The uplink reception and downlink transmission in MU-MIMO for 5G

Aymen Askri

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT006

Thèse de doctorat

TELECOM Paris

IP PARIS

The Uplink Reception and Downlink Transmission in MU-MIMO for 5G

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat: Réseaux, Informations et Communications

Thèse présentée et soutenue à Palaiseau, le 12/04/2021, par

AYMEN ASKRI

Composition du Jury :

Michel Kieffer Professor, CentraleSupélec (Signals and Systems Laboratory)	Président
Samson Lasaulce Director of Research, CNRS (CRAN-CO2 Team)	Rapporteur
Elena Veronica Belmega Associate Professor/Deputy director of ETIS Laboratory (ENSEA-CY Cergy Paris University-CNRS)	Rapporteur
Jean-François Hélard Professor, INSA Rennes (Dept. Signal & Communucations)	Examinateur
Catherine Douillard Professor, IMT Atlantique (Dept. Electronics)	Examinatrice
Ghaya Rekaya-Ben Othman Professor, Télécom Paris (Dept. Comelec)	Directrice de thèse
Philippe Sehier Manager, Nokia Bell Labs (Dept. Physical Research)	Invité

"The most beautiful thing we can experience is the mysterious. It is the source of all true art and science."

Albert Einstein

Abstract

Multiple-input multiple-output (MIMO) technologies were developed to increase system capacity and offer better link reliability. They allow a dense network architecture that will allow many users to connect in the same area without experiencing slowdowns. 5G networks and beyond will use these MIMO technologies with many small antennas allowing the beam to be focused on a given area. Coupled with high-frequency bands, the use of these antennas will significantly increase throughput.

In such systems, multi-user (MU)-MIMO detection in the uplink reception and MU-MIMO precoding in the downlink transmission enable separating user data streams and pre-cancelling interference. However, some challenges have to be met under realistic conditions, such as the reasonable complexity of the decoding and precoding processes, the erroneous channel knowledge, and the adjacent cell interference. This thesis addresses all these limitations above for the uplink reception and the downlink transmission in MU-MIMO systems.

In the uplink reception, we study the well-known sphere decoding (SD) algorithm for MIMO detection. We seek to reduce its complexity which increases exponentially with the number of antennas and the constellation size. Thus, we profit from recent advances in neural networks (NNs) to develop the low-complexity NN assisted SD. We also propose the block recursive MIMO decoding, achieving almost the maximum likelihood (ML) performance. Using deep neural networks (DNNs), we suggest a new and low complex scheme for signal processing and cloud-RAN (C-RAN) detection. This DNN scheme aims to mimic the whole transmission in uplink C-RAN, which considers the quantization constraints at the radio remote units (RRUs) and the corrupted observations at the central processor (CP).

In the downlink transmission, we study the non-linear vector perturbation (VP) precoding. We design the combined VP to serve multiple users with different modulation coding schemes (MCSs). We also introduce the block VP algorithm, which merges linear and non-linear precoding to offer a tunable tradeoff between complexity and performance. To deal with the erroneous channel state information (CSI) in the downlink precoding, we develop the new CSI accuracy indicator reporting to design a novel precoder that is less sensitive to CSI errors.

Résumé

Les technologies à entrées multiples et sorties multiples (MIMO) ont été développées pour augmenter la capacité du système et offrir une meilleure fiabilité de la liaison. Ils permettent une architecture réseau dense qui permettra à de nombreux utilisateurs de se connecter dans la même zone sans subir de ralentissements. Les réseaux 5G et au-delà utiliseront ces technologies MIMO avec de nombreuses petites antennes permettant au faisceau de se concentrer sur une zone donnée. Couplées à des bandes haute fréquence, l'utilisation de ces antennes augmentera considérablement le débit.

Dans ces systèmes, la détection multi-utilisateurs (MU)-MIMO dans la réception de la liaison montante et le précodage dans la transmission de la liaison descendante permettent de séparer les flux de données utilisateur et de pré-annuler les interférences. Cependant, certains défis doivent être relevés dans des conditions réalistes telles que dans des conditions réalistes telles que la complexité raisonnable des processus de décodage et de précodage, la connaissance erronée des canaux et l'interférence des cellules adjacentes. Cette thèse aborde toutes ces limitations ci-dessus pour la réception en liaison montante et la transmission en liaison descendante dans les systèmes MU-MIMO.

Pour la réception sur la liaison montante, nous étudions l'algorithme bien connu de décodage par sphères (SD) pour la détection MIMO. Nous cherchons à réduire sa complexité qui augmente de manière exponentielle avec le nombre d'antennes et la taille de la constellation. Ainsi, nous profitons des récentes avancées dans le domaine des réseaux de neurones (NNs) pour développer le SD assisté par les NNs de faible complexité. Nous proposons également le décodage MIMO récursif par blocs, qui atteint presque la performance de maximum de vraisemblance (ML). En utilisant les réseaux neuronaux profonds (DNNs), nous suggérons un nouveau schéma peu complexe pour le traitement et la détection du signal dans la liaison montante du cloud-RAN (C-RAN). Ce schéma DNN vise à imiter toute la transmission en liaison montante C-RAN, qui prend en compte les contraintes de quantification au niveau des unités radio distantes (RRUs) et les observations corrompues au niveau du processeur central (CP).

Dans la transmission en liaison descendante, nous étudions le précodage de la perturbation vectorielle (VP) non-linéaire. Nous concevons le VP combiné pour servir plusieurs utilisateurs avec différents schémas de codage de modulation (MCSs). Nous introduisons également l'algorithme VP par blocs, qui fusionne le précodage linéaire et non-linéaire pour offrir un compromis accordable entre complexité et performance. Pour traiter les informations erronées sur l'état du canal (CSI) dans le

précodage de la liaison descendante, nous développons le nouvel indicateur de précision CSI pour concevoir un nouveau précodeur moins sensible aux erreurs CSI.

Acknowledgements

I want to express my gratitude to the people who have helped, encouraged, and supported me to carry out this thesis work.

I want to thank from the deep of my heart, my supervisor Ms Ghaya Rekaya-Ben Othman, Professor at Telecom Paris, who opened to me the doors to research and gave me a taste for it. Thank you, Ghaya, for your listening, availability, guidance, precious advice and the continuous trust you have placed in me. Without your bright ideas, experience and competence, this work and its outcome would never have seen the light of day. It was a great pleasure working with you. I am delighted to have you as my thesis supervisor, and I hope this fruitful collaboration will see its future continuity.

I want to thank the jury members very sincerely for participating in the jury of my thesis. I thank Professor Michel Kieffer for acting as the chairman and Professors Samson Lasaulce and Elena Veronica Belmega for reviewing and evaluating my work. Besides, I am grateful to Professors Jean-François Hélard and Catherine Douillard for being my thesis' examiners. I was honoured that Mr Philippe Sehier, the product strategy manager within Nokia Bell Labs in France, participated in the jury as a guest.

I would also like to thank Dr Hadi Ghauch at Telecom Paris to work and discuss together during my thesis.

And a special thanks to Florence Besnard and Chantal Cadiat for services way beyond the call of duty. Let me also thank all my colleagues and PhD friends. I express a huge thanks to Mehdi, Akram, Homa, Mehrasa, and Mustapha, with whom I had a great time that I will never forget.

Above all, my heartfelt love is devoted to my fiancée Asma, whose love, care, help and support encouraged me to work hard. Thank you, Asma, for all the beautiful moments we shared. Also, my gratitude is devoted to my young brothers, Anis and Khalil, whose love is exceptional. Last but not least, I express my deep gratitude to my parents, Abdessattar and Faouzia, for their unconditional love, sacrifices, and encouragements. I hope that this accomplishment makes all of you proud.

Contents

List of Figures

List of Tables

List of Abbreviations

AWGN	**A**dditive **W**hite **G**aussian **N**oise
BC	**B**roadcast **C**hannel
BER	**B**it **E**rror **R**ate
BFS	**B**readth **F**irst **S**earch
BS	**B**ase **S**tation
CF	**C**ompress-and-**F**orward
CP	**C**entral **P**rocessor
C-RAN	**C**loud-**R**adio **A**cces **N**etwork
CSI	**C**hannel **S**tate **I**nformation
CSIAI	**C**hannel **S**tate **I**nformation **A**ccuracy **I**ndicator
CSI-RS	**C**hannel **S**tate **I**nformation **R**eference **S**ignal
CQI	**C**hannel **Q**uality **I**ndicator
CVP	**C**losest **V**ector **P**roblem
DFE	**D**ecision **F**eedback **E**qualizer
DFS	**D**epth **F**irst **S**earch
DL	**D**eep **L**earning
DNN	**D**eep **N**eural **N**etwork
DPC	**D**irty **P**aper **C**oding
FDD	**F**requency **D**ivison **D**uplex
ICI	**I**nter **C**ell **I**nterference
IoT	**I**nternet **o**f **T**hings
ISI	**I**nter **S**ymbol **I**nterference
LTE	**L**ong **T**erm **E**volution
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MCS	**M**odulation **C**oding **S**cheme
MIMO	**M**ultiple **I**nput **M**ultiple **O**utput
ML	**M**aximum **L**ikelihood
MMSE	**M**inimum **M**ean **S**quared **E**rror
MU	**M**ulti **U**ser
NN	**N**eural **N**etwork
NR	**N**ew **R**adio
OAMP	**O**rthogonal **A**pproximate **M**essage **P**assing
PDF	**P**robability **D**ensity **F**unction
PIC	**P**artial **I**nterference **C**ancellation
PMI	**P**recoding **M**atrix **I**ndicator

QAM	**Q**uadrature **A**mplitude **M**odulation
RZF	**R**egularized **Z**ero **F**orcing
SD	**S**phere **D**ecoding
SDMA	**S**pace **D**ivision **M**ultiple **A**ccess
SE	**S**phere **E**ncoder
SGD	**S**tochastic **G**radient **D**escent
SIC	**S**uccessive **I**nterference **C**ancellation
SINR	**S**ignal-to-**I**nterference-plus-**N**oise **R**atio
SMAPE	**S**ymmetric **M**ean **A**bsolute **P**ercentage **E**rror
SNR	**S**ignal-to-**N**oise **R**atio
STBC	**S**pace **T**ime **B**lock **C**oding
SU	**S**ingle **U**ser
SVD	**S**ingular **V**alue **D**ecomposition
TDD	**T**ime **D**ivison **D**uplex
THP	**T**omlinson **H**arashima **P**recoding
UE	**U**ser **E**quipment
VP	**V**ector **P**erturbation
WiFi	**W**ireless **F**idelity
ZF	**Z**ero-**F**orcing

Chapter 1

Introduction

In the last decade, the fourth-generation (4G) [1] has become the standard for mobile users worldwide. While 4G wireless technology has covered new mediums of mobile consumption, it still requires an improved performance due to the rise of massive internet of things (IoT) devices. The 5th generation (5G) of wireless networks [2–4] comes into play to bring the level of performance needed for massive IoT. It promises order-of-magnitude improvements in throughput and latency, along with increased network flexibility and scalability. In the 3^{rd} generation partnership project (3GPP), additional aspects of the 5G New Radio (NR) specification are locked in with every release. Release-15 introduced 5G enhanced mobile broadband (eMBB) use cases [5, 6]. In 2019, Release-16 added support for ultra-reliable low-latency communication (URLLC) [5,7] and massive machine-type communication use cases (mMTC) [8] (see Figure 1.1).

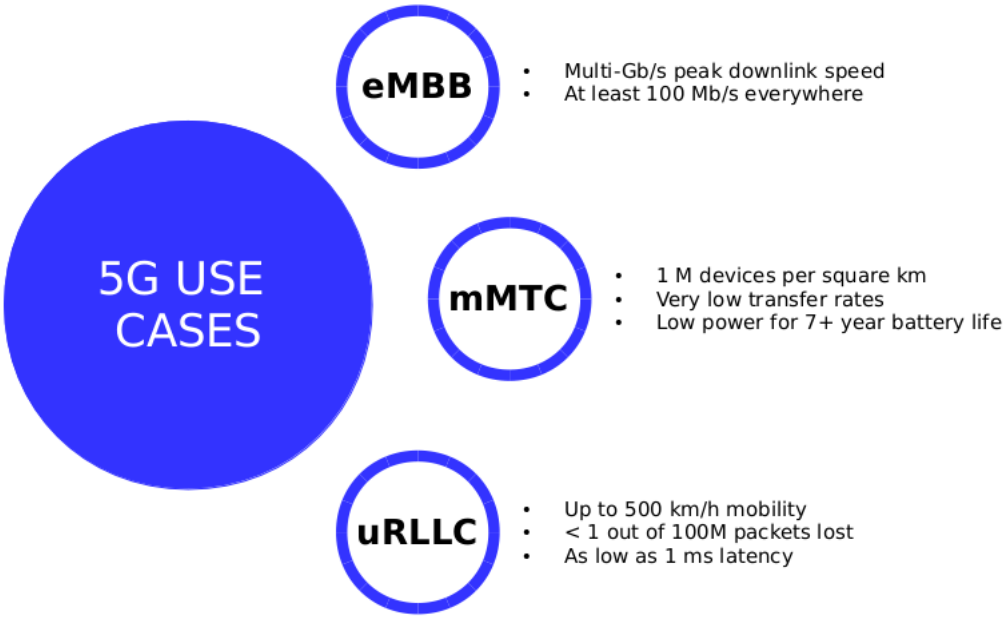


FIGURE 1.1: Primary 5G use cases.

MIMO technology is seen as one of the most promising and efficient solutions in 5G NR as it is developed to answer the spectral efficiency request [9]. MIMO technology is the traditional wireless communication technique for sending and receiving multiple data signals simultaneously over the same radio channel. It plays an imminent role in Wi-Fi communications as well as 3G and 4G networks. MIMO systems were first proposed for point-to-point communications to enhance the throughput and also transmission quality. Indeed, MIMO technology explores spatial diversity by transmitting dependent and independent data on different antennas to enhance signal reliability and boost throughput. In this perspective, space-time block coding (STBC) has been proposed to capture all the degrees of freedom offered by the MIMO system. One can site, for example, the Alamouti code [10] and the Golden code [11], designed for the open-loop case where the channel state information (CSI) is assumed to be only known at the receiver side. The coding operation consists of simple matrix multiplication, and a MIMO decoding is implemented at the receiver side. Over the years, several MIMO detection algorithms have been developed. The trade-off is the good performance in terms of bit error rate (BER) that can require high complexities. Thus, the challenge relates to the complexity and energy consumption of the signal processing at transceiver devices.

Point-to-point MIMO systems have been extended to multi-user (MU)-MIMO, where many users equipped with multiple antennas communicate with an access point also equipped with multiple antennas. We distinguish the uplink case in the multi-user scenario, where the multiple users transmit simultaneously to the base station (BS). The access point identifies and separates the different signals coming from multiple users using multi-user detection techniques. For the downlink case, the BS transmits simultaneously to multiple independent users. In this last case, it performs beamforming or precoding techniques to separate the users spatially. For that, the CSI is needed to be known at the transmitter side. These kinds of systems are called closed-loop ones. The complexity is now reported to the transmitter side, and the receiver architecture for decoding is highly simplified. Precoding enables separating user data streams and pre-cancelling interference such that one or more objective functions are satisfied under one or more constraints. Different precoding techniques exist in the literature to solve that and can be grouped into linear and non-linear. Generally, precoding techniques are equivalent to MIMO decoding but done at the transmitter side subject to the transmission power constraints.

5G NR introduces the concept of massive MIMO, which involves applying MIMO technology for higher network capacity and larger coverage [12, 13]. To transmit massive sums of data in near real-time, new 5G BSs support massive MIMO, which can include tens or even hundreds of antennas, each transmitting a unique data stream. Massive MIMO designs allow the BS to send or receive multiple signals to

or from different users at once, resulting in higher spectral efficiency and a more significant signal-to-interference-plus-noise ratio (SINR). Correspondingly, higher throughput is provided for users, and better coverage is achieved around the cell site. Massive MIMO with smart antenna techniques such as beamforming and precoding [14] is among the crucial technologies for providing higher throughput and capacity gains promised by 5G and beyond. Precoding, also known as "transmit beamforming", exploits the spatial degrees of freedom offered by the multiple transmit antennas to simultaneously serve a plurality of users in a multi-antenna wireless communication system. To implement MIMO beamforming or precoding for 5G BSs, designers must carefully select hardware and software tools to simulate, design, and test highly complex systems containing tens or even hundreds of antenna elements.

Recent years have seen a tremendous resurgence of interest in deep learning (DL) and deep neural networks (DNNs). Thanks to solid learning ability from data, DL today surpasses any other algorithm in the fields of image and speech recognition [15, 16], and can achieve comparable performance to humans on specific tasks. DNNs are currently widely used for many artificial intelligence applications, including computer science. DL techniques have recently been applied to design issues in communication systems such as channel coding, modulation and demodulation schemes, channel estimation, and data detection [17–23]. Motivated by DL technologies' performance, our work leverages recent advances in DNNs to establish several thesis results.

This PhD research study's first objective is to study the complexity and performance trade-offs involved in MIMO decoding for the uplink reception and MIMO precoding for the downlink transmission. The second objective is to design new MIMO decoding and precoding techniques to answer the technical challenges of uplink and downlink. The work is done under a collaborative project with Nokia. The purpose of this collaboration is to design and evaluate new concepts relevant to exploitable IPR (intellectual property rights) in 5G systems standardization and beyond regarding advanced massive MIMO techniques. When comparing research solutions and the current standards, there are several areas for improvements. Thus, we propose new ideas to bring enhancements as possible. Established results have been published and also patented.

The contents and contributions of the thesis chapters are summarized in the following.

First, Chapter 2 presents the fundamentals of MIMO communication systems, which include the single-user MIMO (SU-MIMO) and the multi-user MIMO (MU-MIMO).

The reciprocity of uplink and downlink processing is described as the equivalence between the decoding and precoding operations. Afterwards, relevant background materials on MIMO decoding and precoding techniques are presented, and the performance is evaluated in multi-user wireless communication systems. The presented backgrounds in this chapter help develop reception and transmission schemes in the following chapters.

Chapter 3 focuses on MIMO decoding in the uplink reception. A single cell environment is considered, where only one BS receives signals from different cell users. We are interested in the well-known sphere decoding (SD) algorithm thanks to its efficient performance. The SD requires high computational complexity for decoding; thus, we propose in this chapter two algorithm modifications to reduce its complexity while keeping almost optimal performance. The first algorithm that we propose uses a learning approach to predict the number of lattice points in the sphere and then applied in our proposed neural network (NN) assisted SD. More detail about the learning approach is provided next in the chapter. The second algorithm modification is the block recursive MIMO decoding, which divides the entire MIMO system into blocks. Then it performs sequential decoding to each block subject to some constraints. We show through simulations a complexity reduction for different block divisions than the SD algorithm while offering almost maximum likelihood (ML) performance.

Chapter 4 focuses on the uplink reception for multi-antenna systems considering this time the cloud radio access network (C-RAN) scenarios. Using DNNs, we design an efficient scheme, called QDNet, for fronthaul compression in uplink C-RAN. The proposed architecture includes the processing done at the BSs and the processing completed at the central processor (CP). With sparsely connected layers, QDNet requires less complexity to compute, and it achieves good performance compared to the existing detection algorithms.

Having studied and designed some new schemes in the uplink MU-MIMO reception, Chapter 5 focuses on the downlink MU-MIMO transmission. As precoding provides user-specific spatial channels, we develop in this chapter reliable transmission techniques for multi-user communications in the single-cell environment based on linear and non-linear precoding techniques. The transmission system design should ideally be able to cope with CSI errors, lower or eliminate error floor effects and achieve a close to the sum-capacity limit performance at a realistic computational complexity.

In the first part, we propose a combined vector perturbation (VP) precoding for MU-MIMO downlink systems to answer this challenging question. It enables an adaptive modulation scenario where users apply different modulation coding schemes (MCSs). The second part of the chapter introduces a low-complexity precoding technique which is the block VP algorithm. The proposed scheme allows for obtaining the desired diversity order by fixing the block size. Finally, Chapter 5 presents a novel transmission scheme based on reporting a new CSI accuracy indicator (CSIAI) to deal with channel imperfections at the BS. Based on this new quantity, we develop a downlink precoding technique that is less sensitive to CSI errors and improves the overall system performance.

Finally, Chapter 6 concludes the thesis by summarizing the main results and suggesting some perspectives for possible future work.

Chapter 2

The Uplink and Downlink Processing in MIMO Systems

2.1 Introduction

The explosive development of MIMO systems has granted high data rate services and an expanded variety of applications. IEEE 802.11, 3G, LTE, and 5G NR are some technologies that rely on MIMO systems. In addition to the single-user MIMO, recent progress in wireless communication systems has involved MU-MIMO scenarios' design. These communication systems have the advantage to develop new generations of wireless mobile radio systems in 5G NR and beyond. This chapter gives an insight into MU-MIMO scenarios. We firstly present the fundamental aspects of MIMO communication systems. After that, we focus on the uplink and downlink MU-MIMO schemes and describe the duality between the uplink and the downlink processing. To do so, we present the communication system model for each task. Then we introduce the MIMO decoding for the uplink reception and the MIMO precoding techniques for the downlink transmission.

2.2 Fundamental Aspects of MIMO Communication Systems

Traditional MIMO communication systems are usually referred to as single-user MIMO (SU-MIMO) or also point-to-point MIMO. The access point or the BS, in this case, communicates with only one mobile terminal (user). Both the access point and the user are equipped with multiple antennas as depicted in Figure 2.1.

The transmit antennas $(Tx_1, \ldots, Tx_{N_t})$ respectively send signals $(\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_{N_t})$ to the receive antennas $(Rx_1, \ldots, Rx_{N_r})$. The received signals are respectively denoted by

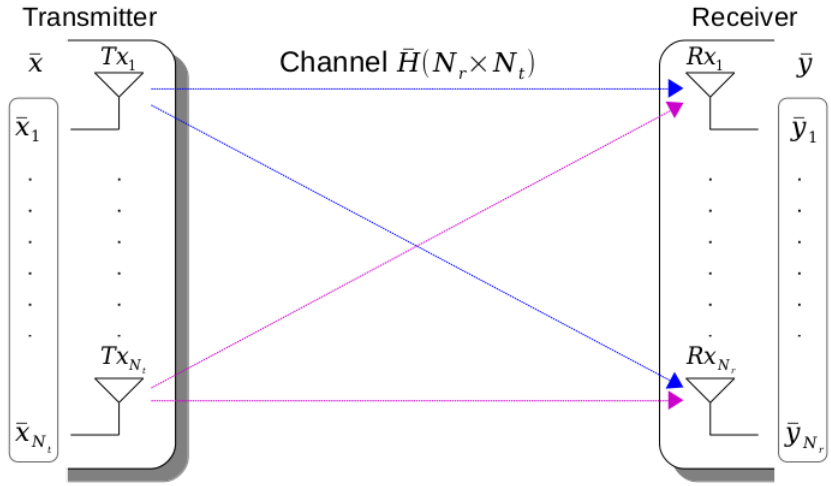


FIGURE 2.1: MIMO system model.

$(\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_{N_r})$. We express the received signal at antenna $Rx_i \; \forall \; i \in \{1, \ldots, N_r\}$ as

$$\bar{\boldsymbol{y}}_i = \sum_{j=1}^{N_t} \bar{\boldsymbol{h}}_{ij} \bar{\boldsymbol{x}}_j + \bar{\boldsymbol{w}}_i \tag{2.1}$$

The MIMO channel model can be described by the following linear system

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{H}} \bar{\boldsymbol{x}} + \bar{\boldsymbol{w}} \tag{2.2}$$

Throughout this thesis, we avoid handling complex-valued variables, and convert (2.2) to its equivalent real-valued representation by using the following convention

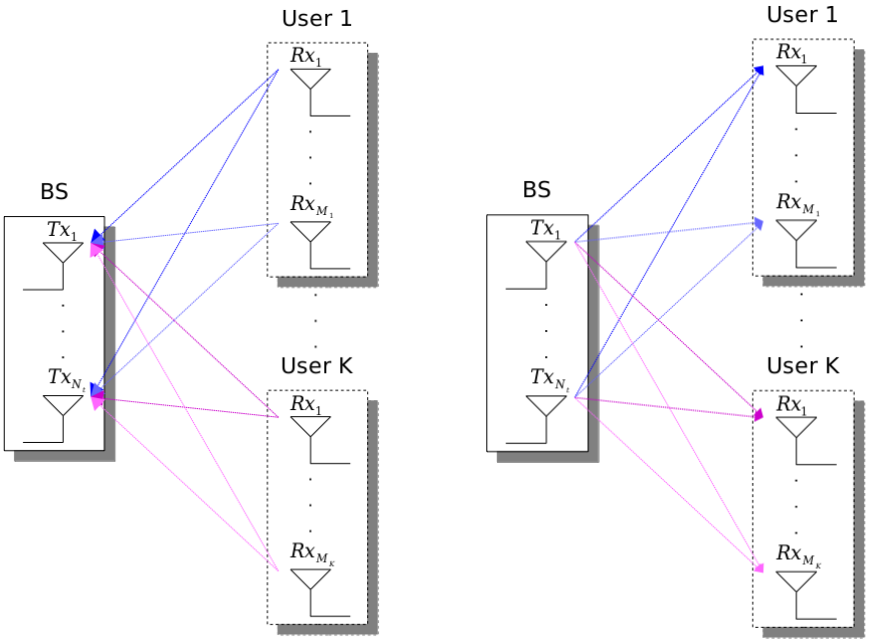
$$\boldsymbol{y} = \boldsymbol{H} \boldsymbol{x} + \boldsymbol{w} \tag{2.3}$$

where

$$\begin{aligned} \boldsymbol{y} &= \begin{bmatrix} \Re(\bar{\boldsymbol{y}}) \\ \Im(\bar{\boldsymbol{y}}) \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} \Re(\bar{\boldsymbol{x}}) \\ \Im(\bar{\boldsymbol{x}}) \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} \Re(\bar{\boldsymbol{w}}) \\ \Im(\bar{\boldsymbol{w}}) \end{bmatrix}, \\ \boldsymbol{H} &= \begin{bmatrix} \Re(\bar{\boldsymbol{H}}) & -\Im(\bar{\boldsymbol{H}}) \\ \Im(\bar{\boldsymbol{H}}) & \Re(\bar{\boldsymbol{H}}) \end{bmatrix} \end{aligned} \tag{2.4}$$

$\Re(.)$ and $\Im(.)$ are defined as the real and imaginary parts of a complex matrix or vector, respectively.

The channel's entries $\boldsymbol{H}$ are assumed to have a Rayleigh distribution that we frequently use to model multi-path fading with no direct line-of-sight (LOS) path. The



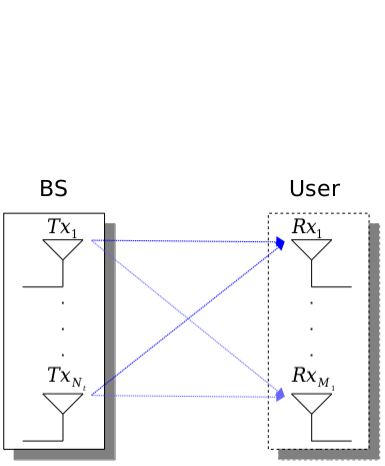
(A) Uplink MU-MIMO (B) Downlink MU-MIMO

FIGURE 2.2: MU-MIMO communication models.

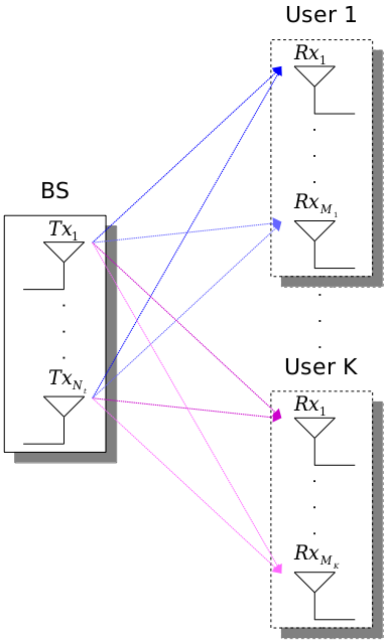
Rayleigh probability density function (PDF) may be constant (flat) or varying (selective) within blocks of transmission. If all frequency components of the signal experience the same magnitude of fading, it is called frequency flat fading. It occurs when the coherence bandwidth of the channel is larger than the signal bandwidth. On the other hand, if all the frequency components of the transmitted signal are affected by different amplitude gains and phase shifts, the fading is frequency-selective. It occurs when the shared signal bandwidth is more significant than the channel's coherence bandwidth.

In contrast to the single-user case where the communication is only with a single user, the BS in MU-MIMO systems can communicate with several mobile terminals. MU-MIMO systems promise to employ multiple receivers to improve communication rate while keeping the same level of reliability. We have the uplink transmission in the multi-user scenario where the multiple users transmit simultaneously to the BS. We also have the downlink transmission in which the BS transmits to numerous independent users. A representation of these systems is depicted in Figure 2.2. More detail is reviewed next in section 2.3.

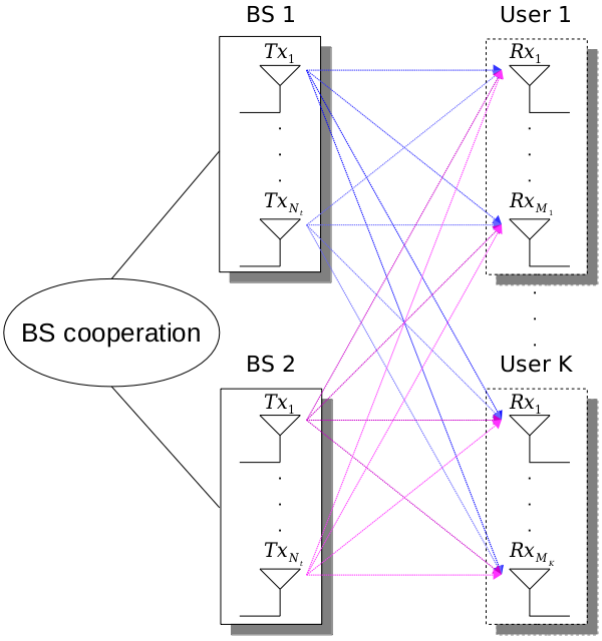
SU-MIMO and MU-MIMO systems are two possible configurations for multi-user communications. We also find the MU-MIMO with cooperation between BSs. Figure 2.3 shows the basic configurations of MIMO systems.



(A) SU-MIMO



(B) MU-MIMO



(C) MU-MIMO with BS cooperation

FIGURE 2.3: MU-MIMO configurations.

2.3 Uplink-Dowlink Duality

Transmission schemes for MU-MIMO systems comprise both uplink MU-MIMO and downlink MU-MIMO. We assume that the BS is equipped with M antennas, and there are K users; each kth user has M_k antennas.

In the uplink transmission, the BS figures out the channel information and decodes the data streams from multiple transmit antennas of users. Let $\boldsymbol{s}_k$ be the transmitted signal vector of user $k \in \{1, \ldots, K\}$, and $\boldsymbol{H}_k$ is the channel matrix from the kth user to the BS. When an additive noise is present, the received signal vector at the BS is expressed by

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{H}_k \boldsymbol{s}_k + \boldsymbol{w} \tag{2.5}$$

where $\boldsymbol{w}$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ_w^2. The number of transmitting antennas from all users is no more than the number of receiving antennas at the BS. In other words, the following constraint should be satisfied in the uplink scenario

$$\sum_{k=1}^{K} M_k \leq M \tag{2.6}$$

In fact, to solve a simultaneous linear system, the number of equations must be greater than or equal to the number of variables. Hence, the BS should have at least $M = \sum_{k=1}^{K} M_k$ receiving antennas, i.e., $M \geq \sum_{k=1}^{K} M_k$.

