Lattice Network Coding in Distributed Massive MIMO Systems

Qinhui Huang

Doctor of Philosophy

University of York

Electronic Engineering

September 2017

Abstract

In this thesis, the uplink of distributed massive MIMO where a large number of distributed access point antennas simultaneously serve a relatively smaller number of users is considered. Lattice network coding (LNC), which comprises compute and forward (C&F) and integer forcing (IF), is employed to avoid the potentially enormous backhaul load. Firstly, novel algorithms for coefficient selection in C&F are proposed. For the first time, we propose a low polynomial complexity algorithm to find the optimal solution for the complex valued case. Then we propose a sub-optimal simple linear search algorithm which is conceptually sub-optimal, however numerical results show that the performance degradation is negligible compared to the exhaustive method. The complexity of both algorithms are investigated both theoretically and numerically. The results show that our proposed algorithms achieve better performance-complexity trade-offs compared to the existing algorithms. Both algorithms are suitable for lattices over a wide range of algebraic integer domains. Secondly, the performance of LNC in a realistic distributed massive MIMO model (including fading, pathloss and correlated shadowing) is investigated in this thesis. By utilising the characteristic of pathloss, a low complexity coefficient selection algorithm for LNC is proposed. A greedy algorithm for selecting the global coefficient matrix is proposed. Comprehensive comparisons between LNC and some other promising linear strategies for massive MIMO, such as small cells (SC), maximum ratio combining (MRC), and minimum mean square error (MMSE) are also provided. Numerical results reveal that LNC not only reduces the backhaul load, but also provides uniformly good service to all users in a wide range of applications. Thirdly, the inevitable loss of information due to the quantisation and modulo operation under different backhaul constraints are investigated. An extended C&F with flexible cardinalities is proposed to adapt to the different backhaul constraints. Numerical results show that by slightly increasing the cardinality, the gap between C&F to the infinite backhaul case can be significantly reduced.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to show my sincere gratitude to my supervisor, Prof. Alister. G. Burr, for his support, supervision and guidance not only on my research but also on career and future development.

I am very grateful to my thesis advisor, Dr. Kanapathippillai. Cumanan, whose insightful discussions and suggestions have benefited me.

I would also like to thank Dr Dong Fang, Dr. Yi Wang, Dr. Tong Peng, Dr. Mehid Molu and Ms Manijeh Bashar and other colleagues in the Communications and Signal Processing Research Group, for useful discussions.

This thesis is dedicated to my parents for their unconditional support, endless love and encouragement.

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, or any other, University. All sources are acknowledged as References. Some of the research presented in this thesis has resulted in some publications. These publications are listed at the end of Chapter 1.

Chapter 1

Introduction

Contents

1.1 Overview

Fading and interference are the two fundamental challenges to the design of wireless communication systems. The former is caused by the propagation characteristics of electromagnetic waves, such as path loss, shadowing by obstacles, and the effect of multipath. The latter is due to the fact that each transmitter-receiver pair can not be regarded as an isolated channel link in wireless transmission. Fading is generally inevitable in wireless environments, while the interference is traditionally avoided by performing some form of orthogonal transmission which enables different users to be distinguished by time or frequency. However, future wireless systems are more interested in increasing spectral efficiency, and this encourages the use of the same time-frequency resources by different users. Thus, the mitigation of fading and interference will still remain the key challenge

in 5G wireless and beyond.

Multiple-input and multiple-output (MIMO) [1] techniques, have been widely applied to combat fading and interference over the past two decades. These employ multiple antennas at both source and destination devices. On the one hand, it mitigates the effect of deep fading by providing multiple effective channel links, which is referred to as the diversity gain. On the other hand, the addtional antenna elements enable the receiver to exploit the interference rather than treating it as pure nuisance. Thus, multiple data streams can be simultaneously transmitted, and this is referred to as the multiplexing gain. Theoretically, the spectral efficiency increases monotonically with the number of antennas, and hence recent focus has shifted towards massive MIMO [2] which employs a large number of antennas to serve a relatively smaller number of users.

An important application of MIMO is in distributed antenna systems (DAS), where multiple distributed access points (APs) are connected and cooperate via backhaul (or 'fronthaul') links to form a giant 'virtual access point'. Typical examples are 'network' MIMO [3], and 'cloud' radio access networks [4]. Compared to the collocated MIMO, the distributed topology reduces the correlation between the access points, which effectively increases the diversity order. The distributed topology also has the advantage in serving the cell edge users, since the access poins are distributed more evenly across the coverage area.

Distributed massive MIMO which combines the concepts of both 'network' MIMO and 'massive' MIMO has attracted a lot interest very recently. It benefits from the lower path loss while maintaining the diversity gain achieved by employing a large scale 'virtual antenna array'. However, these benefits are typically achieved at the expense of enormous backhaul load. For example, assume signal bandwidth 100MHz; 2x oversampling; 8 bits quantisation; 4 antennas; 2 polarisations; the backhaul required per AP is $100 \times 2 \times 8 \times 4 \times 2 = 12.8$ Gbps.

Lattice network coding (LNC) is considered as a promising technique to avoid the enormous backhaul load for distributed massive MIMO systems. It allows each AP to infer and decode a linear combination of the transmitted codewords. The nested structure in lattice coding ensures the linear combination of the codewords is a codeword itself,

hence the cardinality expansion is avoided. In this thesis, we will study the applications of LNC in distributed massive MIMO systems. The objective of this thesis is to find the answers to the following questions:

- *How to select the optimal mapping function at each AP?*
  The performance of LNC greatly depends on the selected mapping function at each AP. Finding the optimal coefficients in LNC has commonly been treated as a shortest vector problem (SVP), which is N-P hard. Fortunately, this problem can be much simpler in a fully distributed system (single antenna at APs), and the corresponding LNC scheme for this case is called compute-and-forward (C&F). Our objective is to propose low polynomial complexity algorithms for C&F to find the optimal mapping function which maximises the achievable rate at each AP.

- *How will LNC perform in realistic distributed massive MIMO systems?*
  The research on LNC so far focusses on the baseline Gaussian or Rayleigh fading model. Our objective is to evaluate the performance of LNC in a realistic distributed massive MIMO model [5] which considers the effects of both large scale and small scale fading. We consider two applications of LNC: C&F in fully distributed deployment and integer-forcing (IF) in partially distributed deployment. For both cases, we aim to propose low complexity algorithms to select both the locally and globally optimal coefficients, and provide comprehensive comparisons with the benchmarks.

- *How much information will inevitably be lost in LNC?*
  LNC employs some quantisation and modulo operations to achieve the minimum backhaul load, and this will definitely lead to some performance degradation. Thus, we aim to find the inevitable gap between LNC to the ideal case (infinite backhaul), and investigate the tradeoff between backhaul capacity and performance.

1.2 Contributions

Based on the discussions above, the main contributions of this thesis are listed as follows:

- For the real valued case, two novel low complexity approaches are proposed to select the locally optimal equation for C&F: reduced candidate set aided algorithm and linear search algorithm. The former reduces the number of candidate vectors in the exhaustive search; the latter employs sampled values as candidates to acquire some sub-optimal solutions. In terms of the computation rate, both proposed algorithms outperform the LLL, and have negligible performance degradations compared to the exhaustive search. In terms of the complexity, both algorithms have much lower complexity compared to the exhaustive search, and lower than the LLL except at very high SNR.

- For the complex valued case, there are also two local selection algorithms proposed for C&F: complex-exhaustive-II algorithm and the linear search algorithm. The former extends the exhaustive-II algorithm in [6] to complex integer based lattices, and it ensures the optimality of the acquired equation; the latter applies a pre-defined step size to both real and imaginary parts for sampling, in order to discard the 'unnecessary' candidates while avoiding performance degradation. The optimised step sizes are related to the number of users and SNR. Theoretical analysis indicates that both our proposed algorithms have low polynomial complexities. Numerical results illustrate that both proposed algorithms have better complexity-performance tradeoff, compared to the existing algorithms. In addition, it is possible to extend our proposed algorithms to some other algebraic integer based lattices, without enhancing the complexity.

- As previously mentioned, we consider two applications of LNC in realistic distributed massive MIMO systems: C&F and IF. For both cases, we propose novel algorithms to reduce the complexity for selecting the locally optimal coefficients by exploiting the properties of pathloss, and their complexity reductions are investigated numerically. We also propose a greedy algorithm to select the globally optimal integer coefficients, and this approach is applicable for both C&F and IF. The system behaviour of LNC is evaluated in the following aspects: 1) the probabilities of rank deficiency corresponding to different numbers of the locally forwarded candidates; 2) the achievable rates corresponding to different numbers of antennas on each AP; 3) the achievable rates under different levels of interference. We consider small cells (SC), maximal ratio combining (MRC), and the centralised MMSE as benchmarks to provide comprehensive comparisons. Numerical results show that

LNC always attains higher capacity than small cells, and outperforms MRC and the centralised MMSE for a wide range of applications. Their respective complexities and required backhaul are also discussed qualitatively.

- We propose an 'extended' network coding scheme for distributed massive MIMO to allow a flexible backhaul load. Similar to standard C&F, it also comprises lattice quantisation and the modulo operation, but with a flexible modulus to the modulo operation. We evaluate the outer bound of this scheme under different backhaul loads by employing jointly optimised scaling factors at APs. The loss of information due to the quantisation and the modulo operation is calculated by two approaches: Monte Carlo integration and an approximate closed form expression. The former aims to acquire an accurate outer bound, while the latter is proposed to select the scaling factors from a practical perspective. We show that the gap between the standard C&F scheme and the infinite-backhaul case can be significantly reduced by slightly increasing the cardinality of the modulo operation at APs.

1.3 Thesis Outline

The structure of the thesis is listed as follows:

- In Chapter 2, a literature review of the techniques that are relevant to this thesis are presented. The review begins with the concepts and properties of 'network MIMO' and massive MIMO, followed by an introduction to the principles of physical layer network coding, particularly lattice based schemes: C&F and integer forcing. Some basic knowledge of nested lattices is also introduced.

- In Chapter 3, three novel coefficient selection algorithms for C&F are presented: 1) reduced candidate set aided algorithm for the real valued case; 2) complex-exhaustive-II algorithm for the complex valued case; 3) a linear search algorithm which can be applied to both cases. For all proposed algorithms, the corresponding complexities are investigated both theoretically and numerically. The achievable rate comparisons with the benchmarks are also given.

- In Chapter 4, both C&F and IF strategies are investigated under a realistic distributed massive MIMO model. Low complexity algorithms for selecting both locally and globally optimal coefficients are presented. Comprehensive comparisons with the benchmarks are also given.

- In Chapter 5, we evaluate the inevitable information loss due to the quantisation and modulo operation. We present two method to calculate the mutual information: 1) Monte Carlo integration; 2) closed form expression. Then we analyse the tradeoff between the performance and backhaul load.

- In Chapter 6, the conclusions and possible future work are presented based on the content of the thesis.

1.4 Notation

Unless noted, we use plain letter, boldface lowercase letters and boldface uppercase letters to denote scalars, vectors, and matrices respectively. All vectors are column vectors. Some special symbols and operations are listed as follows:

$\mathbb{R}$	Real numbers
$\mathbb{C}$	Complex numbers
$\mathbb{Z}$	Real integers
$\mathbb{Z}[i]$	Gaussian integers
$\mathbb{Z}[\omega]$	Eisenstein integers
P	Average power constraint
$\mathbf{x}$	Vector channel input
$\mathbf{y}$	Vector channel output
$\mathbf{z}$	Vector Gaussian noise
N_{r}	Number of antennas per access point
L	Number of users
$\mathcal{N}(\mu, \sigma^2)$	Real Gaussian random variable with mean μ and variance σ^2
$\mathcal{CN}(0, \sigma^2)$	Circularly symmetric complex Gaussian random variable: both real and imaginary parts are i.i.d. $\mathcal{N}(0, \sigma^2/2)$
$\mathcal{Q}(\cdot)$	Quantisation

$\Pr\{A\}$	Probability of an event A
$\mathrm{diag}\{a_1, \cdots, a_n\}$	Diagonal matrix with the diagonal entries equal to $a_1, \cdots, a_n$
mod	Modulo operation
$(\cdot)^T$	Transpose
$(\cdot)^H$	Conjugate transpose
$(\cdot)^\dagger$	Pseudo inverse
Λ	Lattice
$\mathcal{V}(\lambda)$	Voronoi region of a lattice point λ
$\cup$	Union of sets
$\cap$	Intersection of sets
$\cong$	Congruence/Isomorphic
$\bigoplus$	Summation over finite field
erfc	Complementary error function
$\mathbf{I}_N$	$N \times N$ identity matrix
$\|\cdot\|$	Euclidean norm/Euclidean distance
$\|\cdot\|$	Cardinality

1.5 Publication List

Journal Papers

1. Q. Huang and A. Burr, "Low Complexity Coefficient Selection Algorithms for Compute-and-Forward", *IEEE Access.*, vol. 5, pp. 19182-19193, Aug. 2017.

2. Q. Huang, A. Burr, K. Cumanan and F. Dong, "Lattice Network Coding in Cell-Free Massive MIMO: Good Performance with Low Backhaul Load", *ready for submitting to IEEE Transactions on Vehicular Technology*.

3. Y. Wang, A. Burr, Q. Huang, M. Molu "A Multilevel Framework for Lattice Network Coding", *IEEE Transactions on Information Theory*., (under review).

Conference Papers

1. Q. Huang and A. Burr, “Compute-and-forward in cell-free massive MIMO: Great performance with low backhaul load”, *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, Paris, May. 2017.

2. Q. Huang and A. Burr, “The capacity of Cloud-RAN: Outer bound with quantisation and constrained fronthaul load”, *2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Shenzhen, Nov. 2016.

3. Q. Huang and A. Burr, “Low Complexity Coefficient Selection Algorithms for Compute-and-Forward”, *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, May. 2016.

4. A. Burr and Q. Huang, “On quantization and network coding in wireless relay networks”, *2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM),*, Rio de Janerio, Jul. 2016.

5. Y. Wang, A. Burr, Q. Huang “Dense Wireless Cloud Network via Physical Layer Network Coding”, *2015 IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)* , Guildford, Sep. 2015.

Chapter 2

Literature Review

Contents

2.1 Introduction

In this chapter, we review the fundamental techniques utilised throughout the thesis. We first introduce the concept of distributed massive MIMO which is referred to as a combination of 'network' MIMO and massive MIMO, an introduction to the properties of both aspects are presented. Then we review the principles and recent developments of physical layer network coding, particularly the compute and forward scheme. Some essential preliminaries in abstract algebra are also presented.

2.2 Network MIMO

2.2.1 Architectures of Network MIMO

Network MIMO: Conventional View

A conventional cellular network is formed by distinct cells, and each user is served by a unique base station, that means the cell edge users are poorly served due to the existence of inter cell (in some form of co-channel) interference. The concept of 'network' MIMO [3], or coordinated multipoint (CoMP) [7] was originally proposed for dealing with such problem. It allows a cluster of BSs to form a 'virtual' BS via backhaul links to serve a group of users simultaneously, as illustrated by the blue text in Figure 2.1. Joint processing can be performed by the CPU and inter cell interference becomes intra cell interference.

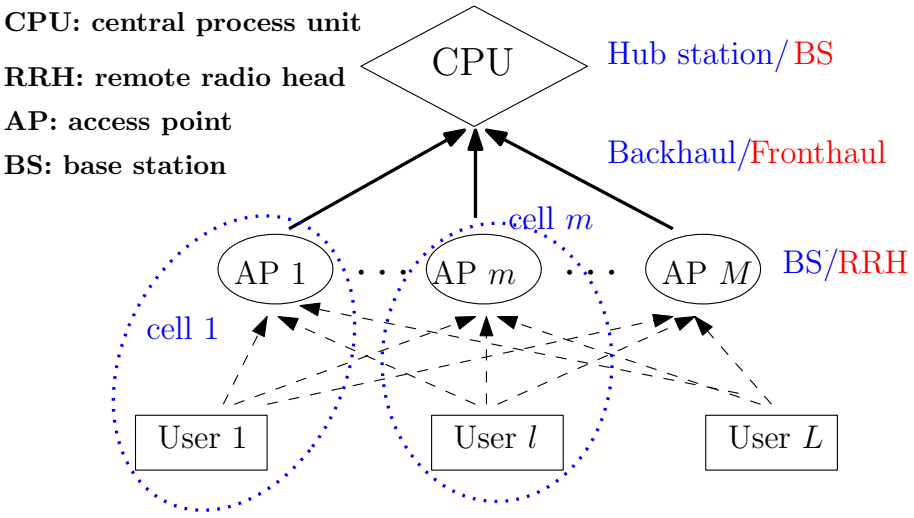

Figure 2.1: A general model of the uplink Network MIMO (or 'Cloud' RAN)

Cloud-RAN

Compared to the traditional view, the relatively newer concept of 'Cloud' radio access networks (C-RAN) [4] can also be treated as a 'network' MIMO architecture. It deploys distributed antenna elements, namely, remote radio heads (RRHs) within a conventional

cell, and RRHs are connected to the BS via fronthaul links. In this architecture, RRHs and BS act as APs and CPU respectively, as illustrated by the red text in Figure 2.1. The main idea of both 'network' MIMO and C-RAN is to exploit additional spatial degrees of freedom from other APs, which means each user is effectively served by multiple APs. As the 'degree of cooperation' increases (larger coverage of the 'virtual BS', denser APs and higher bandwidth of backhaul links), the concept of the 'cell' is not as important as it used to be. The cellular network tends to be 'cell-free', as will be described later.

2.2.2 Relaying Strategies in Network MIMO

The performance of cooperative MIMO systems greatly depends on the relaying strategy performed at the intermediate APs. The ideal relaying strategy is obviously that each AP forwards the received signal precisely to the CPU via perfect backhaul (infinite bandwidth, noise-free), hence the distributed system becomes effectively collocated, and the optimal performance can be achieved by performing joint maximum likelihood (ML) decoding at the CPU. Regardless of the infeasible assumption of perfect backhaul, there are three relaying strategies commonly used in practice.

- Decode-and-Forward (DF): each AP decodes one or more user messages independently, and the recovered messages are then re-encoded for the next transmission. The DF scheme provides near-optimal performance for the Wyner model, in which each AP only see three input signals and decodes the intended user from the two neighbouring users [8]. For a more general scenario, DF is nowadays performed by small cell systems. The cooperative gain can be achieved by employing AP selection for a specific user. The main drawback of this scheme is the poor interference mitigation due to the independent decoding performed at the APs [9, 10].

- Compress-and-Forward: the received signal at each AP is vector quantised and then forwarded to the hub station. Clearly, the performance depends on the quantisation level, finer quantisation leads to lower information loss, however, more backhaul load is required. Since no error correction coding is applied at intermediate APs, noise is therefore accumulated during retransmission [8, 9, 11]. Another approach to compress data is to employ Wyner-Ziv coding [9, 12] at the relays (the quantised

signals at the relays are correlated since they arise from the same data), when side information is available.

- Amplify-and-Forward (AF): each AP simply acts as a repeater and forwards an amplified version of received signal to compensate the propagation loss. The simplicity of AF gives significant adavantage to systems with symple relay units. However, similar to compress-and-forward, noise is also accumulated during retransmission. The error performance of such a scheme is analysed in [13]

2.3 Massive MIMO

Massive MIMO was originally proposed by Marzetta in his landmark paper [2] which considers a system model with an unlimited number of base station antennas. Following the pioneering work in [2], many developments have been carried out during the past few years. Intuitively, massive MIMO implies that the MIMO system goes to a larger scale. It comprises two aspects:

- Denser antenna elements: both of the number of users and APs (more precisely the number of total antennas on APs) are significantly increased within a certain area.
- Larger ratio of AP/user antennas: the main point of massive MIMO is to employ a large number of receiver antennas to serve a relatively smaller number of users.

In this section we briefly review the basic characteristics of Massive MIMO. More details can be found in [2, 14–16].

2.3.1 Asymptotic Properties of Massive MIMO

We consider a system with L single antenna users and M AP antennas (in this section, M denotes the total number of AP antennas), where $L \ll M$. When the scale of the antenna arrays increases, certain asymptotic properties of random matrices become deterministic, as we will discuss below. For simplicity, we employ small scale fading channels to present

the general properties of scaling up the dimension of channel matrix. The effect of large scale fading (including pathloss and shadowing) depends on the specific deployment of AP antennas, which will be introduced in section 2.3.2 and investigated in Chapter 4.

Asymptotic Capacity of Massive MIMO

We assume the channel coefficient between the lth user and the mth AP antenna is an independent identically distributed (i.i.d) complex Gaussian variable with zero mean and unit variance, denoted as $h_{m,l} \sim \mathcal{CN}(0,1)$. Let $\mathbf{h}_l = [h_{1,l}, h_{2,l}, \cdots, h_{M,l}]^T$ denote the channel vector corresponding to user l, and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L]$ be the channel matrix. The transmitted and received signal vectors are represented by $\mathbf{x} = [x_1, x_2, \cdots, x_L]^T$ and $\mathbf{y} = [y_1, y_2, \cdots, y_M]^T$ respectively. Assuming that the channel state information (CSI) is perfectly known at the receiver, we have

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \quad \mathbf{y}, \mathbf{z} \in \mathbb{C}^{M\times 1}, \tag{2.1}$$

where $\mathbf{z}$ denotes the additive white Gaussian noise (AWGN) vector with each element $z_m \in \mathcal{CN}(0, \sigma^2)$. Let $P = \mathbb{E}[|x_l|^2]$ denote the power constraint of the transmitted symbol, and $\mathrm{SNR} = P/\sigma^2$ denote the signal to noise ratio. The receiver applies an equalisation matrix $\mathbf{B}^{L\times M}$ to recover $\mathbf{x}$ from $\mathbf{y}$, expressed as

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{B}\mathbf{y} \\ &= \mathbf{B}\mathbf{H}\mathbf{x} + \mathbf{B}\mathbf{z} \\ &= \mathbf{x} + (\mathbf{B}\mathbf{H} - \mathbf{I}_L)\mathbf{x} + \mathbf{B}\mathbf{z},\end{aligned} \tag{2.2}$$

where $\mathbf{I}_L$ denotes an $L \times L$ identity matrix. The achievable rate of the lth user for an instantaneous channel $\mathbf{H}$ can be expressed as

$$\mathcal{R}_l(\mathbf{H}, \mathbf{B}) = \log_2\Big(1 + \frac{\mathrm{SNR}|\mathbf{b}_l^T\mathbf{h}_l|^2}{\mathrm{SNR}\sum_{l'\neq l}|\mathbf{b}_l^T\mathbf{h}_{l'}|^2 + \|\mathbf{b}_l^T\|^2}\Big) \tag{2.3}$$

where $\mathbf{b}_l^T$ denotes the lth row of $\mathbf{B}$. The most commonly used equalisers in massive MIMO are zero-forcing (ZF) and maximum ratio combining (MRC), denoted by $\mathbf{B}_{\mathrm{ZF}} = \mathbf{H}^{\dagger}$ and $\mathbf{B}_{\mathrm{MRC}} = \mathbf{H}^H$ respectively. By substituting $\mathbf{B}_{\mathrm{ZF}}$ and $\mathbf{B}_{\mathrm{MRC}}$ for $\mathbf{B}$, the expression

(2.3) can be rewritten as

$$\mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{ZF}}) = \log_2\Big(1 + \frac{\mathrm{SNR}}{\|\mathbf{h}_l^\dagger\|^2}\Big) \tag{2.4}$$

and

$$\mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{MRC}}) = \log_2\Big(1 + \frac{\mathrm{SNR}\|\mathbf{h}_l\|^4}{\mathrm{SNR}\sum_{l'\neq l}|\mathbf{h}_l^H\mathbf{h}_{l'}|^2 + \|\mathbf{h}_l^H\|^2}\Big) \tag{2.5}$$

for ZF and MRC respectively, where $\mathbf{h}_l^\dagger$ denotes the lth row of $\mathbf{H}^\dagger$. Since $h_{m,l}$ is i.i.d. and $h_{m,l} \sim \mathcal{CN}(0,1)$, some relevant expected values are given by [15]

$$\begin{aligned} &\mathbb{E}[\|\mathbf{h}_l\|^2] = M,\ \mathbb{E}[\|\mathbf{h}_l\|^4] = M^2 + M, \\ &\mathbb{E}[\mathbf{h}_l^H\mathbf{h}_{l'}] = 0,\ \mathbb{E}[|\mathbf{h}_l^H\mathbf{h}_{l'}|^2] = M, \end{aligned} \tag{2.6}$$

$$\mathbb{E}\Big[\frac{1}{\|\mathbf{h}_l^\dagger\|^2}\Big] = M - L + 1. \tag{2.7}$$

It can be observed from (2.6) that the effective channel $\mathbf{H}^H\mathbf{H}$ tends to a scaled identity matrix $M\mathbf{I}_L$ as $M \to \infty$, which means even $\mathbf{B}_{\mathrm{MRC}}$ is able to translate the network to an interference free state as long as the number of AP antennas is extremely large. By combining (2.4)$\sim$(2.7), the corresponding asymptotic capacities of $\mathbf{B}_{\mathrm{ZF}}$ and $\mathbf{B}_{\mathrm{MRC}}$ can be respectively written as

$$\mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{ZF}})_{M\to\infty} \approx \log_2\big(1 + \mathrm{SNR}(M - L + 1)\big) \tag{2.8}$$

and

$$\mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{MRC}})_{M\to\infty} \approx \log_2\Big(1 + \frac{\mathrm{SNR}M}{\mathrm{SNR}(L-1)+1}\Big) \tag{2.9}$$

as long as SNR is finite. Clearly, the capacity for an instantaneous random $\mathbf{H}$ becomes deterministic as $M \to \infty$, hence the expressions (2.8) and (2.9) also represent the ergodic capacities as

$$\begin{aligned} \mathcal{R}_l(\mathbf{B}_{\mathrm{ZF}}) &= \mathbb{E}\big[\mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{ZF}})\big] = \mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{ZF}})_{M\to\infty} \\ \mathcal{R}_l(\mathbf{B}_{\mathrm{MRC}}) &= \mathbb{E}\big[\mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{MRC}})\big] = \mathcal{R}_l(\mathbf{H}, \mathbf{B}_{\mathrm{MRC}})_{M\to\infty}. \end{aligned} \tag{2.10}$$

Average Noise Enhancement

On the one hand, massive MIMO nulls the interference components $\sum_{l' \neq l} |\mathbf{h}_l^H \mathbf{h}_{l'}|^2$ in MRC. On the other hand, it also averages the noise enhancement among users in ZF. For example, equation (2.7) indicates that the noise scaling factor $\mathbf{B}_{\mathrm{ZF}}$ of a random user is asymptotically deterministic. From a more general perspective, the sum rate of an open-loop system can be expressed as [17]

$$\begin{aligned} \mathcal{R}_{\mathrm{sum}} &= \log_2 \det \left(\mathbf{I}_L + \mathrm{SNR} \mathbf{H}^H \mathbf{H} \right) \\ &= \log_2 \det \left(\mathbf{I}_L + \mathrm{SNR} \mathbf{V} \mathbf{D} \mathbf{V}^H \right) \qquad (2.11) \\ &= \sum_{l=1}^{L} \log_2 (1 + \mathrm{SNR} \lambda_l), \qquad (2.12) \end{aligned}$$

where (2.11) comes from the eigenvalue decomposition of $\mathbf{H}^H \mathbf{H}$, and λ_l denotes the lth corresponding eigenvalue. The condition number $\mathrm{Cond}(\mathbf{H}^H \mathbf{H})$ which denotes the ratio of the maximum and minimum eigenvalues ($\frac{\lambda_{\max}}{\lambda_{\min}}$), approaches 1 as $M \to \infty$, and hence the probability that the channel is ill-conditioned is negligible. This property brings significant benefit for the applications that all users transmit with the same rate, and the throughput of the system is limited by the poorest user.

ZF tends to be Optimal in Massive MIMO

Let $\gamma_{\mathrm{A,u}} = \frac{M}{L}$ denote the ratio of AP/user antennas. Another characteristic of massive MIMO is that the ZF scheme tends to be optimal as $\gamma_{\mathrm{A,u}} \to \infty$. The performance of a MU-MIMO system can be roughly measured by the diversity-multiplexing tradeoff (DMT). The multiplexing gain r and diversity gain d are said to be achieved if [18]

$$\begin{aligned} \lim_{\mathrm{SNR} \to \infty} \frac{\mathcal{R}(\mathrm{SNR})}{\log \mathrm{SNR}} &= r \\ \lim_{\mathrm{SNR} \to \infty} \frac{\log p_{\mathrm{e}}(\mathrm{SNR})}{\log \mathrm{SNR}} &= -d, \end{aligned} \qquad (2.13)$$

where p_{e} denotes the error probability. The DMT achieved by the joint ML and the ZF schemes can be expressed as

$$d_{\mathrm{opt}}(r) = M\Big(1 - \frac{r}{L}\Big), \tag{2.14}$$

and

$$d_{\mathrm{ZF}}(r) = (M - L + 1)\Big(1 - \frac{r}{L}\Big) \tag{2.15}$$

respectively. Assuming that a fixed rate transmission is employed (r=0), the diversity orders achieved by ZF and optimal scheme are $M - L + 1$ and M respectively. Hence the gap of ZF to the joint ML scheme scales as $\frac{\gamma_{\mathrm{A,u}}}{\gamma_{\mathrm{A,u}}-1}$, and tends to 0dB as $\gamma_{\mathrm{A,u}} \to \infty$.

2.3.2 Deployment of Massive MIMO

In practice, massive MIMO can be implemented in either a collocated or distributed manner. In this section, we present an intuitive comparison between these two schemes. For the sake of fairness, we assume a common total number of AP antennas $N_{\mathrm{r,total}}$ for both centralised and distributed schemes.

Centralised Massive MIMO

Centralised deployment utilises $N_{\mathrm{r,total}}$ antennas on one AP, hence $M = 1$ and $N_{\mathrm{r,total}} = N_{\mathrm{r}}$. Apparently, the main advantage of this deployment is that the signals observed by different antennas share a common physical location, and hence joint processing can be performed without extra backhaul. The drawbacks are as follows:

- For a highly compacted antenna array, the propagation paths for a particular user are subject to the same large scaling fading, hence the users far away from the AP are poorly served.
- Correlation between AP antennas.
- The antenna array is expected to be implemented in a 2-dimensional or 3-dimensional structure in massive MIMO, which increases the effect of mutual coupling between antennas [14].

Distributed Massive MIMO

Distributed massive MIMO separates $N_{\mathrm{r,total}}$ antennas over M distributed APs. It brings APs closer to cell edge users, and averages the large scale fading vector among users, hence uniformly good service for all users is expected to be achieved. Compared to the centralised case, the distributed structure mitigates the drawbacks above. However, it also brings two negative effects: 1) additional payloads for backhaul transmissions; 2) latency caused by processes performed at the intermediate APs. The comparison between centralised and distributed schemes is in some sense a performance-backhaul tradeoff.

The authors in [5, 19] proposed a fully distributed cell-free massive MIMO model with $N_{\mathrm{r}} = 1$ and $M = N_{\mathrm{r,total}}$. Each AP employs a locally estimated channel vector to perform MRC, and forwards the processed signal to the CPU via an unlimited capacity backhaul network. It has been shown in [5, 19] that the cell free scheme outperforms the uncoordinated small cell scheme for cell edge users. The energy efficiency for the cell-free scheme is investigated in [20]. Massive MIMO can also be implemented in a partially distributed deployment in which $N_{\mathrm{r}} \geq 1$ and $N_{\mathrm{r,total}} = MN_{\mathrm{r}}$.

2.4 Physical Layer Network Coding

Previously we have reviewed that extra backhaul is required to attain the performance improvement provided by employing distributed deployment. Reducing the backhaul load is usually a big challenge in distributed massive MIMO systems. Physical layer network coding is a backhaul reduction technique which employs the theoretically minimum backhaul load to achieve unambiguous transmission.

2.4.1 Principle of PNC

In this section, we employ the simplest 2 way relay channel (2-WRC) model to interpret the principle of network coding (NC) and physical layer network coding (PNC). In a wired network with multiple sources, the conventional routing strategy to avoid data collisions

is that each node only transmits the data packet of an individual user via a unique time slot. As shown in Figure 2.2(a), the data exchange between 2 users are performed in 4 orthogonal time slots.

Network coding is original proposed in [21] to increase the network throughput by allowing relays to transmit a function of their observed packets rather than repeating them individually. As presented in Figure 2.2(b), the relay broadcasts the exclusive or (XOR) of x_A and x_B instead of occupying two time slots to forward the individual data, and x_B can be acquired by user A by performing $x_A \oplus x_A \oplus x_B$, where x_A acts as the side information. The throughput per time slot is thus improved.

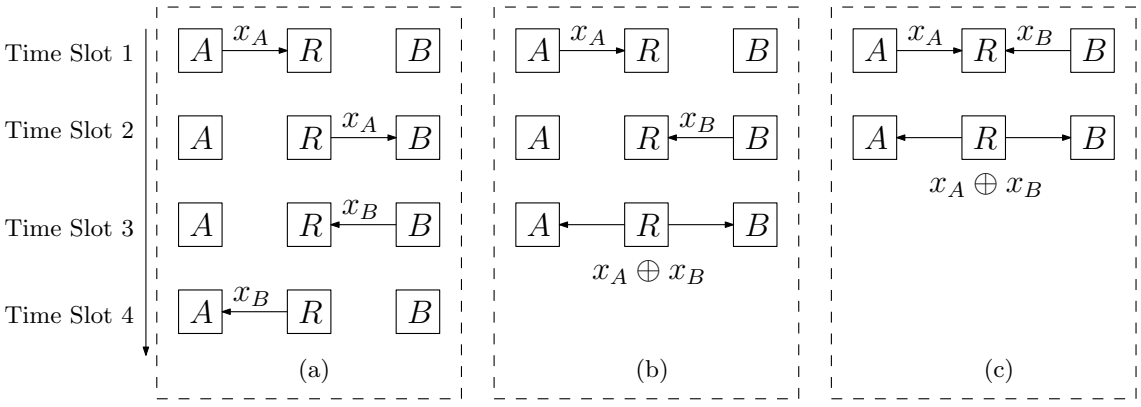

Figure 2.2: Transmission schemes in 2-WRC: (a) TDMA (b) NC (c) PNC

Unlike network coding, the PNC [22] strategy directly on the electromagnetic signal, and it comprises the multiple-access channel (MAC) phase and the broadcast channel (BC) phase. All users transmit their signals to the relay simultaneously in the MAC phase, and the relay naturally maps the superimposed signal to a network coded symbol for broadcasting, as shown in Figure 2.2(c). The XOR operation can be treated as the simplest linear mapping strategy, and the main task of PNC is to design proper mapping strategies to adapt to the time variant channel.

2.4.2 Properties of PNC

In this section, we employ a general multi-way relay channel (M-WRC) to illustrate the basic properties of PNC. We assume a total of L users, and that each user intends to acquire the data from the other $L-1$ users via one relay. The signal vector and channel

vector are denoted by $\mathbf{x} = [x_1, x_2, \cdots, x_L]^T$ and $\mathbf{h} = [h_1, h_2, \cdots, h_L]^T$ respectively.

Cardinality Reduction

Let $\mathcal{X}_l$ denote the set of modulated symbols of the lth user. The cardinality of a set is defined as the number of elements of the set, denoted as $|\cdot|$. For example, BPSK modulation indicates that $\mathcal{X}_l = \{+1, -1\}$ and $|\mathcal{X}_l| = 2$. Let

$$x_{ss} = \mathbf{x}^T \mathbf{h} \tag{2.16}$$

denote the noise-free superimposed signal, and $\mathcal{X}_{\rm ss}$ denote the set of values of $x_{\rm ss}$. PNC employs a mapping defined as

$$\mathcal{M} : x_{\rm ss} \to x_{\rm PNC}, \tag{2.17}$$

where $x_{\rm PNC}$ is the network coded symbol. Let $\mathcal{X}_{\rm PNC}$ denote the set of $x_{\rm PNC}$. The PNC mapping (2.17) equivalently dividing the set $\mathcal{X}_{\rm ss}$ into several subsets, where each subset corresponds to a unique network coded symbol. Clearly, the number of subsets is the cardinality of $\mathcal{X}_{\rm PNC}$, with $|\mathcal{X}_{\rm PNC}| < |\mathcal{X}_{\rm ss}|$, and the cardinality is thus reduced. It can be obtained from (2.16) that the maximum of $|\mathcal{X}_{\rm ss}|$ for an arbitrary $\mathbf{h}$ is the possible combinations of $\{x_1, x_2, \cdots, x_L\}$, which is the product of $|\mathcal{X}_l|$, denoted as $\max|\mathcal{X}_{\rm ss}| = \prod_{l=1}^{L} |\mathcal{X}_l|$. Since user l has x_l as side information, assuming a common modulation is employed for all users, the theoretically minimum cardinality required by performing PNC can be expressed as

$$\min|\mathcal{X}_{\rm PNC}| = \frac{\max|\mathcal{X}_{\rm ss}|}{|\mathcal{X}_l|}. \tag{2.18}$$

Note that more side information can be utilised to design PNC mapping when multiple relays are employed, which enables a larger value to substitute $|\mathcal{X}_l|$ in (2.18), and hence $|\mathcal{X}_{\rm PNC}|$ can be further reduced.

Exclusive Law

In order to achieve unambiguous data recovery, the PNC mapping has to meet the exclusive law which is given by (2.19). There are L equations in (2.19), and each

corresponds to a specific user l. Clearly, the exclusive law requires $\mathcal{X}_{\mathrm{PNC}}$ to include at least $\prod_{l' \neq l} |\mathcal{X}_{l'}|$ elements to distinguish all possible combinations of the other $L-1$ users, which results in the same cardinality constraint as in (2.18).

$$\begin{aligned}
&\mathcal{M}(x_1, x_2, x_3 \cdots, x_L) \neq \mathcal{M}(x_1, x_2', x_3' \cdots, x_L'), \forall (x_2, x_3, \cdots, x_L) \neq (x_2', x_3', \cdots, x_L') \\
&\mathcal{M}(x_1, x_2, x_3 \cdots, x_L) \neq \mathcal{M}(x_1', x_2, x_3' \cdots, x_L'), \forall (x_1, x_3, \cdots, x_L) \neq (x_1', x_3', \cdots, x_L') \\
&\qquad\qquad\qquad\qquad \vdots \\
&\mathcal{M}(x_1, x_2, x_3 \cdots, x_L) \neq \mathcal{M}(x_1', x_2', x_3' \cdots, x_L), \forall (x_1, x_2, \cdots, x_{L-1}) \neq (x_1', x_2', \cdots, x_{L-1}').
\end{aligned} \tag{2.19}$$

Achievable Rate Region of PNC

Let $y = x_{\mathrm{ss}} + z$ denote the received signal by relay, with z is the Gaussian noise. The achievable rate region of PNC is defined as the maximum rate which can be employed in the MAC phase, expressed as

$$\mathcal{R}_1 = \mathcal{R}_2 = \cdots = \mathcal{R}_L \leq I(x_{\mathrm{PNC}}; y), \tag{2.20}$$

where $I(x_{\mathrm{PNC}}; y)$ denotes the mutual information between x_{PNC} and y.

Singular Fading

A channel fading value $\mathbf{h}$ is said to be singular if

$$|\mathcal{X}_{\mathrm{ss}}| < \prod_{l=1}^{L} |\mathcal{X}_l|, \tag{2.21}$$

which means that some points on the noise-free constellation diagram are merged, which poses potential difficulties to meet the exclusive law. Therefore, choosing the optimal PNC mapping for such singular fading states acts as a significant component in PNC research.

For a 2-WRC with QPSK modulation, singular fading occurs when $\frac{h_1}{h_2} =$

$\{0, \pm 1, \pm j, \frac{1}{2}(\pm 1 \pm j)\}$. By employing the minimum cardinality $\min |\mathcal{X}_{\mathrm{PNC}}| = 4$ (see equation 2.18), the authors in [23] proposed an optimal 4-ary denoise-and forward mapping strategy to deal with the cases of $\frac{h_1}{h_2} = \{\pm 1, \pm j\}$. Also in [23], a 5-ary non-linear mapping approach is presented to deal with the cases of $\frac{h_1}{h_2} = \frac{1}{2}(\pm 1 \pm j)$ which are unsolvable with 4-ary mappings; however the cardinality is increased. The singular fading states of 3-WRC and 4-WRC with QPSK modulation are investigated in [24] and [25] respectively, and their corresponding PNC mappings are also given in [24, 25]. For the more general multi-user multi-relay system shown in Figure 2.1, the authors in [26, 27] proposed a binary matrix based PNC approach which selects the optimal mapping by evaluating all possible binary matrices. Additionally, the exclusive law in (2.19) is translated to a 'full rank constraint' in a multi-relay system, and global selection is thus required at the central hub station to ensure the overall matrix is full rank.

2.5 Compute and Forward

Singular fading states vary with different modulation types and different numbers of users. Traditional PNC strategies reviewed so far analyse singular fading case by case. A flexible PNC approach which adaptively fits more general system models is thus preferred.

Compute and forward (C&F), as proposed in [28] has attracted much interest in recent years. It employs structured lattice codes for PNC. Each relay infers and forwards a linear combination of the transmitted codewords of all users. The lattice structure makes C&F stand out among many PNC schemes. In this section, we will review the fundamental principle of C&F, and illustrate connections between C&F and traditional PNC.

2.5.1 Preliminaries of Lattices

We first introduce some essential concepts and definitions in C&F.

Nested Lattice

Definition 2.5.1. An n-dimensional real lattice Λ is a discrete subset of n-space $\mathbb{R}^n$. It is generated by all integer combinations of the basis vectors, expressed as:

$$\Lambda = \{\mathbf{G}_\Lambda \mathbf{r} : \mathbf{r} \in \mathbb{Z}^{k\times 1}\}, \tag{2.22}$$

where $\mathbf{G}_\Lambda = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_k] \in \mathbb{R}^{n\times k}$ is a generator matrix for Λ, and each $\mathbf{g} \in \mathbb{R}^{n\times 1}$ is a basis vector of $\mathbf{G}_\Lambda$.

