO will also be discussed. We will see that the centralized approach can resolve some aspects of this issue. Numerical results for validating our analysis and evaluating the performance of our proposed scheme are then given.

3.1 The General Concept

We start in this section with describing what we mean principally by centralized Cell-Free massive MIMO and elaborate the difference to the original cell-free massive MIMO that uses the distributed approach. To be concise, we specify formally the centralized cell-free massive MIMO in the following definition.

Definition 3.1 (Centralized Cell-Free massive MIMO)**.**
We say a cell-free massive MIMO network is centralized if all the following conditions are fulfilled.

1. *The CSI between all active UEs and all APs is available at the CPU.*
2. *The CPU utilizes the globally obtained CSI for joint data processing.*

We may notice that the above definition emphasizes the importance of CSI. This is reflecting Marzetta's view which says *"CSI isn't everything: it's the only thing!"* [50]. In contrast, the transfer of CSI has conventionally been seen as the major source of the increase in the fronthaul load, and therefore to be avoided. To that end, the CSI is utilized locally at the APs, whereas the coding/decoding of the data-bearing signal takes place at the CPU. This approach can perform well based on the underlying assumption that the favorable channel condition is also fulfilled in cell-free massive MIMO as in colocated massive MIMO. As a result, a low complexity processing such as conjugate beamforming is sufficient at the APs. However, it was later shown in [51] that the channel in cell-free massive MIMO is not as favorable as in colocated massive MIMO. That is, the channel orthogonality in cell-free massive MIMO tends to be poor, especially when the APs use a single antenna. This is because in cell-free massive MIMO a user is expected to be closer to some APs rather than to the other APs. Hence, the large scale fading coefficients to these closer APs are obviously smaller and the channel will experience a kind of spatial correlation, where some components in the channel matrix are stronger. Further, the work in [8, 10] revealed the limited performance of conjugate beamforming for cell-free massive MIMO with single-antenna APs. It is confirmed in [8] that the upper bound of the signal to interference noise ratio (SINR) per user can be achieved if only if the single-antenna APs are colocated.

The rationale behind the centralized approach is the fact that cell-free massive MIMO does not have the same desired channel orthogonality as in the colocated massive MIMO counterpart [51]. As a consequence, the conjugate beamforming does not give a satisfactory performance in cell-free massive MIMO even if power control is used. To improve the performance, centralized cell-free massive MIMO allows the transfer of CSI to the CPU. Although it costs additional fronthaul load, we will see that it can be compensated by the fronthaul load reduction from the data-bearing signal and the use of an appropriate CSI acquisition strategy. We will further show that the CSI exchange is not something to be avoided, and the effort for bringing CSI to the CPU is often very well repaid by a significant throughput improvement.

This centralized approach may lead us also to think about cloud-radio access networks (C-RAN), where the radio access processing of multiple base stations is pooled in one baseband unit (BBU) [52]. Although C-RAN and centralized cell-free massive MIMO have a close connection in centralizing the baseband processing, they exhibit some apparent differences. In C-RAN, radio access is treated as a cloud service which is provided by the BBU pool through software-defined radio (SDR). By doing this, it is expected that the deployment and operational cost can be reduced due to the flexible implementation

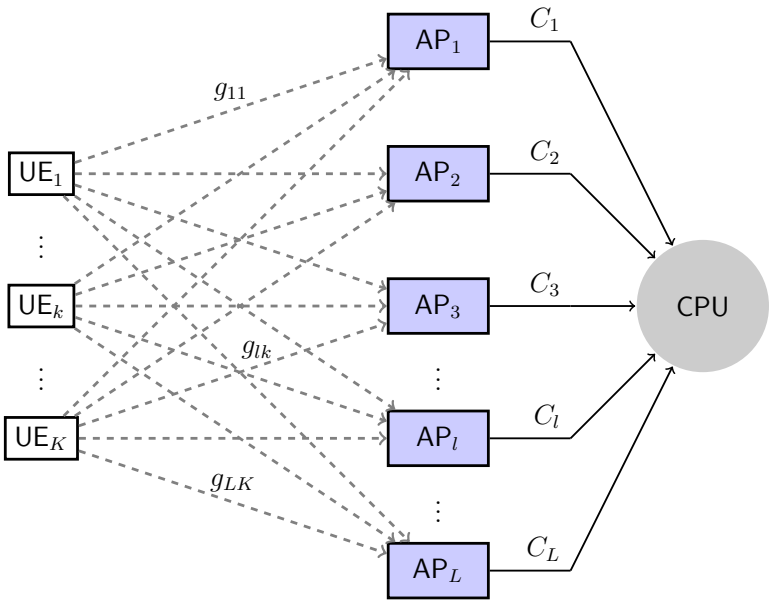


Fig. 3.1 The Schematic diagram of the centralized Cell-Free Massive MIMO with L access points, K users and L capacity-limited fronthaul links.

of the networks. Therefore, C-RAN is independent of radio access technology. Moreover, C-RAN has been mostly studied by this time in the setting of a cellular system. On the other hand, cell-free massive MIMO, particularly the centralized approach, was motivated by different aspects namely to uniformly provide a high data throughput in a large coverage area. In this case, cell-free massive MIMO in the physical layer sticks to include the assumption of channel hardening and favourable propagation. Further, cell-free massive MIMO has an asymmetric setup between uplink and downlink due to the larger number of APs to the number of users. Therefore, the achievable rate duality that holds in C-RAN might not be the case in cell-free massive MIMO [53].

3.2 System Model

We consider in this chapter the uplink transmission of a cell-free massive MIMO system with K single-antenna users (UEs) and L single-antenna APs, which are randomly placed in a wide area. All UEs can be served simulataneously by all APs in the same time-frequency resource. We assume that the APs are sufficiently separated greater than the half wavelength, such that we do not observe any spatial channel correlation between the APs. To jointly process the signal from all UEs, the APs are connected to a CPU by L error-free fronthaul links. The main processing for the L APs are centralized at the CPU by making use of the globally obtained CSI of the served UEs. The communication

between all APs and the CPU is carried out coherrently in baseband. To gain more insight, the schematic diagram of the centralized cell-free massive MIMO is depicted in Figure 3.1. This is actually similar to the original cell-free massive MIMO [6] described in Section 2.1.2, where the number of UEs K is smaller than the number of serving APs L, except that the fronthaul link connecting the l-th AP with the CPU has limited capacity of C_l.

As in canonical massive MIMO described in Section 2.1.1, the network is operated using the TDD protocol. However, to keep the later analysis simple we consider the whole coherence interval to be used only for pilot and uplink data transmission. Suppose that a length τ_c coherence interval is available, we use a fraction τ_p for pilot and a fraction of τ_u for payload data where $\tau_c = \tau_p + \tau_u$. The channel we consider in this chapter is also similar to the one described in Section 2.1.2. Recall that the channel between the k-th user and the l-th AP is specified by

$$g_{lk} = h_{lk}\beta_{lk}^{1/2}, \tag{3.1}$$

where the coefficient h_{lk} models the small-scale fading between the k-th user and the l-th AP with the assumption that it is i.i.d. $\sim \mathcal{CN}(0,1)$. The large-scale fading is denoted by β_{lk} which is likely to be different for each user k and each AP l due to the distributed configuration. The channel from all K users to all L APs can then be expressed as the element-wise product of the small-scale fading matrix $\mathbf{H} \in \mathbb{C}^{L\times K}$ and the large-scale fading matrix $\mathbf{D} \in \mathbb{R}^{L\times K}$ given by

$$\mathbf{G} = \mathbf{H} \odot \mathbf{D}^{1/2}, \text{ where } [\mathbf{H}]_{lk} = h_{lk} \text{ and } [\mathbf{D}]_{lk} = \beta_{lk}. \tag{3.2}$$

3.3 Fronthaul Quantization

As mentioned earlier, we are dealing throughout the thesis with capacity-limited fronthaul links. We also mentioned that the links should be error-free in the sense that any bit stream we transmit through the fronthaul link can be reproduced at the CPU with arbitrary small error probability. However, the way to achieve this error-free transmission is a discussion of channel coding, and should be beyond the scope of this thesis. Similarly, we put beyond our scope the problem of the delay imposed by the transmission through the fronthaul. It is sufficient for us to think that a reliable transmission can be carried out as long as we send a bit stream less than the channel capacity. Therefore for efficient transmission, the signal source at the input of the fronthaul link should be represented in

less than the fronthaul capacity. Since we receive analog signal at the APs but transmit through the fronthaul in digital form, we need to convert the signal by ADC before the transmission. In this case, the fronthaul capacity dictates the resolution we use in the ADC. Further, we consider using a fixed-rate lossless coding such that we may also say that the fronthaul has a resolution of the fronthaul capacity in bits per channel use.

3.3.1 Quantization Model

To simplify our analysis, we consider fronthaul links with capacity of $C_l = C$ bits per channel use for all $l \in \{1, \ldots, L\}$. We consider first in this chapter a scalar quantization in our ADC with resolution $R \leq C$ bits, which is optimally adjusted at fronthaul resolution C. The ADC resolution is related to the quantization level S by $R = \log_2 S$. More precisely, we apply an S-level scalar quantizer Q at each AP with

$$Q(x) = \sum_{s=0}^{S-1} q_s T_s(x), \text{ where } T_s(x) = \begin{cases} 1 & \text{if } x_s < x \leq x_{s+1} \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

We consider Q to be a uniform quantizer in which we have a fixed step size $\Delta = x_{s+1} - x_s$ for $s = 1, \ldots, S-1$, and set the decision threshold $x_0 = -\infty$ and $x_S = \infty$. The reconstruction value is given by $q_s = (s - \frac{S-1}{2})\Delta$. For a complex-valued signal $x \in \mathbb{C}$ we quantize the real and imaginary part separately. In this case, whenever we have $x_s < \mathbb{R}\text{e}\{x\} \leq x_{s+1}$ and $x_{s'} < \mathbb{I}\text{m}\{x\} \leq x_{s'+1}$ for $(s, s') \in \{0, \ldots, S-1\}$, we obtain

$$x_q = Q(x) = Q(\mathbb{R}\text{e}\{x\}) + iQ(\mathbb{I}\text{m}\{x\}) \tag{3.4}$$

$$= q_s^R + iq_{s'}^I, \tag{3.5}$$

where q_s^R and $q_{s'}^I$ are respectively the reconstruction values of the real and imaginary part with the pair $(q_s^R, q_{s'}^I) \in \{q_0^R, \ldots, q_{S-1}^R\} \times \{q_0^I, \ldots, q_{S-1}^I\}$. Moreover, the quantization operation should apply elementwise for a vector valued input. We assume that the large scale fading β_{lk} is relatively constant over a long period and known at the APs. Thus, we can scale the input-output signal of the quantizer according to β_{lk} and approximate the normalised input as normally distributed.

Subsequently, we would like to put constraint on the fronthaul resolution which is expected to be low for some practical reason. As explained in Section 2.2, keeping the resolution low implies the quantization process becomes strongly nonlinear. As for small S the function Q is nonlinear, it is not appropriate to model the quantization using the traditional additive noise model. The assumption in that model that the quantization

noise is uncorrelated with the input signal is then no longer valid. We therefore use a more accurate model to analyse our quantization, based on the Bussgang decomposition [43] (see also Section 2.2.4). Accordingly, for a nonlinear function $Q(x)$ we can write it as

$$x_q = Q(x) = \alpha_q x + d, \tag{3.6}$$

where the distortion term d is uncorrelated to the input signal x. Here, we denote the proportional factor of the quantizer Q by α_q. Recall from Section 2.2.4 that based on the Bussgang's theorem [43] the factor α_q depends on the characteristic of the quantizer Q and the distribution $f(x)$ of the input signal x. It is given by

$$\alpha_q = \frac{1}{P_x} \int_x x Q^*(x) f(x) dx, \tag{3.7}$$

where $P_x = \mathbb{E}\{|x|^2\}$ is the power of x. Further, we define the power ratio of the input x and the output x_q as

$$\begin{aligned} \lambda_q &\triangleq \frac{\mathbb{E}\{|x_q|^2\}}{\mathbb{E}\{|x|^2\}} \\ &= \frac{1}{P_x} \int_x |Q(x)|^2 f(x) dx. \end{aligned} \tag{3.8}$$

3.3.2 Optimum Quantization

The problem of finding the optimum quantizer has long been addressed by Lloyd and Max [54, 55] for a mean squared error distortion measure. We have seen in Section 2.2.1, that the conditions for which a scalar quantizer may achieve the optimum thresholds $[x_{s+1} x_s)$ and optimum reconstruction value q_s are the nearest neighbourhood and the centroid condition. Since the problem has no closed form solution, Lloyd and Max proposed to solve it numerically by iteratively altering between the two conditions (see Algorithm 1). For the uniform quantizer we consider in this chapter, Max has also provided a numerical result in [54] for the step size Δ that gives a minimum distortion in terms of mean squared error.

Similar to [26], we present in this subsection another way to find an optimum uniform quantizer based on our Bussgang model. Given a Gaussian input signal x, we are interested in finding Δ for which the distortion d in (3.6) is minimized. Instead of minimizing the mean squared error, we consider here to choose the step size Δ that maximizes the signal to distortion noise ratio (SDNR) at the output of the quantizer

defined as

$$\text{SDNR} = \frac{\mathbb{E}\{|\alpha_q x|^2\}}{\mathbb{E}\{|d|^2\}}. \tag{3.9}$$

The power of the distortion can be calculated from (3.6) and (3.8) as

$$\begin{aligned}\mathbb{E}\{|d|^2\} &= \mathbb{E}\{|x_q - \alpha x|^2\} \\ &= (\lambda_q - \alpha_q^2)\mathbb{E}\{|x|^2\}\end{aligned} \tag{3.10}$$

such that the SDNR can be written as

$$\begin{aligned}\text{SDNR} &= \frac{\alpha_q^2}{\lambda_q - \alpha_q^2} \\ &= \frac{\alpha_q^2/\lambda_q}{1 - \alpha_q^2/\lambda_q}.\end{aligned} \tag{3.11}$$

To find Δ that maximizes SDNR we need first to find an expression of α_q and λ_q as a function of variable Δ. The closed form expression of them are given in Proposition 3.3.1. Using equations (3.12) and (3.13) we characterize the Bussgang decomposition such that it is directly related to Δ and S. This will be useful for the analysis and numerical evaluation of the quantization process.

Proposition 3.3.1 (Bussgang decomposition scaling factors [26, 56])**.**
Consider a uniform mid-rise quantizer Q given in (3.3) with Gaussian signal input and unit variance. If the quantizer Q is modelled by the Bussgang decomposition given in (3.6), then the linear factor α_q and the power scaling factor λ_q can be written as a function of the step size Δ by

$$\alpha_q = \frac{\Delta}{\sqrt{2\pi}}\left(1 + 2\sum_{s=1}^{S/2-1}\exp(-s^2\Delta^2)\right) \quad \textit{and} \tag{3.12}$$

$$\lambda_q = \Delta^2\left(\frac{1}{4} + 4\sum_{s=1}^{S/2-1} s(1 - \Phi(s\Delta))\right). \tag{3.13}$$

Proof. By substituting the scalar quantization function $Q(x)$ given in (3.3) into (3.7) we obtain

$$\alpha_q = \frac{1}{P_x}\int_x xQ^*(x)f(x)dx$$

$$
\begin{aligned}
&= \frac{1}{P_x} \sum_{s=0}^{S-1} q_s \int_{x_s}^{x_{s+1}} x f(x) dx \\
&= \frac{1}{P_x} \sum_{s=0}^{S-1} \left(s - \frac{S-1}{2} \right) \Delta \int_{x_s}^{x_{s+1}} x f(x) dx \qquad (3.14) \\
&= -\frac{1}{P_x} \left(\frac{S-1}{2} \right) \Delta \int_{-\infty}^{x_1} x f(x) dx \\
&\quad + \frac{1}{P_x} \sum_{s=1}^{S-2} \left(s - \frac{S-1}{2} \right) \Delta \int_{x_s}^{x_{s+1}} x f(x) dx \\
&\quad + \frac{1}{P_x} \left(\frac{S-1}{2} \right) \Delta \int_{x_{S-1}}^{\infty} x f(x) dx, \qquad (3.15)
\end{aligned}
$$

where the first term cancels the last term due to the symmetry from the tails of the Gaussian function. Then, by evaluating the integral with Gaussian distribution $f(x)$ we obtain

$$
\begin{aligned}
\alpha_q &= \frac{1}{P_x} \sum_{s=1}^{S-2} \left(s - \frac{S-1}{2} \right) \Delta \int_{x_s}^{x_{s+1}} x f(x) dx \qquad (3.16) \\
&= \frac{1}{P_x} \sum_{s=1}^{S-2} \left(s - \frac{S-1}{2} \right) \Delta \left[\sqrt{\frac{2}{\pi}} P_x \left[-\frac{1}{2} \exp\left(-t^2\right) \right]_{x_s}^{x_{s+1}} \right] \qquad (3.17) \\
&= \frac{\Delta}{\sqrt{2\pi}} \sum_{s=1}^{S-2} \left(s - \frac{S-1}{2} \right) \left[\exp\left(-s^2\Delta^2\right) - \exp\left(-(s+1)^2\Delta^2\right) \right]. \qquad (3.18)
\end{aligned}
$$

Since we have a uniform step size Δ for $s = 1, \ldots, S-1$, we can substitute the decision threshold x_s and x_{s+1} respectively by $s\Delta$ and $(s+1)\Delta$ to obtain the last equation. We then expand the expression in the bracket such that we have

$$
\begin{aligned}
\alpha_q &= \frac{\Delta}{\sqrt{2\pi}} \sum_{s=1}^{S-2} \left(s - \frac{S-1}{2} \right) \left[\exp\left(-s^2\Delta^2\right) \right] \\
&\quad - \frac{\Delta}{\sqrt{2\pi}} \sum_{s'=1}^{S-2} \left((s'-1) - \frac{S-1}{2} \right) \left[\exp\left(-s'^2\Delta^2\right) \right] \qquad (3.19) \\
&= \frac{\Delta}{\sqrt{2\pi}} \sum_{s=1}^{S-2} \left(s - \frac{S-1}{2} - \left((s-1) - \frac{S-1}{2} \right) \right) \left[\exp\left(-s^2\Delta^2\right) \right] \qquad (3.20) \\
&= \frac{\Delta}{\sqrt{2\pi}} \sum_{s=1}^{S-2} \exp\left(-s^2\Delta^2\right) \qquad (3.21) \\
&= \frac{\Delta}{\sqrt{2\pi}} \left(1 + 2 \sum_{s=1}^{S/2-1} \exp(-s^2\Delta^2) \right), \qquad (3.22)
\end{aligned}
$$

Table 3.1 Optimum step size and power distortion

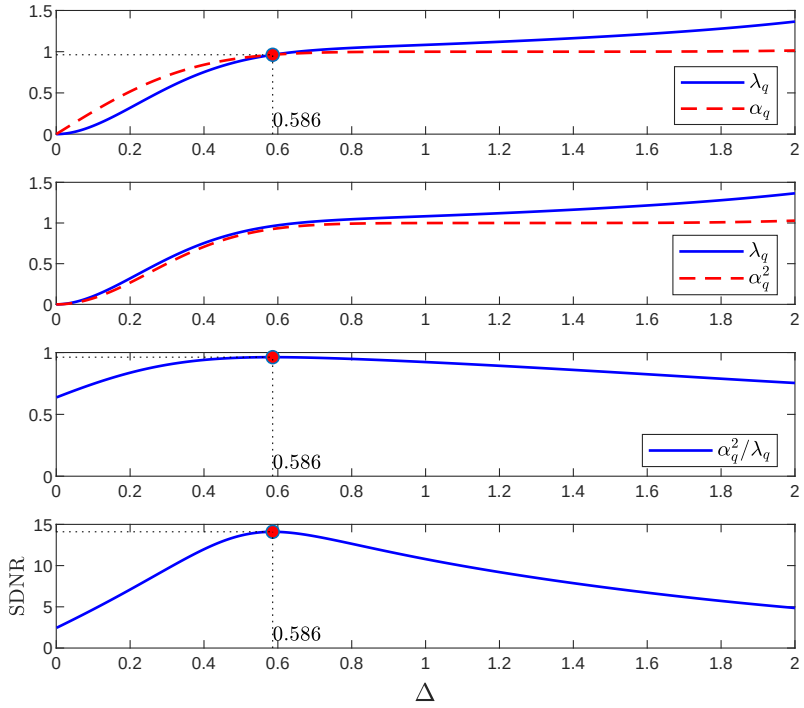
C [bits]	Δ_{opt} [54]	MSE [54]	$\mathbb{E}\{\lvert d\rvert^2\}$	α_q
1	1.5960	0.3634	0.231401	0.6367
2	0.9957	0.1188	0.104722	0.8812
3	0.5860	0.03744	0.036037	0.9626
4	0.3352	0.01154	0.011409	0.9885
5	0.1881	0.003490	0.003483	0.9965

where the last equation follows from the symmetry at the decision threshold $x_s = 0$ or $s = S/2$ while the equation (3.20) follows from the substitution $s' = s + 1$ with some abuse of notation. Similarly, we can express the power scaling factor as

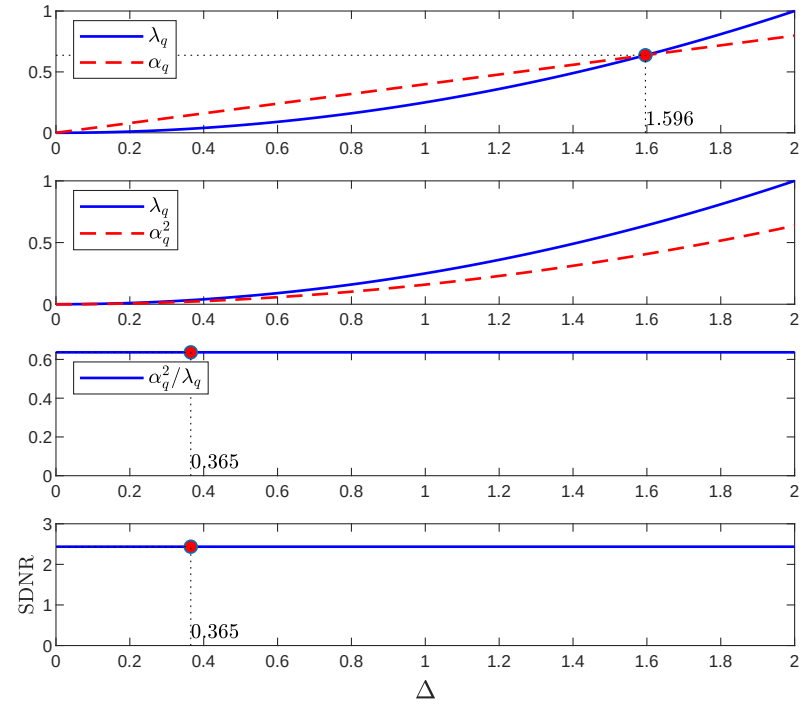
$$\begin{aligned}\lambda_q &= \frac{1}{P_x}\int_x |Q(x)|^2 f(x)dx \\ &= \frac{1}{P_x}\sum_{s=0}^{S-1} q_s^2 \int_{x_i}^{x_{s+1}} f(x)dx \\ &= \Delta^2\left(\frac{1}{4} + 4\sum_{s=1}^{S/2-1} s(1-\Phi(s\Delta))\right),\end{aligned} \tag{3.23}$$

where Φ is the Gaussian cumulative distribution function. ■

Having expressed α_q and λ_q as given in (3.12) and (3.13) allows us now to find the optimum step size Δ_{opt} that maximizes the SDNR. It has previously shown in [46] for a memoryless nonlinear system that maximizing SDNR should give the same result as minimizing the MSE when the linear scale α_q is equal to the power scale λ_q. However, from (3.10) we can observe that plugging (3.12) and (3.13) into (3.10) then directly minimizing distortion may result in the trivial solution of $\Delta_{\text{opt}} = 0$. Therefore, we prefer here to maximize the SDNR rather than to minimize the distortion. We note that the power distortion can not be negative in practice and hence should be equal or greater than zero. It turns out that we obtain from (3.10) $\lambda_q \geq \alpha_q^2$ and $(\alpha_q^2/\lambda_q) \leq 1$. In this case from (3.11), maximizing the SDNR is equivalent to maximizing α_q^2/λ_q. In Figure 3.2 an example of the maximization problem is illustrated for $S = 8$ and $S = 2$, where the required condition $\lambda_q \geq \alpha_q^2$ and $(\alpha_q^2/\lambda_q) \leq 1$ are shown to be fulfilled. The optimum



(a) $S = 8$



(b) $S = 2$

Fig. 3.2 The optimum step size obtained from maximizing SDNR [dB].

Δ_{opt} is then obtained from

$$\begin{aligned}\Delta_{\text{opt}} &= \arg\max_{\Delta} \quad \left[\frac{\alpha_q^2}{\lambda_q}\right] \\ &= \arg\max_{\Delta} \quad \left[\frac{\left(\frac{\Delta}{\sqrt{2\pi}}\left(1+2\sum_{s=1}^{S/2-1}\exp(-s^2\Delta^2)\right)\right)^2}{\Delta^2\left(\frac{1}{4}+4\sum_{s=1}^{S/2-1}s(1-\Phi(s\Delta))\right)}\right].\end{aligned} \tag{3.24}$$

For a real Gaussian input signal x with unit variance, we evaluate the problem in (3.24) numerically where the results correspond to [54] for $S > 2$. For $S = 2$, the problem (3.24) however does not have a unique solution as shown in Figure 3.2b. Thus, we choose $\Delta_{\text{opt}} = 1.5960$ at $\alpha_q = \lambda_q$ which is also equal to the result in [54]. If Δ_{opt} is chosen, the resulting power distortion and the proportional factor α_q are listed in Table 3.1 for the first 5 bits resolution.

3.4 Channel State Information Acquisition Strategies

As given in Definition 3.1, one of the key ingredients of centralized cell-free massive MIMO is the availability of the CSI at the CPU. In this section we present the methods how to make this CSI available when we constrain the fronthaul resolution to be low. In this regard, we will consider two CSI acquisition strategies that take into account the low-resolution fronthaul by utilizing the quantization model described in the previous section.

We use the common approach where the CSI is acquired based on the estimation of known pilots transmitted by the users. The k-th user transmits $\sqrt{\tau_p}\boldsymbol{\varphi}_k$ as its pilot, where a specific random sequence $\boldsymbol{\varphi}_k \in \mathbb{C}^{\tau_p \times 1}$ is taken from an orthonormal basis satisfying $|\langle \boldsymbol{\varphi}_k, \boldsymbol{\varphi}'_k \rangle| = \delta_{kk'}$ and $\|\varphi_k\|^2 = 1$. The sequence length τ_p is assumed to be less than or equal to the coherence interval τ_c. The l-th AP observes the received pilot $\mathbf{y}_l$ from all K users as

$$\mathbf{y}_{p,l} = \sqrt{\tau_p \rho_p} \sum_{k=1}^{K} g_{lk}\boldsymbol{\varphi}_k + \mathbf{w}_p, \tag{3.25}$$

where ρ_p is the normalized transmit power of the pilot and the vector $\mathbf{w}_p \sim \mathcal{CN}(0, \mathbf{I}_K)$ is an additive noise vector with zero mean and identity covariance. To ensure that all pilots are orthogonal for all K users, one should only allow $K \leq \tau_p$ users to transmit

their pilots simultaneously. In this case, the transmitted pilots satisfy

$$\Theta^H\Theta = \tau_p\rho_p\mathbf{I}_K, \text{ where } \Theta = \sqrt{\tau_p\rho_p}[\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_K]. \tag{3.26}$$

Let us consider for a moment the ideal case where the fronthaul is perfect. Then, the channel g_{lk} can be estimated at the AP and sent to the CPU to provide a global CSI without any impairments. In this case, the received pilot $\mathbf{y}_{p,l}$ at the l-th AP is projected onto $\boldsymbol{\varphi}_k^H$ giving

$$\begin{aligned} r_{p,lk} &= \boldsymbol{\varphi}_k^H\mathbf{y}_{p,l} \\ &= \sqrt{\tau_p\rho_p}g_{lk} + \sqrt{\tau_p\rho_p}\sum_{k'\neq k}^{K} g_{lk'}\boldsymbol{\varphi}_k^H\varphi'_k + \boldsymbol{\varphi}_k^H\mathbf{w}_p. \end{aligned} \tag{3.27}$$

To obtain the estimate of g_{lk} we use the Linear Minimum Mean Squared Error (LMMSE) estimator given by

$$\hat{g}_{lk} = c_{lk}r_{p,lk}. \tag{3.28}$$

In this case, we choose c_{lk} that minimizes the Mean Squared Error (MSE)

$$\begin{aligned} \epsilon_{lk} &= \mathbb{E}\{|g_{lk} - \hat{g}_{lk}|^2\} \\ &= \mathbb{E}\{|g_{lk}|^2\} + \mathbb{E}\{|\hat{g}_{lk}|^2\} - 2\,\mathbb{R}\mathrm{e}\{\mathbb{E}\{g_{lk}\hat{g}_{lk}^*\}\} \\ &= \mathbb{E}\{|g_{lk}|^2\} + c_{lk}^2\mathbb{E}\{|r_{p,lk}|^2\} - 2c_{lk}\,\mathbb{R}\mathrm{e}\{\mathbb{E}\{g_{lk}r_{p,lk}^*\}\}. \end{aligned} \tag{3.29}$$

The unique minimum is obtained by taking the derivative of ϵ_{lk} and setting it equal to zero expressed as

$$0 = \frac{\partial\epsilon_{lk}}{\partial c_{lk}} = 2c_{lk}\mathbb{E}\{|r_{p,lk}|^2\} - 2\,\mathbb{R}\mathrm{e}\{\mathbb{E}\{g_{lk}r_{p,lk}^*\}\} \tag{3.30}$$

such that

$$c_{lk} = \frac{\mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,lk}^*g_{lk}\}\}}{\mathbb{E}\{|r_{p,lk}|^2\}} \tag{3.31}$$

$$= \frac{\sqrt{\tau_p\rho_p}\beta_{lk}}{\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2 + 1}, \tag{3.32}$$

where the last equation follows from (3.27). Further, we use γ_{lk} to denote the mean square of the channel estimate given by

$$\begin{aligned}\gamma_{lk} &\triangleq \mathbb{E}\{|\hat{g}_{lk}|^2\} \\ &= c_{lk}^2 \mathbb{E}\{|r_{p,lk}|^2\} \\ &= c_{lk}\, \mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,lk}^* g_{lk}\}\} \\ &= c_{lk} \sqrt{\tau \rho_p} \beta_{lk} \\ &= \frac{\tau_p \rho_p \beta_{lk}^2}{\tau_p \rho_p \sum_{k'=1}^{K} \beta_{lk'} |\boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_{k'}|^2 + 1}.\end{aligned} \tag{3.33}$$

After substituting the optimal coefficient c_{lk} given by (3.31) into (3.29), the minimum mean squared error can be expressed then as

$$\begin{aligned}\epsilon_{lk} &= \mathbb{E}\{|g_{lk}|^2\} - \frac{(\mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,lk}^* g_{lk}\}\})^2}{\mathbb{E}\{|r_{p,lk}|^2\}} \\ &= \mathbb{E}\{|g_{lk}|^2\} - \frac{\mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,lk}^* g_{lk}\}\}\ \mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,lk}^* g_{lk}\}\}}{\mathbb{E}\{|r_{p,lk}|^2\}} \\ &= \mathbb{E}\{|g_{lk}|^2\} - c_{lk}\ \mathbb{R}\mathrm{e}\{\mathbb{E}\{r_{p,lk}^* g_{lk}\}\} \\ &= \beta_{lk} - c_{lk}\ \sqrt{\tau \rho_p} \beta_{lk} \\ &= \beta_{lk} - \gamma_{lk}.\end{aligned} \tag{3.34}$$

We may see (3.34) as the minimum achievable acquisition error which can be used for comparison with our acquisition strategies with low-resolution fronthaul.