Users in the uplink have no channel knowledge, while the BS is alone exploiting the MIMO capacity. Hence, a complicated algorithm is required to decode the received signal at the BS. For that reason, MIMO decoding techniques have been studied extensively in the literature. One straightforward way to estimate the transmitted signal $\boldsymbol{s} = (\boldsymbol{s}_1^T, \ldots, \boldsymbol{s}_K^T)^T$ from the received signal $\boldsymbol{y}$ is to multiply $\boldsymbol{y}$ with an inverse channel matrix, such as zero-forcing (ZF) or minimum mean squared error (MMSE) equalizers. However, this is not the optimal detection that we can achieve using the ML criterion. Optimum decoding which gives ML performance is reviewed later in section 2.4.

In the downlink transmission, the K users are simultaneously receiving signals from the BS. The transmitted signal vector $\boldsymbol{x}$ is expressed as the sum of signals intended to users

$$\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{d}_k \tag{2.7}$$

The channel matrix between the kth user and the BS is denoted by $\boldsymbol{H}_k$. The received signal vector at each kth user is given by

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{x} + \boldsymbol{w}_k \; ; \; k \in \{1, \ldots, K\} \tag{2.8}$$

where $\boldsymbol{w}_k$ is the AWGN noise. Equation (2.8) can be also written as

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{d}_k + \sum_{j \neq k}^{K} \boldsymbol{H}_k \boldsymbol{d}_j + \boldsymbol{w}_k \; ; \; k \in \{1, \ldots, K\} \tag{2.9}$$

The sum term in equation (2.9) represents the interference signal by the cause of multiple users. Processing techniques are used at the BS, such as beamforming and precoding, to mitigate the multi-user interference and improve system performance. In the context of MU-MIMO systems, precoding enables emitting multiple data streams from the transmit antennas with independent and appropriate weightings such that the link throughput is maximized at the receiver output. Precoding, also known as transmit beamforming, exploits the spatial degrees of freedom offered by the multiple transmit antennas to serve users' plurality in a multi-antenna communication system simultaneously. The system's effective signal-to-noise ratio (SNR) is increased, and the receiver architecture is potentially simplified. The complexity is reported to the transmitter side, and the decoding is no more involved at the receiver side. For example, with perfect knowledge of the channel $\boldsymbol{H}$, the BS can perform downlink precoding, as shown in the following equation

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{s} \tag{2.10}$$

where $\boldsymbol{x}$ is the transmitted signal, $\boldsymbol{s}$ is the data symbol vector, and $\boldsymbol{F}$ is the precoding matrix designed to suppress the channel effect. Here, the main problem is how to obtain the channel knowledge at the transmitter side. Most current wireless standards allocate a feedback channel to transmit CSI reference signals (CSI-RS) to the BS. This feedback solution can work in FDD and TDD systems.

We notice duality between the uplink and downlink processing at the BS after reception or before transmission through multiple antennas. Both communication schemes utilize channel information in order to suppress its effect. Decoding operations are similar to that of precoding, and most techniques perform an inverse channel for detection or precoding. We primarily distinguish linear and non-linear processing. Figure 2.4 summarizes the primary equivalent decoding and precoding techniques covered in the literature for uplink and downlink scenarios. We can note, for example, the ZF and the MMSE in both decoding and precoding operations. We also find the decision feedback equalizer (DFE) in MIMO decoding, equivalent to the Tomlinson Harashima precoding (THP). An exemplary of optimal detection

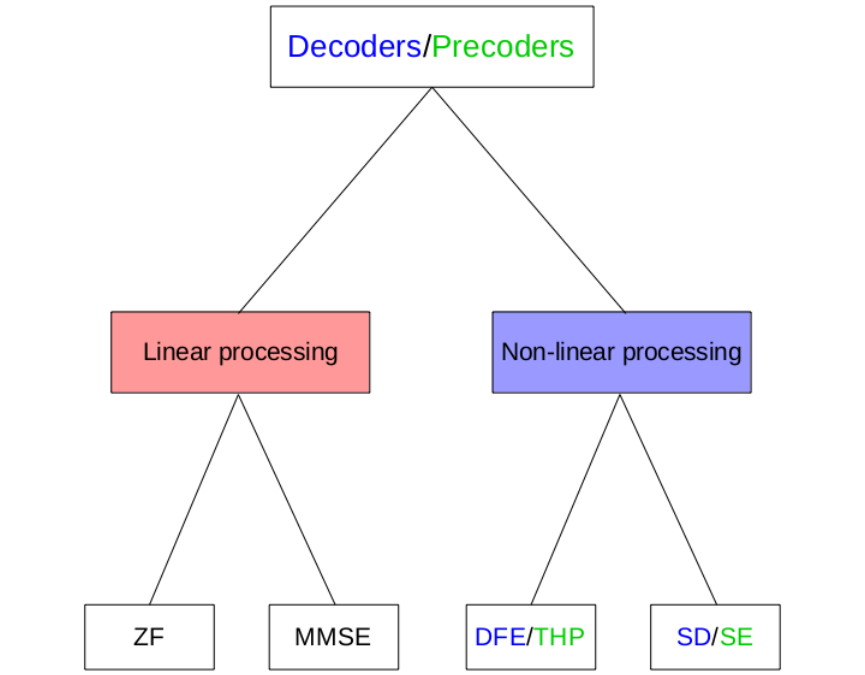


FIGURE 2.4: Similarity of primary decoding and precoding schemes.

algorithms is the SD which gives the best system performance. An equivalent algorithm for precoding is the sphere encoder (SE) which is the same version of the SD, performed at the transmitter side.

2.4 MIMO Decoding Techniques

Several MIMO detection algorithms have been proposed over the years and covered extensively in the literature. On the one hand, we categorize the sub-optimum decoding, which includes both linear and non-linear techniques such as the ZF, the MMSE and the DFE algorithms. On the other hand, we categorize the optimum decoding, which gives ML detection. It includes techniques based on lattice representation and sequential algorithms. We recognize the well-known SD algorithm using the depth-first search (DFS) and the stack decoder algorithm using the breadth-first search (BFS). Figure 2.5 summarizes the primary MIMO decoding techniques, which mainly fall into two categories. These methods are described in detail next.

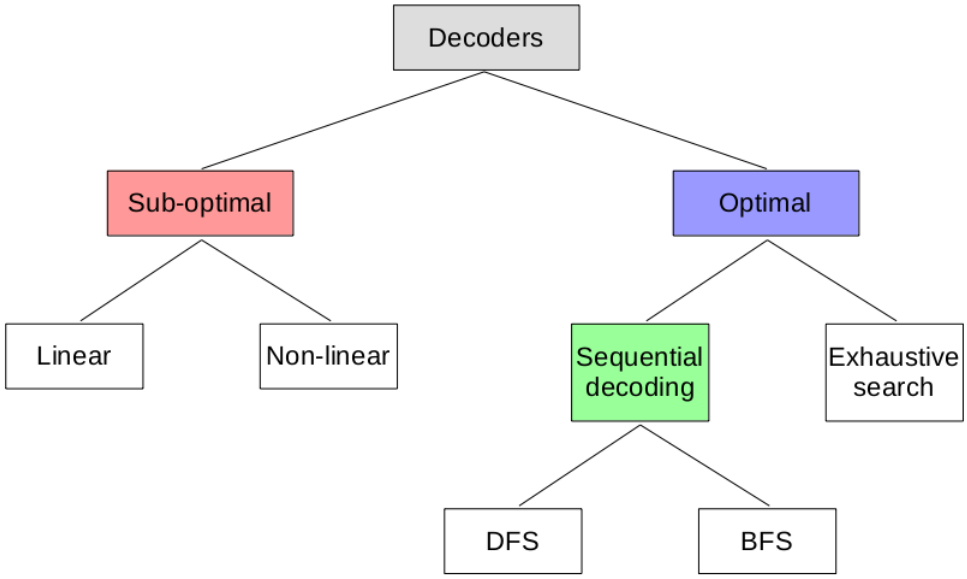


FIGURE 2.5: Primary MIMO decoding techniques.

2.4.1 Sub-optimal decoding

Consider K single antenna users transmitting signals to an N antenna BS. The received signal from all users at the BS can be expressed as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{w} \tag{2.11}$$

The purpose behind MIMO decoding is to find an estimation of the transmitted signal vector $\boldsymbol{s}$. The main decoding techniques can be summarized as follows.

Zero-forcing

The traditional ZF receiver applies the pseudo-inverse of the channel to bring down the inter-symbol interference (ISI) to zero. Knowing the channel $\boldsymbol{H}$, it multiplies the received signal with

$$\boldsymbol{H}^{\dagger} = (\boldsymbol{H}^{H}\boldsymbol{H})^{-1}\boldsymbol{H}^{H} \tag{2.12}$$

Then the estimated symbol vector is detected using some quantification operation to find out the constellation point with the nearest Euclidean metric. The ZF receiver removes all ISI and is excellent when the channel is orthogonal. When there is noise, the ZF equalizer will amplify that noise greatly by cause of the ill-conditioned $\boldsymbol{H}^{H}\boldsymbol{H}$.

The covariance matrix of the resulting noise $\tilde{\boldsymbol{w}} = \boldsymbol{H}^{\dagger}\boldsymbol{w}$ becomes then

$$\begin{aligned} \boldsymbol{K}_{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}} &= \boldsymbol{H}^{\dagger}\boldsymbol{K}_{\boldsymbol{w}\boldsymbol{w}}(\boldsymbol{H}^{\dagger})^{T} \\ &= \sigma_w^2(\boldsymbol{H}^H\boldsymbol{H})^{-1} \end{aligned} \tag{2.13}$$

Consider the eigenvalue decomposition of $\boldsymbol{H}^H\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^H$, where $\boldsymbol{Q}$ is the square unitary matrix, and $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues ($\boldsymbol{\Lambda}_{ii} = \lambda_i > 0$). The covariance matrix of $\tilde{\boldsymbol{w}}$ can be written as

$$\boldsymbol{K}_{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}} = \sigma_w^2\boldsymbol{Q}\ \mathrm{diag}(\lambda_1^{-1},\ldots,\lambda_K^{-1})\ \boldsymbol{Q}^H \tag{2.14}$$

Now it is clear from (2.14) that the noise amplification is caused by the small eigenvalues of the ill-conditioned matrix $\boldsymbol{H}^H\boldsymbol{H}$.

Zero-forcing DFE

A decision feedback equalizer (DFE) is more effective than linear receivers. What characterizes a DFE is past symbol decisions to eliminate the ISI caused by previously detected symbols on the current symbol being estimated. The DFE equalizer comprises two filters. The first one is called a feed-forward filter, similar to a linear equalizer. Its input consists of the samples of the received signal. The second one is called a feedback filter that is also working to remove ISI, operates on noiseless quantized levels, and thus its output is free of channel noise.

The ZF criterion is applied in the following to select the taps. The received signal $\boldsymbol{y}$ in equation (2.11) can be written as

$$\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{s} + \boldsymbol{w} \tag{2.15}$$

where $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{R}$ is the "QR" decomposition of the channel $\boldsymbol{H}$, $\boldsymbol{R}$ is an upper triangular matrix, and $\boldsymbol{Q}$ is an unitary one, i.e., $\boldsymbol{Q}^H\boldsymbol{Q} = \boldsymbol{I}$. First off, the received signal is multiplied by $\boldsymbol{Q}^H$ to obtain

$$\tilde{\boldsymbol{y}} = \boldsymbol{Q}^H\boldsymbol{y} = \boldsymbol{R}\boldsymbol{s} + \underbrace{\tilde{\boldsymbol{w}}}_{=\boldsymbol{Q}^H\boldsymbol{w}} \tag{2.16}$$

Then the ZF-DFE estimates successively the symbol vector $\hat{\boldsymbol{s}} = (\hat{\boldsymbol{s}}_1^T,\ldots,\hat{\boldsymbol{s}}_K^T)^T$ as

$$\hat{\boldsymbol{s}}_k = \begin{cases} \mathcal{Q}\left(\frac{\tilde{\boldsymbol{y}}_K}{\boldsymbol{R}_{KK}}\right) & k = K \\ \mathcal{Q}\left(\frac{1}{\boldsymbol{R}_{kk}}(\tilde{\boldsymbol{y}}_k - \sum_{j=k+1}^{K}\boldsymbol{R}_{kj}\hat{\boldsymbol{s}}_j)\right) & k \in \{1,\ldots,K-1\} \end{cases} \tag{2.17}$$

where $\mathcal{Q}(.)$ is the quantification operation used to estimate the constellation point with the nearest Euclidean metric.

Minimum Mean Squared Error

The ZF receiver amplifies the noise to invert the channel completely. A more balanced receiver, in this case, is the MMSE, which does not usually eliminate ISI completely but instead minimizes the total power of the noise and ISI components. The MSE which we aim to minimize in this method can be expressed as

$$\mathbb{E}_{\boldsymbol{w}}\left(\|\hat{\boldsymbol{s}}-\boldsymbol{s}\|^2 \mid \boldsymbol{H},\boldsymbol{s}\right) \tag{2.18}$$

where $\mathbb{E}_{\boldsymbol{w}}(.)$ denotes the expectation over $\boldsymbol{w}$, and $\hat{\boldsymbol{s}} = \boldsymbol{F}\boldsymbol{y}$ is the received signal multiplied by the linear equalizer $\boldsymbol{F}$. Deriving (2.18), the optimum filter $\boldsymbol{F}_{\text{MMSE}}$ minimizing the MSE can be written as

$$\boldsymbol{F}_{\text{MMSE}} = \left(\boldsymbol{H}^H\boldsymbol{H} + \frac{\sigma_w^2}{\sigma_s^2}I\right)^{-1}\boldsymbol{H}^H \tag{2.19}$$

where $\sigma_s^2 = \mathbb{E}(|\boldsymbol{s}_i|^2)$, is the average power of the $\boldsymbol{s}$ vector components. The MMSE criterion allows us to achieve better performances than the ZF, particularly for low-to-moderate SNR range. However, at high SNR regime, the MMSE receiver is similar to that of the ZF

$$\boldsymbol{F}_{\text{MMSE}} = \begin{cases} \frac{\sigma_s^2}{\sigma_w^2}\boldsymbol{H}^H & \text{SNR} \to 0 \\ \\ \boldsymbol{F}_{\text{ZF}} & \text{SNR} \to \infty \end{cases} \tag{2.20}$$

2.4.2 Optimal decoding

By considering uniformly distributed constellation points, optimal detection can be achieved with the ML criterion. For a known $\boldsymbol{H}$, ML can be implemented by finding the transmitted signal vector that minimizes the Euclidean distance to the received signal vector $\boldsymbol{y}$ as shown in the following equation

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\text{argmin}}\ \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|^2 \tag{2.21}$$

Unfortunately, ML's computational complexity is exponential with the number of antennas and the constellation size, making the exhaustive search unsuitable for practical implementations. Search tree algorithms significantly reduce this complexity, whereby they offer ML performance. We investigate the following two widely used optimum techniques, the SD [24], and the stack decoder [25].

Sphere decoding

The mathematical analysis of the SD algorithm was presented by Fincke and Pohst in [26], the geometric interpretation by Vieterbo and Biglieri in [27], and the practical implementation for fading channels by Vieterbo and Boutros in [24].

Consider equation (2.11) which gives the lattice representation of the MIMO system

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{w}$$

The principle goal of the SD is to look for the nearest point in a sphere of a given radius, centred on the received point, i.e.,

$$\begin{aligned} \underset{\boldsymbol{s}}{\text{minimize}} \quad & \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2 \\ \text{subject to} \quad & \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2 \leq r \end{aligned} \tag{2.22}$$

If no candidate has been found, the sphere radius r is enlarged, and the search is restarted. We look over points from the sphere surface inwards.

Let us go back to equation (2.16) which represents the new coordinate system

$$\tilde{\boldsymbol{y}} = \boldsymbol{Q}^H\boldsymbol{y} = \boldsymbol{R}\boldsymbol{s} + \underbrace{\tilde{\boldsymbol{w}}}_{=\boldsymbol{Q}^H\boldsymbol{w}}$$

$\boldsymbol{Q}$ is a unitary matrix, i.e., $\boldsymbol{Q}^H\boldsymbol{Q} = \boldsymbol{I}$, and the multiplication by $\boldsymbol{Q}^H$ does not fluctuate the previous system. Now we have an equivalent Euclidean distance to minimize to the signal vector $\tilde{\boldsymbol{y}}$ in the lattice $\boldsymbol{\Lambda}_R$ subject to the constellation constraints

$$\begin{aligned} \underset{\boldsymbol{s}}{\text{minimize}} \quad & \|\tilde{\boldsymbol{y}} - \boldsymbol{R}\boldsymbol{s}\|^2 \\ \text{subject to} \quad & \|\tilde{\boldsymbol{y}} - \boldsymbol{R}\boldsymbol{s}\|^2 \leq r \end{aligned} \tag{2.23}$$

Let $\boldsymbol{\rho}$ and $\boldsymbol{\xi}$ be defined as follows

$$\begin{aligned} \boldsymbol{\rho} &= \boldsymbol{R}^{-1}\tilde{\boldsymbol{y}} \\ \boldsymbol{\xi} &= \boldsymbol{\rho} - \boldsymbol{s} \end{aligned} \tag{2.24}$$

By substituting in (2.23), we get

$$\|\boldsymbol{R}\boldsymbol{\xi}\|^2 = \boldsymbol{q}_{KK}\boldsymbol{\xi}_K^2 + \sum_{i=1}^{K-1} \boldsymbol{q}_{ii}\Big(\boldsymbol{\xi}_i + \sum_{j=i+1}^{K} \boldsymbol{q}_{ij}\boldsymbol{\xi}_j\Big)^2 \leq r \tag{2.25}$$

where

$$\begin{aligned} \boldsymbol{q}_{ii} &= \boldsymbol{R}_{ii}^2,\ i \in \{1,\ldots,K\} \\ \boldsymbol{q}_{ij} &= \frac{\boldsymbol{R}_{ij}}{\boldsymbol{R}_{ii}},\ j \in \{i+1,\ldots,K\} \end{aligned} \tag{2.26}$$

Based on some mathematical analysis as described in [28], an interval $I_i = \left[b_{\text{inf}}^{(i)}, b_{\text{sup}}^{(i)}\right]$ is derived to represent the bounds of $\boldsymbol{s}_i$. To find the closest point, the SD algorithm visits the components of I_i starting from $i = K$. The cumulative weight $w(\boldsymbol{s}_i)$ is evaluated for each visited node $\boldsymbol{s}_i$

$$w(\boldsymbol{s}_i) = \sum_{j=i}^{K} w_j(\boldsymbol{s}_j) \tag{2.27}$$

where $w_j(\boldsymbol{s}_j)$ is the weight of $\boldsymbol{s}_j$ which can be written as

$$w_j(\boldsymbol{s}_j) = \begin{cases} \boldsymbol{q}_{KK}\boldsymbol{\xi}_K^2 & j = K \\ \boldsymbol{q}_{jj}\left(\boldsymbol{\xi}_j + \sum_{k=j+1}^{K} \boldsymbol{q}_{jk}\boldsymbol{\xi}_k\right)^2 & j \in \{1,\ldots,K-1\} \end{cases} \tag{2.28}$$

A node located at the bottom level, i.e., $i = 1$, is called a leaf node. The SD algorithm is based on a DFS strategy. If a leaf node is reached not satisfying the metric constraint, the path leading to that leaf node downs to that parent node, and the SD algorithm continuous the search tree to reach the other child nodes. When a leaf node is reached, satisfying the metric constraint, the sphere radius is updated to the found distance. Then the SD algorithm finds the new bounds and restarts the search tree with the new metric constraint. Figure 2.6 shows a simplified diagram of the search tree that would be performed in a 2×2 system with 4-QAM (± 1) modulation. The curve indicates an initial metric constraint, and the dashed lines indicate the discarded paths due to that metric constraint.

Figure 2.7 shows the BER performances of the different sub-optimal decoding techniques compared to the SD algorithm. We consider $K = 8$ single antenna users transmitting to a BS with $M = 8$ antennas. The MIMO channel is Rayleigh fading with 16-QAM input alphabets. We plot the BER performance versus the SNR which is expressed as

$$\text{SNR} = 10\log_{10}\left(\frac{K * \sigma_s^2}{log_2(q)\sigma_w^2}\right) \tag{2.29}$$

where σ_s^2 is the average power of the $\boldsymbol{s}$ vector components belonging to the q-QAM constellation. It is well-observed from Figure 2.7 that the SD algorithm significantly outperforms the linear receivers and the DFE equalizer due to the diversity gain.

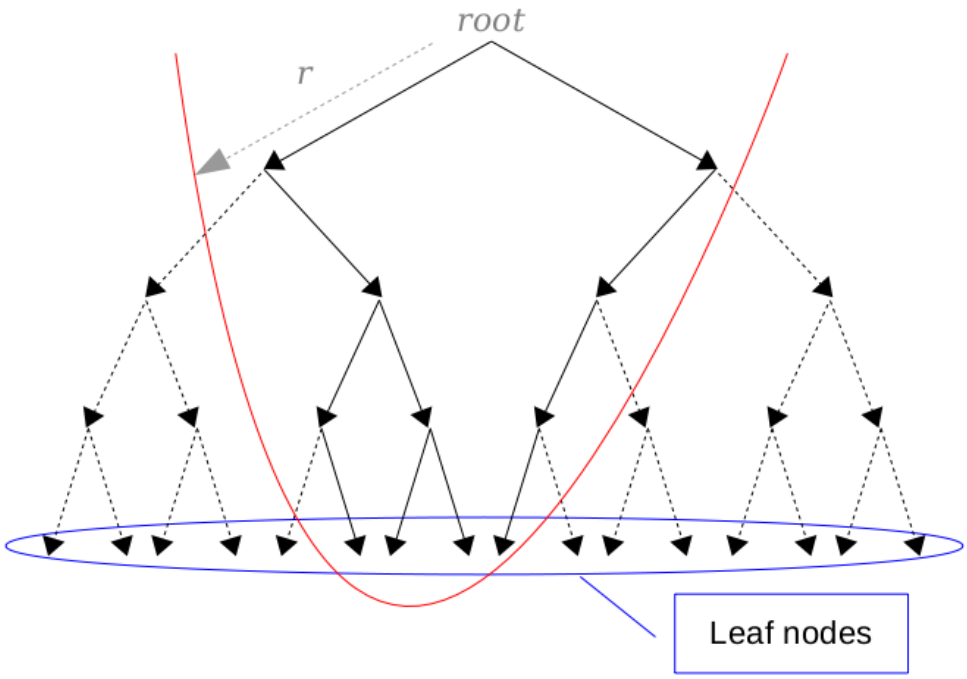


FIGURE 2.6: Diagram of the search tree in the SD algorithm for a 2 $\times$ 2 system with 4-QAM constellation.

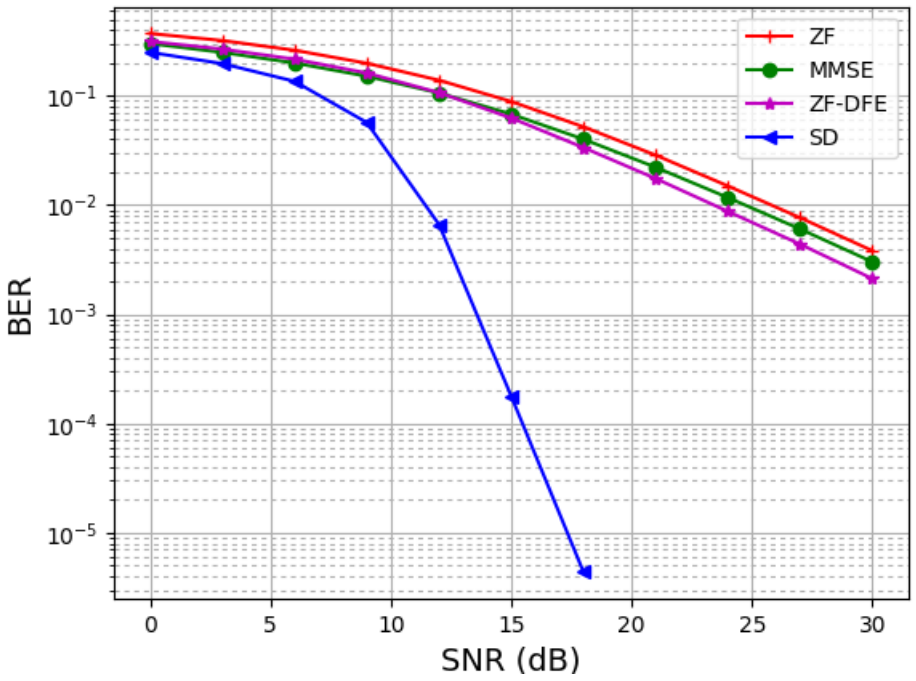


FIGURE 2.7: BER performance of MIMO decoding techniques for an 8 $\times$ 8 MIMO system with 16-QAM constellation.

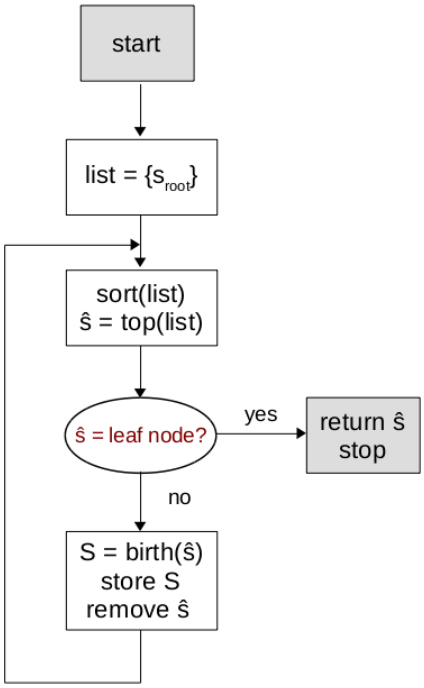


FIGURE 2.8: Flowchart of the stack decoding algorithm.

Stack decoder

The stack decoder is based on a BFS strategy. It starts by generating all child nodes at the top level, i.e., $i = K$, and computes their respective weights. The nodes are then stored in a stack in increasing order of their weights. After that, the stack decoder takes the top node of the stack having the best cost, generates its children, computes their costs, stores them in the stack, and removes the top node just being expanded. The stack is then reordered, and the same processing is performed until a leaf node is on the top of the stack. The path corresponding to the leaf node represents the ML solution. Figure 2.8 shows a flowchart of the stack decoder algorithm.

For an increasing constellation size, the traditional stack decoder requires a high computational complexity since the algorithm goes on all the nodes. Many of these nodes have a high cost and can not lead to the optimal solution; thus, they should not be visited. The spherical-bound (SB) stack decoder has been proposed in [29] to reduce the complexity. It combines the original stack decoder with the search region of the SD algorithm. At each level of the tree, bounds are imposed on the children weights to be stored in the stack. Nodes that do not satisfy the metric constraint are discarded, and hence the number of visited nodes is reduced while preserving ML performance. We should note that the spherical bounds are made larger if no child satisfying the metric constraint is found.

2.5 MIMO Precoding Techniques

The sum capacity of an MU-MIMO broadcast channel (BC) is achieved using the dirty paper coding (DPC) technique. However, the DPC method has very high complexity. Therefore, many precoding alternatives are proposed offering reasonable complexity. These precoding techniques can be grouped into two categories based on linear or non-linear processing.

Exemplary linear precoding techniques include the ZF and the regularized-ZF (RZF). ZF precoding's ability to entirely cancel out multi-user interference makes it useful for high SNR regimes at the expense of losing some signal gain. However, ZF precoding performs far from optimal in the noise-limited regime, particularly when the number of served users approaches transmitting antennas. When using ZF precoding, the transmitted vector is filtered using the channel matrix's pseudo-inverse, which requires a high transmission power, especially when it is ill-conditioned. Non-linear precoding schemes have been proposed to improve the performance of linear precoding. Tomlinson-Harashima precoding (THP) and vector perturbation (VP) are two well known non-linear schemes.

2.5.1 Linear precoding

Consider an MU-MIMO BC composed of an M antenna BS and a group of $K(\leq M)$ non-cooperative single antenna users. Let the channel vector from the BS to the kth user be $\boldsymbol{H}_k^T = (\boldsymbol{h}_{k1}, \ldots, \boldsymbol{h}_{kM})^T$ where $\boldsymbol{h}_{kj}$ denotes the channel gain between the jth transmit antenna and the kth user. We suppose the channel gains to be independent and identically distributed (i.i.d) complex Gaussian random variables with zero-mean and unit variance. When $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_M^T)^T$ is the transmit vector and $\boldsymbol{w}_k$ is the zero-mean complex Gaussian noise with variance σ_w^2, the received signal at the kth user is given by

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{x} + \boldsymbol{w}_k;\ k \in 1, \ldots, K \tag{2.30}$$

Considering all users, the composite channel can be written as $\boldsymbol{H} = [\boldsymbol{H}_1^T, \ldots, \boldsymbol{H}_K^T]^T$. Then the received signal of all users $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_K^T)^T$ can be expressed as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{w} \tag{2.31}$$

where $\boldsymbol{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_K^T)^T$ is the AWGN vector. The transmit signal $\boldsymbol{x}$ is built by multiplying the data symbol vector $\boldsymbol{s}$ with the precoding matrix $\boldsymbol{F}$; thus $\boldsymbol{x}$ can be written as

$$\boldsymbol{x} = \sqrt{\frac{P}{\gamma}} \boldsymbol{F}\boldsymbol{s} \tag{2.32}$$

where γ is the power scaling factor chosen to maintain the power constraint at the BS such that $\|\boldsymbol{x}\|^2 = P$. Thus γ can be defined as

$$\gamma = \|\boldsymbol{F}\boldsymbol{s}\|^2 \tag{2.33}$$

ZF precoding

The ZF precoder [30] can be easily found by taking the pseudo-inverse of $\boldsymbol{H}$

$$\boldsymbol{F} = \boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H)^{-1} \tag{2.34}$$

It has been shown in [31] that ZF precoding performance is relatively low. The sum rate of ZF saturates over increasing K and does not improve. Besides, the achieved diversity without channel coding is shown to be $N - K + 1$. It can be observed that γ characterizes the poor performance of ZF precoding. A considerable quantity of the transmitted power is consumed by the smallest eigenvalues of $\boldsymbol{H}\boldsymbol{H}^H$.

Regularized-ZF precoding

The RZF precoding was proposed in [31] to surmount the issue due to the ill-conditioned $\boldsymbol{H}\boldsymbol{H}^H$. Instead of building $\boldsymbol{F}$ as in (2.34), the RZF precoding uses

$$\boldsymbol{F} = \boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H + \alpha\boldsymbol{I})^{-1} \tag{2.35}$$

where α is the regularization coefficient. By introducing α, the detrimental effect caused by the smallest eigenvalues can be controlled, leading to better system performance. It is proved in [31] that $\alpha = K\sigma_w^2/P$ is the optimal regularization coefficient in the sense of maximizing the SINR. The RZF converges to the ZF precoder at a high SNR level, and thus, the diversity order achieved is $N - K + 1$. We refer to the RZF with an optimal regularization term as a linear MMSE precoder.

2.5.2 Non-linear precoding

Compared to linear processing, non-linear precoding has been shown to achieve better sum-rate performances. The gain is achieved by an additional signal processing manageable at the BS in the downlink scenario. THP [32] and VP [33], two known non-linear precoding schemes are investigated in the following. It is proven in [34] that VP precoding achieves full diversity equal to N.