For example, $\Lambda = \{\alpha\mathbb{Z} : \alpha \in \mathbb{R}\}$ is a simple 1-dimensional lattice. The authors in [29] employ the concept of the principal ideal domain (PID) to extend real lattices to complex lattices for C&F. Let $\mathbb{A}$ denote a PID, which is an integer domain in which every ideal is principal. The most commonly used PIDs are the integers $\mathbb{Z}$, the Gaussian integers $\mathbb{Z}[i]$ and the Eisenstein integers $\mathbb{Z}[\omega]$. Formal definitions of $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ may be expressed as

$$\begin{aligned} \mathbb{Z}[i] =& \{a + bi : a, b \in \mathbb{Z}\} \\ \mathbb{Z}[\omega] =& \{a + b\omega : a, b \in \mathbb{Z} \text{ and } \omega = \frac{1}{2}(-1 + \sqrt{3}i)\}. \end{aligned} \tag{2.23}$$

Hence an n-dimensional lattice over $\mathbb{A}$ can be expressed as $\Lambda = \{\mathbf{G}_\Lambda \mathbf{r} : \mathbf{r} \in \mathbb{A}^{k\times 1}\}$ with $\mathbf{G}_\Lambda \in \mathbb{C}^{n\times k}$. The group property implies that any translation $\Lambda + \lambda$ by a lattice point $\lambda \in \Lambda$ is Λ again, hence Λ is geometrically uniformly distributed. The decision region of a lattice point $\lambda \in \Lambda$ is called the Voronoi region of λ, denoted as $\mathcal{V}_\Lambda(\lambda)$. Let $\mathcal{Q}_\Lambda(\mathbf{x})$ denote a quantiser which finds the closest λ to $\mathbf{x} \in \mathbb{C}^n$, hence $\mathcal{V}_\Lambda(\lambda)$ can be expressed as

$$\mathcal{V}_\Lambda(\lambda) = \{\mathbf{x} \in \mathbb{C}^n : \mathcal{Q}_\Lambda(\mathbf{x}) = \lambda\}. \tag{2.24}$$

The Voronoi region of $\mathcal{V}_\Lambda(\lambda = 0)$ is commonly used to represent the Voronoi region of Λ, denoted as $\mathcal{V}_\Lambda$. Note that $\mathcal{V}_\Lambda(\lambda)$ for all $\lambda \in \Lambda$ are congruent.

Definition 2.5.2. A lattice Λ' is said to be nested in a lattice Λ if Λ' is a sub-lattice in Λ, denoted as $\Lambda' \subseteq \Lambda$, Λ and Λ' are termed as the fine lattice and the coarse lattice respectively. The quotient ring Λ/Λ' is termed a lattice partition. A sequence of lattices $\Lambda_1, \Lambda_2, \cdots, \Lambda_L$ are nested if $\Lambda_1 \subseteq \Lambda_2 \cdots \subseteq \Lambda_L$.

Definition 2.5.3. A nested lattice code $\mathcal{C}(\Lambda, \Lambda')$ is defined as the set of all coset leaders

in Λ/Λ', expressed as

$$\mathcal{C}(\Lambda, \Lambda') = \Lambda \bmod \Lambda' = \{\lambda \bmod \Lambda' : \lambda \in \Lambda\}. \tag{2.25}$$

where $\bmod$ denotes the modulo operation. The Voronoi region of the coarse lattice $\mathcal{V}_{\Lambda'}$ is usually referred to as the shaping region of $\mathcal{C}(\Lambda, \Lambda')$. Note that when applied to a lattice Λ, the modulo operation is defined as

$$\mathbf{x} \bmod \Lambda = \mathbf{x} - \mathcal{Q}_{\Lambda}(\mathbf{x}). \tag{2.26}$$

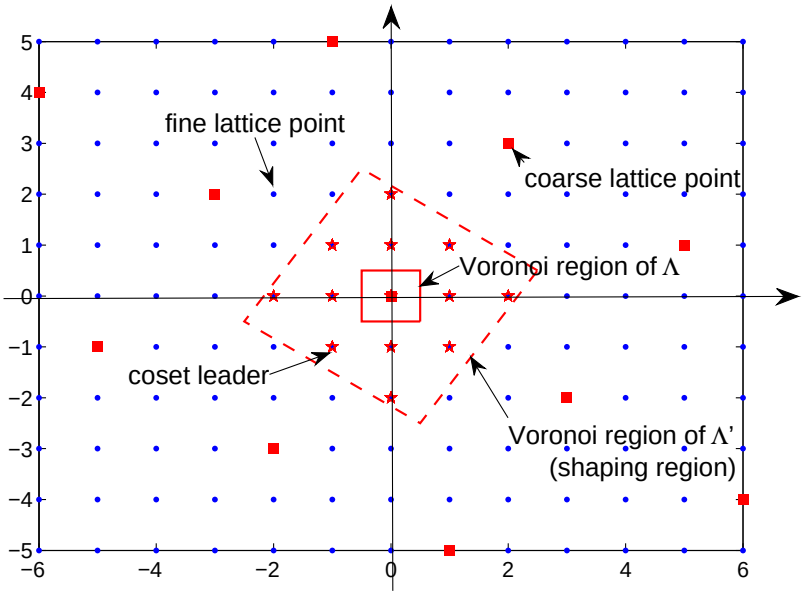

Figure 2.3: An example of lattice partition: $\Lambda/\Lambda' = \mathbb{Z}[i]/(2+3i)\mathbb{Z}[i] \cong \mathbb{F}_{13}$

Figure 2.3 demonstrates an example of a 2-dimensional lattice partition. The fine lattice $\Lambda = \mathbb{Z}[i]$ and the coarse lattice $\Lambda' = (2+3i)\mathbb{Z}[i]$ are represented by blue dots and red squares respectively. The fine lattice Λ is divided into 13 cosets by Λ/Λ', and each coset can be represented by a coset leader which is located within the Voronoi region of Λ'.

Lattice Construction A

Let π denote a prime in $\mathbb{A}$, and $\mathcal{C}(n, k)$ be a linear code of length n defined on $\mathbb{A}/\pi\mathbb{A}$, where $k < n$. For example, the binary code is defined on $\mathbb{Z}/2\mathbb{Z}$ and 2 in a prime in $\mathbb{Z}$. Let

$\mathbf{G} \in (\mathbb{A}/\pi\mathbb{A})^{n\times k}$ be the generator matrix of $\mathcal{C}(n,k)$, the codebook can be expressed as

$$\mathcal{C} = \{\mathbf{G}\mathbf{x} : \mathbf{x} \in (\mathbb{A}/\pi\mathbb{A})^k\}. \tag{2.27}$$

Hence, the codebook $\mathcal{C}$ contains $|\mathbb{A}/\pi\mathbb{A}|^k$ codewords. $|\mathbb{A}/\pi\mathbb{A}|$ denotes the number of cosets in $\mathbb{A}/\pi\mathbb{A}$ which can be treated as the symbolwise cardinality. Based on $\mathcal{C}$, a lattice $\Lambda_\mathcal{C}$ can be generated by employing lattice construction A, which can be expressed as

$$\Lambda_\mathcal{C} = \{\lambda \in \mathbb{A}^n : (\lambda \bmod \pi) \in \mathcal{C}\}, \tag{2.28}$$

where the modulo operation is applied to each component of λ. Alternatively, (2.28) can be expressed as

$$\Lambda_\mathcal{C} = \pi\mathbb{A}^n + \mathcal{C} = \bigcup_{\mathbf{c}\in\mathcal{C}} (\pi\mathbb{A}^n + \mathbf{c}), \tag{2.29}$$

where $\mathbf{c}$ denotes the codeword in $\mathcal{C}$. Let $\Lambda = \Lambda_\mathcal{C}$ be the fine lattice, and $\Lambda' = \pi\mathbb{A}^n$ be the coarse lattice. The quotient ring Λ/Λ' has $|\mathbb{A}/\pi\mathbb{A}|^k$ cosets, and each coset corresponds to a codeword in $\mathcal{C}$, and this exactly corresponds to the definition 2.5.3. Since the codewords in $\mathcal{C}(\Lambda,\Lambda')$ are located within the fundamental region of Λ', the intuitive function of construction A is to generate shifted duplications of $\mathcal{C}$, and thus Λ can be distributed over the entire message space $\mathbb{A}^n$.

Geometrical Parameters of Lattices

The main geometrical parameters of an n-dimensional lattice Λ are as follows [30, 31]:

- The minimum squared Euclidean distance between lattice points, denoted as $d^2_{\min}(\Lambda)$. It also represents the minimum weight (defined in an Euclidean sense rather in a Hamming sense) among codewords.
- The volume of the Voronoi region $\mathcal{V}_\Lambda$, denoted as $V(\Lambda)$.
- The nominal coding gain $\gamma_\mathrm{c}(\Lambda)$: defined as

$$\gamma_\mathrm{c}(\Lambda) = \frac{d^2_{\min}(\Lambda)}{V(\Lambda)^{2/n}}, \tag{2.30}$$

where the quantity $V(\Lambda)^{2/n}$ can be treated as the normalised volume of Λ per two dimension (or per one complex dimension).

The main geometrical parameters of an n-dimensional nested lattice code $\mathcal{C}(\Lambda, \Lambda')$ are as follows:

- The volume of the Voronoi region of Λ' (also known as the shaping region), denoted as $V(\Lambda')$. It can be calculated by $V(\Lambda) = \int_{\mathcal{V}_{\Lambda'}} d\mathbf{x}$.
- The second moment (average power) per dimension $\sigma^2(\Lambda')$, expressed as [32]

$$\sigma^2(\Lambda') = \frac{1}{n} \frac{\int_{\mathcal{V}_{\Lambda'}} \|\mathbf{x}\|^2 d\mathbf{x}}{V(\Lambda')}. \tag{2.31}$$

- The normalised second moment $G(\Lambda')$, defined as

$$G(\Lambda') = \frac{\sigma^2(\Lambda')}{V(\Lambda')^{2/n}} = \frac{1}{n} \frac{\int_{\mathcal{V}_{\Lambda'}} \|\mathbf{x}\|^2 d\mathbf{x}}{V(\Lambda')^{1+2/n}}. \tag{2.32}$$

- The shaping gain $\gamma_{\mathrm{s}}(\Lambda') = \frac{1/12}{G(\Lambda')}$ which measures the decrease in average energy of Λ' relative to a hypercubic region.

2.5.2 Standard Compute and Forward

Now we introduce the fundamental principle of C&F.

System Model

We assume that the system includes L users and M relays which are all equipped with a single antenna. Without loss of generality, we employ complex valued channels with lattices over complex integer domain $\mathbb{A}$ in this section, although C&F was originally presented with real valued channels and $\mathbb{Z}$-lattices [28]. The process of C&F can be divided into 3 stages.

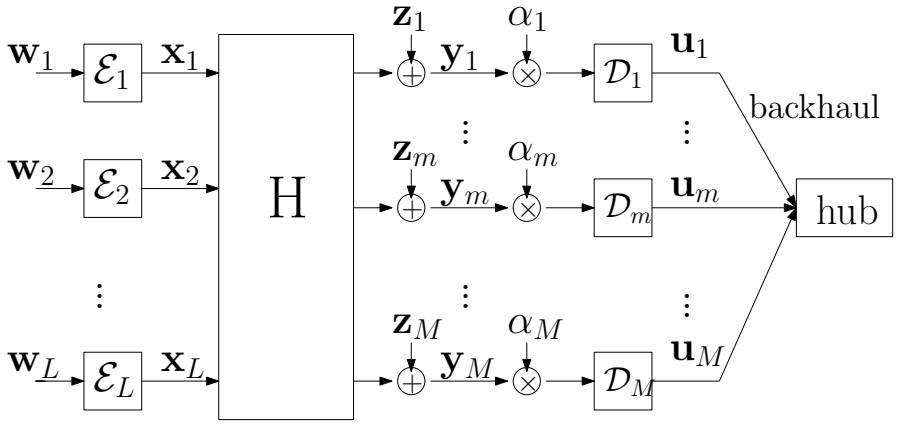

Figure 2.4: System diagram of compute and forward

Encoding: the length k_l original message of user l is denoted as $\mathbf{w}_l \in (\mathbb{A}/\pi\mathbb{A})^{k_l}$. The quotient ring $\mathbb{A}/\pi\mathbb{A}$ is usually isomorphic to a finite field $\mathbb{F}_p$, expressed as $\mathbb{A}/\pi\mathbb{A} \cong \mathbb{F}_p$, where p is a prime in $\mathbb{Z}$ and $|\pi|^2 = p$. For example, the lattice partition $\mathbb{Z}[i]/(2+3i)\mathbb{Z}[i]$ illustrated in Figure 2.3 is an isomorphism to $\mathbb{F}_{13}$.

C&F requires the lattices generated by all users to have a common structure, hence the message vectors $\mathbf{w}_l$ are zero-padded to a common length $k = \max_l k_l$ before transmission. The encoding process for user l is expressed as:

$$\mathbf{x}_l = \mathcal{E}_l(\mathbf{w}_l),\ \mathbf{x}_l \in \Lambda \bmod \Lambda'. \tag{2.33}$$

The encoder $\mathcal{E}_l$ employs a linear code $\mathcal{C}(n,k)$ with a generator matrix $\mathbf{G} \in (\mathbb{A}/\pi\mathbb{A})^{n\times k}$ to map $\mathbf{w}_l$ to a length n codeword $\mathbf{x}_l$, and the corresponding Λ and Λ' can be obtained by employing lattice construction A which is expressed in (2.27)$\sim$(2.29). The message rate of user l is

$$\mathcal{R}_l = \frac{k_l}{n}\log|\pi|^2 \tag{2.34}$$

per complex dimension. Note that the generator matrix $\mathbf{G}$ is not unique for a specific lattice, hence users can utilise different encoders to obtain the desired structure Λ/Λ', as we will discuss in section 2.6.2.

In quantisation theory [33], a random dither vector $\mathbf{d}_l \in \mathbb{C}^n \bmod \Lambda'$ might be employed to make the discrete input signal $\mathbf{x}_l$ continuously uniform, written as $\mathbf{x}_l = (\mathbf{x}_l + \mathbf{d}_l) \bmod \Lambda'$. The dither vector is known by both transmitter and receiver, hence can be easily removed at the relays. Since we are more interested in the lattice structure of C&F, the dither

process is omitted here for simplicity.

Coefficient Selection: this is also a core component in C&F. Again, we use $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_M]$, $\mathbf{h}_m = [h_{m,1}, \cdots, h_{m,L}]^T$ and $h_{m,l} \sim \mathcal{CN}(0,1)$ to denote channel coefficients. The received superimposed signal at the m-th relay can be expressed as:

$$\mathbf{y}_m = \sum_{l=1}^{L} h_{m,l}\mathbf{x}_l + \mathbf{z}_m,\ \mathbf{y}_m \in \mathbb{C}^n, \tag{2.35}$$

where the noise $\mathbf{z}_m \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_n)$ is a length n circularly symmetrical complex Gaussian random vector. Let $\frac{1}{n}\mathbb{E}(\|\mathbf{x}_l\|^2) \leq P$ as the power constraint per channel use. The signal to noise ratio is thus $\mathrm{SNR} = P/\sigma^2$.

The received signal vector $\mathbf{y}_m$ is first scaled by a factor $\alpha_m \in \mathbb{C}$. The relay attempts to choose an integer linear combination of the transmitted codewords to represent the scaled received signal, written as

$$\mathcal{Q}_\Lambda(\alpha_m\mathbf{y}_m) = \sum_{l=1}^{L} a_{ml}\mathbf{x}_l, \tag{2.36}$$

where $\mathcal{Q}_\Lambda$ quantises $\alpha_m\mathbf{y}_m$ to its closest fine lattice point in Λ. The quantisation error contributes to the effective noise at relay m, expressed as

$$\mathbf{z}_{\mathrm{eff},m} = \sum_{l=1}^{L}(\alpha_m h_{ml} - a_{ml})\mathbf{x}_l + \alpha_m\mathbf{z}_m, \tag{2.37}$$

where a_{ml} is an integer in $\mathbb{A}$, and hence the integer combination $\sum_{l=1}^{L}\alpha_{ml}\mathbf{x}_l$ is also a lattice point in Λ. Let $\mathbf{a}_m = [a_{m1}, a_{m2}, \cdots, a_{mL}]^T$ denote the coefficient vector of the linear function. The scaling factor α_m aims to force the scaled channel vector $\alpha_m\mathbf{h}_m$ to approximate an integer vector $\mathbf{a}_m$. The effective noise comprises 2 components:

- Self noise: $\mathbf{z}_{\mathrm{self},m} = \sum_{l=1}^{L}(\alpha_m h_{ml} - a_{ml})\mathbf{x}_l$: caused by the mismatch between the selected integer vector and the scaled channel.
- Scaled Gaussian noise $\mathbf{z}_{\mathrm{sg},m} = \alpha_m\mathbf{z}_m$: the Gaussian noise at the m-th relay is scaled by the scaling factor α_m.

For a given coefficient vector $\mathbf{a}_m$, the achievable computation rate per complex dimension

is given by [28]

$$\mathcal{R}_m(\mathbf{h}_m, \mathbf{a}_m) = \max_{\alpha \in \mathbb{C}} \log^+ \left(\frac{P}{\alpha_m^2 \sigma^2 + P\|\alpha_m \mathbf{h}_m - \mathbf{a}_m\|^2} \right) \tag{2.38}$$

bits per channel use, where $\log^+(\cdot) = \max(\log(\cdot), 0)$, and

$$\sigma_{\text{eff}}^2 = \alpha_m^2 \sigma^2 + P\|\alpha_m \mathbf{h}_m - \mathbf{a}_m\|^2 \tag{2.39}$$

denotes the variance of the effective noise. The Minimum Mean Square Error (MMSE) solution of α_m for a given $\mathbf{a}_m$ is

$$\alpha_{\text{MMSE}}(\mathbf{a}_m) = \frac{\text{SNR}\mathbf{h}_m^H \mathbf{a}_m}{1 + \text{SNR}\|\mathbf{h}_m\|^2}, \tag{2.40}$$

and hence (2.38) can be rewritten as

$$\mathcal{R}_m(\mathbf{h}_m, \mathbf{a}_m) = \log^+ \left(\frac{1}{\mathbf{a}_m^H \mathbf{M} \mathbf{a}_m} \right), \tag{2.41}$$

where $\mathbf{M} = \mathbf{I}_L - \frac{\text{SNR}}{\text{SNR}\|\mathbf{h}_m\|^2 + 1} \mathbf{h}_m \mathbf{h}_m^H$. The target of each relay is to find the integer vector $\mathbf{a}_m$ which maximises the computation rate, expressed as

$$\mathbf{a}_{\text{opt},m} = \underset{\mathbf{a}_m \in \mathbb{A}^L \setminus \{\mathbf{0}\}}{\text{argmax}} \ \mathcal{R}(\mathbf{h}_m, \mathbf{a}_m). \tag{2.42}$$

Note that the all zero vector $\{\mathbf{0}\}$ is not allowed as it contains no information. Specific algorithms for coefficient selection will be investigated in Chapter 3.

Decoding: with the optimal integer vector $\mathbf{a}_{\text{opt},m}$ and its corresponding α_m, the decoding process comprises 3 steps as follows:

- $\alpha_m \mathbf{y}_m \to \Lambda$: decode an integer combination $\sum_{l=1}^{L} a_{m,l} \mathbf{x}_l$ from $\alpha_m \mathbf{y}_m$, as shown in (2.36). The combination is a lattice point in Λ.
- $\Lambda \to \Lambda/\Lambda'$: map the integer combination to a codeword in $\mathcal{C}(\Lambda, \Lambda')$ which is equivalent to shifting the $\sum_{l=1}^{L} a_{m,l} \mathbf{x}_l$ into $\mathcal{V}_{\Lambda'}$, written as $\sum_{l=1}^{L} a_{m,l} \mathbf{x}_l \bmod \Lambda'$.
- $\Lambda/\Lambda' \to (\mathbb{A}/\pi\mathbb{A})^k$: map the length-$n$ codeword back to the length k message, and the integer combination of the codewords becomes the linear combination of the

transmitted messages, expressed as

$$\sum_{l=1}^{L} a_{m,l}\mathbf{x}_l \rightarrow \mathbf{u}_m = \bigoplus_{l=1}^{L} q_{m,l}\mathbf{w}_l \tag{2.43}$$

with $q_{m,l} \in \mathbb{A}/\pi\mathbb{A}$. Note that the calculation of $\bigoplus_{l=1}^{L} q_{m,l}\mathbf{w}_l$ are operated over $\mathbb{A}/\pi\mathbb{A}$. Both $a_{m,l}$ and $q_{m,l}$ can be regarded as coefficients in C&F, the former corresponds to an integer domain $\mathbb{A}$, and the latter corresponds to a lattice partition $\mathbb{A}/\pi\mathbb{A}$ (or a finite finite $\mathbb{F}_p$).

Message Recovery

The last step in C&F is that the hub station recovers all individual $\mathbf{w}_l$ from the M linear equations (2.43) provided by the M relays. Let $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_M]$ be the overall coefficient matrix, with $\mathbf{q}_m = [q_{m,1}, q_{m,2}, \cdots, q_{m,L}]^T$ as the coefficient vector selected by relay m. The messages of all users can be recovered if and only if $\mathrm{Rank}(\mathbf{Q}) = L$. Particularly, when $M = L$, it is equivalent to say the determinant of $\mathbf{Q}$ has to be non-zero. Note that the elements in $\mathbf{Q}$ are defined on a quotient ring, and 'non-zero' here implies that $\det(\mathbf{Q})$ cannot be any zero-divisor of $\mathbb{A}/\pi\mathbb{A}$ [14, 34].

Connections with Traditional PNC

In this section, we review C&F from the traditional PNC perspective (section 2.4.2).

- Cardinality: traditional PNC schemes focus on the scenario that L users intend to exchange data via one relay, and hence the minimum cardinality required at the relay is $\frac{\prod_l |\mathcal{X}_l|}{|\mathcal{X}_l|}$. In C&F, the minimum number of relays required is the number of users. For a particular relay, the other $L-1$ relays provide side information, hence the minimum cardinality of the network coded symbol is $\frac{\prod_l |\mathcal{X}_l|}{\prod_{l' \neq l} |\mathcal{X}_{\mathrm{PNC},l'}|}$. Due to the nested lattice structure in C&F, the cardinality of the network coded symbol is equal to the cardinality of the modulated symbol, expressed as $|\mathcal{X}_{\mathrm{PNC}}| = |\mathcal{X}_l|$, hence C&F employs the theoretical minimum cardinality.

- Exclusive law: in traditional PNC schemes, unambiguous decodability is ensured by the exclusive law. In C&F, the exclusive law can be simply regarded as the rank-constraint of the overall mapping matrix $\mathbf{Q}$.

- Singular fading: as we discussed previously, singular fading states greatly depend on the number of users, and hence traditional PNC strategies have to deal with this issue case by case. In C&F, the scaled channel vector $\alpha\mathbf{h}$ is forced to an integer vector $\mathbf{a}$, and $\sum_{l=1}^{L} a_l\mathbf{x}_l \bmod \Lambda' \in \mathcal{C}(\Lambda, \Lambda')$ holds for any values of L. Hence, a main benefit of C&F is that it adaptively fits for an arbitrary number of users.

- Achievable Rate: the achievable rate of PNC is defined as the mutual information between the received signal and the network coded symbol, as expressed in (2.20). Based on this, the achievable rate of C&F can be expressed as

$$\begin{aligned}\mathcal{R}_m &= \frac{1}{n} I\big(\sum_{l=1}^{L} a_{ml}\mathbf{x}_l \bmod \Lambda'; \alpha_m\mathbf{y}_m\big) \\ &= \frac{1}{n} H\big(\sum_{l=1}^{L} a_{ml}\mathbf{x}_l \bmod \Lambda'\big) - \frac{1}{n} H\big(\sum_{l=1}^{L} a_{ml}\mathbf{x}_l \bmod \Lambda' | \alpha_m\mathbf{y}_m\big)\end{aligned} \tag{2.44}$$

  bits per channel use. Since $\sum_{l=1}^{L} a_{ml}\mathbf{x}_l \bmod \Lambda'$ and $\mathbf{x}_l$ have exactly the same constellation, hence $H(\sum_{l=1}^{L} a_{ml}\mathbf{x}_l \bmod \Lambda') = H(\mathbf{x}_l)$. It is known that for a sufficiently large n, the projection of a uniform probability distribution over an n-sphere $\mathcal{V}_{\Lambda'}$ onto 1 or 2 dimensions is a nonuniform probability distribution that approaches a Gaussian distribution [32], hence (2.44) can be rewritten as

$$\mathcal{R}_m = \log(2\pi eP) - \log(2\pi e\sigma_{\text{eff}}^2) = \log(\frac{P}{\sigma_{\text{eff}}^2}), \tag{2.45}$$

  which results in the same expression as (2.38).

2.5.3 Compute and Forward over Eisenstein Integers

Lattices over Eisenstein Integers

Previously we have mentioned that C&F can be applied in any PID $\mathbb{A}$, however early work in C&F associated with complex valued channels focusses on cubic or hypercubic lattices

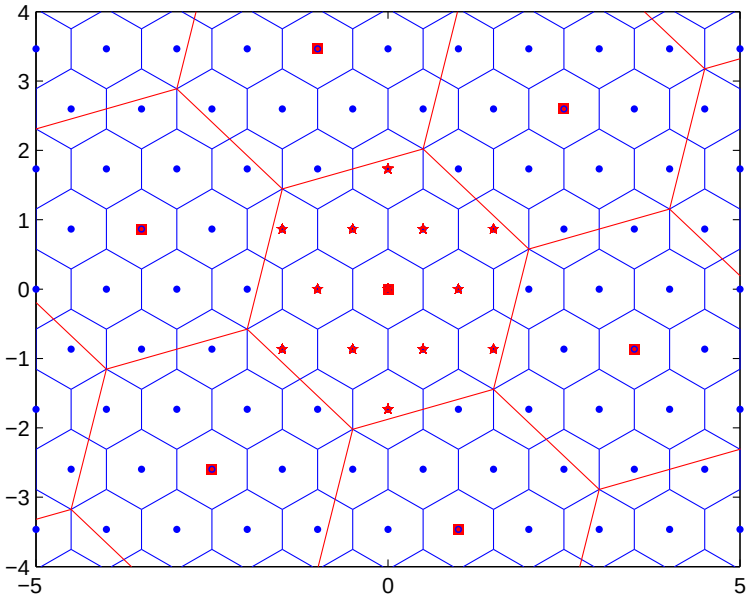

Figure 2.5: Lattice over Eisenstein integers: $\mathbb{Z}[\omega]/(4+3\omega)\mathbb{Z}[\omega] \cong \mathbb{F}_{13}$

based on $\mathbb{Z}[i]$ only. Recently, more focus is given to the Eisenstein integer $\mathbb{Z}[\omega]$-based lattices [35–38]. Recall that $\mathbb{Z}[\omega]$ takes form of $a+b\omega$ where both a and b are integers and $\omega = \frac{-1+\sqrt{3}i}{2}$. Figure 2.5 illustrates an example of a $\mathbb{Z}[\omega]$-lattice. Again, the fine lattices and the coarse lattices are represented by blue points and red squares respectively. The Eisenstein integers have a hexagonal shape which is the densest packing structure in the 2-dimensional complex plane [30]. The denser packing structure implies that $\mathbb{Z}[\omega]$-lattices are more energy efficient. Numerical results in [36] and [35] reveal that $\mathbb{Z}[\omega]$-based C&F outperforms $\mathbb{Z}[i]$-based C&F in terms of computation rate and symbol error performance respectively.

Eisenstein Integers vs Gaussian Integers

The performance of a nested lattice code $\mathcal{C}(\Lambda, \Lambda')$ is usually measured by the nominal coding gain $\gamma_{\mathrm{c}}(\Lambda)$ and the shaping gain $\gamma_{\mathrm{s}}(\Lambda')$, which are defined in section 2.5.1. We take a simple example to compare $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ lattices. Let $\Lambda_1/\Lambda_1' = (\mathbb{Z}[i]/\pi_1\mathbb{Z}[i])^n$ and $\Lambda_2/\Lambda_2' = (\mathbb{Z}[\omega]/\pi_2\mathbb{Z}[\omega])^n$ respectively denote lattice partitions over $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$, with $|\pi_1| = |\pi_2|$. It is known that both $\gamma_{\mathrm{c}}(\Lambda_1)$ and $\gamma_{\mathrm{s}}(\Lambda_1')$ are zero for cubic lattices.

For $\mathbb{Z}[\omega]$ lattices, the minimum distance and volume are $d^2_{\min}(\Lambda_2) = 1$ and $V(\Lambda_2) = (\frac{\sqrt{3}}{2})^n$ respectively, and hence the nominal coding is $\gamma_{\rm c}(\Lambda_2) = \frac{\sqrt{3}}{2} = 0.625$ dB. The shaping region $\mathcal{V}_{\Lambda_2'}$ is a product of n hexagons, and hence $V(\Lambda_2') = (\frac{\sqrt{3}}{2}|\pi_2|^2)^n$. By applying continuous approximation, the second moment per complex dimension is given by [35]

$$\sigma^2(\Lambda_2') = \frac{1}{n}\frac{\int_{\mathcal{V}_{\Lambda_2'}} \|\mathbf{x}\|^2 d\mathbf{x}}{V(\Lambda_2')} = \frac{5}{36n}|\pi_2|^2. \tag{2.46}$$

By combining (2.46) and (2.32), the normalised second moment $G(\Lambda_2')$ can be calculated as $\frac{5}{18\sqrt{3}}$, and hence the shaping gain $\gamma_{\rm s}(\Lambda_2')$ is $\frac{1/6}{G(\Lambda_2')} = 0.167$ dB. The examples in Figure 2.3 and Figure 2.5 employ a common constellation size $|\pi_1|^2 = |\pi_2|^2 = 13$, and both have $d_{\min}(\Lambda) = 1$. It can be easily calculated that the average power of the constellation points in $\mathbb{Z}[\omega]$ case is less than in the $\mathbb{Z}[i]$ case.

Note that to achieve these gains above, the system only needs to utilise a $\mathbb{Z}[\omega]$-quantiser instead of a $\mathbb{Z}[i]$-quantiser, and this will slightly increase the cardinality due to the irrational component in $\mathbb{Z}[\omega]$.

2.6 Integer Forcing Linear Receiver

The C&F scheme deals with distributed systems with single antenna APs: hence each linear combination is calculated based on one signal observation. In this section, we review a centralised version of C&F, namely, integer forcing (IF) [39]. It combines multiple received signals to obtain the integer coefficients.

2.6.1 System Model

Let L denote the number users and $N_{\rm r}$ be the number of AP antennas, with $N_{\rm r} \geq L$. Let $\mathbf{H} \in \mathbb{C}^{N_{\rm r}\times L}$ and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L]^T$ denote the channel and signal matrix respectively, where each $\mathbf{x}_l \in \mathbb{C}^{n\times 1}$ is a length n codeword. Let $\mathbf{Z} \in \mathbb{C}^{N_{\rm r}\times n}$ denote the noise matrix; we have $\mathbf{Y} = \mathbf{HX} + \mathbf{Z}$ and $\mathbf{Y} \in \mathbb{C}^{N_{\rm r}\times n}$. As shown in Figure 2.6, the AP applies an

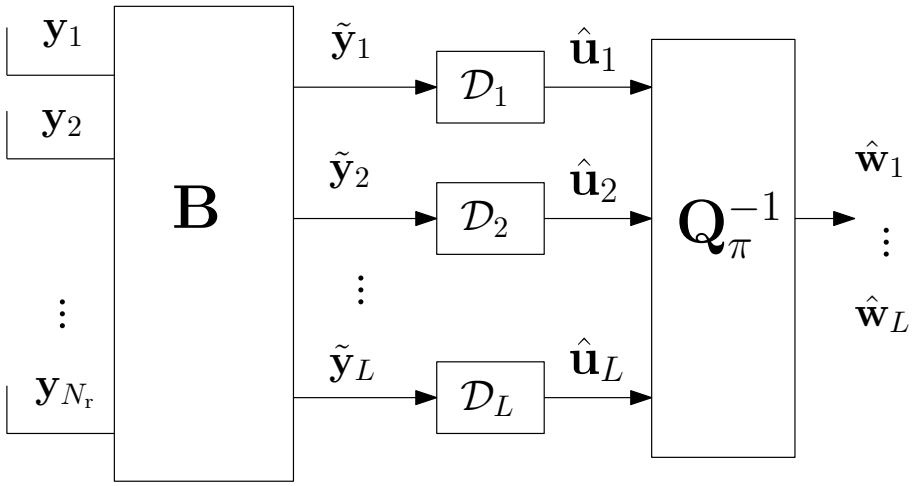

Figure 2.6: System diagram of integer forcing linear receiver

equaliser $\mathbf{B} \in \mathbb{C}^{L \times N_\mathrm{r}}$ to obtain

$$\begin{aligned}\tilde{\mathbf{Y}} &= \mathbf{BY} = \mathbf{BHX} + \mathbf{BZ} \\ &= \mathbf{AX} + (\mathbf{BH} - \mathbf{A})\mathbf{X} + \mathbf{BZ},\end{aligned} \tag{2.47}$$

where $\mathbf{A} = \{a_{l',l}\} \in \mathbb{A}^{L \times L}$ represents the integer coefficient matrix, comprising L linearly independent equations over $\mathbb{A}$. Letting $\tilde{\mathbf{y}}_{l'}^T$, $\mathbf{b}_{l'}^T$ and $\mathbf{a}_{l'}^T$ denote the l'-th rows of $\tilde{\mathbf{Y}}$, $\mathbf{B}$ and $\mathbf{A}$ respectively, we have

$$\tilde{\mathbf{y}}_{l'}^T = \mathbf{a}_{l'}^T\mathbf{X} + (\mathbf{b}_{l'}^T\mathbf{H} - \mathbf{a}_{l'}^T)\mathbf{X} + \mathbf{b}_{l'}^T\mathbf{Z}, \tag{2.48}$$

where the 2nd and 3rd terms in (2.48) correspond to the self noise and the scaled Gaussian noise respectively. Each data stream $\mathbf{y}_{l'}^T$ is sent to a lattice decoder $\mathcal{D}_{l'}$ which aims to recover a linear combination

$$\mathbf{u}_{l'} = \Big(\sum_{l=1}^{L} a_{l',l}\mathbf{w}_l\Big) \bmod \pi = \bigoplus_{l=1}^{L} q_{l',l}\mathbf{w}_l, \quad q_{l',l} \in \mathbb{A}/\pi\mathbb{A}, \tag{2.49}$$

and the corresponding achievable rate of the l'th data stream is [39]

$$\begin{aligned}\mathcal{R}(\mathbf{H}, \mathbf{a}_{l'}, \mathbf{b}_{l'}) &= \log^+\Big(\frac{\mathrm{SNR}}{\sigma_{\text{eff}}^2}\Big) \\ &= \log^+\Big(\frac{\mathrm{SNR}}{\|\mathbf{b}_{l'}\|^2 + \mathrm{SNR}\|\mathbf{b}_{l'}^T\mathbf{H} - \mathbf{a}_{l'}^T\|^2}\Big).\end{aligned} \tag{2.50}$$

Similar to C&F, there exists an MMSE solution of $\mathbf{b}_{l'}$ for given $\mathbf{a}_{l'}$, expressed as

$$\mathbf{b}_{\mathrm{MMSE}}^{T}(\mathbf{a}_{l'}) = \frac{\mathrm{SNR}\mathbf{a}_{l'}^{T}\mathbf{H}^{H}}{\mathbf{I}_{N_{\mathrm{r}}} + \mathrm{SNR}\mathbf{H}\mathbf{H}^{H}}, \tag{2.51}$$

and substituting (2.51) back in (2.50), we have

$$\mathcal{R}(\mathbf{H}, \mathbf{a}_{l'}) = \log^{+}\Big(\frac{1}{\mathbf{a}_{l'}^{H}\mathbf{M}\mathbf{a}_{l'}}\Big), \tag{2.52}$$

where

$$\mathbf{M} = \mathbf{I}_{L} - \mathrm{SNR}\mathbf{H}^{H}(\mathbf{I}_{N_{\mathrm{r}}} + \mathrm{SNR}\mathbf{H}\mathbf{H}^{H})^{-1}\mathbf{H}. \tag{2.53}$$

Both of C&F and integer forcing consider the case that all user messages are encoded with a common rate, hence the achievable rate of the system is determined by the worst data stream. Thus the AP is required to find a proper matrix $\mathbf{A}$ to maximise

$$\mathcal{R}(\mathbf{H}) = \max_{\mathbf{A}\in\mathbb{A}^{L\times L}} \min_{l'=1}^{L} \mathcal{R}(\mathbf{H}, \mathbf{a}_{l'}). \tag{2.54}$$

Note that unambiguous decodability requires that the coefficient matrix $\mathbf{Q}_{\pi} = \mathbf{A} \bmod \pi$ has to be full rank over $\mathbb{A}/\pi\mathbb{A}$.

Integer Forcing vs Zero-Forcing

It is known that zero-forcing employs $\mathbf{H}^{\dagger}$ as the equaliser, hence it can be regarded as a special case of IF setting the integer matrix $\mathbf{A} = \mathbf{I}_{L}$. More precisely, the MMSE solution in (2.51) makes the case of $\mathbf{A} = \mathbf{I}_{L}$ equivalent to MMSE equalisation (also known as the regularised zero-forcing).

2.6.2 Lattice Reduction

It can be observed from (2.41) and (2.52) that maximising the computation rate is equivalent to minimising $\mathbf{a}^{H}\mathbf{M}\mathbf{a}$. Since $\mathbf{M}$ is a positive definite matrix, which means that $\mathbf{M}$ has a unique Cholesky decomposition $\mathbf{M} = \mathbf{L}\mathbf{L}^{H}$, where $\mathbf{L}$ is a lower triangular matrix. Thus, the optimisation problem in C&F is actually to minimise $\|\mathbf{L}^{H}\mathbf{a}\|$, which is equivalent to

finding the shortest vector in the lattice

$$\Lambda = \{\mathbf{L}^H \mathbf{a} : \ \mathbf{a} \in \mathbb{A}^L\}, \tag{2.55}$$

where $\mathbf{L}^H = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_L] \in \mathbb{C}^{L \times L}$ is the generator matrix of Λ. Correspondingly, the problem for integer forcing is actually a shortest independent vectors problem (SIVP).

Basis Reduction

It is known that the generator matrix $\mathbf{G}_\Lambda$ is not unique for a given Λ. The basic idea of lattice reduction is that for a given $\mathbf{G}_\Lambda$, we employ a linear transform $\mathcal{T} \in \mathbb{A}^{L \times L}$ to construct a new generator matrix $\mathbf{G}' = \mathbf{G}_\Lambda \mathcal{T}$ which comprises a shorter and more orthogonal basis. As shown in Figure 2.7, Λ is a 2-dimensional lattice over $\mathbb{Z}$. It can be generated by either $\mathbf{G}_\Lambda = [\mathbf{g}_1, \mathbf{g}_2] = [2\ 3; 2\ 2]$ (red arrows) or $\mathbf{G}'_\Lambda = [\mathbf{g}'_1, \mathbf{g}'_2] = [1\ 0; 0\ 2]$ (black arrows). Clearly, $\mathbf{G}'_\Lambda$ is preferred as it has the shortest basis to generate Λ. Now we briefly outline the process of acquiring $\mathbf{G}'_\Lambda$ based on $\mathbf{G}_\Lambda$.

1. Sort $\mathbf{g}_1$, $\mathbf{g}_2$ in ascending order according to their Euclidean-norm.
2. Let $\eta = \frac{\langle \mathbf{g}_1, \mathbf{g}_2 \rangle}{\|\mathbf{g}_1\|^2}$, and then $\mathbf{g}_2 = \mathbf{g}_2 - \lfloor \eta \rceil \mathbf{g}_1$.
3. If $\mathbf{g}_2 < \mathbf{g}_1$, swap $\mathbf{g}_1$, $\mathbf{g}_2$ and then go back to step 2, otherwise stop.

Let θ denote the angle between $\mathbf{g}_1$ and $\mathbf{g}_2$, and $\langle \mathbf{g}_1, \mathbf{g}_2 \rangle$ as their inner product. Note that $\frac{\langle \mathbf{g}_1, \mathbf{g}_2 \rangle}{\|\mathbf{g}_1\|} = \frac{\|\mathbf{g}_1\| \|\mathbf{g}_2\| \cos(\theta)}{\|\mathbf{g}_1\|} = \|\mathbf{g}_2\| \cos\theta$ is the projection of $\mathbf{g}_2$ on $\mathbf{g}_1$, hence by ignoring the rounding opearation, the algorithm above can be regarded as the Gram-Schmidt process [40]. This 2-Dimensional reduction method can be extended to other integer domains by replacing the rounding operation in step 2 by other quantisations $\mathcal{Q}_\mathbb{A}$ [35].

LLL Algorithm

For the case of large L, it is known that this problem is N-P hard [41], which means that there is no polynomially solvable algorithm which ensures the optimal solution. The

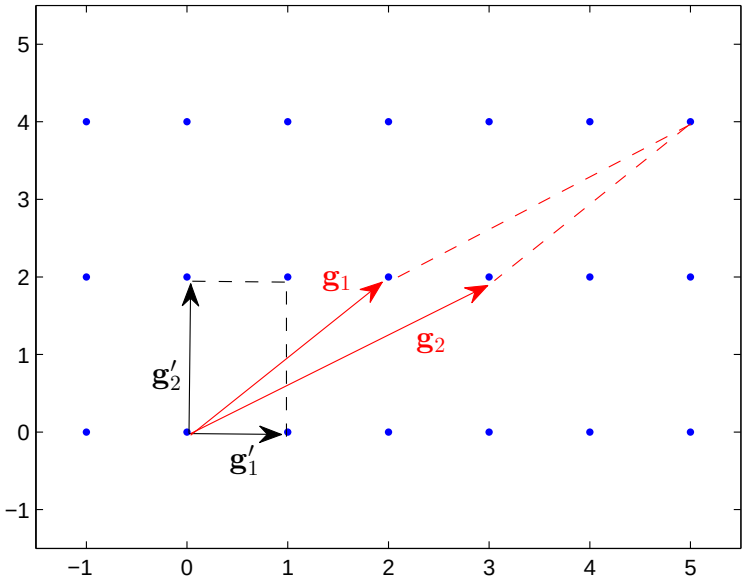

Figure 2.7: An example of reduced basis

Lenstra-Lenstra-Lovász (LLL) lattice reduction algorithm is commonly used to obtain a sub-optimal reduced basis $\mathbf{G}_\Lambda = [\mathbf{g}_1, \cdots, \mathbf{g}_L]$, with $\|\mathbf{g}_1\| \leq \|\mathbf{g}_2\| ... \leq \|\mathbf{g}_L\|$ and $\det(\mathbf{G}_\Lambda) = \pm 1$. Similar to the above algorithm, the LLL algorithm is also based on the Gram-Schmidt process, and performs QR decomposition after each swap; a detailed description can be found in [41, 42]. Unlike the standard Gram-Schmidt process, the quantisation $\mathcal{Q}_{\mathbb{A}}(\eta)$ implies that the reduced basis is not fully orthogonal.

Definition 2.6.1. A basis $\{\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_L\}$ of Λ is called a reduced basis if the orthogonal basis vectors $\{\mathbf{g}_1^*, \mathbf{g}_2^*, \cdots, \mathbf{g}_L^*\}$ calculated by the Gram-Schmidt process satisfy $\|\mathbf{g}_l^*\|^2 \leq \|\mathbf{g}_{l+1}^*\|^2$ for all $1 \leq l < L$ [42].