3.4.1 Estimate-and-Quantize

We now return to the case of imperfect fronthaul with low resolution. We consider first the more straight forward strategy to acquire CSI at the CPU. In this scheme we estimate the channel coefficient g_{lk} first as given in (3.28). In order that it may be sent via limited fronthaul to the CPU, the estimated channel $\hat{g}_{lk}$ is quantized at each AP. Therefore, we call this strategy estimate-and-quantize. Since we send the quantized version $\hat{g}_{lk}^{eq}$ to the CPU, the amount of CSI overhead resulting from this scheme is proportional to the number of users K. For a symbol frame of length τ_c the portion of CSI overhead is then K/τ_c. After transferring via the fronthaul the CPU receives $\hat{g}_{lk}^{eq}$, which can be decomposed due to Bussgang (3.6) as

$$\hat{g}_{lk}^{eq} = Q(\hat{g}_{lk}) = \alpha_{eq} \hat{g}_{lk} + d_{eq}. \tag{3.35}$$

To see how large the performance loss is due to this strategy we are interested to find the MSE after the quantization. We formulate in Lemma 3.4.1 the MSE of this EQ acquisition scheme in relation to the CSI of perfect fronthaul given in (3.34).

Lemma 3.4.1 (The MSE of Estimate-and-Quantize)**.**
Suppose that the CSI for centralized cell-free massive MIMO with a single antenna AP is acquired using the Estimate-and-Quantize strategy under a low-resolution fronthaul constraint. The MSE of the CSI at the CPU is given by

$$\epsilon_{lk}^{eq} = \beta_{lk} - (2\alpha_{eq} - \lambda_{eq})\gamma_{lk}. \tag{3.36}$$

Proof. Following the definition of MSE, we write the MSE of the channel estimate after quantization as

$$\epsilon_{lk}^{eq} = \mathbb{E}\{|g_{lk} - \hat{g}_{lk}^{eq}|^2\} \tag{3.37}$$

$$= \mathbb{E}\{|g_{lk}|^2\} + \mathbb{E}\{|\hat{g}_{lk}^{eq}|^2\} - 2\,\mathrm{Re}\{\mathbb{E}\{g_{lk}^* \hat{g}_{lk}^{eq}\}\}. \tag{3.38}$$

Applying (3.35) allows us to express the expectation in the last term as follows

$$\begin{aligned}\mathbb{E}\{g_{lk}^* \hat{g}_{lk}^{eq}\} &= \alpha_{eq}\mathbb{E}\{g_{lk}^* \hat{g}_{lk}\} + \mathbb{E}\{g_{lk}^* d_{eq}\} \\ &= \alpha_{eq}\mathbb{E}\{g_{lk}^* \hat{g}_{lk}\},\end{aligned} \tag{3.39}$$

where the second term vanishes because g_{lk}^* is uncorrelated with d_{eq}. This follows because $\mathbb{E}\{\hat{g}_{lk} d_{eq}\} = 0$ and our use of a linear MMSE estimator means that the estimation error prior to quantization is also uncorrelated with $\hat{g}_{lk}$ and hence also with d_{eq}. We then obtain

$$\begin{aligned}\epsilon_{lk}^{eq} &= \mathbb{E}\{|g_{lk}|^2\} + \lambda_{eq}\mathbb{E}\{|\hat{g}_{lk}|^2\} - 2\alpha_{eq}\,\mathrm{Re}\{\mathbb{E}\{g_{lk}^* \hat{g}_{lk}\}\}. \\ &= \mathbb{E}\{|g_{lk}|^2\} + \lambda_{eq}\gamma_{lk} - 2\alpha_{eq}\gamma_{lk} \\ &= \beta_{lk} - (2\alpha_{eq} - \lambda_{eq})\gamma_{lk},\end{aligned} \tag{3.40}$$

which completes the proof. ■

The MSE expression in Lemma 3.4.1 should hold for general scalar quantization applied at the APs. We can observe that the EQ acquisition strategy achieves its minimum error when the term $(2\alpha_{eq} - \lambda_{eq})$ is maximized. Compared to the CSI acquisition with perfect fronthaul given in (3.34), the maximum is achieved at value 1. For uniform scalar quantization we can substitute the term α_q and λ_q with the expression in Proposition

3.3.1, and find the optimum Δ numerically as shown in Figure 3.3. As a result, the same Δ should be chosen as given in the Table 3.1.

In the practical implementation of this scheme the channel estimation does not have to be performed at low resolution: the channel can be estimated at the AP at high precision, in the same way as CSI quantization in the coordinated multipoint (CoMP) scenario, and the estimate subsequently quantized at a lower resolution, in order to reduce the fronthaul load. If we prefer to implement a low-cost AP, the drawback of this two-stage quantization remains however the additional complexity and power consumption. In the next subsection we introduce a simpler strategy where the AP does not require to estimate the channel and needs only a single-stage low-resolution quantization.

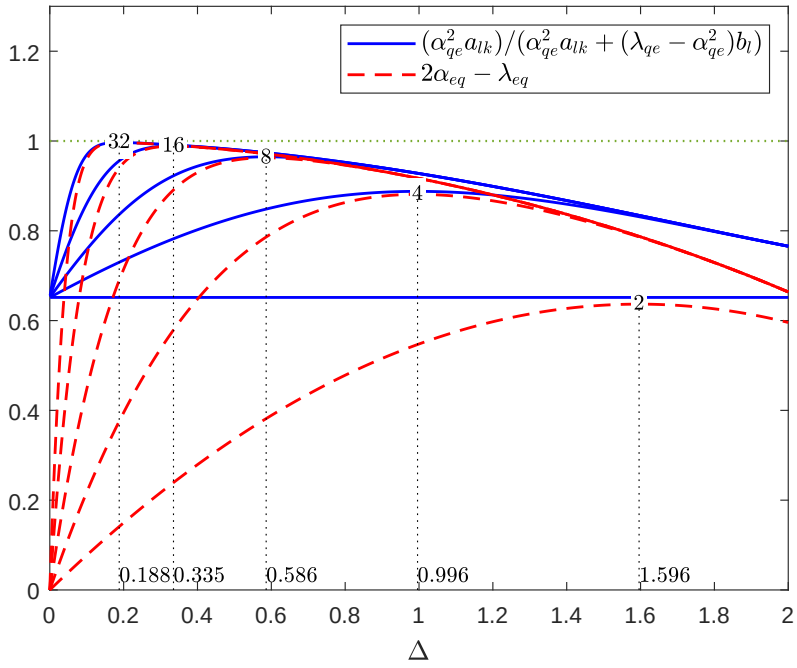


Fig. 3.3 The scaling term specifying the CSI accuracy of EQ strategy $(2\alpha_{eq} - \lambda_{eq})$ and QE strategy $(\alpha_{qe}^2 a_{lk})/(\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2) b_l)$ in relation to the step size Δ for different quantization level $S \in \{2, 4, 8, 16, 32\}$

3.4.2 Quantize-and-Estimate

Unlike the previous scheme, here we quantize the received pilot first and then send it to the CPU to estimate g_{lk}. In this case, at the l-th AP we quantize the signal $\mathbf{y}_{p,l}$ from equation (4.32), that is the superposition of pilot sequences received from K users. After sending its quantized representation through the fronthaul, we obtain at the CPU

the quantized received pilots which once again may be decomposed using the Bussgang decomposition as

$$\mathbf{y}_{p,l}^{q} = Q(\mathbf{y}_{p,l}) = \alpha_{qe}\mathbf{y}_{p,l} + \mathbf{d}_{qe}. \tag{3.41}$$

We aim next to estimate the channel at the CPU from this noisy quantized observation $\mathbf{y}_{p,l}^{q}$ for $l = 1, \ldots, L$. Since we assume that there is no channel correlation between the APs, we can estimate from each AP separately without any performance loss. Thus, we do a projection of $\mathbf{y}_{p,l}^{q}$ onto $\boldsymbol{\varphi}_k^H$ which gives

$$\begin{aligned} r_{p,lk}^{q} &= \boldsymbol{\varphi}_k^H \mathbf{y}_{p,l}^{q} \\ &= \alpha_{qe}\boldsymbol{\varphi}_k^H \mathbf{y}_{p,l} + \boldsymbol{\varphi}_k^H \mathbf{d}_{qe} \\ &= \alpha_{qe} r_{p,lk} + \boldsymbol{\varphi}_k^H \mathbf{d}_{qe}. \end{aligned} \tag{3.42}$$

We then apply the LMMSE estimator to obtain the quantize-and-estimate channel coefficient $\hat{g}_{lk}^{qe}$ given by

$$\hat{g}_{lk}^{qe} = c_{lk}^{qe} r_{p,lk}^{q}, \tag{3.43}$$

where we choose c_{lk}^{qe} that minimizes the MSE.

Lemma 3.4.2 (The MSE of Quantize-and-Estimate)**.**
Suppose that the CSI for a centralized cell-free massive MIMO with single antenna AP is acquired using the Quantize-and-Estimate strategy under a low-resolution fronthaul constraint. The optimum coefficient c_{lk}^{qe} for the LMMSE estimator (3.43) is given by

$$c_{lk}^{qe} = c_{lk} \frac{\alpha_{qe} a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2) b_l}, \quad \textit{where} \tag{3.44}$$

$$a_{lk} \triangleq \tau_p \rho_p \sum_{k'=1}^{K} \beta_{lk'} |\boldsymbol{\varphi}_k^H \boldsymbol{\varphi}_{k'}|^2 + 1, \quad \textit{and } b_l \triangleq \rho_p \sum_{k=1}^{K} \beta_{lk} + 1.$$

The MSE of the CSI at the CPU is given by

$$\epsilon_{lk}^{qe} = \beta_{lk} - \left(\frac{\alpha_{qe}^2 a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2) b_l} \right) \gamma_{lk}. \tag{3.45}$$

Proof. For the estimator in (3.43), the MSE is given by

$$\epsilon_{lk}^{qe} = \mathbb{E}\{|g_{lk} - \hat{g}_{lk}^{qe}|^2\}$$

$$= \mathbb{E}\{|g_{lk}|^2\} + c_{lk}^{qe^2}\mathbb{E}\{|\hat{g}_{lk}|^2\} - 2c_{lk}^{qe}\,\mathrm{Re}\{\mathbb{E}\{g_{lk}^*\hat{g}_{lk}\}\}. \tag{3.46}$$

We obtain from

$$0 = \frac{\partial \epsilon_{lk}^{qe}}{\partial c_{lk}^{qe}} = 2c_{lk}^{qe}\mathbb{E}\{|r_{p,lk}^q|^2\} - 2\,\mathrm{Re}\{\mathbb{E}\{r_{p,lk}^{q^*}g_{lk}\}\} \tag{3.47}$$

the coefficient c_{lk}^{qe} that minimizes the MSE given by

$$c_{lk}^{qe} = \frac{\mathrm{Re}\{\mathbb{E}\{r_{p,lk}^{q^*}g_{lk}\}\}}{\mathbb{E}\{|r_{p,lk}^q|^2\}}. \tag{3.48}$$

Using (3.42) we can express the numerator of c_{lk}^{qe} as

$$\begin{aligned}\mathrm{Re}\{\mathbb{E}\{r_{p,lk}^{q^*}g_{lk}\}\} &= \alpha_{qe}\,\mathrm{Re}\{\mathbb{E}\{r_{p,lk}^*g_{lk}\}\} + \mathrm{Re}\{\mathbb{E}\{\boldsymbol{\varphi}_k^H\mathbf{d}_{qe}g_{lk}\}\} \\ &= \alpha_{qe}\sqrt{\tau_p\rho_p}\beta_{lk},\end{aligned} \tag{3.49}$$

where the second term vanishes due to uncorrelation. Likewise we can express the denominator as

$$\mathbb{E}\{|r_{p,lk}^q|^2\} = \alpha_{qe}^2\mathbb{E}\{|r_{p,lk}|^2\} + \mathbb{E}\{|\boldsymbol{\varphi}_k^H\mathbf{d}_{qe}|^2\}, \tag{3.50}$$

where the first term is given by

$$\alpha_{qe}^2\mathbb{E}\{|r_{p,lk}|^2\} = \alpha_{qe}^2\left(\tau_p\rho_p\sum_{k'=1}^{K}\beta_{mk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2 + 1\right) \tag{3.51}$$

and the second term is given by

$$\begin{aligned}\mathbb{E}\{|\boldsymbol{\varphi}_k^H\mathbf{d}_{qe}|^2\} &= \|\boldsymbol{\varphi}_k^H\|^2\mathbb{E}\{|\mathbf{d}_{qe}|^2\} \\ &\overset{(3.10)}{=} (\lambda_{qe} - \alpha_{qe}^2)\mathbb{E}\{|\mathbf{y}_{p,l}|^2\} \\ &= (\lambda_{qe} - \alpha_{qe}^2)\left(\rho_p\sum_{k=1}^{K}\beta_{lk} + 1\right).\end{aligned} \tag{3.52}$$

Let a_{lk} and b_l denote the following expressions

$$a_{lk} \triangleq \tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2 + 1, \text{ and } b_l \triangleq \rho_p\sum_{k=1}^{K}\beta_{lk} + 1,$$

then we obtain

$$\begin{aligned} c_{lk}^{qe} &= \frac{\alpha_{qe}\sqrt{\tau_p\rho_p}\beta_{lk}}{\alpha_{qe}a_{lk}} \frac{\alpha_{qe}a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2)b_l} \\ &= c_{lk} \frac{\alpha_{qe}a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2)b_l}. \end{aligned} \tag{3.53}$$

Further, we obtain from substituting c_{lk}^{qe} into (3.46) the MSE given by

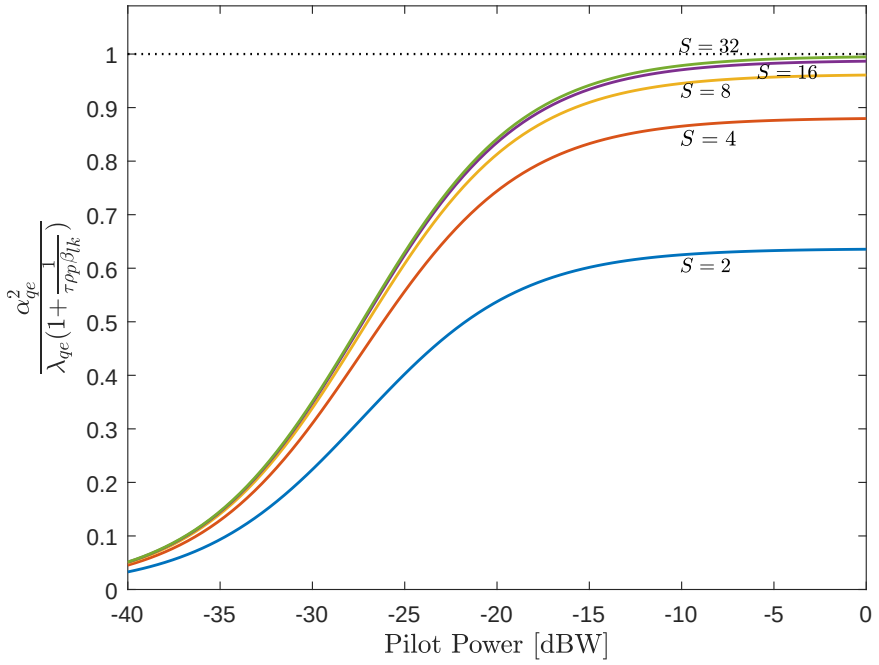
$$\begin{aligned} \epsilon_{lk}^{qe} &= \mathbb{E}\{|g_{lk}|^2\} - \frac{(\mathbb{E}\{r_{p,lk}^{q^*} g_{lk}\})^2}{\mathbb{E}\{|r_{p,lk}^{q}|^2\}} \\ &= \beta_{lk} - \frac{\alpha_{qe}^2 \tau\rho_p \beta_{lk}^2}{\alpha_{qe}^2 a_{lk}} \frac{\alpha_{qe}^2 a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2)b_l} \\ &= \beta_{lk} - \gamma_{lk} \frac{\alpha_{qe}^2 a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2)b_l}, \end{aligned} \tag{3.54}$$

which completes the proof. ■

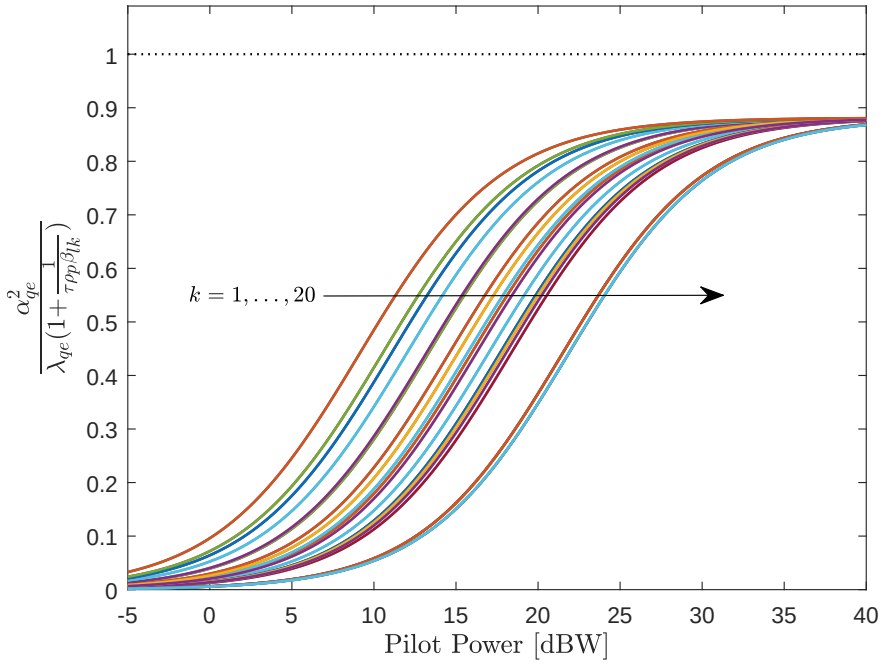
In contrast to the scaling term $(2\alpha_{eq} - \lambda_{eq})$ of the EQ strategy in (3.36), the scaling term of the QE strategy in (3.45) depends on many variables. In addition to the step size Δ, this term depends also on the family of pilot sequences, pilot power, pilot length, large scale fading, and can have a different value for different users and APs. However, it is the only term appearing in the MSE expression that depends on the step size Δ. To see how it behaves in response to varying step size Δ we plot the scaling term $(\alpha_{qe}^2 a_{lk})/(\alpha_{qe}^2 a_{lk} + (\lambda_{qe} - \alpha_{qe}^2)b_l)$ in Figure 3.3 for a certain case with given parameters. From Figure 3.3 at least we can observe that the values of Δ given in Table 3.1 are also a good choice for this strategy.

We note that the MSE expression for the QE strategy in Lemma 3.4.2 is valid for the general case of scalar quantization and pilot sequences. We did not put any any constraint on the pilot sequences where they can be orthogonal or non-orthogonal. Lemma 3.4.2 should also be valid whether the available pilot sequences are underused or overused. For the special case of fully-loaded orthogonal pilots, we can simplify the MSE expression as given in proposition 3.4.1. This assumption is commonly used in the scenario where no pilot contamination is present.

Proposition 3.4.1 (The MSE of QE for Fully-Loaded Orthogonal Pilot)**.**
If a unique pilot sequence is assigned to each user and all available orthogonal pilot sequences are used (fully loaded), then the MSE at the CPU for the QE acquisition



(a) For different quantization level $S = \{2, 4, 8, 16, 32\}$.



(b) For different users with $K = 20$ and $S = 4$.

Fig. 3.4 The behaviour of the term Γ_{lk} with respect to the pilot power.

strategy in Lemma 3.4.2 is given by

$$\epsilon_{lk}^{qe} = \beta_{lk}\left(1 - \frac{\alpha_{qe}^2}{\lambda_{qe}\left(1+\frac{1}{\tau_p\rho_p\beta_{lk}}\right)}\right). \tag{3.55}$$

Proof. Taking β_{lk} outside, we obtain the expression

$$\epsilon_{lk}^{qe} = \beta_{lk}\left(1 - \underbrace{\frac{\gamma_{lk}}{\beta_{lk}}\frac{\alpha_{qe}^2 a_{lk}}{\alpha_{qe}^2 a_{lk} + (\lambda_{qe}-\alpha_{qe}^2)b_l}}_{\Gamma_{lk}}\right). \tag{3.56}$$

Substituting a_{lk}, b_l and γ_{lk} from (3.33), the term Γ_{lk} is then given by

$$\begin{aligned}
\Gamma_{lk} &= \frac{\frac{\tau_p\rho_p\beta_{lk}^2}{\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+1}}{\beta_{lk}}\frac{\alpha_{qe}^2\left(\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+1\right)}{\alpha_{qe}^2\left(\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+1\right)+(\lambda_{qe}-\alpha_{qe}^2)\left(\rho_p\sum_{k=1}^{K}\beta_{lk}+1\right)} \\
&= \frac{\alpha_{qe}^2\tau_p\rho_p\beta_{lk}}{\alpha_{qe}^2\left(\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+1\right)+(\lambda_{qe}-\alpha_{qe}^2)\left(\rho_p\sum_{k=1}^{K}\beta_{lk}+1\right)} \\
&= \frac{\alpha_{qe}^2\tau_p\rho_p\beta_{lk}}{\alpha_{qe}^2\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+\lambda_{qe}\rho_p\sum_{k=1}^{K}\beta_{lk}+\lambda_{qe}-\alpha_{qe}^2\rho_p\sum_{k=1}^{K}\beta_{lk}} \\
&= \frac{1}{\sum_{k'=1}^{K}\frac{\beta_{lk'}}{\beta_{lk}}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+\frac{\lambda_{qe}}{\alpha_{qe}^2}\frac{K}{\tau_p}+\frac{\lambda_{qe}}{\alpha_{qe}^2}\frac{1}{\tau_p\rho_p\beta_{lk}}-\frac{K}{\tau_p}}.
\end{aligned} \tag{3.57}$$

Due to the orthogonality and fully loaded case $K = \tau_p$, we obtain

$$\begin{aligned}
\Gamma_{lk} = \frac{1}{\underbrace{\sum_{k'=1}^{K}\frac{\beta_{lk'}}{\beta_{lk}}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2}_{1}+\frac{\lambda_{qe}}{\alpha_{qe}^2}+\frac{\lambda_{qe}}{\alpha_{qe}^2}\frac{1}{\tau_p\rho_p\beta_{lk}}-1} &= \frac{1}{\frac{\lambda_{qe}}{\alpha_{qe}^2}+\frac{\lambda_{qe}}{\alpha_{qe}^2}\frac{1}{\tau_p\rho_p\beta_{lk}}} \\
&= \frac{\alpha_{qe}^2}{\lambda_{qe}\left(1+\frac{1}{\tau_p\rho_p\beta_{lk}}\right)},
\end{aligned}$$

Substituting back Γ_{lk} in (3.56), we obtain

$$\epsilon_{lk}^{qe} = \beta_{lk}\left(1 - \frac{\alpha_{qe}^2}{\lambda_{qe}\left(1+\frac{1}{\tau_p\rho_p\beta_{lk}}\right)}\right), \tag{3.58}$$

which proves the proposition. ■

Expressing the MSE as in Proposition 3.4.1 gives us now more insight into how the MSE depends on other parameters such as the pilot power ρ_p. It is interesting to see

that the term $\alpha_{qe}^2/\lambda_{qe}$ also minimizes the MSE for the solution of maximizing problem in (3.24). The remaining parameters that can be tuned are the pilot length and pilot power. To better understand the effect of pilot power on the CSI accuracy, we show in Figure 3.4a the behaviour of the term Γ_{lk} against the pilot power ρ_p for a fixed pilot length τ_p, an arbitrarily given β_{lk} and different values of the quantization level S. As expected, the term Γ_{lk} increases as the pilot power increases for all quantization level, but it then reaches a certain limit as the pilot power goes beyond a certain level. It can be observed that the limit depends on S where the value is higher for large S. However, the power level at which Γ_{lk} reaches its limit is independent of S. Next, in Figure 3.4b we consider the CSI of $K = 20$ users each of which connects simultaneously to $L = 100$ APs with quantization level $S = 4$. We compute Γ_k for each user $k = 1, \ldots, K$ to its furthest AP using the respective minimum of β_{lk} over L APs to which it connects. As shown in Figure 3.4b, every users has the same limit of Γ_k but requires a different power level to achieve this limit. This observation suggests that each user should transmit with different pilot power to obtain a good CSI accuracy with high energy efficiency. In this respect, it might be worth applying a pilot power control.

3.5 Data Transmission

Having the CSI available from all active UEs at the CPU, we are ready now to discuss the process of data transmission in centralized cell-free massive MIMO with low-resolution fronthaul. We showcase in this section the data transmission in uplink direction where the k-th user, $k = 1, \ldots, K$, aims to send its data-bearing signal $x_{u,k} \in \mathbb{C}$ with $\mathbb{E}\{|x_{u,k}|^2\} = 1$ to the CPU with the help of L single-antenna APs. All users send their data simultaneously and the l-th AP receives them as

$$y_{u,l} = \sqrt{\rho_u} \sum_{k=1}^{K} g_{lk} x_{u,k} + w_{u,l}. \tag{3.59}$$

For a simple implementation, the AP is supposed to be oblivious in the sense that it does not know the required beamforming vector and the codebook of the users. Consequently, the AP can not detect nor decode the data signal. This is in contrast to the original cell-free massive MIMO in Section 2.1.2 where the estimate of data signal $\hat{x}_{u,k}$ is detected from the received signal $y_{u,l}$ locally at the AP and then the data is decoded at the CPU. Instead of detecting the data signal, we only coarsely quantize $y_{u,l}$ at the AP to be represented by a few bits and then sent to the CPU. Specifically, the uplink data signal received at the l-th APs can be described after the quantization by the Bussgang

decomposition as

$$r_{u,l} = Q(y_{u,l}) = \alpha_{qu} y_{u,l} + d_u. \tag{3.60}$$

To detect jointly the reconstruction value $r_{u,l}$ from all L APs it is convenient to put them together as a vector $\mathbf{r}_u \in \mathbb{C}^L$ such that from (3.60) we can write

$$\mathbf{r}_u = Q(\mathbf{y}_u) = \alpha_{qu} \mathbf{y}_u + \mathbf{d}_u, \tag{3.61}$$

where

$$\mathbf{y}_u = \sqrt{\rho_u} \mathbf{G} \mathbf{x}_u + \mathbf{w}_u, \tag{3.62}$$

$\mathbf{x}_u \in \mathbb{C}^K$ is the transmitted data from all K users, $\mathbf{G}$ is the channel matrix defined in (3.2) and $\mathbf{w}_u \sim \mathcal{CN}(0, \mathbf{I}_L)$ is an additive noise vector. Note that the quantization in equation (3.61) should be understood as an elementwise operation. Further, by plugging (3.62) into (3.61) the reconstructed signal $\mathbf{r}_u$ can be expressed as

$$\begin{aligned} \mathbf{r}_u &= \sqrt{\rho_u} \alpha_{qu} \mathbf{G} \mathbf{x}_u + \alpha_{qu} \mathbf{w}_u + \mathbf{d}_u \\ &= \sqrt{\rho_u} \alpha_{qu} \hat{\mathbf{G}} \mathbf{x}_u + \sqrt{\rho_u} \alpha_{qu} \tilde{\mathbf{G}} \mathbf{x}_u + \alpha_{qu} \mathbf{w}_u + \mathbf{d}_u, \end{aligned} \tag{3.63}$$

where $\tilde{\mathbf{G}}$ is the channel estimation error including the quantization error. In this case, we have the relation

$$\mathbf{G} = \hat{\mathbf{G}} + \tilde{\mathbf{G}}, \text{ where } [\hat{\mathbf{G}}]_{lk} = \hat{g}_{lk}^{eq} \text{ or } [\hat{\mathbf{G}}]_{lk} = \hat{g}_{lk}^{qe} \tag{3.64}$$

from (3.35) and (3.43) depending on the used CSI acquisition strategy. Further, we treat $\hat{\mathbf{G}}$ as the true channel and treat the second term and so forth of equation (3.63) as an effective noise $\mathbf{z}$ such that we can write it as

$$\mathbf{r}_u = \sqrt{\rho_u} \alpha_{qu} \hat{\mathbf{G}} \mathbf{x}_u + \mathbf{z} \tag{3.65}$$

To detect the transmitted data we may use a linear detection matrix $\mathbf{A}$ that is constructed from the channel estimate $\hat{\mathbf{G}}$. The following options are now feasible

$$\mathbf{A} = \begin{cases} \hat{\mathbf{G}} & \text{MRC} \\ \hat{\mathbf{G}} (\hat{\mathbf{G}}^H \hat{\mathbf{G}})^{-1} & \text{ZF} \\ \hat{\mathbf{G}} (\hat{\mathbf{G}}^H \hat{\mathbf{G}} + \frac{1}{\rho_u} \mathbf{I}_k)^{-1} & \text{MMSE.} \end{cases} \tag{3.66}$$

where ZF and MMSE detection may be preferable due to its ability to suppress interference. Using the detection matrix given in (3.66) we then obtain the estimated data as

$$\begin{aligned}\hat{\mathbf{x}}_u &= \mathbf{A}^H \mathbf{r}_u \\ &= \sqrt{\rho_u}\alpha_{qu}\mathbf{A}^H\hat{\mathbf{G}}\mathbf{x}_u + \mathbf{A}^H\mathbf{z}.\end{aligned} \tag{3.67}$$

Subsequently, we suppose that the uplink data is sent to the CPU in the same time frame as the CSI is transferred. To maximize the rate, the same proportion of power can be allocated to the pilot and the data as in the training-based scheme of general MIMO system [57]. Let ρ and ρ_u denote the total transmit power and the transmit power for the uplink data respectively, the power allocation for pilot of length τ_p and for data of length τ_u follows

$$\rho_u\tau_u = \frac{\rho\tau_c}{2} \text{ and } \rho_p\tau_p = \frac{\rho\tau_c}{2}, \text{ where } \tau_c = \tau_p + \tau_u. \tag{3.68}$$