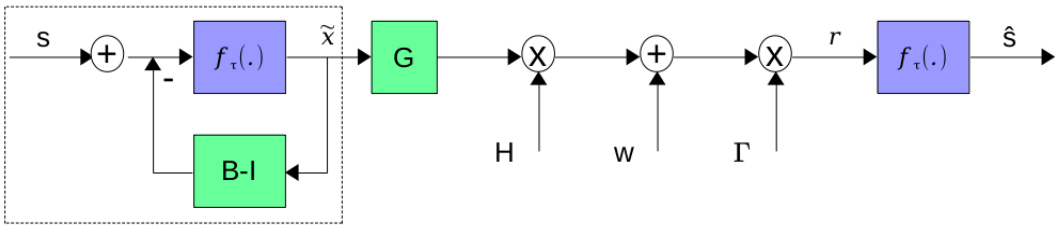


FIGURE 2.9: THP system model.

Tomlinson-Harashima precoding

Figure 2.9 shows a simple system model of THP where the modulo function $f_\tau(.)$ will be described next. The symbol vector $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_1^T, \ldots, \tilde{\boldsymbol{x}}_M^T)^T$ is generated successively. $\boldsymbol{B} - \boldsymbol{I}$ is chosen to be a strictly lower triangular matrix with one's values in the main diagonal. Without the modulo function $f_\tau(.)$, we have

$$\tilde{\boldsymbol{x}} = \boldsymbol{s} - (\boldsymbol{B} - \boldsymbol{I})\tilde{\boldsymbol{x}} = \boldsymbol{B}^{-1}\boldsymbol{s} \tag{2.36}$$

$\boldsymbol{B}$ depends on the random channel matrix $\boldsymbol{H}$, thus the magnitude of some $\tilde{\boldsymbol{x}}_m$, $m \in \{1, \ldots, M\}$, can be large. To circumvent this issue, we apply the modulo operation $f_\tau(.)$ to transfer each large magnitude $\tilde{\boldsymbol{x}}_n$ to the original constellation boundary. The function $f_\tau(.)$ is defined as

$$f_\tau(a) \triangleq a - \left\lfloor \frac{a + \tau/2}{\tau} \right\rfloor \tau \tag{2.37}$$

where $\lfloor . \rfloor$ denotes the greatest integer less than or equal to its input. We note here that $f_\tau(.)$ is performed separately on both real and imaginary components. The parameter τ is chosen to provide symmetric decoding regions around constellation points and therefore, it depends on the employed modulation scheme

$$\tau = 2(|c|_{\max} + \Delta/2) \tag{2.38}$$

where $|c|_{\max}$ is the absolute value of real or imaginary elements of the constellation symbols with the largest magnitude, and Δ is the spacing between the constellation points. Now, in the presence of $f_\tau(.)$, the symbol vector at the transmitter side can be expressed as

$$\tilde{\boldsymbol{x}} = \boldsymbol{s} + \tau\boldsymbol{v} - (\boldsymbol{B} - \boldsymbol{I})\tilde{\boldsymbol{x}} \tag{2.39}$$

where v is an integer vector obtained as a result of the modulo function. From Figure 2.9 which represents the system model of THP, we have the following relation

$$\boldsymbol{r} = \boldsymbol{\Gamma}\boldsymbol{H}\boldsymbol{G}\tilde{\boldsymbol{x}} + \boldsymbol{\Gamma}\boldsymbol{w} \tag{2.40}$$

where $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_K)$, represents the scaling factors. We consider the "LQ" decomposition of the channel matrix $\boldsymbol{H} = \boldsymbol{S}\boldsymbol{F}^H$, where $\boldsymbol{F}$ is an unitary, and $\boldsymbol{S}$ is a lower triangular matrix. Now we have the equivalent equation

$$\boldsymbol{r} = \boldsymbol{\Gamma}\boldsymbol{S}\boldsymbol{F}^H\boldsymbol{G}\boldsymbol{B}^{-1}\boldsymbol{s} + \boldsymbol{\Gamma}\boldsymbol{w} \tag{2.41}$$

THP is implemented by choosing

$$\begin{aligned} \boldsymbol{\Gamma} &= [\text{diag}(\boldsymbol{S})]^{-1} \\ \boldsymbol{B} &= \boldsymbol{\Gamma}\boldsymbol{S} \\ \boldsymbol{G} &= \boldsymbol{F} \end{aligned} \tag{2.42}$$

We can see that this choice allows us to eliminate the intra-cell interference to form user-specific spatial channels. After the transmission, each kth user's receiver is able to apply $f_\tau(.)$ independently on $\boldsymbol{r}_k$ to remove the component $\tau\boldsymbol{v}_k$ from the received signal.

Although THP design reduces the required transmit power compared to linear precoding techniques, better performance can be obtained by optimally perturbing the data symbol vector $\boldsymbol{s}$, so that further reduction in the transmit power is obtained. Figure 2.10 shows the BER performances of the different linear precoding techniques compared to the non-linear schemes. We consider a BS with $M = 8$ antennas serving $K = 8$ single-antenna users. The MIMO channel is Rayleigh fading with 16-QAM input alphabets. We plot the BER performance versus the SNR, which is expressed as

$$\text{SNR} = 10\log_{10}\left(\frac{P}{\sigma_w^2}\right) \tag{2.43}$$

where P is the transmit power at the BS. We can see from Figure 2.10 that the VP precoding has the best performance in terms of BER compared to the other precoding techniques, which perform far from optimal when the number of served users approaches that of transmit antennas.

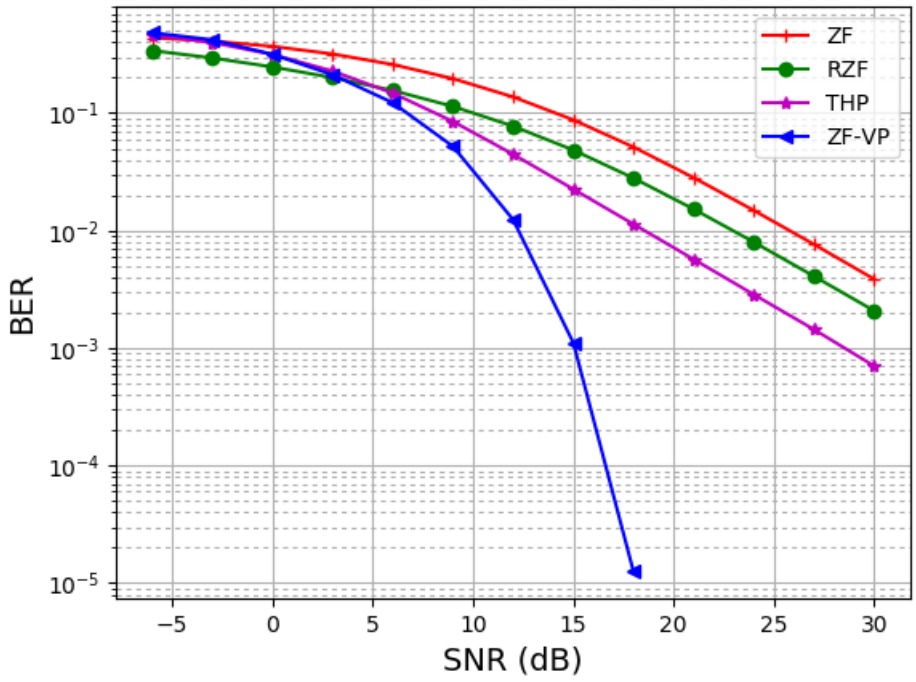


FIGURE 2.10: BER performance of MIMO precoding techniques for an 8×8 MIMO system with 16-QAM constellation.

Vector perturbation precoding

The main purpose of VP precoding is to reduce the transmit power while perturbing the data symbol vector $\boldsymbol{s}$ by an additional signal processing $\boldsymbol{v}$ referred as perturbation vector. More precisely, using $\boldsymbol{s}$ and $\boldsymbol{v}$, the symbol vector $\tilde{\boldsymbol{s}}$ is built as

$$\tilde{\boldsymbol{s}} = \boldsymbol{s} + \tau \boldsymbol{v} \tag{2.44}$$

Then $\tilde{\boldsymbol{s}}$ is precoded using linear precoding such as ZF or RZF before transmission. In other words, the transmit vector $\boldsymbol{x}$ is formed by multiplying $\tilde{\boldsymbol{s}}$ with the precoding matrix $\boldsymbol{F}$

$$\boldsymbol{x} = \sqrt{\frac{P}{\gamma}} \boldsymbol{F} \tilde{\boldsymbol{s}} = \sqrt{\frac{P}{\gamma}} \boldsymbol{F}(\boldsymbol{s} + \tau \boldsymbol{v}) \tag{2.45}$$

Similar to THP, τ in (2.39) is also used by VP which can be represented as an integer-lattice search. At the transmitter side, $\boldsymbol{v}$ is chosen such that γ is minimized. Thus the optimal perturbation vector $\boldsymbol{v}^*$ is given by

$$\boldsymbol{v}^* = \underset{\boldsymbol{v} \in \mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}} \|\boldsymbol{F}(\boldsymbol{s} + \tau \boldsymbol{v})\|^2 \tag{2.46}$$

The arising issue is an integer least-squares problem, and the SE algorithm can solve

this at the transmitter side. The SE is the similar version of the SD algorithm implemented at the receiver side for detection.

The received signal of all users can be expressed as

$$\boldsymbol{y} = \sqrt{\frac{P}{\gamma}}\boldsymbol{H}\boldsymbol{F}\left(\boldsymbol{s} + \tau\boldsymbol{v}^*\right) \tag{2.47}$$

If we consider the VP precoding with the ZF precoder $\boldsymbol{F} = \boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H)^{-1}$, we have

$$\boldsymbol{y} = \sqrt{\frac{P}{\gamma}}\left(\boldsymbol{s} + \tau\boldsymbol{v}^*\right) \tag{2.48}$$

With the knowledge of P and γ, each kth user scales its received signal with $\sqrt{\gamma/P}$ and applies the modulo function $f_\tau(.)$ to remove $\tau\boldsymbol{v}_k$ without knowing its value. The estimated symbol vector is then detected as the constellation point with the nearest Euclidean metric.

The idea of VP precoding is developed using the MSE minimization to propose the MMSE-VP [35,36]. In this approach, the precoding matrix and the optimal perturbation vector are found jointly by minimizing the end-to-end MSE. Thus the MMSE-VP achieves better performances than the VP with ZF or RZF precoder in the entire SNR region.

2.6 Summary

In this chapter, the background material related to MIMO communication systems is presented. At first, we present the different MU-MIMO configurations in both the uplink and downlink schemes. Secondly, we describe the duality between the uplink reception and the downlink transmission. Finally, the main MIMO decoding techniques are described, and the primary precoding approaches are equivalent. Simulation results are shown to evaluate the system performance and enable comparison between the different techniques.

Chapter 3

MIMO Decoding in the Uplink Reception

3.1 Introduction

This chapter focuses on MIMO decoding in the uplink reception. Multiple antennas allow the BS to receive data from multiple users simultaneously. Several MIMO transceiver architectures are covered in the literature. Some of these are linear receivers with or without successive cancellation, and the complexity is mainly at the receiver. This chapter is interested in optimal decodings, such as the well-known SD algorithm, which requires high complexity. Our purpose is to propose new decoding algorithms based on the SD to offer a reduced complexity.

The first part of the chapter presents a DL model for regression to predict the number of lattice points in the sphere, which depends primarily on the sphere centre, the lattice generator matrix and the sphere radius. Some essential definitions associated with lattices are presented in this section to describe the predictive learning approach afterwards. Based on the NN model, we introduce a systematic approach of sphere radius design and control to improve the decoding sphere's initial radius. The learning approach can reduce the number of visited points during the search phase, and thus, the processing time for decoding is decreased.

The second part of the chapter introduces the block recursive MIMO decoding, which divides the whole MIMO system into small sub-systems. The challenge of this technique is to decrease the complexity while preserving near-ML performance essentially. The idea consists of splitting the received symbol vector into smaller vectors. Accordingly, the channel matrix is split into blocks of sizes equal to that of small vectors. This scheme is feasible for any number of blocks with variable sizes. We obtain a significant complexity reduction coupled with an achieved diversity order.

3.2 Counting Lattice Points in the Sphere using NN

Among the many problems in connection with lattices, many remain open. In this work, we focus on the problem of counting lattice points in the sphere.

3.2.1 Definitions and properties of lattices

In this section, we are going to review some essential concepts about lattices and explain some of its main useful parameters in the following. A lattice is a discrete (additive) subgroup of $\mathbb{R}^n$. An equivalent definition is that a lattice $\mathbf{\Lambda}$ of rank p in $\mathbb{R}^n$ $(p \leq n)$ consists of all integral linear combinations of linearly independent vectors $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)$ in $\mathbb{R}^n$, i.e.,

$$\mathbf{\Lambda} = \{z_1\boldsymbol{b}_1 + \ldots + z_p\boldsymbol{b}_p \mid z_1 \ldots z_p \in \mathbb{Z}\} \tag{3.1}$$

Such a set of vectors $\boldsymbol{b}_i$'s is called a lattice basis $\boldsymbol{B}$, i.e.,

$$\boldsymbol{B} = (\boldsymbol{b}_1 \ldots \boldsymbol{b}_p) \tag{3.2}$$

and so $\boldsymbol{B}$ can be defined as the generator matrix of the lattice $\mathbf{\Lambda}$. Now we can define the Gram matrix of $\mathbf{\Lambda}$ as

$$\boldsymbol{G} = \boldsymbol{B}^H\boldsymbol{B} \tag{3.3}$$

and the determinant of $\mathbf{\Lambda}$ as

$$\det(\mathbf{\Lambda}) = \sqrt{\det(\boldsymbol{G})} \tag{3.4}$$

This determinant corresponds to the n-dimensional volume of the parallelotope spanned by the $\boldsymbol{b}_i$'s and defined by

$$V = \left\{ \sum_{j=1}^{n} c_j\boldsymbol{b}_j \mid c_j \in \mathbb{R},\ 0 \leq c_j < 1\ ;\ 1 \leq j \leq n \right\} \tag{3.5}$$

V is the fundamental domain of $\mathbf{\Lambda}$ in the sense that each $\boldsymbol{x} \in \mathbb{R}^n$ has a unique representation $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ with $\boldsymbol{y} \in \mathbf{\Lambda}$ and $\boldsymbol{z} \in V$. Figure 3.1 shows a two-dimensional lattice and the fundamental parallelotope determined by the basis $(\boldsymbol{b}_1, \boldsymbol{b}_2)$.

3.2.2 Learning approach

Let us denote with $\mathcal{B}_r$ the n-dimensional Euclidean ball of radius r centred at the origin, i.e.,

$$\mathcal{B}_r \triangleq \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq r\} \tag{3.6}$$

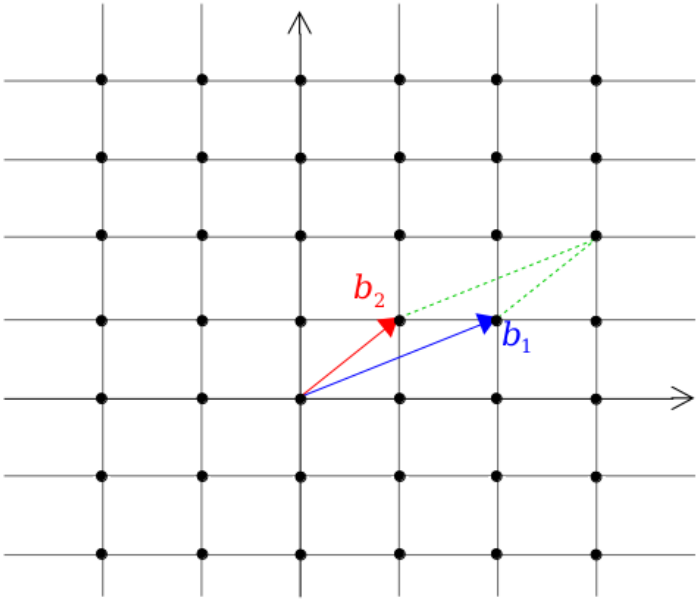


FIGURE 3.1: Fundamental parallelotope of a 2 dimensional lattice.

According to [37], we can make the volume approximation of a sphere containing N_p lattice points which is equal to the volume of N_p fundamental parallelotopes. Hence the number of lattice points lying within a sphere of radius r may be approximated as

$$N_p = \#\{\boldsymbol{x} \in \boldsymbol{\Lambda} : \|\boldsymbol{x}\|_2 \leq r\} \approx \frac{\text{vol}\,(\mathcal{B}_r)}{\det\,(\boldsymbol{\Lambda})} \tag{3.7}$$

where $\text{vol}\,(\mathcal{B}_r) = r^n \pi^{n/2}/\Gamma(n/2+1)$. As we can see, the number of lattice points inside the n-dimensional sphere is proportional to its volume. However, one does not know the density of lattice points inside a given sphere. The approximation is not tight, and the error term is not negligible [38, 39]. This is an interesting problem in pure geometry, and it has considerable practical importance, specifically when the dimension n is larger than three (see [40]). The problem is also deeply connected to complexity theory, particularly to the closest vector problem (CVP), and the Fincke and Pohst variant used for MIMO detection [26]. In fact, decoding a received signal $\boldsymbol{y}$ in $\mathbb{R}^n$ means finding the lattice point closest to $\boldsymbol{y}$, i.e., finding $\tilde{\boldsymbol{v}}$ of some lattice $\boldsymbol{\Lambda}$, such that $|\boldsymbol{y} - \tilde{\boldsymbol{v}}| \leq |\boldsymbol{y} - \boldsymbol{v}|$ for all $\boldsymbol{v} \in \boldsymbol{\Lambda}$. The SD algorithm is well-known for doing this detection [24]. This kind of algorithms requires high complexity scaling exponentially in the lattice dimension. Our work addresses this issue by leveraging recent advances in DNNs to reduce the computational complexity of several lattice problems.

To that end, we train a feed-forward fully-connected DNN to predict the number of lattice points falling inside a given sphere. The chief advantage of such an approach

is that it drastically reduces the exponential complexity of CVP problems. Thus we make use of DNNs to learn and predict the inherent characteristics of lattices. We highlight related work in the literature that describes a DNN based SD [41]. However, this approach learns only the minimum radius and not the number of lattice points falling inside the sphere, which is the quantity of interest. Our work is fundamentally different since we describe a mechanism for learning and predicting the number of lattice points in the n-dimensional sphere with some radius centred at the origin. We formulate the problem as a DNN regression task, which we use to predict the number of lattice points. The training data is obtained via list SD implementations that give the correct number of points falling inside the sphere. This is, in turn, obtained from the radius r and the upper triangular matrix $\boldsymbol{R}$, which is derived from the "QR" decomposition of the lattice generator matrix.

Considering these aspects, the DNN is trained using a set of input-output vector pairs $(\boldsymbol{x}, N_p)$ where $\boldsymbol{x}$ is the input vector, and N_p is the actual number of lattice points inside the sphere. We set $\boldsymbol{x}$ in the form of

$$\boldsymbol{x} = \frac{1}{r}\Big[\boldsymbol{R}_{11}, \ldots, \boldsymbol{R}_{nn}\Big]^T \tag{3.8}$$

where $\boldsymbol{R}_{ij}$ are the coefficients of $\boldsymbol{R}$ matrix $(1 \leq i \leq j \leq n)$. The DNN predicts the number of lattice points at its output layer as

$$\hat{N}_p = f(\boldsymbol{x}; \boldsymbol{\theta}) \tag{3.9}$$

where $\boldsymbol{\theta}$ is the vector of DNN parameters. We use the well known rectified linear unit (ReLU) as an activation function for each layer in the NN

$$\sigma(u) = u^+ = \max(0, u) \tag{3.10}$$

To optimize the parameter vector $\boldsymbol{\theta}$ of the NN, which consists of the weights and biases between input and output layers, we use the mean absolute percentage error (MAPE) as a loss function which results in the following formula

$$\tilde{L}(\theta) = \frac{1}{\#\{S_t\}} \sum_{i \in S_t} \left| \frac{N_p^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})}{N_p^{(i)}} \right| \tag{3.11}$$

where $N_p^{(i)}$ is the desired output when $\boldsymbol{x}^{(i)}$ is used as an input and $\tilde{L}(\boldsymbol{\theta})$ is the loss function. Moreover, S_t denotes the training mini-batch. By choosing a small mini-batch of samples instead of the whole training set, the gradient complexity is significantly reduced. However, the variance of the gradient estimate is inevitably increased. As an optimization method to adjust parameters, we use Adadelta [42], which adapts the learning rate to the parameters. We opted not to use methods such

TABLE 3.1: Structure for the NN.

DNN parameters	Value
Dimension of input variables	$\frac{n(n+1)}{2}$
Dimension of output variables	1
Number of hidden layers	1 to 2
Number of hidden neurons	128

as Adam due to the exponential complexity resulting from hyper-parameter tuning.

3.2.3 Simulation results

In this section, we evaluate the performance of the proposed NN model through several simulation experiments. We consider different sizes of systems, i.e., lattices of dimensions n, where n is varied from 5 to 10. Elements of the generator matrix are modelled as i.i.d zero-mean Gaussian random variables with unit variance. The upper triangular matrix $\boldsymbol{R}$ is obtained using the "QR" factorization of the generator matrix. Moreover, we choose the sphere radius at random and fix it during the simulation results. The number of training samples is equal to 50000, and the mini-batch size of S_t used for stochastic gradient descent (SGD) is 10. The test set consists of $T = 10000$ samples generated independently of the training set. The detailed parametrization of the DNN is shown in Table 3.1.

To evaluate our model, we check the accuracy metric to help us determine whether the estimations generated by our model are close to the correct values. The accuracy metric that we use for the prediction error is the MAPE that usually expresses accuracy as a percentage, and is defined over T samples as

$$MAPE = \frac{100\%}{T} \sum_{t=1}^{T} \left| \frac{N_p^{(t)} - \hat{N}_p^{(t)}}{N_p^{(t)}} \right| \tag{3.12}$$

where $N_p^{(t)}$ and $\hat{N}_p^{(t)}$ are the actual and predicted values respectively, for a sample t of the test set. The MAPE is commonly used as a loss function for regression problems and model evaluation thanks to its intuitive interpretation in relative error. Dividing by the actual value $N_p^{(t)}$ instead of the predicted value $\hat{N}_p^{(t)}$ leads to a different result. Thus the MAPE is not symmetric in the sense that interchanging $N_p^{(t)}$ and $\hat{N}_p^{(t)}$ does not lead to the same answer. This issue has been raised in [43] and [44]. The symmetric MAPE (SMAPE) has been proposed to provide symmetry and robustness against outliers by dividing the absolute loss by the arithmetic mean of the actual $N_p^{(t)}$ and

TABLE 3.2: Accuracy experiment for arbitrary lattices in $\mathbb{R}^n$.

Dimension n	MAPE %	SMAPE %
6	14.575 (14.700)	7.055 (7.052)
7	15.958 (16.290)	8.700 (8.842)
8	17.021 (17.147)	9.028 (9.036)
9	16.377 (16.724)	8.463 (8.701)
10	17.078 (17.078)	9.247 (9.809)

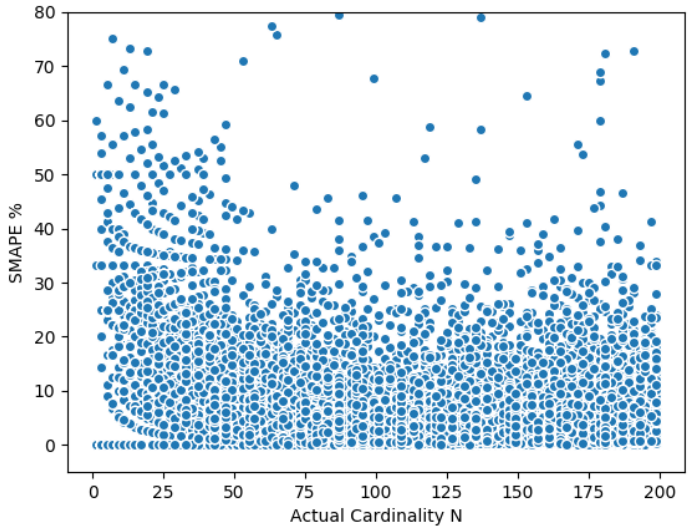
the predicted $\hat{N}_p^{(t)}$

$$SMAPE = \frac{100\%}{T} \sum_{t=1}^{T} \frac{|N_p^{(t)} - \hat{N}_p^{(t)}|}{|N_p^{(t)} + \hat{N}_p^{(t)}|} \tag{3.13}$$

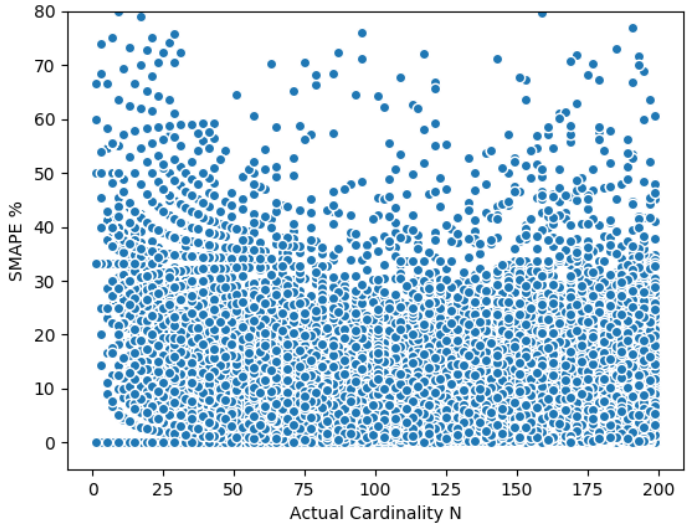
We plot the accuracy of our predictions for dimension $n = 10$ as shown in Figure 3.2, which visualizes the SMAPE metric versus the actual number of points. We observe that the percentage error is concentrated below 20%. This indicates that our model fits the number of lattice points significantly well and can proceed with an accurate prediction. Some accuracies of high errors happen for some unusual characteristics of generator matrices.

Figure 3.3 shows the error histogram of the DNN. We plot the left figure for the training set error and the right one for the test set error. We present results for lattices of dimension 10, but we generally observe similar results for all dimensions between 5 and 10. We can see a high percentage of points whose SMAPE is below 10%, indicating that our model fits the number of points significantly well. In Table 3.2, we present the MAPE and the SMAPE for each lattice dimension on the training set and on the test set whose data occurs within parentheses. We observe the similarity between the training error and the test error, which indicates that our model avoids both under-fitting and over-fitting thanks to the use of ℓ_1 and ℓ_2 regularization techniques.

To validate our DNN model, we use some known lattices. We used benchmarks known as upper bounds on the number of lattice points in a small sphere of radius equal to the "covering radius" for lattices A_n, D_n and E_n (see [40] for a definition of these lattices). In [45], the number of lattice points contained in a small sphere, centred anywhere in $\mathbb{R}^n$, is upper bounded using two methods, via spherical codes and Gaussian measures. The first method resembles the one used by Conway and Sloane in [40] to upper bound the kissing number of a lattice, i.e., the number of its shortest nonzero vectors. It is shown that lattice points in $\mathcal{B}_r$ can be rearranged



(A) Training set.



(B) Test set.

FIGURE 3.2: The SMAPE versus the actual number of points for the dimension $n = 10$.

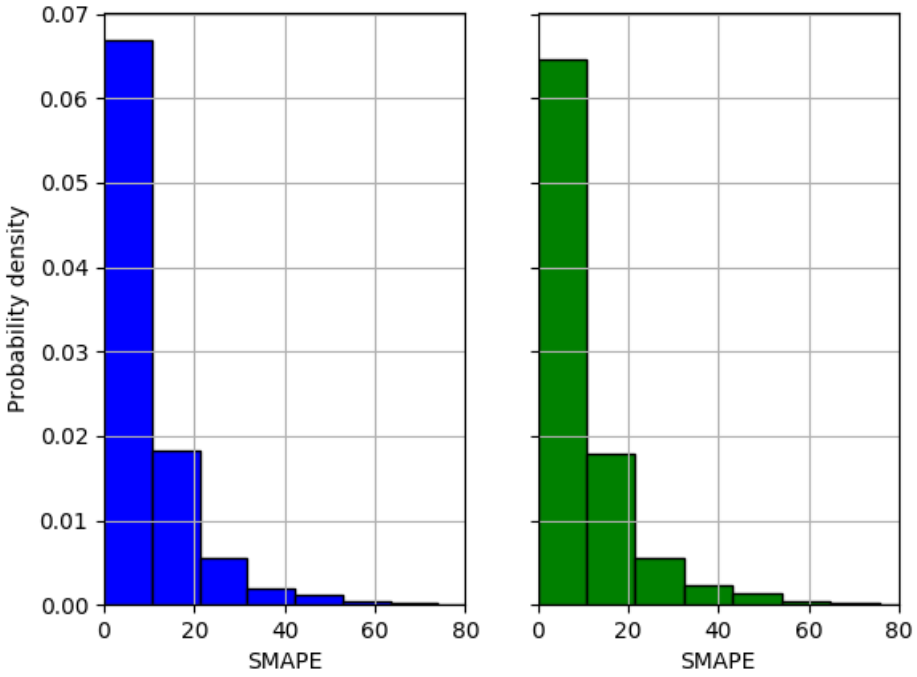


FIGURE 3.3: Plots of the SMAPE histogram of the DNN model over the dimension $n = 10$, the left figure corresponds to the training error and the right one to the test error.

inside $\mathcal{B}_r$ as a spherical code with an absolute minimum angle. This gives rise to upper bounds on N_p since various methods to upper bound the cardinality of a spherical code with a given minimum angle.

To also obtain upper bounds on $|\mathcal{B}_r(z) \cap \mathbf{\Lambda}|$ where

$$\begin{aligned} &\mathcal{B}_r(\boldsymbol{z}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{z}\|_2 \leq r\} \\ &\min_{\boldsymbol{a} \neq 0 \in \Lambda} \|\boldsymbol{a}\|_2 < r \end{aligned} \tag{3.14}$$

authors in [45] present a different approach based on Gaussian-like measures on $\mathbf{\Lambda}$. For each $\boldsymbol{z} \in \mathbb{R}^n$, they give a positive real number $\gamma_{r,\boldsymbol{\lambda},\boldsymbol{z}}$ such that $|\mathcal{B}_r(z) \cap \mathbf{\Lambda}| \leq \gamma_{r,\boldsymbol{\lambda},\boldsymbol{z}}$. Based on worst-case assumptions on $\boldsymbol{z}$, they also obtain a universal upper bound, i.e., a positive real $\gamma_{r,\boldsymbol{\lambda}}$ such that

$$\sup_{\boldsymbol{z} \in \mathbb{R}^n} |\mathcal{B}_r(z) \cap \mathbf{\Lambda}| \leq \gamma_{r,\boldsymbol{\lambda}} \tag{3.15}$$

We adopt these bounds here to enable direct comparisons. The comparison is shown in Table 3.3. We can see through the results that the prediction is accurate as it is almost the same as the actual number of points. The analytical upper bound is no tighter as lattice dimensions increase. The last column of the table shows the SMAPE

TABLE 3.3: Results for some known lattices.

Type	n	Spherical bound	Gaussian bound	Actual number	Predicted number	SMAPE %
A	5	≤ 24	26	7	6	7.69
	6	≤ 54	47	9	10	5.26
	7	≤ 140	99	13	12	4.00
	8	–	188	41	32	12.33
	9	–	391	69	64	3.76
	10	–	758	119	125	2.46
D	5	16	20	7	6	7.69
	6	≤ 37	42	9	10	5.26
	7	≤ 88	88	11	11	0.00
	8	240	183	77	62	10.79
	9	–	595	103	88	7.85
	10	–	1211	133	130	1.14
E	8	16	77	17	12	17.24

value indicating the normalized symmetric absolute deviation from the actual number of points. We notice that in some dimensions, our predicted values are produced with some errors.