It is known that $\|\lambda\| \geq \min\{\|\mathbf{g}_1^*\|, \|\mathbf{g}_2^*\|, \cdots, \|\mathbf{g}_L^*\|\}$ for any non-zero $\lambda \in \Lambda$ [41]. Let $\lambda_{\min}$ denote the actual shortest vector in Λ. From definition 2.6.1, we have $\|\mathbf{g}_l^*\|^2 \geq 2^{-(l-1)}\|\mathbf{g}_1^*\|^2 \geq 2^{-(L-1)}\|\mathbf{g}_1^*\|^2, \forall l$. Thus, $\|\lambda_{\min}\| \geq 2^{\frac{-(L-1)}{2}}\|\mathbf{g}_1^*\|^2$. Since $\|\mathbf{g}_1^*\| = \|\mathbf{g}_1\|$, it can be concluded that the length of the LLL-acquired shortest vector $\mathbf{g}_1$ is at most $2^{\frac{L-1}{2}}$ times the length of the actual $\lambda_{\min}$.

2.6.3 Integer Forcing vs Lattice Reduction aided Detection

Lattice reduction aided (LRA) detection [43] treats the channel matrix $\mathbf{H}$ as the lattice basis, and employs a linear transformation $\mathcal{T}$ to obtained a reduced basis $\mathbf{H}\mathcal{T}$. The received signal vector is hence expressed as $\mathbf{y} = (\mathbf{H}\mathcal{T})(\mathcal{T}^{-1}\mathbf{x}) + \mathbf{z}$. Thus, the effective channel and input signal become $\mathbf{H}\mathcal{T}$ and $\mathcal{T}^{-1}\mathbf{x}$ respectively. Clear, the effective channel $\mathbf{H}\mathcal{T}$ has better condition than $\mathbf{H}$, which means that by employing ZF or MMSE equalisation, the noise enhancements $(\mathbf{H}\mathcal{T})^{-1}$ are more evenly distributed across the data streams.

It can be observed that by regarding $\mathcal{T}^{-1}$ as the integer matrix $\mathbf{A}$, the LRA strategy seems similar to IF. Both aim to recover the linearly transformed $\mathbf{x}$, in order to prevent noise enhancement, and both achieve the optimal multiplexing-diversity tradeoff [39, 44]. However, there are three distinctions between them.

- LRA deals with symbols while IF deals with codewords, hence LRA can be treated as a special case of IF when lattice coding is not utilised.
- $\mathcal{T}$ is a unimodular matrix ($\det(\mathcal{T}) = \pm 1$) over $\mathbb{A}$, while the coefficient matrix $\mathbf{Q}_\pi$ in IF is a full rank matrix over a finite field $\mathbb{A}/\pi\mathbb{A}$. It has been proved that such difference results in IF outperforming the LRA for some channel realisations [39].
- LRA considers the system with $N_\mathrm{r} \geq L$ while there is no such constraint in IF.

A detailed comparison is avaliable in [39, 45, 46].

2.7 Beyond Baseline Lattice Network Coding

Based on the fundamental work in [28], the authors in [47], [29] and [35] propose practical code designs for C&F over $\mathbb{Z}$, $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ respectively. A reverse C&F scheme is proposed for the downlink scenario in [48, 49]. The authors in [50] show that with limited feedback, C&F can adaptively choose the best ring to match the instantaneous channel coefficients. C&F based on multilevel codes is proposed in [51, 52]. The main idea is that a large ring can be expressed as the direct sum of some co-prime finite fields, hence

the decoding complexity is significantly reduced. The successive version of C&F and IF are proposed in [53] and [54] respectively, where the previously decoded functions can be used as side information to generate a new effective channel.

Chapter 3

Low Complexity Coefficient Selection Algorithms for Compute-and-Forward

Contents

3.1 Introduction

Due to their very high density, the next generation of wireless communication systems will require enormous backhaul load to support the data transmission between the access points and the central hub station. In Chapter 2, we have reviewed that PNC has been proposed as a promising strategy to reduce the backhaul load. Among many PNC schemes, C&F has attracted the most interest. The core aspect which dominates the performance of C&F is the selection of the coefficient vectors. The process of selecting the optimal coefficient consists of two stages:

- local selection: each AP selects an integer vector to maximise its computation rate $\mathcal{R}_m(\mathbf{h}_m, \mathbf{a}_m)$ locally.

- global selection: in order to recover the data of all individual users without ambiguity, the vectors provided by the APs have to form a matrix whose rank is equal to the number of users.

Much work has been carried out in the last few years on both stages. For the local selection, the original paper of C&F [28] provides a bound for the coefficient vectors, and the optimal solution can be obtained by performing an exhaustive search within that boundary. The authors in [29] state that the coefficient selection issue is actually a shortest vector problem (SVP). Any lattice reduction algorithm, such as the Lenstra-Lenstra-Lovasz (LLL) algorithm [42], possibly followed by the Fincke-Pohst algorithm [55] can be utilised to acquire sub-optimal solutions. There are two main drawbacks of these lattice reduction algorithms: 1) the complexity increases very rapidly as the number of users increases; 2) it becomes less accurate for large numbers of users. Recently, a polynomially optimal algorithm proposed by Sahraei and Gastpar [6] significantly reduced the number of candidate vectors of [28]. It translated the optimisation problem from multiple variables to one variable. Based on the idea of [6], some improvements are proposed in [56, 57] to further reduce the complexity.

Unfortunately, the methods in [6, 56, 57] are suitable for real valued channels and integer lattices ($\mathbb{Z}$-lattice) only. Finding the optimal solution in polynomial time over complex integer based lattices is still an open problem. For the Gaussian integer based lattices, the sub-optimal lattice reduction based algorithms, such as the complex-LLL [58] and its extensions [59], [41] still work. However, they have the same drawbacks as in real channel scenarios. Recently, more focus is given to the Eisenstein integer based lattice: which has the densest packing strcuture in the 2-dimensional complex plane. A lattice reduction method over the $\mathbb{Z}[\omega]$-lattice is proposed in [35], though for a two way relay system only. An extended version of the algorithm in [6] for both $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ is proposed in [37], however it might sometimes miss the optimal solution. The latest research in [50] illustrated that C&F can operate over many algebraic number fields not restricted to PIDs. Unfortunately, efficient approaches for coefficient selection over these non-cubic lattices are not available in the existing literature.

For the second stage, the most commonly used approach to meet the requirement of unambiguous decodability is for each relay to forward more than one linear equation to the hub. The global optimal full rank matrix is selected by the hub and then fed back to the APs [60]. An alternative approach is that the integer vector provided by each relay is forced to include at least two users. This can significantly reduce the probability of rank deficiency [61].

In this Chapter, we focus on local selection which finds the optimal integer vector for a single AP. We consider both real and complex valued channels. For the real valued scenario, we propose two novel approaches. One aims to reduce the number of candidate vectors in the exhaustive search. The other finds the best scaling factor iteratively, and its corresponding integer vector can be acquired immediately. There are two main benefits of our proposed methods:

- In term of the computation rate: both of our algorithms outperform the LLL, and achieve the same rate compared to the exhaustive search.
- In term of the complexity: our algorithms have much lower complexities compared to the exhaustive search, and lower than the LLL except at very high SNR.

For the complex valued scenario, the main contributions are as follows:

- We propose a low polynomial complexity algorithm to ensure the optimal integer vector can be acquired for both $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ lattices. We also derive a theoretical upper bound on the complexity.
- We propose a suboptimal linear search algorithm for coefficient selection which has lower complexity. Compared to the optimal approach above, it aims to discard the 'unnecessary' candidates by employing a pre-defined step size which is related to the number of users and SNR. The theoretical complexity is also investigated.
- We evaluate numerically the performance and complexity of our two proposed algorithms, and compare them with existing approaches. Simulation results indicate that our proposed algorithms have better complexity-performance tradeoff.

- Our proposed algorithms can be easily extended to lattices over any other algebraic integers without additional complexity.

The remainder of this Chapter is organised as follows: we briefly review the existing selection algorithms as benchmarks in section 3.2. Then we propose novel selection approaches for the real valued case and the complex valued case in section 3.3 and 3.4 respectively. In section 3.5, we present the numerical results in terms of both achievable rate and complexity for different types of lattices.

3.2 Existing Coefficient Selection Algorithms

3.2.1 Exhaustive-I Algorithm

In the original paper of C&F [28], the authors state that the Euclidean norm of the optimal coefficient vector has an upper bound, expressed as $\|\mathbf{a}_{\text{opt}}\| \leq \Phi = \sqrt{1+\text{SNR}\|\mathbf{h}\|^2}$, hence an exhaustive search over all possible $\mathbf{a}$ within that range can be employed to obtain $\mathbf{a}_{\text{opt}}$. The time complexity of this algorithm is $\mathcal{O}(\Phi^{2L})$.

3.2.2 Exhaustive-II Algorithm (Real-valued only)

The authors in [6, 62] propose an exhaustive search algorithm with polynomial complexity. They state that it suffices to search over the integer vectors generated by $\lfloor \alpha \mathbf{h} \rceil$ only rather than considering all possible $\mathbf{a} \in \mathbb{Z}^L$. Therefore, the optimisation problem with an L-dimensional variable $\mathbf{a}$ is translated to an optimisation problem over the one-dimensional variable α. The candidate vectors can be obtained by dividing all possible $\alpha \in \mathbb{R}$ into several intervals, where each interval corresponds to a unique candidate $\mathbf{a}$. The time complexity of this algorithm is $\mathcal{O}(L\Phi\log(L\Phi))$. For convenience, we name this method 'Exhaustive-II' to distinguish it from the original exhaustive algorithm above.

3.2.3 Lattice Reduction Algorithm

In section 2.6, we have reviewed that the maximisation of the computation rate is equivalent to minimising $\mathbf{a}_m^H \mathbf{M} \mathbf{a}_m = \|\mathbf{L}^H \mathbf{a}\|^2$. The LLL and Complex-LLL lattice reduction algorithms are most commonly used for dealing with such SVPs in the $\mathbb{Z}$-lattice and $\mathbb{Z}[i]$-lattice respectively. However, these algorithms only ensure the selected vector is less than $2^{\frac{L-1}{2}}$ times the actual optimal solution. Hence, they become less accurate as the number of users increases.

3.2.4 Quantised Search

For $\mathbb{Z}[i]$-lattices, an intuitive approach for coefficient selection is to employ some quantised (sampled) values of α to generate the candidate set of $\mathbf{a}$, expressed as $\mathbf{a} = \mathcal{Q}_{\mathbb{Z}[i]}(\alpha \mathbf{h})$. The question is how to choose the quantiser. Since $\alpha \in \mathbb{C}$, the authors in [63] allocate step sizes for both the magnitude and the phase of α. Clearly, this method is equivalent to the exhaustive search when both of the step sizes tend to zero. However, zero step size is definitely infeasible in practice. The core aspect of this quantised algorithm is the choice of the step size, which is not analysed in [63].

The method described above leads to an oversampling for small magnitudes and undersampling for large magnitudes. In section 3.4, we will propose a uniform quantiser and describe how to choose the optimal step size.

3.2.5 L-L Algorithm

Very recently, Liu and Ling proposed an efficient algorithm (denoted here as the L-L algorithm) for complex valued channels in [37]. The authors adapt the idea in [6] directly for the complex integer based lattices. However, the algorithm in [37] might miss the optimal solution for some channel realisations, though with a very small probability. A detailed discussion of this approach will be presented in section 3.4.

3.3 Proposed Algorithms for Real Valued Channels

3.3.1 Real Valued Channel Model

We consider a general multiple access system with L users and one AP in this Chapter. In real valued channels, the original message $\mathbf{w}_l \in \mathbb{F}_p^k \cong (\mathbb{Z}/p\mathbb{Z})^k$ and the length n codeword $\mathbf{x}_l \in \mathbb{R}^n$ rather than $\mathbb{C}^n$. The noise vector $\mathbf{z} \in \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, and the channel vector $\mathbf{h} = [h_1, h_2, \cdots, h_L]^T$ with $h_l \in \mathcal{N}(0,1)$. Note that the computation rate in the complex valued channel is measured in bits per complex dimension. For the real channel case, we use bits per real dimension instead, and hence the expressions (2.38) and (2.41) are rewritten as

$$\mathcal{R}(\mathbf{h}, \mathbf{a}) = \max_{\alpha \in \mathbb{R}} \frac{1}{2} \log^+ \left(\frac{P}{\alpha^2\sigma^2 + \|\alpha\mathbf{h} - \mathbf{a}\|^2} \right) \tag{3.1}$$

and

$$\mathcal{R}(\mathbf{h}, \mathbf{a}) = \frac{1}{2} \log^+ \left(\frac{P}{\mathbf{a}^T\mathbf{M}\mathbf{a}} \right) \tag{3.2}$$

respectively, with $\mathbf{M} = \mathbf{I}_L - \frac{\mathrm{SNR}}{\mathrm{SNR}\|\mathbf{h}\|^2+1}\mathbf{h}\mathbf{h}^T$. Note that the scaling factor $\alpha \in \mathbb{R}$ and integer vector $\mathbf{a} = \mathbb{Z}^{L\times 1}$.

3.3.2 Reduced Candidate Set aided Algorithm

We have introduced the exhaustive-II algorithm which employs fewer candidates $\mathbf{a}$ compared to the exhaustive-I. In this section, we propose an algorithm to further reduce the number of candidates. The reduction comes from three aspects:

- A tighter upper bound of $\|\mathbf{a}\|$ can be acquired according to the bound of $|\alpha|$.
- Set a threshold to distinguish the 'unnecessary candidates', and these candidates can be discarded without loss of performance. The threshold can be drawn from an off-line table.
- Set a break condition for the on-line search.

1st Stage: Obtain the Candidate Set

To the best of our knowledge, the existing exhaustive search algorithms for C&F are all based on the following bound:

$$|a_l| \leq \|\mathbf{a}\| < \sqrt{1+\mathrm{SNR}\|\mathbf{h}\|^2},\ \forall l, \tag{3.3}$$

that is, all elements in $\mathbf{a}$ share the same bound. Actually, it is more straightforward to consider the bound of α rather than $\mathbf{a}$. The expression of $\mathcal{R}(\mathbf{h},\mathbf{a})$ in (3.1) can be equivalently rewritten as a function of α, which is

$$\mathcal{R}(\mathbf{h},\alpha) = \max_{\mathbf{a}\in\mathbb{Z}^L} \frac{1}{2}\log^+\Big(\frac{P}{\alpha^2\sigma^2+\|\alpha\mathbf{h}-\mathbf{a}\|^2}\Big). \tag{3.4}$$

For a given α, maximising $\mathcal{R}(\mathbf{h},\alpha)$ in (3.4) is equivalent to minimising $\|\alpha\mathbf{h}-\mathbf{a}\|^2$. Clearly, the MMSE solution of $\mathbf{a}$ is given by

$$\mathbf{a}_{\mathrm{MMSE}} = \lfloor \alpha\mathbf{h} \rceil, \tag{3.5}$$

where the rounding operation is applied to each component of $\alpha\mathbf{h}$. Thus, we have

$$\begin{aligned}\mathcal{R}(\mathbf{h},\alpha) &= \frac{1}{2}\log^+\Big(\frac{P}{\alpha^2\sigma^2+\|\alpha\mathbf{h}-\lfloor\alpha\mathbf{h}\rceil\|^2}\Big)\\ &\leq \frac{1}{2}\log^+\Big(\frac{P}{\alpha^2\sigma^2}\Big),\end{aligned} \tag{3.6}$$

where the equality holds if and only if the selected integer vector is a perfect approximation of the scaled channel. Clearly, the computation rate is zero when $\alpha \geq \sqrt{P/\sigma^2} = \sqrt{\mathrm{SNR}}$. By substituting $\lfloor\alpha\mathbf{h}\rceil$ for $\mathbf{a}$, the effective noise in (2.39) can be rewritten as

$$\sigma^2_{\mathrm{eff}}(\alpha) = \alpha^2\sigma^2 + P\|\alpha\mathbf{h}-\lfloor\alpha\mathbf{h}\rceil\|^2 \tag{3.7}$$

which is an even function of α, and hence it suffices to consider $\alpha > 0$. The bound of α is therefore expressed as

$$0 < \alpha < \sqrt{\mathrm{SNR}}, \tag{3.8}$$

and the bound of each element of $\mathbf{a}$ can be obtained immediately, expressed as:

$$a_l = \lfloor \alpha h_l \rceil \in \begin{cases} \left(0, \lceil \sqrt{\mathrm{SNR}} h_l \rceil\right] \cap \mathbb{Z} & \text{if } h_l \geq 0 \\ \left[\lfloor \sqrt{\mathrm{SNR}} h_l \rfloor, 0\right) \cap \mathbb{Z} & \text{if } h_l < 0 \end{cases} \tag{3.9}$$

Clearly, (3.9) gives a tighter bound than (3.3). For example, assume $\mathbf{h} = [-0.5, 1]$ and SNR=100. By employing (3.3), we have $a_1, a_2 \in \{[-11, 11] \cap \mathbb{Z}\}$. By employing (3.9), we have $a_1 \in \{(0, 10] \cap \mathbb{Z}\}$ and $a_2 \in \{[-5, 0) \cap \mathbb{Z}\}$. Now we address the detailed procedure of generating a complete candidate set for $\mathbf{a}$:

1. Given the channel vector $\mathbf{h} = [h_1, h_2, \cdots, h_L]^T$ and SNR, find all possible integers for each a_l by employing (3.9). We use $\Phi_l = \{a^*_{l,1}, a^*_{l,2}, \cdots, a^*_{l,K_l}\}$ to denote the set of all possible integers a_l, with cardinality K_l.

2. For each Φ_l, establish a set for α, denoted as $\mathcal{S}_l = \{\frac{a^*_{l,1}-0.5}{h_l}, \frac{a^*_{l,2}-0.5}{h_l}, \cdots, \frac{a^*_{l,K_l}-0.5}{h_l}\} = \{\alpha^*_{l,1}, \alpha^*_{l,2}, \cdots, \alpha^*_{l,K_l}\}$. The elements of $\mathcal{S}_l$ represent the discontinuities of $f(\alpha) = \lfloor \alpha h_l \rceil$.

3. Form a new set $\mathcal{S} = \cup_{l=1}^{L} \mathcal{S}_l$, and sort the elements of $\mathcal{S}$ in ascending order. Let $\bar{K}$ denote the the cardinality of $\mathcal{S}$, clearly $\bar{K} = \sum_{l=1}^{L} K_l$.

4. The elements in $\mathcal{S}$ are the discontinuities of $f(\alpha) = \lfloor \alpha \mathbf{h} \rceil$. In other words, the range $\alpha \in (0, \sqrt{\mathrm{SNR}})$ is partitioned into $\bar{K} + 1$ intervals by these cut-off points, and each interval corresponds to a unique $\mathbf{a}$. Note that the very left interval $(0, \alpha_1)$ should be discarded as its corresponding $\mathbf{a} = \lfloor \alpha \mathbf{h} \rceil$ is an all zero vector. Hence, each $\alpha^*_{\bar{k}} \in \mathcal{S}$ can be utilised as the representative of its right interval $[\alpha^*_{\bar{k}}, \alpha^*_{\bar{k}+1})$ to generate the corresponding $\mathbf{a}$, and this requires

$$\lfloor \alpha^*_{\bar{k}} \mathbf{h} \rceil = \lfloor \alpha \mathbf{h} \rceil, \;\; \alpha \in [\alpha^*_{\bar{k}}, \alpha^*_{\bar{k}+1}). \tag{3.10}$$

   Since $\alpha^*_{\bar{k}}$ is a discontinuity, there exists h_l that $\alpha^*_{\bar{k}} h_l$ is precisely a half integer. In order to cancel the uncertainty of $\lfloor \alpha^*_{\bar{k}} \mathbf{h} \rceil$ and make (3.10) valid for all intervals, we define a round operation $\lfloor \cdot \rceil_\infty$

$$\lfloor x \rceil_\infty = \begin{cases} \mathrm{sgn}(x) \lceil x \rceil & \text{if } x \in \mathbb{Z} - \frac{1}{2} \\ \lfloor x \rceil & \text{else} \end{cases} \tag{3.11}$$

which rounds the half integers away from zero (towards infinity). The candidate set can be obtained immediately, denoted as

$$\mathcal{I} = \{\mathbf{a} : \mathbf{a} = \lfloor \alpha_{\tilde{k}}^{*} \mathbf{h} \rceil_{\infty},\ \forall \alpha_{\tilde{k}}^{*} \in \mathcal{S}\} \tag{3.12}$$

2nd stage: Discard Unnecessary Candidates

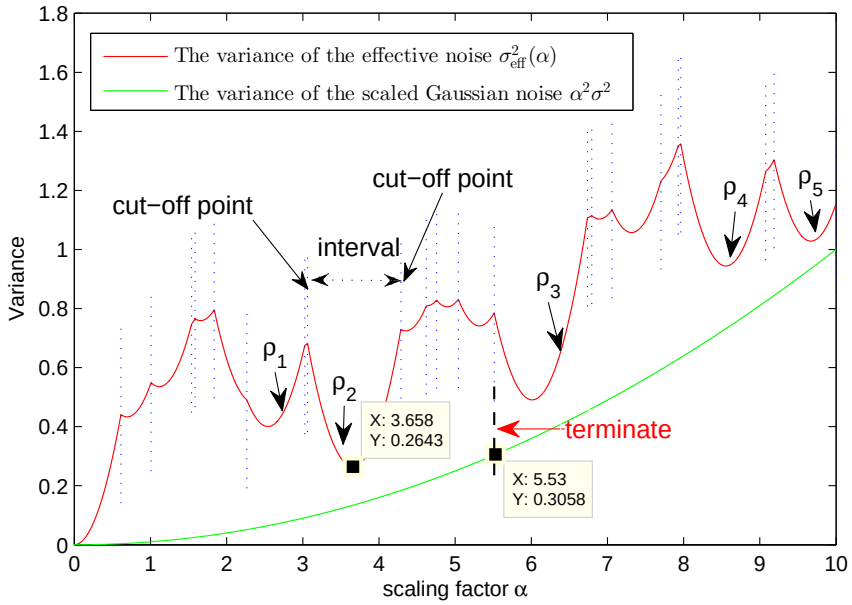

Figure 3.1: Sketch diagram of the reduced set aided algorithm

In the first stage, we have obtained the complete candidate set $\mathcal{I}$. In this stage, we select a subset of $\mathcal{I}$ which is referred to as the 'necessary candidate set'. We take a simple example to present how to select 'necessary candidates'. Assume a 5-user system with $\mathbf{h} = [-0.3246, -0.8164, -0.3153, 0.2208, 0.4958]^T$ and $\text{SNR} = 20\text{dB}$. As shown in Figure 3.1, the red line represents the variance of the effective noise calculated by (3.7). The range of α is divided into 21 intervals with 21 cut-off points which are illustrated by the blue dotted lines. The term $\lfloor \alpha \mathbf{h} \rceil_{\infty}$ is invariant within each interval of α, which means that the 1st and 2nd order derivative of σ_{eff}^2 can be expressed as:

$$\frac{d\sigma_{\text{eff}}^2}{d\alpha} = 2\alpha\sigma^2 + 2\alpha\|\mathbf{h}\|^2 - 2\mathbf{h}^T \lfloor \alpha \mathbf{h} \rceil \tag{3.13}$$

and

$$\frac{d^2\sigma_{\text{eff}}^2}{d\alpha^2} = 2\sigma^2 + 2\|\mathbf{h}\|^2 \geq 0 \tag{3.14}$$

respectively. Clearly, within each interval, σ^2_{eff} is a convex function with a local minimum, as shown in Figure 3.1. At first glance, the global minimum is more likely to be located in one of the largest parabolas, labelled with $\rho_{1\sim5}$, and it therefore seems possible to consider the 'wider' intervals only rather than all of them.

Let $d_{\bar{k}}$ denote the length of the $\bar{k}$th valid interval, and d_{opt} be the length of the interval which includes the global minimum. We define $d_{\max} = \max(d_1, d_2, \cdots, d_{\bar{K}})$ as the widest length among all intervals. We define $\gamma_{\text{opt}} = \frac{d_{\text{opt}}}{d_{\max}}$ and $\gamma_{\bar{k}} = \frac{d_{\bar{k}}}{d_{\max}}$ as the normalised d_{opt} and $d_{\bar{k}}$ respectively. Clearly, we have $0 < \gamma_{\text{opt}}, \gamma_{\bar{k}} \leq 1$. Our target is to find a threshold of γ_{opt}, denoted as γ_{thre} to distinguish the 'wider' intervals whose $\gamma_{\bar{k}} \geq \gamma_{\text{thre}}$. The threshold can be determined by investigating the distribution of γ_{opt}.

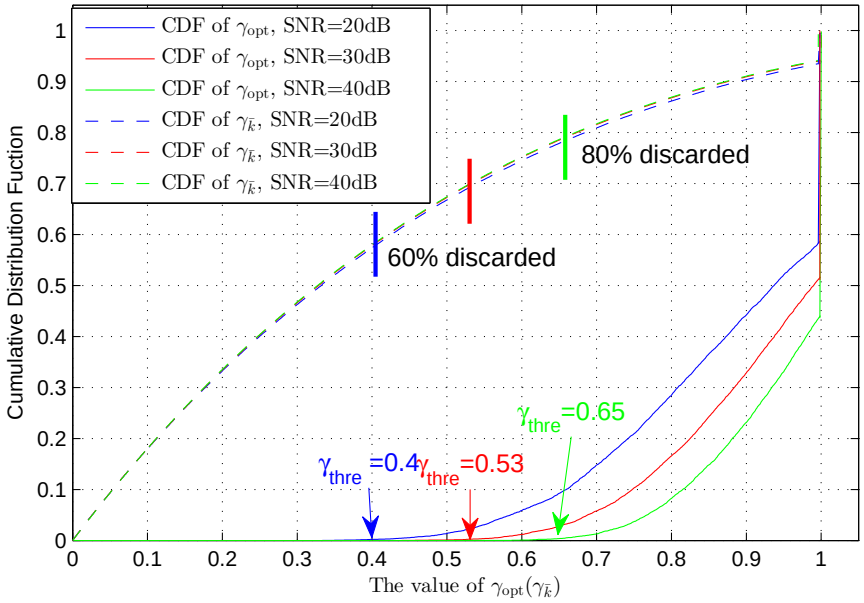

Figure 3.2: The cumulative distribution of γ_{opt} and $\gamma_{\bar{k}}$

The cumulative distribution curves in Figure 3.2 are acquired over 1000 channel realisations, the solid lines illustrate the CDF of γ_{opt} in a 5-user system. The blue, red and green lines represent the scenarios with SNR=20dB, 30dB and 40dB, respectively. It can be seen that $\gamma_{\text{opt}} \geq 0.4$ for all channel trials when SNR = 20dB. Therefore, we can treat 0.4 as the threshold of γ_{opt} to define the 'wider' interval. The corresponding γ_{thre} at SNR = 30dB and SNR = 40dB can be determined in a same way, as labelled with arrows. It can be observed that γ_{thre} increases monotonically with SNR, for convenience, the value of γ_{thre} for SNR = 20dB can be used in the region of 20dB$\leq$SNR$<$30dB, this modification does not affect the accuracy.

The value of γ_{thre} is only related to the number of users L and SNR, it does not depend on the instantaneous channel realisation. Hence, we can establish a table Θ to store the values of γ_{thre} corresponding to each L and SNR region, and this table can be obtained off-line, and therefore does not bring extra complexity to the on-line search. Table 3.1 demonstrates a part of such a table.

Table 3.1: A partial table of Θ with $L = 5, 8, 10$ (real valued case)

L \ SNR (dB)	<5	[5 10)	[10 15)	[15 20)	[20 25)	[25 30)	[30 35)	[35 40)	$\cdots$
5	0.05	0.07	0.15	0.32	0.4	0.48	0.53	0.58	$\cdots$
8	0.04	0.04	0.05	0.12	0.2	0.31	0.42	0.46	$\cdots$
10	0.03	0.03	0.04	0.05	0.15	0.22	0.35	0.39	$\cdots$

Algorithm 1 Reduced set aided algorithm

Output: the optimal coefficient vector $\mathbf{a}_{\text{opt}}$

Offline Search: obtain table Θ

For given L and SNR
for $trial = 1 : 1000$ **do**
  generate $\mathbf{h}_{trial}$ ($h_l \in \mathcal{N}(0,1)$) and obtain the corresponding candidate set $\mathcal{I}$.
  obtain $\mathbf{a}_{\text{opt},trial} = \text{argmax}_{\mathbf{a}\in\mathcal{I}} \mathcal{R}(\mathbf{h}_{trial}, \mathbf{a})$ based on (3.2).
  calculate the normalised width γ_{opt} for $\mathbf{a}_{\text{opt}}$.
end for
set $\min_{trial=1}^{1000} \gamma_{\text{opt},trial} \rightarrow \gamma_{\text{thre}}(L, \text{SNR})$

Online Search: obtain $\mathbf{a}_{\text{opt}}$ for instantaneous channel $\mathbf{h}$

select γ_{thre} according to Θ, and obtain the candidate set $\mathcal{I}$ for $\mathbf{h}$.
for all $\mathbf{a} \in \mathcal{I}$ **do**
  find $\gamma_{\bar{k}} \geq \gamma_{\text{thre}}$ and $\widetilde{\mathcal{S}} = \{\widetilde{\alpha}_1^*, \widetilde{\alpha}_2^*, \cdots, \widetilde{\alpha}_J^*\}$.
end for
$j = 2$, $\mathbf{a}_{\text{opt}} = \lfloor \widetilde{\alpha}_1^* \mathbf{h} \rceil_\infty$, $\sigma_{\text{opt}}^2 = \mathbf{a}_{\text{opt}}^T \mathbf{M} \mathbf{a}_{\text{opt}}$ (3.2).
while $(\widetilde{\alpha}_j^* \sigma)^2 \leq \sigma_{\text{opt}}^2$ and $j \leq J$ **do**
  $\mathbf{a}_j = \lfloor \widetilde{\alpha}_j^* \mathbf{h} \rceil_\infty$, $\sigma_{\text{eff}}^2 = \mathbf{a}_j^T \mathbf{M} \mathbf{a}_j$, $\sigma_{\text{opt}}^2 = \min(\sigma_{\text{opt}}^2, \sigma_{\text{eff}}^2)$, $\mathbf{a}_{\text{opt}} = \underset{\mathbf{a}\in\{\mathbf{a}_{\text{opt}}, \mathbf{a}_j\}}{\text{argmin}} \mathbf{a}^T \mathbf{M} \mathbf{a}$.
  $j = j + 1$
end while
Return $\mathbf{a}_{\text{opt}}$

From Figure 3.2, we can also estimate the complexity reduction resulting from this stage. The dashed lines represent the CDF of $\gamma_{\bar{k}}$ in the cases of SNR = 20dB, 30dB and 40dB. Since the wider intervals are selected according to $\gamma_{\bar{k}} \geq \gamma_{\text{thre}}$, therefore, about 60% of the intervals are discarded ($\gamma_{\bar{k}} < 0.4$) when SNR=20dB, and the complexity is

reduced by up to 80% at high SNR (SNR≥40dB). It can be observed that the CDF of $\gamma_{\bar{k}}$ is invariant for different values of SNR, which implies that the intervals of α are randomly uniformly distributed within the valid range.

3rd Stage: On-line Search

In this stage, we will find $\mathbf{a}_{\text{opt}}$ based on the 'wider' intervals acquired in the second stage. Let $\widetilde{\mathcal{S}} = \{\widetilde{\alpha}_1^*, \widetilde{\alpha}_2^*, \cdots, \widetilde{\alpha}_J^*\}$ denote the cut-off points of the 'wider' intervals, with $\widetilde{\mathcal{S}} \subseteq \mathcal{S}$. As the values of $\widetilde{\alpha}_j^*$ have an ascending order, we can set a break condition for the searching procedure. When the scaled Gaussian noise of the current interval is already larger than the minimum effective noise of the previous vectors, we terminate the procedure.
As shown in Figure 3.2, the green line represents the variance of the scaled Gaussian noise $\sigma_{\text{sg}}^2 = \alpha^2\sigma^2$. The selected 'wider' intervals are $\rho_1 \sim \rho_5$, and $\widetilde{\alpha}_1^* \sim \widetilde{\alpha}_5^*$ are their corresponding left endpoints. We terminate at ρ_3 since it is impossible to find any $\lfloor \alpha\mathbf{h} \rceil_\infty$ (with $\alpha \geq \widetilde{\alpha}_3^* = 5.53$) whose effective noise is smaller than the local minima of ρ_2 which is 0.2643, even with a zero self-noise. The whole procedure is outlined in algorithm 1.

3.3.3 Linear Search Algorithm for Real Lattices

In this section, we propose a conceptually simple method which employs a linear search over values of α using a fixed step size Δ for α, calculating for each value the corresponding $\sigma_{\text{eff}}^2(\alpha)$, and selects the one which minimises $\sigma_{\text{eff}}^2(\alpha)$. Compared to the quanised search algorithm in [63], we perform the following processes to make it more efficient.

- we set an initial value of α as $\frac{0.5+\xi}{\mathbf{max}(|h_1|,|h_2|,\cdots,|h_L|)}$, where the symbol ξ is an arbitrarily small number in order to ensure $\lfloor \alpha\mathbf{h} \rceil$ is not an all zero vector.
- we employ the same break condition used in the previous section to reduce the 'valid range' of α.

Compared to the reduced set aided algorithm, this method employs sampled values of α rather than determining the intervals based on $\mathbf{h}$, and without calculating the MMSE

Algorithm 2 Linear Search Algorithm for Real Lattices

Input: channel vector $\mathbf{h}$, step size Δ, thermal noise σ^2, ξ
Output: the optimal coefficient vector $\mathbf{a}_{\text{opt}}$
1: Initialise: $\alpha = \frac{0.5+\xi}{\max(|h_1|,|h_2|,\cdots,|h_L|)}$, $\sigma^2_{\text{opt}} = \sigma^2_{\text{eff}}(\alpha)$. Equation (3.7)
2: then $\alpha_{\text{sample}} = \alpha + \Delta$, $\sigma^2_{\text{sg}} = \alpha^2_{\text{sample}}\sigma^2$.
3: **while** $\sigma^2_{\text{opt}} > \sigma^2_{\text{sg}}$ and $\alpha_{\text{sample}} \leq \sqrt{\text{SNR}}$ **do**
4: **if** $\sigma^2_{\text{eff}}(\alpha_{\text{sample}}) < \sigma^2_{\text{opt}}$ **then**
5: $\alpha = \alpha_{\text{sample}}$, $\sigma^2_{\text{opt}} = \sigma^2_{\text{eff}}(\alpha_{\text{sample}})$
6: **end if**
7: $\alpha_{\text{sample}} = \alpha_{\text{sample}} + \Delta$, $\sigma^2_{\text{sg}} = \alpha^2_{\text{sample}}\sigma^2$
8: **end while**
9: Return $\mathbf{a}_{\text{opt}} = \lfloor \alpha \mathbf{h} \rceil$

solution for each candidate $\mathbf{a}$. The performance of this algorithm greatly depends on the choice of the step size. Theoretically, the step size should be small enough to 'visit' each of the 'wider' intervals. Since we use sampled values of α rather than α_{MMSE} to calculate $\sigma^2_{\text{eff}}(\alpha)$ during each comparison (line 4 in algorithm 2), it may sometimes miss the global minimum. The worst case is that α_{MMSE} is precisely the midpoint of two consecutive sampled points and $\lfloor \alpha \mathbf{h} \rceil$ takes the same value for both. Since the second derivative of $\sigma^2_{\text{eff}}(\alpha)$ is a constant within a specific interval (3.14), the worst case error can be upper bounded by

$$e_{\text{ub}} = \sigma^2_{\text{eff}}(\alpha_{\text{MMSE}} + \frac{\Delta}{2}) - \sigma^2_{\text{eff}}(\alpha_{\text{MMSE}}) = \frac{\Delta^2}{4}(\sigma^2 + \|\mathbf{h}\|^2). \tag{3.15}$$

Note that this method is mainly motivated by applications which employ complex lattices over $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$. For complex lattices, α is complex, and hence its intervals become a set of Voronoi regions over a complex plane and the problem becomes much more complicated. This will be investigated in the following section.

3.4 Proposed Algorithms for Complex Valued Channels

In this section, we propose two algorithms for the complex valued scenario. One extends the exhaustive-II algorithm in [6] to complex lattices, which ensures the optimal solution. The other employs linear search with optimised step sizes to obtain the necessary candidates.

3.4.1 Exhaustive-II in Complex Valued Channel

Since the Exhaustive-II algorithm selects $\mathbf{a}_{\text{opt}}$ with low polynomial complexity in the real channel case, it is worthwhile to investigate the feasibility of Exhaustive-II in the complex case. This section comprises three parts: we first propose the complex exhaustive-II algorithm, then gives the complexity analysis, and finally a comparison with the L-L method is provided.

Complex Exhaustive-II Algorithm

Previously we have discussed that the computation rate can be expressed as a function of α. For the complex scenario, we rewrite (3.6) as

$$\mathcal{R}(\mathbf{h},\alpha)=\log^{+}\Big(\frac{P}{\alpha^2\sigma^2+P\|\alpha\mathbf{h}-\mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h})\|^2}\Big), \tag{3.16}$$

where $\alpha\in\mathbb{C}$. Again, the quantisation $\mathcal{Q}_{\mathbb{A}}(\cdot)$ is applied to each αh_l. Let α_{opt} denote the globally optimal α which maximises $\mathcal{R}(\mathbf{h},\alpha)$ in (3.16). Now we provide the valid range of α_{opt}.

Proposition 1. The amplitude of α_{opt} is upper bounded by $\sqrt{\text{SNR}}$, and it suffices to restrict the phases of α to $0\sim\frac{\pi}{2}$ and $0\sim\frac{\pi}{3}$ for $\mathbb{Z}[i]$-lattice and $\mathbb{Z}[\omega]$-lattice respectively.

Proof. The proof of $|\alpha_{\text{opt}}|<\sqrt{\text{SNR}}$ is omitted here: see (3.4)$\sim$(3.6).
Define u as a unit in $\mathbb{A}$: we have:

$$\begin{aligned}\mathcal{R}(\mathbf{h},\alpha)&=\log^{+}\Big(\frac{P}{\alpha^2\sigma^2+P\|\alpha\mathbf{h}-\mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h})\|^2}\Big)\\&=\log^{+}\Big(\frac{P}{(u\alpha)^2\sigma^2+P\|u\alpha\mathbf{h}-\mathcal{Q}_{\mathbb{A}}(u\alpha\mathbf{h})\|^2}\Big)\\&=\mathcal{R}(\mathbf{h},u\alpha)\end{aligned} \tag{3.17}$$

Hence the complex plane of α is divided into several 'equivalent regions' due to the existence of units. As the corresponding units in $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ are $\{\pm1,\pm i\}$ and $\{\pm1,\pm\omega,\pm\omega^2\}$ respectively, hence the number of equivalent regions in $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ are 4 and 6 respectively, and hence it suffices to restrict the phase within $0\sim\frac{2\pi}{4}$ and

$0 \sim \frac{2\pi}{6}$ respectively. □

Recall that for the real valued case: the range of $\alpha \in \mathbb{R}$ is divided into several intervals. The quantised value $\lfloor \alpha \mathbf{h} \rceil$ is invariant within each interval. Hence each interval corresponds to a unique candidate vector $\mathbf{a}$, and the candidates can be acquired by choosing a representative of α for each interval.

For the complex case, $\alpha \in \mathbb{C}$, and hence the intervals of α become a set of Voronoi regions over a complex plane. We use $h_{\max}$ to denote the channel coefficient with the largest amplitude in $\mathbf{h}$. Let $\mathcal{V}_0$ denote the fundamental region of $\mathbb{A}$, and $\mathcal{A}_{\mathcal{V}_0}$ be the area of $\mathcal{V}_0$. Define a lattice as

$$\Lambda_l \triangleq \{\lambda : \lambda = \frac{a_l}{h_l}, a_l \in \mathbb{A}\}. \tag{3.18}$$

Clearly, the generator matrix of Λ_l is $\frac{1}{h_l}\mathbf{B}_{\mathbb{A}}$, where $\mathbf{B}_{\mathbb{A}}$ is the generator matrix of $\mathbb{A}$. For example, $\mathbf{B}_{\mathbb{Z}[i]} = [1\ 0; 0\ 1]$ and $\mathbf{B}_{\mathbb{Z}[\omega]} = [1\ 0; 0\ \omega]$. According to (3.18), the Voronoi region of a specific lattice point $\frac{a_l}{h_l} \in \Lambda_l$, denoted as $\mathcal{V}_{l,a_l}$, can be expressed as

$$\mathcal{V}_{l,a_l} = \{\alpha : \alpha \in \mathbb{C},\ \mathcal{Q}_{\mathbb{A}}(\alpha h_l) = a_l\}, \tag{3.19}$$

which means all values of $\alpha \in \mathcal{V}_{l,a_l}$ result in the same quantised integer a_l. Due to the lattice property, the Voronoi regions of $\mathcal{V}_{l,a_l}$ are congruent for all $a_l \in \mathbb{A}$. Hence we use $\mathcal{A}_{\mathcal{V}_l}$ to denote the area of $\mathcal{V}_{l,a_l},\ \forall a_l$. Since $\mathcal{A}_{\mathcal{V}_0}$ and $\mathcal{A}_{\mathcal{V}_l}$ can be calculated by $\det(\mathbf{B}_{\mathbb{A}})$ and $\det(\frac{1}{h_l}\mathbf{B}_{\mathbb{A}})$ respectively, clearly we have

$$\mathcal{A}_{\mathcal{V}_l} = \mathcal{A}_{\mathcal{V}_0}/|h_l|^2, \tag{3.20}$$

and hence we have the following results:

Proposition 2. The valid range of α is divided into several convex polygonal regions $\mathcal{V}_{\mathbf{a}}$, and each region corresponds to a unique candidate vector $\mathbf{a}$. The area of each region is upper bounded by $\mathcal{A}_{\mathcal{V}_0}/|h_{\max}|^2$.