3.6 Achievable Rate

In this section, we now try to determine what data rate can be achieved by the centralized cell-free massive MIMO when the fronthaul resolution is low. In particular, we derive the expression of the achievable rates when the CPU performs ZF detection. The choice of ZF is usually made for the reason of a good trade-off between performance and complexity. Using the ZF detection matrix given in (3.66) the estimated data in the equation (3.67) is simplified to

$$\hat{\mathbf{x}}_u = \sqrt{\rho_u}\alpha_{qu}\mathbf{x}_u + \mathbf{A}^H\mathbf{z}, \tag{3.69}$$

since $(\hat{\mathbf{G}}^H\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H\hat{\mathbf{G}} = \mathbf{I}_K$. The SINR for ZF is then given by

$$\text{SINR}^{\text{ZF}} = \frac{\rho_u\alpha_{qu}^2}{\mathbb{E}\{|\mathbf{A}^H\mathbf{z}|^2\}}. \tag{3.70}$$

By substituting the effective noise $\mathbf{z}$ with the three last terms in (3.63), the denominator of (3.70) can be written as

$$\begin{aligned}\mathbb{E}\{|\mathbf{A}^H\mathbf{z}|^2\} &= \mathbb{E}\{\mathbf{A}^H\mathbf{z}\mathbf{z}^H\mathbf{A}\} \\ &= \mathbb{E}\{\mathbf{A}^H(\sqrt{\rho_u}\alpha_{qu}\tilde{\mathbf{G}}\mathbf{x}_u + \alpha_{qu}\mathbf{w}_u + \mathbf{d}_u)(\sqrt{\rho_u}\alpha_{qu}\mathbf{x}_u^H\tilde{\mathbf{G}}^H + \alpha_{qu}\mathbf{w}_u^H + \mathbf{d}_u^H)\mathbf{A}\} \\ &= \rho_u\alpha_{qu}^2\mathbb{E}\{\mathbf{A}^H\tilde{\mathbf{G}}\underbrace{\mathbf{x}_u\mathbf{x}_u^H}_{\mathbf{I}_K}\tilde{\mathbf{G}}^H\mathbf{A}\} + \alpha_{qu}^2\mathbb{E}\{\mathbf{A}^H\underbrace{\mathbf{w}_u\mathbf{w}_u^H}_{\mathbf{I}_L}\mathbf{A}\} + \mathbb{E}\{\mathbf{A}^H\mathbf{d}_{qu}\mathbf{d}_{qu}^H\mathbf{A}\}\end{aligned}$$

$$= \rho_u \alpha_{qu}^2 \mathbb{E}\{\mathbf{A}^H \left[\sum_{k=1}^{K} \tilde{\mathbf{g}}_k \tilde{\mathbf{g}}_k^H\right] \mathbf{A}\} + \alpha_{qu}^2 \mathbb{E}\{\mathbf{A}^H \mathbf{A}\} + \sigma_{d_u}^2 \mathbb{E}\{\mathbf{A}^H \mathbf{A}\}, \tag{3.71}$$

where for ZF

$$\begin{aligned}\mathbb{E}\{\mathbf{A}^H \mathbf{A}\} &= \mathbb{E}\{\underbrace{(\hat{\mathbf{G}}^H \hat{\mathbf{G}})^{-1} \hat{\mathbf{G}}^H \hat{\mathbf{G}}}_{\mathbf{I}_K} (\hat{\mathbf{G}}^H \hat{\mathbf{G}})^{-1}\} \\ &= \mathbb{E}\{(\hat{\mathbf{G}}^H \hat{\mathbf{G}})^{-1}\}. \end{aligned} \tag{3.72}$$

We can further express the denominator as

$$\mathbb{E}\{|\mathbf{A}^H \mathbf{z}|^2\} = \rho_u \alpha_{qu}^2 \mathbb{E}\{\mathbf{A}^H \left[\sum_{k=1}^{K} \tilde{\mathbf{g}}_k \tilde{\mathbf{g}}_k^H\right] \mathbf{A}^H\} + (\alpha_{qu}^2 + \sigma_{d_u}^2) \mathbb{E}\{(\hat{\mathbf{G}}^H \hat{\mathbf{G}})^{-1}\} \tag{3.73}$$

such that the SINR for the k-th user is given by

$$\mathrm{SINR}_k^{\mathrm{ZF}} = \frac{\rho_u \alpha_{qu}^2}{\left[\rho_u \alpha_{qu}^2 \mathbb{E}\{\mathbf{A}^H \left[\sum_{k=1}^{K} \tilde{\mathbf{g}}_k \tilde{\mathbf{g}}_k^H\right] \mathbf{A}^H\} + (\alpha_{qu}^2 + \sigma_{d_u}^2) \mathbb{E}\{(\hat{\mathbf{G}}^H \hat{\mathbf{G}})^{-1}\}\right]_{k,k}}. \tag{3.74}$$

However, due to the nature of the matrix $\mathbf{G}$ in the case of distributed massive MIMO, which tends to have independent large scale fading coefficients, the closed form SINR expression for ZF is difficult to deal with. To obtain a rather simpler SINR expression for our quantized CF massive MIMO we follow the approximation derived in [58]. We then use the SINR expression to formulate the ergodic achievable rate which is the lower bound of the ergodic capacity of discrete memoryless interference channel [see 33, Corollary 1.3]. The resulting achievable rates for the k-th user using this approximation is summarized in Proposition 3.6.1. The the ergodic rate in (3.75) is based on the assumption that the channel fading is a stationary and ergodic random process. This implies that we may use a coherence time as a set of channel realization to represent the whole process [59]. In this case, we perform the coding with a fixed codebook and take the average in (3.75) over different coherence time.

Proposition 3.6.1 (The achievable rate of Zero-Forcing)**.**
Suppose that K users are served by L single-antenna APs in uplink Cell-Free massive MIMO. The ergodic achievable rate of the k-th user with Zero-Forcing and low-resolution fronthaul is determined by

$$R_{u,k}^{\mathrm{ZF}} = \mathbb{E}\left\{\log_2\left(1 + \mathrm{SINR}_k^{\mathrm{ZF}}\right)\right\}. \tag{3.75}$$

The $\mathrm{SINR}_k^{\mathrm{ZF}}$ can be approximated by

$$\mathrm{SINR}_k^{\mathrm{ZF}} \approx \rho_u \alpha_{qu}^2 \left(\frac{L - K + 1}{L}\right) \hat{\mathbf{g}}_k^H \mathbf{\Lambda}^{-1} \hat{\mathbf{g}}_k \tag{3.76}$$

with L dimensional matrix

$$\mathbf{\Lambda} = \text{diag}\{\Lambda_1, ..., \Lambda_L\} \tag{3.77}$$

where

$$\Lambda_l = \sigma_{d_u}^2 + \alpha_{qu}^2\sigma_n^2 + \rho_u\alpha_{qu}^2\sum_{k=1}^{K}\epsilon_{lk}^q. \tag{3.78}$$

The data distortion variance caused by quantization and the noise variance are denoted respectively by $\sigma_{d_u}^2 = (\lambda_{qu} - \alpha_{qu}^2)\sigma_y^2$ *and* σ_n^2*. The estimation error is given by* $\epsilon_{lk}^q \in \{\epsilon_{lk}^{eq}, \epsilon_{lk}^{qe}\}$ *depending on the CSI acquisition scheme.*

Proof. The main notion is to involve the effective noise $\mathbf{z}$ in the ZF detector matrix. In this case, we use

$$\bar{\mathbf{A}}^H = (\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1/2}, \text{ where} \tag{3.79}$$

$$\mathbf{\Lambda} = \mathbb{E}\{\mathbf{z}\mathbf{z}^H\} \text{ and } \bar{\mathbf{A}}^H\mathbf{\Lambda}^{-1/2}\hat{\mathbf{G}} = \mathbf{I}_K. \tag{3.80}$$

Due to the independent realization of the additive noise and estimation error at each AP we may assume that the effective noise $\mathbf{z}$ is uncorrelated over L APs. Thus, we can express $\mathbf{\Lambda}$ as

$$\begin{aligned}
\mathbf{\Lambda} &= \mathbb{E}\{\mathbf{z}\mathbf{z}^H\} \\
&= \mathbb{E}\{(\sqrt{\rho_u}\alpha_{qu}\tilde{\mathbf{G}}\mathbf{x}_u + \alpha_{qu}\mathbf{w}_u + \mathbf{d}_u)(\sqrt{\rho_u}\alpha_{qu}\mathbf{x}_u^H\tilde{\mathbf{G}}^H + \alpha_{qu}\mathbf{w}_u^H + \mathbf{d}_u^H)\} \\
&= \rho_u\alpha_{qu}^2\mathbb{E}\{\tilde{\mathbf{G}}\mathbf{x}_u\mathbf{x}_u^H\tilde{\mathbf{G}}^H\} + \alpha_{qu}^2\mathbb{E}\{\mathbf{w}_u\mathbf{w}_u^H\} + \mathbb{E}\{\mathbf{d}_u\mathbf{d}_u^H\} \\
&= \rho_u\alpha_{qu}^2\sum_{k=1}^{K}\mathbb{E}\{\tilde{\mathbf{g}}_k\tilde{\mathbf{g}}_k^H\} + \alpha_{qu}^2\sigma_n^2\mathbf{I}_L + \sigma_{d_u}^2\mathbf{I}_L \\
&= \text{diag}\{\Lambda_1, ..., \Lambda_L\}
\end{aligned} \tag{3.81}$$

where

$$\Lambda_l = \sigma_{d_u}^2 + \alpha_{qu}^2\sigma_n^2 + \rho_u\alpha_{qu}^2\sum_{k=1}^{K}\epsilon_{lk}^q,$$

$\sigma_{d_u}^2$ is the distortion variance resulting from quantizing data, σ_n^2 is the noise variance and $\epsilon_{lk}^q \in \{\epsilon_{lk}^{eq}, \epsilon_{lk}^{qe}\}$ is the estimation error from (3.36) or (3.45) depending on the scheme. To detect the data signal, we apply first a filter $\mathbf{\Lambda}^{-1/2}$ to $\mathbf{r}_d$ to whiten $\mathbf{z}$. After the detection we obtain

$$\begin{aligned}
\hat{\mathbf{x}}_d &= \bar{\mathbf{A}}^H\mathbf{\Lambda}^{-1/2}\mathbf{r}_d \\
&= \sqrt{\rho_u}\alpha_{qu}\mathbf{x}_d + (\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}^H\mathbf{\Lambda}^{-1}\mathbf{z}.
\end{aligned} \tag{3.82}$$

The instantaneous SINR (i.e. the SINR for a specific realization of $\mathbf{z}$) for the k-th user can then be expressed as

$$\mathrm{SINR}_k^{\mathrm{ZF}} = \frac{\rho_u \alpha_{qu}^2}{\left[(\hat{\mathbf{G}}^H \boldsymbol{\Lambda}^{-1} \hat{\mathbf{G}})^{-1} \hat{\mathbf{G}}^H \boldsymbol{\Lambda}^{-1} \mathbf{z}\mathbf{z}^H \boldsymbol{\Lambda}^{-1} \hat{\mathbf{G}} (\hat{\mathbf{G}}^H \boldsymbol{\Lambda}^{-1} \hat{\mathbf{G}})^{-1}\right]_{k,k}}. \tag{3.83}$$

Next, we may approximate $\mathbf{z}\mathbf{z}^H$ in (3.83) by its expectation $\boldsymbol{\Lambda}$ such that it remains

$$\mathrm{SINR}_k^{\mathrm{ZF}} \overset{(3.80)}{\approx} \frac{\rho_u \alpha_{qu}^2}{\left[(\hat{\mathbf{G}}^H \boldsymbol{\Lambda}^{-1} \hat{\mathbf{G}})^{-1}\right]_{k,k}} \tag{3.84}$$

In this way, we can express the SINR_k^{ZF} as [58]

$$\mathrm{SINR}_k^{\mathrm{ZF}} \approx \rho_u \alpha_{qu}^2 \left(\frac{L-K+1}{L}\right) \hat{g}_k^H \boldsymbol{\Lambda}^{-1} \hat{g}_k. \tag{3.85}$$

This completes the proof of the proposition. ■

3.7 Scalability

In this section we discuss now how the AP processing and the fronthaul load of the centralized cell free massive MIMO with single antenna APs are scaled with the number of users. As mentioned previously, future networks are required to handle a large number of users, preferably with low complexity processing and low resource utilization. We call such a network a scalable network referring to the recent work [60], which appeared at the time of writing this thesis. As described in [60], there are many aspects of scalability in cell-free massive MIMO. However, we discuss the most important ones which are the AP processing and the fronthaul load. To begin with, we discuss first the scalability of the original cell-free massive MIMO in the uplink in terms of AP processing and fronthaul load. Then, we compare it with the scalability of the centralized cell-free massive MIMO respectively using EQ and QE strategy described in Section 3.4.

Recall from Section 2.1.2 that in the original cell-free massive MIMO the APs perform the channel estimation and the data detection. Suppose that K users should send their data of length τ_u simultaneously in the uplink. The l-th single-antenna AP must compute first K times channel estimation to obtain $\hat{g}_{lk}$ for $k = 1, \ldots, K$ from the received pilots of length τ_p. In the simplest case of least square channel estimation the l-th AP should perform at least $K\tau_p$ complex scalar multiplications. Assuming that the pilot length τ_p is fixed and independent of the number of users K, then the AP processing for the

channel estimation grows linearly with K. When the number of users $K \to \infty$ then the complexity for the channel estimation in the original cell-free massive MIMO becomes infinite and hence unscalable. To detect the uplink data of length τ_u, the l-th AP needs to perform $K\tau_u$ complex scalar multiplication $\hat{g}_{lk}^* y_{u,l}$ in (2.31) for $k = 1, \ldots, K$. Hence, the AP processing for the data detection is also unscalable when the number of users K gets very large.

In terms of the fronthaul load, the original cell-free massive MIMO does not utilize the fronthaul for transferring the CSI since the estimated CSI has been employed locally at the AP to detect the data. The fronthaul is then utilized only for transmitting the detected payload data $\{\hat{g}_{lk}^* y_{u,l}, k = 1, \ldots, K\}$. This means that all we have to transmit over the fronthaul is the $K\tau_u$ complex data symbols. Using an ADC with resolution of $\log_2[S]$ bits we have then $2\log_2[S]K\tau_u$ bits fronthaul load. However, the fronthaul load will be unlimited when the number of users $K \to \infty$. In this case, the original cell-free massive MIMO is also unscalable in terms of the fronthaul load.

In comparison to the original cell-free massive MIMO, the complexity of the AP processing in centralized cell-free massive MIMO can be significantly reduced. This is due to the migration of the data detection to the CPU which presumably has much more computing power. Recall from (3.60) that the l-th AP needs only to forward the quantized data signal $r_{u,l}$ to the CPU over a fronthaul. Obviously, this process of forwarding has low complexity and independent to the number of users K. Since the the data detection is performed at the CPU, we need also to look at the complexity of the CSI acquisition at the CPU. If the channel estimation is performed at the AP such as in the EQ strategy described in Section 3.4.1, then the l-th AP still needs to compute at least $K\tau_p$ complex scalar multiplications. It turns out that this strategy has an unscalable AP processing for the CSI acquisition. In the opposite, if the QE strategy in Section 3.4.2 is applied, the l-th AP needs only to transmit the quantized received pilots $\mathbf{y}_{p,l}^q$ from equation (3.41). As we assumed previously, the length of the received pilots τ_p is fixed and independent of the number of users K. Therefore, we can say that the centralized cell-free massive MIMO with single-antenna AP using QE strategy is scalable in terms of the AP processing.

To determine the fronthaul load required by the centralized cell-free massive MIMO we need to take into account the load for CSI acquisition and the load for transmitting the data signal. If the EQ strategy is used, then we need to transmit τ_u received data symbols and K channel estimates $\{\hat{g}_{lk}$, for $k = 1, \ldots, K\}$ over the fronthaul. In total, the fronthaul load is $2\log_2[S](K + \tau_u)$ where an ADC with resolution $\log_2[S]$ bits is used. Indeed, the fronthaul load for this strategy is also unscalable, but for large τ_u, this still

Table 3.2 The scalability of cell-free massive MIMO with single-antenna AP

	Min. AP Processing [Scalar multiplication]	Min. Fronthaul Load [bits]
Original CF maMIMO (MRC)	$K\tau_p + K\tau_u$	$2\log_2[S]K\tau_u$
Centralized CF maMIMO (EQ)	$K\tau_p$	$2\log_2[S](K+\tau_u)$
Centralized CF maMIMO (QE)	-	$2\log_2[S](\tau_p+\tau_u)$

grows more slowly with K than $2\log_2[S]K\tau_u$ in case of the original cell-free massive MIMO. In contrast, if the QE strategy is used for centralized cell-free massive MIMO, only τ_u received data symbols and τ_p received pilot symbols should be transferred over the fronthaul. In this case, we load the fronthaul with $2\log_2[S](\tau_p+\tau_u)$ bits which is independent of the number of users K. As a result, the centralized cell-free massive MIMO with single antenna AP using the QE strategy is also scalable in terms of the fronthaul load. Table 3.2 summarizes the scalability of cell-free massive MIMO with single-antenna AP for different schemes. As shown in the Table 3.2, the multiplier $2\log_2[S]$ is one of determining factor for the increasing of the fronthaul load. If a high resolution ADC is employed, then the fronthaul load can increase very rapidly, especially for the case of the original cell free massive MIMO with a large number of users. On the other hand, a low-resolution ADC combined with centralized cell-free massive MIMO using the QE strategy seems to offer a promising solution for the scalability of the fronthaul load. This observation also suggests that it is feasible in cell-free massive MIMO to employ fronthaul links with low resolution if the centralized approach is used.

3.8 Performance Evaluation

In the following, we provide some numerical results for the centralized cell-free massive MIMO described previously. We do simulations with system parameters similar to [6] where there are $L = 200$ APs and $K = 20$ users distributed uniformly in an area of 1×1 km^2. We assume that this simulation area is wrapped around to avoid the boundary effects as shown in Figure 3.5. For the channel g_{lk} given in (4.3) we model the large scale fading $\beta_{lk} = \mathrm{PL}_{lk} \cdot 10^{(\sigma_{sh} z_{lk})/10}$, where the factor $10^{(\sigma_{sh} z_{lk})/10}$ is the uncorrelated shadowing with standard deviation $\sigma_{sh} = 8$ dB and $z_{lk} \sim \mathcal{N}(0,1)$. The path loss coefficient follows

Table 3.3 Physical parameters used for the simulation:

Area		1×1 km^2
Carrier frequency	f	1.9 GHz
Bandwidth	B	20 MHz
AP antenna height	h_{AP}	15m
UE antenna height	h_u	1.65m
Boltzmann constant	k_b	1.381×10^{-23}
Noise temperature	T_0	290 Kelvin
Noise figure		9 dB

the three-slope model according to

$$\mathrm{PL}_{lk} = \begin{cases} -\mathcal{L} - 35\log_{10}(d_{lk}), d_{lk} > d_1 \\ -\mathcal{L} - 15\log_{10}(d_1) - 20\log_{10}(d_{lk}), d_0 < d_{lk} \leq d_1 \\ -\mathcal{L} - 15\log_{10}(d_1) - 20\log_{10}(d_0), d_{lk} \leq d_0, \end{cases} \tag{3.86}$$

where d_{lk} is the distance between the l-th AP and the k-th user, $d_0 = 0.01$km, $d_1 = 0.05$km, and

$$\mathcal{L} \triangleq 46.3 + 33.9\log_{10}(f) - 13.83\log_{10}(h_{AP}) - (1.1\log_{10}(f) - 0.7)h_u + (1.56\log_{10}(f) - 0.8). \tag{3.87}$$

We choose the carrier frequency $f = 1.9$ GHz, the AP antenna height $h_{AP} = 15$m and the user antenna height $h_u = 1.65$m. In our simulation the normalized transmit SNRs ρ_u and ρ_p are defined as the transmit power divided by the noise power which is $B \times k_b \times T_0 \times$ noise figure, where we denote by B the communication bandwidth, by k_b the Boltzmann constant and by T_0 the noise temperature. The values of the physical parameters used in the simulation are summarized in Table 3.3. To make a fair comparison, our simulation considers the fully-loaded orthogonal case with $\tau_p = K$ where EQ and QE scheme spend the same length of CSI overhead. We allocate 10% of symbols for acquiring CSI where $\tau_p = 20$ symbols are spent for the pilot from overall $\tau_c = 200$.

In Figure 3.6, we first validate with simulations our analytical MSE approximations which are obtained in (3.36) and (3.45) using Bussgang decomposition. Note that in our simulation setup the large scale fading has very small value up to -17 order of magnitude. This boils down to very small channel gain and to very small typical value of MSE. It

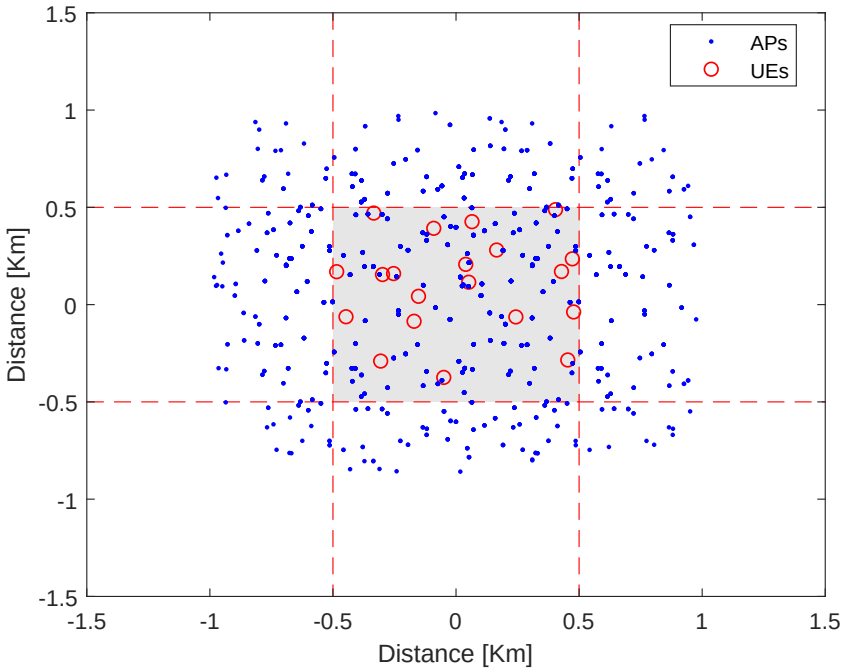


Fig. 3.5 An example of the wrap around model for the simulation of Cell-Free massive MIMO with $L = 100$ and $K = 10$.

is shown in Fig. 3.6 that our analyses for both strategies are quite close to simulations especially for small S and high transmit power. In at least 80% of cases (those with the lower MSE) the QE scheme gives a poorer MSE than EQ and for $S > 2$ this proportion increases. However, it is the larger channel estimate errors that have stronger influence on the rate.

Using the corresponding channel estimation errors we then evaluate the average achievable rates per user given in (3.75). In this case, we compare their performance in terms of their per-user net throughput defined as

$$\mathcal{T}_{u,k}^{ZF} \triangleq B\frac{1-\tau_p/\tau_c}{2}R_{u,k}^{ZF}, \tag{3.88}$$

where the CSI overhead is taken into account by the term $1 - \tau_p/\tau_c$. As shown in Fig. 3.7 the QE scheme achieves higher throughput than the EQ scheme for small S over the whole range of transmit power. The performance gap is decreasing as we increase the quantization level. For small S, the achievable rates computed by our approximation (3.76) has only relatively small deviation from the rate computed by (3.74). This result seems to disagree with the theoretical result [61] which suggests that performing estimation and then compression should be optimal. However, we note that

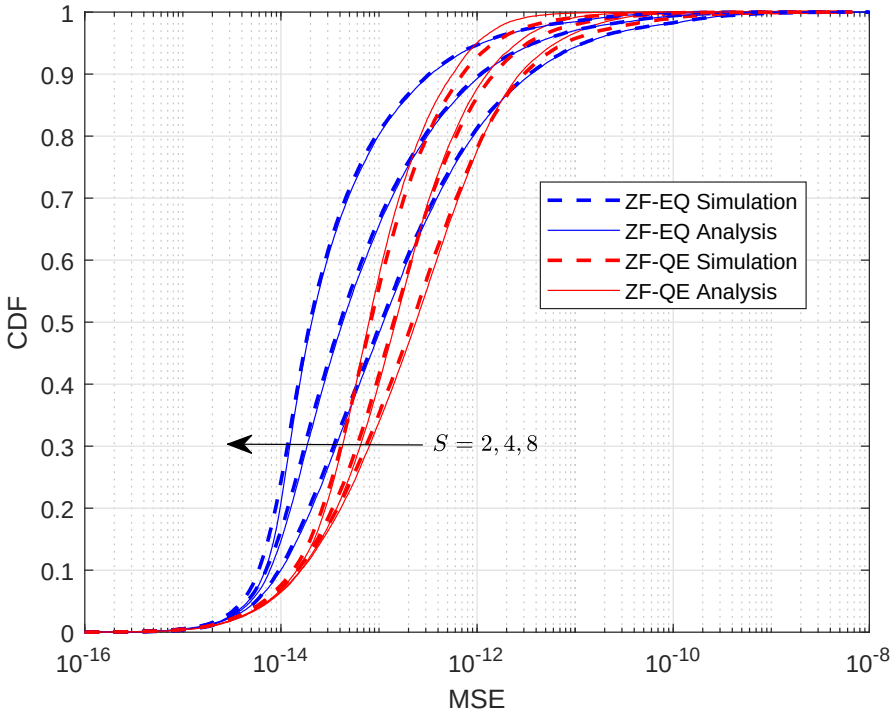


Fig. 3.6 The cumulative distribution of the channel estimation MSE ϵ_{lk}^{q} for the schemes estimate-and-quantize (EQ) in (3.36) and quantize-and-estimate (QE) in (3.45) with $K = 20$, $L = 200$ and Transmit Power $= 0$ dBW

we investigate here a rather unidealized system model, where we consider fading and take into account many interplaying parameters. Therefore, the QE scheme may be better than the EQ scheme in some specific cases. As in our simulation result, this may be caused by the nonlinearity of estimation and quantization in the case of low resolution. Further, we can also clearly observe that ZF with low quantization level $S = 4$ can already outperform MRC even with infinite quantization precision. This demonstrates the great improvement resulting from having global CSI available at the CPU. With $S = 32$ or $R = 5$ bits we are about 5 dB away from ZF with ideal fronthaul to reach 60 Mbits/s/Hz average throughput per user. Meanwhile, the trade off between the increasing throughput and the resulting latency due to CSI overhead is left for future works.

3.9 Summary

In this chapter, we have presented a centralized cell-free massive MIMO system as an alternative to original cell-free massive MIMO that uses a distributed approach. Its

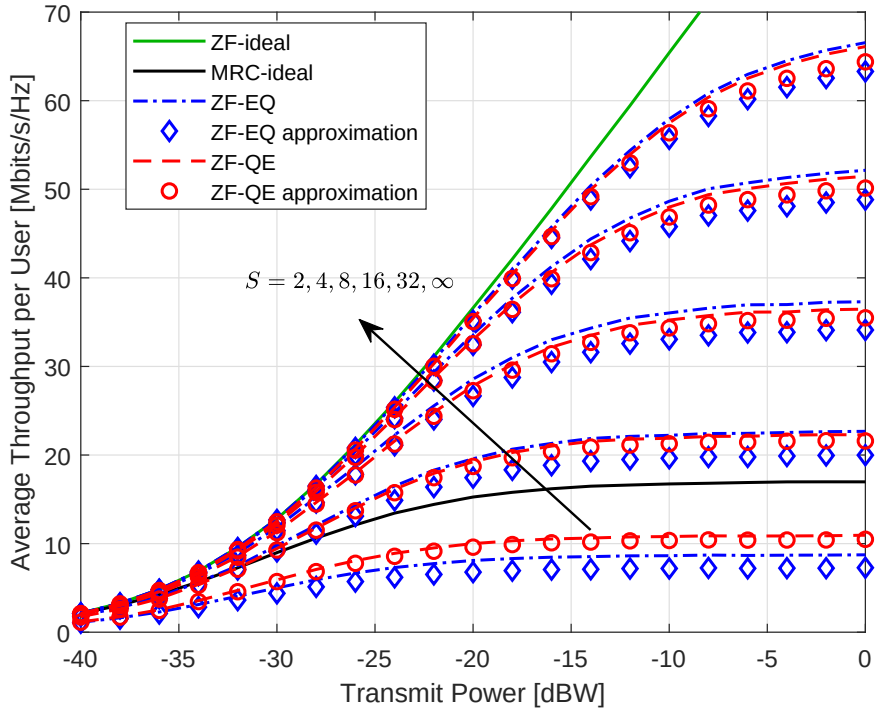


Fig. 3.7 The average per user throughput for different number of quantization level S, transmit power and CSI acquisition schemes for $K = 20$ and $L = 200$.

general concept was elaborated, where the CSI from active UEs is available at the CPU, and in contrast to original cell-free massive MIMO, the CPU utilizes the global CSI for joint data processing. We have studied the system for the case, where the APs are equipped with single-antenna and connected to the CPU by low-resolution fronthauls. Therefore, we characterized a scalar uniform quantization of low-resolution and analysed its optimality condition using Bussgang decomposition.

To enable the centralized approach under limited-capacity fronthaul links, we have developed two strategies for acquiring CSI at the CPU, which are estimate-and-quantize (EQ) and quantize-and-estimate (QE). We have analysed their performance and formulated their MSE expression. Further, we have provided the data transmission scheme in the centralized cell-free massive MIMO under low-resolution fronthauls. For a given low-resolution constraint to the CSI and the data, the achievable rate per user of the single-antenna APs centralized cell-free massive MIMO was derived in case of zero-forcing (ZF) detection. Moreover, we have also derived a more simple SINR expression for ZF with the given low-resolution constraint.

By using numerical simulation, we have evaluated its corresponding average throughput performance for various resolutions. We have also compared its throughput to the

original cell-free massive MIMO for different CSI strategies. The simulation results showed that the centralized approach using ZF with 2 bits resolution can already outperform the original cell-free massive MIMO with ideal fronthaul. Further, the low-complexity scheme ZF-QE outperforms ZF-EQ at low resolution, especially for 1-bit. The scalability issue for the centralized cell-free massive MIMO was also addressed in terms of AP processing and fronthaul load. Our investigation showed that centralized cell-free massive MIMO with single antenna AP is more scalable than the original cell-free massive MIMO, particularly when the QE strategy is used.

Chapter 4

Centralized Cell-Free massive MIMO with Multiple-Antenna Access Points

Improving wireless link capacity by adding more antenna at the receiver or transmitter end has become an appealing and well-known technique. In the context of cell-free massive MIMO, one can consider utilizing more than a single antenna at the APs and expect the same logic to apply to its system performance. As we have seen in the previous chapter, the centralized approach has also been able to improve the user throughput in the cell-free massive MIMO. To increase the throughput even further, we can thus extend the centralized cell-free massive MIMO described in the previous chapter to the case where the APs are allowed to have more than a single antenna.