3.3 Learning assisted SD

We cannot deny it has always been a trade-off between efficiency and complexity; both are proportional. This section is interested in the SD algorithm [24] thanks to its good performance. However, its major problem is still highlighted. It requires a considerable amount of computations compared to other decoding techniques. It can be shown that, both from a worst-case and from an average point of view, the SD requires an exponential complexity [46]. Several algorithm modifications have been

explored in the past to reduce its complexity further. Techniques such as re-ordering the computation sequence, performing a fixed number of operations, modified norm definition, and early termination of search have been proposed [47–56]. Some suggested modifications solve the ML decoding problem exactly, and others sacrifice some performance to reduce complexity.

Some other cited works focused mainly on the change of sphere radius, which is crucial and significantly affects the computational complexity [57]. On the one hand, if the radius is too small, there might not be any lattice point inside the spherical region. On the other hand, a vast radius may result in too many lattice points, increasing the decoding complexity. A simple method (SDIRS) to increase the radius search was proposed [24]. The SD algorithm starts with an initial radius of r_1 based on the noise statistics in this approach. When the search fails, we increase the radius to r_2 $(> r_1)$ and the same procedure is repeated until we find the ML solution.

Since the sphere radius directly affects the search range and complexity, it is important to design. Therefore, we profit from the learning approach, which predicts the number of lattice points inside the sphere to reduce the SD algorithm's computational complexity. However, we will consider in this time the received signal point as an origin of the sphere. Hence, known sequences of the received signal, the generator matrix elements, the sphere radius, and the actual number of lattice points create the training data set. The input vector of the NN is now in the form of

$$\boldsymbol{x} = \left[\boldsymbol{y}^T,\ \boldsymbol{R}_{11}, \ldots, \boldsymbol{R}_{nn},\ r\right]^T \tag{3.16}$$

We should mention here that the radius is arbitrary fixed for each generator matrix. The noise variance is randomly generated in such a way that the SNR is uniformly distributed on $[\mathrm{SNR}_{\min}, \mathrm{SNR}_{\max}]$ where $\mathrm{SNR}_{\min}$ and $\mathrm{SNR}_{\max}$ are the minimal and maximal SNR values over which we used the network. We execute the same predictive learning for counting the number of lattice points in the sphere centred at the received point. This procedure is effected offline, and then we use the NN's updated parameters for the entire communication phase.

3.3.1 NN assisted SD with a dichotomic search of radius

Our principal purpose is to implement the SD algorithm using an enhanced initial radius, leading to a small number of lattice points inside the sphere. In this case, the NN model's number of lattice points is predicted as a function of the received signal point, the generator matrix, and the sphere radius. We start first by predicting the number of lattice points falling inside the sphere with an initial radius equal to that proposed in [24]. Then, if this expected number is large, we update the radius

using a dichotomic search as proposed in [58]. Indeed, we divide the square radius by two, and we predict the number of lattice points with the new radius again. We repeat the same procedure until we attain a predicted number less or equal to a given threshold. Finally, we start the search phase of the SD algorithm with the suited radius. Figure 3.4 shows a flowchart of the proposed NN-SD algorithm.

3.3.2 Smart SD with improved radius

Every time we update the radius, NN computations are necessary to predict the number of lattice points and check if it is still large or not. This leads to an additional average complexity related to NN computations. Therefore, we seek to evaluate the average number of radius updates as a function of the SNR. In other words, we want to evaluate the average number of NN calculations before starting the SD algorithm. With an initial radius $r_0^2 = 2n\sigma_w^2$, we analyze theoretically the number of radius updates using the equation in (3.7) which approximates the number of lattice points N_p. Let the SNR be defined as $\rho = P/\sigma_w^2$. By successively dividing the square radius L times till reaching a small expected number of lattice points, we start the SD algorithm with the radius $r_L^2 = r_0^2/2^L$. Now, if we use the approximate function in (3.7), we determine the number of iterations L as a function of the SNR (ρ expressed in dB)

$$L = a\rho + b \tag{3.17}$$

where

$$\begin{aligned} a &= -\frac{1}{10\log_{10} 2} \\ b &= \frac{2}{n\log_{10} 2}\Big(\log_{10} V_n - \mathbb{E}[\log_{10}\det(\mathbf{\Lambda})] - \mathbb{E}[\log_{10} N_p]\Big) \\ &\quad + \frac{1}{\log_{10} 2}\log_{10} 2nP \end{aligned} \tag{3.18}$$

Proof: see Appendix A.

As we deal with uncorrelated channels, we exploit some properties of Wishart distribution to determine the statistical expectation of log-determinant. The Wishart distribution is a family of distributions for symmetric positive definite matrices. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independently drawn from a p-variate normal distribution $N_p(0, \mathbf{\Sigma})$, and form a $p \times n$ data matrix $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]$. The distribution of a $p \times p$ random matrix $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^T$ is said to have the Wishart distribution with n degrees of freedom and covariance matrix $\mathbf{\Sigma}$.

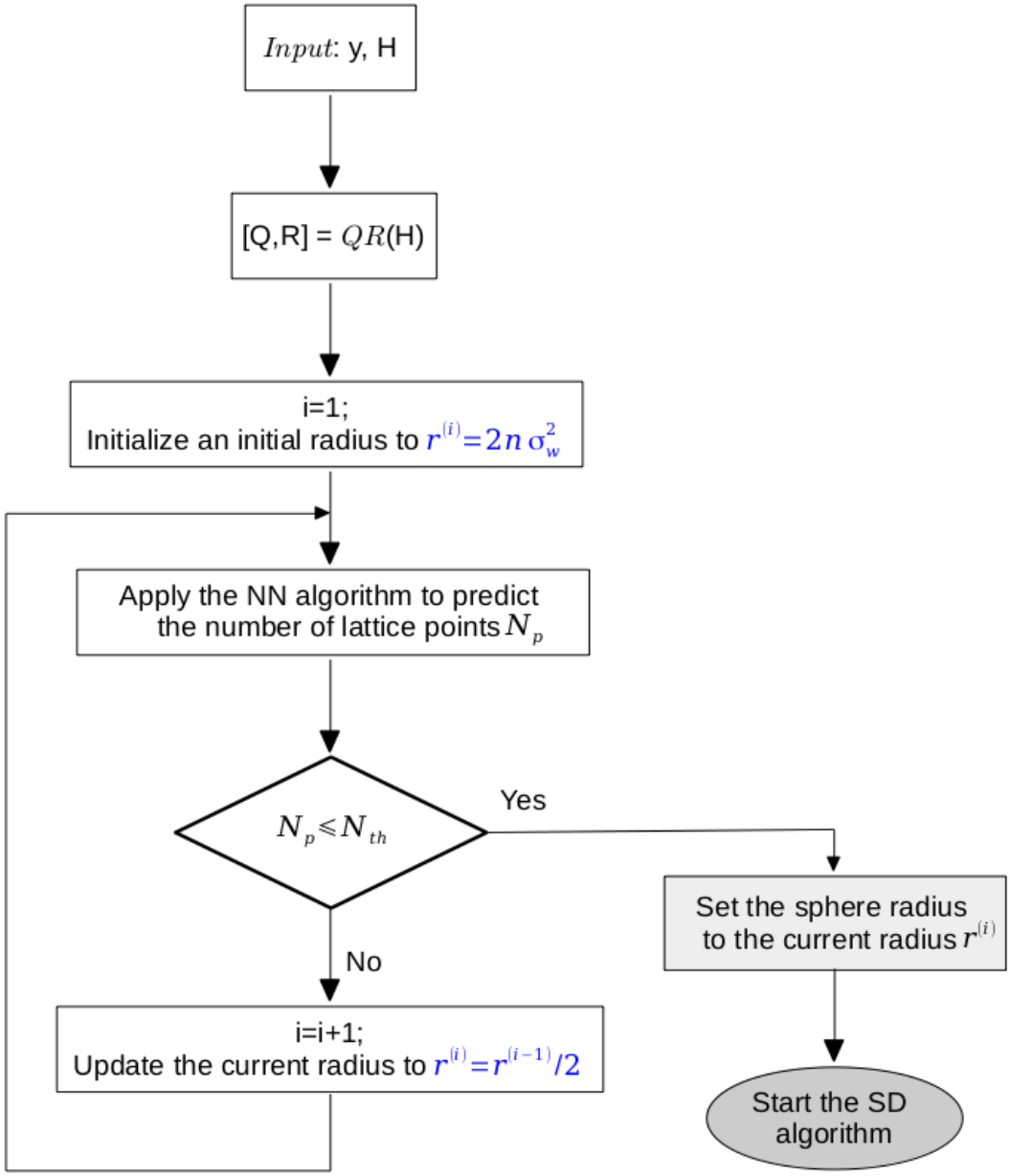


FIGURE 3.4: Flowchart of the NN-SD algorithm.

So let $\mathbf{\Lambda}$ be distributed according to a $p \times p$ dimensional Wishart distribution with n degrees of freedom and covariance matrix $\mathbf{\Sigma}$. The expected value of the log-determinant has the following formula [59]

$$\mathbb{E}[\ln \det(\mathbf{\Lambda})] = \psi_p(\frac{n}{2}) + p\ln(2) + \ln\det(\mathbf{\Sigma}) \tag{3.19}$$

where ψ_p is the multivariate digamma function which represents the derivative of the log of the multivariate gamma function. The latter is a generalization of the gamma function.

Based on equation (3.17), we come up with a new algorithm, smart SD (SSD), without NN computations. We start the search phase of the SD with an improved radius which is equal to $r_0^2/2^L$ with $r_0^2 = 2n\sigma_w^2$, and L is calculated as expressed in (3.17) and (3.18).

3.3.3 NN-SD vs. SSD

In this part, we want to make a fair comparison between the NN-SD and SSD. First off, we should mention that both algorithms find an enhanced initial sphere radius before starting the SD search phase. In the beginning, we proposed the NN-SD algorithm as described in 3.3.1. The main steps are presented in Figure 3.4. The sphere radius is used as an element of the NN-SD input to predict the number of lattice points. Every time the predicted number is high, we reduce the sphere radius, and we predict again. A fixed number of floating-point operations (FLOPS) is required with every prediction due to NN computations (weight matrix multiplications). In this context, we find from simulations that the number of radius updates L in the NN-SD decreases linearly as a function of SNR. Thus, we proposed the SSD algorithm to evaluate L theoretically. The comparison results are similar to that of NN-SD except for the processing complexity, which is reduced. Indeed, SSD has the advantage to avoid NN computations as the enhanced initial sphere radius is determined theoretically. However, we can not deny that if we had not proposed the NN-SD, the idea of the SSD algorithm would not have reached.

3.3.4 Simulation results

In this section, we present computer simulations of the NN-SD compared to the SDIRS algorithm. We evaluate the performance as well as the complexity advantage yielded by the proposed scheme. We consider 8×8 MIMO channel with 16-QAM input alphabet. The performance is evaluated in terms of BER for low-to-moderate SNRs. The computational complexity is measured by counting the number of multiplications or measuring the processing time for decoding with the same computer

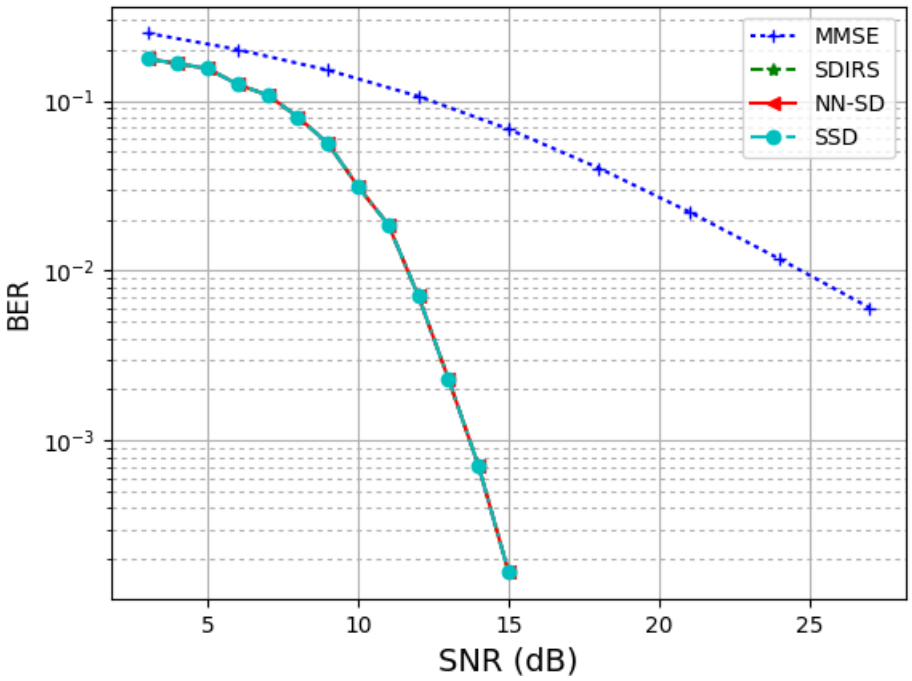


FIGURE 3.5: BER performance of the NN-SD for the 8×8 MIMO system with 16-QAM constellation.

processor. To compare with the SDIRS algorithm, we exploit the same sets of transmitting signals, fading channels, and noises.

In Figure 3.5, we plot the system performance as a function of the SNR. We can see that we have the best BER performance, and we assert perfectly that our proposed NN-SD algorithm is an ML detector contrary to the DL based SD in [41], which loses ML performance at low-to-moderate SNRs. As demonstrated earlier, we also propose the SSD algorithm for decoding with an improved sphere radius. We plot in Figure 3.6 the average number of updates on the radius as a function of the SNR. Correspondingly, we use this result to calculate the initial sphere radius immediately without NN computations. We can see that this number decreases linearly as a function of the SNR (expressed in dB) as proved in 3.3.2 where we analyze this behaviour theoretically. The line slope is independent of the system size, and it is the same as calculated theoretically. However, the intercept depends on the size and the statistical expectation of the log number of lattice points. The latter is not in an explicit closed form. Therefore, it makes it empirically obtained from computer simulations using NNs.

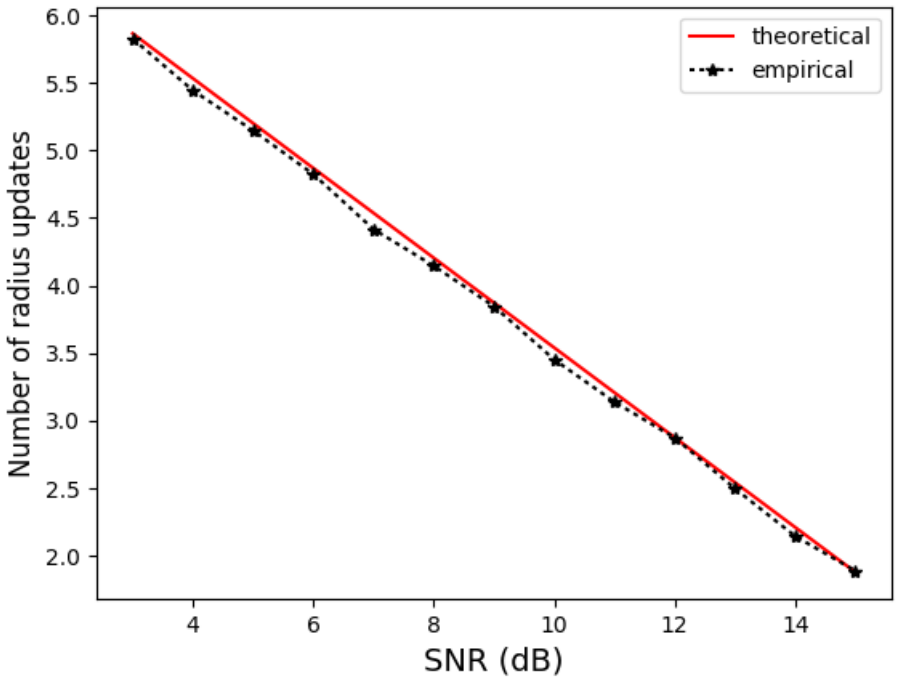


FIGURE 3.6: Average number of radius updates. The solid lines are obtained theoretically, and the dashed lines are obtained empirically from computer simulations.

In Figure 3.7, we plot the average number of multiplication to measure the computational complexity as a function of SNR. It is well observed that the NN-SD significantly reduces the number of operations compared to the SDIRS algorithm. This complexity reduction is explained by choice of an initial sphere radius that allows a small number of lattice points to fall inside the sphere, and thus the search tree size decreases in the average sense. Besides, we plot in Figure 3.8 the average processing time as a function of SNR. The comparison results are similar to that in Figure 3.7. We compare the average decoding time of the SSD to that of the MMSE receiver and the NN-SD. As seen, the decoding time is lower, and the reason for this complexity reduction is that we do not need any more NN computations to update the radius. For example, the average processing time for the SSD at 12 dB SNR is almost 10 times the MMSE decoding time, while it is almost 233 times for the SDIRS algorithm compared to the MMSE receiver.

Figure 3.9 displays the average number of lattice points falling inside the decoding sphere. We can see that this average in the NN-SD is almost constant as a function of SNR, while it is higher in the SDIRS algorithm for low-to-moderate SNRs.

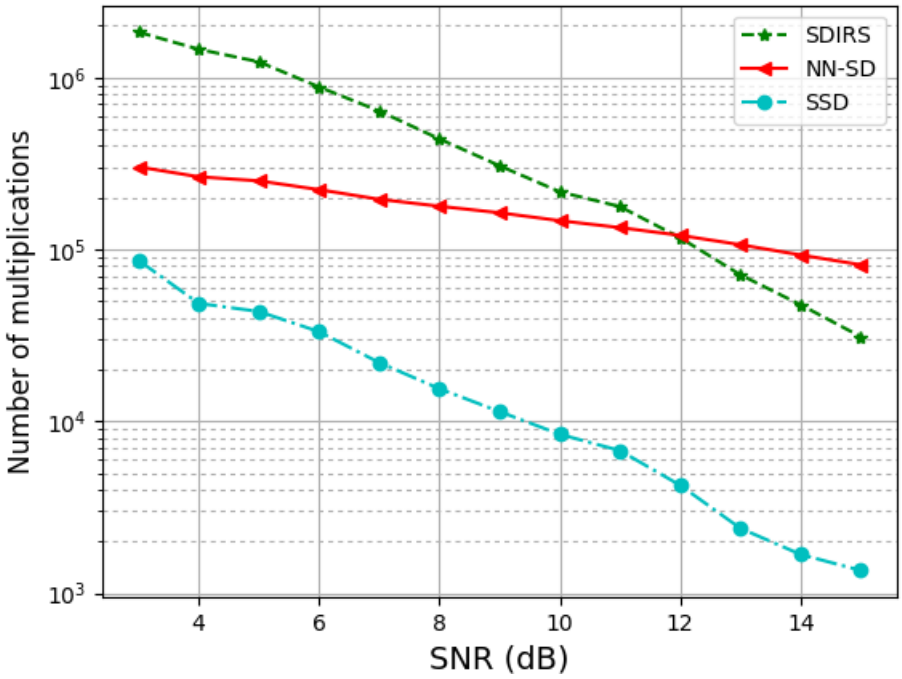


FIGURE 3.7: Average number of multiplications in the decoding process.

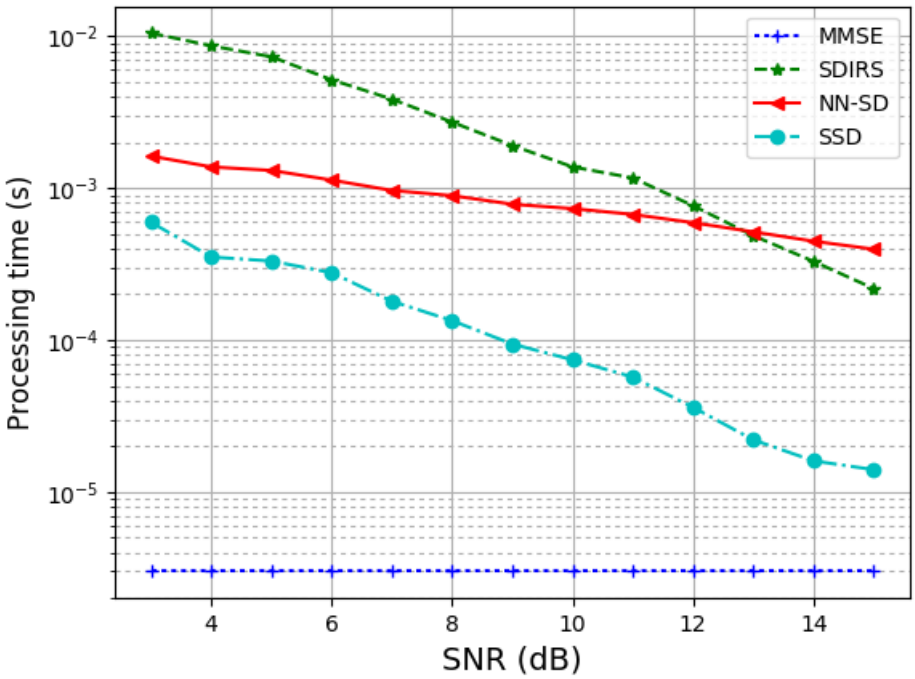


FIGURE 3.8: Average processing time in the decoding process.

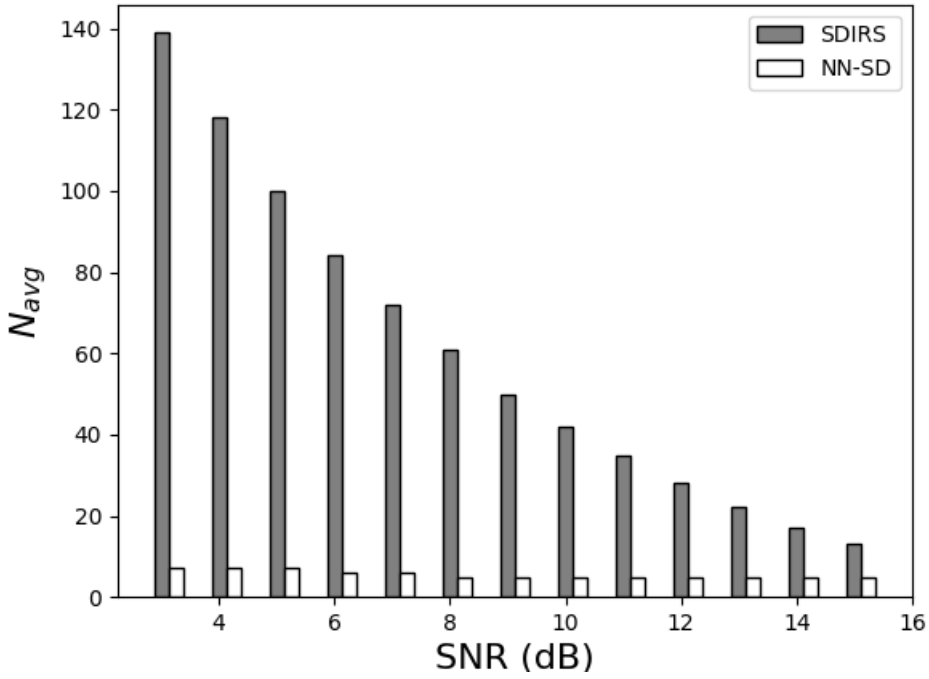


FIGURE 3.9: Average number of lattice points (N_{avg}) falling inside the search sphere.

3.4 Block Recursive MIMO Decoding

Block recursive decoding for space-time coded systems has been recovered in academic literature [60–65], and is shown to slightly reduce the complexity of ML detection. Partitioned signal sets are decoded recursively, taking advantage of the equivalent channel matrix form induced by the code structure. The main issue of block decoding is using, at one step, an exhaustive list over one block, which increases the overall complexity. We can find in the literature two approaches for MIMO block decoding. The first one is based on the division of the channel matrix into two blocks. In [60], an ML decoding scheme is performed on the first block of size p_1; then, a DFE equalizer is applied for the remaining system after subtracting the first ML output from the received signal. It was shown that this scheme could increase the diversity order for the worst channel from 1 to p_1. The second approach, which is the partial interference cancellation (PIC) group decoding [66], consists in splitting the received signal into $S \geq 2$ subsets. A selection of one set likelihood function is performed, informally

1. Perform an exhaustive list of solutions for the selected subset;
2. Cancel the decoded part from the received signal for each candidate;
3. Decode the remaining $S - 1$ subsets using the ZF equalizer;

4. Select the optimal solution from all complete solutions.

The choice of the subset to decode first is crucial since it affects the system performance. Thus empirical and analytical set selection criteria on the equivalent channel matrix are developed. In [61–63], the main set selection criterion is considered a determinant of the channel covariance matrix. This quantity measures the instantaneous SNR of the corresponding linear system. Another criterion is based on minimizing the condition number of the covariance matrix to measure the ZF accuracy. Then the ratio of these quantities should be maximized. Inspired by the works mentioned above, authors in [64] introduce two new low-complexity decoders, namely adaptive conditional ZF (ACZF) and ACZF with successive interference cancellation (SIC). They give two identical sufficient conditions based on STBC characteristics to get full diversity. One sufficient condition is the total rank of at least one of the S sub-matrices.

In our work, we propose a new block decoding strategy, which is a generalization method of the work described in [67], where the MIMO system is only divided into two blocks.

3.4.1 Block division

The idea is to resolve the sub-systems coming from any division into more than two blocks. Let us consider the upper triangular matrix which is divided into k blocks as depicted in Figure 3.10. Let $(p_1, \ldots, p_k)$ be the block sizes satisfying $\sum_{j=1}^{k} p_j = 2n$. $\boldsymbol{R}_i \in \mathbb{R}^{p_i \times p_i}$ and $\boldsymbol{B}_i \in \mathbb{R}^{p_{i+1} \times (\sum_{j=1}^{i} p_j)}$ are the upper triangular and feedback matrices, respectively, where $i \in \{1, \ldots, k\}$. Accordingly, the transmitted and the received signal vectors are split into $(\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(k)})$ and $(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(k)})$, respectively. Our proposed recursive decoding method is implemented with two major steps. Firstly we estimate the information symbols in the first (k-1) blocks using block decoding to obtain a possible trial of the incomplete information symbols. Secondly, we search for the remaining data which minimizes the overall ML metric. To summarize, the decoding process consists of the following steps:

1. Choose the number of blocks k and their corresponding sizes $(p_1, \ldots, p_k)$;
2. Create a list of solutions for the first block using a sequential decoding. This list is composed of the ML solution which minimizes the Euclidean distance $\|\boldsymbol{y}^{(1)} - \boldsymbol{R}_1 \boldsymbol{s}^{(1)}\|^2$, and its neighbors;
3. Create a new list of solutions for each candidate in the previous list to minimize the Euclidean metric $\|\boldsymbol{y}^{(2)} - \boldsymbol{B}_1 s^{(1)} - \boldsymbol{R}_2 \boldsymbol{s}^{(2)}\|^2$;
4. **Repeat the last procedure until achieving the (k-1)th block**;
5. Sort the set of candidates in increasing order of their weights;

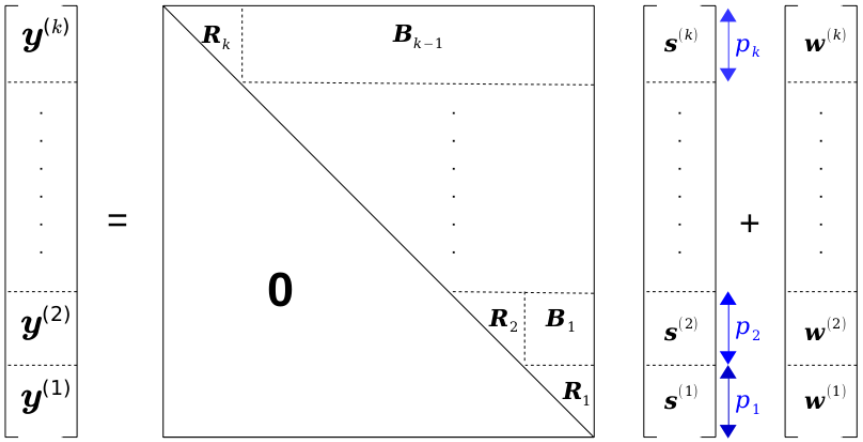


FIGURE 3.10: Block division of the decoding system.

6. Search for the remaining data symbol vector in the kth block by starting with the top of ordered candidates.

The last step ends when an examined full candidate's smallest weight is less than the next one's partial weight to examine. The second, third, and fourth steps need a set of radiuses $(r_1, \ldots, r_{k-1})$ representing the thresholds on weights to create the lists. For that, these radiuses are computed based on the derivation of an upper bound of the frame error rate P_{ef}.

3.4.2 Diversity order analysis

At high SNR level, the frame error is caused by an error on one symbol with high probability. Thus the frame error rate can be approximated by $P_{\text{ef}} = 2nP_{\text{es}}$, where P_{es} is the symbol error rate. In [68], we have derived in detail an upper bound of P_{ef} given by

$$P_{\text{ef}} \leq c\rho^{-n} + \sum_{i=1}^{k-1} \frac{\Gamma_u(\frac{p_i}{2}, \frac{r_i^2}{2\sigma_w^2})}{\Gamma(\frac{p_i}{2})} \tag{3.20}$$

where $c > 0$ is some constant, ρ denotes the SNR expressed in dB, and $\Gamma_u(a, x)$ is the upper Gamma function.

Proof:

Let us denote with

- $\boldsymbol{s} = \left(\boldsymbol{s}^{(k)^T}, \ldots, \boldsymbol{s}^{(1)^T}\right)^T$ the transmitted vector split into k blocks;
- $\hat{\boldsymbol{s}}$ the estimated information symbols;

- $\tilde{\boldsymbol{s}}$ the event that $\boldsymbol{s}$ was visited during the search;
- $\tilde{\boldsymbol{s}}^{(i)}$ the event that $\boldsymbol{s}^{(i)}$ was visited during the search;

By using the conditional probability rule, the frame error rate $P_{ef} = \Pr(\boldsymbol{s} \neq \hat{\boldsymbol{s}})$ could be written as

$$\begin{aligned} P_{ef} &= \Pr(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \cap \tilde{\boldsymbol{s}}) + \Pr(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \cap \tilde{\boldsymbol{s}}^c) \\ &= \Pr(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \tilde{\boldsymbol{s}}) \Pr(\tilde{\boldsymbol{s}}) + \underbrace{\Pr(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \tilde{s}^c)}_{=1} \Pr(\tilde{\boldsymbol{s}}^c) \\ &= \Pr\left(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \tilde{\boldsymbol{s}}\right) \Pr(\tilde{\boldsymbol{s}}) + \Pr(\tilde{\boldsymbol{s}}^c) \\ &\leq \Pr\left(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \tilde{\boldsymbol{s}}\right) + \Pr(\tilde{\boldsymbol{s}}^c) \end{aligned} \tag{3.21}$$

We start by deriving $\Pr\left(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \tilde{\boldsymbol{s}}\right)$

$$\Pr\left(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \boldsymbol{s}\right) = \mathbb{E}_{\boldsymbol{R}} \mathbb{E}_{\boldsymbol{s}} \sum_{\substack{\boldsymbol{s}' \in \mathcal{S} \\ \boldsymbol{s}' \neq \boldsymbol{s}}} \Pr\left(\| \boldsymbol{y} - \boldsymbol{R}\boldsymbol{s}' \|^2 \leq \| \boldsymbol{y} - \boldsymbol{R}\boldsymbol{s} \|^2\right) \tag{3.22}$$

where $\mathcal{S}$ is the set of visited candidates during the search, and we have

$$\Pr\left(\| \tilde{\boldsymbol{y}} - \boldsymbol{R}\boldsymbol{s}' \|^2 < \| \tilde{\boldsymbol{y}} - \boldsymbol{R}\boldsymbol{s} \|^2\right) \leq \mathcal{Q}\left(\sqrt{\frac{\| \boldsymbol{R}\left(\boldsymbol{s}' - \boldsymbol{s}\right) \|^2}{\sigma_w^2}}\right) \tag{3.23}$$

where $\mathcal{Q}(.)$ is the Gaussian Q-function. After performing the QR decomposition of $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{R}$, we derive the distributions of $\boldsymbol{R}_{ii}^2$ for $i \in \{1, \ldots, 2n\}$. The diagonal entries $\boldsymbol{R}_{ii}^2$ are obtained from the Bartlett decomposition [69] of the following random matrix

$$\begin{bmatrix} \Re(\bar{\boldsymbol{H}}) \\ \Im(\bar{\boldsymbol{H}}) \end{bmatrix}^T \begin{bmatrix} \Re(\bar{\boldsymbol{H}}) \\ \Im(\bar{\boldsymbol{H}}) \end{bmatrix} \tag{3.24}$$

where $\begin{bmatrix} \Re(\bar{\boldsymbol{H}}) \\ \Im(\bar{\boldsymbol{H}}) \end{bmatrix}$ has i.i.d. entries according to $\mathcal{N}(0, 1/2)$; thus $\boldsymbol{R}_{ii}^2 \sim \chi^2(2n - i + 1, 1/2)$ for $i \in \{1, \ldots, n\}$. Additionally, $\boldsymbol{R}_{2n2n}^2 \sim \chi^2(2, 1/2)$ and the approximate PDF of $\boldsymbol{R}_{ii}^2$ for $i \in \{n+1, 2n-1\}$, is derived in [70].