Proof. Let $\mathcal{V}_{\mathbf{a}}$ denote the Voronoi region of α for given $\mathbf{a}$, expressed as

$$\begin{aligned}\mathcal{V}_{\mathbf{a}} &= \{\alpha : \mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h}) = \mathbf{a}\} \\ &= \{\alpha : \bigcap_{l=1}^{L} \mathcal{Q}_{\mathbb{A}}(\alpha h_l) = a_l\}.\end{aligned} \tag{3.21}$$

It can be obtained from (3.19) and (3.21) that $\mathcal{V}_{\mathbf{a}}$ is the intersection region of $\mathcal{V}_{l,a_l}$ for all l, denoted as $\mathcal{V}_{\mathbf{a}} = \cap_l \mathcal{V}_{l,a_l}$. It is known that the intersection of convex sets is also convex. Since each individual $\mathcal{V}_{l,a_l}$ is a convex polygon (square for $\mathbb{Z}[i]$ and hexagon for $\mathbb{Z}[\omega]$), hence $\mathcal{V}_{\mathbf{a}}$ is also a convex polygon. As an intersection region, the area of $\mathcal{V}_{\mathbf{a}}$ can be upper bounded by the smallest size among $\mathcal{V}_{l,a_l},\ \forall l$ which is

$$\min_l \mathcal{A}_{\mathcal{V}_l} = \min_l \frac{\mathcal{A}_{\mathcal{V}_0}}{|h_l|^2} = \frac{\mathcal{A}_{\mathcal{V}_0}}{|h_{\max}|^2}. \tag{3.22}$$

□

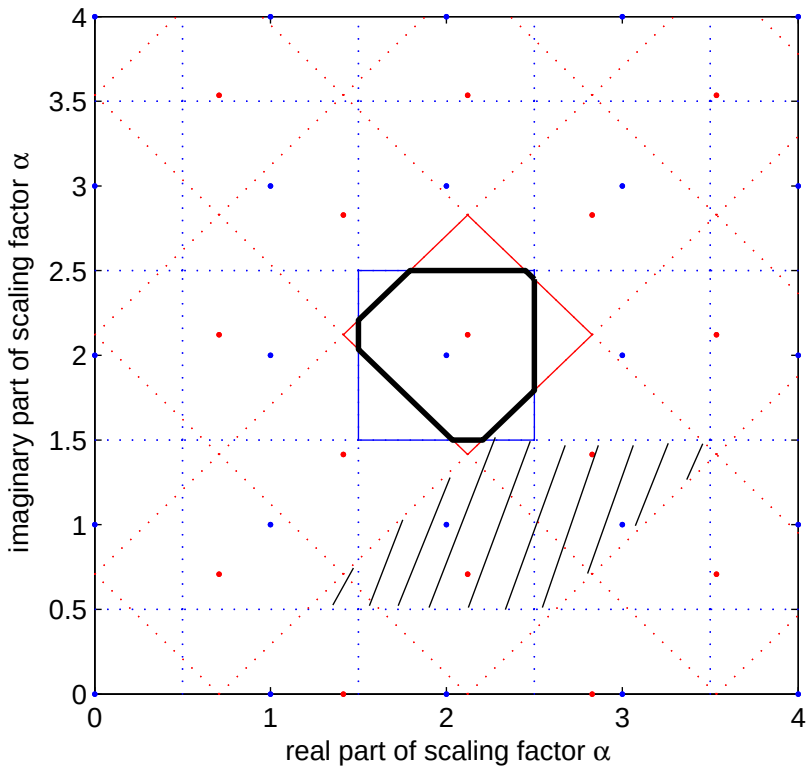

Figure 3.3: An example with $\mathbb{Z}[i]$-lattices: $\mathbf{h} = [1, \frac{1+i}{\sqrt{2}}]^T, \mathbf{a} = [2+2i, 3i]^T$

We take a simple example to interpret the above proposition. We consider a 2 user system employing the $\mathbb{Z}[i]$-lattice, with $\mathbf{h} = [1,\ \frac{1+i}{\sqrt{2}}]^T$. As shown in Figure 3.3, the real

and imaginary parts of α are represented by the x-axis and y-axis respectively. The lattice points in $\Lambda_{l=1}$ and $\Lambda_{l=2}$ are represented by the red and blue dots respectively, and each red (blue) square corresponds to a unique a_1 (a_2) respectively. For example, the red (blue) solid square in the centre corresponds to $\mathcal{V}_{1,a_1=2+2i}$ and $\mathcal{V}_{2,a_2=3i}$ respectively. Hence in order to acquire $\mathcal{Q}_{\mathbb{Z}[i]}(\alpha\mathbf{h}) = \mathbf{a} = [2+2i, 3i]^T$, the value of α has to be chosen within the region $\mathcal{V}_{\mathbf{a}}$ which is the black octagon in the centre.

Let $\widetilde{\Phi}$ denote the range of α_{opt} acquired in Proposition 1. Since $\mathbf{a}_{\text{opt}} = \mathcal{Q}_{\mathbb{A}}(\alpha_{\text{opt}}\mathbf{h})$, thus $\mathbf{a}_{\text{opt}}$ must be in the set

$$\mathbf{a}_{\text{opt}} \in \mathcal{I} = \{\mathbf{a} : \mathbf{a} = \mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h}), \alpha \in \widetilde{\Phi}\}. \tag{3.23}$$

According to Proposition 2, the candidate set $\mathcal{I}$ can be acquired by the following steps:

1. Select at least one representative of α within each $\mathcal{V}_{\mathbf{a}}$. The set of these representatives is denoted as $\mathcal{S}$.
2. Use $\mathcal{S}$ to generate the candidate set $\mathcal{I}$.

Clearly, the main challenge is to select the representatives of α. For the real valued scenario, each $\mathcal{V}_{\mathbf{a}}$ is a 1-dimensional interval which can be simply represented by its endpoint (discontinuity of $f(\alpha) = \lfloor \alpha\mathbf{h} \rceil$). However, in the complex case, each $\mathcal{V}_{\mathbf{a}}$ is a polygon, and the discontinuities become the edges of $\mathcal{V}_{\mathbf{a}}$. Hence the number of discontinuities becomes infinite. The vertices of each $\mathcal{V}_{\mathbf{a}}$ are the most easily calculated points among all discontinuities: can we therefore use these vertices as the representatives?

Assume α_{v} is a vertex of $\mathcal{V}_{\mathbf{a}}$. Clearly, α_{v} is shared by its adjacent polygons, which means that $\alpha_{\text{v}}\mathbf{h}$ is singular to the quantisation operation $\mathcal{Q}_{\mathbb{A}}(\alpha_{\text{v}}\mathbf{h})$ (due to the fact that some components of $\mathbf{a}_{\text{v}}$ are precisely half integers). The singular quantisation is not a problem for the real case. Since each $\mathcal{V}_{\mathbf{a}}$ has two ends, and hence if $\mathcal{Q}_{\mathbb{Z}}(\alpha\mathbf{h})$ is open at one end, then it has to be closed at the other end as long as $\mathcal{Q}_{\mathbb{Z}}$ rounds $\alpha\mathbf{h}$ in the same direction at both ends (see Eq 3.11). This is because each interval has redundancy (2 ends) to compensate the quantisation uncertainty (2 possibilities: round up or down), and they are balanced for all intervals. However, for the complex channel, the redundancy and the quantisation uncertainty are not always balanced. Take the $\mathbb{Z}[i]$-lattice for example:

Algorithm 3 Complex-Exhaustive-II Algorithm

Input: channel vector $\mathbf{h} = [h_1, h_2, \cdots, h_L] \in \mathbb{C}^L$, SNR, integer domain $\mathbb{A}$ ($\mathbb{Z}[i]$,$\mathbb{Z}[\omega]$, etc) with basis $\mathbf{B}_{\mathbb{A}}$

Output: optimal coefficient vector $\mathbf{a}_{\text{opt}}$

$\mathbf{Phase\ 1}$: obtain the representatives of α, stored in set $\mathcal{S}$. The initial $\mathcal{S} = \emptyset$

According to Proposition. 1, calculate the range of α: $\widetilde{\Phi}$.
for $l = 1 : L$ **do**
  Generator Λ_l according to (3.18). Find $\{\lambda : \lambda \in \Lambda_l \cap \widetilde{\Phi}\}$, and store these λ into $\Omega_l = \{\alpha^*_{l,1}, \alpha^*_{l,2}, \cdots, \alpha^*_{l,K_l}\}$
  $\mathcal{S}_l = \emptyset$, $\Psi_l = \emptyset$
  for $k = 1 : K_l$ **do**
    $a_l = \alpha^*_{l,k} h_l$.
    Find the vertices of the corresponding $\mathcal{V}_{l,a_l}$, calculated by $\alpha^*_{l,k} + \frac{1}{h_l}\frac{z_0}{2}$
    $z_0 = \pm 1 \pm i$ for $\mathbb{Z}[i]$
    $z_0 = \pm 1 \pm \frac{\sqrt{3}}{3} i,\ \pm\frac{2\sqrt{3}}{3} i$ for $\mathbb{Z}[\omega]$
    Store these vertices into set $\mathcal{S}_l$
    Calculate the linear equation of each edge of $\mathcal{V}_{l,a_l}$, add them into set Ψ_l
  end for
  $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_l$
end for
for $\bar{l} = 1 : L - 1$ **do**
  for $\hat{l} = \bar{l} + 1 : L$ **do**
    Find all combinations of $\{c_1, c_2\}$, with $c_1 \in \Psi_{\bar{l}}$ and $c_2 \in \Psi_{\hat{l}}$. Calculate the crossing point of c_1 and c_2: the crossing points which are not in $\widetilde{\Phi}$ should be discarded. Store the remaining in set $\mathcal{S}_{\bar{l},\hat{l}}$
    $\mathcal{S}_{\bar{l}} = \mathcal{S}_{\bar{l}} \cup \mathcal{S}_{\bar{l},\hat{l}}$
  end for
  $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_{\bar{l}}$
end for

$\mathbf{Phase\ 2}$: Obtain the candidate set $\mathcal{I}$ and select $\mathbf{a}_{\text{opt}}$

for all representative $\alpha \in \mathcal{S}$ **do**
  $\mathbf{a} = \mathcal{Q}^*_{\mathbb{A}}(\alpha \mathbf{h})$, discard the repeated $\mathbf{a}$, take the remaining $\mathbf{a}$ as the candidate set $\mathcal{I}$
end for
for all $\mathbf{a} \in \mathcal{I}$ **do**
  Calculate $\mathcal{R}(\mathbf{h}, \mathbf{a})$ by equation (2.41)
end for
Return $\mathbf{a}_{\text{opt}} = \text{argmax}_{\mathbf{a} \in \mathcal{I}} \mathcal{R}(\mathbf{h}, \mathbf{a})$

each $\mathcal{Q}_{\mathbb{Z}[i]}(\alpha_{\mathrm{v}}\mathbf{h})$ has four possible values, while the number of vertices of each $\mathcal{V}_{\mathbf{a}}$ is uncertain. Particularly for a triangle $\mathcal{V}_{\mathbf{a}}$, the redundancy (3 vertices) is apparently not able to compensate the quantisation uncertainty. This means if we set the quantiser to round $\alpha_{\mathrm{v}}\mathbf{h}$ in a specific direction for all vertices, we might miss that triangle $\mathcal{V}_{\mathbf{a}}$ (and also the corresponding $\mathbf{a}$). Hence, we define a 'full direction' quantiser $\mathcal{Q}^*_{\mathbb{A}}(\cdot)$ which returns all equal likely $\mathbf{a}$. For example, $\mathcal{Q}^*_{\mathbb{Z}[i]}(0.5+1.5i) = \{1+2i, 1+1i, 0+2i, 0+1i\}$. The modified quantiser ensures that all Voronoi regions $\mathcal{V}_{\mathbf{a}}$ can be 'visited' by considering α_{v} as the representative, thus all candidates vectors $\mathbf{a} \in \mathcal{I}$ can be acquired.

The only issue remaining is to calculate the coordinates of the vertices. Clearly, each vertex is a crossing point of two lines, denoted by c_1 and c_2, and each line is exactly an edge of an individual $\mathcal{V}_{l,a_l}$. Let Ψ_l denote the set which includes the edges of $\mathcal{V}_{l,a_l}$, $\forall a_l$. There are two kinds of vertex which can be acquired separately, which means $\mathcal{S}$ can be divided into two subsets as follows:

- $\mathcal{S}$-I: the vertices of individual $\mathcal{V}_{l,a_l}$. This corresponds to the case that c_1 and c_2 belong to the same Ψ_l,
- $\mathcal{S}$-II: intersection points of two sets of parallel lines, where one set belongs to $\Psi_{\bar{l}}$ and the other belongs to $\Psi_{\hat{l},\hat{l}\neq\bar{l}}$. This corresponds to the case that c_1 and c_2 belong to different Ψ_l, denoted as $c_1 \in \Psi_{\bar{l}}$ and $c_2 \in \Psi_{\hat{l}}$.

Intuitively, the former indicates the vertices of the red/blue squares in Figure 3.3, and these points can be easily acquired based on Λ_l which is defined in (3.18), while the latter indicates the vertices of the parallelograms in Figure 3.3 (labelled by the black shading). The proposed complex-Exhaustive-II algorithm is summarised in Algorithm 3, and the corresponding procedure for acquiring $\mathcal{S}$-I and $\mathcal{S}$-II are outlined in steps 2$\sim$12 and 13$\sim$19 respectively. Starting with the obtained candidate set $\mathcal{I}$, we employ (2.41) to examine $\mathcal{R}(\mathbf{h}, \mathbf{a})$ for all $\mathbf{a} \in \mathcal{I}$, and select the optimal one.

Complexity of Complex-Exhaustive-II Algorithm

The complexity mainly depends on the number of candidates $\mathbf{a}$, and this number is upper bounded by the outputs of $\mathcal{Q}^*_{\mathbb{A}}(\alpha\mathbf{h})$ for all α in $\mathcal{S}$ (step 20$\sim$21 in Algorithm 3). Let ξ denote the number of quantiser outputs for each $\alpha \in \mathcal{S}$, expressed as $\xi = |\mathcal{Q}(\alpha\mathbf{h})|$. Clearly, each representative $\alpha \in \mathcal{S}$-I is shared by 4 and 3 adjacent Voronoi regions for $\mathbb{Z}[i]$-lattices and $\mathbb{Z}[\omega]$-lattices respectively, thus $\xi = 4$ (3) for $\mathbb{Z}[i]$ ($\mathbb{Z}[\omega]$) respectively. In principle, the probability of more than 2 independent lines intersect at the same point is infinitesimal, hence $\xi = 4$ for both $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ cases if $\alpha \in \mathcal{S}$-II. Therefore, ξ can be regarded as a constant, and hence the number of candidates $\mathbf{a}$ is bounded by $\xi|\mathcal{S}|$. Since the area of the valid range of α is bounded by SNR, the total number of $\mathcal{V}_{l,a_l}$ for all l is therefore expected to be

$$\sum_{l} \mathbb{E}\Big[\frac{\mathrm{SNR}}{\mathcal{A}_{\mathcal{V}_l}}\Big] = \sum_{l} \frac{\mathrm{SNR}\mathbb{E}[|h_l|^2]}{\mathcal{A}_{\mathcal{V}_0}} = \frac{\mathrm{SNR}L}{\mathcal{A}_{\mathcal{V}_0}}, \tag{3.24}$$

where $\mathcal{A}_{\mathcal{V}_0}$ is a constant for given integer domain $\mathbb{A}$ as described previously, and the second equality is due to the assumption that $h_l \sim \mathcal{CN}(0,1)$.

For each pair of sets of parallel lines from $\Psi_{\bar{l}}$ and $\Psi_{\hat{l},\hat{l}\neq\bar{l}}$, the expected number of parallelograms for $\mathbb{Z}[i]$-lattices is

$$\mathbb{E}\Big[\frac{\mathrm{SNR}}{\mathcal{A}_{para}}\Big] = \frac{\mathrm{SNR}\mathbb{E}[|h_{\bar{l}}||h_{\hat{l}}|\sin(\theta_{\bar{l},\hat{l}})]}{\mathcal{A}_{\mathcal{V}_0}} \tag{3.25}$$

$$= \frac{\mathrm{SNR}\mathbb{E}[|h_{\bar{l}}|]\mathbb{E}[|h_{\hat{l}}|]\mathbb{E}[\sin(\theta_{\bar{l},\hat{l}})]}{\mathcal{A}_{\mathcal{V}_0}} \tag{3.26}$$

$$= 0.5\frac{\mathrm{SNR}}{\mathcal{A}_{\mathcal{V}_0}}. \tag{3.27}$$

Here $\mathcal{A}_{para}$ denotes the area of the parallelograms, and $\theta_{\bar{l},\hat{l}}$ denotes the intersection angle of the two sets of parallel lines which is randomly distributed within $0 \sim \frac{\pi}{2}$, hence $\mathbb{E}[\sin(\theta_{\bar{l},\hat{l}})] = \frac{2}{\pi}$. The expression (3.26) comes from the independence of the channel components $h_{\hat{l}}$ and $h_{\bar{l}}$. Since the expected value of $|h|$ equals $\sqrt{\frac{\pi}{4}}$ with $h \sim \mathcal{CN}(0,1)$, the simplified expression is therefore written as (3.27). Similarly, we expect the number

of parallelograms as

$$\mathbb{E}\Big[\frac{\mathrm{SNR}}{\mathcal{A}_{para}}\Big] = \frac{\frac{\sqrt{3}}{2}\mathrm{SNR}\mathbb{E}[|h_{\bar{l}}||h_{\hat{l}}|\sin(\theta_{\bar{l},\hat{l}})]}{(\frac{1}{2})^2\mathcal{A}_{\mathcal{V}_0}} = \sqrt{3}\frac{\mathrm{SNR}}{\mathcal{A}_{\mathcal{V}_0}}. \tag{3.28}$$

for $\mathbb{Z}[\omega]$-lattices. Since there are respectively 2 (3) sets of parallel lines for $\mathbb{Z}[i]$ ($\mathbb{Z}[\omega]$) in each Ψ_l, the total number of parallelograms is therefore expected to be

$$\binom{L}{2}\binom{2}{1}\binom{2}{1}\mathbb{E}\Big[\frac{\mathrm{SNR}}{\mathcal{A}_{para}}\Big] \text{ and } \binom{L}{2}\binom{3}{1}\binom{3}{1}\mathbb{E}\Big[\frac{\mathrm{SNR}}{\mathcal{A}_{para}}\Big] \tag{3.29}$$

for $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ respectively. The expressions (3.24) and (3.29) can be roughly regarded as the the expected values of $|\mathcal{S}\text{-I}|$ and $|\mathcal{S}\text{-II}|$. Since the total number of candidates is $\xi(|\mathcal{S}\text{-I}| + |\mathcal{S}\text{-II}|)$ and the computation rate can be calculated in $\mathcal{O}(L)$ for each candidate $\mathbf{a}$, hence the overall time complexity can be expressed as

$$\mathcal{O}(\mathrm{SNR}L^2(L-1)) + \mathcal{O}(\mathrm{SNR}L^2). \tag{3.30}$$

Note that the constant components are omitted in (3.30), and their effect will be evaluated numerically in section 3.5.

L-L Algorithm vs Complex-Exhaustive-II Algorithm

The L-L algorithm in [37] is described as an optimal deterministic algorithm. Actually, it does not ensure the optimal solution for some channel realisations. In this section, we will present a comparison between the L-L algorithm and our proposed complex exhaustive-II algorithm.

The main difference between these two algorithms is the representatives of α. The exhaustive-II algorithm considers both $\mathcal{S}$-I and $\mathcal{S}$-II, while the L-L algorithm considers the individual $\mathcal{V}_{l,a_l}$ only. Specifically, the vertices and the midpoints of edges of individual $\mathcal{V}_{l,a_l}$ are considered for L-L, hence the representatives of α can be regarded as an extended version of $\mathcal{S}$-I (though the L-L algorithm is not interpreted in this manner in [37]). However, $\mathbf{a}_{\mathrm{opt}}$ is sometimes generated by $\mathcal{Q}^*_{\mathbb{A}}(\alpha\mathbf{h})$ with $\alpha \in \mathcal{S}$-II. In this case, $\mathbf{a}_{\mathrm{opt}}$ is missed

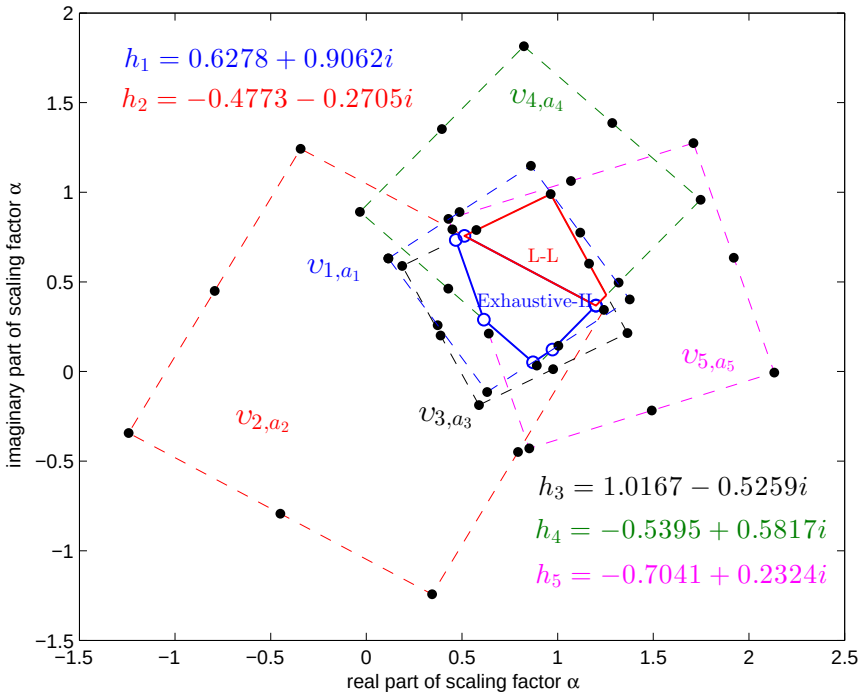

Figure 3.4: Comparison of L-L and complex exhaustive-II

by the L-L algorithm. In Figure 3.4, we present an intuitive comparison of these two algorithms. A $\mathbb{Z}[i]$-lattice based system is considered, with $L = 5$ and SNR = 10dB. The channel components h_l and their corresponding $\mathcal{V}_{l,a_l}$, $\forall l$ are denoted by different colours. The representatives of α utilised in L-L are marked by the black dots, which result in the 'optimal' solution $\hat{\mathbf{a}}_{\text{opt}} = [1i, -1i, 1, -1, -1]$ with $\mathcal{R}(\hat{\mathbf{a}}_{\text{opt}}, \mathbf{h}) = 0.585$. However, the actual optimal solution acquired by employing the complex exhaustive-II algorithm is $\mathbf{a}_{\text{opt}} = [1i, 1i, 1, -1, -1]$ with $\mathcal{R}(\mathbf{a}_{\text{opt}}, \mathbf{h}) = 0.702$. Let $\mathcal{V}_{\text{opt}}$ denote the corresponding Voronoi region of $\mathbf{a}_{\text{opt}}$, which is represented by the blue solid polygon (labelled as Exhaustive-II) in Figure 3.4. Clearly, $\mathcal{V}_{\text{opt}}$ can be 'visited' by the blue points marked with blue circles from $\mathcal{S}$-II. By contrast, none of the black dots are located within $\mathcal{V}_{\text{opt}}$, and hence $\mathbf{a}_{\text{opt}}$ is not considered as a candidate in the L-L algorithm.

3.4.2 Linear Search Algorithm for Complex Lattices

The complex-exhaustive-II method requires the calculation of the vertices of all $\mathcal{V}_{\mathbf{a}}$ in order to obtain the complete candidate set $\mathcal{I}$. In this section we extend the linear search

in section 3.3.3 to the complex case, in which we simply employ the sampled values of α to acquire the candidate vectors. Recall that the candidates $\mathbf{a}$ which correspond to wide intervals are referred to as the necessary candidates in the real valued scenario. In this section, we will similarly show that the candidates $\mathbf{a}$ correspond to some large $\mathcal{V}_{\mathbf{a}}$ can be regarded as the necessary candidates.

Note that the reduced candidate method in section 3.3.2 requires the length of all intervals for each instantaneous $\mathbf{h}$. For the complex case, calculating the area of each $\mathcal{V}_{\mathbf{a}}$ will bring enormous complexity to the online search. Therefore, we alternatively combine the algorithms in 3.3.2 and 3.3.3 in this section. We still investigate the threshold of necessary candidates, and then take the threshold as the step size for sampling, in order to achieve a good performance-complexity tradeoff.

Off-line Search: Obtain the Optimal Step Size

As previously discussed, the corresponding $\mathcal{V}_{\mathbf{a}}$ of $\mathbf{a}$ are uniformly distributed over the valid range of α with random sizes. Hence we utilise the simplest uniform sampler to generate α as

$$\alpha_{\text{sample}} = \Delta(k_1 + k_2 i),\ k_1, k_2 \in \mathbb{Z}, \tag{3.31}$$

where k_1 and k_2 are non-negative integers and not both zero. The positive real number Δ denotes the step size which controls the sampling rate. The key factor is to choose a proper step size.

Figure 3.5 gives an intuitive view of determining the step size. We adopt the same channel and axis labelling as in Figure 3.3. Again the x-axis and y-axis denote the real and imaginary parts of α respectively, and the corresponding effective noise calculated by

$$\sigma^2_{\text{eff}}(\alpha) = \|\alpha\|^2\sigma^2 + P\|\alpha\mathbf{h} - \mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h})\|^2 \tag{3.32}$$

is shown in the colour bar. The 1st and 2nd order derivatives of $\sigma^2_{\text{eff}}(\alpha)$ for complex lattices are given by

$$\frac{d\sigma^2_{\text{eff}}}{d\alpha} = 2\alpha\sigma^2 + 2P\alpha\|\mathbf{h}\|^2 - 2\mathbf{h}^H\mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h}). \tag{3.33}$$

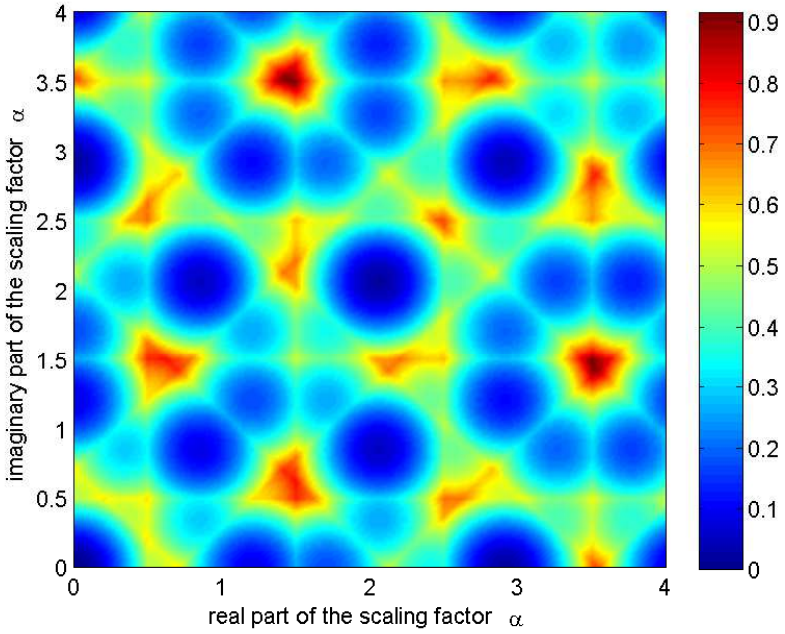

Figure 3.5: The corresponding $\sigma^2_{\text{eff}}(\alpha)$ of the example in Figure 3.3, SNR=30dB

and

$$\frac{d^2\sigma^2_{\text{eff}}}{d\alpha^2} = 2\sigma^2 + 2P\|\mathbf{h}\|^2 \geq 0. \tag{3.34}$$

respectively. Similar to the real case, $\mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h})$ is invariant within each $\mathcal{V}_{\text{a}}$, and the 2nd derivative is the same for all candidates $\mathbf{a}$. Thus, the global minimum is more likely to be located in one of the larger $\mathcal{V}_{\mathbf{a}}$. As shown in Figure 3.5, the dark blue regions correspond to the large $\mathcal{V}_{\mathbf{a}}$ in Figure 3.3. Their corresponding $\mathbf{a}$ can be regarded as 'necessary candidates' since they have lower effective noise.

Let g denote the number of edges of $\mathcal{V}_{\text{opt}}$. For the example demonstrated in Figure 3.5, the Voronoi region of the optimal vector $\mathcal{V}_{\text{opt}}$ corresponds to the black octagon labelled in Figure 3.3. Let d_{opt} denote the width of the largest square (with all sides vertical or horizontal) that fits in $\mathcal{V}_{\text{opt}}$, as shown in Figure 3.6. It can be observed that $\mathcal{V}_{\text{opt}}$ will definitely be visited if $\Delta \leq d_{\text{opt}}$. Finding the largest square in $\mathcal{V}_{\text{opt}}$ is a convex optimisation problem described as:

$$\begin{aligned} \underset{\mathbf{m}}{\text{maximise}} \quad & \mathbf{m}^T\mathbf{Q}\mathbf{m} \\ \text{subject to} \quad & \mathbf{A}\mathbf{m} \leq \mathbf{b} \\ & \text{and } m_1 + m_2 = m_3 + m_4, \end{aligned}$$

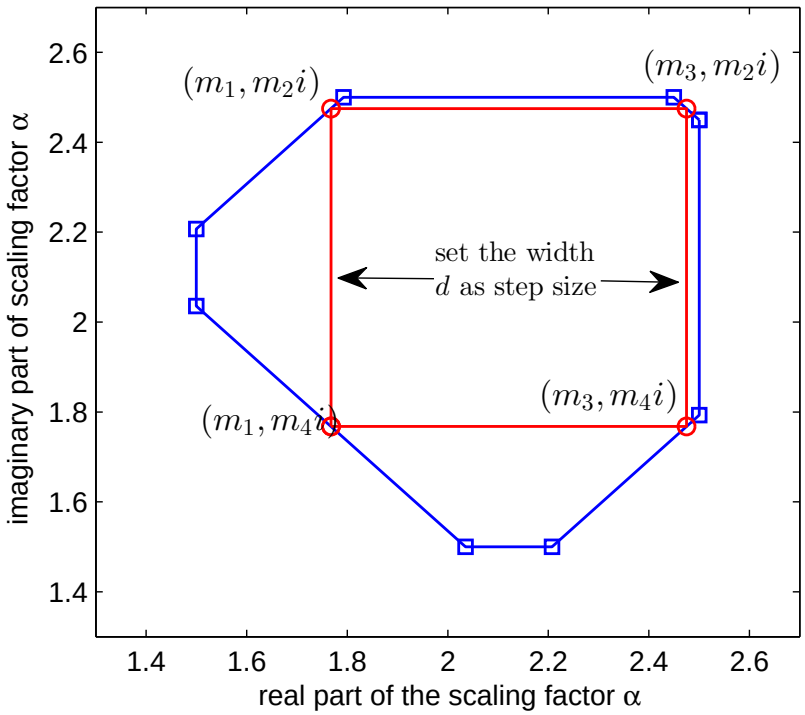

Figure 3.6: Finding the step size according to $\mathcal{V}_{\text{opt}}$

where $\mathbf{Q} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix}$, and $\mathbf{m} = [m_1, m_2, m_3, m_4]^T$, as labelled in Figure 3.6. The restriction $\mathbf{Am} \leq \mathbf{b}$ comprises $4g$ linear equations which correspond to the condition that the 4 vertices of the square should be located within the g-edge convex polygon. Such an optimisation problem is linearly solvable, with complexity $\mathcal{O}(g)$.

Similar to the real valued case: given SNR and L, the threshold of large Voronoi regions $\mathcal{V}_{\mathbf{a}}$ can be determined by exploiting the statistical characteristic of d_{opt}. As described in Proposition 2, the area of $\mathcal{V}_{\text{opt}}$ is upper bounded by $\mathcal{A}_{\mathcal{V}} = \mathcal{A}_{\mathcal{V}_0}/|h_{\max}|^2$ for a particular $\mathbf{h}$, hence we define the normalised d_{opt} as $\gamma_{\text{opt}} = d_{\text{opt}}/\sqrt{\mathcal{A}_{\mathcal{V}}}$. Let $\mathcal{V}_{\text{rand}}$ denote a random $\mathcal{V}_{\mathbf{a}}$ within the valid range of α, and d_{rand} be the width of the largest square that fits in $\mathcal{V}_{\text{rand}}$. Similarly $\gamma_{\text{rand}} = d_{\text{rand}}/\sqrt{\mathcal{A}_{\mathcal{V}}}$. Clearly, $0 < \gamma_{\text{opt}}, \gamma_{\text{rand}} \leq 1$. Figure 3.7 illustrates the cumulative distribution of γ_{opt} in a 5-user, $\mathbb{Z}[i]$-lattice based system. Again, the results are acquired over 1000 channel trials. It can be observed that $\gamma_{\text{thre}} = 0.28$, $\gamma_{\text{thre}} = 0.39$, and $\gamma_{\text{thre}} = 0.47$ can be assigned to SNR=20dB, 30dB and 40dB respectively. As same as in the real valued case: the threshold γ_{thre} increases monotonically

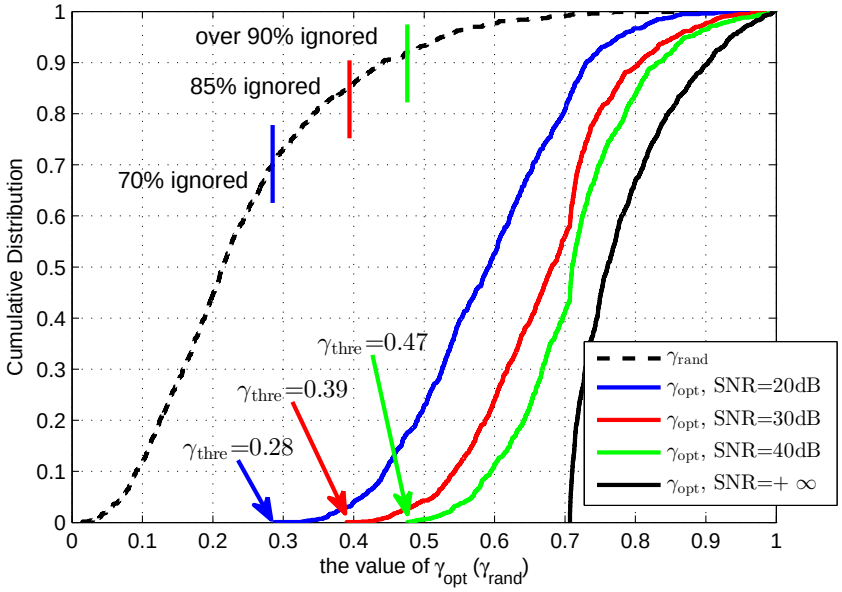

Figure 3.7: The cumulative distribution of γ_{opt} and γ_{rand}

with SNR, and hence the values of γ_{thre} for SNR=20dB can be used in the region of 20dB$\leq$SNR$<$25dB without the loss of accuracy; the threshold of γ_{opt} also depends on SNR and L only, and not on any particular instantaneous channel, hence an off-line table Θ can be established to store the values of γ_{thre} corresponding to each L and SNR region, which will not bring extra complexity to the online search.

Given a instantaneous channel $\mathbf{h}$, we set the step size as

$$\Delta = \gamma_{\text{thre}}\sqrt{\mathcal{A}_{\mathcal{V}}} = \gamma_{\text{thre}}\sqrt{\frac{\mathcal{A}_{\mathcal{V}_0}}{|h_{\max}|^2}} \tag{3.35}$$

to capture the necessary candidates. Again, the complexity reduction resulting from this stage can be estimated by investigating the distribution of d_{rand}. We can see that the probability of $d_{\text{rand}} \leq 0.47$ is roughly 70%, which means that about 70% of the candidates examined in the complex-exhaustive-II search are ignored by the sampled values. Thus, the complexity is potentially reduced by 70% when SNR=20dB, and this reduction increases with SNR. Note that the meaning of the 'ignored candidates' here is slightly different to the 'discarded candidates' in Figure 3.2. Unlike the reduced candidate set aided algorithm, some of the ignored $\mathcal{V}_{\mathbf{a}}$ here might still be visited by the sampling values of α, hence the proportions labelled in Figure 3.7 are referred to as a potential complexity reduction.

Table 3.2: A partial table of Θ with $L = 5, 8, 10, 20$ (complex valued case)

	SNR(dB) / L	<5	[5 10)	[10 15)	[15 20)	[20 25)	[25 30)	[30 35)	[35 40)	$\cdots$	$+\infty$
$\mathbb{Z}[i]$	5	0.06	0.09	0.12	0.21	0.28	0.33	0.39	0.44	$\cdots$	0.71
	8	0.04	0.05	0.07	0.13	0.16	0.25	0.32	0.38	$\cdots$	0.71
	10	0.04	0.05	0.06	0.10	0.12	0.17	0.22	0.29	$\cdots$	0.71
	20	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.06	$\cdots$	0.71
$\mathbb{Z}[\omega]$	5	0.06	0.10	0.12	0.20	0.29	0.33	0.40	0.44	$\cdots$	0.71
	8	0.05	0.05	0.08	0.11	0.16	0.24	0.32	0.37	$\cdots$	0.71
	10	0.04	0.05	0.06	0.09	0.13	0.18	0.23	0.28	$\cdots$	0.71
	20	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.05	$\cdots$	0.71

Algorithm 4 Linear search algorithm

Output: optimal coefficient vector $\mathbf{a}_{\text{opt}}$

Offline Search: obtain table of Θ

Given particular L and SNR

1: **for** $trail = 1 : 1000$ **do**
2: generate $\mathbf{h}_{trial} \in \mathbb{C}^L$
3: obtain $\mathbf{a}_{\text{opt},trial} = \text{argmax}\, \mathcal{R}(\mathbf{h}_{trial}, \mathbf{a})$ by exhaustive-II, and acquire its corresponding $\mathcal{V}_{\text{opt},trial}$
4: calculate the normalised width $\gamma_{\text{opt},trial}$ for $\mathcal{V}_{\text{opt},trial}$
5: **end for**
6: set $\min_{trial=1}^{1000} \gamma_{\text{opt},trial} \rightarrow \gamma_{\text{thre}}(L, \text{SNR})$

Online Search: obtain $\mathbf{a}_{\text{opt}}$ for a given $\mathbf{h}$

7: $\Delta = \gamma_{\text{thre}} \sqrt{\frac{\mathcal{A}_{\mathcal{V}_0}}{|h_{\max}|^2}}$ (Eq. 3.35)
8: Generate α_{sample} within the valid range of α, and stored these samples in $\mathcal{S}$.
9: Generate $\mathbf{a} = \mathcal{Q}_{\mathbb{A}}(\alpha_{\text{sample}}\mathbf{h}),\ \forall \alpha_{\text{sample}} \in \mathcal{S}$, and discard the repeated $\mathbf{a}$.
10: Sort the remaining $\mathbf{a}$ in ascending order to their corresponding $|\alpha_{\text{sample}}|$, and store these remaining $\mathbf{a}$ in $\mathcal{I} = \{\mathbf{a}_1, \cdots, \mathbf{a}_J\}$, and α_{sample} in $\widetilde{\mathcal{S}} = \{\widetilde{\alpha}_1, \cdots, \widetilde{\alpha}_J\}$.
11: $j = 2$, $\mathbf{a}_{\text{opt}} = \mathbf{a}_1$, $\sigma^2_{\text{opt}} = \mathbf{a}^H_{\text{opt}} \mathbf{M} \mathbf{a}_{\text{opt}}$ (Eq. 2.41).
12: **while** $\widetilde{\alpha}_j^2 \sigma^2 \leq \sigma^2_{\text{opt}}$ and $j \leq J$ **do**
13: $\sigma^2_{\text{eff}} = \mathbf{a}_j^H \mathbf{M} \mathbf{a}_j$, $\sigma^2_{\text{opt}} = \min(\sigma^2_{\text{opt}}, \sigma^2_{\text{eff}})$, $\mathbf{a}_{\text{opt}} = \underset{\mathbf{a} \in \{\mathbf{a}_{\text{opt}}, \mathbf{a}_j\}}{\text{argmin}}\ \mathbf{a}^H \mathbf{M} \mathbf{a}$.
14: $j = j + 1$
15: **end while**
16: Return $\mathbf{a}_{\text{opt}} = \mathcal{Q}_{\mathbb{A}}(\alpha_{\text{opt}}\mathbf{h})$

Online Search: Obtain the Optimal Integer Vector

Based on the table Θ, the step size for a instantaneous channel $\mathbf{h}$ is determined by (3.35). The task of the online search is to check $\mathbf{a} = \mathcal{Q}_{\mathbb{A}}(\alpha_{\text{sample}}\mathbf{h})$ one by one (the repeated

candidates are discarded directly), and select $\mathbf{a}_{\text{opt}}$. As outlined in Algorithm 4, we employ the following processes to make it more efficient.

- The values of α_{sample} are sorted in ascending order to their amplitudes (line 10, Algorithm 4). Note that the step size Δ changes the scale of α_{sample} only, and the order of α_{sample} is invariant to $\mathbf{h}$. Thus, no extra complexity is required for sorting α_{sample} when $\mathbf{h}$ varies.
- We employ the same break condition as described in section 3.3: Let σ^2_{opt} be the minimum effective noise obtained from the preceding samples, the search terminates when $\widetilde{\alpha}_j^2\sigma^2 \geq \sigma^2_{\text{opt}}$ (Line 12, Algorithm 4).

Note that we will not apply this online process to the complex-exhaustive-II algorithm for the following reasons:

- In the complex-exhaustive-II algorithm, the representatives of $\alpha \in \mathcal{S}$ do not have a fixed meshed structure as the sampled valued in (3.31), and the order of these $\alpha \in \mathcal{S}$ varies with the instantaneous channel $\mathbf{h}$. Thus, a complexity of $n\log(n)$ is required for sorting these α, assuming that $n = |\mathcal{S}|$ is the total number of the representatives.
- In the reduced candidate set aided algorithm, each representative $\alpha^*_{\bar{k}}$ is the left endpoint of the corresponding interval $[\alpha^*_{\bar{k}}, \alpha^*_{\bar{k}+1})$. Since $\alpha > 0$, $\alpha^*_{\bar{k}}$ has the smallest amplitude in $[\alpha^*_{\bar{k}}, \alpha^*_{\bar{k}+1})$. Thus, $(\alpha^*_{\bar{k}}\sigma)^2 \geq \sigma^2_{\text{opt}}$ is a sufficient condition of $\alpha^2\sigma^2 \geq \sigma^2_{\text{opt}},\ \alpha \in [\alpha^*_{\bar{k}}, \alpha^*_{\bar{k}+1})$. In contrast, each representative employed in the complex-exhaustive-II method α_{v} is a vertex of $\mathcal{V}_{\mathbf{a}}$, having $\alpha^2_{\text{v}}\sigma^2 \geq \sigma^2_{\text{opt}}$ is not sufficient to obtain $\alpha^2\sigma^2 \geq \sigma^2_{\text{opt}},\ \alpha \in \mathcal{V}_{\mathbf{a}}$. Therefore, this break condition might lose the optimality of the complex-exhaustive-II algorithm.