The advantages of adding more antennas at the AP have been reported previously in some studies such as in [51, 62]. As investigated in [62] in the case of conjugate beamforming and zero-forcing, adding more antennas can in fact increase the throughput if the same number of APs is deployed. But if the total number of antennas in the system should remain constant, then the single-antenna scheme with more APs can outperform the multi-antenna scheme in terms of the outage rate. A further investigation is given in [51] by examining the channel hardening and favorable propagation. It is shown that those properties appear much weaker in cell-free massive MIMO with single antenna APs compared to a massive MIMO system with co-located antennas. While the favorable propagation can be improved by many factors, the channel hardening can evidently be improved by deploying multiple antennas at the APs. On the other hand, increasing the number of antennas at each AP while reducing the number of APs will decrease the macro diversity which is one of the most advantageous features in cell free massive MIMO. Therefore, there is a tension in cell free massive MIMO between deploying many single-antenna APs versus fewer multi-antenna APs. The tension will be stronger if we

include some other factors such as the fronthaul load and energy efficiency. The known results so far show that the average rate decreases from many single-antenna APs to fewer multi-antenna APs as given in [51].

In this regard, we study in this chapter centralized cell-free massive MIMO with multiple-antenna APs considering the limited capacity of the fronthaul links. At the time of writing this thesis, we found a comprehensive work independently investigating a centralized approach with multiple antenna APs and correlated channels [11] but without assuming limited-capacity fronthaul. On the other hand, the work in [9] investigated a centralized approach using multiple antenna APs with a limited-capacity fronthaul but considering uncorrelated channels and separate scalar quantization. Unlike the other works mentioned above, we consider both in this chapter where we study the centralized approach with multi-antenna APs with correlated channel and limited-capacity fronthaul.

After describing the considered system model including the spatial channel correlation model, we will consider two schemes of AP processing in the next two sections. In the first scheme, we will handle the received signals across the multi-antenna AP individually, where we use independent scalar quantization at each antenna similar to the work in [9]. Since the resulted fronthaul load from multi-antenna AP is larger than single-antenna AP, we propose in the second scheme to jointly process the received signal at the APs. By exploiting the fact that the channels across the AP's antennas are correlated, we employ Vector Quantization (VQ) with a small number of bits per dimension. After describing the quantization model, we will also describe how the CSI is acquired in the joint processing scheme and how much data rate can be achieved by this scheme. Before we evaluate its system performance in the subsequent section, we will discuss the scalability and the possibility to extend the centralized cell-free massive MIMO with multi-antenna APs into a larger network. Then, we will show using simulations that the joint processing can deliver a higher data rate than its separate processing counterpart at the same fronthaul resolution. In the opposite situation, we can set the joint processing to the same throughput performance as the individual processing, which compensates with a lower requirement of the fronthaul load. Although we have not resolved the problem of finding the optimum number of antennas and APs with limited-capacity fronthaul, the results in this chapter may give some insights into the problem.

4.1 System Model

As in the previous chapter, we consider in this chapter also a cell-free massive MIMO system for the uplink transmission, where we have K single-antenna users (UEs) intending

to send their data to the Central Processing Unit (CPU) with the help of L Access Points (APs). The processing of the signals received at the APs is virtualized at the CPU, which is connected to the L APs by L error-free fronthaul links which carry the signals in digitally encoded form. The distinction from the model in the previous chapter is that the APs are now equipped with $N \geq 1$ antennas. We fix the total number of AP antennas in the system at $M = LN$.

4.1.1 Channel Model

We denote the channel between the k-th user and the m-th antenna of the l-th AP by g_{mk} where $m = (l-1)N + 1, \ldots, lN$ for $l = 1, \ldots, L$, and $k = 1, \ldots, K$. For a given l and k the channel is specified by the $N \times 1$ vector $\mathbf{g}_{lk} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{\Sigma}_{lk})$ where $\mathbf{\Sigma}_{lk} \in \mathbb{C}^{N \times N}$ is the covariance matrix including the large scale fading and the spatial correlation given by

$$\mathbf{\Sigma}_{lk} = \beta_{lk} \mathbf{R}_{lk}. \tag{4.1}$$

The large scale fading β_{lk} is a path-loss dependent coefficient whereas the correlation matrix $\mathbf{R}_{lk} \in \mathbb{C}^{N \times N}$ is dependent on the particular environment between AP and UE. In this case, we follow the local scattering model given in [33], where any user k at the azimuth angle θ to the AP l is surrounded by scatterers causing correlation to the multipath signal components received between the antennas of the AP. This channel model is illustrated in Figure 4.1. Accordingly, the correlation coefficient can be specified by an angle of arrival $\bar{\theta}$ which is treated as a random variable with probability density function $f(\bar{\theta})$ and the entries of the correlation matrix $\mathbf{R}_{lk}$ are then determined by

$$[\mathbf{R}_{lk}]_{a,b} = \int e^{j2\pi d_H (a-b) sin(\bar{\theta})} f(\bar{\theta}) d\bar{\theta}, \tag{4.2}$$

where d_H is the spacing between antennas $1 \leq a, b \leq N$. Further, $\bar{\theta}$ can be expressed as $\bar{\theta} = \theta + \delta$, where δ is a random deviation from the nominal angle with Gaussian distribution and standard deviation σ_δ.

Using the Karhunen-Loeve representation we can describe the correlated channel vector as

$$\mathbf{g}_{lk} = \beta_{lk}^{1/2} \mathbf{U}_l \Lambda_l^{1/2} \mathbf{h}_{lk}, \tag{4.3}$$

where the vector $\mathbf{h}_{lk} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{I_N})$ models the small scale fading between the k-th user and the l-th AP. The unitary matrix $\mathbf{U} \in \mathbb{C}^{N \times r}$ and the diagonal matrix $\Lambda \in \mathbb{R}^{r \times r}$ comprise respectively the eigenvectors and the associated eigenvalues of the correlation

matrix $\mathbf{R}_{lk}$ with rank r. The channel vector of the k-th user to all L APs is then given by $\mathbf{g}_k \sim \mathcal{CN}(\mathbf{0}_M, \mathbf{\Sigma}_k)$, where $\mathbf{\Sigma}_k = \text{diag}\,(\mathbf{\Sigma}_{1k}, \ldots, \mathbf{\Sigma}_{Lk})$. Further, we stack the channel from K users to all L APs in the columns of the $M \times K$ matrix $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_K]$, such that under the assumption of perfect fronthaul the received signal at the CPU can be modeled as

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \tag{4.4}$$

where $\mathbf{x} \in \mathbb{C}^K$ is the channel input from all K users and $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}_M, \mathbf{I}_M)$ is the i.i.d. additive Gaussian white noise at APs. Later, we will remove the assumption of perfect fronthaul and assume that the l-th fronthaul link connecting the l-th AP to the CPU can transmit quantized signals reliably at a maximum rate of C_l.

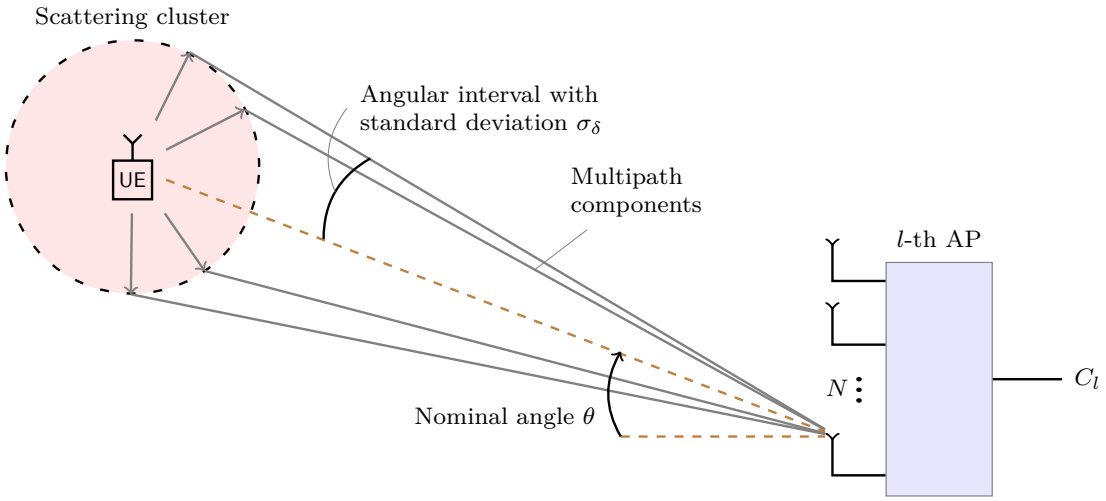


Fig. 4.1 An illustration of the local scattering model for Non Line of Sight (NLoS) channel between an UE and an AP.

4.2 Individual Processing at Multiple-Antenna APs

We investigate first in this section a simple case where the APs process the received signal at the multi-antenna individually. Although the received signals to be processed at the APs are clearly different from the case of single-antenna APs described in the previous chapter, initially in this section we will not change the processing at each AP antenna, but will assuming separate processing of the signal at each AP antenna. We will investigate this naive approach particularly for the fronthaul quantization and will later compare this with a more advanced approach that we propose. Thus, the description in

this section may serve as a baseline for comparison with the scheme we propose in the subsequent section.

4.2.1 Fronthaul Quantization

Due to the limited capacity of the fronthaul link and the high load of the digitally encoded signal we need to compress the received signal at the AP for efficient transmission to the CPU. In this section, we consider quantizing the received signals using independent scalar quantizers followed by independent fixed-rate lossless coding. For N AP antennas we apply correspondingly N scalar quantizer as illustrated in Figure 4.2. The individual scalar quantizer we use in this section follows the model described in Section 3.3.1. However, the quantized signals are sent together via the same fronthaul links. This results in an increased fronthaul load compared to the case of a single-antenna AP. To simplify our analysis, we consider fronthaul links with equal capacity of $C_l = C$ bits for all $l \in \{1, \ldots, L\}$. Thus, the m-th scalar quantizer at the N-antenna AP should set the number of quantization level smaller than the quantizer in the single-antenna AP by $S_m = 2^{R_m}$ with $R_m = \frac{C}{N}$ to meet the fronthaul capacity.

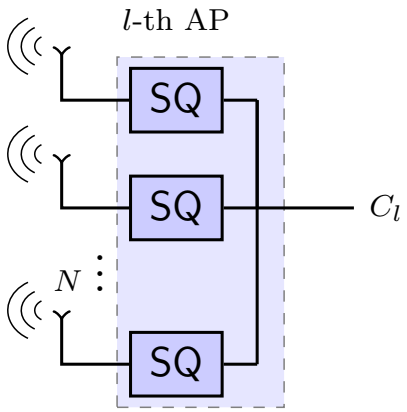


Fig. 4.2 Illustration of the individual processing using scalar quantization

4.2.2 CSI Acquisition

For the CSI acquisition we adopt the method from the previous chapter where we may apply the Estimate-and-Quantize (EQ) or Quantize-and-Estimate strategy (QE) for the respective channel between the k-th user and the m-th antenna for $m = (l-1)N + 1, \ldots, lN$. Because we have now N antennas at each AP, the l-th AP observes the

received pilot $\mathbf{Y}_{p,l} \in \mathbb{C}^{N\times\tau}$ from all K users as

$$\mathbf{Y}_{p,l} = \sqrt{\tau\rho_p}\sum_{k=1}^{K}\mathbf{g}_{lk}\boldsymbol{\varphi}_k^H + \mathbf{W}_l \tag{4.5}$$

where the k-th user sends $\sqrt{\tau}\boldsymbol{\varphi}_k$ as its pilot with transmit power ρ_p. The superposition of the pilots from all K users is corrupted at the l-th AP by an additive noise matrix $\mathbf{W}_l$ whose entries are uncorrelated with zero mean and unit variance.

Recall from Section 3.4.1 that for the EQ strategy we first estimate at the l-th AP the channel coefficient g_{mk} by projecting each row of $\mathbf{Y}_{p,l}$ onto $\boldsymbol{\varphi}_k^H$ and then weighting appropriately to minimize the MSE to obtain the estimate

$$\hat{\mathbf{g}}_{lk} = \left(\frac{\sqrt{\tau_p\rho_p}\beta_{lk}}{\tau_p\rho_p\sum_{k'=1}^{K}\beta_{lk'}|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}|^2+1}\right)\mathbf{Y}_{p,l}\boldsymbol{\varphi}_k^H \tag{4.6}$$

The estimated channel vector is then quantized separately for each element using a scalar quantizer. Since the quantization processes are carried out independently, we may decompose each element of the vector using the Bussgang decomposition and write the quantized channel estimate as

$$\hat{\mathbf{g}}_{lk}^{eq} = Q(\hat{\mathbf{g}}_{lk}) = \alpha_{eq}\hat{\mathbf{g}}_{lk} + \mathbf{d}_{eq}. \tag{4.7}$$

The quantized vectors $\{\hat{\mathbf{g}}_{lk}^{eq}$ for $k = 1,\ldots,K\}$ are then sent together via the l-th fronthaul link to be gathered at the CPU. For the QE strategy we refer to Section 3.4.2, where we apply it to each row of the received pilot matrix in (4.5). By treating the elements of the matrix $\mathbf{Y}_{p,l}$ independently the Bussgang decomposition of the quantized pilots are given by

$$\mathbf{Y}_{p,l}^{q} = Q(\mathbf{Y}_{p,l}) = \alpha_{qe}\mathbf{Y}_{p,l} + \mathbf{D}_{qe}. \tag{4.8}$$

The CPU receives $\mathbf{Y}_{p,l}^{q}$ at the output of the l-th fronthaul link, which is then projected onto $\boldsymbol{\varphi}_k^H$ and weighted by the LMMSE coefficient c_{lk}^{qe} from Lemma 3.4.2. We then obtain

$$\hat{\mathbf{g}}_{lk}^{qe} = c_{lk}^{qe}\mathbf{Y}_{p,l}^{q}\boldsymbol{\varphi}_k^H. \tag{4.9}$$

4.2.3 Data Transmission

Consider that all users send their data $x_{u,k}$ with $\mathbb{E}\{|x_{u,k}|^2\} = 1$ simultaneously, then the l-th AP receives them at the N multiple antenna as

$$\mathbf{y}_{u,l} = \sqrt{\rho_u} \sum_{k=1}^{K} \mathbf{g}_{lk} x_{u,k} + \mathbf{w}_{u,l}. \tag{4.10}$$

For a moderately large K we can assume that $\mathbf{y}_{u,l}$ is distributed as a multivariate Gaussian variable. Following the centralized approach, the l-th AP does not perform the data detection but immediately quantizes the received signal vector $\mathbf{y}_{u,l} \in \mathbb{C}^N$ as

$$\mathbf{r}_{u,l} = Q(\mathbf{y}_{u,l}) \tag{4.11}$$

$$= \alpha_{qu} \mathbf{y}_{u,l} + \mathbf{d}_{u,l}, \tag{4.12}$$

where each element of $y_{u,l}$ is quantized independently allowing a Bussgang decomposition to each element in the second equation. The CPU collects then the quantized data from L APs in a stack as a signal vector $\mathbf{r}_u$ of length LN.

$$\underbrace{\begin{bmatrix} \mathbf{r}_{u,1} \\ \vdots \\ \mathbf{r}_{u,L} \end{bmatrix}}_{\triangleq \mathbf{r}_u} = \alpha_{qu} \underbrace{\begin{bmatrix} \mathbf{y}_{u,1} \\ \vdots \\ \mathbf{y}_{u,L} \end{bmatrix}}_{\triangleq \mathbf{y}_u} + \underbrace{\begin{bmatrix} \mathbf{d}_{u,1} \\ \vdots \\ \mathbf{d}_{u,L} \end{bmatrix}}_{\triangleq \mathbf{d}_u}. \tag{4.13}$$

Although the received signal $\mathbf{r}_u$ has the same expression as in the case of the single-antenna AP, it has a rather different structure due to the received signal

$$\mathbf{y}_u = \sqrt{\rho_u} \mathbf{G} \mathbf{x}_u + \mathbf{w}_u, \tag{4.14}$$

which stems from the different characteristic of the channel $\mathbf{G}$. Expressing the channel matrix as $\mathbf{G} = \mathbf{H} \odot \mathbf{D}^{1/2}$, the large-scale fading matrix $\mathbf{D}$ has now a block structure where every N rows have the same components due to the deployment of N antennas at each AP. This will affect the capability of the linear detector matrix to suppress the interference more effectively. Similarly as given in (3.66), we can construct at the CPU a linear detector matrix $\mathbf{A}$ from the estimated channel $\hat{\mathbf{G}}$, where $[\hat{\mathbf{G}}]_{lk} = \hat{g}_{lk}^{eq}$ or $[\hat{\mathbf{G}}]_{lk} = \hat{g}_{lk}^{qe}$. By decomposing the channel $\mathbf{G} = \hat{\mathbf{G}} + \tilde{\mathbf{G}}$ as a sum of estimated channel and channel

estimation error, the estimated data can then be obtained from

$$\begin{aligned}\hat{\mathbf{x}}_u &= \mathbf{A}^H \mathbf{r}_u \\ &= \mathbf{A}^H(\sqrt{\rho_u}\alpha_{qu}\hat{\mathbf{G}}\mathbf{x}_u + \sqrt{\rho_u}\alpha_{qu}\tilde{\mathbf{G}}\mathbf{x}_u + \alpha_{qu}\mathbf{w}_u + \mathbf{d}_u) \\ &= \sqrt{\rho_u}\alpha_{qu}\mathbf{A}^H\hat{\mathbf{G}}\mathbf{x}_u + \mathbf{A}^H\mathbf{z}, \end{aligned} \tag{4.15}$$

where the effective noise $\mathbf{z}$ is given by

$$\mathbf{z} = \sqrt{\rho_u}\alpha_{qu}\tilde{\mathbf{G}}\mathbf{x}_u + \alpha_{qu}\mathbf{w}_u + \mathbf{d}_u. \tag{4.16}$$

4.2.4 Achievable Rate

Similarly to the previous chapter, we also use the capacity lower bound to specify the data rate achievable by this scheme. In this case, we can apply a formula in the same way to [9, Th. 1] and [11, Pr. 1] due to the similar setting we have in this section. The only difference is that we consider spatially correlated channels and capacity-limited fronthaul altogether. In the previous works [9, 11], only one of them was considered respectively. This then changes the value of the estimated channels, the detector matrix and the estimation errors in the SINR formula. We recall that the signal from the k-th user after the detection with a general linear detector $\mathbf{A}^H$ can be written from (4.15) as

$$\begin{aligned}\hat{x}_{u,k} = \sqrt{\rho_u}\alpha_{qu}\mathbf{a}_k^H\hat{\mathbf{g}}_k x_{u,k} + \sqrt{\rho_u}\alpha_{qu}\sum_{\substack{i=1\\i\neq k}}^{K}\mathbf{a}_k^H\hat{\mathbf{g}}_i x_{u,i} \\ + \sqrt{\rho_u}\alpha_{qu}\sum_{i=1}^{K}\mathbf{a}_k^H\tilde{\mathbf{g}}_i x_{u,i} + \alpha_{qu}\mathbf{a}_k^H\mathbf{w}_u + \mathbf{a}_k^H\mathbf{d}_u,\end{aligned} \tag{4.17}$$

where $\mathbf{a}_k$ and $\hat{\mathbf{g}}_k$ are respectively the k-th columns of matrices $\mathbf{A}$ and $\hat{\mathbf{G}}$. The first term in (4.53) is the signal from the k-th user that is desired to decode. All of the other terms can be regarded as an interference term

$$\nu_k = \sqrt{\rho_u}\alpha_{qu}\sum_{\substack{i=1\\i\neq k}}^{K}\mathbf{a}_k^H\hat{\mathbf{g}}_i x_{u,i} + \sqrt{\rho_u}\alpha_{qu}\sum_{i=1}^{K}\mathbf{a}_k^H\tilde{\mathbf{g}}_i x_{u,i} + \alpha_{qu}\mathbf{a}_k^H\mathbf{w}_u + \mathbf{a}_k^H\mathbf{d}_u, \tag{4.18}$$

which has zero mean and variance

$$\sigma_{\nu,k}^2 = \rho_u\alpha_{qu}^2\sum_{\substack{i=1\\i\neq k}}^{K}|\mathbf{a}_k^H\hat{\mathbf{g}}_i|^2 + \mathbf{a}_k^H\left(\rho_u\alpha_{qu}^2\sum_{i=1}^{K}\mathbb{E}\{\tilde{\mathbf{g}}_i\tilde{\mathbf{g}}_i^H\} + \alpha_{qu}^2\sigma_n^2\mathbf{I}_M + \sigma_{d_u}^2\mathbf{I}_M\right)\mathbf{a}_k. \tag{4.19}$$

Thus, we can we can use the concept of treating interference as noise and formulate the achievable rate of this scheme as given in the following proposition.

Proposition 4.2.1.
Consider that K users are served by L APs which are equipped with N antennas and connected to the CPU by low-resolution fronthaul. If the APs perform individual processing for each antenna, then the ergodic achievable rate of the k-th user in the uplink of centralized cell-free massive MIMO is determined by

$$R_{u,k}^{\mathrm{In}} = \mathbb{E}\left\{\log_2\left(1+\mathrm{SINR}_k^{\mathrm{In}}\right)\right\}. \tag{4.20}$$

The $\mathrm{SINR}_k^{\mathrm{In}}$ *is given by*

$$\mathrm{SINR}_k^{\mathrm{In}} = \frac{\rho_u \alpha_{qu}^2 |\mathbf{a}_k^H \hat{\mathbf{g}}_k|^2}{\rho_u \alpha_{qu}^2 \sum_{\substack{i=1 \\ i\neq k}}^{K} |\mathbf{a}_k^H \hat{\mathbf{g}}_i|^2 + \mathbf{a}_k^H \boldsymbol{\Lambda} \mathbf{a}_k} \tag{4.21}$$

with the M=LN dimensional matrix

$$\boldsymbol{\Lambda} = \mathrm{diag}\{\Lambda_1, ..., \Lambda_M\}, \tag{4.22}$$

where

$$\Lambda_m = \sigma_{d_u}^2 + \alpha_{qu}^2 \sigma_n^2 + \rho_u \alpha_{qu}^2 \sum_{i=1}^{K} \epsilon_{li}^q, \tag{4.23}$$

The data distortion variance and the noise variance are denoted respectively by $\sigma_{d_u}^2 = (\lambda_{qu} - \alpha_{qu}^2)\sigma_y^2$ *and* σ_n^2. *The estimation error is given by* $\epsilon_{lk}^q \in \{\epsilon_{lk}^{eq}, \epsilon_{lk}^{qe}\}$ *depending on the CSI acquisition scheme from (3.36) or (3.45).*

Proof. The sketch of the proof is the same as the proof in [33, Th. 4.1], where the input is given here by $x = x_{u,k}$, the output by $y = \hat{x}_{u,k}$, the channel response by $h = \mathbf{a}_k^H \hat{\mathbf{g}}_k$, the random realization affecting the interference is given by $u = \{\hat{\mathbf{g}}_k\}$ and the interference term ν is given by (4.18). It remains to show that the input is conditionally uncorrelated to the interference term ν. That is

$$\mathbb{E}\{x^*\nu|h,u\} = \mathbb{E}\{x^*\nu|\{\hat{\mathbf{g}}_k\}\} = 0. \tag{4.24}$$

We have uncorrelation from the first term of ν due to the assumption that there is no correlation between the channels of user $i \neq k$. From the second term we have uncorrelation due to the fact that the channel estimate is uncorrelated with the channel estimation error. The third term is due to the independent realization of the additive noise and the last term is due to the Bussgang decomposition. ■

4.3 Joint Processing at Multiple-Antenna APs

In practical scenarios, the wireless channel of a multi-antenna system tends to be spatially correlated. It can be described by the model given in Section 4.1.1, among others. This spatial correlation is simply omitted by the processing at the AP described in the previous section, where the received signals at each antenna are treated individually. In contrast, we propose in this section to process the signals received at the multiple antennas jointly for the fronthaul quantization. In this case, we use vector quantization (VQ) with a precision of only small number of bits, so that we can simultaneously exploit the channel correlation and meet the low bit requirements of the fronthaul.

4.3.1 Fronthaul with Vector Quantization

We consider compression which consists of vector quantization followed by fixed-rate lossless coding. At each AP a vector quantizer Q is applied as an interface to the fronthaul with

$$Q(\mathbf{x}) = \sum_{s=0}^{S-1} \mathbf{q}_s T_s(\mathbf{x}), \text{ where } T_s(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{C}_s \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

In contrast to the separate scalar quantization in the previous section, we take jointly the received signal from the N antennas at the AP as the input of our vector quantizer Q. Figure 4.3 depicts this joint processing scheme for the l-th AP. In this case, we arrange the N samples from N antennas as an N-dimensional vector $\mathbf{x}$. Whenever the input vector $\mathbf{x} \in \mathbb{R}^N$ falls into the cell $\mathcal{C}_s$, the index s will be transmitted on the fronthaul link, and the reconstruction value $\mathbf{q}_s$ taken from the codebook $\mathcal{Q} = \{\mathbf{q}_s\}_{s=0}^{S-1} \subset \mathbb{R}^N$ will be used at the CPU. The codebook size is chosen in corresponding to the fronthaul capacity by $S = 2^C$. Moreover, to be comparable with the individual processing in the previous section, we allocate a rate of $R_N = \frac{1}{N}\log_2[S] = \frac{C}{N}$ bit per dimension for N-dimensional vector quantization. Here, we keep R_N small, to one or two bits per dimension. For a complex-valued signal we quantize the real and imaginary part separately. We do this because the correlation affects the real and imaginary part of the channel independently.

The optimal codebooks can be found using the Linde Buzo Gray (LBG) algorithm for minimum mean squared error [40]. This algorithm is the counter part of Lloyd algorithm for vector quantization, where the optimal codebook is obtained by iterating between finding the optimal partition using the nearest neighbour criterion and finding the optimal reconstruction values from the centroid condition. By doing so we expect to

adapt the codebooks to the spatial channel correlation. Note that the channel correlation is a large scale statistical parameter. Hence, we should only run the LBG algorithm once every coherence block. For a further detail on the LBG algorithm, we suggest the reader look at Algorithm 2 in Section 2.2.2 or refer to [40].

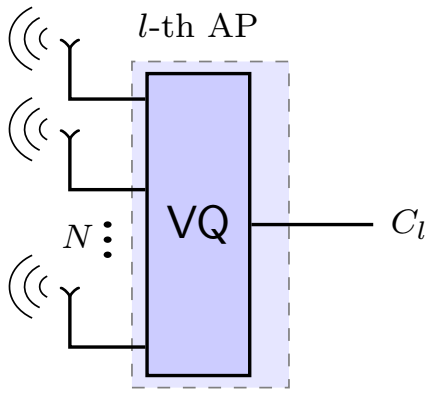


Fig. 4.3 Illustration of the joint processing using vector quantization.

4.3.2 Bussgang Decomposition for Vector Quantization

In the same way as the scalar quantizer, the vector quantizer Q given in (4.25) is also generally non-linear and the error $\mathbf{e} \triangleq \mathbf{x} - Q(\mathbf{x})$ resulting from the quantization process is correlated with the input vector $\mathbf{x}$. Therefore, using the Bussgang theorem [43] we would like to express our quantizer as the following linear model

$$\mathbf{x}_q = Q(\mathbf{x}) = \mathbf{F}\mathbf{x} + \mathbf{d}. \tag{4.26}$$

If we consider that the input $\mathbf{x}$ is Gaussian, then the distortion $\mathbf{d}$ is statistically equivalent to the quantization error $\mathbf{e}$ but uncorrelated with the signal component $\mathbf{x}$. The linear operator $\mathbf{F}$, which depends essentially on the given distortion characteristic of Q, tells us also about the proportional factor between the input-output covariance of the quantizer expressed as [43]

$$\mathbf{C}_{xx_q} = \mathbf{F}\mathbf{C}_{xx}, \text{ where} \tag{4.27}$$

$$\mathbf{C}_{xx_q} = \mathbb{E}\{\mathbf{x}\mathbf{x}_q^H\} \text{ and } \mathbf{C}_{xx} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\}. \tag{4.28}$$

In this case, finding $\mathbf{F}$ can be seen as finding the LMMSE estimator for $\mathbf{x}_q$ from the observation $\mathbf{x}$ [47]

$$\mathbf{F} = \mathbf{C}_{xx_q}\mathbf{C}_{xx}^{-1}, \tag{4.29}$$

where the estimation error $\mathbf{d}$ is then orthogonal to $\mathbf{x}$. For separate scalar quantizers observed in the previous section the matrix $\mathbf{F}$ has the form of a diagonal matrix, which can not be the case for a joint processing using vector quantization. In fact, the closed form expression of $\mathbf{F}$ is not yet known for a general quantizer, particularly for vector quantizers. Therefore, we compute $\mathbf{F}$ numerically whenever it is needed by assuming that we have access to measurements of the input as well as the output of Q. We estimate the covariance matrix $\mathbf{C}_{xx}$ from the sample covariance matrix

$$\hat{\mathbf{C}}_{xx} = \frac{1}{N_t}\sum_{n_t=1}^{N_t}\mathbf{x}[n_t]\mathbf{x}[n_t]^H \tag{4.30}$$

and respectively for $\mathbf{C}_{x_qx_q}$ and $\mathbf{C}_{x_qx}$. The number of observations N_t can be conveniently taken as equal to the number of codebooks' training, where $\hat{\mathbf{C}}_{xx}$ will approach $\mathbf{C}_{xx}$ for large N_t. Using $\mathbf{F}$ given in (4.29) we can then compute the covariance of the distortion $\mathbf{d}$ expressed as

$$\begin{aligned}\mathbf{C}_{dd} &= \mathbb{E}\{(\mathbf{x}_q - \mathbf{F}\mathbf{x})(\mathbf{x}_q - \mathbf{F}\mathbf{x})^H\} \\ &= \mathbf{C}_{x_qx_q} - \mathbf{C}_{x_qx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x_qx}^H.\end{aligned} \tag{4.31}$$

4.3.3 CSI Acquisition with Vector Quantization

In this subsection we address the CSI acquisition problems by utilizing vector quantization assuming a fronthaul resolution of only a few bits. As in the single-antenna AP, we consider two possible strategies which will be described as follows.

4.3.3.1 Estimate-and-Quantize

In this scheme we first estimate the channel at the APs and then quantize the resulting CSI with the quantizer given in (4.25) to meet the fronthaul-capacity limit of C bits. The channel estimation is done based on the transmission of known pilots. Every user uses a specific sequence taken from a set $\mathcal{P}_\varphi$ of orthonormal random sequences $\boldsymbol{\varphi}_k \in \mathbb{C}^{\tau\times 1}$ with $\langle\boldsymbol{\varphi}_k, \boldsymbol{\varphi}'_k\rangle = \delta_{kk'}$ and $\|\boldsymbol{\varphi}_k\|^2 = 1$, where the sequence length τ is assumed to be less than or equal to the coherence interval τ_c. The k-th user sends $\sqrt{\tau}\boldsymbol{\varphi}_k$ as its pilot such

that the l-th AP observes the received pilot $\mathbf{Y}_{p,l} \in \mathbb{C}^{N\times\tau}$ from all K users as

$$\mathbf{Y}_{p,l} = \sqrt{\tau\rho_p}\sum_{k=1}^{K}\mathbf{g}_{lk}\boldsymbol{\varphi}_k^H + \mathbf{W}_l, \tag{4.32}$$

where ρ_p is the normalized transmit power of the pilot and $\mathbf{W}_l$ is an additive noise matrix at the l-th AP whose entries are uncorrelated with zero mean and unit variance.