Hence, from the Chernoff bound, we obtain

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{R}}\left(\Pr\left(\| \boldsymbol{y} - \boldsymbol{R}\boldsymbol{s}' \|^2 \leq \| \boldsymbol{y} - \boldsymbol{R}\boldsymbol{s} \|^2\right)\right) \\ &\leq \frac{2}{\left(1 + \frac{\|\boldsymbol{s}' - \boldsymbol{s}\|^2}{2\sigma_w^2}\right)^n} \leq \frac{2}{\left(1 + \frac{d_{\min}^2}{2\sigma_w^2}\right)^n} \end{aligned} \tag{3.25}$$

where $d_{\min}$ is the distance between the nearest neighbors. Accordingly,

$$\begin{aligned}\Pr\left(\hat{\boldsymbol{s}} \neq \boldsymbol{s} \mid \tilde{\boldsymbol{s}}\right) &\leq \mathbb{E}_{\boldsymbol{s}} \sum_{\substack{\boldsymbol{s}' \in \mathcal{S} \\ \boldsymbol{s}' \neq \boldsymbol{s}}} \frac{2}{\left(1+\frac{d_{\min}^2}{2\sigma_w^2}\right)^n} \\ &\leq \beta_1 \rho^{-n} \end{aligned} \tag{3.26}$$

where β_1 is some positive constant.

Now we derive $\Pr(\tilde{\boldsymbol{s}}^c)$

$$\Pr(\tilde{\boldsymbol{s}}^c) = \sum_{i=1}^{k} \Pr\left(\tilde{\boldsymbol{s}}^{(i)^c} \mid \left(\tilde{\boldsymbol{s}}^{(i-1)} \cap \ldots \cap \tilde{\boldsymbol{s}}^{(1)}\right)\right) \times \Pr\left(\tilde{\boldsymbol{s}}^{(i-1)} \cap \ldots \cap \tilde{\boldsymbol{s}}^{(1)}\right) \tag{3.27}$$

Denoting by $\mathcal{E}_i = \Pr\left(\tilde{\boldsymbol{s}}^{(i)^c} \mid (\tilde{\boldsymbol{s}}^{(i-1)} \cap \ldots \cap \tilde{\boldsymbol{s}}^{(1)})\right)$ for $i \in \{1, \ldots, k\}$, it follows that

$$\Pr(\tilde{\boldsymbol{s}}^c) \leq \sum_{i=1}^{k} \mathcal{E}_i \tag{3.28}$$

We derive $\mathcal{E}_i$ for $i \in \{1, \ldots, k-1\}$ and we let the derivation when $i = k$ later. Recall that $\mathcal{E}_i$ implies that the weight of $\boldsymbol{s}^{(i)}$ given the right partial transmitted message $\left(\boldsymbol{s}^{(i-1)^T}, \ldots, \boldsymbol{s}^{(1)^T}\right)^T$, falls over a certain fixed threshold r_i

$$\begin{aligned}\mathcal{E}_i &= \mathbb{E}_{\boldsymbol{R}_i} \sum_{\boldsymbol{s}^{(i)} \in \mathcal{A}^{p_i}} \Pr\left(\boldsymbol{s}^{(i)}\right) \times \Pr\left(\parallel \tilde{\boldsymbol{y}}^{(i)} - \boldsymbol{R}_i \boldsymbol{s}^{(i)} \parallel^2 > {r_i}^2\right) \\ &= \Pr\left(\parallel \boldsymbol{z}^{(i)} \parallel^2 > r_i^2\right) \end{aligned} \tag{3.29}$$

where $\tilde{\boldsymbol{y}}^{(i)} = \boldsymbol{y}^{(i)} - \boldsymbol{B}_{i-1}\left(\boldsymbol{s}^{(i-1)^T}, \ldots, \boldsymbol{s}^{(1)^T}\right)^T$. Since $\boldsymbol{z}^{(i)}$ is a vector of p_i Gaussian entries characterised by $\mathcal{N}(o, \sigma_w^2/2)$, then $\frac{\|\boldsymbol{z}^{(i)}\|^2}{\sigma_w^2/2}$ is distributed according to the $\chi^2\left(p_i\right)$

$$\begin{aligned}\mathcal{E}_i &= \Pr\left(\frac{\parallel \boldsymbol{z}^{(i)} \parallel^2}{\sigma_w^2/2} > \frac{r_i^2}{\sigma_w^2/2}\right) \\ &= \int_{\frac{r_i^2}{\sigma_w^2/2}}^{\infty} f_i(x)dx = \frac{\Gamma_u(\frac{p_i}{2}, \frac{r_i^2}{\sigma_w^2})}{\Gamma(\frac{p_i}{2})} \end{aligned} \tag{3.30}$$

where f_i is the PDF of χ^2 with p_i degrees of freedom, and $\frac{\Gamma_u(a,x)}{\Gamma(a)}$ is the regularized upper Gamma function.

The last term $\mathcal{E}_k$ in (3.27) depends on the decoding scheme that we apply for the

last block. We propose to perform an ML decoder such that the SD algorithm after eliminating the interference caused by the previously detected symbols. We have

$$\begin{aligned}\mathcal{E}_k &\leq \mathbb{E}_{\boldsymbol{R}_k}\mathbb{E}_{\boldsymbol{s}^{(k)}} \sum_{\boldsymbol{s}'^{(k)}\neq\boldsymbol{s}^{(k)}} \Pr(\| \tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{R}_k\boldsymbol{s}'^{(k)} \|^2 \leq \| \tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{R}_k\boldsymbol{s}^{(k)} \|^2) \\ &\leq \mathbb{E}_{\boldsymbol{s}^{(k)}} \sum_{\boldsymbol{s}'^{(k)}\neq\boldsymbol{s}^{(k)}} \frac{2}{\left(1+\frac{\|\boldsymbol{s}'^{(k)}-\boldsymbol{s}^{(k)}\|^2}{2\sigma_w^2}\right)^n} \\ &\leq \mathbb{E}_{\boldsymbol{s}^{(k)}} \sum_{\boldsymbol{s}'^{(k)}\neq\boldsymbol{s}^{(k)}} \frac{2}{\left(1+\frac{d_{\min}^2}{2\sigma_w^2}\right)^n} \end{aligned} \tag{3.31}$$

Similarly to (3.25), the entry of the first column of $\boldsymbol{R}_k$ is distributed according to $\chi^2(2n, 1/2)$ which explains the last two inequalities in (3.31). Hence, we can write

$$\mathcal{E}_k \leq \beta_2 \rho^{-n} \tag{3.32}$$

for some positive constant β_2. Now, from (3.28), we have

$$\Pr(\widetilde{\boldsymbol{s}}^c) \leq \beta_2\rho^{-n} + \sum_{i=1}^{k-1} \frac{\Gamma_u(\frac{p_i}{2}, \frac{r_i^2}{\sigma_w^2})}{\Gamma(\frac{p_i}{2})} \tag{3.33}$$

Finally, by combining (3.26) and (3.33), an upper bound of (3.21) is

$$P_{ef} \leq (\beta_1+\beta_2)\rho^{-n} + \sum_{i=1}^{k-1} \frac{\Gamma_u(\frac{p_i}{2}, \frac{r_i^2}{\sigma_w^2})}{\Gamma(\frac{p_i}{2})} \tag{3.34}$$

The diversity order that could be achieved by this decoding scheme is controlled by the second term given that the first one achieves full diversity. To guarantee an overall diversity order of at least $d \in \{1,\ldots,n\}$, each term of the sum should decrease at the order of ρ^{-d}. This goes back to find for each ith block the minimum threshold r_i such that

$$\frac{\Gamma_u(\frac{p_i}{2}, \frac{r_i^2}{2\sigma_w{}^2})}{\Gamma(\frac{p_i}{2})} \leq \delta\rho^{-d},\ i \in \{1,\ldots,k-1\} \tag{3.35}$$

for some positive constant δ that controls the SNR gain. In [67], r_i's are calculated numerically for two blocks. Now, in this work as described in [68], we give the analytical calculus of r_i. Indeed, the inequality on r_i is solved based on the asymptotic inversion of incomplete Gamma functions [71]. We are interested in the x-value that solves the following equation at each ith level

$$\mathcal{Q}(a,x) = \frac{\Gamma_u(a,x)}{\Gamma(a)} = q \tag{3.36}$$

where $a = p_i/2$, $x = r_i^2/2\sigma_w^2$ and $q = \delta\rho^{-d} \in [0, 1]$.

The approximations are obtained by using uniform asymptotic expansions of the incomplete Gamma functions in which an error function is a dominant term

$$\mathcal{Q}(a, x) = \frac{1}{2}\text{erfc}(\eta\sqrt{a/2}) + R_a(\eta) \tag{3.37}$$

The real parameter η is defined by

$$\frac{1}{2}\eta^2 = \lambda - 1 - \ln\lambda,\ \lambda = x/a,\ \text{sign}(\eta) = \text{sign}(\lambda - 1) \tag{3.38}$$

We denote the solution of the above equation by $\eta(q, a)$. The inversion problem starts by inverting the error function considering $R_a(\eta)$ in (3.37) as a perturbation. Thus, we define the number $\eta_0 = \eta_0(q, a)$ as the real number that satisfies the equation

$$\frac{1}{2}\text{erfc}(\eta_0\sqrt{a/2}) = q \tag{3.39}$$

We write

$$\eta(q, a) = \eta_0(q, a) + \epsilon(q, a) \tag{3.40}$$

and we determine the function ϵ that appears in the form

$$\epsilon(q, a) \sim \frac{\epsilon_1}{a} + \frac{\epsilon_2}{a^2} + \frac{\epsilon_3}{a^3} + \ldots, \tag{3.41}$$

The coefficients ϵ_i can be found in [71] as functions of η_0 using the Taylor expansions. In our work, only the first and second-order terms, i.e., ϵ_1 and ϵ_2, are considered.

3.4.3 Simulation results

This section presents numerical results of the proposed recursive decoder compared to the original SD algorithm. We consider a 10×10 MIMO channel with a 16-QAM input alphabet and a block division on 2 and 3 blocks. The figure legend indicates the block sizes $(p_1, \ldots, p_k)$.

In Figure 3.11, we plot the decoder performance as a function of SNR. We can see that, by fixing the target diversity to 10, the block decoder achieves almost ML performance for all considered divisions. In Figure 3.12, we plot the average processing time for decoding as a function of SNR. We use the same computer processor and the same programming language C to measure the processing time. It is well observed

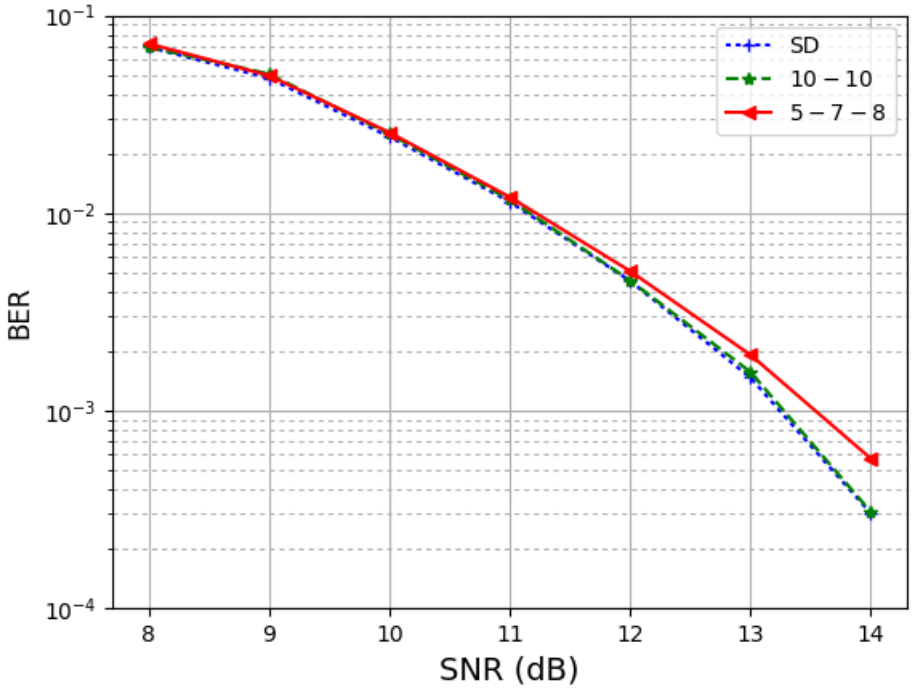


FIGURE 3.11: BER performance of the block decoder for the 10×10 MIMO system with 16-QAM constellation.

that the SD algorithm has a more considerable time of processing than the recursive block decoder.

The computer simulations are obtained with different values of δ, which refers to the factor gain appearing in the determination of r_i's. Since it is not in an explicit closed-form, this factor requires a numerical optimization obtained from simulations. We should mention its crucial effect on the trade-off between the proposed block recursive decoder's performance and complexity.

3.5 Summary

To reduce the complexity of the well-known SD. At first, we train a NN model to predict the number of lattice points falling inside the n-dimensional sphere with some radius centred at the origin. We argued that the proposed model is more reliable than the analytical upper bounds covered in the literature for some known lattices. For the general case, our model can proceed with an accurate approximation. Secondly, we use the learning approach to propose the NN assisted SD, which gives the ML performance significantly lower complexity. Besides, we propose another decoding algorithm without NN computations, and we obtain more significant complexity reduction. Finally, we propose the block recursive MIMO decoder. The latter achieves

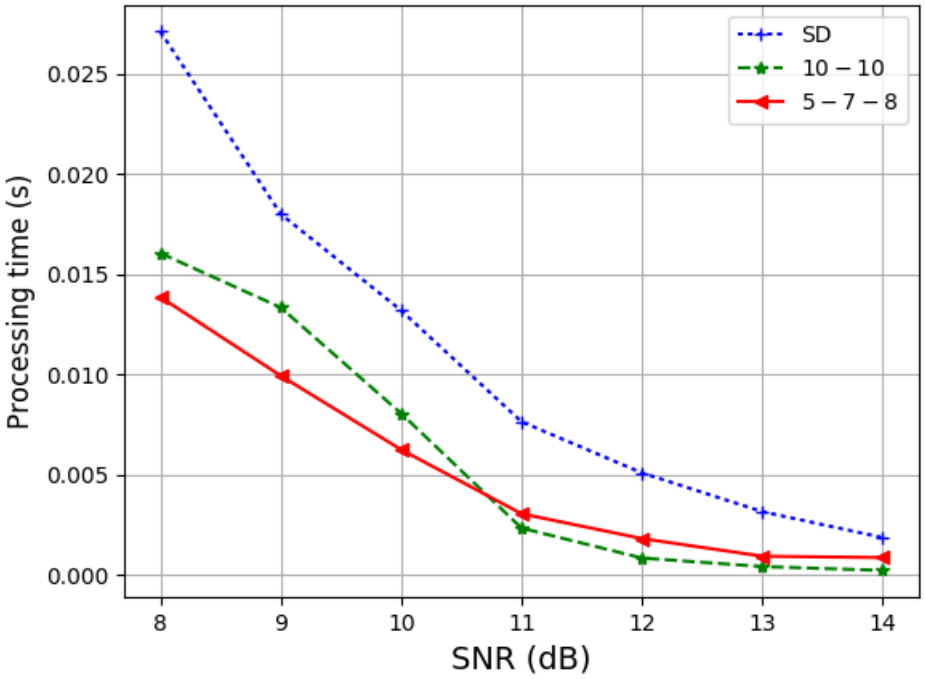


FIGURE 3.12: Average processing time in the block decoder compared to the SD algorithm.

a complexity reduction coupled with the desired diversity order. The general case of multiple blocks with different sizes is studied. Some structures of block sizes reduce the complexity with a guarantee of almost ML performance. As the goal is to lower the complexity, the main issue of this approach is the unknown size of blocks to divide the MIMO system and obtain the minimum complexity. We focus on the problem of dividing the MIMO system in such a way that the gain in complexity reduction achieved by the block decoding is maximized. One does not know the best blocks' sizes that provide the minimum complexity. This is an exciting problem, and it has considerable importance since it affects computational complexity. A novel strategy can be presented to predict the optimal blocks' sizes to resolve the mentioned problem. To that end, we can train, for example, a feed-forward NN to predict blocks' sizes allowing us to achieve the minimum complexity compared to all other divisions.

Chapter 4

Learning assisted Fronthaul Compression for Uplink C-RAN

4.1 Introduction

In this chapter, the MIMO decoding in the uplink reception remains to be investigated. However, this time, it will be reviewed in the cloud-RAN, sometimes referred to as centralized-RAN. C-RAN is an architecture for cellular networks consisting of many radio remote units (RRUs) or BSs connected to a central processor (CP). Due to the prohibitive complexity of computations, the most efficient uplink C-RAN schemes are challenging to be implemented in practical systems. Using DNNs, we propose a new and low complex method for uplink C-RAN subject to some quantization rules. This is the first work that uses DNNs to mimic the C-RAN system to the best of our knowledge. Our architecture's objective, called QDNet, is to jointly optimize the processing done at the BSs and the processing done at the CP side. Our goal is not to solve signal detection in multi-antenna systems. Instead, the goal is to mimic the whole transmission in uplink C-RAN, which considers the quantization constraints at the BSs and the corrupted observations at the CP. Inspired by the projected gradient descent algorithm, QDNet is designed as a distributed DNN with sparse connections. Experiment results are provided and show that our scheme outperforms linear receivers such as the ZF) equalizer and achieves near-optimal performance compared to the SD algorithm.

4.2 Background

With the extensive use of advanced wireless devices and video streaming, smaller-sized cellular networks emerge to meet the increasing data rates demand. Consequently, the distance between BSs becomes small, inducing higher inter-cell interference (ICI). C-RAN has been seen as a valuable model to cope with the dominant ICI by enabling the joint precoding in the downlink transmission and the joint decoding in the uplink reception at the CP side [72, 73]. The compress-and-forward (CF) is

one of the most studied techniques in the uplink C-RAN. Each BS first transforms, quantizes, and then transmits the quantized signal to the CP via the finite capacity fronthaul. However, it is challenging to find the optimal compression scheme that minimizes the distortion error within capacity constraints. This can be more problematic if a high quantity of observations is available. Under certain assumptions such as Gaussian signals or channels, this problem has been formulated in [74, 75] with a single BS and Wyner-Ziv compression.

Nevertheless, due to the computational complexity and latency induced by an infinite blocklength coding, this compression scheme is troublesome to be implemented in practical communication systems. To avoid the delay caused by long block length coding, we deal with quantization schemes that are exceptional cases of source coding with a fixed block length. Thus, the objective of BSs turns into finding a transformation scheme that can better adapt the given quantization rules. Besides, we aim to find an adequate decoding scheme at the CP side to mitigate the quantization noise impact efficiently.

It is not easy to solve this joint optimization problem with classical mathematical tools as the quantization noise is not simple to characterize. Therefore, we resort to recent advances in DNNs to mimic the whole transmission chain in uplink C-RAN. Several recent works focus on MIMO detection by using DNNs to improve the widely used iterative algorithms. In [76], authors proposed a DNN architecture called DetNet and demonstrated good performance on i.i.d. Gaussian channels. [77] proposed a sparsely connected NN which can reduce the computational complexity while ensuring good performance. Inspired by the orthogonal approximate message passing (OAMP) and iterative soft-thresholding algorithms, OAMPNet [78], and MMNet [79] are developed, respectively. They are shown to be efficient in reconstructing the transmitted messages via i.i.d. Gaussian and 3GPP channels. However, all the algorithms mentioned above are proposed for a point-to-point MIMO configuration. The NN architectures are used to ensure the decoding at the receiver side based on the received signals.

Motivated by DL technologies' performance, we resort to recent advances in DNNs to mimic the whole transmission chain in uplink C-RAN. Due to the finite capacity fronthaul, we aim to find an efficient way to forward the quantized signals at the BSs and find a correspondent decoding scheme at the CP side. This is the first work that uses DNNs to mimic the C-RAN system to the best of our knowledge. So the challenge is to design a distributed NN which is jointly optimized at the BSs and the CP. The aim of our work is not to solve signal detection in MIMO systems. Instead, it aims to mimic the whole transmission in uplink C-RAN where the CP observations are corrupted due to the quantization. Our work is different from previously proposed NN-assisted MIMO receivers in the literature since they are assuming perfect knowledge of observations contrary to our work which deals with corrupted observations.

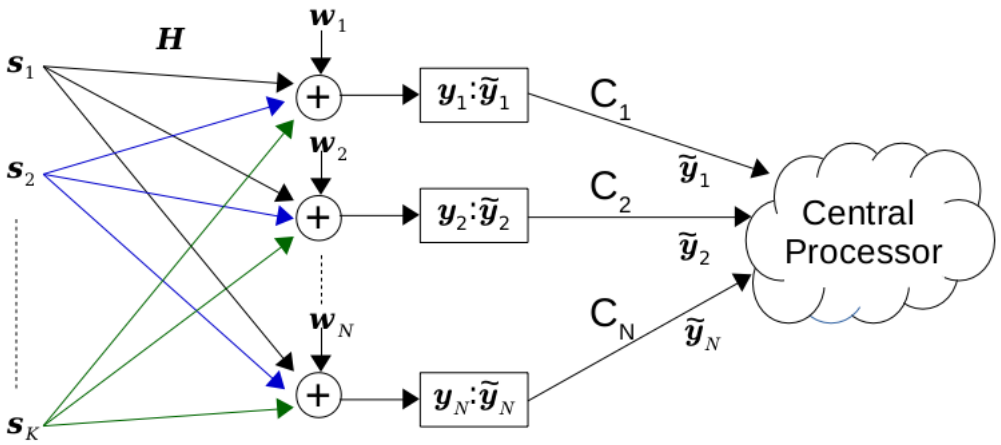


FIGURE 4.1: An uplink C-RAN system with a finite capacity fronthaul.

4.3 System Model and Problem Conception

We consider an uplink C-RAN model as shown in Figure 4.1, where K remote users with N_t transmit antennas emit their messages independently to N remote BSs. Each nth BS is equipped with N_r receiving antennas and connected to the CP via a noiseless fronthaul link with a limited capacity $C_n \; \forall \; n \in \{1, \ldots, N\}$. The transmitted message sent by the ith user is denoted as $\bar{\boldsymbol{s}}_i$ and belongs to a finite constellation $\bar{\mathcal{S}}$. In practice, we assume that the constellation set $\bar{\mathcal{S}}$ is given by a QAM modulation. All constellations are normalized to unit average power (e.g., 4-QAM constellation is represented by $\{\pm\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}\}$). The received signal $\bar{\boldsymbol{y}}_n$ at the nth BS can be expressed as

$$\bar{\boldsymbol{y}}_n = \sum_{i=1}^{K} \bar{\boldsymbol{H}}_{ni} \bar{\boldsymbol{s}}_i + \bar{\boldsymbol{w}}_n \tag{4.1}$$

where $\bar{\boldsymbol{H}}_{in} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix between the ith user and the nth BS, and $\bar{\boldsymbol{w}}_n \sim \mathcal{CN}(0, 2\sigma_w^2 I_{N_r})$ is the AWGN correspondent to the nth BS. The main challenge in MIMO detection is the use of complex-valued signals which are less common in machine learning. Thus, by using the convention in chapter 2, we convert (4.1) to its equivalent real-valued representation

$$\boldsymbol{y}_n = \sum_{i=1}^{K} \boldsymbol{H}_{in} \boldsymbol{s}_i + \boldsymbol{w}_n \tag{4.2}$$

Let us denote with $\boldsymbol{H} = (\boldsymbol{H}_1^T, \ldots, \boldsymbol{H}_N^T)^T$ the global channel in the C-RAN architecture where $\boldsymbol{H}_n \; \forall \; n \in \{1, \ldots, N\}$ is the channel to the nth BS when considering all K users. So the received signal at all BSs can be expressed as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{w} \tag{4.3}$$

where $\boldsymbol{s} = (\boldsymbol{s}_1^T, \ldots, \boldsymbol{s}_K^T)^T$ is the transmitted message sent from all users in the system and $\boldsymbol{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_K^T)^T$ is the AWGN noise.

Since there are no direct links among the BSs, the received signals at the BSs need to be transformed and quantized in a distributed way. The quantization schemes that can be applied to $\boldsymbol{y}_n \; \forall \; n \in \{1, \ldots, N\}$ such as the scalar quantization, the Lloyd-Max quantization [80], or Grassmannian [81], induce severe degradation due to the great size of $\boldsymbol{y}_n$. For instance, if we assume that the scalar quantization is applied and the quantization resources are uniformly allocated, each element of $\boldsymbol{y}_n$ is going to have $C_n/2N_r$ capacity to be exploited. In a massive MIMO system with high N_r, the distortion error induced by the scalar quantization can lead to poor performance at the CP side. Hence, we should use a transformed version $\boldsymbol{r}_n$ of $\boldsymbol{y}_n$ which has a smaller dimension and so can be quantized with less deterioration. It is possible to make the best use of the quantizer to improve the decoding process at the CP side. In this perspective, our work aims to find a useful transformation scheme at the BS side before quantization as long as a correspondent decoding scheme at the CP side which takes into account the considered quantization scheme. Mathematically, if we assume that $\tilde{\boldsymbol{r}}_n$ is the quantized version of $\boldsymbol{r}_n = T_n(\boldsymbol{y}_n; \boldsymbol{H}_n)$, and $\hat{\boldsymbol{s}} = D(\tilde{\boldsymbol{r}}_1; \ldots; \tilde{\boldsymbol{r}}_N; \boldsymbol{H})$ is the estimated message at the CP side, then the optimization problem can be expressed as

$$\begin{aligned} \underset{T_1(.), \ldots, T_N(.), D(.)}{\text{minimize}} \quad & \mathbb{E}_{\boldsymbol{s}} \left[\left\| \boldsymbol{s} - \hat{\boldsymbol{s}} \right\|^2 \right] \\ \text{subject to} \quad & R_q(\tilde{\boldsymbol{r}}_n) \leq C_n \; \forall \; n \in \{1, \ldots, N\} \end{aligned} \tag{4.4}$$

where $T_n(.)$ is the transformation done at the nth BS $\forall \; n \in \{1, \ldots, N\}$, $R_q(.)$ is the required number of bits subject to the quantization rule q, and $D(.)$ is the decoding process done at the CP side.

There are $N + 1$ unknown mappings and functions to be found in the optimization problem (4.4), which seems challenging to obtain with classical approaches. Thus, we choose to use NNs to reformulate our problem (4.4) as a NN regression task to predict the transmitted message $\boldsymbol{s}$.

4.4 QDNet Design

In this section, we present our NN architecture designed for uplink C-RAN consisting of multiple BSs, and called QDNet (Quantization-and-Decoding Network).

The prediction of the transmitted message $\boldsymbol{s}$ of all users is primarily based on the received signal $\boldsymbol{y}_n$ at each nth BS and the global channel $\boldsymbol{H}$. Considering these aspects, the NN is trained using a set of input-output vector pairs $(\boldsymbol{e}, \boldsymbol{s})$ where $\boldsymbol{e} = ((\boldsymbol{y}_1; \boldsymbol{H}_1), \ldots, (\boldsymbol{y}_N; \boldsymbol{H}_N))$ is the input vector. The NN predicts the estimated message $\hat{\boldsymbol{s}}$ at its output layer as

$$\hat{\boldsymbol{s}} = f(\boldsymbol{e}; \boldsymbol{\theta}) \tag{4.5}$$

where $\boldsymbol{\theta}$ is the vector of NN parameters. To optimize $\boldsymbol{\theta}$, we use the mean squared error (MSE) as a loss function which results in the following formula

$$\min_{\boldsymbol{\theta}} \mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{s}^{(m)} - \hat{\boldsymbol{s}}^{(m)}\right\|^2\right) \tag{4.6}$$

where M indicates the number of training examples, $\boldsymbol{s}^{(m)}$ and $\hat{\boldsymbol{s}}^{(m)}$ are the desired target vector and the output vector, respectively, of the mth example.

Our proposed scheme is inspired by the iterative projected gradient descent algorithm. For a given observation $\boldsymbol{y}$, the probability $p(\boldsymbol{y}|\boldsymbol{s})$ can be proved to be inversely proportional to the distance $\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2$. Correspondingly, a projected gradient descent algorithm based on the ML detection can be expressed as

$$\begin{aligned}\hat{\boldsymbol{s}}_{k+1} &= \Pi\left[\hat{\boldsymbol{s}}_k - \frac{\partial\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2}{\partial \boldsymbol{s}}|_{\boldsymbol{s}=\boldsymbol{s}_k}\right] \\ &= \Pi\left[\hat{\boldsymbol{s}}_k + \delta_k\left(\boldsymbol{H}^T\boldsymbol{y} - \boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k\right)\right]\end{aligned} \tag{4.7}$$

where $\boldsymbol{s}_k$ is the estimate of $\boldsymbol{s}$ in the kth iteration, $\Pi[.]$ is a non-linear projection operator, and δ_k is a step size. Intuitively, each iteration is a linear combination of $\hat{\boldsymbol{s}}_k$, $\boldsymbol{H}^T\boldsymbol{y}$, and $\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k$ followed by a non-linear projection. This hints that two main ingredients in the architecture should be $\boldsymbol{H}^T\boldsymbol{y}$ and $\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k$. Our NN construction is based on mimicking that projected gradient descent like a solution for the maximum likelihood optimization. For massive MIMO systems, the matched filter [82] is widely-used linear detector and seems to be a good solution. It is attractive for practical implementations thanks to its low complexity. So it is clear to recognize the signal $\boldsymbol{H}^T\boldsymbol{y}$ to be forwarded to the CP instead of the received signal $\boldsymbol{y}$ of large dimension. Besides, regarding the fronthaul with limited capacity, it is better to work with smaller dimensions to mitigate the degradation induced by the quantization noise.