Complexity of the Linear Search Algorithm

The complexity of the linear search algorithm can be analysed from two perspectives. On the one hand, the proportion of candidates ignored is quite small ($\gamma_{\text{thre}} \approx 0$) in the low SNR region, and hence the complexity of the linear search can be measured by the

exhaustive-II search. On the other hand, the number of candidates for the high SNR case can be expected to be

$$\frac{\mathrm{SNR}}{E[\Delta^2]} = \frac{\mathrm{SNR}}{\mathcal{A}_{\mathcal{V}_0}} E[\frac{|h_{\max}|^2}{\gamma_{\mathrm{thre}}^2}] = \frac{\mathrm{SNR}}{\gamma_{\mathrm{thre}}^2 \mathcal{A}_{\mathcal{V}_0}} E[|h_{\max}|^2], \tag{3.36}$$

where the first equality comes from (3.35), and the second is due to the fact that the threshold γ_{thre} tends to a constant in the high SNR region: when SNR$\to \infty$, the optimal α is free to be chosen as the least common multiple of $\{\frac{1}{h_l} : \ l = 1 : L\}$. In this case, the centre points of all individual $\mathcal{V}_{l,a_l}$ (see Prop. 2) overlap, and hence the optimal Voronoi $\mathcal{V}_{\mathrm{opt}}$ is very likely to be the smallest individual $\mathcal{V}_{l,a_l}$. By employing the moment generating function of $|h_{\max}|^2$, we have

$$\begin{aligned}
\mathbb{E}[|h_{\max}|^2] &= \frac{1}{\beta}\mathbb{E}[\log e^{\beta|h_{\max}|^2}] \quad (\beta > 0) && (3.37)\\
&\leq \frac{1}{\beta}\log\mathbb{E}[e^{\beta|h_{\max}|^2}] && (3.38)\\
&= \frac{1}{\beta}\log\int_0^\infty \Pr(e^{\beta|h_{\max}|^2} \geq x)dx && (3.39)\\
&\leq \frac{1}{\beta}\log\int_0^\infty \sum_{l=1}^{L} \Pr(e^{\beta|h_l|^2} \geq x)dx && (3.40)\\
&= \frac{1}{\beta}\log\sum_{l=1}^{L} \mathbb{E}[e^{\beta\|h_l\|^2}] && (3.41)\\
&= \frac{1}{\beta}\log\frac{L}{1-2\beta} && (3.42)\\
&= \frac{1}{\beta}(\log L - \log\frac{1}{1-2\beta}) && (3.43)
\end{aligned}$$

where (3.38) comes from Jensen's inequality. (3.39) and (3.41) are based on the relation between the expected value and the survival function. (3.40) is obtained by Boole's inequality (also known as the union bound), and (3.42) is valid because the moment generating function of a chi-square variable $|h_l|^2$ is $\frac{1}{1-2\beta}$. Since (3.42) holds for any $\beta > 0$, we can pick β to tighten this bound. According to Arithmetic Mean-Geometric Mean (AMGM) inequality, the minimum value of (3.43) is attained by setting $\log L = \log\frac{1}{1-2\beta}$. Hence, we have $\mathbb{E}[|h_{\max}|^2] \leq \frac{4\log L}{1-\frac{1}{L}}$. Again, the corresponding $\mathcal{R}(\mathbf{a}, \mathbf{h})$ for each candidate vector $\mathbf{a}$ can be calculated in $\mathcal{O}(L)$, and hence the time complexity for high SNR case is given by

$$\mathcal{O}(\mathrm{SNR}L\frac{\log L}{1-\frac{1}{L}}) \tag{3.44}$$

with the constant components omitted.

3.5 Numerical Results and Discussions

3.5.1 Real Valued Channel

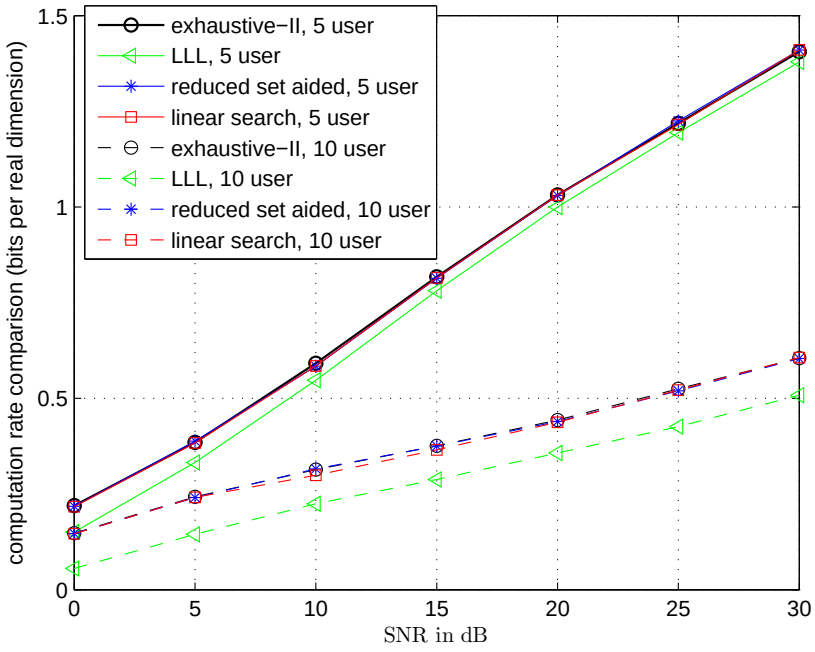

Figure 3.8: Average $\mathcal{R}(\mathbf{h})$ comparison: real valued channel

In this section, we investigate both the computation rate the complexity of our proposed algorithms in section 3.3, compared with the exhaustive-II search [6] and the LLL algorithm [42]. We choose 0.4 as the step size of the linear search. The values of γ_{thre} are drawn from Table 3.1. We consider two scenarios in which 5 users and 10 users are employed respectively. All results are the average over 10000 channel realisations.

Figure 3.8 shows the rate comparison of the 4 algorithms. We use solid and dashed lines to represent the case of 5 users and 10 users respectively. We use the exhaustive-II search (marked with circles) as the benchmark since it ensures the optimal solution. Compared to this, the reduced set aided algorithm attains the same computation rate, which proves the reduced set is sufficient for finding the best vector. The linear search

algorithm also achieves the optimal throughput with 5 users, and performs slightly worse in low SNR with 10 users. This is because γ_{thre} is very small in such a scenario (see Table 1), and therefore $\Delta = 0.4$ is not small enough to visit every wide interval of α. Both of our proposed methods outperform the LLL in both scenarios, and the gap increases with the number of users.

The complexity is measured by the average number of the floating point operations per channel realisation. In our simulation, each addition or multiplication is counted as 1, and the rounding operations and comparisons are ignored. The results are demonstrated in Figure 3.9. Unsurprisingly, the complexity of the exhaustive-II approach is much higher than that of the other three. Among the remaining three, our proposed algorithms have similar complexities. Since the LLL algorithm has a complexity of $\mathcal{O}(L^4 \log L)$ which is not related to SNR, hence the comparison between our proposed algorithms with the LLL is an SNR-L tradeoff. Compared with the LLL, both our algorithms have lower complexities within a wide range of SNR, and the range gets wider as the number of users increases. Actually, the most common SNR region in practical wireless environments (10dB$\sim$30dB) is covered by these ranges.

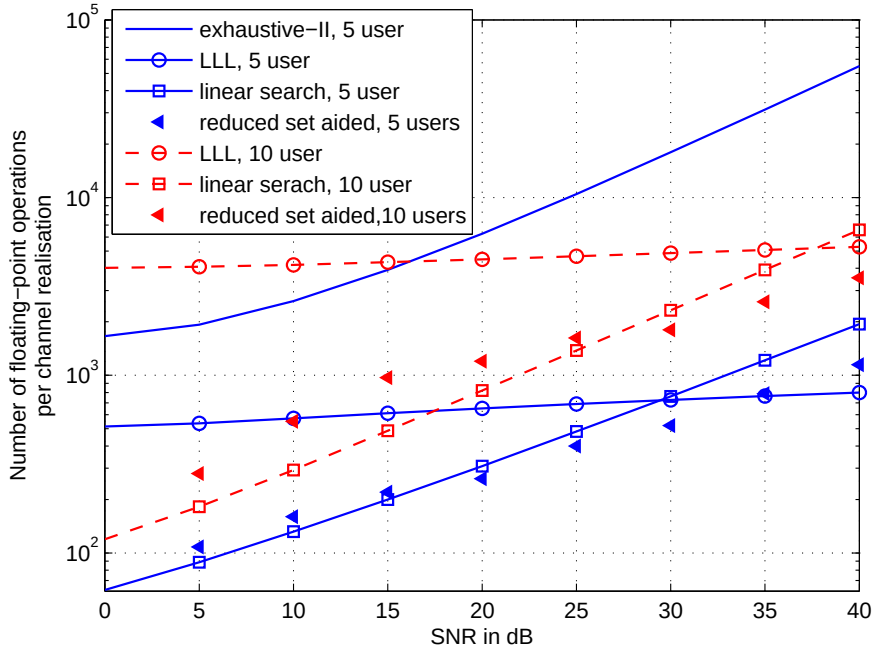

Figure 3.9: Average complexity comparison: real valued channel

3.5.2 Complex Valued Channel

In this section, we investigate both the computation rate and the complexity of our proposed algorithms in section 3.4, compared with the CLLL method [58] and the L-L [37] algorithm. Again, all results are acquired over 10000 channel realisations.

Computation Rate Comparison

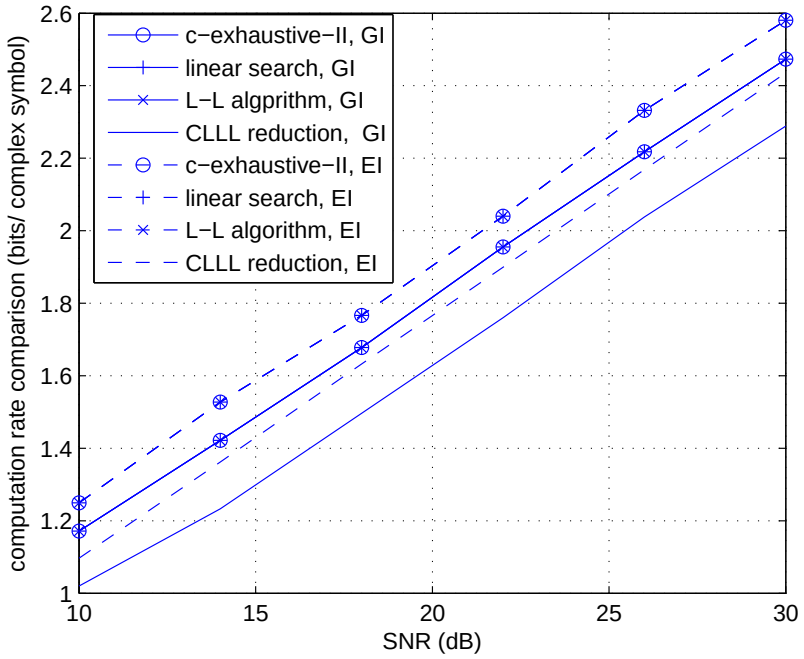

Figure 3.10: Average $\mathcal{R}(\mathbf{h})$ comparison: complex valued channel, 5 users

Figure 3.10 shows the average $\mathcal{R}(\mathbf{h})$ of a 5 user scenario. We use solid and dashed lines to represent the case of $\mathbb{Z}[i]$ (denoted as GI) and $\mathbb{Z}[\omega]$ (denoted as EI) based lattices respectively. Unsurprisingly, the denser structure of $\mathbb{Z}[\omega]$ leads to a better performance than the $\mathbb{Z}[i]$ based lattice. Previously we have established that both the L-L algorithm and the linear search method might sometimes miss the optimal solution. However, the numerical results reveal that the probability of missing $\mathbf{a}_{\text{opt}}$ is quite small. The gaps to the exhaustive-II algorithm are negligible for both algorithms, and they all outperform the CLLL method. Similarly, Figure 3.11 reveals the rate comparison of a 10 user scenario. Compared to the case of $L = 5$, the advantage of our proposed algorithms to the CLLL is increased.

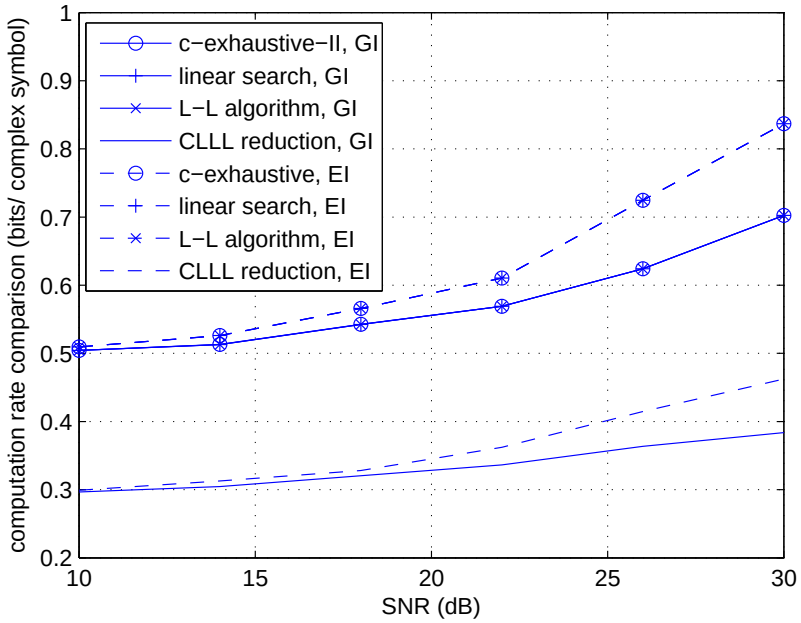

Figure 3.11: Average $\mathcal{R}(\mathbf{h})$ comparison: complex valued channel, 10 users

Complexity Comparison

Again, we investigate the complexity by counting the flops. The number of flops required for each complex addition and multiplication are 2 and 6 respectively, and the rounding operations are ignored in the simulation. It suffices to consider $\mathbb{Z}[i]$-based lattice only (any other non-cubic lattices have a similar result). By considering $h_l \sim \mathcal{CN}(0,1)$ and $\mathbb{E}[\|\mathbf{h}\|^2] = L$, the complexity of the L-L algorithm in [37] can be rewritten as

$$\mathcal{O}\big(L^2(\mathrm{SNR}L + \sqrt{\mathrm{SNR}L} + 2)\big). \tag{3.45}$$

Compared with (3.30), we can see that the L-L algorithm and the exhaustive-II algorithm have almost the same theoretical complexity, both being dominated by $\mathcal{O}(L^3\mathrm{SNR})$. However, numerical results in Figure 3.12, 3.13, and 3.14 reveal that our proposed exhaustive-II algorithm has less complexity than the L-L algorithm. The reasons are as follows:

- the L-L algorithm considers the bound of candidate $\mathbf{a}$ as

$$|a_l| \leq \sqrt{1 + \mathrm{SNR}\|\mathbf{h}\|^2}, \tag{3.46}$$

while our complex exhaustive-II considers

$$|a_l| = \mathcal{Q}_{\mathbb{A}}(\alpha h_l) \tag{3.47}$$

$$\leq \mathcal{Q}_{\mathbb{A}}(\sqrt{\mathrm{SNR}} h_l) \tag{3.48}$$

$$\leq \lceil \sqrt{\mathrm{SNR}} h_l \rceil_{\mathbb{A}}, \tag{3.49}$$

where (3.48) is based on Proposition 1 and $\lceil \cdot \rceil_{\mathbb{A}}$ denotes the ceiling operation. Clearly, (3.49) gives a tighter bound than (3.46).

- In section section 3.4.1, we have established that the $\mathcal{S}$-II set in the exhaustive-II is not considered in the L-L algorithm. However, many of the candidates $\mathbf{a}$ generated by $\lfloor \alpha \mathbf{h} \rceil, \alpha \in \mathcal{S}$-II are duplicates of the candidates generated from the set $\mathcal{S}$-I. These duplicates will not participate in the calculation of $\mathcal{R}(\mathbf{h}, \mathbf{a})$. Hence the actual complexity of the exhaustive-II is slightly less than the expression of (3.30).

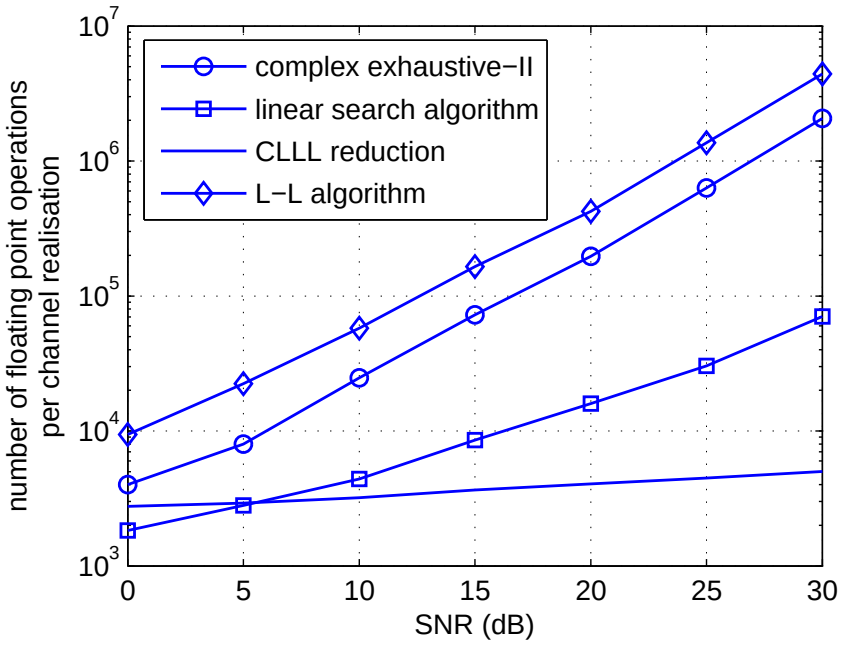

Figure 3.12: Average complexity comparison: complex valued channel, 5 users

As discussed previously, the linear search reduces the complexity from two aspects: 1) optimise Δ to ignore the unnecessary candidates; 2) employ online process to narrow the valid range of α. We can see that the linear search has less complexity than the L-L and exhaustive-II in all figures. Since the the complexity of the linear search varies (see Table 3.2), the gap increases as with SNR. It can be also observed form Table 3.2 that for a given

SNR, γ_{thre} decreases monotonically as L increases, hence the complexity advantage of the linear search becomes less significant for large L, as shown in Figure 3.14. However, the linear search algorithm is proposed for the application that all users have the same fading statistic ($h_l \sim \mathcal{CN}(0,1), \forall l$). In a practical distributed MIMO scenario, where the large scale fading is considered, it is not usual to have more than 20 users with the same level of signal power.

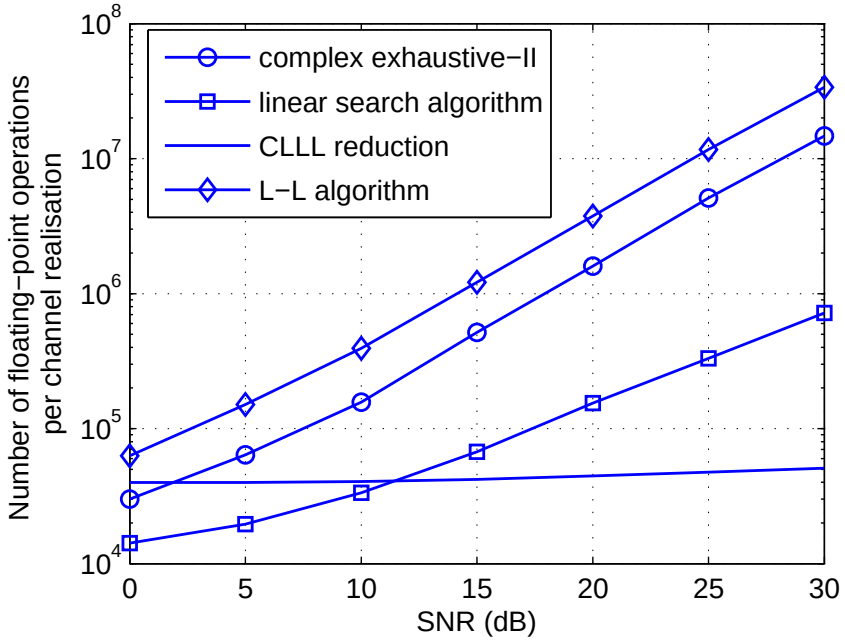

Figure 3.13: Average complexity comparison: complex valued channel, 10 users

The comparison of the LLL and the other three is a tradeoff between L and SNR. In the high SNR region, the LLL algorithm has the complexity advantage while for a large number of users, our proposed algorithms have less complexity.

3.6 Concluding Remarks

In this Chapter, we have given novel coefficient selection algorithms for C&F in both real and complex valued scenarios. A tighter bound of the candidate vectors is derived by considering the valid range of the scaling factor, and this valid range is further reduced by employing online process. We have determined threshold to discard (ignore) the 'narrow' intervals and 'small size' Voronoi regions for the real and complex valued case respectively. The complex-exhaustive-II algorithms extended the idea of interval partition to

Voronoi region partition to ensure the acquired coefficients are optimal. We have shown the theoretical complexity for both algorithms. Numerical comparisons with other existing algorithms are also given. We have shown all of our proposed approaches have good performance-complexity tradeoff.

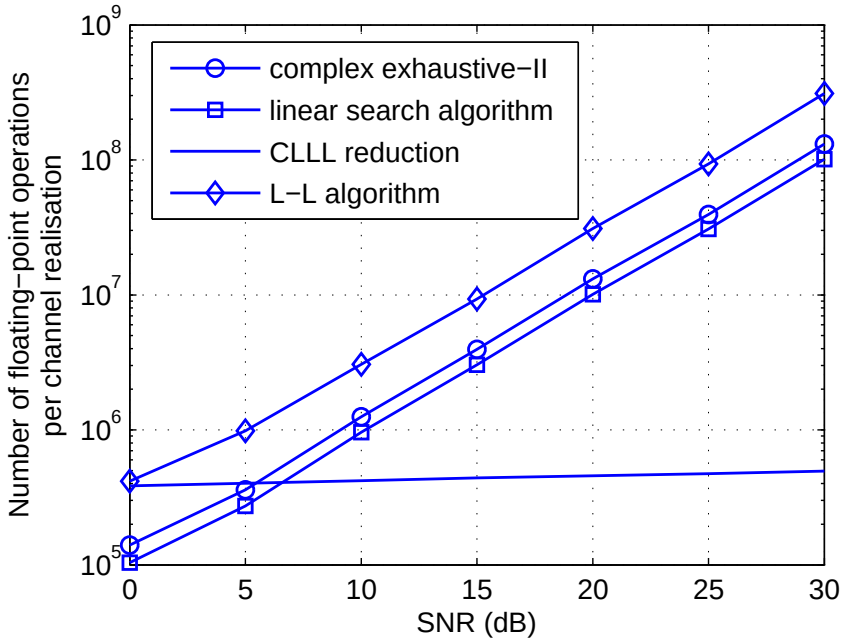

Figure 3.14: Average complexity comparison: complex valued channel, 20 users

Chapter 4

Lattice Network Coding in distributed Massive MIMO: Good Performance with Minimum Backhaul Load

Contents

4.1 Introduction

Previously, we have reviewed massive MIMO as a promising technique to meet the capacity density requirement in 5G wireless. By increasing the ratio of BS antennas to users, wireless networks tend to a quasi-orthogonal, 'interference free' state [2, 14]. Recently, distributed massive MIMO has attracted a lot interest [64–66]. Compared to collocated massive MIMO, the distributed version brings the APs much closer to the the 'cell edge' users, which leads to a uniformly good service for all users. A traditional way to perform

such distributed massive MIMO is by means of a small-cell [67] deployment, where users benefit from selection between denser APs. Recently, the authors in [5, 19] proposed a 'cell-free' model, where a large number of randomly distributed single antenna APs serve a much smaller number of users using the same time-frequency resource, hence there is no cell partition. It is proved [5, 19] that by employing simple maximal ratio combining (MRC), cell free massive MIMO gives better performance for the 'cell edge' users compared to the small cell scheme. Further improvement can be attained by employing zero forcing based algorithms for both uplink [68] and downlink [69] scenarios.

However, the existing performance evaluations of the distributed model are all based on infinite backhaul capacity, which is clearly infeasible in practice. Previously we have introduced C&F as an efficient approach to reduce the backhaul load. In Chapter 3, we have studied the coefficient selection algorithms for a single AP. In this Chapter, we will investigate the performance of both C&F and integer forcing in a distributed massive MIMO model from the following aspects.

- Coefficient selection for large scale systems.
- Different deployment of C&F: distributed and partially distributed.
- The effect of realistic fading parameters on C&F, such as pathloss, shadowing etc.
- The comparison of C&F to other MU-MIMO detection strategies.

The research into C&F so far mainly focusses on two scenarios: general MAC models and small scale MU-MIMO systems. The former corresponds to the local selection, and the corresponding low-complexity algorithms for large L have been studied in the last Chapter. For the latter scenario, the original paper of C&F [28] studied a 2×2 case, and evaluated the probability of rank-deficiency. Recall that the standard C&F considers $L = M$, hence each AP usually needs to forward a set of candidates for $\mathbf{a}_m$ to reduce the probability of such rank-defciency [60]. For the case of $L < M$, a greedy algortihm [49] and a game-theory based algorithm [70] can be applied to the AP selection to maximise the system throughput, however, both [49] and [70] consider a small number of users and single antenna APs. For the case of multiple-antenna APs, integer-forcing can be applied to each AP, and forwards multiple linearly combined data streams (although IF is

originally proposed as a centralised scheme). For simplicity, both C&F and IF are referred to as lattice network coding (LNC).

The coefficent selection for lattice network coding greatly depends on the channel state itself, and i.i.d Gaussian and Rayleigh distributed variables are usually utilised as the baseline channel model. Beyond the baseline model, [49] employs a Bernoulli distributed variable to simulate the effect of shadow fading. However, more comprehensive realistic factors, such as the free space pathloss and the shadowing correlation are not considered in [49].

It has been proved that with sufficiently high bandwidth backhaul, the area spectrum efficiency increases as the distribution factor increases [65, 66]. However, very little literature focusses on the restricted bandwidth backhaul scenario, particularly for a large number of users. In this case, the system throughput is dominated by the interference rather than the thermal noise, and whether the minimum cardinality strategy C&F is still able to effectively cancel the interference is unknown.

Since the primary advantage of distributed massive MIMO is to provide uniformly good service for all users, we focus on the symmetric scenario where all users transmit with a common rate. The main contributions of this Chapter are as follows:

- Exploiting the properties of pathloss, we propose a novel coefficient selection algorithm to reduce the complexity for both C&F and IF.
- We propose a greedy algorithm to select the globally optimal integer coefficients.
- We study the performance of LNC in distributed massive MIMO systems from three perspectives: 1) the probability of rank deficiency. 2) the outage probability for a given rate. 3) the achievable throughput.
- We provide a comprehensive comparison between LNC, small cells, MRC and centralised MMSE. We show that LNC achieves better performance than small cells and MRC, and outperforms the centralised MMSE for some applications. Their respective complexities and required backhaul are also discussed.

The chapter is organised as follows. We introduce the wrap-around distributed model

and the benchmarks in Section 4.2. Then we propose low complexity algorithms for selecting both locally and globally optimal coefficients in Section 4.3. Numerical results and discussions are given in Section 4.4. We conclude the chapter in Section 4.5.

4.2 Preliminaries

4.2.1 System Model

We consider an uplink distributed massive MIMO scheme which comprises access links and backhaul links. There are M APs and L ($M > L$) users randomly deployed in a large square with size $D\text{km} \times D\text{km}$, and all devices share the same time-frequency resource. We make two changes to the original cell free model [5, 19]: 1) Each AP is equipped with $N_\text{r} \geq 1$ antennas. 2) Limited backhaul links are utilised.

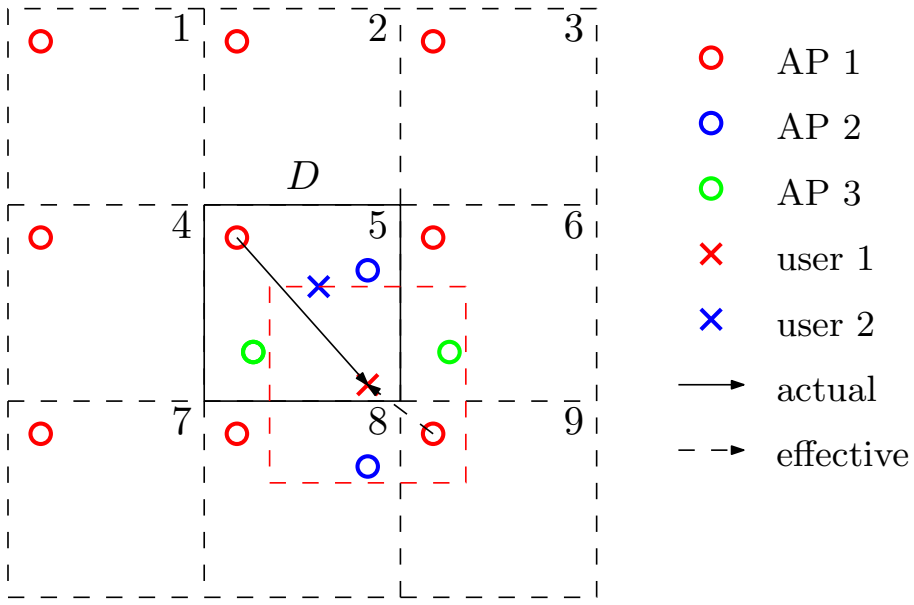

Figure 4.1: System model of distributed massive MIMO with wrap around topology

We consider a wrap-around topology [5,19,71] to avoid the boundary effect. As shown in Figure 4.1, there exists an actual cell in the middle and some mirror cells surrounding it. The mirror cells have exactly the same topology as the actual cell. It can be observed that the cell edge user 1 is far away from AP 1 in the actual cell. However, there exists an effective AP 1 in cell 9 which is much nearer to user 1, and hence the effective distance is

significantly reduced (as denoted by the solid and dashed arrows). Letting the location of actual AP m be the origin point, the set of equivalent APs can be represented as a lattice generated by $\Lambda = D(a + bi)$, where $a, b \in \mathbb{Z}$. Thus, the effective distance between AP m and user l can be simply determined by

$$\mathrm{Loc}(\mathrm{user_l}) - D \times \mathcal{Q}_{\mathbb{Z}[i]}\Big(\frac{\mathrm{Loc}(\mathrm{user}_l) - \mathrm{Loc}(\mathrm{AP}_m)}{D}\Big), \tag{4.1}$$

where $\mathrm{Loc}(\cdot)$ denotes the corresponding location in the actual cell. Similarly, we obtain the effective AP 2 and 3 in cell 8 and 6 respectively. Such wrap-around model enables the effective serving area of user l can be regarded as a D width square with user l in the centre (marked by the red dashed square), and hence the boundary effect is eliminated.

We still use $\mathbf{x}_l \in \mathbb{C}^{n\times 1}$ and $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_L]^T$ to denote the codewords and signal matrix respectively, with $\mathbb{E}[\|\mathbf{x}_l\|]^2 \leq nP$. The channel vector $\mathbf{g}_{m,l} \in \mathbb{C}^{N_\mathrm{r}\times 1}$ between the mth AP and the lth user is expressed as

$$\mathbf{g}_{m,l} = \beta_{m,l}^{1/2}\mathbf{h}_{m,l}, \tag{4.2}$$

where $\beta_{m,l}$ and $\mathbf{h}_{m,l}$ denote the large scale fading and the small scale fading respectively. Each component in $\mathbf{h}_{m,l}$ is i.i.d Rayleigh distributed with unit variance. $\beta_{m,l}$ is a scalar for a given AP-user pair regardless of the number of antennas, and expressed as

$$\beta = \mathrm{PL}_{m,l}\mathrm{SH}_{m,l}, \tag{4.3}$$

where $\mathrm{PL}_{m,l}$ and $\mathrm{SH}_{m,l}$ respectively denote the pathloss and shadow fading, and the calculation of both parameters will be described in section 4.4. We assume the local channel matrix $\mathbf{G}_m = [\mathbf{g}_{m,1}, \cdots, \mathbf{g}_{m,L}]$ is perfectly known at each AP m. The received signal of the mth AP can be expressed as

$$\mathbf{Y}_m = \mathbf{G}_m\mathbf{X}^T + \mathbf{Z}_m, \quad \mathbf{Y}_m \in \mathbb{C}^{N_\mathrm{r}\times n}, \tag{4.4}$$

where each noise component in $\mathbf{Z}_m \in \mathbb{C}^{N_\mathrm{r}\times n}$ has variance σ^2, and $\mathrm{SNR} = P/\sigma^2$.

As previously mentioned, we consider two applications of LNC in this Chapter: C&F in fully distributed models and IF in partial distributed models. To be fair, we assume the

same total number of AP antennas for different deployments, denoted as $N_{\mathrm{r,total}}$. For both cases, each AP selects the locally optimal F integer vectors $\mathbf{a}_m$ according to

$$\begin{aligned}\mathbf{a}_{m,\mathrm{opt}} &= \underset{\mathbf{a}_m \in \mathbb{A}^{L\times 1}\setminus\{\mathbf{0}\}}{\operatorname{argmax}} \mathcal{R}_{m,\mathrm{LNC}}(\mathbf{G}_m, \mathbf{a}_m) \\ &= \underset{\mathbf{a}_m \in \mathbb{A}^{L\times 1}\setminus\{\mathbf{0}\}}{\operatorname{argmax}} \log^+\Big(\frac{1}{\mathbf{a}^H\mathbf{M}\mathbf{a}}\Big),\end{aligned} \tag{4.5}$$

where

$$\mathbf{M} = \mathbf{I}_L - \mathrm{SNR}\mathbf{G}_m^H(\mathbf{I}_{N_\mathrm{r}} + \mathrm{SNR}\mathbf{G}_m\mathbf{G}_m^H)^{-1}\mathbf{G}_m. \tag{4.6}$$

The locally selected coefficients $\mathbf{A}_m = [\mathbf{a}_{m,1}, \cdots, \mathbf{a}_{m,F}]^T \in \mathbb{A}^{F\times L}$ are then forwarded to the hub, and the hub performs global optimisation to select L from the MF vectors.

4.2.2 Benchmarks

In this section, we introduce some existing strategies as benchmarks. To make a comprehensive comparison, our benchmarks include both finite and infinite backhaul techniques, as well as the centralised strategies.

Centralised Massive MIMO with MMSE Equalisation

In the collocated scenario, the global channel matrix $\mathbf{G}^{N_\mathrm{r}\times L}$ is available at the AP. It is well studied that zero-forcing and MMSE both achieve near-optimal performance in massive MIMO. We consider the MMSE equaliser $\mathbf{B}_{\mathrm{MMSE}}^{L\times N_\mathrm{r}}$ as a benchmark, expressed as

$$\mathbf{B}_{\mathrm{MMSE}} = \frac{\mathrm{SNR}\mathbf{G}^H}{\mathbf{I}_{N_\mathrm{r}} + \mathrm{SNR}\mathbf{G}\mathbf{G}^H}. \tag{4.7}$$

According to the equalised signal matrix $\hat{\mathbf{Y}} = \mathbf{B}\mathbf{Y}$, the achievable rate for user l can be expressed as

$$\mathcal{R}_{l,\mathrm{MMSE}} = \log\Big(1 + \frac{\mathrm{SNR}|\mathbf{b}_l^T\mathbf{g}_l|^2}{\mathrm{SNR}\sum_{l'\neq l}|\mathbf{b}_l^T\mathbf{g}_{l'}|^2 + \|\mathbf{b}_l\|^2}\Big), \tag{4.8}$$

where $\mathbf{b}_l^T$ and $\mathbf{g}_l$ denote the lth row in $\mathbf{B}_{\mathrm{MMSE}}$ and lth column in $\mathbf{G}$ respectively. Since we focus on the symmetric scenario where all users transmit with a common rate, the system

throughput per user is determined by the worst user, denoted as

$$\mathcal{R}_{\mathrm{MMSE}} = \min_{l=1,\cdots,L} \mathcal{R}_{l,\mathrm{MMSE}}. \tag{4.9}$$

Small Cells

As for C&F, the small cell scheme is also a minimum backhaul strategy. We consider the baseline non-cooperative small cell scheme in which each user is decoded and forwarded by one AP only. The corresponding AP for the lth user is selected based on the strength of the channel vector $\mathbf{g}_{m,l}$

$$\underset{\text{available } m}{\operatorname{argmax}} \ \|\mathbf{g}_{m,l}\|^2. \tag{4.10}$$

We assume AP selection is performed user by user with random priorities. To minimise latency, an AP is selected by at most N_{r} users. If an AP has been chosen by N_{r} users, this AP becomes unavailable. For the case of $N_{\mathrm{r}} = 1$, users are served by APs in a one-to-one manner. Let m_l denote the index of the AP allocated to the lth user, then its achievable rate can be expressed as

$$\mathcal{R}_{l,\mathrm{SC}} = \log\Big(1 + \frac{\mathrm{SNR}|g_{m_l,l}|^2}{1 + \mathrm{SNR}\sum_{l' \neq l} |\mathbf{g}_{m_l,l'}|^2}\Big), \quad N_{\mathrm{r}} = 1. \tag{4.11}$$

For the case of $N_{\mathrm{r}} > 1$, each AP might be selected by more than one users. We assume MMSE equalisation and parallel decoding are performed at each AP. Given the local channel matrix $\mathbf{G}_{m_l}$, the achievable rate of user l is expressed as

$$\mathcal{R}_{l,\mathrm{SC}} = \log\Big(1 + \frac{\mathrm{SNR}\|\mathbf{b}_{m_l,l}^T \mathbf{g}_{m_l,l}\|^2}{\mathrm{SNR}\sum_{l' \neq l} \|\mathbf{b}_{m_l,l}^T \mathbf{g}_{m_l,l'}\|^2 + \|\mathbf{b}_{m_l,l}\|^2}\Big), \quad N_{\mathrm{r}} > 1, \tag{4.12}$$

where $\mathbf{b}_{m_l,l}^T$ denotes the lth row in $\mathbf{B}_{m_l,\mathrm{MMSE}} = \frac{\mathrm{SNR}\mathbf{G}_{m_l}^H}{\mathbf{I}_{N_{\mathrm{r}}} + \mathrm{SNR}\mathbf{G}_{m_l}\mathbf{G}_{m_l}^H}$. Again, the achievable rate per user is determined by the worst user, denoted as

$$\mathcal{R}_{\mathrm{SC}} = \min_{l=1,\cdots,L} \mathcal{R}_{l,\mathrm{SC}}. \tag{4.13}$$

Note that in some works in the literature, the AP selection is based on large scale fading. However, the coefficient selection in C&F is performed during the coherence time of small

scale fading, hence we assume the AP allocation in the small cell system is also performed during the small scale fading coherence time, in order to provide a fair comparison.

Maximum Ratio Combining (MRC)

In the original paper of cell free massive MIMO [5, 19], the received signal at the mth AP is multiplied by the conjugate transpose of the local channel matrix $\mathbf{G}_m$; then the precise signal matrix of $\mathbf{Y}_{m,\mathrm{MRC}} = \mathbf{G}_m^H \mathbf{Y}_m$ is forwarded to the hub via infinite backhaul. The hub combines the equalised signals from all APs, denoted as

$$\mathbf{Y}_{\mathrm{MRC}} = \sum_m \mathbf{G}_m^H \mathbf{Y}_m = \mathbf{G}^H \mathbf{Y}, \tag{4.14}$$

where $\mathbf{G} \in \mathbb{C}^{MN_{\mathrm{r}} \times L} = [\mathbf{G}_1^T, \cdots, \mathbf{G}_M^T]^T$ denotes the global channel matrix. Let $\mathbf{g}_l$ denote the lth column of $\mathbf{G}$, then the achievable rate of the lth user can be expressed as

$$\mathcal{R}_{l,\mathrm{MRC}} = \log\Big(1 + \frac{\mathrm{SNR}|\mathbf{g}_l^H \mathbf{g}_l|^2}{\|\mathbf{g}_l^H\|^2 + \mathrm{SNR}\sum_{l' \neq l} |\mathbf{g}_l^H \mathbf{g}_{l'}|^2}\Big), \tag{4.15}$$

and the system rate per user is

$$\mathcal{R}_{\mathrm{MRC}} = \min_{l=1,\cdots,L} \mathcal{R}_{l,\mathrm{MRC}}. \tag{4.16}$$

4.3 Lattice Network Coding in Distributed Massive MIMO

In this section, we propose low complexity algorithms for selecting both locally and globally optimal coefficients, and then provide a qualitative comparison with the benchmarks.

4.3.1 Local Selection

Previously we have stated that the computation rate in C&F can be expressed as a function of the scaling factor α. It can be easily proved that the statement still holds for the IF scenario ($N_{\mathrm{r}} > 1$). For a given scaling vector $\mathbf{b}_m$, the optimal integer vector $\mathbf{a}$ can also be expressed as $\mathcal{Q}_{\mathbb{A}}(\mathbf{b}_m^T \mathbf{G}_m)$, and hence the optimisation problem in (4.5) is equivalent to

$$\begin{aligned} \mathbf{b}_{m,\mathrm{opt}} &= \underset{\mathbf{b}_m \in \mathbb{C}^{N_{\mathrm{t}} \times 1}}{\operatorname{argmax}} \mathcal{R}(\mathbf{G}_m, \mathbf{b}_m) \\ &= \underset{\mathbf{b}_m \in \mathbb{C}^{N_{\mathrm{r}} \times 1}}{\operatorname{argmax}} \log^+\Big(\frac{P}{\|\mathbf{b}_m\|^2 \sigma^2 + \|\mathbf{G}_m^T \mathbf{b}_m - \mathcal{Q}_{\mathbb{A}}(\mathbf{G}_m^T \mathbf{b}_m)\|^2 P}\Big). \end{aligned} \tag{4.17}$$

Note that the quantisation is applied to each component of $\mathbf{G}_m^T \mathbf{b}_m$, denoted as $a_{m,l} = \mathcal{Q}_{\mathbb{A}}(\mathbf{g}_{m,l}^T \mathbf{b}_m)$.

The local selection described in (4.5) and (4.17) are based on a MAC with L users. Now we propose a simple algorithm to reduce the network size. Similar to C&F, we can easily obtain a bound for the scaling vector $\mathbf{b}_m$ from (4.17), denoted as

$$\|\mathbf{b}_{m,\mathrm{opt}}\| < \sqrt{P/\sigma^2} = \sqrt{\mathrm{SNR}}. \tag{4.18}$$

For the users which are located far away from the mth AP, particularly the strength of their corresponding channel vectors $\|\mathbf{g}_{m,l}\| \leq \frac{1}{2\sqrt{\mathrm{SNR}}}$, we have

$$\|\mathbf{g}_{m,l}^T \mathbf{b}_m\| < \|\mathbf{g}_m\| \sqrt{\mathrm{SNR}} \leq \frac{1}{2}, \tag{4.19}$$

which means that the corresponding $a_{m,l} = \mathcal{Q}_{\mathbb{A}}(\mathbf{g}_{m,l}^T \mathbf{b}_m)$ for these users are definitely zero, no matter what $\mathbf{b}_m$ is selected. Therefore, these users are not able to contribute to the linear equations of the mth AP, can be regarded as trivial users. Hence the L-dimensional MAC is reduced to $L_{\mathrm{eff},m}$-dimensions, expressed as

$$L_{\mathrm{eff},m} = |\Phi_m|, \ \Phi_m = \{\mathbf{g}_{m,l} : \ \|\mathbf{g}_{m,l}\| > \frac{1}{2\sqrt{\mathrm{SNR}}}\}, \tag{4.20}$$

where $L_{\mathrm{eff},m}$ denotes the number of effective users, and Φ_m denotes the set of effective channel vectors.

C&F in Reduced Network

The trivial users can be simply treated as interference, which adds to the thermal noise. In C&F, the channel vector $\mathbf{g}_{m,l}$ is actually a scalar $g_{m,l}$, which results in the effective SNR is

$$\mathrm{SNR}_{\mathrm{eff},m} = \frac{P}{\sigma^2 + P\sum_{g_{m,l}\notin\Phi_m}|g_{m,l}|^2}, \tag{4.21}$$

with the updated SNR, we repeat the operation in (4.20) to further reduce the elements in Φ until convergence. Then, the complex-exhaustive-II algorithm can be directly applied to the finally reduced network.