To allow all pilots to be orthogonal for all K users, only $K \leq \tau$ users may transmit their pilots simultaneously. In this case, the transmitted pilots satisfy

$$\Phi^H\Phi = \tau\rho_p\mathbb{I}_K, \text{ where } \Phi = \sqrt{\tau\rho_p}[\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_K]. \tag{4.33}$$

The channel vector $\mathbf{g}_{lk}$ can be estimated at the APs where the received pilot $\mathbf{Y}_{p,l}$ is projected onto $\boldsymbol{\varphi}_k$ expressed as

$$\begin{aligned}\mathbf{r}_{p,lk} &= \frac{1}{\sqrt{\tau\rho_p}}\mathbf{Y}_{p,l}\boldsymbol{\varphi}_k \\ &= \mathbf{g}_{lk} + \sum_{k'\neq k}^{K}\mathbf{g}_{lk'}\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k + \frac{1}{\sqrt{\tau\rho_p}}\mathbf{W}_l\boldsymbol{\varphi}_k.\end{aligned} \tag{4.34}$$

To obtain the estimate of $\mathbf{g}_{lk}$ we use the LMMSE estimator given by

$$\hat{\mathbf{g}}_{lk} = \boldsymbol{\Gamma}_{lk}\mathbf{r}_{p,lk} \tag{4.35}$$

with the gain matrix $\boldsymbol{\Gamma}_{lk}$ is given by

$$\boldsymbol{\Gamma}_{lk} = \boldsymbol{\Sigma}_{lk}\left(\boldsymbol{\Omega}_{lk}\right)^{-1}, \tag{4.36}$$

where

$$\boldsymbol{\Sigma}_{lk} = \mathbb{E}\{\mathbf{g}_{lk}\mathbf{g}_{lk}^H\} \text{ and } \boldsymbol{\Omega}_{lk} = \mathbb{E}\{\mathbf{r}_{p,lk}\mathbf{r}_{p,lk}^H\} = \boldsymbol{\Sigma}_{lk} + \frac{1}{\tau\rho_p}\mathbf{I}_N. \tag{4.37}$$

After accomplishing the channel estimation the APs quantize the channel estimate $\hat{\mathbf{g}}_{lk}$. We assume that the large scale fading β_{lk} is relatively constant over a long period and known at the APs. Thus, we may scale the input to the vector quantizer accordingly with β_{lk} and approximate the distribution as multivariate Gaussian. Consequently, the quantized channel estimate can be written using Bussgang decomposition as

$$\hat{\mathbf{g}}_{lk}^{eq} = \mathbf{F}_{g,l}\hat{\mathbf{g}}_{lk} + \mathbf{d}_{g,l}. \tag{4.38}$$

The channel estimation error after the quantization is given by $\tilde{\mathbf{g}}_{lk}^{eq} = \mathbf{g}_{lk} - \hat{\mathbf{g}}_{lk}^{eq}$ with covariance

$$\begin{aligned}\boldsymbol{\Psi}_{lk}^{eq} &= \mathbb{E}\{\tilde{\mathbf{g}}_{lk}^{eq}(\tilde{\mathbf{g}}_{lk}^{eq})^H\} \\ &= \mathbb{E}\{(\mathbf{g}_{lk} - \hat{\mathbf{g}}_{lk}^{eq})(\mathbf{g}_{lk} - \hat{\mathbf{g}}_{lk}^{eq})^H\} \\ &= \boldsymbol{\Sigma}_{lk} - \mathbf{F}_{g,l}\mathbf{C}_{\hat{g}g} + (\mathbf{F}_{g,l} - \mathbf{I}_N)(\mathbf{F}_{g,l}\mathbf{C}_{\hat{g}g})^H + \mathbf{C}_{d_g d_g},\end{aligned} \tag{4.39}$$

where $\mathbf{C}_{\hat{g}g}$ is the cross covariance matrix between the channel $\mathbf{g}_{lk}$ and its estimate $\hat{\mathbf{g}}_{lk}$, and $\mathbf{C}_{d_g d_g}$ is the covariance of the distortion from (4.38). The last equation (4.39) follows from doing some algebra and the fact that the channel $\mathbf{g}_{lk}$ and channel estimate $\hat{\mathbf{g}}_{lk}$ are uncorrelated with the distortion $\mathbf{d}_{g,l}$. The channel estimate and estimation error of the k-th user can be expressed then as a stack of column vectors from all L AP given respectively by

$$\hat{\mathbf{g}}_k^{eq} = \begin{bmatrix} \hat{\mathbf{g}}_{1k}^{eq} \\ \vdots \\ \hat{\mathbf{g}}_{Lk} \end{bmatrix} \text{ and } \tilde{\mathbf{g}}_k^{eq} = \mathbf{g}_k - \hat{\mathbf{g}}_k^{eq} \sim \mathcal{NC}(\mathbf{0}, \boldsymbol{\Psi}_k^{eq}), \tag{4.40}$$

where $\boldsymbol{\Psi}_k^{eq} = \text{diag}(\boldsymbol{\Psi}_{1k}^{eq}, \ldots, \boldsymbol{\Psi}_{Lk}^{eq})$.

4.3.3.2 Quantize-and-Estimate

Instead of transferring the quantized CSI we consider here an alternative CSI acquisition strategy where we first quantize the received pilots at the APs and then estimate the channel from the quantized pilots at the CPU. To be more specific, the l-th AP quantizes the received pilots at the N antennas jointly as

$$\begin{aligned}\mathbf{y}_{qp,l}^{(t)} = Q(\mathbf{y}_{p,l}^{(t)}) &= Q\left(\sqrt{\tau\rho_p}\sum_{k=1}^{K}\mathbf{g}_{lk}\boldsymbol{\varphi}_k^{(t)^*} + \mathbf{w}_l^{(t)}\right) \\ &= Q\left(\sqrt{\tau\rho_p}\mathbf{G}_l\boldsymbol{\varphi}^{(t)^H} + \mathbf{w}_l^{(t)}\right)\end{aligned} \tag{4.41}$$

where the superscript $t = \{1, \ldots, \tau\}$ denotes the index of the pilot sequence. Accordingly $\mathbf{y}_{p,l}^{(t)}$ is the t-th column of $\mathbf{Y}_{p,l}$ in (4.32) and $\boldsymbol{\varphi}^{(t)}$ is the t-th row of Φ in (4.33).

Applying the Bussgang decomposition to (4.41) we obtain

$$\begin{aligned}\mathbf{y}_{qp,l}^{(t)} &= \mathbf{F}_{p,l}\mathbf{y}_{p,l}^{(t)} + \mathbf{d}_{p,l}^{(t)} \\ &= \sqrt{\tau\rho_p}\mathbf{F}_{p,l}\mathbf{G}_l\boldsymbol{\varphi}^{(t)^H} + \mathbf{F}_{p,l}\mathbf{w}_l^{(t)} + \mathbf{d}_{p,l}^{(t)}.\end{aligned} \tag{4.42}$$

The CPU receives from all L APs as a stack of (4.42)

$$\mathbf{y}_{qp}^{(t)} = \begin{bmatrix} \mathbf{y}_{qp,1}^{(t)} \\ \vdots \\ \mathbf{y}_{qp,L}^{(t)} \end{bmatrix} = \begin{bmatrix} \sqrt{\tau\rho_p}\mathbf{F}_{p,1}\mathbf{G}_1{\boldsymbol{\varphi}^{(t)}}^H + \mathbf{F}_{p,1}\mathbf{w}_1^{(t)} + \mathbf{d}_{p,1}^{(t)}. \\ \vdots \\ \sqrt{\tau\rho_p}\mathbf{F}_{p,L}\mathbf{G}_L{\boldsymbol{\varphi}^{(t)}}^H + \mathbf{F}_{p,L}\mathbf{w}_L^{(t)} + \mathbf{d}_{p,L}^{(t)}. \end{bmatrix} \tag{4.43}$$

which we can concisely rewrite as a $M \times \tau$ matrix for τ-length sequences of quantized received pilots given by

$$\mathbf{Y}_{qp} = \begin{bmatrix} \mathbf{y}_{qp,1}^{(1)} & \cdots & \mathbf{y}_{qp,1}^{(\tau)} \\ \vdots & \ddots & \vdots \\ \mathbf{y}_{qp,L}^{(1)} & \cdots & \mathbf{y}_{qp,L}^{(\tau)} \end{bmatrix} = \sqrt{\tau\rho_p}\mathbf{F}_p\mathbf{G}\Phi^H + \mathbf{F}_p\mathbf{W} + \mathbf{D}, \tag{4.44}$$

where the matrix $\mathbf{F}_p$ is a $M \times M$ diagonal matrix with $\mathbf{F}_{p,l} \in \mathbb{R}^{N\times N}$ in its block diagonal entries. With similar structure to $\mathbf{Y}_{qp} \in \mathbb{C}^{M\times\tau}$ the matrices $\mathbf{W}$ and $\mathbf{D}$ denote respectively the noise and distortion matrices.

We can then project $\mathbf{Y}_{qp}$ onto $\boldsymbol{\varphi}_k$ and expressed the result as

$$\begin{aligned} \mathbf{r}_{qp,k} &= \frac{1}{\sqrt{\tau\rho_p}}\mathbf{Y}_{qp}\boldsymbol{\varphi}_k \\ &= \mathbf{F}_p\mathbf{G}\Phi^H\boldsymbol{\varphi}_k + \frac{1}{\sqrt{\tau\rho_p}}(\mathbf{F}_p\mathbf{W} + \mathbf{D})\boldsymbol{\varphi}_k \\ &= \mathbf{F}_p\mathbf{g}_k + \mathbf{F}_p\sum_{k'\neq k}^{K}\mathbf{g}_{k'}\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k + \frac{1}{\sqrt{\tau\rho_p}}\left(\mathbf{F}_p\mathbf{W} + \mathbf{D}\right)\boldsymbol{\varphi}_k. \end{aligned} \tag{4.45}$$

We estimate $\mathbf{g}_k$ using an LMMSE estimator such that the channel estimate for the k-th user using the QE strategy is given by

$$\hat{\mathbf{g}}_k^{qe} = \boldsymbol{\Gamma}_{qp,k}\mathbf{r}_{qp,k}. \tag{4.46}$$

Following the known result in estimation theory [33, Lemma B.17] we use the gain matrix $\boldsymbol{\Gamma}_{qp,k}$ given by

$$\begin{aligned} &\boldsymbol{\Gamma}_{qp,k} = \boldsymbol{\Sigma}_k\mathbf{F}_p^H\left(\boldsymbol{\Omega}_{qp,k}\right)^{-1}, \text{ where } \boldsymbol{\Sigma}_k = \mathbb{E}\{\mathbf{g}_k\mathbf{g}_k^H\} \text{ and} \\ &\boldsymbol{\Omega}_{qp,k} = \mathbb{E}\{\mathbf{r}_{qp,k}\mathbf{r}_{qp,k}^H\} = \mathbf{F}_p\boldsymbol{\Sigma}_k\mathbf{F}_p^H + \frac{1}{\tau\rho_p}\left(\mathbf{F}_p\mathbf{F}_p^H + \mathbf{D}\mathbf{D}^H\right) \end{aligned} \tag{4.47}$$

The covariance of the channel estimation error $\tilde{\mathbf{g}}_k^{qe} = \mathbf{g}_k - \hat{\mathbf{g}}_k^{qe}$ is then given by

$$\boldsymbol{\Psi}_k^{qe} = \mathbb{E}\{\tilde{\mathbf{g}}_k^{qe}(\tilde{\mathbf{g}}_k^{qe})^H\} = \boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k \mathbf{F}_p^H \boldsymbol{\Omega}_{qp,k}^{-1} \mathbf{F}_p \boldsymbol{\Sigma}_k. \tag{4.48}$$

Note that the received signals at the APs are uncorrelated over l and t such that the Gram matrix $\mathbf{F}_p\mathbf{F}_p^H$ and $\mathbf{D}\mathbf{D}^H$ have a block diagonal structure. Their submatrices are positive definite since $\mathbf{F}$ and $\mathbf{d}$ in (4.26) are positive definite for a large number of observations in the sample covariance matrix (4.30). Thus, the matrix $\boldsymbol{\Omega}_{qp,k}$ is invertible.

Further, we can optimize the codebook $\mathcal{Q}$ of each AP off-line for any CSI acquisition strategy and we need only to update it as the β_{lk} changes. As demonstrated in Figure 4.4a, 4.4b, 4.4c for the EQ strategy and in Figure 4.4d, 4.4e, 4.4f for the QE strategy, the codebooks can exploit the spatial channel correlation effectively. Due to the LBG algorithm the reconstruction points $\{\mathbf{q}_s\}$ are placed more densely in the region where the input signals occur with high probability. As the correlation increases, the reconstruction points get closer to the diagonal to optimally represent the dependency between input signals. Thus, the distance from the input signals to the points $\{\mathbf{q}_s\}$ becomes smaller resulting in a smaller average distortion.

4.3.4 Data Transmission

Consider, as previously, that all users send their data $x_{u,k}$ simultaneously in the uplink with $\mathbb{E}\{|x_{u,k}|^2\} = 1$ and the l-th AP receives them as

$$\mathbf{y}_{u,l} = \sqrt{\rho_u} \sum_{k=1}^{K} \mathbf{g}_{lk} x_{u,k} + \mathbf{w}_l. \tag{4.49}$$

We investigate now the case where the data-bearing signal $\mathbf{y}_{u,l} \in \mathbb{C}^N$ is jointly processed using vector quantization without first performing the data detection at the AP. One of the advantages of doing this is the possibility to approximate the received signal $\mathbf{y}_{u,l}$ as a multivariate Gaussian variable, particularly when $\mathbf{y}_{u,l}$ is a superposition of a very large number K of user signals. If in contrast we detect the user data signal first, then the detected user data signal can not be well approximated as Gaussian. Suppose that we have a rather large K, then we may approximate $\mathbf{y}_{u,l}$ as Gaussian and hence we can make use of the Bussgang decomposition for the vector quantization given in (4.26). In this case, we can express $\mathbf{y}_{u,l}$ after the joint vector quantization as

$$\mathbf{r}_{u,l} = Q(\mathbf{y}_{u,l}) = \mathbf{F}_{u,l}\mathbf{y}_{u,l} + \mathbf{d}_{u,l}, \tag{4.50}$$

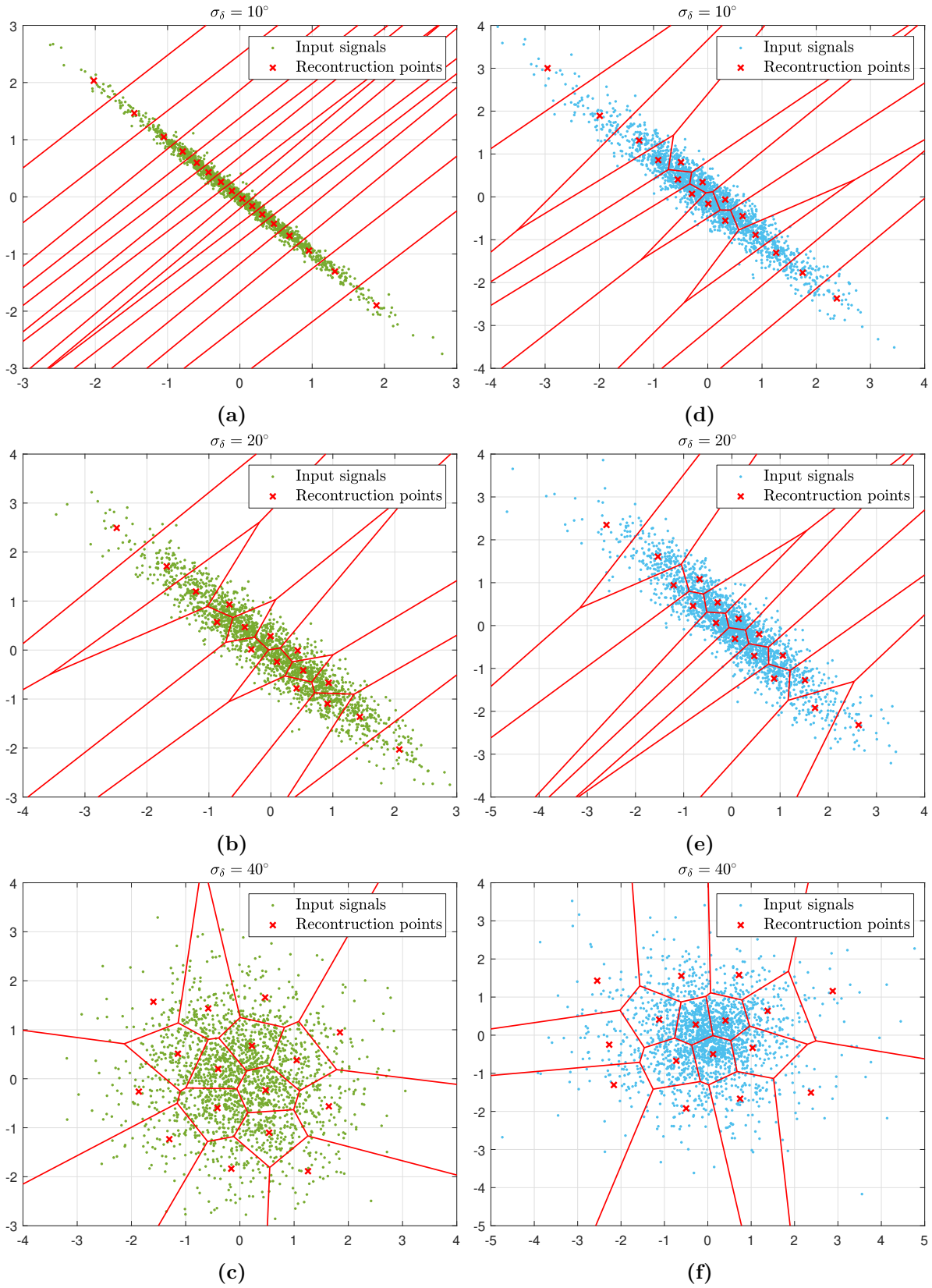


Fig. 4.4 The Voronoi region of 2-dimensional codebooks $\mathcal{Q}$ for EQ (a,b,c) and QE (d, e, f) with different degree of correlation (in this case for random angular spread δ with Gaussian distribution and different standard deviation $\sigma_\delta = 10°, 20°, 40°$).

and the CPU collects from L APs the quantized data signal as

$$\underbrace{\begin{bmatrix}\mathbf{r}_{u,1}\\ \vdots \\ \mathbf{r}_{u,L}\end{bmatrix}}_{\triangleq \mathbf{r}_u} = \mathbf{F}_u \underbrace{\begin{bmatrix}\mathbf{y}_{u,1}\\ \vdots \\ \mathbf{y}_{u,L}\end{bmatrix}}_{\triangleq \mathbf{y}_u} + \underbrace{\begin{bmatrix}\mathbf{d}_{u,1}\\ \vdots \\ \mathbf{d}_{u,L}\end{bmatrix}}_{\triangleq \mathbf{d}_u}, \tag{4.51}$$

where $\mathbf{F}_u = \text{diag}(\mathbf{F}_{u,1}, \ldots, \mathbf{F}_{u,L}) \in \mathbb{C}^{M\times M}$ with $M = LN$. Suppose now that the CPU constructs a linear detector matrix $\mathbf{A}$ from the jointly vector-quantized channel estimate $\hat{\mathbf{G}}$ given in (4.40) or (4.46), then the estimated data can be obtained from

$$\begin{aligned}\hat{\mathbf{x}}_u &= \mathbf{A}^H \mathbf{r}_u \\ &= \mathbf{A}^H(\mathbf{F}_u \mathbf{y}_u + \mathbf{d}_u) \\ &= \mathbf{A}^H(\mathbf{F}_u(\sqrt{\rho_u}\mathbf{G}\mathbf{x}_u + \mathbf{w}_u) + \mathbf{d}_u) \\ &= \sqrt{\rho_u}\mathbf{A}^H\mathbf{F}_u\hat{\mathbf{G}}\mathbf{x}_u + \sqrt{\rho_u}\mathbf{A}^H\mathbf{F}_u\tilde{\mathbf{G}}\mathbf{x}_u + \mathbf{A}^H\mathbf{F}_u\mathbf{w}_u + \mathbf{A}^H\mathbf{d}_u\end{aligned} \tag{4.52}$$

4.3.5 Achievable Rate

Similar to the individual processing scheme, we are now interested to determine the data rate achievable if the APs jointly process the received signal using vector quantization. Since we take into account the dependency of the quantization of all M antennas, the estimated data for the k-th user can be written as

$$\begin{aligned}\hat{x}_{u,k} = \sqrt{\rho_u}\sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\hat{\mathbf{g}}_k x_{u,k} + \sqrt{\rho_u}\sum_{\substack{i=1\\ i\neq k}}^{K}\sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\hat{\mathbf{g}}_i x_{u,i} \\ + \sqrt{\rho_u}\sum_{i=1}^{K}\sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\tilde{\mathbf{g}}_i x_{u,i} + \sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\mathbf{w}_u + \mathbf{a}_k^H\mathbf{d}_u,\end{aligned} \tag{4.53}$$

where the first term comprises the data for the k-th user to be decoded. The remaining terms, given by

$$\nu_k = \sqrt{\rho_u}\sum_{\substack{i=1\\ i\neq k}}^{K}\sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\hat{\mathbf{g}}_i x_{u,i} + \sqrt{\rho_u}\sum_{i=1}^{K}\sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\tilde{\mathbf{g}}_i x_{u,i} + \sum_{m=1}^{M}\mathbf{a}_k^H\mathbf{f}_m\mathbf{w}_u + \mathbf{a}_k^H\mathbf{d}_u, \tag{4.54}$$

can be seen as interference with variance

$$\sigma^2_{\nu,k} = \rho_u \sum_{\substack{i=1 \\ i\neq k}}^{K} \sum_{m=1}^{M} |\mathbf{a}_k^H \mathbf{f}_m \hat{\mathbf{g}}_i|^2 + \sum_{m=1}^{M} \mathbf{a}_k^H \mathbf{f}_m \left(\rho_u \sum_{i=1}^{K} \boldsymbol{\Psi}_i + \sigma_n^2 \mathbf{I}_M \right) \mathbf{a}_k \mathbf{f}_m^H + \mathbf{a}_k^H \mathbf{C}_{d_u d_u} \mathbf{a}_k, \quad (4.55)$$

where $\mathbf{C}_{d_u d_u}$ is the covariance of the data distortion due to quantization and $\boldsymbol{\Psi}_i$ is the covariance of the estimation error. Using the same principle of treating interference as noise we can formulate the achievable rate for this scheme as summarized in the following proposition.

Proposition 4.3.1.
Assume that K users are served by L APs which are equipped with N antennas and connected to the CPU with low-resolution fronthaul. If the APs perform joint processing using vector quantization, then the ergodic achievable rate of the k-th user in the uplink of centralized cell-free massive MIMO is given by

$$R_{u,k}^{\text{Jo}} = \mathbb{E}\left\{ \log_2 \left(1 + \text{SINR}_k^{\text{Jo}} \right) \right\}. \quad (4.56)$$

The $\text{SINR}_k^{\text{Jo}}$ *is given by*

$$\text{SINR}_k^{\text{Jo}} = \frac{\rho_u \sum_{m=1}^{M} |\mathbf{a}_k^H \mathbf{f}_m \hat{\mathbf{g}}_k|^2}{\rho_u \sum_{\substack{i=1 \\ i\neq k}}^{K} \sum_{m=1}^{M} |\mathbf{a}_k^H \mathbf{f}_m \hat{\mathbf{g}}_i|^2 + \sum_{m=1}^{M} \mathbf{a}_k^H \mathbf{f}_m \boldsymbol{\Lambda} \mathbf{a}_k \mathbf{f}_m^H + \mathbf{a}_k^H \mathbf{C}_{d_u d_u} \mathbf{a}_k} \quad (4.57)$$

where

$$\boldsymbol{\Lambda} \triangleq \rho_u \sum_{i=1}^{K} \boldsymbol{\Psi}_i + \sigma_n^2 \mathbf{I}_M, \quad (4.58)$$

$\mathbf{C}_{d_u d_u}$, σ_n^2 *are respectively the covariance of the data distortion and the noise variance,* $\boldsymbol{\Psi}_i \in \{\boldsymbol{\Psi}_i^{eq}, \boldsymbol{\Psi}_i^{qe}\}$ *is the covariance of the estimation error from (4.40) or (4.48) depending on the CSI acquisition scheme.*

Proof. The proof is similar to the proof of Proposition 4.2.1 which follows [33, Th. 4.1] apart from replacing the useful signal term with the first term in (4.53) and the interference term with (4.54). ■

In the ideal case of perfect fronthaul, the covariance $\mathbf{C}_{d_u d_u}$ will be a zero matrix and $\mathbf{F}_u$ will be an identity matrix, such that the SINR expression in (4.57) will be equal to the SINR expression in [11, Eq. (12)] given by

$$\text{SINR}_k = \frac{\rho_u |\mathbf{a}_k^H \hat{\mathbf{g}}_k|^2}{\rho_u \sum_{\substack{i=1 \\ i\neq k}}^{K} |\mathbf{a}_k^H \hat{\mathbf{g}}_i|^2 + \mathbf{a}_k^H \left(\rho_u \sum_{i=1}^{K} \boldsymbol{\Psi}_i + \sigma_n^2 \mathbf{I}_M \right) \mathbf{a}_k} \quad (4.59)$$

Table 4.1 The scalability of cell-free massive MIMO with Multi-antenna AP

	Min. AP Processing [Scalar multiplication]	Min. Fronthaul Load [bits]
Original CF maMIMO (MRC)	$NK\tau_p + NK\tau_u$	$2\log_2[S]K\tau_u$
Centralized CF maMIMO (EQ)	$NK\tau_p$	$2N\log_2[S](K+\tau_u)$
Centralized CF maMIMO (QE)	-	$2N\log_2[S](\tau_p+\tau_u)$

4.4 Scalability

Since we now have multiple antennas at each AP, we discuss briefly in this section how this affects the scalability of centralized cell-free massive MIMO in terms of AP processing and fronthaul load. We then discuss another aspect of scalability which is the CPU processing load when we have a very large number of users distributed in a very large coverage area. A natural question is whether the centralized approach can still handle a large number of users in such a scenario.

Suppose that each AP has N antennas and receives signals from K users simultaneously. If we employ the original cell-free massive MIMO which uses MRC and least square estimation, the AP requires to process N times more scalar multiplications for data detection and channel estimation. But we can apply the same scalar coefficient to all antennas at the AP, resulting in just one weighted signal per user and data symbol. Thus, the minimum fronthaul load is $2\log_2[S]K\tau_u$ which is equal to that in the case of a single-antenna AP. On the other hand, in centralized cell-free massive MIMO the fronthaul load scales with N either for EQ or QE CSI acquisition as given in Table 4.1. In this regard, one might think that the centralized approach requires a greater fronthaul load. However, since K and τ_u are typically much larger than N, the centralized approach may require much less fronthaul load than the original cell-free massive MIMO. In this case, although the required fronthaul load increases with N, the centralized approach with multiple antennas is still scalable in terms of AP processing and fronthaul load, particularly when the QE strategy is applied.

So far we have assumed that in centralized cell-free massive MIMO the CPU has a very high computing power to perform all the required processing from acquiring the CSI to decoding the user data. However, for practical consideration such as an implementation in a big city, the CPU processing may become excessive. When the number of users and the service area to be covered increases, then the number of APs required will be very high, and the CPU processing will become unscalable. To deal with this issue, it is therefore necessary to mention that the centralized concept can be

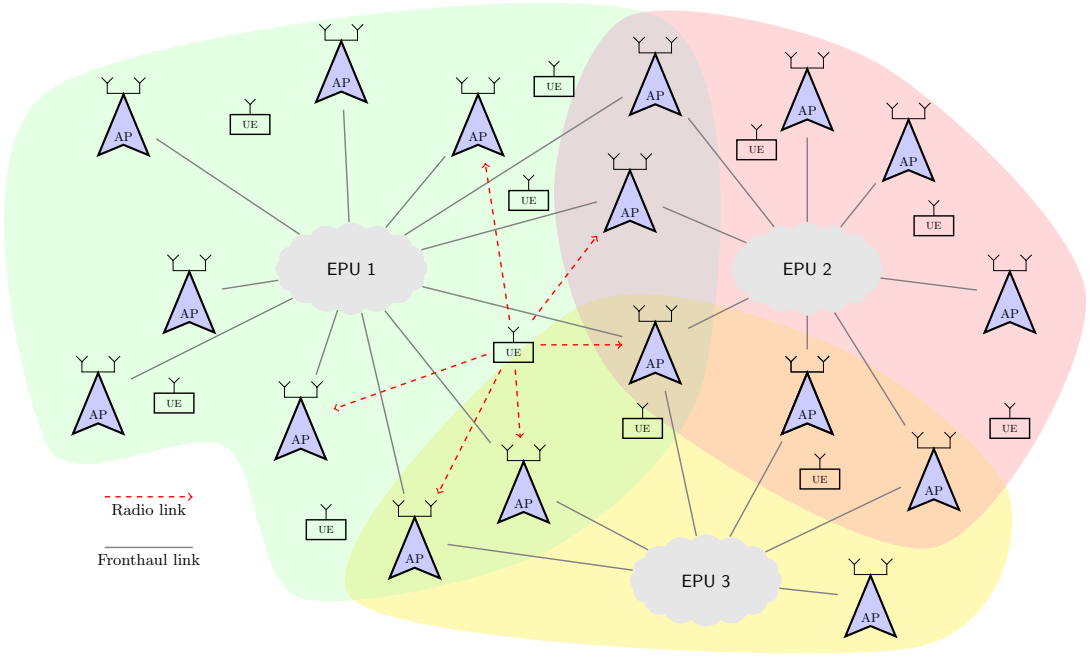


Fig. 4.5 The architecture of "Fog-massive MIMO" (F-MaMIMO) from [27] for instance of 3 Edge Processing Units (EPUs). The centralized cell-free massive MIMO can be extended to form this architecture.

extended to a larger network where the APs in the service area are connected to more than one CPU. This is illustrated in Figure 4.5 which is studied in [27] under name of Fog massive MIMO (F-maMIMO). As shown in the figure, the network is a combination of overlapping cell-free massive MIMO networks. The task of one single CPU is now delegated to several CPUs which are here called Edge Processing Units (EPUs) due to their place on the network edge. These EPUs are coordinated to serve the users using the APs which have the best channel condition. In this case, using the centralized approach the EPUs have only to know the CSI from their selected APs. Based on the acquired CSI, the data processing is performed at multiple EPUs which can be attained among others by making consensus between adjacent EPUs to find the optimal detection matrix. A similar network of cell-free massive MIMO with multiple CPUs has also recently been studied in [60] by exploiting the dynamic cooperation clustering (DCC). The concept is also possible to be combined with the centralized approach described in this thesis. However, further investigation is beyond the scope of this thesis.