Let us remark that the terms $\boldsymbol{H}^T\boldsymbol{y}$ and $\boldsymbol{H}^T\boldsymbol{H}$ can be rewritten as

$$\boldsymbol{H}^T\boldsymbol{y} = \sum_{n=1}^{N}\boldsymbol{H}_n^T\boldsymbol{y}_n \tag{4.8}$$

$$\boldsymbol{H}^T\boldsymbol{H} = \sum_{n=1}^{N}\boldsymbol{H}_n^T\boldsymbol{H}_n \tag{4.9}$$

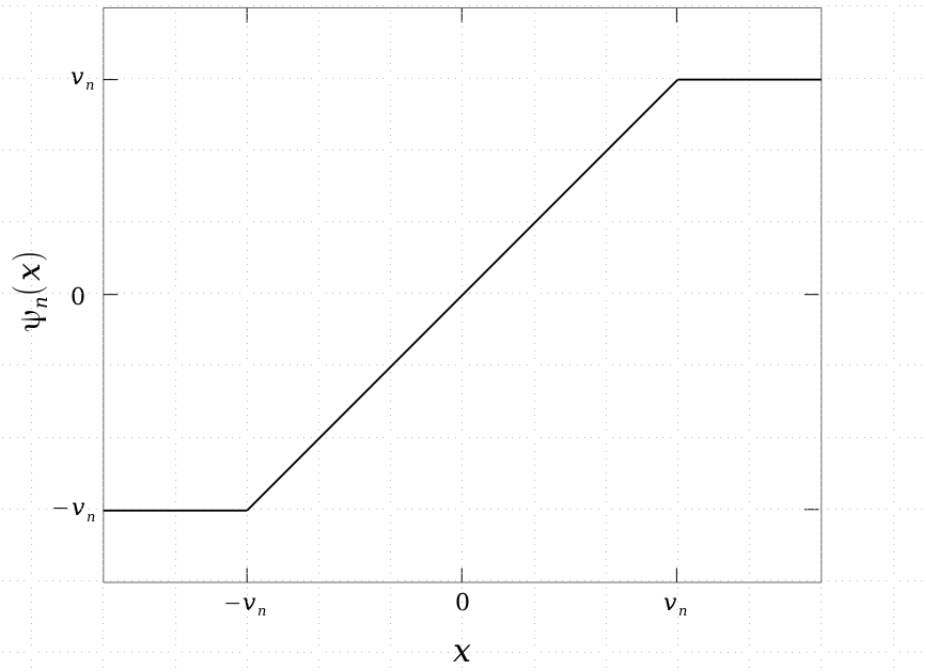


FIGURE 4.2: The clipping function $\psi_n(.)$.

Although each BS transforms and quantizes its observation in a distributed way, the term $\boldsymbol{H}^T\boldsymbol{y}$ can be approximated at the CP side by taking the sum of all the transmitted signals sent by the BSs to the CP. Each BS has its transformation and quantization processes, whereas the decoding process is done at the CP side with a shared network.

4.4.1 Quantization Process at the BS side

Without loss of generality, we present the architecture corresponding to the nth BS employing the scalar quantization. Before being quantized, the signal $\boldsymbol{H}_n^T\boldsymbol{y}_n$ is clipped at the nth BS by involving a piece-wise linear soft sign operator $\psi_n(.)$ defined as

$$\psi_n(x) = -v_n + \rho(x + v_n) - \rho(x - v_n) \tag{4.10}$$

where $\rho(x) = \max\{0, x\}$ and $v_n > 0$ is the clipping threshold parameter to be optimized during the training phase. The operator is plotted in Figure 4.2, and the NN structure is illustrated in Figure 4.3.

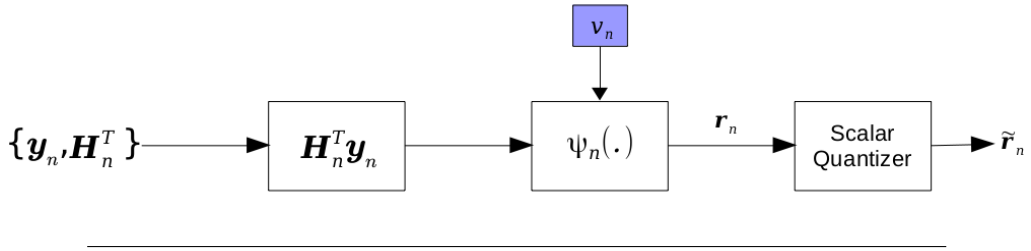


FIGURE 4.3: NN structure at each nth BS.

4.4.2 Decoding Process at the CP side

At this time, $\tilde{r}_n$ is the nth BS's output which will be transmitted to the CP. Once all signals are received from all BSs, the CP adds up these signals to obtain

$$\tilde{r} = \sum_{n=1}^{N} \tilde{r}_n \tag{4.11}$$

$\tilde{r}$ represents a degraded version of $H^T y$ due to the distortion induced by the transformation and quantization processes. To take into account this distortion, we modify the gradient descent algorithm in (4.7) to this form:

$$\hat{s}_{k+1} = \Pi\left[\hat{s}_k + \delta_k\left(\tilde{r} - H^T H \hat{s}_k\right)\right] \tag{4.12}$$

In the first step, we enhance these iterations in (4.12) by changing the step size δ_k for each kth iteration by $\theta_k^{(1)}$ and $\theta_k^{(2)}$ corresponding to $\tilde{r}$ and $H^T H \hat{s}_k$, respectively, that is

$$\hat{s}_{k+1} = \Pi\left[\hat{s}_k + \theta_k^{(1)}\tilde{r} - \theta_k^{(2)} H^T H \hat{s}_k\right] \tag{4.13}$$

In the second step, to mimic the non-linear projection operator $\Pi[.]$, a non-linear denoiser $\zeta_k(.)$ is applied to z_k to produce $\hat{s}_{k+1}$ where $z_k = \hat{s}_k + \theta_k^{(1)}\tilde{r} - \theta_k^{(2)} H^T H \hat{s}_k$. Together, the linear and denoising steps aim to recover an improved estimate $\hat{s}_k$ from one iteration to another. Figure 4.4 illustrates each iteration of the estimation algorithm which assumes $\hat{s}_0 = 0$.

The denoiser is a non-linear function $\zeta_k : \mathbb{R} \longrightarrow \mathbb{R}$ applied to each element of z_k. Many existing MIMO detection schemes [83, 84] assume that the noise $z_k - s$ at the input of the denoiser has an i.i.d Gaussian distribution with a variance σ_k^2. In this perspective, an optimal element-wise denoising function is given by

$$\zeta(z; \sigma_k^2) = \frac{1}{Z} \sum_{s_i \in \mathcal{S}} s_i \exp\left(-\frac{\|s_i - z\|^2}{\sigma_k^2}\right) \tag{4.14}$$

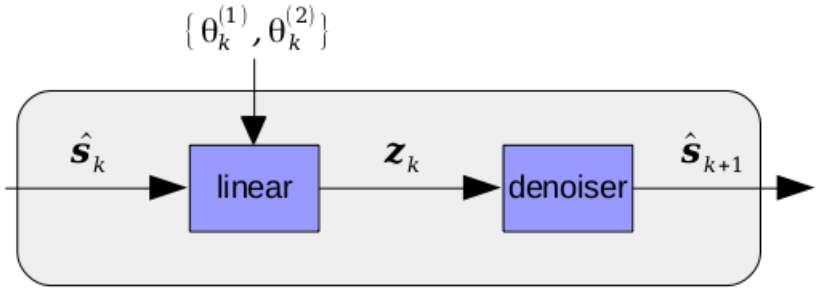


FIGURE 4.4: One block of an iterative estimation.

where $Z = \sum_{s_i \in \mathcal{S}} \exp\left(-\frac{\|s_i - z\|^2}{\sigma_k^2}\right)$. The noise $z_k - s$ consists of three different noise kinds: the channel noise, the contribution of the clipping and quantization noise, and the residual error caused by deviation of s_k from the true value of s. As the clipping and quantization noise are challenging to characterize, we predict the variance σ_k^2 as a function of the channel noise as follows

$$\sigma_k^2 = \theta_k^{(4)} \times \left(\sigma_w^2 + \theta_k^{(3)}\right) \tag{4.15}$$

where $\theta_k^{(3)} > 0$ and $\theta_k^{(4)} > 0$ are the parameters to optimize in the kth iteration during the training phase. As we can see, the standard deviation of the input noise at the denoisers, σ_k, varies from iteration to another and depends on the linear steps in each iteration. Figure 4.5 shows a flowchart representing a single layer of QDNet which corresponds to the kth iteration of the estimation process. The model has only four parameters per layer: $\theta_k^{(1)}$, $\theta_k^{(2)}$, $\theta_k^{(3)}$, and $\theta_k^{(4)}$. These parameters are optimized during the training phase over randomly sampled i.i.d Gaussian channels. The training is done offline, and then the optimized parameters of the NN are used for the entire communication phase

Unlike DetNet introduced in [76], a sparsely connected NN is used in our architecture to highly reduce the computational loads as proposed in [77]. The intuition behind our NN architecture is twofold. On the first hand, the contribution of $\theta_k^{(1)}$ and $\theta_k^{(2)}$ is useful to compensate for the degradation induced by the corrupted observation $\tilde{r}$. On the other hand, the introduction of the denoising function $\zeta_k(.)$ is convenient to improve the quality of the estimate s_k from one iteration to the next. The goal of QDNet is to mimic the transmission chain in the uplink C-RAN such that we can find appropriate transformation and decoding schemes that minimize the estimation error at the CP side while adhering to some quantization rules at the BSs. Thus, two different tasks need to be accomplished by QDNet, namely, optimize the parameters $\{v_n\}_{n=1}^{N}$ at the BSs, and also optimize the parameters $\{\theta_k^{(1)}, \theta_k^{(2)}, \theta_k^{(3)}, \theta_k^{(4)}\}_{k=0}^{L-1}$ where L is the number of layers used for decoding at the CP side. Figure 4.6 summarizes

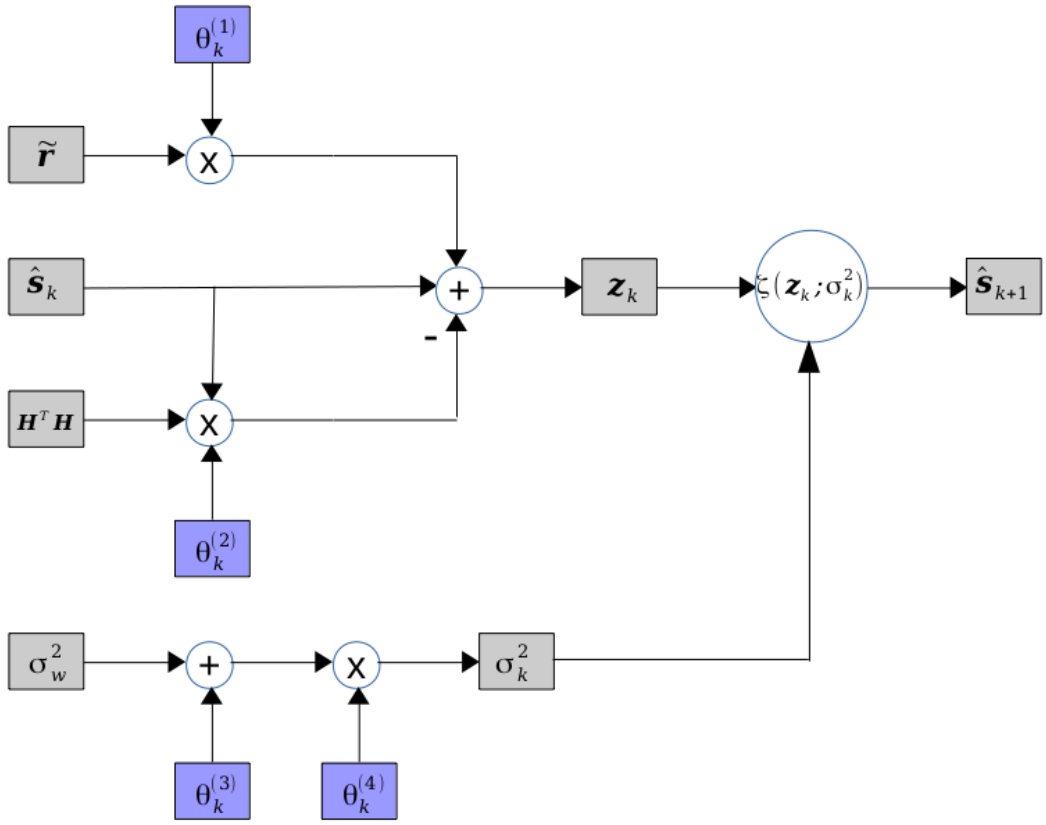


FIGURE 4.5: A flowchart representing a single layer of QDNet at the CP side.

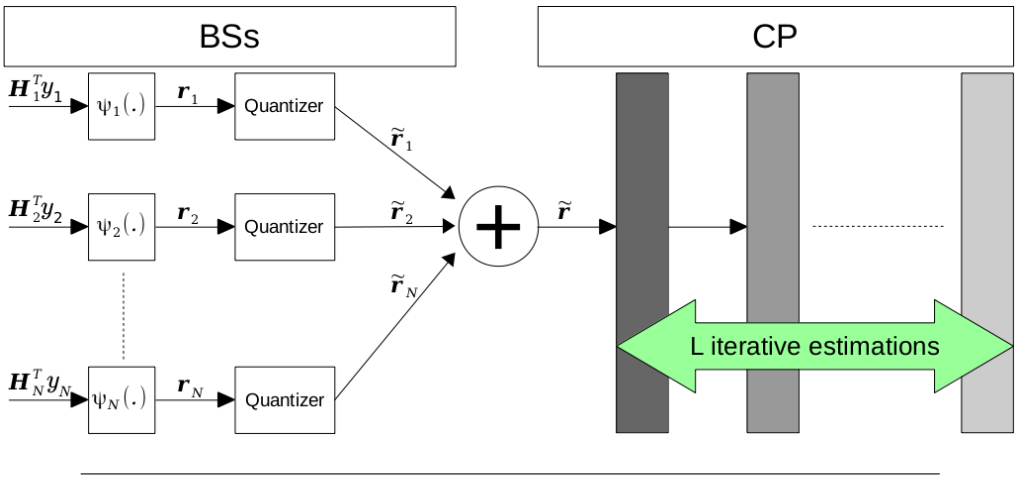


FIGURE 4.6: Illustration of the QDNet architecture for uplink C-RAN.

the transmission system in uplink C-RAN as modelled by the QDNet design. The received signals at BSs are quantized and sent in parallel to the CP. Once signals are reinforced at the CP side, the shared network of QDNet estimates all users' transmitted message.

4.4.3 QDNet Complexity

The QDNet is proposed to simplify the detection network model in uplink C-RAN, and the simplification runs throughout the entire data processing. First, the network connection structure is simplified by working with sparse connectivity instead of full connectivity, which reduces the number of required FLOPS. Each node of input is only connected to one node of the output. Second, the input/output variables and the loss function are optimized to avoid irreversible problems with the channel matrix. Moreover, the gain in processing complexity is not negligible as no matrix inversion is needed to be calculated. For example, the ZF or MMSE detector's matrix inversion using Gauss-Jordan elimination requires $O((KN_t)^3)$ elementary operations. This complexity is skipped in the QDNet scheme. Based on the above improvements, the network complexity is reduced to $O((KN_t)^2)$. This can be seen from the flowchart representing a single layer in Figure 4.5, which shows the operations of multiplications and the element-wise denoising function. Hence, the $O((KN_t)^3)$ complexity of classical linear detectors is reduced to $O((KN_t)^2)$ with the proposed QDNet model.

4.5 Experiments

In this section, we show the performance and advantages of our proposed QDNet using computer simulations. The performance evaluation of QDNet is given for i.i.d Gaussian channels.

4.5.1 Implementation details

In our simulation, the QDNet is implemented in Keras. The number of layers, i.e., the number of iterations in the estimation algorithm is set to $L = 10$. Training and test data are generated through the model described in section 4.3. They consist of randomly generated sources: the signal $\boldsymbol{s}$, the channel matrix $\boldsymbol{H}_n$ correspondent to each nth BS in the C-RAN system, and the channel noise $\boldsymbol{w}_n \; \forall \; n \in \{1, \ldots, N\}$. The transmitted data $\boldsymbol{s}$ is generated randomly and uniformly from Q-ary QAM modulation symbols. All users in the system are assumed to use the same modulation. All simulation channels are given fast fading channels randomly generated with i.i.d $\mathcal{CN}(0, 1/N_r)$ elements. During training, the SNR is uniformly distributed on $[\text{SNR}_{\text{min}}, \text{SNR}_{\text{max}}]$ where SNR_{min} and SNR_{max} are the minimal and maximal SNR values over which we used the network. The SNR of the system is used to measure the noise level and is defined as

$$\text{SNR(dB)} = 10 \log_{10} \left(\frac{K \times N_t}{N_r \times 2\sigma_w^2} \right) \tag{4.16}$$

We train the QDNet network with 10000 iterations with a batch size of 500 samples. For the optimization algorithm in training, Adam [85] method is employed, and the learning rate is set to 0.001. In our experiment settings, we choose the MSE as defined in (4.6) as the cost function.

4.5.2 Competing schemes

The system performance depends on the considered MIMO detector. Therefore, we have tested the performance of the following detection schemes:

- **QZF**: ZF detector [86] with quantized observations of $\boldsymbol{H}^T \boldsymbol{y}$.
- **QSD**: SD algorithm [86] with quantized observations of $\boldsymbol{U}^T \boldsymbol{y}$ where $\boldsymbol{U}$ is an $2N_r \times 2KN_t$ semi-unitary matrix. The latter results from the singular value decomposition (SVD) of the channel matrix $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.
- **QDNet**: Our proposed NN algorithm is described in section 4.4. We implement this NN with 2 layers at the BS side for transformation and quantization, and $L = 10$ layers at the CP side for decoding. The layer for transformation

of $\boldsymbol{H}_n^T \boldsymbol{y}_n \; \forall \; n \in \{1, \ldots, N\}$ contains only one trainable parameter which represents the clipped threshold v_n. In contrast, each kth layer for decoding at the CP side contains four trainable parameters $\{\theta_k^{(1)}, \theta_k^{(2)}, \theta_k^{(3)}, \theta_k^{(4)}\}$. Intuitively, the total number of parameters in the QDNet network is equal to $N + 4L$. This demonstrates that the complexity of QDNet is not high compared to other NN architectures.

In our work, we choose to compare the QDNet results to well-known decoders for detection: the sub-optimal ZF and the optimal SD. The comparison is fair since the quantization process is the same for all detection schemes. In this perspective, we should mention that our proposed algorithm, QDNet, is not optimal when considering complete observations, i.e., when no quantizer is applied. Indeed, the SD algorithm has been developed to attain low complexity with the ML performance [24]. Thus, the QSD scheme outperforms QDNet as we move to a high number of quantization bits. However, QDNet performance is significant as we address fewer quantization bits, and even QDNet can outperform the QSD.

4.5.3 Quantization model

The scalar quantization encoding with R_q bits is performed in all detection algorithms before sending observations from BSs to the CP. Although, as one would expect, this is not ideal and will not approach any theoretical limits, scalar quantization is a relatively simple technique commonly implemented in hardware architecture. We have used the uniform quantization in our work, which represents the simplest form of scalar quantization. The peak-to-average power ratio of the signal can be limited by clipping the signal amplitude. This would help to reduce later the quantization error. The clipped signal is defined by

$$\tilde{x}(t) = \begin{cases} x(t) & |x(t)| < v_{th} \\ v_{th} & x(t) \geq v_{th} \\ \text{-}v_{th} & x(t) \leq \text{-}v_{th} \end{cases} \tag{4.17}$$

where $v_{th} > 0$ is the clipping threshold.

As the clipping process introduces additional noise, a trade-off between the clipping and quantization noise must be found. Therefore, we resort to machine learning as introduced in the QDNet network to find the optimal transformation by optimizing the clipping threshold v_n, $\forall \; n \; \in \{1, \ldots, N\}$ at each nth BS. For the QZF and QSD schemes, it is challenging to find this clipping threshold straightforwardly. Thus, we use a grid search to get this optimal factor to achieve the best performance subject to the number of quantization bits.

To construct the uniform quantizer in the QDNet, we employ the rounding function to the clipped signals. However, rounding is a fundamentally non-differentiable

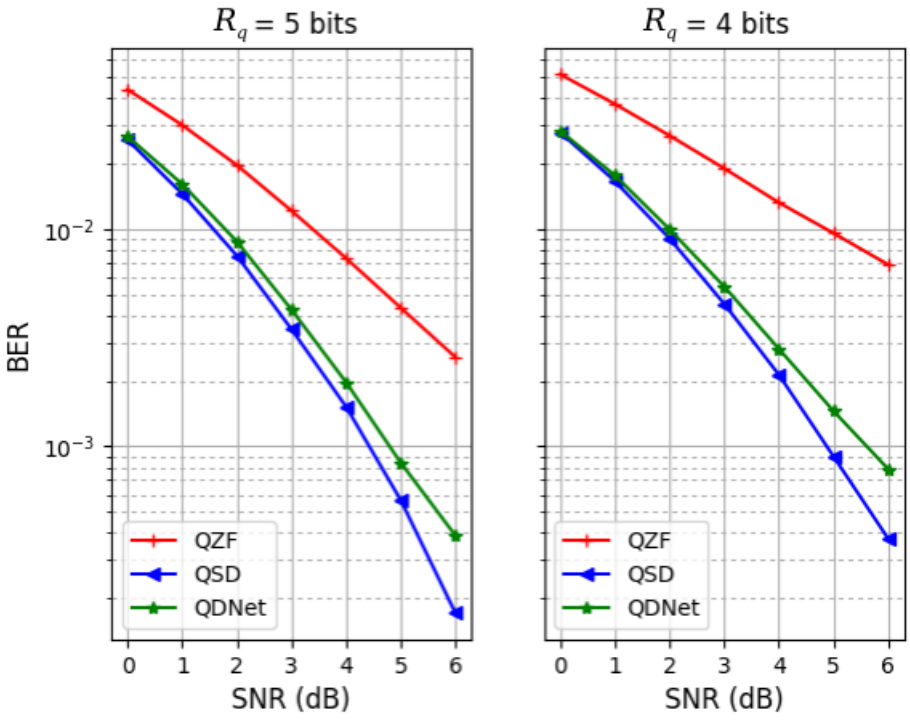


FIGURE 4.7: BER vs. SNR of different schemes for single BS scenario and 4-QAM modulation (4 transmitters and 10 receivers).

function; thus, we force its gradient in the training process equal to that of the identity function.

4.5.4 Experiment results

Single BS scenario

We consider the single BS case where only one BS is sending its observations to the CP. The experiments address a MIMO channel with an input of size $K = 4$ single antenna users and an output of $N_r = 10$ receiving antennas. We plot the bit error rate (BER) performance versus the SNR. Figures 4.7 and 4.8 show that the QDNet outperforms the QZF in the entire SNR region. It is also observed that relative performance is achieved to that of the optimal detector QSD as we move to a high number of quantization bits.

Multi-BS scenario

For the multi-BS case, we consider $N = 2$ and $N = 3$ BSs in the C-RAN system. The MIMO channel correspondent to each BS has an input of size $K = 4$ single antenna

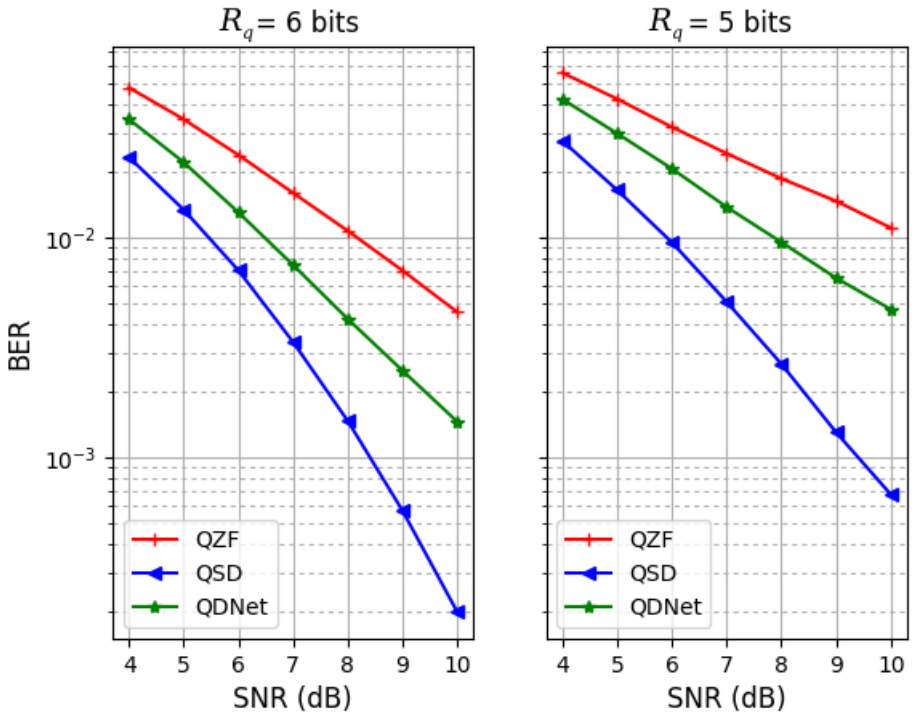


FIGURE 4.8: BER vs. SNR of different schemes for single BS scenario and 16-QAM modulation (4 transmitters and 10 receivers).

users and an output of $N_r = 6$ receiving antennas. In this setting, the QZF and QSD perform joint detection with a high degree of cooperation among BSs to exchange information. This can be regarded as one large distributed MIMO system with an output of size equal to $N \times N_r$ receiving antennas.

Intuitively, the QZF and QSD schemes perform detection established on the quantized version of the signal $\boldsymbol{H}^T\boldsymbol{y}$ instead of the signal $\tilde{\boldsymbol{r}}$ which is equal to the sum of all quantized signals sent by BSs to the CP. For that reason, the two schemes have the advantage of getting a reduced quantization noise compared to the QDNet network. However, this is the task of NNs to learn from the corrupted observations, and so can result in equivalent or improved performance. Thus, we adopt the QZF and QSD schemes to enable only direct comparisons, but we should mention that these two schemes can not be applied in practical systems as long as quantized signals are sent from all BSs in the C-RAN system.

Figure 4.9 shows the BER performance versus SNR of the different schemes for 4-QAM modulation. We can see that QDNet performs well for detection as long as quantized signals from BSs are used constructively at the CP side in the QDNet architecture. The gain in QDNet performance is significant as the number of BSs increases in the C-RAN system. This indicates that the proposed NN architecture is

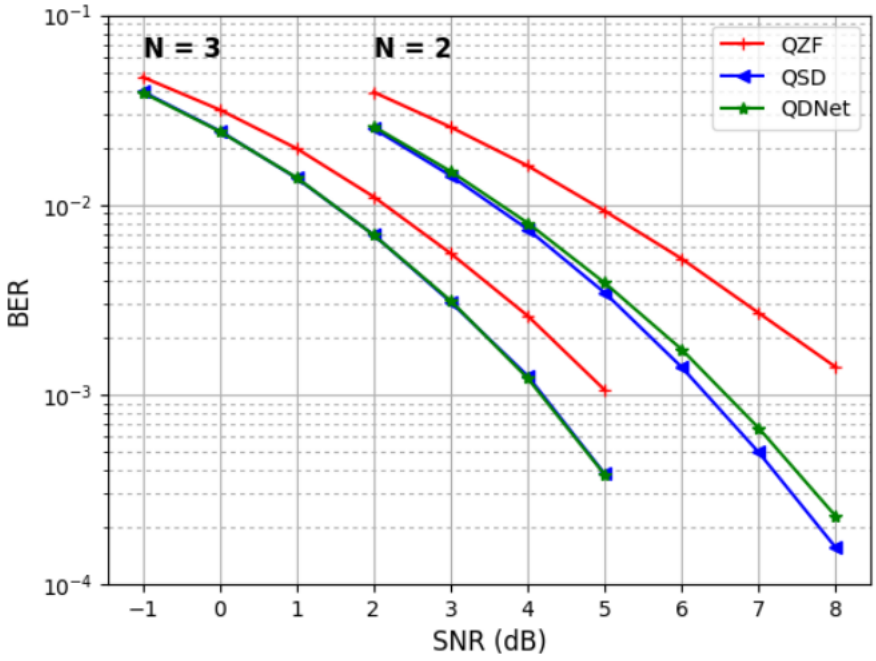


FIGURE 4.9: BER vs. SNR of different schemes for 4-QAM modulation and 5 quantization bits.

reliable for likely fronthaul compression for multi-antenna uplink C-RAN. It is more suitable for scenarios with a large scale of antennas or a high number of BSs. This is also approved in Figure 4.10 for 16-QAM modulation. Hence, we consider in the following $N = 3$ BSs to see the quantization noise effect.

By varying the number of quantization bits, we plot the BER of the different schemes to see the performance degradation at high quantization noise. For fixed SNRs, Figure 4.11 shows the BER as a function of R_q for 4-QAM modulation. It is well observed that the QDNet performance outperforms that of the QSD scheme as we move to less number of quantization bits. This improvement in BER can be explained by the fact that the QDNet has well learned from the corrupted observations, especially at high quantization noise levels. Figure 4.12 also shows that QDNet behaviour is confirmed for 16-QAM modulation.

4.6 Summary

This chapter proposed a distributed DNN architecture, called QDNet, to design an efficient scheme for fronthaul compression in multi-antenna uplink C-RAN. QDNet includes the quantization process at the BSs and the decoding process at the CP. So the challenge was to design a distributed NN which is jointly optimized at the

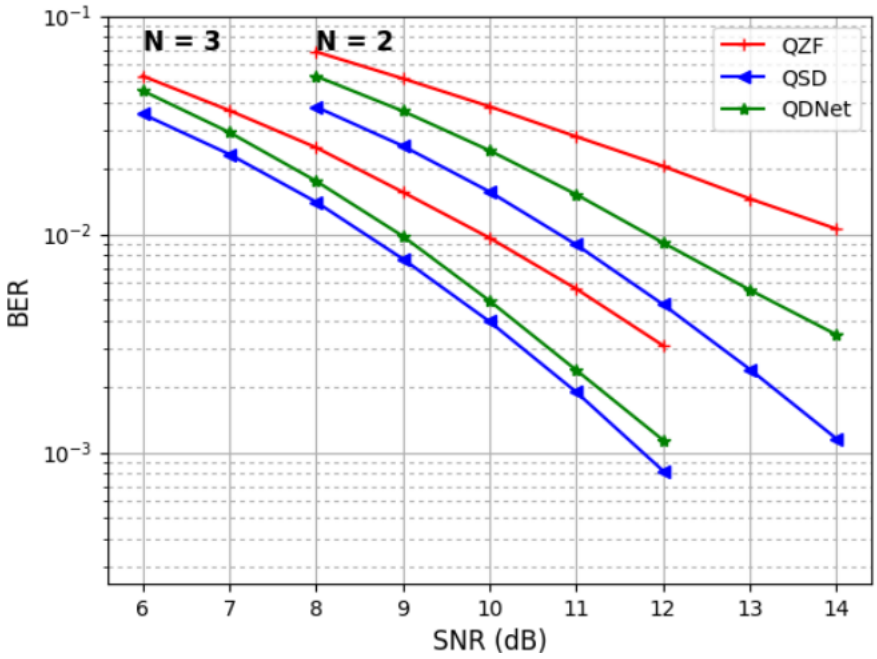


FIGURE 4.10: BER vs. SNR of different schemes for 16-QAM modulation and 5 quantization bits.