IF in Reduced Network

Although the complex-exhaustive-II and the L-L algorithms can be extended to IF, the complexity for both methods increases exponentially with N_{r}. Thus, we apply the sub-optimal complex-LLL algorithm to the size reduced network for IF. Let $\mathbf{G}_{\mathrm{eff},m} \in \mathbb{C}^{N_{\mathrm{r}}\times L_{\mathrm{eff}}}$ denote the effective channel which comprises $\mathbf{g}_{m,l} \in \Phi_m$, and $\mathbf{G}_{\mathrm{tri},m} \in \mathbb{C}^{N_{\mathrm{r}}\times(L-L_{\mathrm{eff}})}$ as the trivial components. Let $\mathbf{a}_{\mathrm{eff},m} \in \mathbb{A}^{L_{\mathrm{eff}}\times 1}$ denote the integer vector of the reduced network. Hence, the $\sigma^2_{\mathrm{eff},m}$ term in (4.17) can be rewritten as

$$\begin{aligned}
\sigma^2_{\mathrm{eff},m} &= \|\mathbf{b}_m\|^2\sigma^2 + P\|\mathbf{G}_m^T\mathbf{b}_m - \mathbf{a}_m\|^2 \\
&= \sigma^2\mathbf{b}_m^H\mathbf{b}_m + P\|\mathbf{G}_{\mathrm{eff},m}^T\mathbf{b}_m - \mathbf{a}_{\mathrm{eff},m}\|^2 + P\|\mathbf{G}_{\mathrm{tri},m}\mathbf{b}_m\|^2 \\
&= \sigma^2\mathbf{b}_m^H\mathbf{b}_m + P(\mathbf{b}_m^H\mathbf{G}_{\mathrm{eff},m}\mathbf{G}_{\mathrm{eff},m}^H\mathbf{b}_m - 2\mathbf{b}_m^T\mathbf{G}_{\mathrm{eff},m}\mathbf{a}_{\mathrm{eff},m} + \mathbf{a}_{\mathrm{eff},m}^H\mathbf{a}_{\mathrm{eff},m} + \mathbf{b}_m^H\mathbf{G}_{\mathrm{tri},m}\mathbf{G}_{\mathrm{tri},m}^H\mathbf{b}_m) \\
&= \mathbf{b}_m^H\big(\sigma^2\mathbf{I}_{N_{\mathrm{r}}} + P(\mathbf{G}_{\mathrm{eff},m}\mathbf{G}_{\mathrm{eff},m}^H + \mathbf{G}_{\mathrm{tri},m}\mathbf{G}_{\mathrm{tri},m}^H)\big)\mathbf{b}_m - 2P\mathbf{b}_m^T\mathbf{G}_{\mathrm{eff},m}\mathbf{a}_{\mathrm{eff},m} + P\mathbf{a}_{\mathrm{eff},m}^H\mathbf{a}_{\mathrm{eff},m} \\
&= \mathbf{b}_m^H(\sigma^2\mathbf{I}_{N_{\mathrm{r}}} + P\mathbf{G}_m\mathbf{G}_m^H)\mathbf{b}_m - 2P\mathbf{b}_m^T\mathbf{G}_{\mathrm{eff},m}\mathbf{a}_{\mathrm{eff},m} + P\mathbf{a}_{\mathrm{eff},m}^H\mathbf{a}_{\mathrm{eff},m}.
\end{aligned} \tag{4.22}$$

The 1st order derivative of $\sigma^2_{\mathrm{eff},m}$ with respect to $\mathbf{b}_m$ is

$$\frac{d\sigma^2_{\mathrm{eff},m}}{d\mathbf{b}_m} = 2(\sigma^2\mathbf{I}_{N_{\mathrm{r}}} + P\mathbf{G}_m\mathbf{G}_m^H)\mathbf{b}_m - 2P\mathbf{G}_{\mathrm{eff},m}\mathbf{a}_{\mathrm{eff},m}, \tag{4.23}$$

hence the MMSE solution of $\mathbf{b}_m$ for given $\mathbf{a}_{\mathrm{eff},m}$ is

$$\mathbf{b}_{\mathrm{opt},m}^T(\mathbf{a}_{\mathrm{eff},m}) = \frac{\mathrm{SNR}\mathbf{a}_{\mathrm{eff},m}\mathbf{G}_{\mathrm{eff},m}^H}{\mathbf{I}_{N_{\mathrm{r}}} + \mathrm{SNR}\mathbf{G}_m\mathbf{G}_m^H}. \tag{4.24}$$

We combine (4.22) and (4.24) to obtain

$$\sigma^2_{\text{opt},m} = \text{SNR}\mathbf{a}^H_{\text{eff},m}\big(\mathbf{I}_{L_{\text{eff},m}} - \text{SNR}\mathbf{G}^H_{\text{eff},m}(\mathbf{I}_{N_\text{r}} + \text{SNR}\mathbf{G}_m\mathbf{G}_m^H)^{-1}\mathbf{G}_{\text{eff},m}\big)\mathbf{a}_{\text{eff},m}. \quad (4.25)$$

Hence, the optimisation in (4.5) can be rewritten as

$$\begin{aligned}\mathbf{a}_{\text{eff},m,\text{opt}} &= \underset{\mathbf{a}_{\text{eff},m}\in\mathbb{A}^{L_\text{eff}\times 1}}{\text{argmax}} \log^+\Big(\frac{P}{\sigma^2_{\text{eff},m}}\Big) \\ &= \underset{\mathbf{a}_{\text{eff},m}\in\mathbb{A}^{L_\text{eff}\times 1}}{\text{argmax}} \log^+\Big(\frac{1}{\mathbf{a}^H_{\text{eff},m}\mathbf{M}_{\text{eff},m}\mathbf{a}_{\text{eff},m}}\Big),\end{aligned} \quad (4.26)$$

where $\mathbf{M}_{\text{eff},m} = \mathbf{I}_{L_{\text{eff},m}} - \text{SNR}\mathbf{G}^H_{\text{eff},m}(\mathbf{I}_{N_\text{r}} + \text{SNR}\mathbf{G}_m\mathbf{G}_m^H)^{-1}\mathbf{G}_{\text{eff},m}$ is an $L_{\text{eff},m}$ dimensional square matrix which is positive definite for any finite SNR, and hence it has a unique Cholesky decomposition $\mathbf{M}_{\text{eff},m} = \mathbf{L}_{\text{eff},m}\mathbf{L}^H_{\text{eff},m}$. We employ the CLLL algorithm to find F locally optimal $\mathbf{a}_{\text{eff},m,f}$ which minimise $\|\mathbf{L}^H_{\text{eff},m}\mathbf{a}_{\text{eff},m}\|$, then the corresponding $\mathbf{a}_{m,f}$ can be acquired immediately by zero-padding the $L - L_\text{eff}$ trivial components. Clearly, the size-reduced network enables the lattice reduction to be performed with a lower dimensional basis, which reduces the complexity. The procedure of the local selection with the size-reduced network is outlined in Algorithm 5.

Algorithm 5 Local Selection with Size-Reduced Network

Input: local channel vector $\mathbf{G}_m$, SNR
Output: F optimal integer vectors $\mathbf{a}_{m,f}$

Compute and Forward with $N_\text{r} = 1$:

$\Phi_m = \{g_{m,1}, g_{m,2}, \cdots, g_{m,L}\}$, $L_\text{eff} = L$, $\text{SNR}_\text{eff} = \text{SNR}$, $\Delta\Phi_m = \Phi_m$
while $\Delta\Phi_m \neq \emptyset$ **do**
  $\Delta\Phi_m = \{g_{m,l} : g_{m,l} \in \Phi_m \text{ and } |g_{m,l}| \leq \frac{1}{2\sqrt{\text{SNR}_\text{eff}}}\}$
  $\Phi_m = \Phi_m \setminus \Delta\Phi_m$
  $\text{SNR}_{\text{eff},m} = \dfrac{\text{SNR}}{1 + \text{SNR}\sum_{g_{m,l}\notin\Phi_m}|g_{m,l}|^2}$
end while
Based on the effective channel $\mathbf{g}_m \in \mathbb{C}^{L_\text{eff}}$ with $g_{m,l} \in \Phi_m$, and the effective SNR_eff, employ complex-exhaustive-II algorithm to obtain $\mathbf{a}_{m,f}$.

Integer Forcing with $N_\text{r} > 1$:

$\Phi_m = \{\mathbf{g}_{m,l} : \|\mathbf{g}_{m,l}\| > \frac{1}{2\sqrt{\text{SNR}}}\}$, $L_{\text{eff},m} = |\Phi_m|$.
Based on the effective channel $\mathbf{G}_{\text{eff},m} \in \mathbb{C}^{N_\text{r}\times L}$, employ CLLL to obtain $\mathbf{a}_{\text{eff},m,f}$ according to equation (4.26).
$\mathbf{a}_{m,f} = \mathbf{a}_{\text{eff},m,f} \cup \{\mathbf{0}\}^{L-L_\text{eff}}$

4.3.2 Global Selection

Each AP forwards the locally selected coefficients $\mathbf{A}_m = \{\mathbf{a}_{m,f}\}$ and their corresponding $\mathcal{R}(\mathbf{a}_{m,f})$ to the hub, and hence the hub has an integer matrix $\mathbf{A} = [\mathbf{A}_1^T, \cdots, \mathbf{A}_M^T]^T$. Let $J = MF$ denote the total number of equations in $\mathbf{A}$. We sort these J equations in descending order according to their corresponding computation rates $\mathcal{R}_{\mathrm{LNC}}(\mathbf{a}_{m,f})$, which results in an equivalent $J \times L$ matrix $\hat{\mathbf{A}} = \{\hat{\mathbf{a}}_j\}$, where $\hat{\mathbf{a}}_j^T$ denotes the jth row in $\hat{\mathbf{A}}$. Let m_j denote the index of the source AP of $\hat{\mathbf{a}}_j$, which means the candidate vector $\hat{\mathbf{a}}_j$ is forwarded by the m_j th AP, denoted as $\mathbf{a}_j \in \mathbf{A}_{m_j}$.

The hub intends to select an $L \times L$ submatrix $\hat{\mathbf{A}}_{\mathrm{sub}}$ which is formed by taking rows of $\hat{\mathbf{A}}$ with indices in $\mathcal{S} \subset \{1, 2, \cdots, J\}$, with $|\mathcal{S}| = L$. Let $\Omega_m = \{\hat{\mathbf{a}}_j : m_j = m,\ j \in \mathcal{S}\}$ denote the globally selected equations from $\mathbf{A}_m$, and $F_m = |\Omega_m|$ be the number of equations in Ω_m. Clearly, $F_m \leq F$ for all m and $\sum F_m = L$. Note that Ω_m is also the candidate set to be fed back to the mth AP. The values of F_m do not have to be the same for all m, which means that some APs are 'passive' during the backhaul transmission if $\Omega_m = \emptyset$. As previously mentioned, the globally optimal sub-matrix $\hat{\mathbf{A}}_{\mathrm{sub,opt}}$ has to be full rank, and the computation rate of the worst row (worst equation) in $\hat{\mathbf{A}}_{\mathrm{sub}}$ should be maximised. This max-min optimisation problem is formally expressed as

$$\underset{\hat{\mathbf{A}}_{\mathrm{sub,opt}}}{\text{maximise}} \quad \min_{j \in \mathcal{S}} \mathcal{R}_{\mathrm{LNC}}(\hat{\mathbf{a}}_j) \tag{4.27}$$

$$\text{subject to} \quad \mathrm{Rank}(\hat{\mathbf{A}}_{\mathrm{sub}}) = |\mathcal{S}| = L \tag{4.28}$$

$$F_m \leq N_{\mathrm{r}},\ \forall m. \tag{4.29}$$

Rigorously, the rank constraint in (4.28) should be $\mathrm{Rank}(\mathbf{Q}) = L$, where $\mathbf{Q} = \hat{\mathbf{A}}_{\mathrm{sub}} \bmod \pi$. However, the assessment of $\mathbf{Q}$ relies on a specific finite field $\mathbb{A}/\pi\mathbb{A}$. Hence the integer matrix over $\mathbb{A}$ is commonly used instead of $\mathbf{Q}$ for the performance evaluation in a general case. The constraint (4.29) is due to the fact that the maximum number of data streams that can be simultaneously transmitted via the backhaul is limited by N_{r} (although each AP is able to decode more than N_{r} streams). When $F_m > N_{\mathrm{r}}$, the F_m decoded linear combinations $\mathcal{D}(\hat{\mathbf{a}}_j^T \mathbf{X}), \hat{\mathbf{a}}_j \in \Omega_m$ require more than 1 time slot to be transmitted, which results in unnecessary latency.

We propose a simple greedy algorithm to solve the problem above, and the optimality of $\hat{\mathbf{A}}_{\text{sub,opt}}$ is ensured. Since the equations in $\hat{\mathbf{A}}$ are sorted in descending order, we set the first row (highest rate equation) $\hat{\mathbf{a}}_1^T$ as the initial $\hat{\mathbf{A}}_{\text{sub,opt}}$, and then add $\hat{\mathbf{a}}_j^T$ to $\hat{\mathbf{A}}_{\text{sub,opt}}$ one by one. For each $\hat{\mathbf{a}}_j^T$, if its presence increases the rank of the current $\hat{\mathbf{A}}_{\text{sub,opt}}$, then we keep it and save it to Ω_m with $m = m_j$, otherwise, we discard it. Note that when $|\Omega_m|$ is already equal to N_{r}, the remaining $\hat{\mathbf{a}}_j \in \mathbf{A}_m$ would be directly skipped. The procedure terminates when $\hat{\mathbf{A}}_{\text{sub,opt}}$ meets the rank requirement. Alternatively, we can set $\hat{\mathbf{A}}$ as the initial $\hat{\mathbf{A}}_{\text{sub,opt}}$, and check $\hat{\mathbf{a}}_j$ one by one from the bottom row. For each $\hat{\mathbf{a}}_j$, if its absence would not change the rank, then it can be discarded.

Both the approaches above end up with the same solution, and the same criterion is applied to both methods: whether a row is useful or not depends on whether it can replaced by higher rate equations. The only difference between them is that the former intends to keep L necessary rows while the latter discards $MF - L$ unnecessary rows. Since $MF >> L$ in massive MIMO, we employ the former one, which has lower complexity. The greedy algorithm for the global selection is outlined in Algorithm 6.

Algorithm 6 Greedy Algorithm for Global Selection

Input: $J \times L$ integer matrix $\mathbf{A}$, computation rate $\mathcal{R}_{\text{LNC}}(\mathbf{a}_{m,f})\ \forall m, f$
Output: selected equations to be fed back to APs: $\Omega_m,\ \forall m$

Initialisation : $\Omega_m = \emptyset\ \forall m$, $F_m = 0\ \forall m$, $j = 1$.

1: Sort the equations in $\mathbf{A}$ in descending order to acquire $\hat{\mathbf{A}}$ with respect to their corresponding $\mathcal{R}_{\text{LNC}}(\mathbf{a}_{m,f})$.
2: $\hat{\mathbf{A}}_{\text{sub,opt}} = \hat{\mathbf{a}}_j$, $m = m_j$.
3: $\Omega_m = \Omega_m \cup \hat{\mathbf{a}}_j$ and $F_m = F_m + 1$
4: $j = 2$
5: **while** $\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}}) < L$ and $j \leq J$ **do**
6: $\quad m = m_j$
7: $\quad$ **if** $F_m = N_{\text{r}}$ **then**
8: $\quad\quad$ break if loop and go to line 14.
9: $\quad$ **else if** $\text{Rank}\big([\hat{\mathbf{A}}_{\text{sub,opt}};\ \hat{\mathbf{a}}_j]\big) = \text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}})$ **then**
10: $\quad\quad$ break if loop and go to line 14.
11: $\quad$ **else**
12: $\quad\quad$ $\hat{\mathbf{A}}_{\text{sub,opt}} = [\hat{\mathbf{A}}_{\text{sub,opt}};\ \hat{\mathbf{a}}_j]$; $\Omega_m = \Omega_m \cup \hat{\mathbf{a}}_j$; $F_m = F_m + 1$
13: $\quad$ **end if**
14: $\quad j = j + 1$
15: **end while**
16: Return $\hat{\mathbf{A}}_{\text{sub,opt}}$ and $\Omega_m,\ \forall m$

Note that there are two applications of the algorithm above:

- For closed loop systems, the globally selected coefficients are fed back to the APs at the beginning of each coherence time. Let R_0 denote the common rate for the access links, hence the total backhaul load is LR_0 in this case.

- For open loop systems (feedback unavailable), each AP alternately transmits all F decoded data streams to the hub, and then the hub performs the greedy algorithm to select L equations for message recovery. The total backhaul load is JR_0 in this case.

Since the system throughput is determined only by which equations are selected, the two applications above have the same performance if $F \leq N_{\mathrm{r}}$. The former minimises the latency, while the latter minimises backhaul.

4.3.3 Qualitative Analysis

In this section, we compare LNC to the benchmarks from three aspects: backhaul load, complexity and the system throughput.

Backhaul

Clearly, this is the primary advantage of lattice network coding compared to any other linear processing schemes. Both small cells and LNC employ LR_0 bits per channel use for the backhaul transmissions, which is the theoretically minimum backhaul required to achieve lossless transmission in a distributed system. Note that the extra bits needed to convey the integer coefficients $\mathbf{A}$ and Ω_m are negligible.

Complexity

Compared to the benchmark schemes, the extra complexity of lattice network coding arises from the coefficient selection. In a quasi-static case, this additional complexity is

negligible compared to channel coding and decoding. The number of decoders required is equal to the number of intended data streams, which is L for all schemes.

Throughput

In small cells, each user is served by a unique AP (a unique equation more precisely). In LNC, each equation provided by one AP might involve one or more users, and each user might take part in multiple equations from many APs. Hence we conclude that small cell transmission is a special case of LNC in which only one user is involved in each equation, and we therefore expect LNC to achieve higher throughput than small cells.

MRC enables each user to be served by all APs. Unlike the collocated massive MIMO, the channel strength for a specific user varies at different APs. MRC eliminates inter-user interference only asymptotically, as the ratio of antennas to users tends to infinity. In contrast, provided a full rank matrix is formed, LNC allows all users to be recovered, analogously to zero-forcing, but without noise enhancement. Hence LNC may also outperform MRC.

4.4 Numerical Results and Discussions

In this section, we provide numerical results of LNC performance on distributed massive MIMO systems, and compare it with the benchmarks.

4.4.1 Simulation Setup

We adopt the parameter settings in [5, 19] as the basis to establish our simulation model. The pathloss $\mathbf{PL}_{m,l}(d_{m,l}) \propto d_{m,l}^{-\gamma}$ is a continuous function of $d_{m,l}$ (in km), where $d_{m,l}$ denotes the effective distance calculated by (4.1). The exponent γ is a 'three-slope' variable which relies on $d_{m,l}$. It is equal to 0, 2 and 3.5, for $d_{ml} \leq d_0$, $d_0 < d_{ml} \leq d_1$ and $d_1 < d_{m,l}$ respectively. The Hata-COST231 model [72] is employed to characterise the

propagation, hence $\mathbf{PL}_{m,l}$ are given by [5]

$$\begin{aligned}\mathbf{PL}_{m,l}(d_{m,l} > d_1) &= \mathbf{PL}_{m,l}(1) - 35\log_{10} d_{m,l} \\ \mathbf{PL}_{m,l}(d_0 < d_{m,l} \leq d_1) &= \mathbf{PL}_{m,l}(1) - 35\log_{10} d_1 - 20\log_{10}(\frac{d_{m,l}}{d_1}),\end{aligned} \tag{4.30}$$

where $d_0 = 0.01$km and $d_1 = 0.05$km, and $\mathbf{PL}_{m,l}(1)$ denotes the 1km pathloss which is given by [72]

$$\mathbf{PL}_{m,l}(1) = -46.3 - 33.9\log_{10} f_{\mathrm{c}} + 13.82\log_{10} \mathrm{h}_{\mathrm{AP}} + (1.1\log_{10} f_{\mathrm{c}} - 0.7)\mathrm{h}_{\mathrm{u}} - (1.56\log_{10} f_{\mathrm{c}} - 0.8). \tag{4.31}$$

where f_{c}, h_{AP}, and h_{u} denote the carrier frequency, the height of APs and the height of users respectively.

We consider the shadow fading for both uncorrelated and correlated cases. The uncorrelated shadowing is denoted as

$$\mathbf{S}_{m,l} = 10^{\frac{\sigma_{\mathrm{sh}} c_{m,l}}{10}}, \tag{4.32}$$

where the standard deviation σ_{sh} is set to 8dB [5], and $c_{m,l} \sim \mathcal{N}(0,1)$. For the correlated case, $c_{m,l}$ is denoted as [73]

$$c_{m,l} = \sqrt{0.5}c_m + \sqrt{0.5}c_l \tag{4.33}$$

where $c_m \sim \mathcal{N}(0,1)$ and $c_l \sim \mathcal{N}(0,1)$ denote the shadowing caused by the obstacles in the vicinity of the mth AP and the lth user respectively. $\{c_m\}$ and $\{c_l\}$ are generated according to their covariance functions, denoted as

$$\mathbb{E}\{c_m c_{m'}\} = 2^{-\frac{d_{m,m'}}{0.1}},\ \mathbb{E}\{c_l c_{l'}\} = 2^{-\frac{d_{l,l'}}{0.1}}, \tag{4.34}$$

where $d_{m,m'}$ (in km) denotes the effective distance of the mth and m'th APs, and $d_{l,l'}$ (in km) denotes the effective distance of the lth and l'th users. Clearly, the correlation increases monotonically as the distance decreases. For the extreme case, the same geographical locations result in the same shadowing.

In all examples, we assume i.i.d small scale fading with $h_{ml} \sim \mathcal{CN}(0,1)$, and equal

power allocation for all users. To maintain the comparability to the original work of cell-free massive MIMO, we choose the following parameters as same as in [5]. The carrier frequency $f_{\mathrm{c}} = 1900$MHz. The height of APs and user are $\mathrm{h_{AP}} = 15$m and $\mathrm{h_u} = 1.65$m respectively. The system bandwidth is 20MHz, and the transmit power and the thermal noise density are 200mW and -174dBm/Hz respectively.

4.4.2 Complexity Reduction of the Local Selection

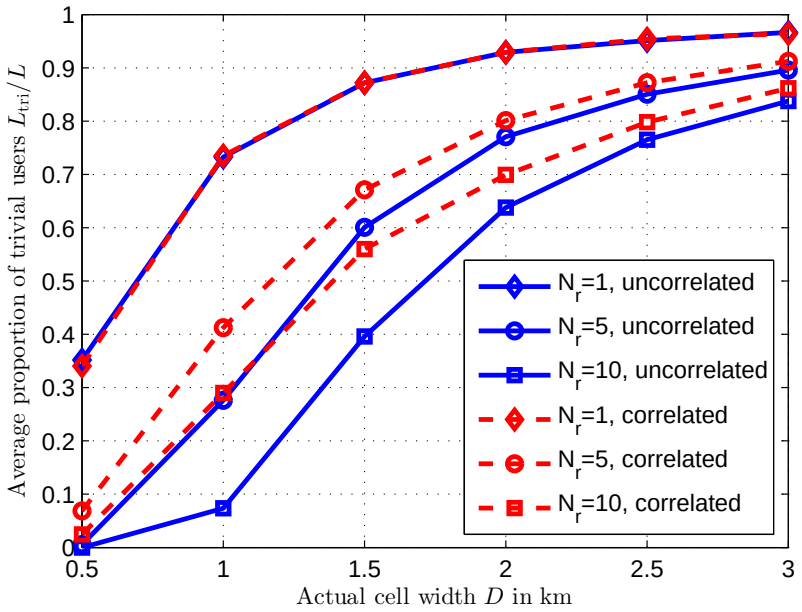

Figure 4.2: Average proportion of the trivial users $\mathbb{E}[\frac{L_{\mathrm{tri}}}{L}]$: $L = 40$, $N_{\mathrm{r,total}} = 100$

Algorithm 5 in section 4.3.1 enables L_{tri} trivial users to be selected which do not participate in the selection procedure. Thus, the proportion of such trivial users, denoted as $\frac{L_{\mathrm{tri}}}{L}$, can be used to measure the complexity reduction compared to the original network. We consider 3 cases corresponding to $N_{\mathrm{r}} = 1, 5$ and 10. The total number of AP antennas is 100 for all cases, and the average results are obtained over 2000 channel realisations. As shown in Figure 4.2, $\mathbb{E}[\frac{L_{\mathrm{tri}}}{L}]$ increases monotonically as the cell width D increases: this is simply due to the larger pathloss for the bigger cell. Recall that the criterion for trivial users is $\|\mathbf{g}_{m,l}\| \leq \frac{1}{2\sqrt{\mathrm{SNR}}}$. Since $\|\mathbf{g}_{m,l}\|$ is expected to increase with N_r, hence $\mathbb{E}[\frac{L_{\mathrm{tri}}}{L}]$ decreases monotonically as N_{r} increases. Compared to the uncorrelated case, the correlated shadowing means that the component c_m in (4.33) is the same for all N_{r} antennas, which effectively reduces the 'receiver-diversity'. Thus, the correlation increases

the probability of $\|\mathbf{g}_{m,l}\| \leq \frac{1}{2\sqrt{\mathrm{SNR}}}$, as illustrated in Figure 4.2. Note that $L = 40$ is considered in all cases, but the same results can be acquired by applying different L, since the density of users only affect the absolute number of L_{tri} rather than the proportion. It can be observed that Algorithm 5 achieves considerable complexity reduction, especially for the fully distributed scenario, and this reduction is caused by considering the range of scaling factor $|\alpha|$ rather than the integer element $|a_l|$.

4.4.3 C&F in Fully Distributed Massive MIMI Model

Complexity-Performance Tradeoff of the Global Selection

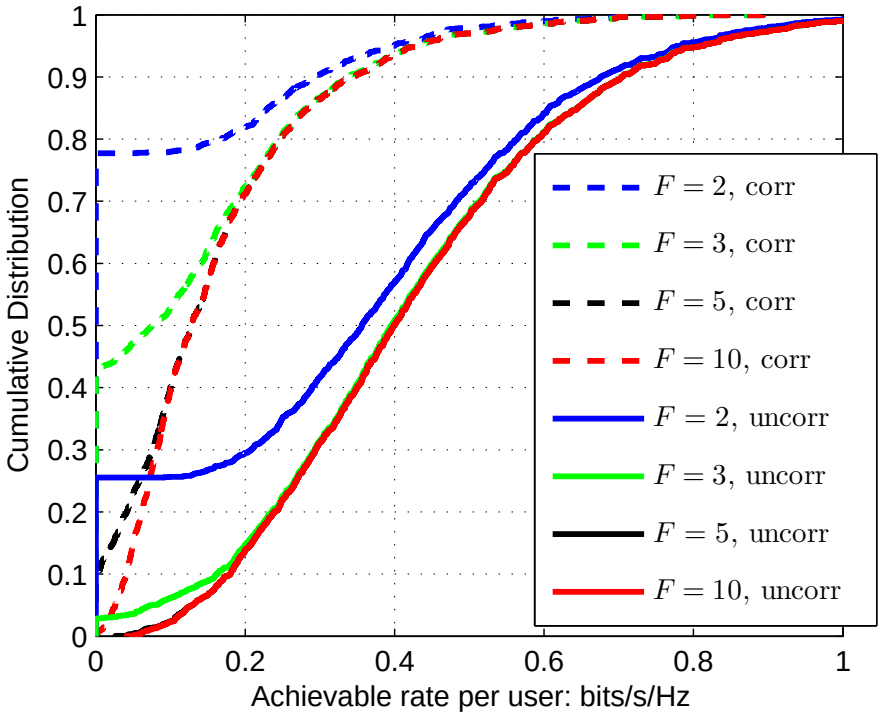

Figure 4.3: Throughput-F tradeoff of C&F: $M = 100$, $L = 40$, $N_{\mathrm{r}} = 1$, $D = 1$km

The rank-deficient probability of the globally selected matrix $\hat{\mathbf{A}}_{\mathrm{sub,opt}}$ greatly depends on the number of locally forwarded candidates F. Figure 4.3 demonstrates the cumulative distribution of the system throughput per user which is the computation rate of the worst equation in $\hat{\mathbf{A}}_{\mathrm{sub,opt}}$. Again, the results are obtained over 2000 channel trials. The zero-rate channel trials mainly correspond to the cases that the hub is not able to select a full rank matrix from the MF candidates. As expected, this rank-deficiency probability

decreases monotonically as F increases, however the selection complexity also increases. We can also observe that the curves of $F = 5\&10$ in the uncorrelated scenario are indistinguishable, which means that if F is sufficiently large to make $\hat{\mathbf{A}}_{\text{sub,opt}}$ full rank, the additional gain by further increasing F is negligible. As previously mentioned, the correlation effectively decreases the 'receiver-diversity', and hence the rank-deficiency problem is more serious for the correlated case. The choice of F depends on the system preference. We focus on the performance preferred systems which employ relatively large F to ensure $\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}}) = L$.

Achievable Rate for Given Outage Probability

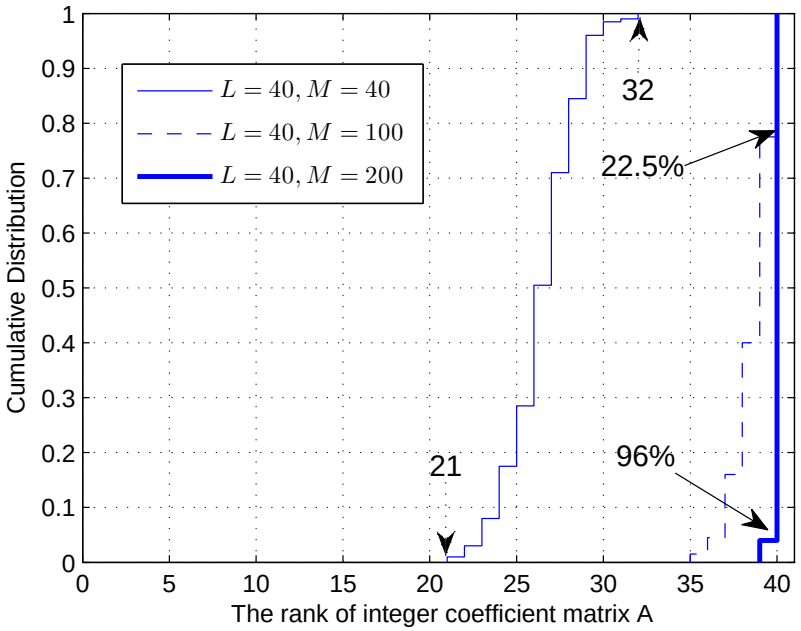

Figure 4.4: CDF of $\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}})$: $F = 1$, $N_{\text{r}} = 1$, $D = 1$km, uncorrelated

In this section, we briefly investigate the performance of the complexity preferred systems which employ relatively smaller F, especially $F = 1$. It can be seen from Figure 4.3 that the probability of rank-deficiency is quite large for small F. Fortunately, rank-deficiency does not imply that none of the L users is decodable, and the number of decodable users is actually determined by $\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}})$. Figure 4.4 shows the cumulative distribution of $\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}})$ with $F = 1$ over 200 channel realisations. For the traditional MU-MIMO scenario with $M = L = 40$ (solid thin line), there are 100% of channel trails end up with rank-deficiency ($\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}}) < 40$), and this proportion is still as

high as 77.5% for the distributed massive MIMO with $M = 100$ (dashed line). However, the minimum number of guaranteed decodable user is significantly increased from 21 to 35. This implies that many more users can be recovered from $\hat{\mathbf{A}}_{\text{sub,opt}}$ regardless of the high probability of rank-deficiency. Therefore, if we allow a small group of users to be passive for each transmission, C&F is still able to provide considerable performance even with $F = 1$.

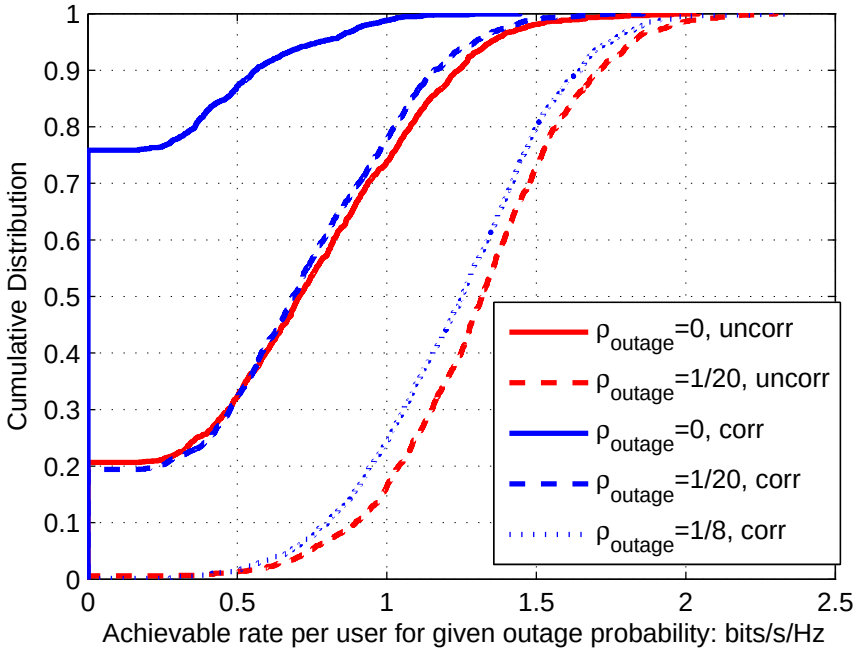

Figure 4.5: CDF of achievable rate for given ρ_{outage}, $L = 40$, $M = 200$, $N_{\text{r}} = 1$, $D = 1$km, $F = 1$

Let R_0 and R_l respectively denote the common transmission rate and the individual achievable rate, and N_{outage} be the number of users in outage (or passive users) during each transmission. We define the outage probability to be the expected value of N_{outage}/L, denoted as

$$\rho_{\text{outage}}(R_0) \triangleq \Pr_{\text{outage}}(R < R_0) = \mathbb{E}[\frac{N_{\text{outage}}}{L}], \tag{4.35}$$

and hence the achievable rate for a given target outage probability can be expressed as

$$R_{\text{outage}}(\rho) \triangleq \sup\{R : \rho_{\text{outage}}(R) \leq \rho\}, \tag{4.36}$$

where sup stands for 'supremum'. In this case, the rank constraint is relaxed to $\text{Rank}(\hat{\mathbf{A}}_{\text{sub,opt}}) \geq L(1 - \rho_{\text{outage}})$. For example, if we allow an outage probability 5% with $L = 40$, then we only need to schedule $40 \times 0.95 = 38$ users and APs for each trans-

mission. The achievable rate for these active users is determined by the computation rate of the 38th best equation in $\hat{\mathbf{A}}_{\text{sub,opt}}$. Again, we treat the equations corresponding to insufficient rank as zero rate equations. Figure 4.5 illustrates the achievable rate $R_{\text{outage}}(\rho)$ for different ρ_{outage}. As expected, the probability of rank-deficiency is significantly reduced with the relaxed condition. The achievable symmetrical rate can be greatly improved by dropping a small proportion of users in each transmission for both correlated and uncorrelated scenarios.

Performance Comparison with Benchmarks

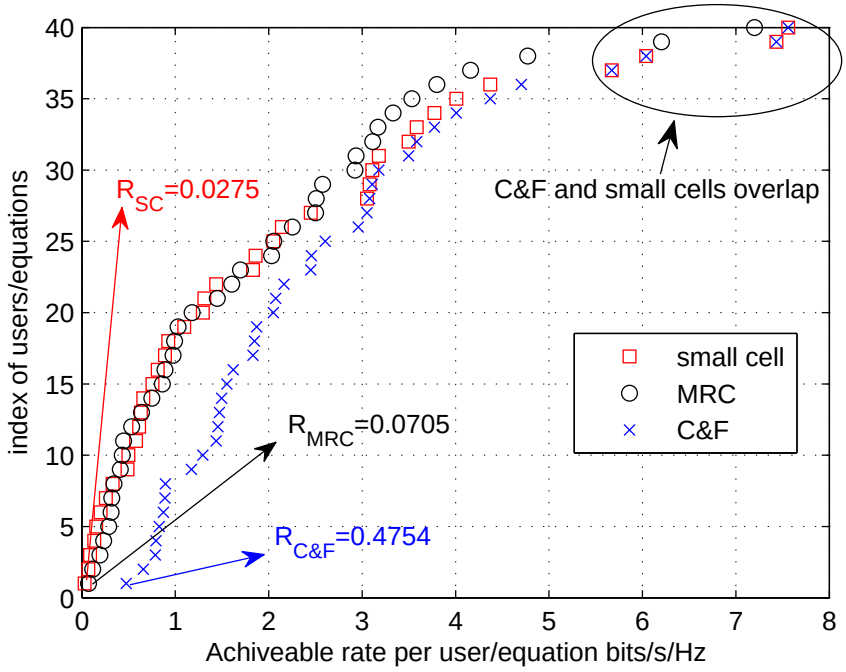

Figure 4.6: Rate comparison for an example channel realisation: $L = 40$, $M = 100$, $N_{\text{r}} = 1$, $D = 1$km, $F = 5$, uncorrelated

As previously mentioned, we focus on the performance preferred systems. Thus, we consider the target outage probability $\rho_{\text{outage}} = 0$ for the rest of this Chapter. We now compare the performance of C&F, MRC and small cells. Figure 4.6 illustrates their achievable rates under an example channel realisation. The black circles calculated by (4.15) and red squares calculated by (4.11) represent the individual rates of 40 users for MRC and small cells respectively. The blue crosses denote the corresponding computation rates of equations in $\hat{\mathbf{A}}_{\text{sub,opt}}$ (if $\text{rank}(\hat{\mathbf{A}}_{\text{sub.opt}}) < 40$, use zero to denote the remaining $40 - \text{rank}(\hat{\mathbf{A}}_{\text{sub.opt}})$ computation rates). All rates are sorted in ascending order, and

the y-axis denotes the index of users (or equations for C&F).

We can see that for both C&F and small cell schemes, the corresponding rates of the top 4 users (or equations) are exactly the same (shown in the right top corner of Figure 4.6). This is due to the fact that when a user is located very close to a specific AP, the equation provided by that AP is very likely to contain that user only, hence the C&F is equivalent to small cells. For other APs, C&F provides higher rate equations compared to the 'single user access' in a small cell. Thus small-cell is a special case of C&F as discussed in section 4.3.3. It is also observed that C&F gives the best performance for 'cell edge' users (or lower rate users), and C&F achieves the highest symmetrical rate (labelled in Figure 4.6) among these schemes.

Figure 4.7 shows the cumulative distributions of their corresponding achievable rates. Again, their achievable rates are determined by the worst user/equation, respectively given in (4.11), (4.16), and (4.27). We can see that with sufficiently large F, C&F provides better performance compared to the benchmarks. For both uncorrelated and correlated cases, the throughput of C&F is roughly 4 and 8 times better than MRC and small cells respectively. Again, a relatively larger F is required for the correlated shadowing case to meet the rank requirement.