4.5 Performance Evaluation

We provide in this section some numerical simulations to assess the performance of the schemes considered above. Unless otherwise stated, we performed our simulations with M AP antennas in total, $L = M/N$ APs and $K = 20$ users distributed uniformly in an area of 1×1 km^2. This area is wrapped around by its copies so that it resembles a network with infinite area. The channel $\mathbf{g}_{lk}$ in (4.3) is modelled with the large scale fading β_{lk} given as

$$\beta_{lk} = \mathrm{PL}_{lk} \cdot 10^{\frac{\sigma_{sh} z_{lk}}{10}}, \tag{4.60}$$

where the factor $10^{\frac{\sigma_{sh} z_{lk}}{10}}$ is the uncorrelated shadowing with standard deviation $\sigma_{sh} = 8$ dB and $z_{lk} \sim \mathcal{N}(0, 1)$. The path loss coefficient follows the three-slope model which is similar to the model used in Chapter 3 given by (3.86). We use also for the simulation the physical parameters given in Table 3.3.

We evaluate first the performance of the CSI acquisition for the individual and joint processing respectively using the EQ and QE strategy. We use the MSE as the performance metric, defined as

$$\mathrm{MSE} = \frac{1}{MK} \mathbb{E}\{\|\mathbf{G} - \hat{\mathbf{G}}\|^2\}, \tag{4.61}$$

where $\hat{\mathbf{G}}$ is the channel matrix estimate depending on the acquisition scheme employed. The MSE of different schemes is evaluated by Monte Carlo simulation, where the transmission of orthogonal pilots of length $\tau = K = 20$ is repeated over a sufficient number of independent realizations. For each large scale fading realization we carry out off-line training with $N_t = 100$ over random small scale fading to approach the optimal codebooks for our vector quantizers.

Fig. 4.6 shows the MSE of different acquisition schemes against transmit power for $L = 30$, $\tau = K = 20$, $N = 4$ and $R_N = 2$ bits/dim, equivalent to a fronthaul capacity C of 8 bits. Along with VQ-EQ and VQ-QE we also present two other schemes as baselines in which uniform Scalar Quantization (SQ) and estimation are performed at the individual antennas of the APs for both EQ and QE. For each scheme we plot three curves with different angular spread standard deviation $\sigma_\delta = 10°, 20°, 40°$, with Gaussian distributed δ. It is expected that the correlation becomes weaker as σ_δ increases. The angles of arrival θ are assumed to be random uniformly distributed in $[-\pi, \pi]$ according to the distribution of users.

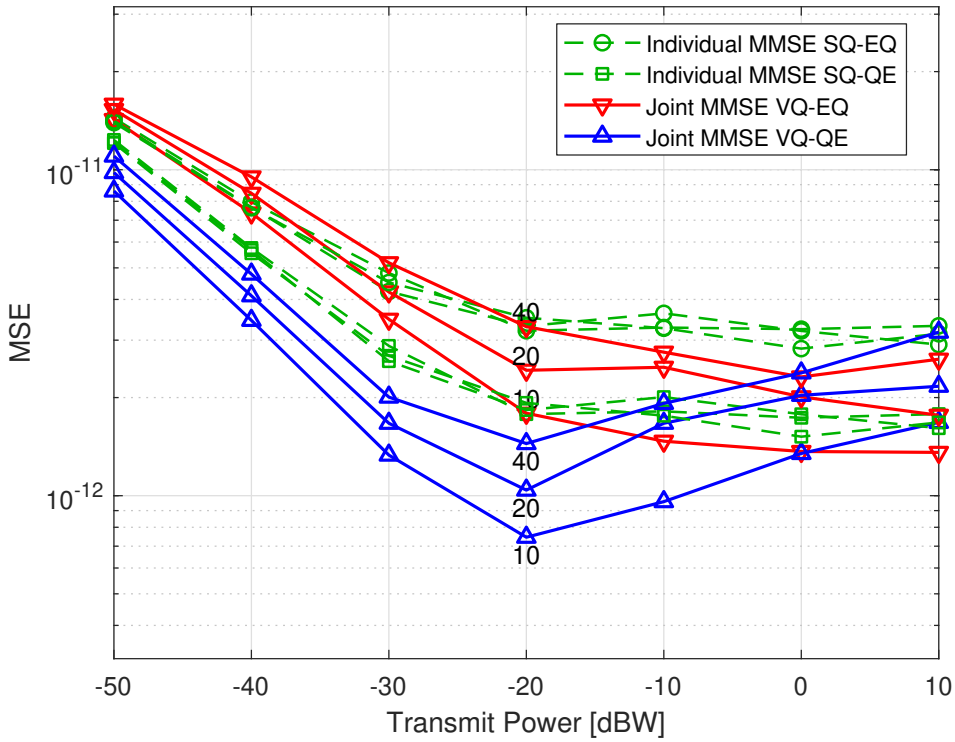


Fig. 4.6 The MSE versus transmit power for $M = 120$, $N = 4$, $K = 20$, $R_N = 2$ bits/dim and $\sigma_\delta = 10°, 20°, 40°$.

As can be observed in Figure 4.6, the joint processing schemes VQ-EQ and VQ-QE can generally provide improvements to the baseline schemes in which the individual processing is performed at the APs . For both joint processing schemes the channel estimate becomes more accurate as the channel correlation increases. It is not the case for the baseline schemes, where the estimate performance is relatively constant. As the transmit power increases up to -20 dB all schemes show improving performance as expected. It should be noted that in our simulation set-up the path losses are large which leads to small channel gains and small typical values of MSE. In all cases, the lowest MSE can be achieved by VQ-QE at -20 dB transmit power when strong spatial correlation is present. Above this power the MSE performance of the other schemes remains constant, but that of VQ-QE degrades. In this regime the quantization noise dominates the additive noise so that increasing the transmit power has little effect.

The estimator of VQ-QE is derived based on the simplifying assumption that the quantizer input and the respective quantization noise is Gaussian. Below -20 dB the additive noise is still able to make the effective noise appear Gaussian so that the estimator for VQ-QE performs quite well. But in the regime above -20 dB the correlated quantization noise across the antennas is significantly different from Gaussian which

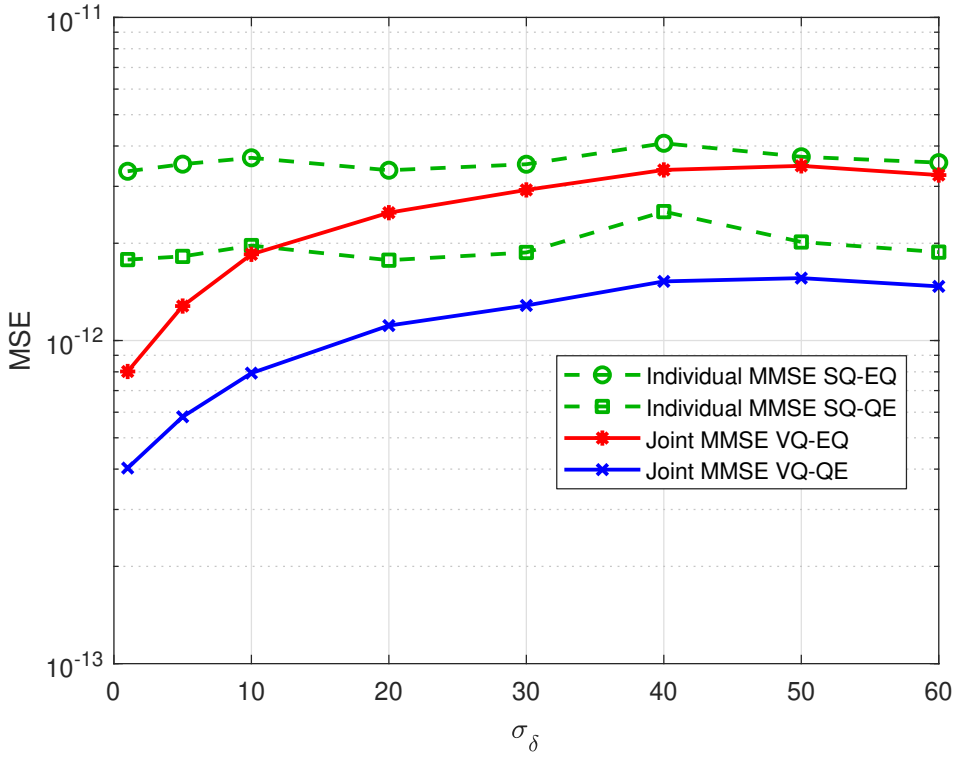


Fig. 4.7 The MSE versus angular spread standard deviation σ_δ for Gaussian distributed δ, $M = 120$, $K = 20$, $R_N = 2$ bits/dim, and TxPower=-20dB

leads to a mismatch in the estimation. This behaviour shows that an appropriate portion of noise might help to enhance the estimation accuracy in the case of model mismatch. This condition has been studied previously for instance in [63]. The Similar behaviour is also observed in [64], where one-bit channel estimation is performed for co-located massive MIMO with spatial and temporal correlation. The effect is also explained by the mismatch of the quantization noise to the Gaussian assumption of the estimator: this is more significant above -20 dB. A full discussion of this effect is however beyond the scope of the present work. Although the MSE of VQ-QE increases at higher transmit power, it is still roughly equal to the SQ-EQ scheme in the asymptotic regime.

Fig. 4.7 shows the dependence of MSE on spatial correlation (i.e. σ_δ). This figure confirms that the proposed schemes, at least at moderate SNR, can effectively exploit the strong channel correlation to achieve more accurate CSI. In the asymptotic regime, where the channels are uncorrelated, the VQ schemes can still achieve a considerable gain due to the space filling advantage obtained from the dense packing codebooks of VQ.

In addition, we investigate in Figure 4.8 the relationship of the MSE to the number of antennas per AP. We let the number of antennas N increase whereas the number of bits per antenna and the total number of antennas are fixed respectively to $R_N = 1$ and

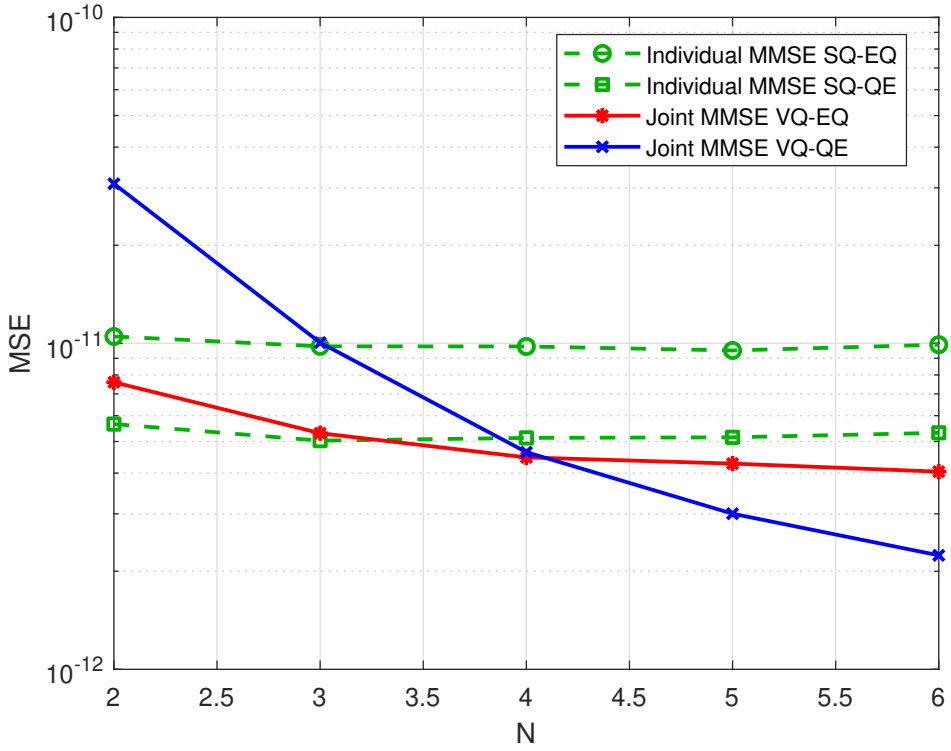


Fig. 4.8 The MSE versus number of antennas per AP N for $M = 120$, $K = 20$, $R_N = 1$ bits/dim, TxPower=-20dB and $\sigma_\delta = 10°$.

$M = 120$. In this case, increasing N means also increasing the fronthaul capacity C per AP. As we can observe in Figure 4.8, the MSE performance of the baseline schemes is independent of N. In contrast, vector quantization, especially VQ-QE, is able to exploit the correlation of the antennas at an AP to improve the CSI accuracy.

Having evaluated the CSI acquisition performance, we further evaluate numerically the achievable rate of the individual and joint processing in terms of per-user net throughput defined as

$$\mathcal{T}_{u,k} \triangleq B\frac{1-\tau_p/\tau_c}{2}R_{u,k}, \tag{4.62}$$

where the rate $R_{u,k} \in \{R^{\text{In}}_{u,k}, R^{\text{Jo}}_{u,k}\}$ is obtained from (4.20) or (4.56). In the simulation, two simple beamforming vectors $\mathbf{a}_k$ are considered for the data detection at the CPU, namely MRC and ZF. We use the same parameters as before except the total number of antenna M in the system. In this case, we use greater M to achieve higher rate in order to better distinguish the simulation results among the different schemes.

We first investigate the distribution of the user throughput for $M = 200, \sigma_\delta = 10°$ and transmit power -20 dB at which the lowest MSE is achieved by VQ-QE in Figure

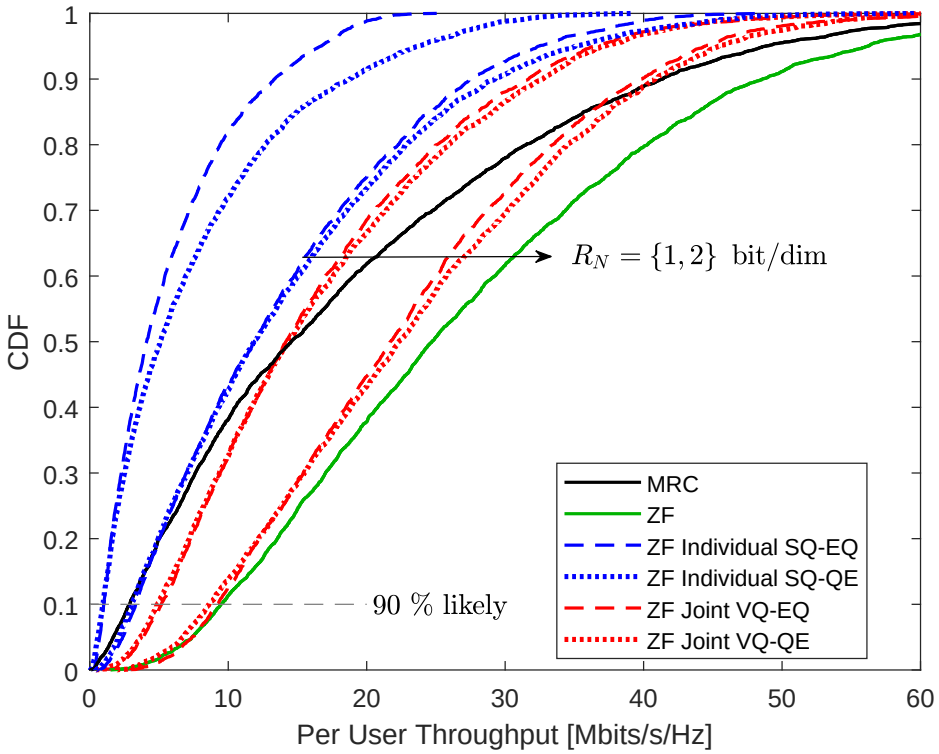


Fig. 4.9 The CDF of per user throughput for different AP processing schemes and different quantization rate per dimension with $M = 200, K = 20, N = 4$, TxPower= -20dB and $\sigma_\delta = 10^\circ$.

4.6. In this setting, the Cumulative Distribution Function (CDF) of the user throughput is shown in Figure 4.9 for different schemes including ideal MRC and ideal ZF. For the case of limited-capacity fronthaul, we use codebooks with the rate $R_N = 1, 2$ bits per dimension for the joint processing with vector quantization, and with the rate per antenna $R_m = 1, 2$ bits for the individual processing with scalar quantization. As can be seen from the CDF, the joint processing with VQ can clearly outperform the individual processing. For both schemes, the QE strategy provides the majority of the users with better throughput than the EQ strategy. However, in terms of the 90 % likely throughput the EQ strategy slightly outperforms the QE strategy. We can also observe from the figure that using the joint processing scheme at resolution 2 bit/dim each user can already achieve at least 10 Mbits/s/Hz with 90% likelihood.

In Figure 4.10, we show the simulation results for the average per-user throughput against the transmit power. As in the previous figure, the results are given for different schemes and different strategies at rates 1 bit and 2 bits per dimension. For all low-resolution schemes and strategies, the average throughputs increases with the number

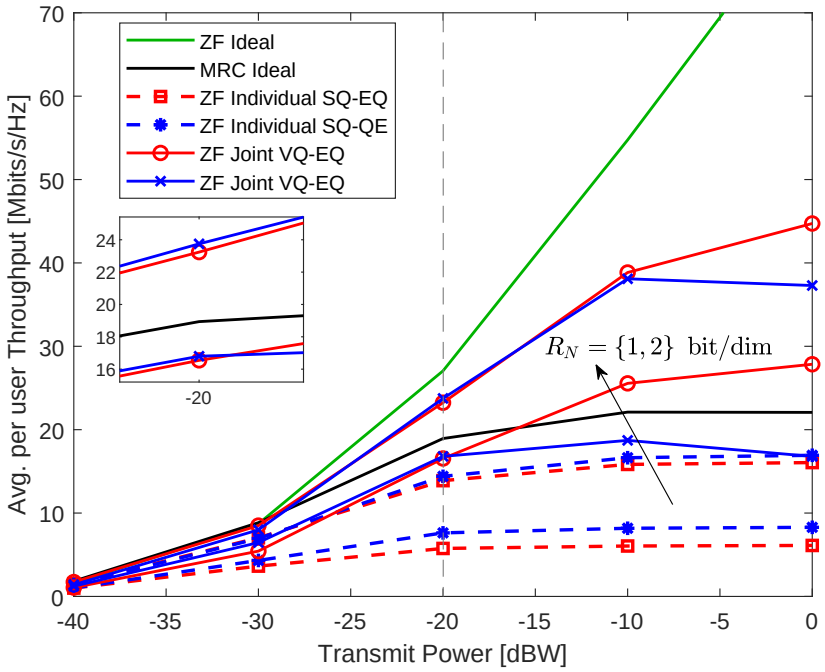


Fig. 4.10 The average per user throughput against the transmit power for different AP processing schemes and different rate per dimension with $M = 200, K = 20, N = 4$ and $\sigma_\delta = 10°$.

of bits per dimension, and increase as the transmit power increases up to certain point. Then, except the joint processing scheme with the QE strategy, the average throughputs become flat due to the dominant effect of the quantization noise either from the CSI acquisition or from the data transmission. In the other case of joint processing with the QE strategy, we see a slight decline in the average throughput with the increasing of the transmit power. We might suspect that this effect arises from the CSI acquisition of the QE strategy which has a degraded accuracy in the high power regime as previously discussed from Figure 4.6. However, the losses due to this effect are insignificant compared to the gain obtained by switching from individual processing to joint processing. As can be observed in Figure 4.10, joint processing can more than double the average throughput per user compared to the individual processing scheme at the same rate per dimension.

In the final simulation, we aim to determine the variation of the average per-user throughput with the number of antennas at APs when the number of the total of antenna in the system should be fixed. In this case, we perform a simulation with $K = 20$ users, a total of $M = 240$ antennas, at transmit power -20dB, at quantization resolution $R_N = 2$ bit/dim, and under a correlated channel with angular spread standard deviation $\sigma_\delta = 10°$. The simulation results are given in Figure 4.11 with the corresponding number

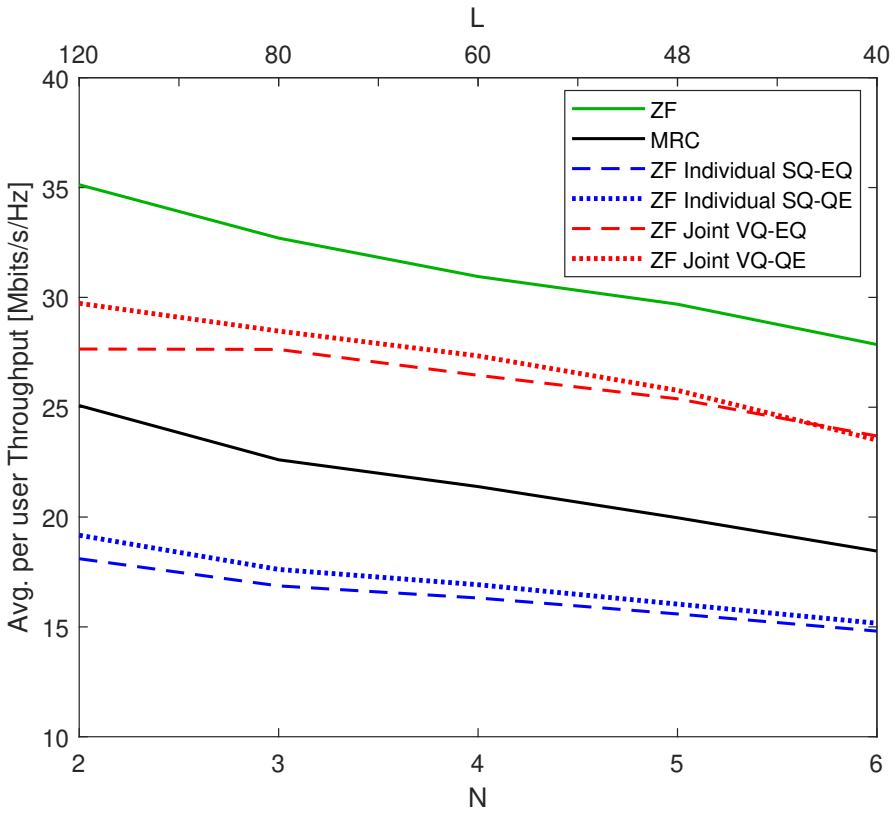


Fig. 4.11 The average per user throughput against the number of antenna per AP N (i.e. the number of APs L) for different AP processing schemes with $M = 240, K = 20$, $R_N = 2$ bit/dim, TxPower= -20dB and $\sigma_\delta = 10°$.

of APs L on the upper horizontal axis. Accordingly, as the number of antennas per AP N increases, we need only deploy a smaller number of APs L in the same service area. As shown in the figure, the average throughput of the limited-capacity schemes have a similar tendency to the ideal schemes, where the throughput decreases as the number of antenna per APs N increases. Therefore, the same explanation in the ideal scheme might underlie this similar behaviour in the limited-capacity scheme, which is the reduction of the macro-diversity due to the smaller number of APs. Nevertheless, we have observed in Figure 4.10 that the throughput can be increased by increasing the fronthaul resolution. Hence, we might need a higher resolution for the smaller number of multiple-antenna APs to maintain the same throughput. As a compensation, these fewer multi-antenna APs require less cost for the infrastructure deployment.

4.6 Summary

In this chapter, we have extended the centralized cell-free massive MIMO from the previous chapter to the case where the APs are equipped with multi antennas. To this network, we considered a channel with spatial correlation across the multi antennas, which obeys the local scattering model. We considered two schemes for the processing at the APs. First, we treated the received signal across the multi-antenna individually and used a scalar quantization for each antenna. Second, we processed the received signals across the multi-antenna jointly using vector quantization.

Under low-resolution fronthauls, we developed the CSI acquisition strategies; estimate-and-quantize (QE) and quantize-and-estimate (EQ), respectively for both schemes by making use of Bussgang decomposition. Subsequently, we derived the achievable rate for both schemes, when the CSI acquisition and the data transmission are constrained with low-resolution fronthauls. To asses the performance of our proposed scheme, we have numerically evaluated the MSE of the CSI acquisition and per-user data throughput. The results showed that the joint processing with vector quantization can improve the CSI accuracy as well as the data throughput in spatially correlated channels. We have also investigated the scalability issue for multi-antenna APs. As long as the number of user and the length of data payload are greater than the number of APs' antenna, the centralized approach with QE strategy is still scalable.

Chapter 5

Lattice Vector Quantization for Multiple-Antenna Access Points

While delivering high data rates has been the main driver for the development of wireless communication from one generation to the next, the interest to achieve this in high-mobility and high-density scenario has appeared only recently due to the new "killer apps" such as autonomous vehicles. Requirements for this new application would be high reliability and low latency in addition to high data throughput [8, 12]. A straightforward strategy such as the quantize-and-estimate with joint processing scheme discussed in the previous chapter might enable us to meet this demand as it can provide the users with high data throughput with relatively uncomplicated processing at access points. However, the primary difficulty of the proposed scheme resides in finding the optimum codebook for vector quantization, especially in a high dimension. To address this, a reasonable assumption was made in chapter 4 - that is, that the coherence time of the channel is long enough such that the optimal codebook can be obtained from sufficient data training. This optimal codebook then needs to be updated in every coherence block.

However, the above assumption can no longer be made in a high mobility scenario. This is because the coherence time becomes shorter as the users move rather rapidly. Hence, although optimum quantization can be achieved by utilizing vector quantization with optimum codebooks constructed by the LBG algorithm, it is not suitable for users with high mobility and hence has short coherence time. In this chapter, we are motivated to provide high data throughput with high mobility, where it is preferable to avoid a training based method to minimize the processing time. For this purpose, we look in this chapter at lattice vector quantization, which is intended to be applied at the multi-antenna APs to jointly quantize the received signals following the scheme in the previous chapter.

Therefore, the focus of this chapter is to devise a codebook for the above-mentioned application. In this case, we consider constructing a codebook from lattice points. Although the constructed codebook is suboptimal, as we will see later, by using a lattice we aim to find a good trade-off between performance and complexity. After giving a brief introduction to lattices in the subsequent section, we will describe the codebook design problem using a lattice. Then, we propose two procedures for constructing a lattice codebook which is fast and near-optimal. These are intended for uncorrelated and correlated channel applications, respectively. Finally, we give some numerical results which evaluate the performance of our proposed procedures.

5.1 Background on Lattices

We begin in this section with some basic concept and notation of lattices. A more comprehensive discussion of lattices can be found in [65, 66]. An N-dimensional lattices Λ is a subset of infinite points in $\mathbb{R}^N$, in which the reflection and addition operations will give another point in Λ [66]. It can be defined as

$$\Lambda = \{\lambda = \sum_{n}^{N} i_n \mathbf{b}_n : i_n \in \mathbb{Z}\} \tag{5.1}$$

$$= \{\lambda = \mathbf{B}\mathbf{i} : \mathbf{i} \in \mathbb{Z}^N\}, \tag{5.2}$$

where $\mathbf{i} = [i_1, \ldots, 1_N]^T$ is an N-dimensional integer column vector and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$ is an $N \times N$ generator matrix with its columns as linearly independent basis vectors. Because the same Λ can be generated by many different generator matrices, sometimes we explicitly write $\Lambda(\mathbf{B})$ to show that Λ is generated by matrix $\mathbf{B}$. We omit the notation $\mathbf{B}$ when it is clear from the context. For some integer matrix $\mathbf{T}$ with $|\det(\mathbf{T})| = 1$, we say $\Lambda(\mathbf{B}') = \Lambda(\mathbf{B})$ if and only if $\mathbf{B}' = \mathbf{B}\mathbf{T}$. For example, an integer or cubic lattice $\Lambda(\mathbf{I}) = \mathbb{Z}^N$ is generated by an identity matrix $\mathbf{I}$. It turns out that any lattice can be obtained from the linear transformation of the integer lattice as $\Lambda(\mathbf{B}) = \mathbf{B}\Lambda(\mathbf{I})$.

To each lattice point we can associate a non-overlapping congruent cell such that it covers the whole $\mathbb{R}^N$. The cell associated with the origin ($\lambda = \mathbf{0}$) is called the fundamental cell where a shifting of the lattice points creates a partition of $\mathbb{R}^N$. With respect to a lattice $\Lambda(\mathbf{B})$, the partition can have different shapes corresponding to the considered fundamental cell. One simple fundamental cell is a fundamental parallelotope which can be described using the generator matrix as

$$\mathcal{P}_0(\mathbf{B}) = \{\mathbf{B} \cdot [x_1, \ldots, x_N]^T : 0 \leq x_1, x_2, \ldots, x_N \leq 1\}. \tag{5.3}$$

Another common fundamental cell is a fundamental Voronoi cell which is defined for a lattice $\Lambda(\mathbf{B})$ as

$$\mathcal{V}_0(\mathbf{B}) = \{\mathbf{x} : \|\mathbf{x}\|^2 \leq \|\mathbf{x} - \lambda\|^2, \forall \lambda \in \Lambda(\mathbf{B})\}. \tag{5.4}$$

It is a set of points in $\mathbb{R}^N$ that are closer or at least equally distant to the origin rather than any other lattice point. Despite the fact that different cells make different partition, the volume of the cell is given by the determinant of its generator matrix which is independent of their shapes. In this case, we have $\text{Vol}(\mathcal{P}_0(\mathbf{B})) = \text{Vol}(\mathcal{V}_0(\mathbf{B})) = |\det(\mathbf{B})|$.

To measure how efficiently the lattice cells can cover the space, the ratio of the second moment per dimension of a uniformly distributed random variable to the volume of the cell is usually computed. It is given by

$$G(\mathcal{P}_{\text{ol}}) = \frac{1}{N} \cdot \frac{\int_{\mathcal{P}_{\text{ol}}} \|\mathbf{x}\|^2 d\mathbf{x}}{\text{Vol}(\mathcal{P}_{\text{ol}})^{1+2/N}} \tag{5.5}$$

for a polytope $\mathcal{P}_{\text{ol}}$ and called the normalized second moment. The minimum possible value of the normalized second moment

$$G_N = \min_{\mathcal{P}_{\text{ol}}} G(\mathcal{P}_{\text{ol}}) \tag{5.6}$$

is achieved by an N-dimensional sphere. For a lattice $\Lambda(\mathbf{B})$ we can compute the normalized second moment by

$$G(\Lambda) = \frac{1}{N} \cdot \frac{\int_{\mathcal{V}_0} \|\mathbf{x}\|^2 d\mathbf{x}}{\text{Vol}(\mathcal{V}_0(\mathbf{B}))^{1+2/N}}. \tag{5.7}$$

5.2 Lattice Quantizer Design

We recall that quantization can be seen as a mapping process from an input signal to the elements of a countable set known as the codebook by which transmission is more efficient. Therefore, the most important aspect of designing a quantizer consists of defining the mapping rules and choosing the codebook appropriately. In lattice quantization, we aim to represent the input signal using the lattice points as the codewords in our codebook. The primary reason underlying this choice is usually due to the highly-ordered structure possessed by the lattice. It makes a lattice quantizer less complex for the implementation, especially in high dimensions.