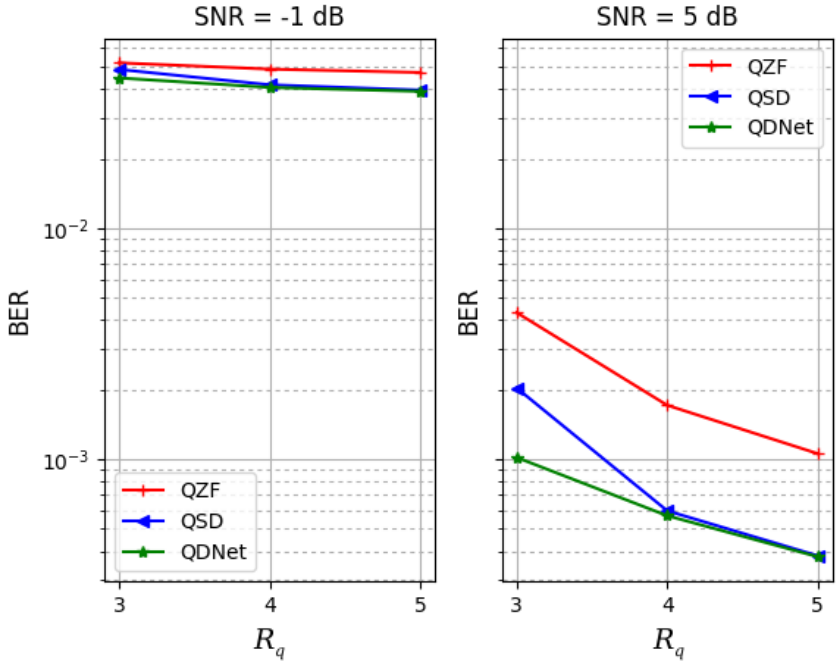


FIGURE 4.11: BER vs. number of quantization bits for 4-QAM modulation.

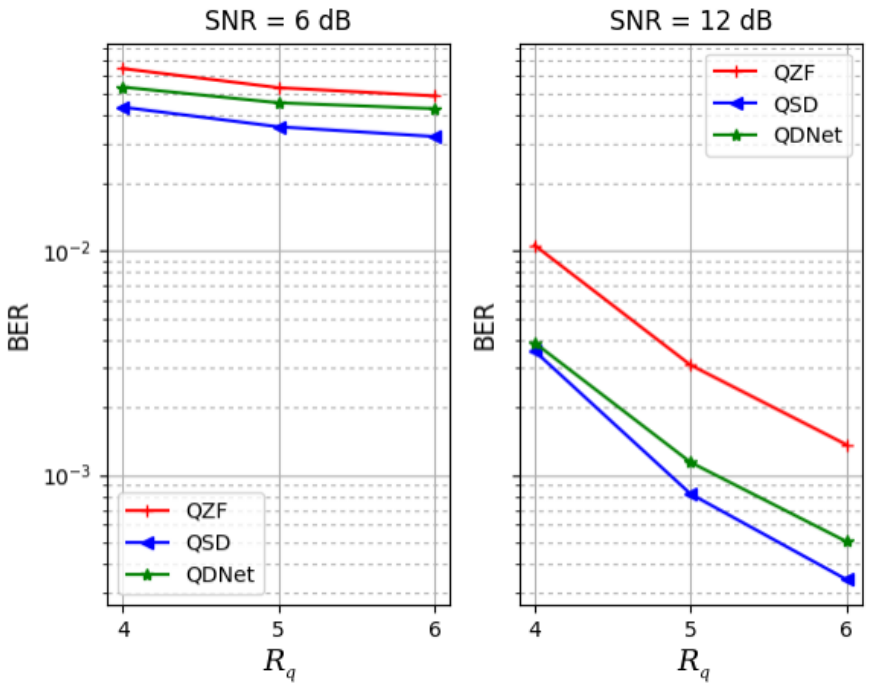


FIGURE 4.12: BER vs. number of quantization bits for 16-QAM modulation.

BSs and the CP. Our goal was not to solve the signal detection in multi-antenna systems but to mimic the whole transmission in uplink C-RAN. We compare the proposed NN results to well-known receivers for detection: the sub-optimal ZF and the optimal SD. We did not compare NN-assisted receivers in the literature since all works assume perfect knowledge of observations contrary to our work which deals with corrupted ones. Our paper has used the uniform quantization encoding with R_q bits in all detection algorithms before sending observation from BSs to the CP. Hence, the capacity constraint is defined by the number of quantization bits R_q. We have used a sparsely connected NN to reduce computational loads highly while maintaining outstanding performance than existing detection algorithms.

Chapter 5

MU-MIMO Precoding in the Downlink Transmission

5.1 Introduction

In Chapters 3 and 4, we have studied the MIMO decoding in the uplink reception (many-to-one). This chapter shifts the focus to the downlink scenario (one-to-many) and studies the MU-MIMO precoding techniques. We develop reliable transmission schemes for multi-user communications in the single-cell environment based on linear and non-linear precoding techniques.

The first part of this chapter presents a combined VP (Comb-VP) precoding for MU-MIMO downlink systems, taking into account different MCSs. It enables an adaptive modulation scenario where users apply different MCSs. The perturbation vector's search is combined for all users like in the conventional VP (Conv-VP), but different modulations are used simultaneously. The performance of the combined VP is optimal compared to existing solutions. We propose the combined MMSE VP in this setting, which achieves better error rate performance by minimizing the MSE criterion. Besides, we suggest an ordering of users according to their modulation size. Indeed, by starting with the highest modulation order in the search tree of the SE algorithm, this has the advantage of reducing the complexity by minimizing the size of the search tree in the average sense.

The second part of the chapter introduces a low-complexity block VP (Block-VP) precoding, which has the advantage of choosing the desired diversity order by fixing the size of blocks. The idea of Block-VP is to apply perturbation for each block resulting from the QR-decomposition of the precoding matrix, taking into account the feedback information from previously perturbed blocks. The block division may be interpreted as a user grouping method in such scenarios of adaptive modulation,

and its performance is better than other precoding techniques covered in the literature.

Finally, this chapter considers the downlink precoding in MU-MIMO when CSI errors are present in cellular networks due to imperfect estimation and quantization. These CSI impairments can degrade significantly the performance of precoders used to mitigate the intra-cell interference. Most importantly, several users are present in the network with different CSI accuracy, making the precoder more sensitive to CSI errors. Thus, we propose a new feedback quantity referred to as the CSI accuracy indicator (CSIAI). This quantity will be transmitted by the user equipment (UE) to the BS to cope with CSI errors. The design of an appropriate precoder based on the CSIAI reporting has the advantage of achieving a better performance in the downlink transmission. The system design can lower or eliminate the ceiling effects. Simulation results show that an improvement in the average symbol error rate (SER) performance is achieved.

5.2 MU-MIMO Precoding for Adaptive Modulation

Conventional VP precoding does not exploit the fact that users use different MCSs, depending on the SINR. In [87], block diagonalization and VP are combined to propose the block diagonalized VP (BD-VP). The latter enables different users to apply various modulation schemes. Besides, authors in [88] propose the user grouping VP (UG-VP), which improves BD-VP performance. These existing solutions are suboptimal since VP is applied for each user or group independently. To eliminate the performance loss, authors in [89] propose a joint VP algorithm applied to adaptive modulation scenarios. By scaling the constellation, the modulo base for different modulation types is made the same; thus, the joint VP reaches a comparable performance with the conventional VP.

To keep also the performance advantage of conventional VP, we propose in our work, a combined VP to mitigate downlink interference between users applying different MCSs. We also introduce the combined MMSE VP that achieves the best performance by minimizing the end-to-end MSE. Indeed, the new design of the combined VP includes a diagonal matrix $\boldsymbol{T}$ instead of the scalar modulo base τ which has a constant value:

$$\boldsymbol{T} = \begin{bmatrix} \tau_1 & & \\ & \ddots & \\ & & \tau_K \end{bmatrix} \tag{5.1}$$

The data symbol vector $\boldsymbol{s}$ is perturbed by adding a perturbation signal $\boldsymbol{T}\boldsymbol{t}$, where $\boldsymbol{T}$ is a diagonal matrix of elements equal to the modulo bases relative to each modulation

type, and $\boldsymbol{t}$ is the K-dimensional integer vector. Then, the transmit signal can be expressed as

$$\boldsymbol{x} = \frac{1}{\sqrt{\beta}}\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{T}\boldsymbol{t}) \tag{5.2}$$

where $\beta = \|\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{T}\boldsymbol{t})\|^2$ to satisfy the unit transmit power. The combined VP can be represented as an integer-lattice search where at the transmitter $\hat{\boldsymbol{t}}$ is chosen such that β is minimized, that is

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t}\in\mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}} \|\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{T}\boldsymbol{t})\|^2 \tag{5.3}$$

With this transformation, the optimal perturbation vector's search tree can be applied within K dimensions so that there will be no performance loss compared to BD-VP and UG-VP, sub-optimal. Elements of the diagonal matrix $\boldsymbol{T}$ may take any order so that τ_i corresponds to the ith user using modulation type $\mathcal{M}_i$. Regarding this setting, we propose a user ordering which allows complexity to decrease. Without loss of generality, we assume there are Q modulation types applied, denoted as $\mathcal{M}_1, \ldots, \mathcal{M}_Q$. With the same value of the initial sphere radius, we find that starting the search in the SE algorithm with the highest modulation order is less complex than starting with any other order. The reduction in complexity comes from the fact that the SE algorithm is based on a DFS strategy. If a leaf node is reached not satisfying the metric constraint linked to the sphere radius, i.e., it is outside the search sphere, the tree's path leads to that leaf node downs to that parent node, and the SE continuous the search tree to reach the other child nodes. It should be noted that the order in which the child nodes originating from the same parent node are visited is relevant to the final complexity of the SE. When a leaf node is reached to satisfy the metric constraint, the SE updates the radius and restarts the search tree with the new metric constraint. Hence, when we consider the lowest modulation order at the bottom level of the tree, this will faster the search by reducing the SE's total number of paths. Figure 5.1 shows a simplified diagram of the search tree that would be performed in a 2×2 system with 4-QAM (± 1) and 16-QAM ($\pm 1, \pm 3$) modulations. The curve indicates the initial metric constraint, and the dashed lines indicate the discarded paths due to that metric constraint. During the search, the nodes corresponding to 4-QAM modulation (at the bottom) will be more frequently visited than the nodes at the top level. This has the advantage of reducing the size of the search tree in the average sense.

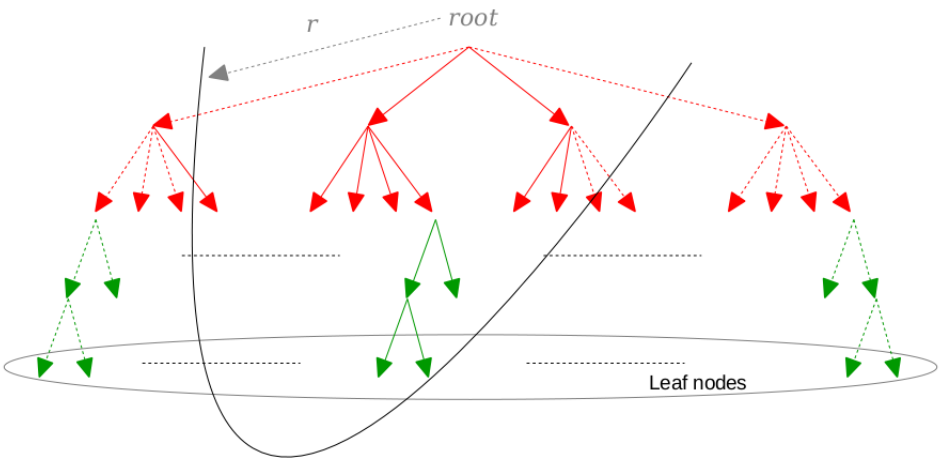


FIGURE 5.1: Search tree diagram of the SE for a 2×2 system with 16-QAM at the top level and 4-QAM at the bottom level.

Rewrite $\mathbf{F} = [\mathbf{F}_{\mathcal{M}_1}, \ldots, \mathbf{F}_{\mathcal{M}_Q}]$. The transmit signal norm can be written as

$$\begin{aligned} \|\boldsymbol{F}(\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t})\|^2 &= \left\| [\boldsymbol{F}_{\mathcal{M}_1}, \ldots, \boldsymbol{F}_{\mathcal{M}_Q}] \,.\, (\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t}) \right\|^2 \\ &= \left\| \sum_{i=\mathcal{M}_1}^{\mathcal{M}_Q} \boldsymbol{F}_i \,.\, (\boldsymbol{s}_i + \tau_i \boldsymbol{t}_i) \right\|^2 \\ &= \left\| \sum_{i=\mathcal{M}_1}^{\mathcal{M}_Q} \tau_i \boldsymbol{F}_i \,.\, (\tau_i^{-1} \boldsymbol{s}_i + \boldsymbol{t}_i) \right\|^2 \end{aligned} \tag{5.4}$$

where $\boldsymbol{t}_i$ is the perturbation vector corresponding to modulation type $\mathcal{M}_i$.

5.2.1 Combined MMSE-VP

In the literature, the conventional VP idea is extended using the mean squared error (MSE) minimization criterion, and this approach is known as the MMSE-VP [35]. It is worth mentioning that the MSE minimization approach has been successively applied to obtain precoder designs in various multi-user downlink transmission scenarios. In MMSE-VP, the precoder and the optimal perturbation vector are found jointly by minimizing the end-to-end MSE. Indeed, the MMSE-VP seeks a balance between noise enhancement suppression and residual interference mitigation. However, the conventional-VP finds the optimal perturbation vector that minimizes the noise enhancement effect for a given linear precoder such as ZF and regularized-ZF. For these reasons, the MMSE-VP achieves better BER performance compared to conventional-VP in the entire SNR region.

Based on the idea of the combined VP, we propose the combined MMSE-VP for adaptive modulation to minimize the MSE criterion. We define a deviation vector to measure the distortion between the scaled received vector $\sqrt{\beta}\boldsymbol{y}$ and the perturbed vector $\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t}$ as

$$\boldsymbol{d} = \sqrt{\beta}\boldsymbol{y} - (\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t}) = (\boldsymbol{H}\boldsymbol{F}-\boldsymbol{I})(\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t}) + \sqrt{\beta}\boldsymbol{w} \tag{5.5}$$

Given the data vector $\boldsymbol{s}$ and the channel matrix $\boldsymbol{H}$, the MSE is expressed as a function of $\boldsymbol{v} = \boldsymbol{T}\boldsymbol{t}$ and $\boldsymbol{F}$

$$\begin{aligned} e(\boldsymbol{v},\boldsymbol{F}) &= \mathbb{E}_{\boldsymbol{w}}(\|\boldsymbol{d}\|^2|\boldsymbol{H},\boldsymbol{s}) \\ &= \|(\boldsymbol{H}\boldsymbol{F}-\boldsymbol{I})(\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t})\|^2 + K\beta\sigma_w^2 \end{aligned} \tag{5.6}$$

where $\mathbb{E}_{\boldsymbol{w}}(.)$ denotes the noise expectation. The optimal precoding matrix that minimizes (5.6) is obtained referring to [35]

$$\boldsymbol{F}_o = \boldsymbol{H}^H\left(\boldsymbol{H}\boldsymbol{H}^H + K\sigma_w^2\boldsymbol{I}\right)^{-1} \tag{5.7}$$

The optimal perturbation vector can be found as

$$\boldsymbol{t}_o = \underset{\boldsymbol{t}\in\mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}}\ K\sigma_w^2\tilde{\boldsymbol{s}}^H\left(\boldsymbol{H}\boldsymbol{H}^H + K\sigma_w^2\boldsymbol{I}\right)^{-1}\tilde{\boldsymbol{s}} \tag{5.8}$$

where $\tilde{\boldsymbol{s}} = \boldsymbol{s}+\boldsymbol{T}\boldsymbol{t}$ denotes the perturbed data vector. With Cholesky factorization, (5.8) can be rewritten as

$$\boldsymbol{t}_o = \underset{\boldsymbol{t}\in\mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}}\ \|\boldsymbol{L}^H(\boldsymbol{s}+\boldsymbol{T}\boldsymbol{t})\|^2 \tag{5.9}$$

where $\left(\boldsymbol{H}\boldsymbol{H}^H + K\sigma_w^2\boldsymbol{I}_K\right)^{-1} = \boldsymbol{L}\boldsymbol{L}^H$ with $\boldsymbol{L}$ a lower triangular matrix.

5.2.2 Simulation results

This section evaluates the proposed combined VP scheme's performance with conventional VP [33], which considers the highest modulo base for all users. We also compare with UG-VP [88], which outperforms BD-VP when reducing the number of individual searches for the perturbation vector, i.e., the number of different groups. We consider an MU-MISO BC with a BS equipped with $M = 8$ transmit antennas serving $K = 8$ single antenna users simultaneously. We note that the proposed scheme can be applied to any MU-MIMO system with a total number of receive antennas N_r ($\leq M$). The performance is evaluated in terms of BER versus the SNR, and we average the performance through Monte Carlo simulations.

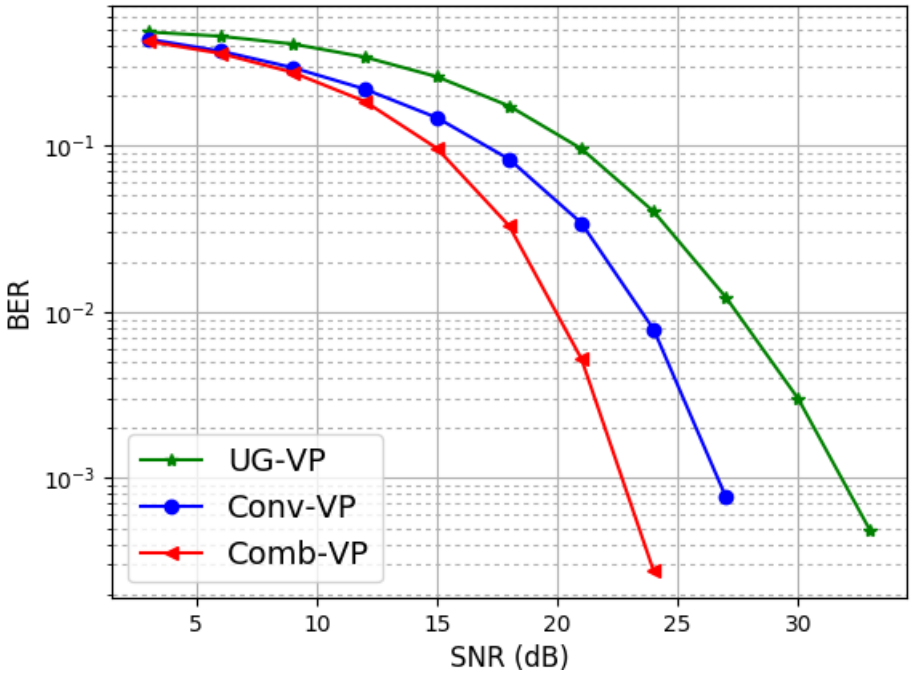


FIGURE 5.2: Averaged BER of all users applying for 3 different modulation types.

Figure 5.2 shows the BER performance averaged over all users of UG-VP, Conv-VP and Comb-VP. It is assumed that users 1, 2 apply 4-QAM, users 3, 4, 5 apply 16-QAM and the other ones 64-QAM. It is well observed that our proposed algorithm Comb-VP outperforms UG-VP and Conv-VP, in which the highest modulo base τ is used for all users. To see the difference in performance per user, we plot in Figure 5.3 the BER of users applying different modulations versus the SNR. For example, the users' BER performance applying 16-QAM with the proposed algorithm is even better than the conventional VP. The modulo base's choice τ is crucial since it provides the decoding region around every signal constellation point. Hence, when we consider only the highest modulo base for all users, τ is made too large for users applying for smaller modulation order. Therefore, the minimization yields a null perturbation vector for these users, independently of their data symbol vectors, and the perturbation technique reduces to simple channel inversion.

For comparison, Figure 5.4 shows that the combined MMSE-VP precoder outperforms the combined ZF-VP. In general, the diversity of these two precoders is the same. However, we observed an SNR gain achieved by the combined MMSE-VP in the entire SNR region for any adaptive modulation scenario.

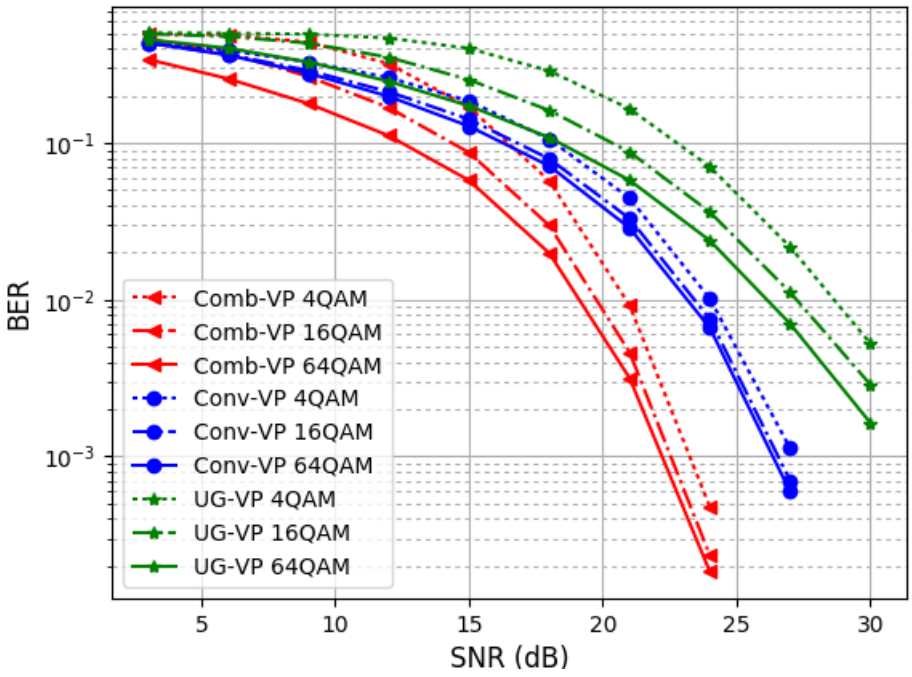


FIGURE 5.3: BER performance of the users per modulation.

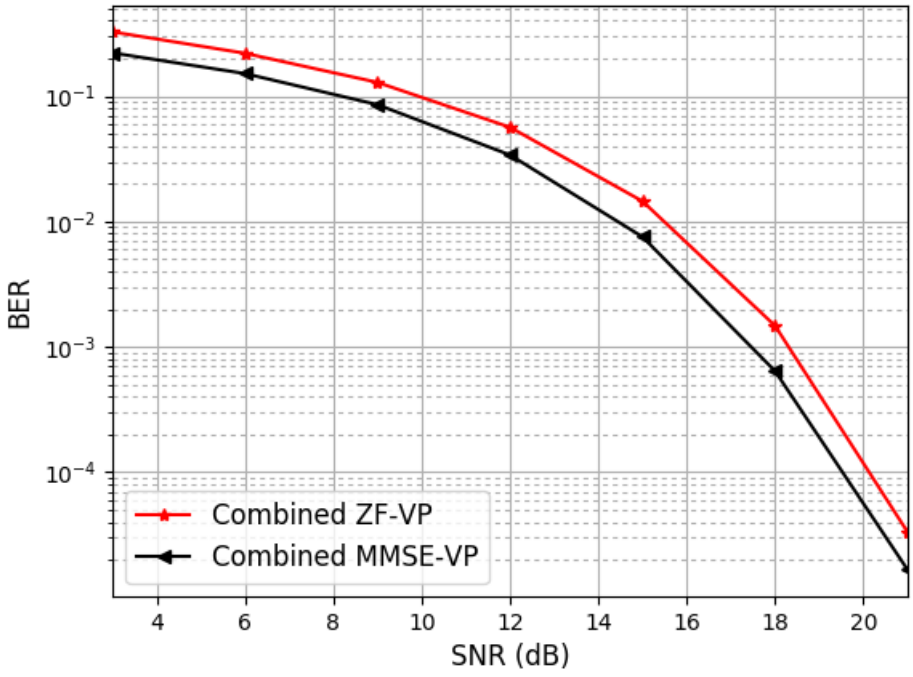


FIGURE 5.4: Performance of combined MMSE-VP precoder.

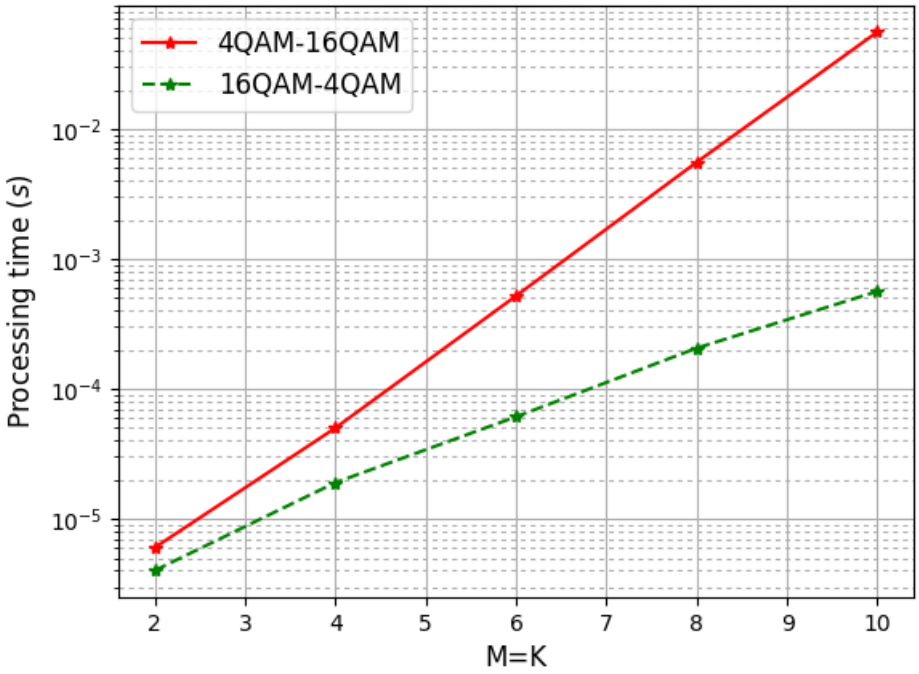


FIGURE 5.5: Ordering effect in the complexity of combined VP.

As mentioned earlier, ordering users from the highest modulation order to the lowest order has the advantage of reducing the complexity. This is shown in Figure 5.5, which visualizes the processing time as a function of the number of antennas. The processing time is measured using the same computer processor for all algorithms. We consider the case where we have $M = K$, the one-half of the users apply 4-QAM and the other half 16-QAM. The complete line shows the running time when we start the SE algorithm's search tree with 4-QAM modulation at the tree's top level. In contrast, the dashed line shows when we start with 16-QAM modulation at the top level, which is much faster to select the desired perturbation vector.

5.3 Block Recursive MU-MIMO Precoding

This section introduces a low-complexity precoding technique called the block VP (Block-VP) algorithm based on the "QR" decomposition of the precoding matrix. VP is applied for each block by taking into account the feedback information of the previously perturbed blocks. Thus the perturbation will not be applied for each group independently as in BD-VP and UG-VP. The proposed scheme allows for achieving the desired diversity order by fixing the size of blocks. We decompose the VP error power concerning the considered block division, and we derive the diversity order of all users. In Block-VP, we consider the block division of the upper triangular matrix $\boldsymbol{R}$, which is developed from the QR decomposition of the precoding

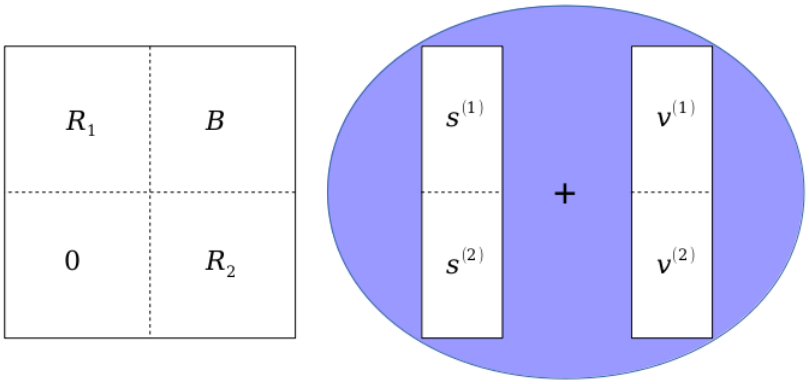


FIGURE 5.6: Block division of the precoding system.

matrix $\boldsymbol{F} = \boldsymbol{QR}$. We divide the MU-MIMO system into two blocks, as depicted in Figure 5.6. The first block is of size l_1, and the second block at the bottom is of size l_2. Accordingly the data symbol and the perturbation vectors are split into $(\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)})$ and $(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)})$, respectively.

5.3.1 Preliminaries

Without loss of generality, we assume single-antenna receivers. We note that all scheduled users are symmetric because they have the same multiplexing gain, equal to one. Hence the user is said to have d as the diversity order if its average bit error probability decays as SNR^{-d}. Moreover, because all users have the same received SNR, they all experience the same diversity order. We introduce some preliminary results, as presented in [90], which will be used in this section to calculate the diversity order when the block VP is performed.

Definition 1 (see [40]): Two lattices are said to be congruent if one can be obtained from the other by a combination of rotations and reflections, i.e., if and only if their generator matrices, $\boldsymbol{A}$ and $\boldsymbol{A}'$ are related by

$$\boldsymbol{A}' = \boldsymbol{QAU} \tag{5.10}$$

where $\boldsymbol{U}$ is a square unimodular matrix and $\boldsymbol{Q}$ is a matrix with orthonormal columns, such that $\boldsymbol{Q}^H\boldsymbol{Q} = \boldsymbol{I}$. The distance between any pair of points is preserved due to the equivalence of congruent matrices. Note that $\boldsymbol{A}$ and $\boldsymbol{A}'$ have the same number of columns but may differ in the number of rows.

Lemma 1: Let $\boldsymbol{A}$ be a generic $M \times K$ matrix, with $K \leq M$, and let $\boldsymbol{B} = (\boldsymbol{A}^H)^-$ be the Moore-Penrose pseudo-inverse of its Hermitian transpose. Let us consider a generic

partitioning of the columns of the two matrices: $\boldsymbol{A} = [\boldsymbol{A}_1|\boldsymbol{A}_2]$ and $\boldsymbol{B} = [\boldsymbol{B}_1|\boldsymbol{B}_2]$, and let us indicate with $\tilde{\boldsymbol{A}}_1$ the projection of $\boldsymbol{A}_1$ in the orthogonal complement of $\boldsymbol{A}_2$ and with $\tilde{\boldsymbol{A}}_2$ the projection of $\boldsymbol{A}_2$ in the orthogonal complement of $\boldsymbol{A}_1$

$$\tilde{\boldsymbol{A}}_1 \triangleq (\boldsymbol{I}_n - \boldsymbol{I}_{A_2})\boldsymbol{A}_1 \tag{5.11}$$

$$\tilde{\boldsymbol{A}}_2 \triangleq (\boldsymbol{I}_n - \boldsymbol{P}_{A_1})\boldsymbol{A}_2 \tag{5.12}$$

where $\boldsymbol{P}_{A_2} \triangleq (\boldsymbol{A}_2^H)^\dagger \boldsymbol{A}_2^H$ and $\boldsymbol{P}_{A_1} \triangleq (\boldsymbol{A}_1^H)^\dagger \boldsymbol{A}_1^H$ are the projection matrices on the column space of $\boldsymbol{A}_2$ and $\boldsymbol{A}_1$, respectively. The following identity holds true

$$\boldsymbol{B}_1 = (\tilde{\boldsymbol{A}}_1^H)^\dagger \tag{5.13}$$

Proof: see [90].