4.4.4 IF in Partially Distributed Massive MIMO Model

In this section, we investigate the performance of IF in partially distributed systems. We set $F \geq N_{\mathrm{r}}$ to fully utilise the antenna elements at each AP. Figure 4.8 and 4.9 illustrate the rate comparisons in the cases of $N_{\mathrm{r}} = 5$ and $N_{\mathrm{r}} = 10$ respectively. For all strategies, the achievable rate depends on both SNR and SIR (signal to interference ratio). Since a common $N_{\mathrm{r,total}}$ is employed, hence a larger N_{r} will lead to sparser APs. That means the SNR at each AP antenna decreases as N_{r} increases for all schemes. The SIR in MRC depends on the correlation between users, and such correlation depends on the distance between users which is invariant with N_{r}. Thus, the larger the N_{r} employed, the worse performance MRC will achieve. By contrast, the SIR in both SC and IF correspond to the non-integer error $\|\mathbf{G}_m^T\mathbf{b}_m - \mathbf{a}_m\|^2$. Theoretically, such mismatch between $\mathbf{b}_m^T\mathbf{G}_m$ and $\mathbf{a}_m$ can always be reduced by scaling up the size of the equaliser $\mathbf{b}_m$. Thus, with sufficiently

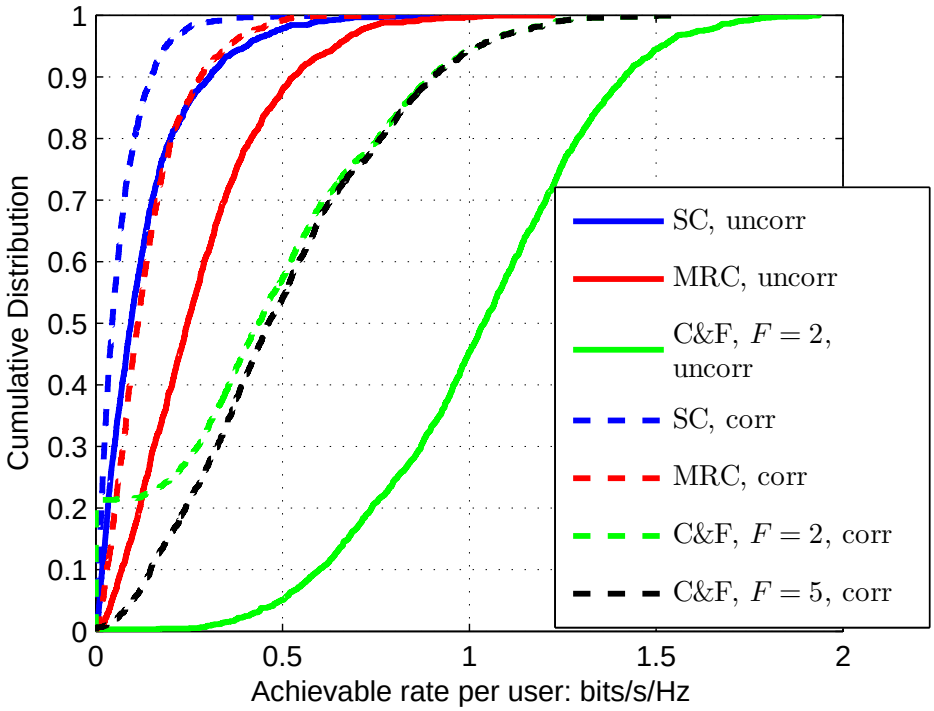

Figure 4.7: CDF of achievable rates: $L = 40$, $M = 200$, $N_{\mathrm{r}} = 1$, $D = 1$km, 2000 channel trials

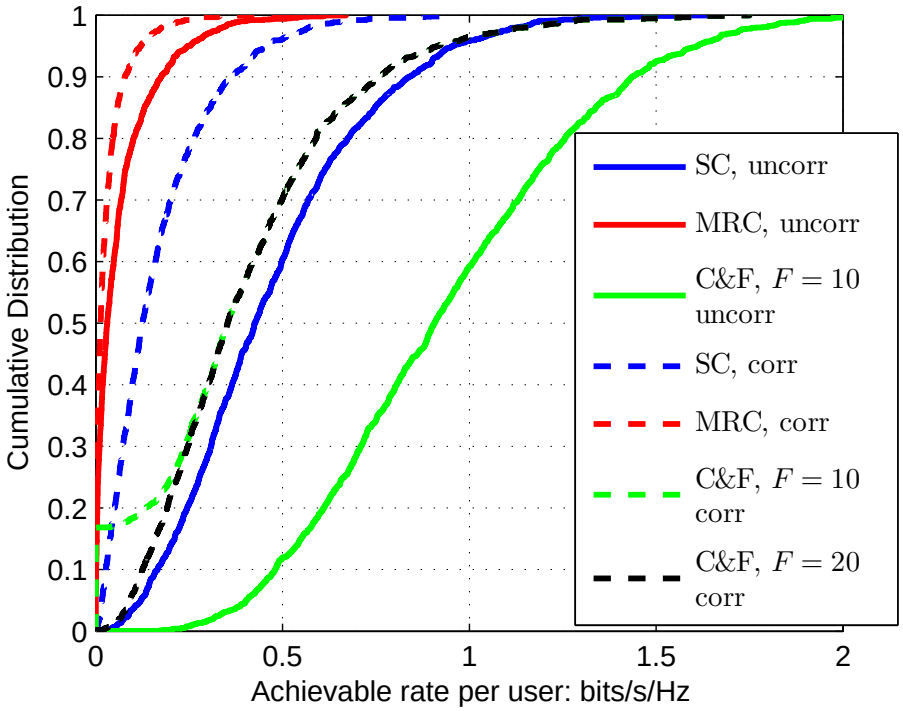

Figure 4.8: CDF of achievable rates: $L = 40$, $N_{\mathrm{r,total}} = 200$, $N_{\mathrm{r}} = 5$, $M = 40$, $D = 1$km, 2000 channel trials

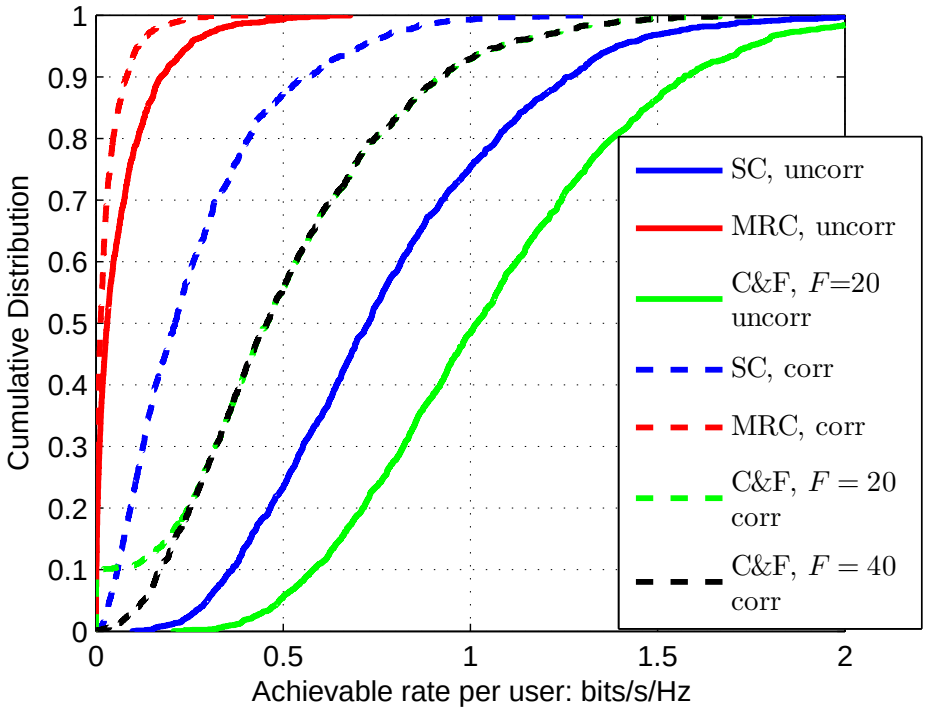

Figure 4.9: CDF of achievable rates: $L = 40$, $N_{\mathrm{r,total}} = 200$, $N_{\mathrm{r}} = 10$ $M = 20$, $D = 1$km, 2000 channel trials

large N_{r}, SC outperforms MRC in both correlated and uncorrelated cases, and the gap increases with N_{r}, as shown in Figure 4.8 and 4.9.

Since the total number of candidate equations MF is the same in Figure 4.7 to 4.9, the performance of IF is mainly determined by the locally selected equations. For each AP, SC is equivalent to MMSE with the selected users. Recall that the gap between IF and MMSE depends on the local channel matrix $\mathbf{G}_m$, and this gap is particularly large for the ill-conditioned $\mathbf{G}_m$ due to the inevitable noise enhancement of MMSE. As discussed in section 2.3.1, the condition number of $\mathbf{G}_m^H\mathbf{G}_m$ is expected to decrease as N_{r} increases. Hence, the gap between IF and SC in Figure 4.9 is smaller than in 4.8. An extreme case is that in centralised massive MIMO ($M = 1$, $N_{\mathrm{r}} \gg L$), this gap is negligible and both schemes approach the performance of joint ML decoding.

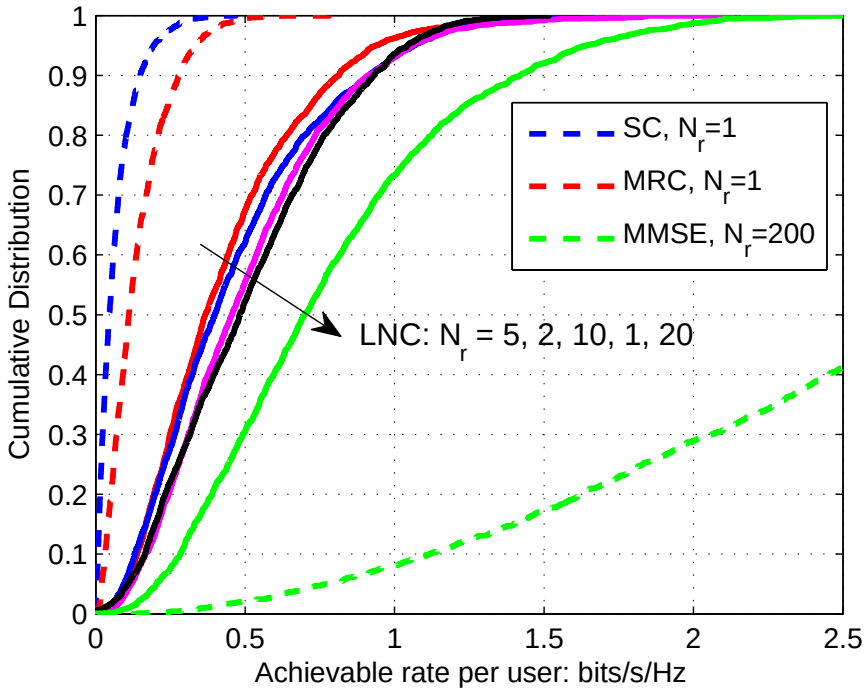

Figure 4.10: Distributed vs centralised massive MIMO: $L = 40$, $N_{\mathrm{r,total}} = 200$, $D = 1$km, correlated shadowing, 2000 channel trials

4.4.5 Distributed vs Centralised Massive MIMO

Now we take the collocated MMSE as the benchmark to provide the performance comparison over different sizes of coverage area. Again, we assume a common $N_{\mathrm{r,total}}$ to make the comparison fair. Since the assumption of uncorrelated shadowing is not realistic for the fully centralised model, we consider the correlated case only in this section. Figure 4.10 shows the rate comparison under a small-coverage region ($D = 1$km). Since the performance in this case is dominated by the interference rather than noise, and centralised MMSE (dashed green curve) is the best for interference cancellation among these schemes, it therefore attains the highest rate. We use solid curves to represent the performance of LNC. As previously discussed, even with a single AP antenna, C&F (black curve) is able to mitigate the interference caused by some 'strong-signal' users (by scaling them to some non-zero integers). By slightly increasing N_{r}, the ability improvement of interference cancellation is limited, while the received SNR is greatly degraded. Thus, the corresponding performance of $N_{\mathrm{r}} = 1$ is better than $N_{\mathrm{r}} = 2\&5$. By further increasing N_{r}, LNC is able to deal with the interference caused by relatively weaker users. In this case, the benefit from the interference cancellation is able to compensate the SNR degradation,

which results in the performance of LNC increasing monotonically with N_{r}, and finally ends with the fully centralised case.

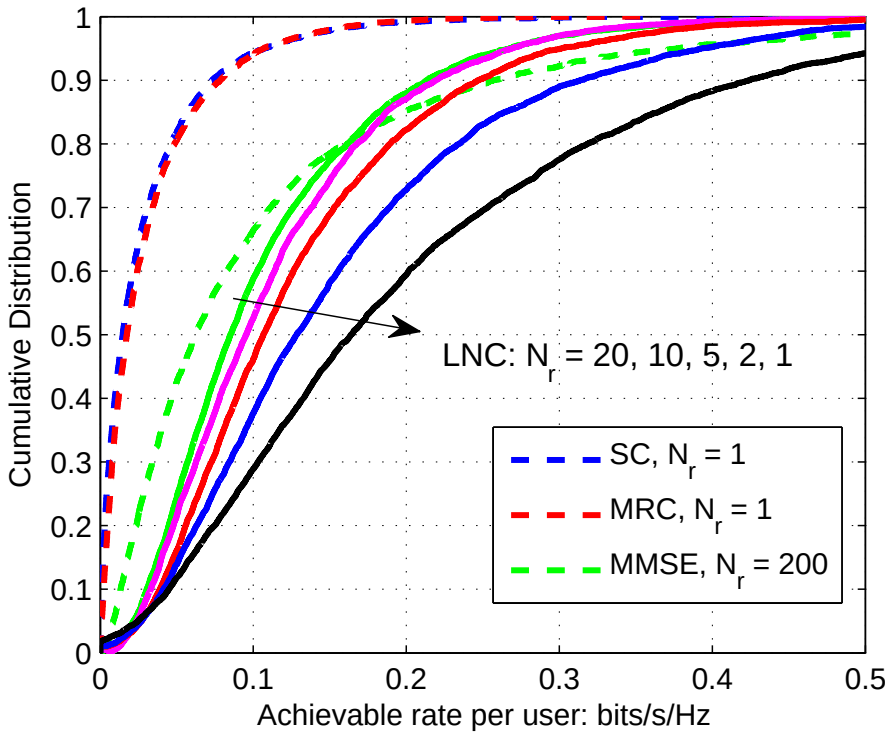

Figure 4.11: Distributed vs centralised massive MIMO: $L = 40$, $N_{\mathrm{r,total}} = 200$, $D = 4$km, correlated shadowing, 2000 channel trials

Figure 4.11 illustrates the rate comparison over a relatively larger region ($D = 4$km). In this case, the interference is not as dominant compared to SNR. The benefit of interference cancellation from increasing N_{r} becomes negligible compared to the SNR degradation. Therefore, LNC outperforms the centralised MMSE, and the performance of LNC monotonically increases as N_{r} decreases, and ends up with the fully distributed C&F.

For a large coverage region ($D = 15$km), the performance is dominated by the thermal noise rather than the interference. Hence the centralised MMSE attains the lowest rate due to the pathloss, as shown in Figure 4.12. Again, the performance of LNC monotonically increases as N_{r} decreases. Unlike the previous two cases, the fully distributed MRC here attains a slightly higher median rate than C&F. This is because the correlation is negligible in such a sparse network and MRC is optimal for maximising the received SNR. However, MRC requires infinite backhaul whereas C&F does not. This tradeoff should be considered for the implementation issue. Additionally, it can be observed from Figure 4.10$\sim$4.12 that the gap between SC and C&F monotonically reduces as the level

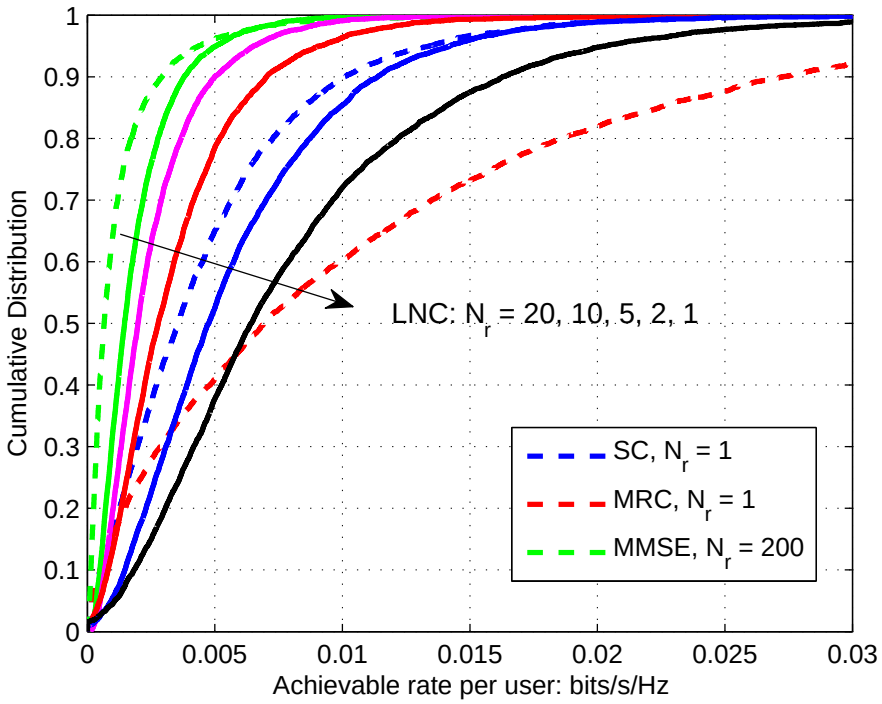

Figure 4.12: Distributed vs centralised massive MIMO: $L = 40$, $N_{\mathrm{r,total}} = 200$, $D = 15$km, correlated shadowing, 2000 channel trials

of interference decreases.

4.4.6 Performance Comparison in One-Slope Pathloss Model

The results obtained so far are based on a three-slope pathloss model as described in (4.30). To simplify the evaluation of the system behaviour, the pathloss is sometimes measured as a one-slope model [71] where a unified exponent $\gamma = 3.5$ is employed throughout the entire coverage area, expressed

$$\mathbf{PL}_{m,l}(d_{m,l}) = \mathbf{PL}_{m,l}(1) - 35 \log_{10} d_{m,l}. \tag{4.37}$$

The distinction between these two models is that for fixed M, L and relative positions, the *noise-to-interference* ratio is invariant with different D in the one-slope model, while this statement does not hold for a certain range of D in the three-slope model. In this section, we will briefly investigate how this issue affects LNC and the benchmark strategies.

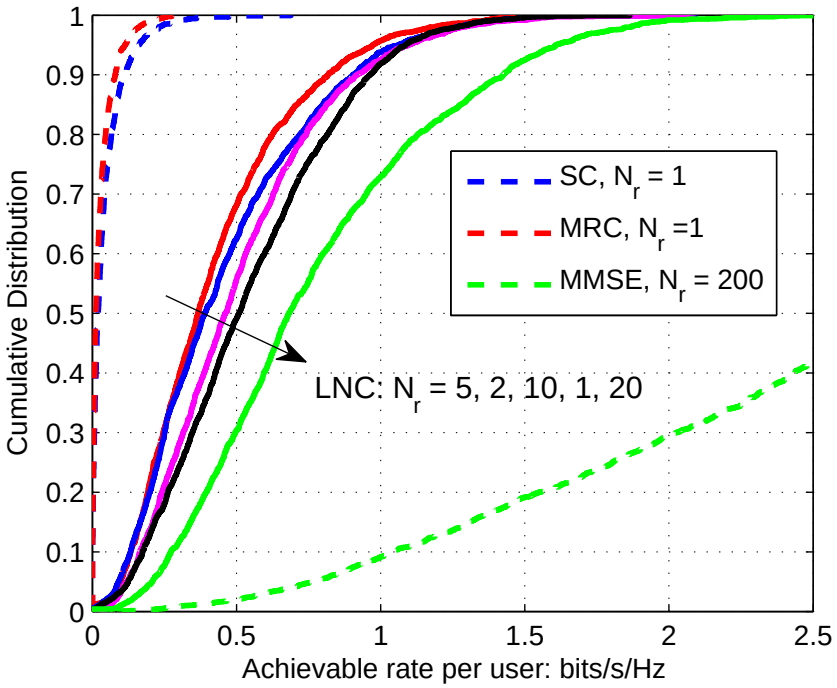

Figure 4.13: Rate comparison in one slope model: $L = 40$, $N_{\mathrm{r,total}} = 200$, $D = 1$km, correlated shadowing, 2000 channel trials

Figure 4.13 applies the same parameters as in Figure 4.10 except the pathloss model. It can be observed that for the LNC, SC and MMSE there is no distinct difference under these two models, while there is a significant degradation for the MRC scheme in the one-slope model. We can see from (4.30) and (4.37) that a unified $\gamma = 3.5$ will result in less pathloss if and only if $d_{m,l} < 0.05$km, which implies that the effect of this modification is more significant on dense networks. Thus, we focus on the comparison in the fully distributed case. Intuitively, the 'good' users which are very close to an AP will benefit from the smaller pathloss. However, as the symmetrical rate is determined by the worst user, the one-slope model might have a negative effect.

Let user l denote such a bad user which has no nearby APs, while there exists a user l' which is very close to AP m. Since MRC combines the received signal from all APs, this strong link $g_{m,l'}$ will inevitably contribute to the interference to user l. In contrast, for SC and LNC, this strong interference component can be avoided by selecting another AP m' instead. Thus, the performance degradation for MRC is more significant while SC and LNC are not sensitive to the change of pathloss model, as shown in Figure 4.14. Figure 4.15 provides the comparison with $D = 15$km: as expected, the curves for all schemes are

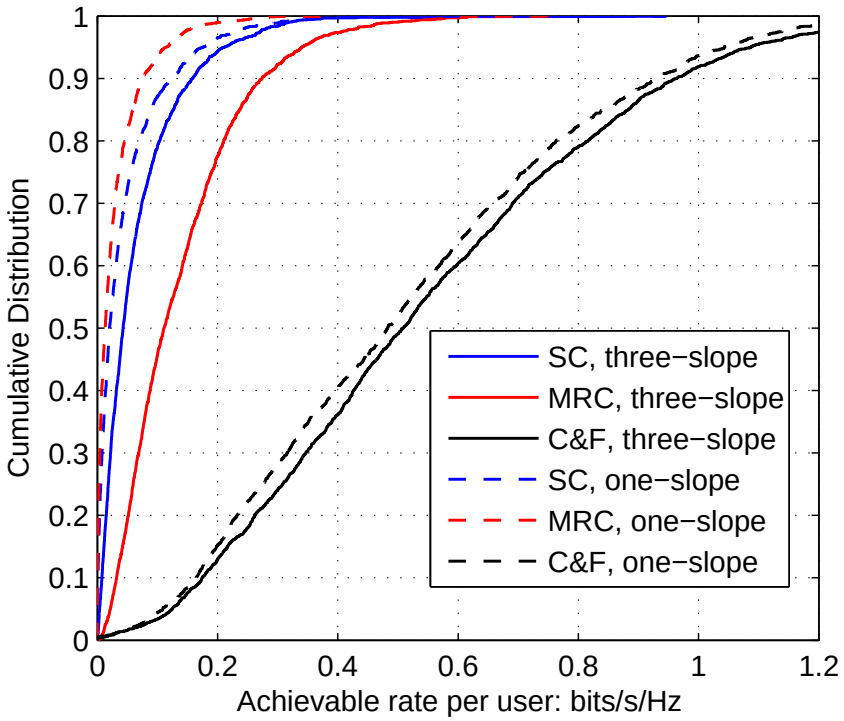

Figure 4.14: 3-slope model vs 1-slope model in fully distributed systems: $L = 40$, $M = 200$, $N_{\mathrm{r}} = 1$, $D = 1$km, correlated shadowing, 2000 channel trials

indistinguishable because the probability that $d_{m,l} < 0.05$km is negligible in such sparse networks.

4.5 Concluding Remarks

In this Chapter, we have applied the LNC scheme to the distributed massive MIMO systems to reduce the backhaul load, and analysed its benefits in such systems. Novel low complexity local selection algorithms for both C&F and IF are proposed. Numerical results have shown that our proposed local selection algorithm significantly reduces the number of effective users, particular for the fully distributed case with C&F. We have also proposed a simple greedy method for the global selection, particularly for the IF case with $N_{\mathrm{r}} > 1$.

We have also given comprehensive comparisons between LNC with the benchmarks. The LNC scheme has slightly higher complexity than the benchmarks, arising from se-

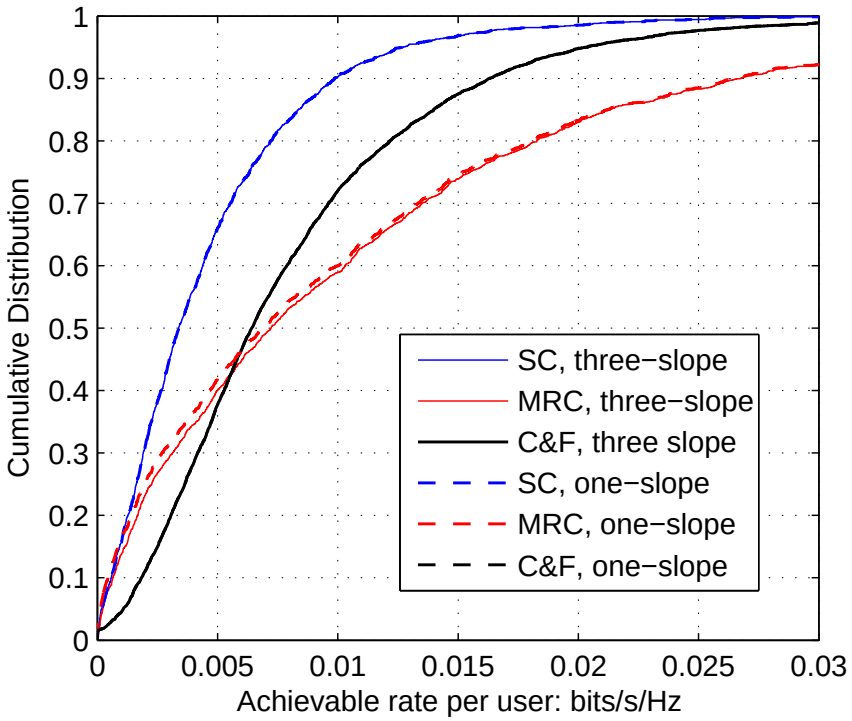

Figure 4.15: 3-slope model vs 1-slope model in fully distributed systems: $L = 40$, $M = 200$, $N_{\mathrm{r}} = 1$, $D = 15$km, correlated shadowing, 2000 channel trials

lecting the optimal integer vectors at the APs, and these processes also lead to some extra latency. In terms of system throughput, numerical results have shown that LNC always outperform small cells, and this gap increases as N_{r} decreases. In addition, the LNC scheme outperforms centralised MMSE and distributed MRC for a wide range of applications.

Chapter 5

Distributed Antenna Systems: Outer Bound with Quantisation and Constrained Backhaul Load

Contents

5.1 Introduction

Motivated by increasingly dense mobile user terminals, the next generation of wireless network systems will require a very large number of access points within a certain area, which might be compared with 'massive MIMO'. In this Chapter, we still focus on the distributed version of massive MIMO, since it allows APs to be located nearer to the users while maintaining the benefits of employing large antenna arrays.

In previous Chapters, we have introduced several promising techniques for future distributed massive MIMO systems, and these schemes can be roughly classified into two types as follows:

- *Quantise-and-Forward* (Q&F) based schemes: each AP forwards a quantised version of the received signal or a quantised version of the equalised signal to the central hub, and then joint processing is performed at the hub. This type includes MRC, Zero-forcing, MMSE, joint ML decoding, etc.
- *Physical-layer Network Coding* (PNC) based schemes: each relay decodes some linear combinations of the source codewords, and then forwards the linearly combined codewords to the hub for message recovery. As previously discussed, we consider decode-and-forward as a special case of PNC. This type includes non-cooperative small cells, C&F, integer forcing, etc.

Clearly, the Q&F based schemes will attain higher spectrum efficiency compared to the centralised massive MIMO, but at the expense of very high backhaul load. The optimal performance can be achieved by setting the quantisation precision of APs as high as possible, which means that the received signals can be precisely forwarded to the hub without loss of information [74]. This corresponds to ideal distributed massive MIMO or ideal C-RAN [75]. In contrast, the PNC based schemes minimise the cardinality with a relatively poorer performance. In Chapter 4, we have shown that the non-cooperative SC usually attains lower rate compared to the centralised MMSE. Although C&F and IF have stronger ability to mitigate the interference, however the system behaviour is still worse than the centralised MMSE when the level of interference is very high. Therefore, it is worthwhile to analyse the trade-off between the performance and the backhaul load. In this Chapter, we will consider an intermediate case between these two types, which means the backhaul load is between the theoretical minimum and infinite.

In practice there is an inevitable loss of information in distributed MIMO systems, arising from the quantisation and modulo operations which are required to restrict backhaul load. Much work has been carried out to find the achievable rate region (inner bound) of such systems [8, 26, 28, 76]. The authors in [77] proposed a 'quantise-and-forward' scheme with joint decoding. However, the rate region in [77] can be tightened by allow-

ing adaptive quantisation rather than a fixed quantiser. Further more, the Gaussian input assumed in [77] is not practical. Previously we have introduced C&F [28] as a promising PNC scheme. Each relay employs lattice decoding which is in effect an adaptive lattice quantiser followed by a modulo lattice operation which minimises the mean square error of the effective noise. Most of related work on C&F so far (this also includes our work in Chapter 3) focus on the achievable rate (inner bound) of one relay. The outer bound, at least a tight outer bound of the overall system is still an open problem. It has been shown in [28, 78] that the classic cut-set upper bound is not tight for C&F. The authors in [76] [26] investigated the overall system performance of many PNC schemes (C&F, binary matrix, etc), and the result in [26] has shown that all of these have significant gaps, of at least 5 dB, compared to the ideal C-RAN. In this Chapter, our objective is to determine the inevitable gap of a backhaul capacity constrained system, at least for a specific quantisation strategy, and hence to tighten the outer bound for such a system. This quantisation strategy is a form of compress-and-forward relaying, but it can be regarded as equivalent to a C&F scheme, albeit with only one-dimensional lattice quantisation and a simple modulo operation. The main contributions of this Chapter are as follows:

- In this Chapter, we employ jointly optimised scaling factors to evaluate the outer bound of a distributed MIMO system with uniform quantisation (equivalent to one dimensional lattice quantisation) under different backhaul load constraints.
- We show that the gap between the standard PNC scheme and ideal distributed MIMO can be significantly reduced by slightly increasing the cardinality of the modulo operation at APs.
- For our proposed adaptive quantisation scheme with extended cardinality, we will give brief guidance on optimising the scale factor of the adaptive quantiser based on the global channel state.

The remainder of this Chapter is organised as follows. We introduce the system model in Section 5.2. Then we provide the outer bound evaluation and implementation discussion in Section 5.3. Numerical results and discussions are given in 5.4. We finally conclude the Chapter in Section 5.5.

5.2 System Model and Preliminaries

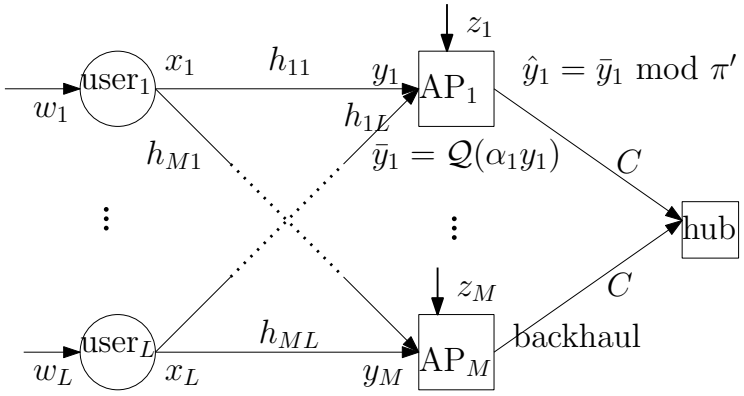

Figure 5.1: Sketch diagram of lattice network coding with extended cardinality.

The distributed model with a modified PNC scheme is illustrated in Figure 5.1. Again, we assume L single antenna users are served by M single antenna APs simultaneously, and the APs are connected via backhaul connections to a hub at which the signals are integrated. We use w_l to denote the original data of the lth user. The channel vector corresponding to the mth AP is represented by $\mathbf{h}_m = [h_{m1}, h_{m2}, \cdots, h_{mL}]^T$. Let $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L]^T$ denote the global channel matrix. $z_m \sim \mathcal{CN}(0, \sigma^2)$ is the complex Gaussian noise added at the mth AP per channel use.

In this Chapter, we are interested in the outer bound calculated by the mutual information for the uncoded scenario. All processes will be described from a one-dimensional lattice perspective. For each user, the original data is modulated onto one of the coset leaders of a quotient ring $\mathbb{A}/\pi\mathbb{A}$, where $\mathbb{A}$ denotes a principal ideal domain and π is a prime number in $\mathbb{A}$. Let $x_l \in \mathbb{A} \bmod \pi$ denote the modulated symbol, and $|\pi|$ be the cardinality of the finite ring. Note that the cardinality is $|\pi^2|$ when $\mathbb{A}$ is the complex integer domain. Thus, the source rate $\mathcal{R}_\mathrm{s}$ can be denoted either as $\log|\pi|$ bits per dimension or $\log|\pi^2|$ bits per complex dimension.

The received signal of the mth relay is expressed as:

$$y_m = \mathbf{x}^T \mathbf{h}_m + z_m, y_m \in \mathbb{C} \tag{5.1}$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_L]^T$ denotes the transmitted symbol vector. The received signal

is scaled by the factor $\alpha_m \in \mathbb{C}$ and then quantised to the nearest integer in $\mathbb{A}$, written as:

$$\bar{y}_m = \mathcal{Q}_{\mathbb{A}}(\alpha_m y_m), \;\; \bar{y}_m \in \mathbb{A} \tag{5.2}$$

We assume the global channel matrix $\mathbf{H}$ is known to the hub, and the scaling factors $\{\alpha_m\}$, $\forall m$ are jointly optimised by the hub, and then fed back to the corresponding APs. In the quasi-static fading scenario, the extra load for sharing $\mathbf{H}$ and $\{\alpha_m\}$ is negligible compared to the data load.

Let $V(\Lambda)$ denote the normalised volume of the fundamental region of Λ. Given α_m, the equivalent quantisation interval is $V(\mathbb{A})/|\alpha_m|$ per dimension. From the average perspective, a fixed larger scaling factor will lead to a lower quantisation error. However, this does not hold for every situation. For example, suppose a sequence of BPSK modulated signals are transmitted through a point to point channel with $h = \frac{2}{3}$. It is clear that the equivalent quantiser $\frac{2}{3}\mathbb{Z}$ (with $\alpha = 1.5$) outperfoms the equivalent quantiser $0.5\mathbb{Z}$ (with $\alpha = 2$).

The final step is to take modulo operation to the quantised signal, expressed as:

$$\hat{y}_m = \bar{y}_m \bmod \pi', \tag{5.3}$$

with $|\pi'| \geq |\pi|$. The modulo operation aims to meet the backhaul constraint. When the cardinality employed at each AP is smaller than the product of source cardinalities ($|\pi'| < |\pi|^L$), it can be regarded as an extended PNC scheme. Particularly when $|\pi'| = |\pi|$, it is exactly the standard PNC. In section 5.5, we will see this modification (increased cardinality) improves the outer bound significantly. We assume the same π' is applied at all APs, in order to ensure that the data recovery at the hub operates over a unified finite ring.

Note that the inner bound of C&F we have studied in Chapters 3 and 4 assumes that both candidate vectors $\mathbf{a}$ and scaling factors α are optimised independently at each AP. The global selection in section 4.3.2 only decides which candidates should be selected without actually changing the values. Thus, we consider these jointly optimised $\{\alpha_m\}$ will lead to an outer bound. We should also note that the joint optimisation here is different to the centralised IF and MMSE. The quantisation and the modulo operation here are applied

to the scaled signal at each individual antenna element $\alpha_m y_m$. Although the equaliser matrix $\mathbf{b}_{\mathrm{IF}}$ and $\mathbf{b}_{\mathrm{MMSE}}$ are also jointly determined at the hub, the quantisation and the modulo operation are applied to the combined signal from all the M antenna elements, denoted as $\mathbf{b}_{\mathrm{MMSE}}^T\mathbf{y}$ with $\mathbf{y} \in \mathbb{C}^{M\times 1}$. That means the equalisation vector $\mathbf{b}_{\mathrm{IF}}$ and $\mathbf{b}_{\mathrm{MMSE}}$ are not applicable for the distributed scenario, due to the backhaul capacity constraint.

5.3 Outer Bound Evaluation

Let $\mathbf{y} = [y_1, y_2, \cdots, y_M]^T$ and $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_M]^T$ denote the input and output respectively of APs, and $\bar{\mathbf{y}} = [\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_M]^T$ be the quantised signal vector. From the expressions in section 5.2, we have:

$$\begin{aligned} \hat{\mathbf{y}} &= \bar{\mathbf{y}} \bmod \pi' \\ &= \mathcal{Q}_{\mathbb{A}}\big(\mathbf{B}(\mathbf{H}\mathbf{x} + \mathbf{z})\big) \bmod \pi', \end{aligned} \tag{5.4}$$

where $\mathbf{z} = [z_1, z_2, \cdots, z_M]^T$, and $\mathbf{B} = \mathrm{diag}([\alpha_1, \alpha_2, \cdots, \alpha_M])$ denotes a diagonal matrix based on the scaling factors. Again, the quantisation $\mathcal{Q}_{\mathbb{A}}$ and the modulo operation mod π' are applied to each element in $\bar{\mathbf{y}}$ and $\hat{\mathbf{y}}$ respectively. Given the channel matrix $\mathbf{H}$, we address the outer bound $\mathcal{R}_{\mathrm{ob}}$ as the maximum mutual information between $\mathbf{x}$ and $\hat{\mathbf{y}}$ corresponding to the optimal scaling vector, expressed as:

$$\mathcal{R}_{\mathrm{ob}}(\mathbf{H}, C) = \underset{\mathbf{B}\in\mathbb{C}^M}{\mathrm{argmax}}\, I(\mathbf{x}; \hat{\mathbf{y}}), \tag{5.5}$$

where $C = \log|\pi'|$ denotes the backhaul capacity in bits per dimension.

5.3.1 Outer Bound of Ideal Distributed MIMO

The outer bound corresponding to the case of infinite backhaul has been investigated in many works. However since we allow adaptive quantisation at APs, it is still necessary to prove that the bound in such a scenario is invariant to scaling.

Proposition 3. The optimal scaling factors of ideal distributed MIMO are $\alpha_m =$

$+\infty$, $\forall m$, and the corresponding outer bound is:

$$\lim_{C\to\infty} \mathcal{R}_{\text{ob}}(\mathbf{H}, C) = I(\mathbf{x};\mathbf{y})$$

Proof. $C \to \infty$ implies the cardinality employed at APs $|\pi'|$ also tends to infinity, and hence the modulo operation $\bar{\mathbf{y}} \bmod \pi'$ has no effect on the quantised signal $\bar{\mathbf{y}}$. In this case, the scaling factor intends to minimise the quantisation error only. Since the quantisation error $\|\mathbf{y} - \bar{\mathbf{y}}\|^2$ is upper bounded by $\sum_m \left(\frac{V(\mathbb{A})}{2\alpha_m}\right)^2$, which decreases monotonically as α_m increases. Thus, by setting $\alpha_m = +\infty$ for all APs, the optimal solution in (5.5) corresponds to the mutual information between $\mathbf{x}$ and $\mathbf{y}$, which is independent of the scaling vector $\mathbf{B}$. □

The calculation of $I(\mathbf{x};\mathbf{y})$ is given as follows:

$$\begin{aligned}\mathcal{R}_{\text{ob}}(H) = I(\mathbf{x};\mathbf{y}) &= H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) \\ &= M\mathcal{R}_{\text{s}} - \sum_{\mathbf{x}} \int_{\mathbf{y}} p(\mathbf{x},\mathbf{y}) \log\big(\frac{1}{p(\mathbf{x}|\mathbf{y})}\big) d\mathbf{y} \qquad (5.6) \\ &= M\mathcal{R}_{\text{s}} - \int_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log\big(\frac{1}{p(\mathbf{x}|\mathbf{y})}\big) \qquad (5.7)\end{aligned}$$

where the conditional probability $p(\mathbf{x}|\mathbf{y})$ is expressed as:

$$\begin{aligned}p(\mathbf{x}|\mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})} \qquad (5.8) \\ &= \prod_{m=1}^{M} \frac{p_{\mathcal{N}}(y_m|\mathbf{x}^T\mathbf{h}_m, \sigma)}{\sum_{\mathbf{x}} p_{\mathcal{N}}(y_m|\mathbf{x}^T\mathbf{h}_m, \sigma)} \qquad (5.9)\end{aligned}$$

The expression (5.8) comes from the equiprobability of the source symbols, and (5.9) comes from the assumption that the corresponding channels of the M APs are independent. The term $p_{\mathcal{N}}(y_m|\mathbf{x}^T\mathbf{h}_m, \sigma)$ denotes the probability density function of a Gaussian variable y_m with mean $\mathbf{x}^T\mathbf{h}_m$ and standard deviation σ, expressed as:

$$p_{\mathcal{N}}(y_m|\mathbf{x}^T\mathbf{h}_m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{\|y_m - \mathbf{x}^T\mathbf{h}_m\|^2}{2\sigma^2}\Big) \qquad (5.10)$$

Due to the continuous nature of the Euclidean space $\mathbf{y} \in \mathbb{C}^M$, there is no closed form of (5.7). Monte Carlo integration is commonly used to evaluate such expressions.

5.3.2 Outer Bound of Constrained Fronthaul

We are more interested in a system constrained by a finite fronthaul load $C = \log|\pi'|$ with $\pi' \neq \infty$. The mutual information of $\mathbf{x}$ and $\hat{\mathbf{y}}$ is written as:

$$\begin{aligned} I(\mathbf{x};\hat{\mathbf{y}}) &= H(\mathbf{x}) - H(\mathbf{x}|\hat{\mathbf{y}}) \\ &= M\mathcal{R}_{\mathrm{s}} - \sum_{\mathbf{x}}\sum_{\hat{\mathbf{y}}} p(\mathbf{x},\hat{\mathbf{y}})\log\big(\frac{1}{p(\mathbf{x}|\hat{\mathbf{y}})}\big) \qquad (5.11) \\ &= M\mathcal{R}_{\mathrm{s}} - \sum_{\hat{\mathbf{y}}} p(\hat{\mathbf{y}}) \sum_{\mathbf{x}} p(\mathbf{x}|\hat{\mathbf{y}})\log\big(\frac{1}{p(\mathbf{x}|\hat{\mathbf{y}})}\big) \qquad (5.12) \end{aligned}$$

with the conditional probability $p(\mathbf{x}|\hat{\mathbf{y}})$ expressed as:

$$\begin{aligned} p(\mathbf{x}|\hat{\mathbf{y}}) &= \frac{p(\hat{\mathbf{y}}|\mathbf{x})p(\mathbf{x})}{p(\hat{\mathbf{y}})} = \frac{p(\hat{\mathbf{y}}|\mathbf{x})}{\sum_{\mathbf{x}} p(\hat{\mathbf{y}}|\mathbf{x})} \qquad (5.13) \\ &= \prod_{m=1}^{M} \frac{p(\hat{y}_m|\mathbf{x})}{\sum_{\mathbf{x}} p(\hat{y}_m|\mathbf{x})} \qquad (5.14) \end{aligned}$$

Recall that the integers are uniformly distributed over the Euclidean space for any PID. For simplicity, we employ a cubic lattice to present the remaining derivation, since the integral of the PDF of the Gaussian noise over a cubic region can be expressed using some Q-functions, which is a 'presentation-friendly' approach. The calculation of non-cubic lattices can also be presented in a similar manner. In addition, since a hypercube lattice can be constructed by a Cartesian product of $\mathbb{Z}$-lattices, and the key geometrical parameters of a lattice are invariant to Cartesian product [79], it suffices to consider a 1-dimensional $\mathbb{Z}$-lattice with real valued channels only. Let Λ_{c}' denote the coarse lattice $\pi'\mathbb{Z}$. For given α_m, the term $p(\hat{y}_m|\mathbf{x})$ in (5.14) can be expressed as:

$$\begin{aligned} p(\hat{y}_m|\mathbf{x}) &= \sum_{\lambda_{\mathrm{c}}'\in\Lambda_{\mathrm{c}}'} p(\bar{y}_m = \hat{y}_m + \lambda_{\mathrm{c}}'|\mathbf{x}) \qquad (5.15) \\ &= \sum_{\lambda_{\mathrm{c}}'\in\Lambda_{\mathrm{c}}'} \int_{\frac{\hat{y}_m+\lambda_{\mathrm{c}}'-0.5}{\alpha_m}}^{\frac{\hat{y}_m+\Lambda_{\mathrm{c}}'+0.5}{\alpha_m}} p(y_m|\mathbf{x})dy_m \qquad (5.16) \\ &= \sum_{\mathbb{Z}} \int_{\frac{\hat{y}_m+\pi'\mathbb{Z}-0.5}{\alpha_m}}^{\frac{\hat{y}_m+\pi'\mathbb{Z}+0.5}{\alpha_m}} p_{\mathcal{N}}(y_m|\mathbf{x}^T\mathbf{h}_m,\sigma)dy_m \qquad (5.17) \\ &= \frac{1}{2}\sum_{\mathbb{Z}} \left(\mathrm{erfc}\Big(\frac{\beta_{m,l}-\mathbf{x}^T\mathbf{h}_m}{\sqrt{2}\sigma}\Big) - \mathrm{erfc}\Big(\frac{\beta_{m,r}-\mathbf{x}^T\mathbf{h}_m}{\sqrt{2}\sigma}\Big)\right) \qquad (5.18) \end{aligned}$$

where the expression (5.15) comes from the reverse process of (5.3), and (5.16) is obtained based on (5.2). Combined with (5.10), the integral in (5.17) is further expressed as the complementary error function in (5.18), where $\beta_{m,l} = (\hat{y}_m + \pi'\mathbb{Z} - 0.5)/\alpha_m$ and $\beta_{m,r} = (\hat{y}_m + \pi'\mathbb{Z} + 0.5)/\alpha_m$ denote the left and right boundaries of the Voronoi region of $(\hat{y}_m + \pi'\mathbb{Z})/\alpha_m$ respectively.