5.2.1 Codebook Construction

Using lattice points for the codebook is however not straight-forward. As there are many types of lattice, there arises first the question of which lattice is good for quantization. For an infinite lattice and uniformly distributed input signal, Conway and Sloane have proposed in [67] to use the normalized second moment $G(\Lambda)$ given by (5.7)as a metric. Since $G(\Lambda)$ can be interpreted as the mean squared quantization error per symbol, the lower the normalized second moment the better the lattice for quantization. Based on this criteria, the best lattices have been found in [67] for quantization in $\mathbb{R}^N$ up to the dimension $N = 10$. They are lattices with the lowest $G(\Lambda)$ where the hexagonal or A_2 lattice for instance is the best in two dimension and D_4 is the best in four dimensions. Further, the normalized second moment of lattice $G(\Lambda)$ is lower bounded by the normalized second moment of the N-sphere G_N. As calculated in [67], G_N decreases as the dimension N increases, and converges to $1/(2\pi e)$ as N goes to infinity. Interestingly, as shown in [66, 68], there exists a lattice Λ_N for which

$$\lim_{N\to\infty} G(\Lambda_N) = G_N = \frac{1}{2\pi e}. \tag{5.8}$$

Thus, using a lattice in a high dimension for the quantization codebook is an attractive option.

To use lattice points in practice, we still need a further treatment. This is because a lattice has intrinsically an infinite number of points, whereas a countable and finite set of codewords is required for an efficient codebook. Hence, we should pick a subset of lattice points to be included in the codebook: this is usually called truncation or shaping of the lattice. It can be done in different ways which can be categorized in at least two methods namely the probabilistic or "soft" shaping and the geometric or "hard" shaping [66]. One parameter determining the size of the shaping is the rate of the codewords. In the first method, the codewords are assigned with variable rate where the distant lattice points with less probability of occurrence are omitted. An entropy coding such as the Huffman code can be made use in this method. For the implementation of this method, it is necessary to think about the limitation of entropy coding which is the need for buffer feedback [40].

In the second method, which we consider in this chapter, the lattice points are selected using a shape that can have any arbitrary geometry. Only the lattice points that lie in the intersection with this shape are included in the codebook. To be more specific, suppose that we use a shape $\mathcal{U} \subset \mathbb{R}^N$ and want to select some lattice points from the

lattice $\Lambda(\mathbf{B})$, then the codebook is given as

$$\mathcal{C} = \mathcal{U} \cap \Lambda(\mathbf{B}). \tag{5.9}$$

In this case, we truncate the lattice points of $\Lambda(\mathbf{B})$ which are not contained in $\mathcal{U}$. Since $\mathcal{U}$ is a finite subset of $\mathbb{R}^N$, an efficient codebook with fixed rate codewords can be obtained. To do so, we require next to specify $\mathcal{U}$ and $\Lambda(\mathbf{B})$, where some care should be taken to deal with the granular and overload distortion. We define overload distortion as the error caused by an input signal that falls in the overload region. That is the region outside the shape $\mathcal{U}$. Conversely, the granular distortion is the distortion caused by the input signal that falls inside the shape $\mathcal{U}$ called granular region.

The problem of minimizing the granular and overload distortion of the geometric shaping codebook has been studied in [69] for non-uniformly distributed input signals. The density and the arrangement of the lattice points inside the granular region are the determining factors that contribute to the distortion. On the other hand, the overload distortion is primary determined by the ability of the shape $\mathcal{U}$ to fit the distribution of the input signal. Since the overload and the granular distortion are coupled, the overall distortion minimization is not easy to find analytically. In [69], the minimum MSE counted in granular and overload distortion were found numerically, but an expression to specify $\mathcal{U}$ and $\Lambda(\mathbf{B})$ for minimum distortion were missing. A further study with more practical perspective has been done in a technical report [70] for a Gaussian input signal. An N-sphere is used in [70] as the shape $\mathcal{U}$ which is later scaled again to fit the desired rate and the input signal variance. The problem of finding the optimum radius of the shape which minimise the overall distortion was posed in the report. For a given dimension N and rate per dimension R_N the optimum radius of the sphere is given as

$$a_{\text{opt}} = \sqrt{\frac{4R_N \ln(2)}{N}}. \tag{5.10}$$

In the next subsections, we will use part of the result in [70], such as the optimum radius, to construct our codebook.

5.2.2 Near-Optimum Codebook for Uncorrelated Channel

As mentioned previously, we wish to devise a vector quantization scheme that has low complexity and sufficiently fast processing for our desired application. For this purpose, we attempt to make use of the structural advantage of a lattice, where we consider the construction of the quantization codebook using geometric shaping. However, we do not

follow the approach given in [70], because the scaling rule proposed in the report is based on the data training. Instead, we construct the codebook in this chapter using a shape $\mathcal{U}$ that has the form of a Voronoi region of a coarse lattice as proposed by Conway and Sloane for the first time in [71].

The main advantage of using the Voronoi shape is the availability of a fast encoding, indexing, and decoding algorithm enabled by exploiting the algebraic structure of the lattice [71, 72]. We note that the terminology encoding and decoding are switched in many works due to the transmission and channel coding perspective. On the other hand, we use the source coding perspective following the description in Section 2.2 as illustrated in Figure 2.2. In this case, encoding means finding the closest lattice point given an arbitrary point in $\mathbb{R}^N$, whereas indexing is a procedure of assigning index to the lattice points and conversely decoding is finding the lattice points from a given index. We will make some modification to the procedure such that a near optimal codebook can be obtained.

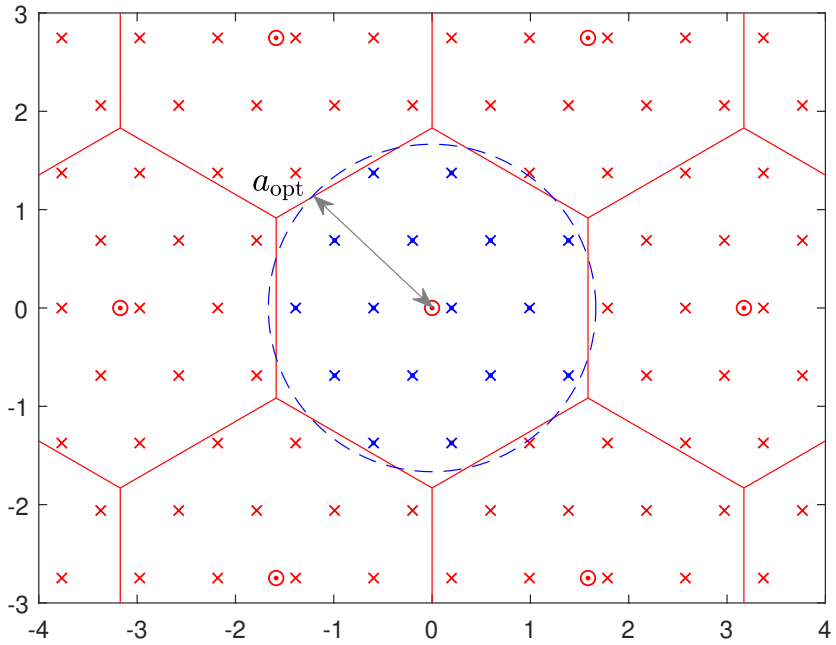


Fig. 5.1 An illustration of constructing a lattice Voronoi codebook using a fine lattice Λ_{f} (cross points) and a coarse lattice Λ_{c} (dots). The optimal sphere radius a_{opt} is shown by the dashed blue lines.

To be more precise, let us assume that we have an independent identically distributed Gaussian input signal $\mathbf{x} \in \mathbb{R}^N$ with zero mean and unit variance. Suppose that we use a codebook $\mathcal{C}_s$ of size S to quantize $\mathbf{x}$. Then, $\mathcal{C}_s$ is constructed by using a fine lattice $\Lambda_{\text{f}}(\mathbf{B})$ and a shape from the Voronoi cell $\mathcal{V}_{\text{c}}(\mathbf{B})$ around a coarse lattice $\Lambda_{\text{c}}(\mathbf{B})$. For a zero-mean

input signal, it is reasonable to use the fundamental cell $\mathcal{V}_{0c}(\mathbf{B})$. Further, it is required for the coarse lattice to be a subset of a fine lattice and satisfy the relation

$$\Lambda_c(\mathbf{B}) = r\,\Lambda_f(\mathbf{B}), \tag{5.11}$$

where r is an integer determining the codebook size $S = r^N$. Hence, the quantization rate per dimension is given by

$$R_N = \frac{1}{N}\log_2 S = \log_2 r \quad [\text{bit/dim}]. \tag{5.12}$$

Unless otherwise stated, we consider subsequently a codebook that is constructed by lattices Λ_c and Λ_f with the same generator matrix. The specific form of our codebook can be then expressed as

$$\mathcal{C}_s = (\mathcal{V}_{0c} + \mathbf{v}) \cap (\mu\,\Lambda_f), \tag{5.13}$$

$$= \{\mathbf{x}_s = \mu\lambda_f : \|\lambda_f\|^2 < \|\lambda_f - (\lambda_c + \mathbf{v})\|^2, \forall\,\lambda_f \in \Lambda_f \text{ and } \lambda_c \in \Lambda_c\}. \tag{5.14}$$

where a small shift $\mathbf{v}$ is sometimes required to avoid some lattice points lying over the shape boundary. The length of the shift $\mathbf{v}$ can be chosen based on many different criteria such as to minimize the total energy as given in [71]. Other than that, we can choose any random number that is small compared to the lattice basis. Here, we choose the latter as suggested in [70] for the sake of simplicity.

Algorithm 3: Codebook Construction (Uncorrelated)

input : Dimension N and rate per dimension R_N
output : Matrix $\mathbf{X}_s$ whose s-th row is a code $\mathbf{x}_s \in \mathcal{C}_s$

1 Choose generator matrix $\mathbf{B}$ with the lowest $G(\Lambda_N)$;
2 Compute the index candidates $\tilde{\mathcal{J}} = \prod_{n=1}^{N} \{0, 1, \ldots, (r-1)\}_n$, where $r = 2^{R_N}$;
3 Compute $\tilde{\mathbf{X}}_s = \mathbf{B}\tilde{\mathcal{J}}^T$ and $\mathbf{Z} = (\tilde{\mathbf{X}}_s - \mathbf{v})r^{-1}$, where $\mathbf{v}$ an arbitrary small shift (5.13);
4 Find the closest lattice point $\lambda_s \in \Lambda_f$ to each point of $\mathbf{Z}$ (using algorithm in [72]);
5 Compute $\mathbf{X}_s = \mu(\tilde{\mathbf{X}}_s - r\boldsymbol{\lambda}_s - \mathbf{v})$, where μ is given by (5.15);

In addition, we scale Λ_f by μ to obtain a truncation within a spherical region. The radius of the sphere is that at which the volume of N-dimensional Voronoi cell is equal to the volume of the N-sphere. In this case, we wish to find a scale μ such that the sphere radius is the optimal radius given by (5.21). This is illustrated in Figure 5.1 for the case of two dimensions, where we use the hexagonal lattice for both Λ_c and Λ_f. The lattice points of Λ_c are shown by the red dot points, whereas the lattice points of Λ_f are

shown by the cross points. We want to set the sphere radius corresponding to $\mathcal{V}_{0c}$ to be the optimum sphere radius shown by the dashed blue line. For this, the scale μ can be calculated from

$$\begin{aligned}
\mathrm{Vol}(\mathcal{V}_{0c}(\mathbf{B})) &= \mathrm{Vol}(N\text{-sphere at } a_{\mathrm{opt}}) \\
\det(\mu\, r\, \mathbf{B}) &= V_N a_{\mathrm{opt}}^N \\
\mu^N r^N \det(\mathbf{B}) &= V_N a_{\mathrm{opt}}^N \\
\mu &= \sqrt[N]{\frac{V_N a_{\mathrm{opt}}^N}{r^N \det(\mathbf{B})}} \\
\mu &= \frac{a_{\mathrm{opt}}}{r} \sqrt[N]{\frac{V_N}{\det(\mathbf{B})}},
\end{aligned} \tag{5.15}$$

where V_N is the N-dimensional unit ball given by

$$V_N = \frac{(\pi)^{N/2}}{(N/2)!} = \frac{2^N (\pi)^{\frac{N-1}{2}} \left(\frac{N-1}{2}\right)!}{N!}. \tag{5.16}$$

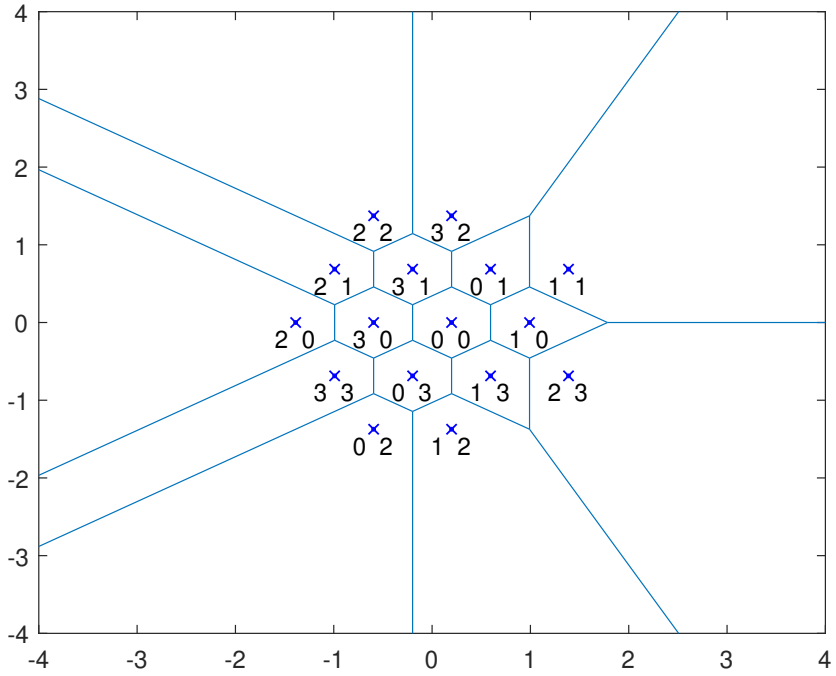


Fig. 5.2 The resulting lattice Voronoi codebook obtained by Algorithm 3 for dimension $N = 2$, codebook size $S = 16$ or rate per dimension $R_N = 2$ bit/dim.

To be mapped into a codeword, each lattice point $\mathbf{x}_s \in \mathcal{C}_s$ should be associated with a unique index $\mathbf{s}$. The set of indices for the codebook $\mathcal{C}_s$ can be written as a matrix

$$\begin{aligned} \mathcal{J} &= [\mathbf{s}_1, \ldots, \mathbf{s}_S]^T, \text{ where} \\ \mathbf{s} &= (s_1, \ldots, s_N) \text{ and } s_n \in \{0, \ldots, (r-1)\}. \end{aligned} \tag{5.17}$$

In Algorithm 3, we give the procedure for constructing a lattice codebook for an uncorrelated input signal. We should note in step 4 that the algorithm for finding the closest λ_s is given in [72] according to the type of lattice. In this case, we should use the algorithm intended for the lattice that we have chosen in step 1. Accordingly, Figure 5.2 depicts an example of the resulting codebook together with the assigned indices for the case of two dimensions with $R_N = 2$ bit/dim. Having constructed the codebook, we can use it for encoding based on the maximum likelihood method with the decision boundary shown in the figure by the blue lines.

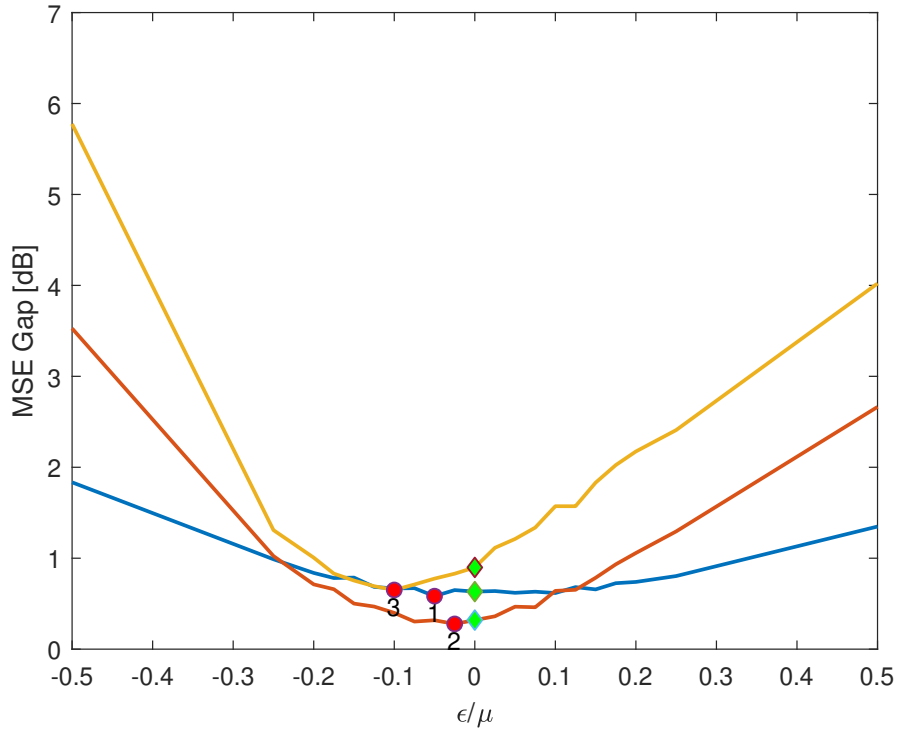


Fig. 5.3 The distortion gap between the lattice quantization using the Voronoi lattice codebook and the optimum quantization (LBG) in relation to a small change of the scale μ by ϵ.

We are interested next to see how far the constructed codebook is from optimal. To that end, we define a metric MSE gap as the absolute MSE difference between the lattice quantization codebook $\mathcal{C}_s$ and the optimal quantization codebook obtained from the

LBG algorithm. Further, we do a variational analysis on the MSE gap in which we make a numerically small change ϵ to the scale μ. The numerical evaluation for the case of two dimensions and for different quantization rates per dimension R_N is depicted in Figure 5.3. The three curves correspond to different numbers of bits per dimension indicated by the numbers $1, 2, 3$. The red dots are the points where the MSE gap has a minimum. As shown in the figure, they have only a relative small difference from the green dots, which are the scale μ obtained from (5.15). Up to 3 bits per dimension the MSE gap is still below 2 dB. As the number of bits per dimension increases, the MSE gap becomes more sensitive to the scale perturbation as shown by the steeper descent of the curves.

5.2.3 Near-Optimum Codebook for Correlated Channel

In the previous subsection, we have designed a codebook which is intended for an uncorrelated input signal. Hence, applying such a codebook might be far from optimal if the channel between the users and the multi-antenna AP is correlated. On the other hand, we have seen from the previous chapter that the codebook obtained by LBG algorithm can exploit the spatial channel correlation very well. Motivated by the previous results, we wish in this subsection to design a codebook for a correlated input signal while keeping the benefit of the lattice. In particular, we aim to exploit the correlation without the need for training.

For the above-mentioned purpose, we consider a two-step process where we perform lattice quantization using an ellipsoid codebook preceded by the Karhunen-Loeve (KL) transform. Given the correlation matrix of the input signal $\mathbf{x} \in \mathbb{R}^N$, denoted by $\mathbf{R}_x$, then the KL-transformed signal is given by [40]

$$\mathbf{y} = \mathbf{V}^T\mathbf{x}, \text{ where } \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N] \tag{5.18}$$

is the KL-transformed matrix whose columns are the eigenvectors of $\mathbf{R}_x$. This transformation linearly maps the input into its orthogonal components which can further simplify the following quantization process. Due to this transformation, we may classify this approach as transformed vector quantization or transformed lattice quantization in our special case. Another type of transformation might be more suitable for another different scenario. After the quantization, we need then to transform back the output signal using the inverse transform matrix $\mathbf{V}^{-1}$.

Since the input signal $\mathbf{x}$ is correlated, the transformed signal $\mathbf{y}$ has some dominant components. To support this, we design a codebook $\mathcal{C}_e$ using a lattice shaping that has an ellipsoidal shaping boundary. We use essentially a similar technique to that described

Algorithm 4: Codebook Construction (Correlated)

input : Dimension N, rate per dimension R_N, Correlation coefficient ρ
output : Matrix $\mathbf{X}_e$ whose e-th row is a code $\mathbf{x}_e \in \mathcal{C}_e$

1 Choose generator matrix $\mathbf{B}$ with the lowest $G(\Lambda_N)$;
2 Find the appropriate vector $\mathbf{r}$ using (5.25) and satisfying constraints in Table 5.1;
3 Compute the index candidates $\tilde{\mathcal{J}} = \prod_{n=1}^{N} \{0, 1, \ldots, (r_n - 1)\}$;
4 Compute $\tilde{\mathbf{X}}_e = \mathbf{B}\tilde{\mathcal{J}}^T$ and $\mathbf{Z} = (\tilde{\mathbf{X}}_e - \mathbf{v}) \div \mathbf{r}$, where $\mathbf{v}$ an arbitrary small shift (5.20);
5 Find the closest lattice point $\lambda_{\mathrm{e}} \in \Lambda_{\mathrm{f}}$ to each point of $\mathbf{Z}$ (using algorithm in [72]);
6 Compute $\mathbf{X}_e = \boldsymbol{\mu} \odot (\tilde{\mathbf{X}}_e - \mathbf{r} \odot \boldsymbol{\lambda}_{\mathrm{e}} - \mathbf{v})$, where $\boldsymbol{\mu}$ is given by (5.21);

Table 5.1 The admissible vector $\mathbf{r}$ [73]

Lattice	Constraint on $\mathbf{r}$
A_2	r_1 and r_2 have the same parity (i.e. $\mathbf{r} \in D_2$)
$D_n\,(n \geq 2)$	$r_1, \ldots, r_N$ all share the same parity (i.e. $\mathbf{r} \in 2D_n^*$)
$2D_n^+\,(n$ even $\geq 4)$	$r_1, \ldots, r_N$ all share the same parity and $\sum_{n=1}^{N} r_n$ is a multiple of 4 ($\mathbf{r} \in 2D_n^+$)

in [73] which is actually a generalization of the Voronoi shaping in [71]. Hence, the legacy of fast processing from [71] is still preserved. To be more specific, consider a lattice

$$\Lambda_{\mathrm{e}} = \mathbf{r} \odot \Lambda_{\mathrm{f}}, \tag{5.19}$$

where $\mathbf{r}$ is a vector with positive integer elements. Subsequently, we denote $\odot$ and $\div$ as a pointwise multiplication and a pointwise division. Then, a Voronoi shaping $\mathcal{V}_{0\mathrm{e}}(\Lambda_{\mathrm{e}})$ is used, and the codebook can be expressed as

$$\mathcal{C}_e = (\mathcal{V}_{0\mathrm{e}}(\Lambda_{\mathrm{e}}) + \mathbf{v}) \cap (\boldsymbol{\mu}\, \Lambda_{\mathrm{f}}), \tag{5.20}$$

where $\mathbf{v}$ is a small shift vector. To ensure an optimal sphere radius for each dimension, we now have $\boldsymbol{\mu}$ as a vector given by

$$\boldsymbol{\mu} = \frac{\mathbf{a}_{\mathrm{opt}}}{\mathbf{r}} \sqrt[N]{\frac{V_N}{\det(\mathbf{B})}}, \text{ where } \mathbf{a}_{\mathrm{opt}} = \sqrt{\frac{4\log_2(\mathbf{r})\, \ln(2)}{N}}. \tag{5.21}$$

In Algorithm 4, we describe the formal procedure for constructing the codebook. We modify the procedure in the previous subsection, but we do not change the key steps.

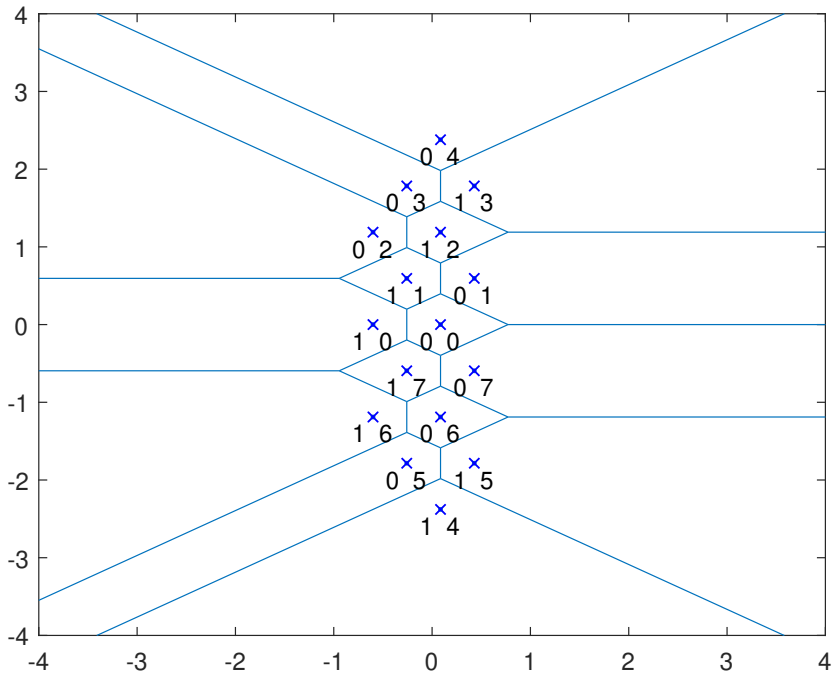


Fig. 5.4 The resulting ellipsoid Voronoi codebook obtained by Algorithm 4 for dimension $N = 2$, Codebook size $S = 16$ and $\mathbf{r} = [2, 8]^T$.

However, extra care should be taken in choosing the appropriate vector $\mathbf{r}$ in the second step. It should satisfy the constraint in Table 5.1 for a given type of lattice. Furthermore, the codebook size is now determined by

$$S = \prod_{n=1}^{N} r_n. \tag{5.22}$$

Hence, the choice of vector $\mathbf{r}$ can only be taken from the set

$$\mathcal{I} = \mathcal{D}(N, D) = \{\mathbf{r}_1, \ldots, \mathbf{r}_I\}, \tag{5.23}$$

which is the set of N-length vectors from all possible divisor combinations

$$\mathcal{D} = \text{divisor}(S) = \{d_1, \ldots, d_D\}. \tag{5.24}$$

The next task is to select $\mathbf{r}_i \in \mathcal{I}$ that best matches the correlation of the input signal. To that end, we compute

$$\mathbf{r} = \underset{\mathbf{r}_i \in \mathcal{I}}{\text{argmin}} \left\| \left(\frac{\mathbf{r}_i}{\|\mathbf{r}_i\|} \right) - \left(\frac{\mathbf{u}}{\|\mathbf{u}\|} \right) \right\|, \text{ where } \mathbf{u} = [u_1, \ldots, u_N]^T \tag{5.25}$$

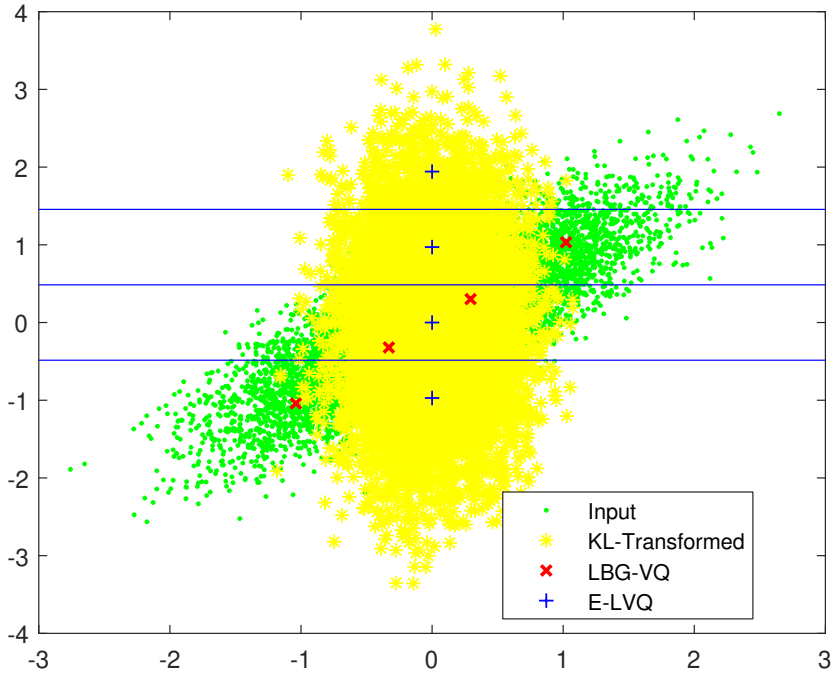
is a stack of singular values from the correlation matrix $\mathbf{R}_x$. We show the resulting codebook in Figure 5.4 for the example of two dimensions, $S = 16$ and $\mathbf{r} = [2, 8]^T$.

5.3 Performance Evaluation

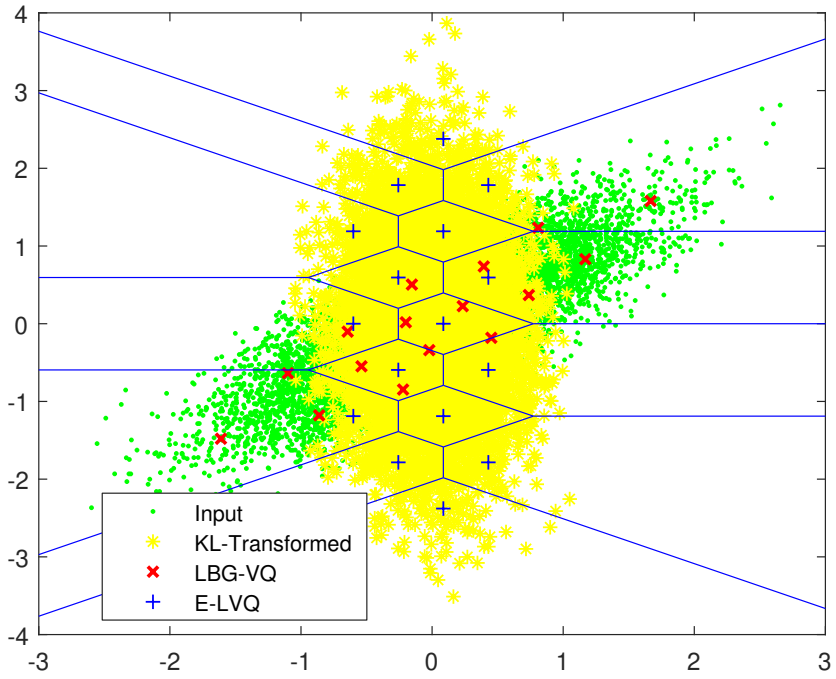
In the following discussion, we evaluate the proposed algorithms by simple simulations, in which quantization in two dimension will be used as a showcase. We consider a correlated Gaussian input signal with different degrees of correlation $0 \leq \rho \leq 1$. A strongly correlated signal is indicated by a large coefficient ρ. This is shown pictorially in Figures 5.5 for $\rho = 0.8$, where the green input signals are relatively concentrated on a diagonal showing a linear dependency between the two dimensions. We compare first in both figures the condition when the input signals are quantized using different codebooks and different rates per dimension.