Note that $p(\hat{y}_m|\mathbf{x})$ is a summation of infinite terms of $p(\bar{y}_m = \hat{y}_m + \lambda'_{\mathrm{c}}|\mathbf{x})$, which makes the Monte Carlo integration unavailable. Fortunately, the probability of a Gaussian variable decreases exponentially with the Euclidean distance. Hence the infinite number of terms of $p(\bar{y}_m = \hat{y}_m + \lambda'_{\mathrm{c}}|\mathbf{x})$ can be reduced to a finite number by discarding those whose corresponding $\lambda'_{\mathrm{c}} \in \Lambda'_{\mathrm{c}}$ are far away from $\alpha_m \mathbf{x}^T \mathbf{h}_m$. In our simulations in section 5.4, we will discard all λ'_{c} with the corresponding $p(\bar{y}_m = \hat{y}_m + \Lambda'_{\mathrm{c}}|\mathbf{x}) < 2.2251 \times 10^{-308}$. Note that 2.2251×10^{-308} is the minimum positive number in Matlab.

Based on the derivations above, we are able to calculate the corresponding $I(\mathbf{x}; \hat{\mathbf{y}})$ for any scaling vector $\mathbf{B} \in \mathbb{R}^M$. Then the outer bound can be obtained by an exhaustive search over all possible $\mathbf{B}$.

- For a very large scaling factor α, the mutual information is dominated by the scaled Gaussian noise αz rather than the interference. Unlike the unlimited backhaul scenario, the folded scaled Gaussian noise $\alpha z \bmod \pi'$ tends to be uniformly distributed for finite π'. In this case, the channel capacity approaches zero since the effective constellation points are indistinguishable. Hence it is sufficient to conduct the search within a region of $\mathbf{B} \in [0, b)^M$ rather than $\mathbb{R}^M$.
- The continuous region of $\mathbf{B}$ can be replaced by a set of discrete points by setting a fixed step size $\Delta\alpha$ per dimension. In this case there are $\lceil \frac{b}{\Delta\alpha} \rceil^M$ candidate vectors of $\mathbf{B}$ to be searched.

It is obvious that a larger b and a smaller $\Delta\alpha$ will lead to a more accurate evaluation. From many observations, we found $\Delta\alpha = 0.01\pi$ and $b = 10\pi'$ are sufficient to obtain accurate results.

5.3.3 Implementation Discussion

The outer bound in section 5.3.2 is obtained based on two assumptions which are not always feasible in practice.

- The selection of the scaling vector is completely done by the hub station. The optimisation of a M-dimensional variable leads to high complexity.
- The mutual information, which is required to optimise the scaling factor, is calculated by Monte-Carlo integration.

As addressed in previous Chapters, the former issue can be solved by allowing each relay to select its best several scaling factors locally, and then forward them to the hub to process global selection. The original optimisation problem is decomposed to M optimisations with 1-dimensional variables, which will end up with the achievable inner bound in Chapter 3 and 4.

To overcome the latter one, we derive an approximate closed form expression for the mutual information in the rest of this section. In section 5.4 we will show that this closed form expression gives almost the same result as the Monte Carlo does.

As discussed in section 5.3.2, most points of $\lambda'_{\mathrm{c}} \in \Lambda'_{\mathrm{c}}$ can be eliminated due to the rapid decline of the the probability for larger distances. Can we keep only the closest point around $\alpha_m \mathbf{x}^T \mathbf{h}_m$ in Λ'_{c} and discard all others?

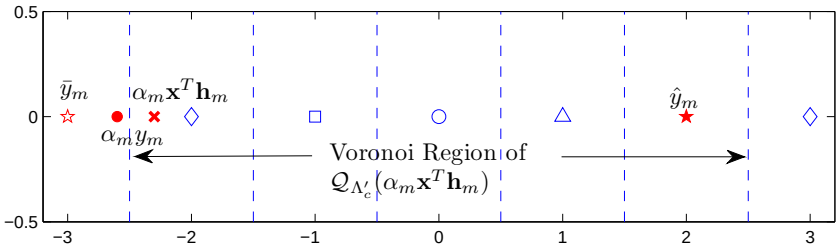

Figure 5.2: illustration of 'wrap-around' error

To answer this question, we need to investigate the additional error caused by the modulo operation. Let $\mathcal{Q}_{\Lambda'_{\mathrm{c}}}(\alpha_m \mathbf{x}^T \mathbf{h}_m)$ denote the closest coarse lattice point around $\alpha_m \mathbf{x}^T \mathbf{h}_m$. If the Gaussian noise moves $\alpha_m y_m$ out of the Voronoi region of $\mathcal{Q}_{\Lambda'_{\mathrm{c}}}(\alpha_m \mathbf{x}^T \mathbf{h}_m)$, the

value of $\hat{y}_m$ will be wrapped to the opposite end after the modulo operation. Figure 5.2 illustrates an example of this wrap around error. We assume $\Lambda'_{\rm c} = 5\mathbb{Z}$, and the Voronoi region of $\mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m)$ is an interval from -2.5 to 2.5. The Gaussian noise moves the signal from $\alpha_m \mathbf{x}^T \mathbf{h}_m$ (red cross) to $\alpha_m y_m$ (red circle), and this leads to an error $\bar{y}_m = -2 \rightarrow \bar{y}_m = -3$. The modulo operation 'amplifies' this error by mapping $\bar{y}_m = -3$ to $\hat{y}_m = 2$ which results in a total error of $\hat{y}_m = -2 \rightarrow \hat{y}_m = 2$. Clearly, when $\hat{y}_m$ is located at the edge of the Voronoi region ($\hat{y}_m = \pm 2$), the corresponding $\bar{y}_m$ is more likely to be $\bar{y}_m = \hat{y}_m + \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m) \mp \pi'$, while the points in the middle ($\hat{y}_m \in \{-1, 0, 1\}$) are less sensitive to this 'wrap around' error. Thus, we consider $\lambda'_{\rm c} = \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m)$, $\lambda'_{\rm c} \in \{\mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m), \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m) + \pi'\}$ and $\lambda'_{\rm c} \in \{\mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m), \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m) - \pi'\}$ for the middle points, left edge point and right edge point respectively. The term $p(\hat{y}_m|\mathbf{x})$ in (5.15) can be approximated by

- If $\hat{y}_m = -\frac{\pi'-1}{2}$ (left edge point)

$$p(\hat{y}_m|\mathbf{x}) = \sum_{\lambda'_{\rm c} \in \{\mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m), \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m) + \pi'\}} p(\bar{y}_m = \hat{y}_m + \lambda'_{\mathbf{c}}|\mathbf{x}) \tag{5.19}$$

- If $\hat{y}_m = \frac{\pi'-1}{2}$ (right edge point)

$$p(\hat{y}_m|\mathbf{x}) = \sum_{\lambda'_{\rm c} \in \{\mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m), \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m) - \pi'\}} p(\bar{y}_m = \hat{y}_m + \lambda'_{\mathbf{c}}|\mathbf{x}) \tag{5.20}$$

- If $\hat{y}_m \neq \pm\frac{\pi'-1}{2}$ (middle points)

$$p(\hat{y}_m|\mathbf{x}) = p\big(\bar{y}_m = \hat{y}_m + \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m)|\mathbf{x}\big) \tag{5.21}$$

and hence the probability $p(\hat{y}_m)$ can be written as:

$$p(\hat{y}_m) = \sum_{\mathbf{x}} \frac{p(\hat{y}_m|\mathbf{x})}{\sum_{\lambda=-(\pi'+1)/2}^{(\pi'+1)/2} p\big(\bar{y}_m = \lambda + \mathcal{Q}_{\Lambda'_{\rm c}}(\alpha_m \mathbf{x}^T \mathbf{h}_m)|\mathbf{x}\big)} \tag{5.22}$$

and the term $p(\hat{\mathbf{y}})$ can be easily obtained by $p(\hat{\mathbf{y}}) = \prod_{m=1}^{M} p(\hat{y}_m)$. Plugging this $p(\hat{\mathbf{y}})$ back into (5.12), the mutual information can be estimated by a closed form expression.

5.4 Numerical Results and Discussions

In this section, we evaluate the throughputs in different scenarios based on the formulae obtained in section III. We apply $\mathbb{A}/\pi\mathbb{A} = \mathbb{Z}/3\mathbb{Z}$ ($x_m \in \{-1, 0, 1\}$) to all source symbols. The *signal-to-noise ratio* (SNR) is defined as $\mathbb{E}[|x_m|^2]/\sigma^2$.

5.4.1 Multiple Access Channel with Fixed Fading

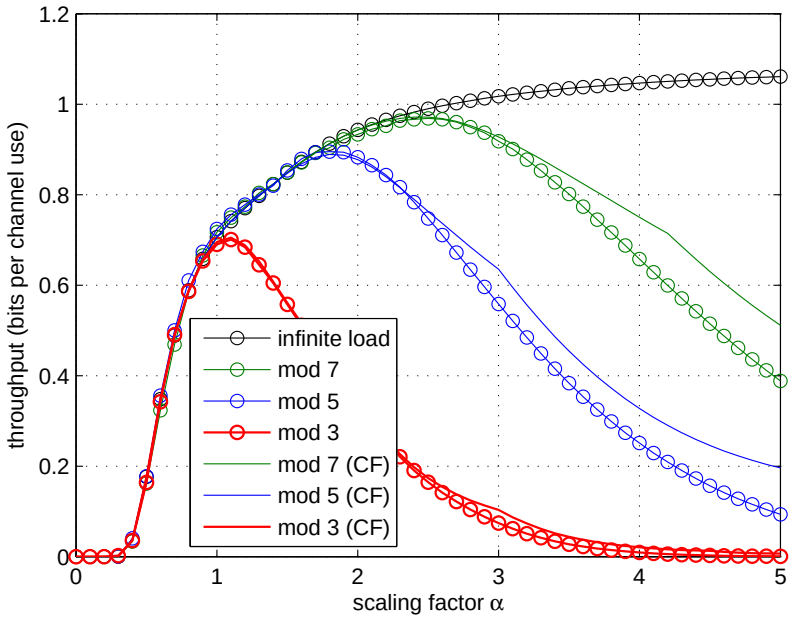

Figure 5.3: $I(\mathbf{x}, \hat{y}_m)$: $\mathbf{h}_m = [\frac{1}{2}, \frac{1}{3}]^T$, SNR=10dB

In many distributed MIMO applications, the local optimisation at each AP plays a fundamental role. In this part, we consider a 2 by 1 multiple access channel with fixed fading $\mathbf{h}_m = [\frac{1}{2}, \frac{1}{3}]^T$. Figure 5.3 illustrates the throughput $I(\mathbf{x}; \hat{y}_m)$ for a low SNR case. The curves with circle markers show the throughputs obtained by the Monte Carlo method in section 5.3.2, and the curves without marker are calculated by the closed form expressions in 5.3.3. We use black, green, blue and red colours to present the cases of $\Lambda'_c = 3\mathbb{Z}$ (standard PNC), $\Lambda'_c = 5\mathbb{Z}$, $\Lambda'_c = 7\mathbb{Z}$ and infinite load (ideal case), respectively.

The throughput of a low SNR system is dominated by the 'wrap around' error. It can be seen that the function of $I(\mathbf{x}; \hat{y}_m)$ has a single maximum. The peak points corresponds

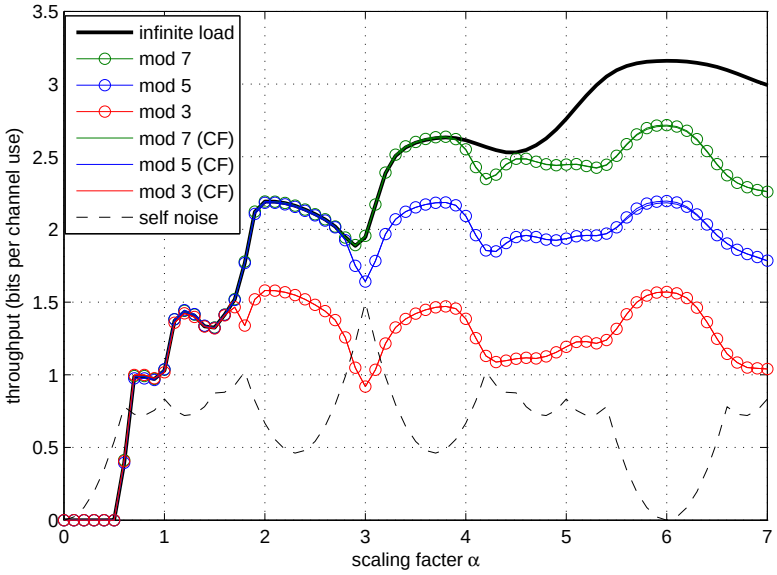

Figure 5.4: $I(\mathbf{x}, \hat{y}_m)$: $\mathbf{h}_m = [\frac{1}{2}, \frac{1}{3}]^T$, SNR=30dB

to the points at which the 'wrap around' error begins to dominate, and the distribution of the 'folded noise' tends to uniform as the scaling factor increases. Additionally, these curves show that our closed form expression provides a good approximation of $I(\mathbf{x}; \hat{y}_m)$ (note that the mismatch in the large α region does not affect the selection of optimal α), and this expression is a summation of a finite number of differentiable terms. Hence, any hill climbing method can be used to find the local optimal α.

Figure 5.4 shows the $I(\mathbf{x}|\hat{y}_m)$ with SNR=30dB. Since the noise is small, the middle points are hardly affected by the 'wrap around', and hence a sufficiently good approximation can be acquired by considering the effect on the edge points only. Therefore, the closed form expressions in (5.19)∼(5.21) give a more accurate approximation compared to the low SNR case. It can be observed that the curves calculated by Monte Carlo and the closed form are indistinguishable in Figure 5.4). Due to the relatively lower probability of 'wrap around', the mutual information is dominated by the quantisation error. The quantisation error can be roughly measured by the variance of the 'self noise', which denotes the mismatch between the scaled channel and its quantised value, written as $\sigma^2_{\text{self}} = \|\alpha\mathbf{h}_m - \mathcal{Q}_{\mathbb{A}}(\alpha\mathbf{h}_m)\|^2$. It is clear that the peak points of $I(\mathbf{x}; \hat{y}_m)$ correspond to the trough values of σ^2_{self} (represented by the dashed black curve), and the corresponding values of α are not sensitive to different cardinalities of $\mathbb{A}/\pi'\mathbb{A}$. Hence, the selection of the local optimal α_m can be implemented by the following 2 steps:

- Find the candidates of α_m according to σ^2_{self}
- For each candidate, calculate the corresponding $I(\mathbf{x}; \hat{y}_m)$ based on the closed form expression, and selects the one with the maximum $I(\mathbf{x}; \hat{y}_m)$

5.4.2 MIMO with Fixed Fading

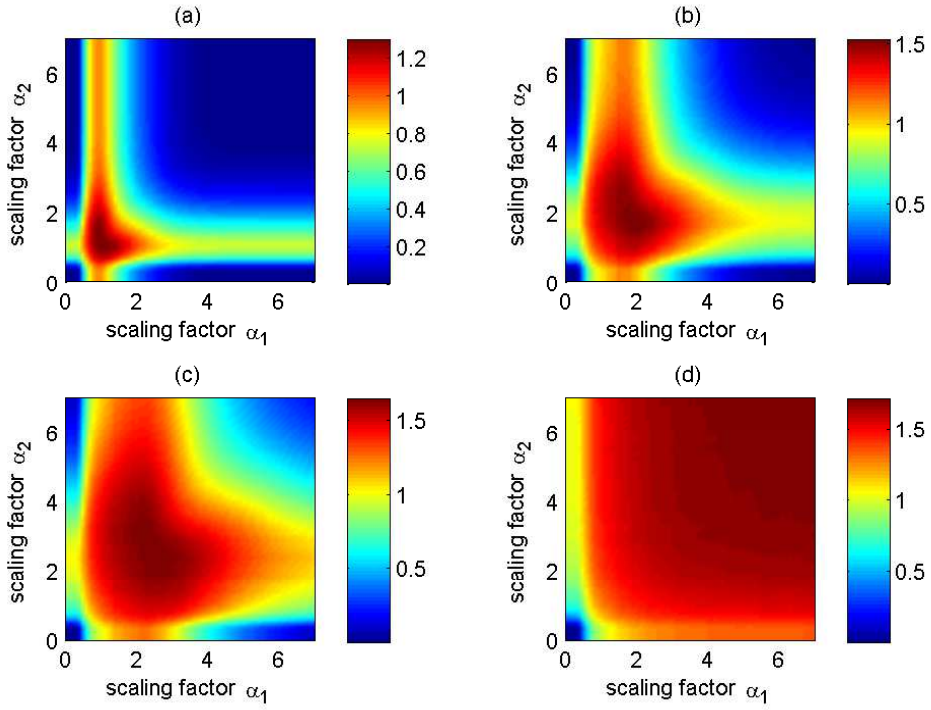

Figure 5.5: SNR = 10dB: (a) $\Lambda'_{\text{c}} = 3\mathbb{Z}$ (b) $\Lambda'_{\text{c}} = 5\mathbb{Z}$ (c) $\Lambda'_{\text{c}} = 7\mathbb{Z}$ (d) infinite backhaul

In this section, we consider a 2 by 2 MIMO system in which the hub selects α_m for both relays. The channel matrix $\mathbf{H} = [\frac{3}{4}, \frac{1}{4}; \frac{1}{2}, \frac{1}{3}]$. Figure 5.5 and 5.6 respectively show the cases of SNR = 10dB and SNR=30dB. In both figures, the x-axis and y-axis denote α_1 and α_2 respectively. The values of $I(\mathbf{x}; \hat{\mathbf{y}})$, calculated by the Monte Carlo integration in section 5.3.2, are represented in the colour bar. It can be observed from Figure 5.5 that the function of $I(\mathbf{x}; \hat{\mathbf{y}})$ has single maximum for the low SNR case which means that the optimal scaling vector $\mathbf{B}$ can still be acquired by employing some hill climbing methods. For high SNR there are many values of $\mathbf{B}$ which correspond to a near-maximum throughput, hence a linear search (similar to Algorithm 4 we have addressed in Chapter 3) with a relatively wider step size can be utilised to reduce the complexity. Again, the performance in high SNR is dominated by the quantisation error, and the quantisation

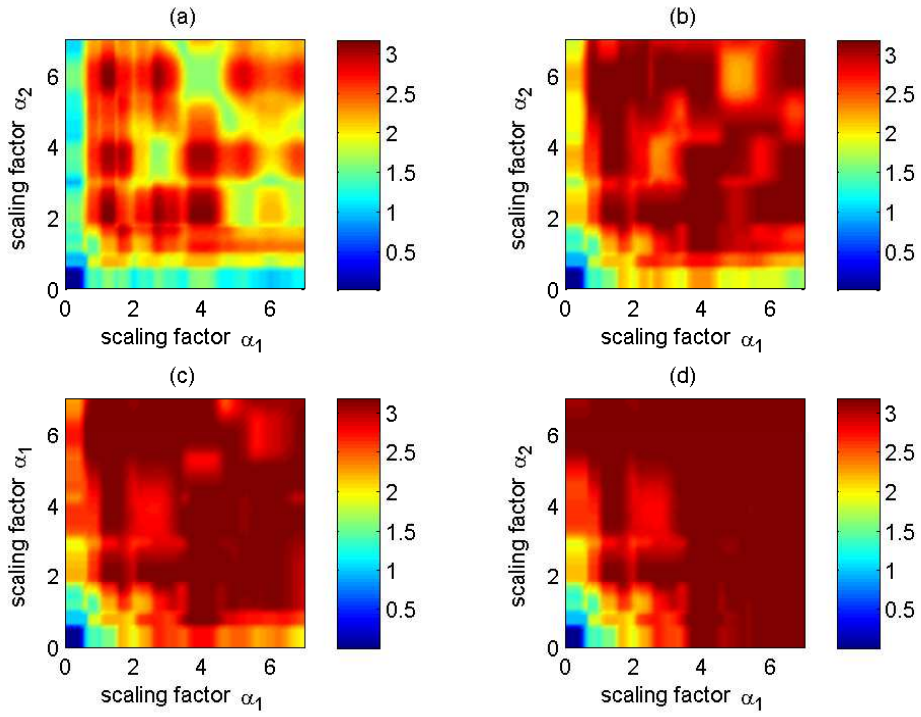

Figure 5.6: SNR = 30dB: (a) $\Lambda'_{\mathrm{c}} = 3\mathbb{Z}$ (b) $\Lambda'_{\mathrm{c}} = 5\mathbb{Z}$ (c) $\Lambda'_{\mathrm{c}} = 7\mathbb{Z}$ (d) infinite backhaul

error can be measured by the non-integer noise. At first glance, $\{\alpha_1 = 4,\ \alpha_2 = 6\}$ is one set of optimum scaling factors since it precisely scales the channel matrix $\mathbf{H}$ to an integer matrix $[3,\ 1; 3,\ 2]$. However, when a $\mathrm{mod}\ 3$ operation is applied, it becomes a rank-deficient matrix $[0,\ 1; 0,\ 2]$ over $\mathbb{F}_3$. This results in a great loss of information as shown in Figure 5.6(a). This issue can be solved by employing extended C&F, as shown in (b), (c), and (d). Note that when $\alpha_1 \approx 3$ and $\alpha_2 \approx 4$, the effective integer channel $\mathcal{Q}_{\mathbb{Z}}(\mathbf{BH}) = [2,\ 1; 2,\ 1]$ is singular no matter which moduli is applied, and hence the corresponding performance is relatively worse than other values of $\mathbf{B}$ in all sub-figures in 5.6 .

5.4.3 Fading Channel

Finally, we evaluate the average $I(\mathbf{x}; \hat{\mathbf{y}})$ of a 2 by 2 system. The results are obtained based on 1000 channel realisations, with each element $h_{m,l} \sim \mathcal{N}(0, 1)$. As shown in Figure 5.6, the outer bound can be significantly improved by slightly increase the cardinality. Specifically, if a half rate FEC code is employed (represents by the dashed red line), there

is a 5dB gap between standard PNC (with circle marker) and ideal distributed MIMO (without marker), and this gap can be reduced to less than 2dB(1dB) if we increase the cardinality to 5(7) respectively. Note that in those cases, the backhaul loads are only increased by $\log 5/\log 3 = 1.46$ and $\log 7/\log 3 = 1.77$ times respectively.

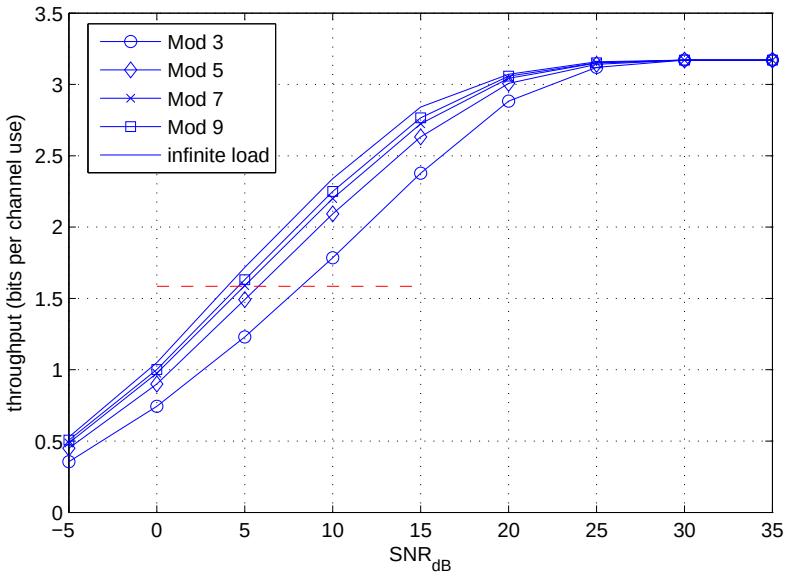

Figure 5.7: Average throughput over 1000 channel realisations

5.5 Concluding Remarks

In this Chapter, we have extended a standard PNC scheme for distributed massive MIMO to allow a flexible backhaul load. The lattice quantisation and modulo operations make it possible to be implemented as a C&F strategy. We have investigated the outer bound of this scheme under different backhaul constraints by maximising the mutual information rather than minimising the mean square error of the quantisation noise. The maximum mutual information is acquired by jointly optimising the scaling factors of all distributed APs. We have proposed two approaches to evaluate this outer bound: Monte Carlo integration based method and an approximately closed form expression based method. The former provided an accurate evaluation and the latter significantly reduced the computational complex. From a practical perspective, the scaling factor can be optimised independently at each AP, and a linear search with relatively wider step sizes can be employed.

Simulation results have shown that the outer bound of this scheme can be significantly improved by slightly increasing the cardinality of the modulo operation.

Chapter 6

Conclusions and Future Work

Contents

6.1 Summary of the Work

In this thesis, we have investigated the applications of LNC in distributed massive MIMO systems, for the purpose of avoiding the enormous backhaul load. Specifically, we have given novel algorithms for the coefficient selection in LNC, and analysed the system behaviour in a realistic distributed massive MIMO model. Additionally, the outer bound evaluated under different backhaul constraints is presented.

In the following, we summarise the work by answering the questions proposed at the very beginning of this thesis.

1) How to select the optimal mapping function at each AP?

For the real valued scenario, we have given two novel algorithms for the coefficient selection in C&F: the reduced candidate set aided algorithm and the linear search algorithm. The former employs pre-determined thresholds to discard the 'narrow' intervals, while the latter utilised a fixed step size for sampling the scaling factor α, and selects the

sample which minimises the effective noise. For the complex valued scenario, complex-exhaustive-II and complex linear search algorithm are proposed. The former extends the interval partition to Voronoi region partition, and selects at least one representative for each region to acquire a complete candidate set. The optimality is ensured by this complete candidate set; the latter applies the optimised step size in both x-axis and y-axis directions, to ignore the 'small size' Voronoi regions. The complexity of these algorithms is analysed both numerically and theoretically. Numerical results have shown that our proposed algorithms have better performance complexity tradeoff, compared to the existing approaches.

2) How will LNC perform in realistic distributed massive MIMO systems?

We have evaluated the performance of LNC in a realistic distributed massive MIMO system, which considers both large scale and small scale fading. Novel algoritms for both local and golbal selections are proposed by exploiting the properties of pathloss. We have compared the LNC scheme with distributed MRC (with infinite backhaul), distributed small cells, and the centralised MMSE scheme. Numerical results have shown LNC always outperform SC since SC can be treated as a special case of LNC; LNC outperforms centralised MMSE when the interference is not very strong, since LNC benefits from the distributed deployment; LNC outperforms MRC except in a thermal noise-dominated system, since LNC has a stronger ability to mitigate the interference.

3) How much information will inevitably be lost in LNC?

We have extended standard C&F to allow a flexible cardinality (slightly larger cardinality) at the APs, in order to meet the different backhaul constraint. The loss of information in LNC is due to some quantisations and modulo operations. We employ jointly optimised scaling factors at the APs to evaluate the inevitable loss of information. A Monte Carlo based calculation is proposed to give an accurate evaluation of the outer bound. A simplified closed form expression is proposed for coefficient selection in this extended C&F, for the purposed of approaching the performance of the ideal distributed MIMO (infinite backhaul).

6.2 Future Work

Future avenues of relevant research are listed as follows:

- The investigation of LNC in this thesis assumes that the channel fading is quasi-static and the electromagnetic waves arrive simultaneously at APs. The impact of fast fading and imperfect synchronisation [80] might be considered in the future.
- The complex-exhaustive-II algorithm proposed in Chapter 3 assumes the integer domain $\mathbb{A}$ is a principal ideal domain (PID). When $\mathbb{A}$ is not a PID, the Voronoi region of each integer element in $\mathbb{A}$ might not be a regular polygon: whether the complex-exhaustive-II still applicable for this case should be investigated in the future.
- The comparisons between the distributed LNC scheme and the centralised MMSE scheme are based on perfect channel estimation and independent small scaling fading. The impact of channel estimation error and correlated small scale fading will be further investigated. On the one hand, the correlated small scale fading will bring more performance degradation to the centralised massive MIMO. On the other hand, the sensitivity of both schemes to the channel estimation error is unknown.
- 'One-bit' massive MIMO [81] approach is proposed very recently for centralised systems. It applies a simple sgn quantiser to each antenna element to reduce the complexity. It can be potentially applied to distributed system to reduce the back-haul load. The comparison between LNC and this 'one-bit' scheme with limited number of AP antennas will be investigated in the future.
- According to the description of the extended C&F in Chapter 5, practical designs with multilevel lattice codes will be explored. A straightforward approach is that each user employs a linear code over $\mathbb{A}/\pi\mathbb{A}$, while the APs decode the message over $\mathbb{A}/\pi^m\mathbb{A}$. In order to increase the flexibility, the decoding process at the APs might be performed over $\mathbb{A}/\pi\pi'\mathbb{A}$, where π and π' are co-prime in $\mathbb{A}$.
- Performance comparison between the lattice quantiser (uniform) and the Lloyd-Max quantiser (non-uniform) will be studied.

Glossary

AF	**A**mplify-and-**F**orward
AP	**A**ccess **P**oint
AWGN	**A**dditional **W**hite **G**aussian **N**oise
BC	**B**roadcast **C**hannel
BER	**B**it-**E**rror-**R**ate
BPSK	**B**inary **P**hase **S**hift **K**eying
CDF	**C**umulative **D**istribution **F**unction
CPU	**C**entral **P**rocess **U**nit
C-RAN	**C**loud **R**adio **A**ccess **N**etwork
C&F	**C**ompute-and-**F**orward
CSI	**C**hannel **S**tate **I**nformation
dB	Decibel
DF	**D**ecod-and-**F**orward
IF	**I**nteger-**F**orcing
Hz	**H**ert**z**
MAC	**M**ultiple **A**ccess **C**hannel
MIMO	**M**ultiple-**I**nput **M**ultiple-**O**utput
ML	**M**aximum **L**ikelihood
MMSE	**M**inimum **M**ean **S**quare **E**rror
MRC	**M**aximum **R**atio **C**ombining
MU-MIMO	**M**ultiple **U**ser-**MIMO**
LNC	**L**attice **N**etwork **C**oding **E**rror
PAM	**P**hase **A**mplitude **M**odulation
PDF	**P**robability **D**ensity **F**unction

PID	**P**rincipal **I**deal **D**omain
PNC	**P**hysical-layer **N**etwork **C**oding
QAM	**Q**uadrature **A**mplitude **M**odulation
SC	**S**mall **C**ells
SIC	**S**uccessive **I**nterference **C**ancellation
SINR	**S**ignal to **I**nterference plus **N**oise **R**atio
SIR	**S**ignal to **I**nterference **R**atio
SNR	**S**ignal to **N**oise **R**atio
SVD	**S**ingular **V**alue **D**ecomposition
SVP	**S**hortest **V**ector **P**roblem
ZF	**Z**ero **F**orcing

Bibliography

[1] E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

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

[3] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[4] C. Mobile, "C-RAN: the road towards green RAN," *White Paper, ver*, vol. 2, 2011.

[5] H. Q. Ngo, A. E. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *CoRR*, vol. abs/1602.08232, 2016. [Online]. Available: http://arxiv.org/abs/1602.08232

[6] S. Sahraei and M. Gastpar, "Compute-and-forward: Finding the best equation," in *Proc. IEEE Allerton*, Sep. 2014, pp. 227–233.

[7] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[8] A. Sanderovich, O. Somekh, and S. Shamai, "Uplink macro diversity with limited backhaul capacity," in *Proc. 2007 IEEE ISIT*, Nice, Jun. 2007, pp. 11–15.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[10] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory.*, vol. 51, no. 9, pp. 3037–3063, Sep. 2005.

[11] L. Zhou and W. Yu, "Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 1981–1993, Oct. 2013.

[12] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory.*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[13] Y. Zhao, R. Adve, and T. J. Lim, "Symbol error rate of selection amplify-and-forward relay systems," *IEEE Commun. Lett.*, vol. 10, no. 11, pp. 757–759, Nov. 2006.

[14] 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 Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[15] Y. G. Lim, C. B. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6827–6842, Dec. 2015.

[16] E. Bjornson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[17] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[18] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory.*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[19] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc IEEE SPAWC*, Stockholm, Jun. 2015, pp. 201–205.

[20] H. Q. Ngo, L. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *CoRR*, vol. abs/1702.07601, 2017. [Online]. Available: http://arxiv.org/abs/1702.07601

[21] R. Ahlswede, N. Cai, S. Y. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory.*, vol. 46, no. 4, pp. 1204–1216, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1109/18.850663

[22] S. Zhang, S. C. Liew, and P. P. Lam, "Hot topic: Physical-layer network coding," in *Proc. ACM Mobicom*, Los Angeles, Sep. 2006, pp. 358–365.

[23] T. Koike-Akino, P. Popovski, and V. Tarokh, "Optimized constellations for two-way wireless relaying with physical network coding," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 5, pp. 773–787, Jun. 2009.

[24] S. Shukla, V. T. Muralidharan, and B. S. Rajan, "Wireless network-coded three-way relaying using latin cubes," in *Proc. IEEE PIMRC*, Sydney, Sep. 2012, pp. 197–203.

[25] S. Shukla and B. S. Rajan, "Wireless network-coded four-way relaying using latin hyper-cubes," in *Proc. IEEE WCNC*, Shanghai, Apr. 2013, pp. 2410–2415.

[26] D. Fang and A. Burr, "The uplink of distributed MIMO: Wireless network coding versus coordinated multipoint," *IEEE Commun. Lett.*, vol. 19, no. 7, pp. 1229–1232, Jul. 2015.

[27] A. Burr, D. Fang, and M. Molu, "Linear wireless physical-layer network coding based on binary matrices for multilayer relay networks," in *Proc. IEEE ISWCS*, Barcelona, Aug. 2014, pp. 982–986.

[28] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory.*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[29] C. Feng, D. Silva, and F. Kschischang, "An algebraic approach to physical-layer network coding," *IEEE Trans. Inf. Theory.*, vol. 59, no. 11, pp. 7576–7596, Nov. 2013.

[30] J. H. Conway, N. J. A. Sloane, and E. Bannai, *Sphere-packings, Lattices, and Groups*. New York, NY, USA: Springer-Verlag New York, Inc., 1987.

[31] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. New York, NY, USA: Cambridge University Press, 2014.

[32] U. Erez and R. Zamir, "Achieving 1/2 log (1+SNR) on the awgn channel with lattice encoding and decoding," *IEEE Trans. Inf. Theory.*, vol. 50, no. 10, pp. 2293–2314, Oct. 2004.

[33] R. Zamir, "Lattices are everywhere," in *Proc. IEEE ITA*, San Diego, Feb. 2009, pp. 392–421.

[34] A. Burr and D. Fang, "Linear physical layer network coding based on rings," in *Proc. IEEE WCNC*, Istanbul, Apr. 2014, pp. 370–375.

[35] Q. Sun, J. Yuan, T. Huang, and K. Shum, "Lattice network codes based on eisenstein integers," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2713–2725, Jul. 2013.

[36] N. E. Tunali, Y. Huang, J. J. Boutros, and K. R. Narayanan, "Lattices over eisenstein integers for compute-and-forward," *IEEE Trans. Inf. Theory.*, vol. 61, no. 10, pp. 5306–5321, Jul. 2015.

[37] W. Liu and C. Ling, "Efficient integer coefficient search for compute-and-forward," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8039–8050, Dec. 2016.

[38] Y. Wang, A. G. Burr, Q. Huang, and M. M. Molu, "A multilevel framework for lattice network coding," *CoRR*, vol. abs/1511.03297, 2015. [Online]. Available: http://arxiv.org/abs/1511.03297

[39] J. Zhan, B. Nazer, U. Erez, and M. Gastpar, "Integer-forcing linear receivers," *CoRR*, vol. abs/1003.5966, 2010. [Online]. Available: http://arxiv.org/abs/1003.5966

[40] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2012.

[41] H. Cohen, *A Course in Computational Algebraic Number Theory*, ser. Graduate Texts in Mathematics. Springer Berlin Heidelberg, 2013. [Online]. Available: https://books.google.co.uk/books?id=5TP6CAAAQBAJ

[42] A. K. Lenstra and H. W. Lenstra, "Factoring polynomials with rational coefficients," *Muth. Ann*, pp. 515–534, 1982.

[43] H. Yao and G. W. Wornell, "Lattice-reduction-aided detectors for MIMO communication systems," in *Proc. IEEE GLOBECOM*, vol. 1, Taipei, Nov. 2002, pp. 424–428 vol.1.

[44] M. Taherzadeh, A. Mobasher, and A. K. Khandani, "LLL reduction achieves the receive diversity in MIMO decoding," *IEEE Trans. Inf. Theory.*, vol. 53, no. 12, pp. 4801–4805, Dec. 2007.

[45] R. Fischer, M. Cyran, and S. Stern, "Factorization approaches in lattice-reduction-aided and integer-forcing equalization," in *Proc IEEE IZS*, Zurich, Mar. 2016, pp. 1–5.

[46] S. Stern and R. F. H. Fischer, "Optimal factorization in lattice-reduction-aided and integer-forcing linear equalization," in *Proc IEEE SCC*, Hamburg, Feb. 2017, pp. 1–6.

[47] O. Ordentlich, J. Zhan, U. Erez, M. Gastpar, and B. Nazer, "Practical code design for compute-and-forward," in *Proc. IEEE ISIT*, St. Petersburg, Jul. 2011, pp. 1876–1880.

[48] S. N. Hong and G. Caire, "Reverse compute and forward: A low-complexity architecture for downlink distributed antenna systems," in *Proc IEEE ISIT*, Cambridge, Jul. 2012, pp. 1147–1151.

[49] S. N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory.*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[50] Y. Huang, K. R. Narayanan, and P. Wang, "Adaptive compute-and-forward with lattice codes over algebraic integers," *CoRR*, vol. abs/1501.07740, 2015. [Online]. Available: http://arxiv.org/abs/1501.07740

[51] Y. C. Huang and K. R. Narayanan, "Multistage compute-and-forward with multilevel lattice codes based on product constructions," in *Proc. IEEE ISIT*, Honolulu, Jun. 2014, pp. 2112–2116.

[52] Y. Huang, K. R. Narayanan, and N. E. Tunali, "Multistage compute-and-forward with multilevel lattice codes based on product constructions," *CoRR*, vol. abs/1401.2228, 2014. [Online]. Available: http://arxiv.org/abs/1401.2228

[53] B. Nazer, "Successive compute-and-forward," in *Proc. IEEE IZS*, Zurich, Mar. 2012, pp. 103–107.

[54] O. Ordentlich, U. Erez, and B. Nazer, "Successive integer-forcing and its sum-rate optimality," in *Proc. IEEE Allerton*, Monticello, Oct. 2013, pp. 282–292.

[55] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of computation*, vol. 44, no. 170, pp. 463–471, 1985.

[56] J. Wen and X. W. Chang, "A linearithmic time algorithm for a shortest vector problem in compute-and-forward design," in *Proc. IEEE ISIT*, Barcelona, Jul. 2016, pp. 2344–2348.

[57] Q. Huang and A. Burr, "Low complexity coefficient selection algorithms for compute-and-forward," in *Proc. IEEE VTC Spring*, Nanjing, May 2016, pp. 1–5.

[58] Y. H. Gan, C. Ling, and W. H. Mow, "Complex lattice reduction algorithm for low-complexity full-diversity MIMO detection," *IEEE Trans. Singal Process.*, vol. 57, no. 7, pp. 2701–2710, Jul. 2009.

[59] I. Stewart and D. Tall, *Algebraic Number Theory and Fermat's Last Theorem: Third Edition*, ser. Ak Peters Series. Taylor & Francis, 2001. [Online]. Available: https://books.google.co.uk/books?id=PIibasv45boC

[60] L. Wei and W. Chen, "Compute-and-forward network coding design over multi-source multi-relay channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3348–3357, Sep. 2012.

[61] M. M. Molu, K. Cumanan, and A. G. Burr, "Low-complexity compute-and-forward techniques for multisource multirelay networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 926–929, May 2016.

[62] S. Sahraei and M. Gastpar, "Polynomially solvable instances of the shortest and closest vector problems with applications to compute-and-forward," *IEEE Trans. Inf. Theory*, vol. PP, no. 99, pp. 1–1, 2017.

[63] A. Sakzad, E. Viterbo, Y. Hong, and J. Boutros, "On the ergodic rate for compute-and-forward," in *Proc. IEEE NetCod*, Cambridge, Jun. 2012, pp. 131–136.

[64] U. Madhow, D. R. Brown, S. Dasgupta, and R. Mudumbai, "Distributed massive MIMO: Algorithms, architectures and concept systems," in *Proc. IEEE ITA*, San Diego, Feb. 2014, pp. 1–7.

[65] G. N. Kamga, M. Xia, and S. Aïssa, "Spectral-efficiency analysis of massive MIMO systems in centralized and distributed schemes," *CoRR*, vol. abs/1601.04829, 2016. [Online]. Available: http://arxiv.org/abs/1601.04829

[66] D. Qiao, Y. Wu, and Y. Chen, "Massive MIMO architecture for 5G networks: Co-located, or distributed?" in *Proc. IEEE ISWCS*, Barcelona, Aug. 2014, pp. 192–197.

[67] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and K. Zheng, "Toward 5G densenets: architectural advances for effective machine-type communications over femtocells," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 134–141, Jan. 2015.

[68] E. Nayebi, A. E. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," *CoRR*, vol. abs/1702.03231, 2017. [Online]. Available: http://arxiv.org/abs/1702.03231

[69] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive MIMO with zero-forcing precoding design," *IEEE Commun. Lett*, vol. 21, no. 8, pp. 1871–1874, Aug. 2017.

[70] K. N. Pappi, P. D. Diamantoulakis, H. Otrok, and G. K. Karagiannidis, "Cloud compute-and-forward with relay cooperation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3415–3428, Jun. 2015.

[71] K. Hosseini, W. Yu, and R. S. Adve, "Large-scale MIMO versus network MIMO for multicell interference mitigation," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 930–941, Oct. 2014.

[72] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge University Press, 2005.

[73] Z. Wang, E. K. Tameh, and A. R. Nix, "Joint shadowing process in urban peer-to-peer radio channels," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 52–64, Jan. 2008.

[74] A. Burr and Q. Huang, "On quantization and network coding in wireless relay networks," in *Proc. IEEE SAM*, Rio de Janeiro, Jul. 2016, pp. 1–5.

[75] Q. Huang and A. Burr, "The capacity of Cloud-RAN: Outer bound with quantisation and constrained fronthaul load," in *Proc. IEEE WPMC*, Shenzhen, Nov. 2016, pp. 497–501.

[76] A. Burr and D. Fang, "Physical layer network coding for distributed massive MIMO," in *Proc. IEEE WSA*, Ilmenau, Mar. 2015, pp. 1–5.

[77] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai, "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. Inf. Theory.*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.

[78] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge University Press, 2012.

[79] G. D. Forney, M. D. Trott, and S.-Y. Chung, "Sphere-bound-achieving coset codes and multilevel coset codes," *IEEE Trans. Inf. Theory.*, vol. 46, no. 3, pp. 820–850, May 2000.

[80] S. Zhang, S. C. Liew, and P. P. Lam, "On the synchronization of physical-layer network coding," in *Proc. IEEE ITW*, Chengdu, Oct. 2006, pp. 404–408.

[81] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.