In Figure 5.5a is first depicted the case when we are asked to quantize an input signal with $\rho = 0.8$ using a quantizer with available codebook size $S = 4$. We see in the figure that the optimal arrangement of the codebook points obtained by the LBG algorithm is, in this case, in the form of a line. The figure depicts also the KL-transformed input signals, which are shown by the yellow dots, overlaying the untransformed input signals, which are shown by the yellow dots. Similarly, the ellipsoidal lattice codebook algorithm places the codebook points in a line by selecting the vector $\mathbf{r} = [1, 4]^T$. By allowing a larger codebook size the LBG algorithm can place the codebook points more densely around the origin following the distribution of the input signal as illustrated in Figure 5.5b. On the other hand, the algorithm for the ellipsoidal lattice codebook can only arrange the codebook points uniformly. This implies that the probability of a larger granular distortion is increased. However, the points are still following the boundary of the transformed input signal, which prevents the overload distortion becoming too large. Through the simulation, we aim to determine how much the sub-optimum arrangement of the codebook points can affect the performance of the quantizer.

Figure 5.6 depicts the simulation results for moderately correlated input signals. The distortion for the ellipsoid lattice codebook vector quantization (E-LVQ) is compared with the optimum codebook vector quantization (LBG-VQ) in terms of the MSE at a different rate per dimension R_N. As a baseline for the simulation, the distortion of the lattice codebook vector quantizer (LVQ) obtained by Algorithm 3 is also depicted in the figure. From Figure 5.6a we can observe that at moderate correlation the gap between E-LVQ and LBG-VQ is very small at low rate, but increases as the rate increases. On the other hand, the gap between LVQ and E-LVQ decreases as the rate increases. As we



(a) $S = 4$, $\mathbf{r} = [1, 4]^T$.



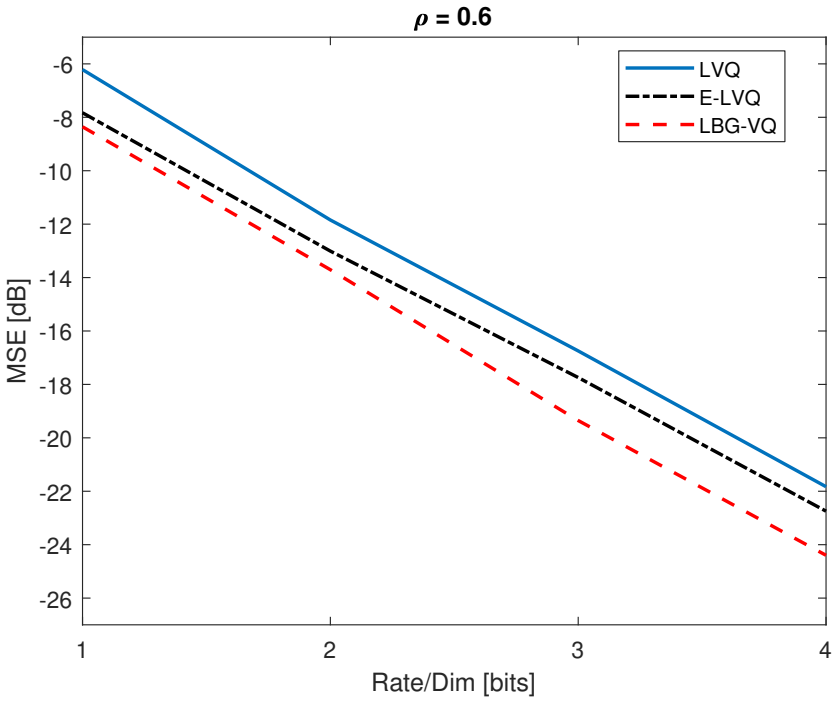
(b) $S = 16$, $\mathbf{r} = [2, 8]^T$.

Fig. 5.5 An illustration of the codebook points arrangement in relation to the input signals for different available codebook size S and for correlation $\rho = 0.8$.

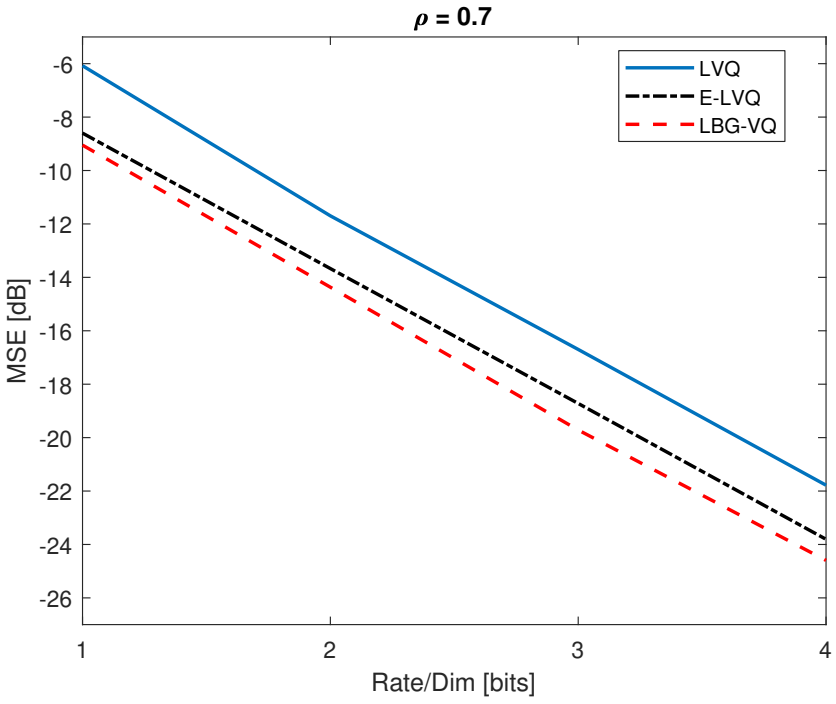
increase the correlation ρ to 0.7 in Figure 5.6b, the gap between E-LVQ and LBG-VQ is smaller at higher rates, and the gap between LVQ and E-LVQ getting larger overall the rate. Further, Figure 5.7 provides the simulation results for a rather strong correlation. At correlation $\rho = 0.8$, an additional MSE improvement is shown in Figure 5.7a for the E-LVQ and LBG-VQ, whereas the MSE of LVQ does not appear to change as previously. Although in the presence of stronger correlation, the MSE of LVQ relatively stays constant as shown in Figure 5.7b. In another case, E-LVQ and LBG-VQ are further improved, although the gap between them becomes larger at a higher rate. Our observation from the figures reveals that the E-LVQ can adapt to the correlation relatively well, while the LVQ fails to exploit the correlation. Further, the performance difference between E-LVQ and LBG-VQ is still acceptable, not exceeding 2 dB, especially at a low rate per dimension. A near-optimum MSE performance can be achieved when the E-LVQ operates at low-resolution.

5.4 Summary

We have identified the problem of LBG vector quantization for multi-antenna APs described in the previous chapter. That is, the processing time is high, which makes it unsuitable for communication in a high mobility scenario. In this chapter, we have suggested the use of a lattice vector quantization for such a scenario. We described the lattice quantizer design problem, which boils down to the codebook construction problem. For the uncorrelated and correlated channel scenarios, we proposed a procedure for constructing the lattice codebook, which is based on Voronoi shaping. In this case, the fast encoding and decoding algorithms from Conway and Sloane [71, 72] were modified. For both scenarios, we have formulated a general expression of a scale factor, which allows the codebook to be near-optimal. We have made use of the ellipsoidal lattice codebook for the correlated channel scenarios and suggested an additional KL transformation in the quantization process. The performance of the proposed codebooks for uncorrelated and correlated input signals were evaluated using numerical simulation. The results showed that the performance gaps of the constructed codebooks to the optimum codebooks are still acceptable. Further, we have demonstrated that the ellipsoidal codebook can utilize the correlation to improve the performance.

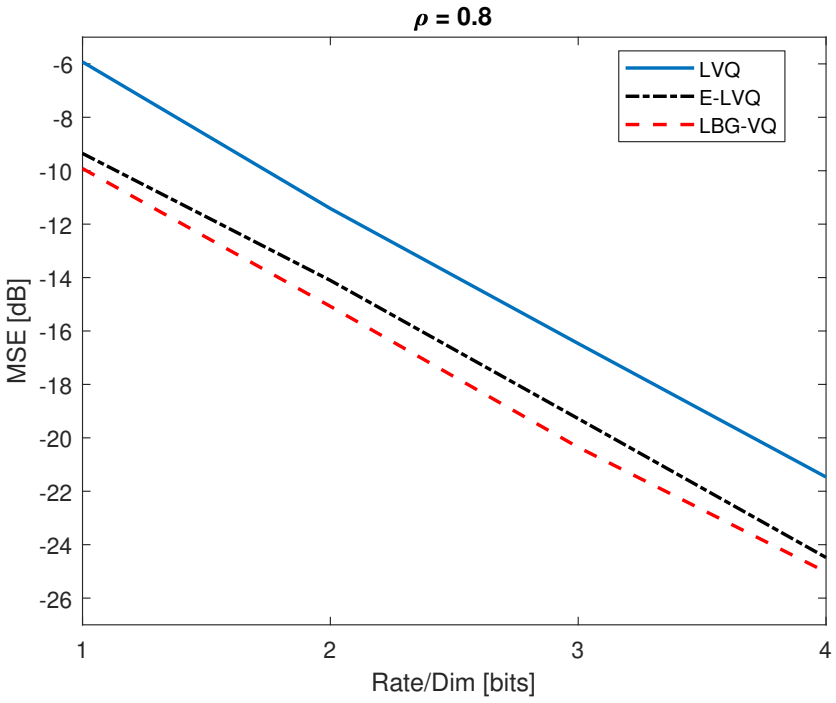


(a)

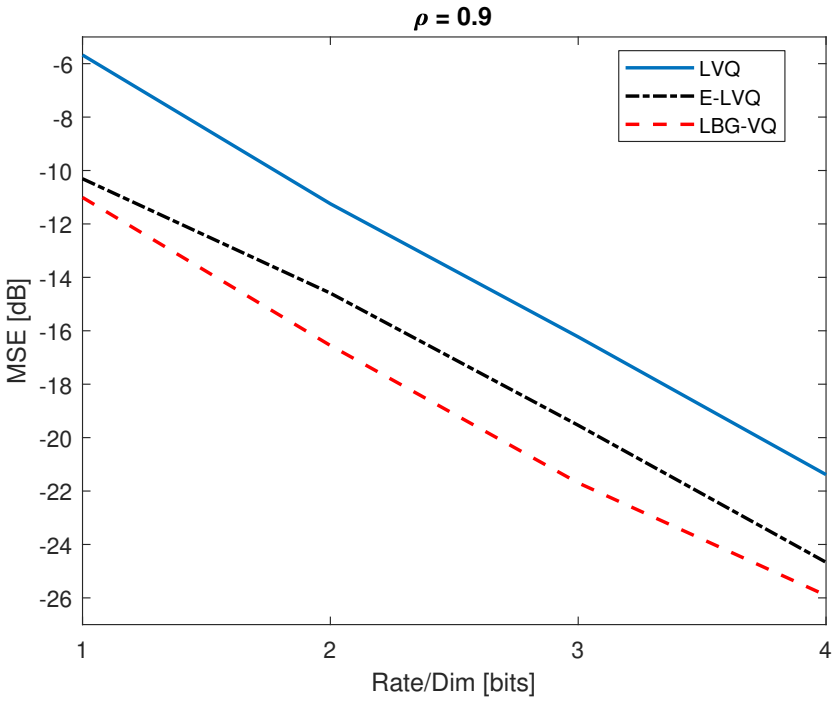


(b)

Fig. 5.6 The MSE against the number of bits/dimension between LBG vector quantization and lattice vector quantization and Ellipsoid vector quantization for dimension 2 and correlation $\rho = 0.6$ and 0.7.



(a)



(b)

Fig. 5.7 The MSE against the number of bits/dimension between LBG vector quantization and lattice vector quantization and Ellipsoid vector quantization for dimension 2 correlation $\rho = 0.8$ and 0.9.

Chapter 6

Conclusion and Future Research

Conclusion

In this thesis, we have studied the centralized approach to cell-free massive MIMO with low-resolution fronthaul links. The significant importance of this study lies in carrying out cell-free massive MIMO to its greatest advantage by taking into account the practical consideration. In general, we have demonstrated that the centralized approach can much improve the throughput performance even if the fronthaul links are subject to limited capacity. We have also revealed the contradiction to the widely accepted premise, which says, that the distributed approach is better than the centralized approach in case of fronthaul signalling.

In particular, we have developed some schemes as well as strategies to enable the centralized approach with low-resolution fronthauls. Their performances have been analysed and assessed using Bussgang decomposition for two cases, which are the single-antenna APs and the multi-antenna APs. We have shown for the first case, that the centralized approach using ZF detection with 2 bits resolution can already outperform the throughput of the original cell-free massive MIMO with ideal fronthaul. Moreover, the ZF scheme using the simpler QE strategy outperforms the EQ strategy at low-resolution, especially for 1-bit. In terms of fronthaul load and AP processing, we have revealed that the centralized approach is more scalable than the distributed approach, and the QE strategy is more scalable than the EQ strategy with an increasing number of users. For the second case of multi-antenna APs, we compare two different schemes for the processing at the APs, namely individual processing and joint processing across the multi antennas. We have demonstrated that the joint processing with vector quantization can improve the CSI accuracy as well as the data throughput in spatially correlated channels. We have also investigated the scalability issue for multi-antenna APs. As long as the

number of user and the length of data payload are greater than the number of APs' antenna, the centralized approach with QE strategy is still scalable.

This thesis has further advocated for the use of lattice vector quantization at the APs to enable high data throughput transmission in high-mobility and high-density scenarios. For this purpose, we have designed a lattice vector quantizer, which reduces to the codebook design problem. Based on Voronoi shaping, our fast constructed codebooks have shown a near-optimal performance in terms of MSE. Further, we have demonstrated that the ellipsoidal codebook can utilize the correlation to improve the performance.

Future Research

We realize that this thesis has left many open questions and open problems outstanding. However, these may also create opportunities for further research. We list below some of the potential research directions which can be carried out for the future.

- **Extension of lattice quantization**: Although the proposed codebook procedures in Chapter 5 should be valid for any dimensions, we have investigated their performance in the case of two dimensions for the reason of perceptibility and simplicity. Therefore, it remains to verify the performance in high dimensions. We have also adopted only the geometric approach for codebook construction. An alternative such as the joint probabilistic and geometric design would be also interesting for further investigation. Moreover, we have designed the codebook based on the source characteristics following the source coding perspective. The fact that the APs act essentially as relay, and the medium of the fronthaul may affect the performance, it might be worth designing the codebook from a source-channel coding perspective or even a secure source-channel coding perspective.

- **Fractional bit resolution**: As mentioned in [40], an interesting feature of VQ particularly at low resolution is the ability to quantize with fractional bit resolution per dimension. Therefore, quantizing the received signal from N multi-antenna AP using $(N-1)$ dimensional VQ is an attractive option to further reduce the fronthaul load and the power consumption.

- **Generalization of Bussgang theorem**: In Chapter 4, we have modelled the low-resolution vector quantization using Bussgang decomposition. Unfortunately, the closed-form expression for the Bussgang model is still missing. Therefore, it might be required to make use of a more general theorem such as Price's theorem

[74] to analyse the correlated distortion of vector quantization or lattice vector quantization.

- **Optimum AP density**: As we have seen in Chapter 4, it is not yet clear what is the optimum AP density and number of antenna per APs in the centralized cell-free massive MIMO subject to the fronthaul resolution, power consumption, and other aspects. To deal with this, we might use the tools from stochastic geometry, which models the randomness in the space in a more systematic way.

- **Pilot design**: Although the QE strategy is more scalable than the EQ strategy, the performance of the QE strategy relies on the design of the pilot sequence and pilot assignment. In this case, a set of robust pilot sequences is required. Research into solving this problem has been already underway.

- **Centralized cell-free Milimeter-wave**: As mentioned previously, cell-free massive MIMO is interesting to support wireless transmission at millimeter-wave. The close distance of the APs to the users and the ability to exploit macro-diversity will help us to deal with the channel impairments in millimeter-wave. One of the challenging issues in millimeter-wave communication is the CSI acquisition, particularly if we would like to implement a centralized approach.

- **Large network with multiple EPUs**: We predict that the future network infrastructure will be somehow in the form of a cell-free network with multiple EPUs and utilize a sort of dynamic cooperation as mentioned in Chapter 4. Therefore, we encourage ourselves in the future to investigate further this system architecture with more practical constraints and more realistic assumptions.

Nomenclature

Abbreviations

5G	Fifth Generation
6G	Sixth Generation
ADC	Analog Digital Converter
AGC	Automatic Gain Control
AP	Access Point
AWGN	Additive White Gaussian Noise
C-RAN	Cloud Radio Access Networks
CDF	Cumulative Distribution Function
CPU	Central Processing Unit
CSI	Channel State Information
DAS	Distributed Antenna System
DCC	Dynamic Cooperation Clustering
E-LVQ	Ellipsoid Lattice Vector Quantization
EPU	Edge Cloud Processing Unit
EQ	Estimate and Quantize
F-maMIMO	Fog massive MIMO
FDD	Frequency Division Duplex

I/Q	In-phase Quadrature
IoT	Internet of Things
KL	Karhunen Loeve
LBG-VQ	Linde Buzo Gray
LBG	Lind Buzo Gray Vector Quantization
LMMSE	Linear Minimum Mean Squared Error
LVQ	Lattice Vector Quantization
MIMO	Mulitple Input Multiple Output
MMSE	Minimum Mean Squared Error
MRC	Maximum Ratio Combining
MSE	Mean Squared Error
PCM	Pulse Code Modulation
QE	Quantize and Estimate
SDNR	Signal to Distortion Noise Ratio
SINR	Signal to Interference Noise Ratio
SNR	Signal to Noise Ratio
SQ	Scalar Quantization
TDD	Time Division Duplex
UE	User Equipment
VQ	Vector Quantization
ZF	Zero Forcing

References

[1] 5G-PPP. (2017). Key Challenges for the 5G Infrastructure PPP, [Online]. Available: https://5g-ppp.eu/.

[2] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010, ISSN: 1536-1276. DOI: 10.1109/TWC.2010.092810.091092.

[3] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013, ISSN: 1053-5888. DOI: 10.1109/MSP.2011.2178495.

[4] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and H. Yang, "Cell-Free Massive MIMO Systems," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Nov. 2015, pp. 695–699. DOI: 10.1109/ACSSC.2015.7421222.

[5] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO: Uniformly Great Service for Everyone," in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Jun. 2015, pp. 201–205. DOI: 10.1109/SPAWC.2015.7227028.

[6] ——, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017, ISSN: 1536-1276. DOI: 10.1109/TWC.2017.2655515.

[7] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of Cell-Free Massive MIMO Systems with MMSE and LSFD Receivers," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 203–207. DOI: 10.1109/ACSSC.2016.7869024.

[8] H. Yang and E. G. Larsson, "Can Massive MIMO Support Uplink Intensive Aplications?" In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Marrakech, Morocco, Apr. 2019.

[9] M. Bashar, H. Q. Ngo, A. G. Burr, D. Maryopi, K. Cumanan, and E. G. Larsson, "On the Performance of Backhaul Constrained Cell-Free Massive MIMO with Linear

Receivers," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct. 2018, pp. 624–628. DOI: 10.1109/ACSSC.2018.8645433.

[10] D. Maryopi, M. Bashar, and A. Burr, "On the Uplink Throughput of Zero Forcing in Cell-Free Massive MIMO With Coarse Quantization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7220–7224, Jul. 2019, ISSN: 0018-9545. DOI: 10.1109/TVT.2019.2920070.

[11] E. Björnson and L. Sanguinetti, "Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation," *Submitted to IEEE Transactions on Wireless Communications*, March 2019. [Online]. Available: https://arxiv.org/pdf/1903.10611.pdf.

[12] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-Wave Vehicular Communication to Support Massive Automotive Sensing," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 160–167, Dec. 2016, ISSN: 1558-1896. DOI: 10.1109/MCOM.2016.1600071CM.

[13] R. H. Walden, "Analog-to-Digital Converter Survey and Analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, Apr. 1999, ISSN: 0733-8716. DOI: 10.1109/49.761034.

[14] R. W. Heath, "Going Toward 6G [From The Editor]," *IEEE Signal Processing Magazine*, vol. 36, no. 3, May 2019, ISSN: 1053-5888.

[15] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the Performance of Cell-Free Massive MIMO Relying on Adaptive NOMA/OMA Mode-Switching," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 792–810, 2020.

[16] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local Partial Zero-Forcing Precoding for Cell-Free Massive MIMO," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[17] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive mimo uplink with optimal uniform quantization," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 4, pp. 971–987, 2019.

[18] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-Centric 5G Cellular Networks: Resource Allocation and Comparison With the Cell-Free Massive MIMO Approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1250–1264, 2020.

[19] C. D'Andrea, A. Garcia-Rodriguez, G. Geraci, L. G. Giordano, and S. Buzzi, "Analysis of UAV Communications in Cell-Free Massive MIMO Systems," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 133–147, 2020.

[20] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and H. V. Poor, "Mixed quality of service in cell-free massive mimo," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1494–1497, 2018.

[21] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "Enhanced max-min sinr for uplink cell-free massive mimo systems," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[22] ——, "On the uplink max–min sinr of cell-free massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2021–2036, 2019.

[23] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, M. Debbah, and J. Kittler, "Exploiting Deep Learning in Limited-Fronthaul Cell-Free Massive MIMO Uplink," *IEEE Journal on Selected Areas in Communications*, Feb. 2020, Accepted for Publication.

[24] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, and M. Debbah, "Deep Learning-Aided Finite-Capacity Fronthaul Cell-Free Massive MIMO with Zero Forcing," *IEEE International Conference on Communications (ICC)*, Jun. 2020, Accepted for Publication.

[25] M. Alageli, A. Ikhlef, F. Alsifiany, M. A. M. Abdullah, G. Chen, and J. Chambers, "Optimal Downlink Transmission for Cell-Free SWIPT Massive MIMO Systems With Active Eavesdropping," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1983–1998, 2020.

[26] A. Burr, M. Bashar, and D. Maryopi, "Cooperative Access Networks: Optimum Fronthaul Quantization in Distributed Massive MIMO and Cloud RAN," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, Jun. 2018, pp. 1–5. DOI: 10.1109/VTCSpring.2018.8417560.

[27] ——, "Ultra-Dense Radio Access Networks for Smart Cities: Cloud-RAN, Fog-RAN and "Cell-Free" massive MIMO," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Workshop CorNer*, Bologna, Italy, Sep. 2018.

[28] D. Maryopi and A. G. Burr, "Few-Bit CSI Acquisition for Centralized Cell-Free Massive MIMO with Spatial Correlation," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Marrakech, Morocco, Apr. 2019.

[29] M. Bashar, A. G. Burr, D. Maryopi, K. Haneda, and K. Cumanan, "Robust Geometry-Based User Scheduling for Large MIMO Systems Under Realistic Channel Conditions," in *European Wireless 2018; 24th European Wireless Conference*, May 2018.

[30] A. Burr and D. Maryopi, "On the Modelling of Coarse Vector Quantization in Distributed Massive MIMO," *IEEE Statistical Signal Processing Workshop*, 2021, Submitted.

[31] J. Choi, D. J. Love, and P. Bidigare, "Downlink Training Techniques for FDD Massive MIMO Systems: Open-Loop and Closed-Loop Training With Memory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 802–814, 2014.

[32] M. Barzegar Khalilsarai, S. Haghighatshoar, and G. Caire, "How to Achieve Massive MIMO Gains in FDD Systems?" In *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 1–5.

[33] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017, ISSN: 1932-8346. DOI: 10.1561/2000000093. [Online]. Available: http://dx.doi.org/10.1561/2000000093.

[34] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable Rates of FDD Massive MIMO Systems With Spatial Channel Correlation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 5, pp. 2868–2882, 2015.

[35] J. Flordelis, F. Rusek, F. Tufvesson, E. G. Larsson, and O. Edfors, "Massive MIMO Performance—TDD Versus FDD: What Do Measurements Say?" *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2247–2261, 2018.

[36] T. Van Chien and E. Björnson, "Massive MIMO communications," in *5G Mobile Communications*, W. Xiang, K. Zheng, and X. (Shen, Eds. Cham: Springer International Publishing, 2017, pp. 77–116, ISBN: 978-3-319-34208-5. DOI: 10.1007/978-3-319-34208-5_4. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-34208-5_4.

[37] V. Saxena, G. Fodor, and E. Karipidis, "Mitigating Pilot Contamination by Pilot Reuse and Power Control Schemes for Massive MIMO Systems," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–6. DOI: 10.1109/VTCSpring.2015.7145932.

[38] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013, ISSN: 0090-6778. DOI: 10.1109/TCOMM.2013.020413.110848.

[39] T. Wiegand and H. Schwarz, "Source coding: Part I of Fundamentals of Source and Video Coding," *Foundations and Trends in Signal Processing*, vol. 4, no. 1-2, pp. 1–222, 2011, ISSN: 1932-8346. DOI: 10.1561/2000000010. [Online]. Available: http://dx.doi.org/10.1561/2000000010.

[40] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991, ISBN: 0-7923-9181-0.

[41] Y. Lu and H. H. Zhou, "Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants," *ArXiv*, vol. abs/1612.02099, 2016.

[42] R. M. Gray, "Fundamental of Quantization," in *Short Course*, Melbourne, Australia, 2006. [Online]. Available: https://ee.stanford.edu/~gray/shortcourse.pdf.

[43] J. Bussgang, "Crosscorrelation Functions of Amplitude-Distorted Gaussian Signals," *RLE Technical Reports*, vol. 216, 1952.

[44] D. Dardari, "Exact Analysis of Joint Clipping and Quantization Effects in High Speed WLAN Receivers," in *IEEE International Conference on Communications, 2003. ICC '03.*, vol. 5, May 2003, 3487–3492 vol.5. DOI: 10.1109/ICC.2003.1204103.

[45] A. Behravan and T. Eriksson, "Analysis of Distortion in a Memoryless Bandpass Nonlinearity," *Proceedings of Nordic Radio Symposium*, 2004.

[46] P. Zillmann, "Relationship Between Two Distortion Measures for Memoryless Nonlinear Systems," *IEEE Signal Processing Letters*, vol. 17, no. 11, pp. 917–920, Nov. 2010, ISSN: 1070-9908. DOI: 10.1109/LSP.2010.2072498.

[47] A. Mezghani and J. A. Nosek, "Capacity Lower Bound of MIMO Channels with Output Quantization and Correlated Noise," *Proc. IEEE Int. Symp.Inf. Theory*, vol. pp. 1-5, 2012.

[48] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and Power Optimization in Cell-Free Massive MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017, ISSN: 1536-1276. DOI: 10.1109/TWC.2017.2698449.

[49] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah, "Cell-free massive mimo with limited backhaul," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–7. DOI: 10.1109/ICC.2018.8422865.

[50] T. L. Marzetta, "Massive MIMO: It Really Works!" In *NYU Wireless*, New York, USA, October, 2017. [Online]. Available: https://wireless.engineering.nyu.edu/presentations/update-3/Marzetta_web_26October_2017.pdf.

[51] Z. Chen and E. Björnson, "Channel Hardening and Favorable Propagation in Cell-Free Massive MIMO With Stochastic Geometry," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018, ISSN: 0090-6778. DOI: 10.1109/TCOMM.2018.2846272.

[52] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks—a technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[53] L. Liu, P. Patil, and W. Yu, "An uplink-downlink duality for cloud radio access network," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 1606–1610.

[54] J. Max, "Quantizing for Minimum Distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960, ISSN: 0096-1000. DOI: 10.1109/TIT.1960.1057548.

[55] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489.

[56] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Max-Min Rate of Cell-Free Massive MIMO Uplink with Optimal Uniform Quantization," *IEEE Transactions on Communications*, pp. 1–1, 2019. DOI: 10.1109/TCOMM.2019.2926706.

[57] B. Hassibi and B. M. Hochwald, "How Much Training is Needed in Multiple-Antenna Wireless Links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003, ISSN: 0018-9448. DOI: 10.1109/TIT.2003.809594.

[58] R. Senanayake, A. Lozano, P. Smith, and J. Evans, "Analytical Handle for ZF Reception in Distributed Massive MIMO," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 16–20. DOI: 10.1109/ACSSC.2016.7868985.

[59] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, 1998.

[60] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *ArXiv*, vol. abs/1908.03119, 2019.

[61] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 406–411, 1970.

[62] A. A. I. Ibrahim, A. Ashikhmin, T. L. Marzetta, and D. J. Love, "Cell-Free Massive MIMO Systems Utilizing Multi-Antenna Access Points," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Oct. 2017, pp. 1517–1521. DOI: 10.1109/ACSSC.2017.8335610.

[63] S. Chlaily, C. Ren, P. Amblard, O. Michel, P. Comon, and C. Jutten, "Information–estimation relationship in mismatched gaussian channels," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 688–692, 2017.

[64] H. Kim and J. Choi, "Channel Estimation for One-Bit Massive MIMO Systems Exploiting Spatio-Temporal Correlations," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–6. DOI: 10.1109/GLOCOM.2018.8647574.

[65] J. H. Conway, N. J. A. Sloane, and E. Bannai, *Sphere-packings, Lattices, and Groups*. Berlin, Heidelberg: Springer-Verlag, 1987, ISBN: 0-387-96617-X.

[66] R. Zamir, B. Nazer, Y. Kochman, and I. Bistritz, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation and Multiuser Information Theory*. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139045520.

[67] J. Conway and N. Sloane, "Voronoi Regions of Lattices, Second Moments of Polytopes, and Quantization," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 211–226, Mar. 1982, ISSN: 1557-9654. DOI: 10.1109/TIT.1982.1056483.

[68] U. Erez, S. Litsyn, and R. Zamir, "Lattices Which Are Good for (Almost) Everything," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3401–3416, Oct. 2005, ISSN: 1557-9654. DOI: 10.1109/TIT.2005.855591.

[69] M. V. Eyuboglu and G. D. Forney, "Lattice and Trellis Quantization with Lattice- and Trellis-Bounded Codebooks-High-Rate Theory for Memoryless Sources," *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 46–59, Jan. 1993, ISSN: 0018-9448. DOI: 10.1109/18.179341.

[70] T. Erikson and E. Agrell, "Lattice-Based Quantization part II," *Dept.Inf. Theory, Chalmers Univ. Technol, Technical Reports*, Oct. 1996.

[71] J. Conway and N. Sloane, "A Fast Encoding Method for Lattice Codes and Quantizers," *IEEE Transactions on Information Theory*, vol. 29, no. 6, pp. 820–824, Nov. 1983, ISSN: 0018-9448. DOI: 10.1109/TIT.1983.1056761.

[72] ——, "Fast Quantizing and Decoding and Algorithms for Lattice Quantizers and Codes," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 227–232, Mar. 1982, ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056484.

[73] S. Ragot and R. Lefebvre, "Near-Ellipsoidal Voronoi Coding," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1815–1820, Jul. 2003, ISSN: 0018-9448. DOI: 10.1109/TIT.2003.813484.

[74] R. Price, "A Useful Theorem for Nonlinear Devices Having Gaussian Inputs," *IRE Transactions on Information Theory*, vol. 4, no. 2, pp. 69–72, Jun. 1958, ISSN: 2168-2712. DOI: 10.1109/TIT.1958.1057444.