Corollary 1: Let $[\boldsymbol{a}_1^*, \ldots, \boldsymbol{a}_K^*]$ be the orthogonalized basis of $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K] = (\boldsymbol{A}^H)^\dagger$. The following identity holds true:

$$\|\boldsymbol{b}_K\| = \frac{1}{\|\boldsymbol{a}_K^*\|} \tag{5.14}$$

Proof: From Lemma 1, we obtain: $\boldsymbol{b}_K = ((\boldsymbol{a}_K^*)^H)^\dagger$, from which the assert follows.

Lemma 2 (see [91]): Let ρ and ξ be functions of the random channel $\boldsymbol{H}$, such that they satisfy the following relationship for any channel realization

$$\rho \geq c\xi \tag{5.15}$$

where c is a constant. Then, if ρ represents the SNR and ξ is a χ^2-distributed random variable with $2k$ degrees of freedom, the diversity order of the system d is at least k, i.e., once denoted with P_e the average BER

$$d \triangleq \lim_{\sigma_w^2 \to 0} \frac{\log P_e}{\log \sigma_w^2} \geq k \tag{5.16}$$

Similarly, if $\rho \leq \xi$, we have $d \leq k$.

Lemma 3: Let the received SNR ρ have the following lower bound

$$\rho \geq c \ \min\{\xi_1, \ldots, \xi_k\} \tag{5.17}$$

where $\rho, \xi_1, \ldots, \xi_k$ are generic functions of the transmit power and the channel realization, and c is a constant. Then, for a given random channel distribution, the error exponent is lower bounded by the minimum of the k error exponents associated with the SNRs $\rho_1 = \xi_1, \ldots, \rho_k = \xi_k$. In formulas, let P_e be the average error

probability provided by ρ and $P_{e,1}, \ldots, P_{e,k}$ the average error probabilities given by $\rho_1, \ldots, \rho_k$, respectively, then

$$\lim_{\sigma_w^2 \to 0} \frac{\log P_e}{\log \sigma_w^2} \geq \min \left\{ \lim_{\sigma_w^2 \to 0} \frac{\log P_{e,1}}{\log \sigma_w^2}, \ldots, \lim_{\sigma_w^2 \to 0} \frac{\log P_{e,k}}{\log \sigma_w^2} \right\} \tag{5.18}$$

Proof: see [90].

5.3.2 Decomposition of the VP error power

Let us partition the $M \times K$ matrix $\boldsymbol{F} = (\boldsymbol{H})^\dagger$ into two parts $\boldsymbol{F} = [\boldsymbol{F}_1 | \boldsymbol{F}_2]$, with $\boldsymbol{F}_1$ of size $M \times l_1$ and $\boldsymbol{F}_2$ of size $M \times l_2$ such that $l_1 + l_2 = K$. Similarly, let us partition $\boldsymbol{s} = \left[{\boldsymbol{s}^{(1)}}^T | {\boldsymbol{s}^{(2)}}^T \right]^T$ and $\boldsymbol{H}^H = [\boldsymbol{H}_1 | \boldsymbol{H}_2]$. Let us indicate with $\tilde{\boldsymbol{H}}_1$, the projection of $\boldsymbol{H}_1$ in the orthogonal complement of $\boldsymbol{H}_1$ and with $\tilde{\boldsymbol{F}}_2$ the projection of $\boldsymbol{F}_2$ in the orthogonal complement of $\boldsymbol{F}_1$. Therefore the subspace $(\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)$ lies in the column space of $\boldsymbol{F}_1$. Besides, by applying Lemma 1, the following identities hold

$$\boldsymbol{F}_1 = (\tilde{\boldsymbol{H}}_1^H)^\dagger \tag{5.19}$$

$$\tilde{\boldsymbol{F}}_2 = (\boldsymbol{H}_2^H)^\dagger \tag{5.20}$$

Correspondingly, we can split the transmitted vector into two orthogonal components, the first one in the space spanned by $\boldsymbol{F}_1$ and the other in its orthogonal complement

$$\begin{aligned} \gamma =& \|\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{v})\|^2 \\ =& \underbrace{\|\tilde{\boldsymbol{F}}_2(\boldsymbol{s}^{(2)} + \boldsymbol{v}^{(2)})\|^2}_{\gamma_2} + \underbrace{\|(\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)(\boldsymbol{s}^{(2)} + \boldsymbol{v}^{(2)}) + \boldsymbol{F}_1(\boldsymbol{s}^{(1)} + \boldsymbol{v}^{(1)})\|^2}_{\gamma_1} \end{aligned} \tag{5.21}$$

By taking the QR-decomposition of the partitioned $\boldsymbol{F}$

$$(\boldsymbol{F}_1 \;\; \boldsymbol{F}_2) = (\boldsymbol{Q}_1 \;\; \boldsymbol{Q}_2) \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{B} \\ 0 & \boldsymbol{R}_2 \end{pmatrix} \tag{5.22}$$

we can write

$$\boldsymbol{F}_1 = \boldsymbol{Q}_1 \boldsymbol{R}_1 \tag{5.23}$$

$$\tilde{\boldsymbol{F}}_2 = \boldsymbol{Q}_2 \boldsymbol{R}_2 \tag{5.24}$$

$$\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2 = \boldsymbol{Q}_1 \boldsymbol{B} \tag{5.25}$$

Based on these identities, the transmitted power reads

$$\gamma = \underbrace{\|\boldsymbol{Q}_2\boldsymbol{R}_2(\boldsymbol{s}^{(2)}+\boldsymbol{v}^{(2)})\|^2}_{\gamma_2} + \underbrace{\|\boldsymbol{Q}_1\boldsymbol{B}(\boldsymbol{s}^{(2)}+\boldsymbol{v}^{(2)})+\boldsymbol{Q}_1\boldsymbol{R}_1(\boldsymbol{s}^{(1)}+\boldsymbol{v}^{(1)})\|^2}_{\gamma_1} \tag{5.26}$$

Let us bound the two terms in (5.21). We start with the first one, denoted as γ_2. Inspired by [34], we consider $\boldsymbol{G} = (\boldsymbol{H}_2^H)^\dagger$ as the generator matrix of an l_2-dimensional lattice in $\mathbb{C}^M$. The covering radius of this lattice is defined as

$$\zeta(\boldsymbol{G}) \triangleq \max_{\boldsymbol{a}\in\mathbb{C}^{l_2}} \min_{\boldsymbol{z}\in\mathbb{Z}\mathbb{C}^{l_2}} \|\boldsymbol{G}(\boldsymbol{a}-\boldsymbol{z})\| \tag{5.27}$$

Introducing the covering radius [40] is useful since it provides the following convenient upper bound to γ_2

$$\gamma_2 = \|\boldsymbol{G}\tilde{\boldsymbol{s}}_2\|^2 = \min_{\boldsymbol{z}\in\mathbb{Z}\mathbb{C}^{l_2}} \|\boldsymbol{G}(\boldsymbol{s}^{(2)}+\tau\boldsymbol{z})\|^2 \leq \tau^2\zeta^2(\boldsymbol{G}) \tag{5.28}$$

where $\tilde{\boldsymbol{s}}^{(2)} = \boldsymbol{s}^{(2)} + \boldsymbol{v}^{(2)}$, and τ is the scalar integer associated to the perturbation vector $\mathbf{v}_2$. Using a result of Banaszczyk [92], it follows that

$$m_1(\boldsymbol{H}_2^H)\zeta((\boldsymbol{H}_2^H)^\dagger) \leq l_2 \tag{5.29}$$

where $m_1(\boldsymbol{H}_2^H)$ denotes the first successive minima or equivalently the shortest lattice vector of the dual lattice generated by $\boldsymbol{H}_2^H$, i.e.,

$$m_1(\boldsymbol{H}_2^H) \triangleq \min_{\boldsymbol{z}\in\mathbb{Z}\mathbb{C}^M\setminus\{0\}} \|\boldsymbol{H}_2^H\boldsymbol{z}\| \tag{5.30}$$

We note that [92] applies to real-valued lattices, but (5.29) follows directly by considering one complex dimension as two real valued dimensions. Combining (5.28) and (5.29) yields

$$\gamma_2 = \|\boldsymbol{G}\tilde{\boldsymbol{s}}_2\|^2 \leq \frac{(\tau l_2)^2}{m_1^2(\boldsymbol{H}_2^H)} \tag{5.31}$$

The implication of (5.31) is that γ_2 cannot be large unless there is in the denominator a short non-zero vector in the lattice generated by $\boldsymbol{H}_2^H$. Note, however, that the existence of such a short non-zero vector is not sufficient for γ_2 to be large. Using (5.31), we obtain

$$Pr(\gamma_2 \geq \kappa) \leq Pr\left(m_1^2(\boldsymbol{H}_2^H) \leq (\tau l_2)^2\kappa^{-1}\right) \tag{5.32}$$

where $\kappa > 0$ is arbitrary. It follows from [34] that

$$Pr(m_1^2(\boldsymbol{H}_2^H) \leq \rho^{-1}) \doteq \rho^{-M} \tag{5.33}$$

where $g(\rho) \doteq h(\rho)$ is used to denote the equality to first order in the exponent

$$\lim_{\rho\to\infty} \frac{\log g(\rho)}{\log \rho} = \lim_{\rho\to\infty} \frac{\log h(\rho)}{\log \rho} \tag{5.34}$$

Now let us turn to the second term in (5.21), denoted as γ_1. The optimal perturbation vector $\boldsymbol{v}^{(1)}$ is found by searching for the point of the lattice $\boldsymbol{F}_1\boldsymbol{\Lambda}^{l_1}$ nearest to the vector $((\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)(\boldsymbol{s}^{(2)} + \boldsymbol{v}^{(2)}) + \boldsymbol{F}_1\boldsymbol{s}^{(1)}) \in \boldsymbol{F}_1\,\mathbb{C}^{l_1}$.

Two approximate solutions to the closest point problem were introduced in [93], based on the lattice reduction algorithm of [94]: the rounding-off and the nearest plane procedures which are also referred to as Babai points. We upper bound the second term in (5.21) under the rounding-off solution. Let us introduce the Lovász-reduced basis $\overline{\boldsymbol{F}}_1 = [\overline{\boldsymbol{f}}_{1,1}, \ldots, \overline{\boldsymbol{f}}_{1,l_1}]$ of the complex lattice $\boldsymbol{F}_1$, such that $\boldsymbol{F}_1 = \overline{\boldsymbol{F}}_1\boldsymbol{U}_1$ with $\boldsymbol{U}_1$ an $l_1 \times l_1$ unimodular matrix. Let us introduce the orthogonalized basis $[\overline{\boldsymbol{f}}^*_{1,1}, \ldots, \overline{\boldsymbol{f}}^*_{1,l_1}]$ of the reduced lattice, such that

$$\overline{\boldsymbol{f}}_{2,i} = \sum_{i\le j} \mu_{j,i}\overline{\boldsymbol{f}}^*_{2,j} \text{ with } \mu_{i,i} = 1 \tag{5.35}$$

Finally, let us denote with r the covering radius [40] of the lattice $\boldsymbol{\Lambda}$. In Appendix B, we show that for $1 \le l_1 \le K$, the second term in (5.21) can be upper bounded as follows

$$\begin{aligned} &\|(\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)\tilde{\boldsymbol{s}}^{(2)} + \boldsymbol{F}_1\tilde{\boldsymbol{s}}^{(1)}\|^2 \le C_{\boldsymbol{\Lambda}^{l_1}}\|\overline{\boldsymbol{f}}^*_{1,l_1}\|^2 \\ &\text{with } C_{\boldsymbol{\Lambda}^{l_1}} = \frac{3}{2} l_1 2^{l_1} r^2 \end{aligned} \tag{5.36}$$

where $\tilde{\boldsymbol{s}}^{(1)} = \boldsymbol{s}^{(1)} + \boldsymbol{v}^{(1)}$. Note that the above upper bound holds for the full lattice search as well because the error introduced by the full sphere search cannot be larger than the error given by the lattice reduction with the Babai approximation.

Let us call $\overline{\boldsymbol{H}}_1 = [\overline{\boldsymbol{h}}_{1,1}, \ldots, \overline{\boldsymbol{h}}_{1,l_1}] = (\overline{\boldsymbol{F}}_2^H)^\dagger$ the pseudo-inverse of the reduced generator matrix. Using Corollary 1, we obtain

$$\|(\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)\tilde{\boldsymbol{s}}^{(2)} + \boldsymbol{F}_1\tilde{\boldsymbol{s}}^{(1)}\|^2 \le \frac{C_{\boldsymbol{\Lambda}^{l_1}}}{\|\overline{\boldsymbol{h}}_{1,l_1}\|^2} \tag{5.37}$$

Now we show the following result.

Proposition: The lattice generated by the $M \times l_1$ random matrix $\overline{\boldsymbol{H}}_1$ is congruent to a lattice generated by a complex Gaussian matrix, say $\boldsymbol{H}_c$, of size $(M - l_2) \times l_1$. In particular, the minimum distances of the lattices $\overline{\boldsymbol{H}}_1$ and $\boldsymbol{H}_c$ have the same distribution.

Proof: see Appendix C.

Accordingly, we can relate the denominator in (5.37) to the minimum distance $m_{\boldsymbol{H}_c}$ of the lattice $\boldsymbol{H}_c\boldsymbol{\Lambda}^{l_1}$ as follows

$$\min\left\{\|\overline{\boldsymbol{h}}_{1,1}\|,\ldots,\|\overline{\boldsymbol{h}}_{1,l_1}\|\right\} \geq \min_{\boldsymbol{t}\in\boldsymbol{\Lambda}^{l_1}}\|\overline{\boldsymbol{H}}_1\boldsymbol{t}\| = m_{\boldsymbol{H}_c} \tag{5.38}$$

from which, we finally obtain

$$\gamma_1 = \|(\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)\tilde{\boldsymbol{s}}^{(2)} + \boldsymbol{F}_1\tilde{\boldsymbol{s}}^{(1)}\|^2 \leq \frac{C_{\boldsymbol{\Lambda}^{l_1}}}{m^2_{\boldsymbol{H}_c}} \tag{5.39}$$

The distribution of $m_{\boldsymbol{H}_c}$ is unknown in general, however, in [34], an upper bound is derived for complex Gaussian matrices with i.i.d. entries of zero-mean and unit-variance. By applying this result to the $(M-l_2)\times l_1$ Gaussian matrix $\boldsymbol{H}_c$, we obtain, for some constant c_1

$$Pr\left[m^2_{\boldsymbol{H}_c} \leq \frac{a}{P}\right] \leq \begin{cases} c_1(\frac{a}{P})^{M-l_2}(\log P)^{M-l_2+1} & K=M,\ P\geq ae \\ c_1(\frac{a}{P})^{M-l_2} & K<M \end{cases} \tag{5.40}$$

where the expression for $K=M$ holds for sufficiently large transmit power with e being the Euler constant.

5.3.3 Diversity order: lower bound

By combining together (5.21), (5.31), and (5.39), we obtain a lower bound on $\rho = \frac{P}{\gamma}$

$$\begin{aligned}\rho &\geq \frac{P}{2\max\{\gamma_1,\gamma_2\}} \\ &\geq \frac{1}{2}\min\left\{\frac{Pm^2_{\boldsymbol{H}_c}}{C_{\boldsymbol{\Lambda}^{l_1}}}, \frac{Pm_1^2(\boldsymbol{H}_2^H)}{(\tau l_2)^2}\right\}\end{aligned} \tag{5.41}$$

From a well-known result due to Bartlett [95] on the element distribution of the "QR" factorization of Gaussian matrices (see for example [96]), we note that the first l terms are independent random variables and χ^2_{M-l+1} distributed. The next step of the derivation is to find the error exponent associated with the term in (5.41)

$$\Gamma = \frac{Pm^2_{\boldsymbol{H}_c}}{C_{\boldsymbol{\Lambda}^{l_1}}} \tag{5.42}$$

Let us bound the bit error probability P_e as follows

$$\begin{aligned} P_e &< N_s \mathbb{E}\left[Q\left(\sqrt{\Gamma \frac{d_{\min}^2}{2}}\right)\right] \\ &\leq \frac{N_s}{2} \int_0^\infty e^{-a\frac{d_{\min}^2}{4}} f_\Gamma(a)\, \mathrm{d}a \\ &= \frac{N_s d_{\min}^2}{8} \int_0^\infty e^{-a\frac{d_{\min}^2}{4}} F_\Gamma(a)\, \mathrm{d}a \end{aligned} \tag{5.43}$$

where $f_\Gamma(a)$ and $F_\Gamma(a)$ are the density and distribution functions of Γ, respectively, $d_{\min}$ is the minimum distance between constellation points, and N_s is the maximum number of nearest neighbor symbols. In (5.43), we have used the upper bound on the Q-function: $Q(x) \leq \exp(-x^2/2)/2$, while the last equation results from the integration by parts. Now, because

$$F_\Gamma(a) = Pr\left[m_{\boldsymbol{H}_c}^2 \leq \frac{C_{\boldsymbol{\Lambda}^{l_1}} a}{P} \right] \tag{5.44}$$

by combining (5.40) and (5.44) with (5.43) and replacing all the terms that do not depend on P with constants c_2 and c_3, we obtain the following bound

$$P_e \leq \begin{cases} c_2 \frac{(\log P)^{M-l_2+1}}{P^{M-l_2}} & K = M,\ P \to \infty \\ c_3 \frac{1}{P^{M-l_2}} & K < M \end{cases} \tag{5.45}$$

Therefore, we obtain for any K

$$\lim_{P\to\infty} -\frac{\log P_e}{\log P} \geq M - l_2 \tag{5.46}$$

Finally from Lemmas 2 and 3 applied to (5.41), we conclude

$$d \geq \min\{M, M-l_2\} = M - l_2 \tag{5.47}$$

5.3.4 Diversity order: upper bound

We will show in this section that the lower bound in (5.47) is also an upper bound, then the diversity order is precisely $M - l_2$. First, we acknowledge that with fixed M and l_2, if we increase K, $l_1 = K - l_2$ is increased, the diversity order in the system cannot increase because of the fundamental trade-off between diversity and multiplexing. Let us call d_1 the diversity order for a system with $l_2 = K - 1$, and $l_1 = 1$. We can derive an upper bound for d_1 by lower bounding the power normalization factor. From (5.21), after noting that $\boldsymbol{F}_1$ is an $M \times 1$ vector and $(\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2) = \boldsymbol{F}_1 \boldsymbol{u}^{(2)} \in \boldsymbol{F}_1 \mathbb{C}$,

where $\boldsymbol{u}^{(2)}$ is a linear combination of the data symbol vector $\boldsymbol{s}^{(2)}$, we obtain

$$\begin{aligned}\gamma &\geq \|\boldsymbol{F}_1(\boldsymbol{u}^{(2)} + \boldsymbol{s}^{(1)} + \boldsymbol{v}^{(1)})\|^2 \\ &= \|\boldsymbol{F}_1\|^2 \left| \mathcal{Q}_{\boldsymbol{\Lambda}}(\boldsymbol{u}^{(2)} + \boldsymbol{s}^{(1)}) - (\boldsymbol{u}^{(2)} + \boldsymbol{s}^{(1)}) \right|^2\end{aligned} \tag{5.48}$$

where $\mathcal{Q}_{\boldsymbol{\Lambda}}$ denotes the quantization processing in the complex lattice $\boldsymbol{\Lambda}$. From Corollary 1 and (5.19), it follows that $\|\boldsymbol{F}_1\|^2 = 1/\|\tilde{\boldsymbol{H}}_1\|^2$, where $\tilde{\boldsymbol{H}}_1$ is the projection of the vector $\boldsymbol{H}_1$ in the orthogonal complement of the $M \times l_2$ matrix $\boldsymbol{H}_2$. By taking the expectation of (5.48) concerning the data, we obtain

$$\begin{aligned}\gamma &\geq \frac{\mathbb{E}\left[|\mathcal{Q}_{\boldsymbol{\Lambda}}(\boldsymbol{u}^{(2)} + \boldsymbol{s}^{(1)}) - (\boldsymbol{u}^{(2)} + \boldsymbol{s}^{(1)})|^2\right]}{\|\tilde{\boldsymbol{H}}_1\|^2} \\ &= \frac{c_4}{\|\tilde{\boldsymbol{H}}_1\|^2},\ c_4 \neq 0\end{aligned} \tag{5.49}$$

for some constant c_4, where the expectation is non zero because $\boldsymbol{u}^{(2)}$ and $\boldsymbol{s}^{(1)}$ are independent. If the numerator of (5.49) is zero then $(\boldsymbol{u}^{(2)} + \boldsymbol{s}^{(1)}) \in \boldsymbol{\Lambda}$. Therefore, if $\boldsymbol{\Lambda} = \tau\mathcal{Z}[j]$, there exists a Gaussian integer $(k_R + jk_I) \in \mathcal{Z}[j]$, such that $\boldsymbol{u}^{(2)} = \tau(k_R + jk_I) - \boldsymbol{s}^{(1)}$, which implies that $\boldsymbol{u}^{(2)}$ and $\boldsymbol{s}^{(1)}$ are not independent. Finally, we obtain an upper bound on $\frac{P}{\gamma}$, which reads

$$\frac{P}{\gamma} \leq \frac{P\|\tilde{\boldsymbol{H}}_1\|^2}{c_4} \tag{5.50}$$

From the Bartlett decomposition in [96] on the $M \times (1 + l_2)$ Gaussian matrix $\boldsymbol{H}^H$, we know that, by construction, $\|\tilde{\boldsymbol{H}}_1\|^2$ is χ^2-distributed with $2(M - l_2)$ degrees of freedom. Therefore, from Lemma 2, we obtain

$$d \leq d_1 \leq M - l_2 \tag{5.51}$$

By combining (5.47) and (5.51), we can conclude that in an M transmit antennas MIMO Gaussian BC with K ($\leq M$) single-antenna users, the diversity order for each user achieved by the Block-VP is

$$d = M - l_2 \tag{5.52}$$

where $0 \leq l_2 \leq K - 1$ is the size of the block at the bottom. Correspondingly, we can say that in terms of diversity, the size of the first block at the top $l_1 = K - l_2$ defines the target diversity order $d = M - K + l_1$. Consequently, the proposed block VP achieves a complexity reduction coupled with the desired diversity order.

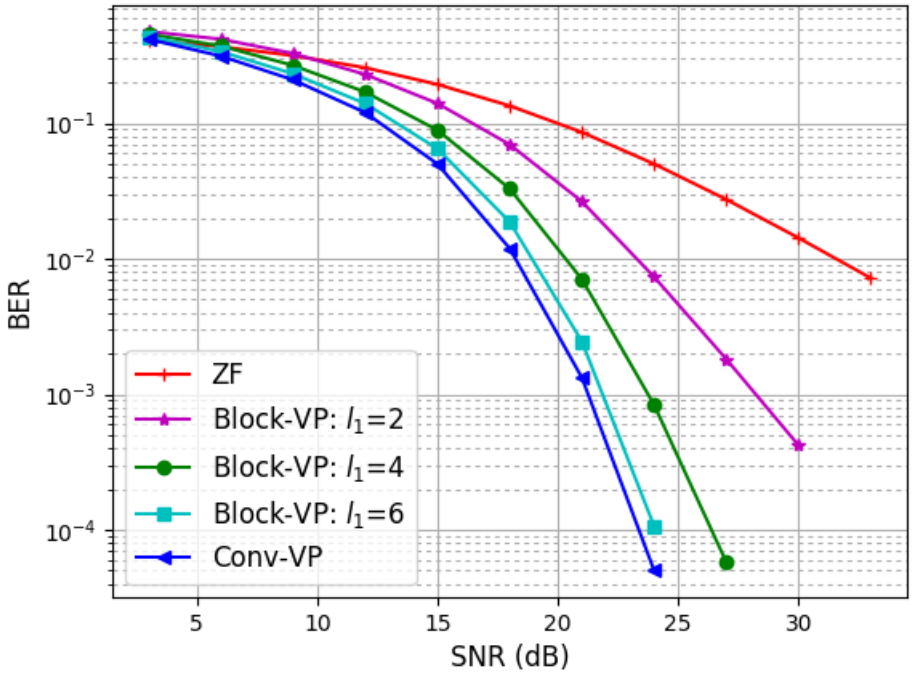


FIGURE 5.7: Block VP results for an 8×8 system with variable block sizes.

5.3.5 Simulation results

To validate our analysis for the proposed Block-VP, we calculate the error exponent numerically from the BER versus SNR curves. Figure 5.7 shows results of the Block-VP, for a square system with $M = K = 8$ and a block division into two blocks with variable size $l_1 = 2, 4$ and 6. The modulation scheme is 16-QAM. Simulation results confirm that the diversity order is $M - K + l_1$ in all cases.

Figure 5.8 shows the curves for the scenario where we have two modulation types, 4-QAM and 16-QAM. Users $1, 2, 3, 4$ apply 4-QAM and the other ones 16-QAM. In this case, we consider a block division or a user grouping based on the type of modulations. Users with the same modulation order are allocated into the same group, i.e., the same block. It is well observed that the Block-VP outperforms the UG-VP algorithm and has comparable performance to the conventional VP, which uses the modulo base τ relative to the highest modulation order. The Block-VP advantage is that the complexity is reduced compared to the conventional VP since the perturbation vector's search is divided into searches with smaller sizes. However, we should note that the Comb-VP always has optimal performance compared to all schemes.

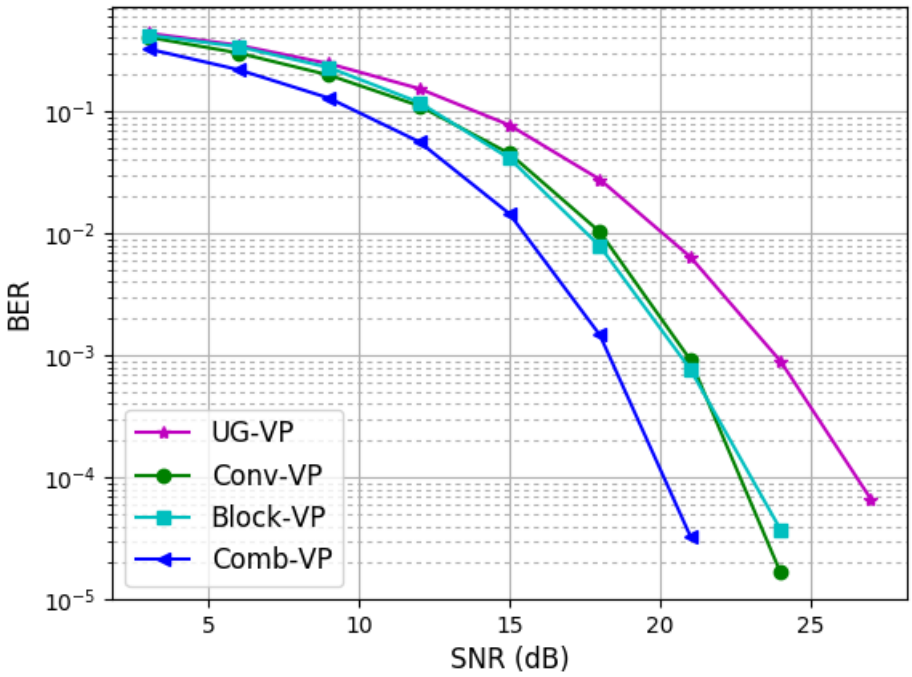


FIGURE 5.8: Block VP for an 8×8 system with two modulation orders.

5.4 Precoding for Users with Different CSI Accuracy

MU-MIMO downlink precoding is a closed-loop transmission where the knowledge about CSI is needed at the BS. In a time-division duplexing (TDD) system, CSI can be acquired at the BS by exploiting its reciprocity channels. In a frequency-division duplexing (FDD) system, the BS can have quantized CSI, obtained via finite-rate feedback from users. Unfortunately, the CSI obtained at the transmitter is not perfect due to quantization and feedback errors. In TDD, ambient noise and time variation make the BS's CSI imperfect. In FDD, quantization noise causes noise in the BS's downlink CSI knowledge. For these reasons, the CSI available at the BS is never perfect in practice. Therefore the impact of CSI errors in MU-MIMO BC is an important issue that needs to be addressed. Analysis of the impact channel estimation errors on the performance can be found in many studies [97–101].

Under perfect or imperfect CSI, two VP precoding schemes, namely conventional-VP [33], and MMSE VP [35, 102], are generally used. In [103] and [99], conventional-VP was performed under the CSI Gaussian error model. The SER performance shows high error floor levels, especially in high SNR regions. The idea of conventional-VP using the MSE minimization criterion is known as the MMSE-VP. The latter seeks

a balance between noise enhancement suppression and residual interference mitigation. Thus it achieves better SER performance compared to conventional-VP in the entire SNR region under perfect CSI assumption. Motivated by this, the MMSE-VP under the Gaussian error model was investigated in [102].

5G networks employ MU-MIMO techniques for managing the intra-cell interference. It has been shown that these techniques can achieve high spectral efficiency and support high data rate user services. However, its performance depends on the channel response's accuracy, which is the leading performance-limiting factor in such scenarios. In 3GPP, the channel quality indicator (CQI) is usually reported for wireless communication systems. We propose a new approach based on reporting a new CSIAI that measures each user's channel estimation error to deal with the channel imperfections. This indicator represents the CSI accuracy and can be calculated as a function of the wireless channel conditions, e.g., quantized SNR measurement, UE SINR, and UE mobility. Based on this CSIAI reporting, we develop a downlink precoding technique that is less sensitive to CSI errors and has improved performance.

5.4.1 CSI accuracy indicator reporting

In various studies as in [99, 104] and references therein, the channel coefficients estimated at the BS deviate from the real channel coefficients by a Gaussian error. This model captures various scenarios such as errors due to channel estimation, feedback delay, channel quantization in FDD systems, and reciprocity mismatch in TDD systems. When the BS has only an estimate $\hat{\boldsymbol{H}} = [\hat{\boldsymbol{h}}_1, \ldots, \hat{\boldsymbol{h}}_K]^T$ of the channel $\boldsymbol{H}$, then the relation between $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}$ is given by

$$\boldsymbol{H} = \hat{\boldsymbol{H}} + \boldsymbol{E} \tag{5.53}$$

where we assume that the error matrix $\boldsymbol{E}$ has i.i.d. zero mean Gaussian random components. The CSI accuracy is characterized by the error matrix which is assumed to be independent of $\boldsymbol{H}$. Also, we assume that $\boldsymbol{E}$ is independent of the data vector $\boldsymbol{s}$ and the Gaussian noise $\boldsymbol{w}$. In our channel model assumption, the transmitter precoding matrix needs to be designed based on the knowledge of the estimated channel matrix $\hat{\boldsymbol{H}}$. In fact, given $\hat{\boldsymbol{H}}$, we are interested in designing the precoding matrix $\boldsymbol{F}$ at the BS such that the MSE signal at each user receiver is minimized. Therefore the optimization problem to solve is

$$\boldsymbol{F}_{\text{opt}} = \underset{\text{Tr}(\boldsymbol{F}^H\boldsymbol{F})\leq P}{\text{argmin}} \; \mathbb{E}\left[\left\|\boldsymbol{H}\boldsymbol{F}\boldsymbol{s} + \sqrt{\gamma}\boldsymbol{w} - \boldsymbol{s}\right\|^2 \middle| \hat{\boldsymbol{H}}\right] \tag{5.54}$$

which minimizes the above MSE objective function.

In the literature, the random components of $\boldsymbol{E}$ are assumed to have the same error variance σ_e^2. However, this could be far from a realistic scenario, where the channel estimation errors could be different for each user. For example, this can be related to the user estimation technique defined by the UE vendor based on reference signals. This can also be related to the number of bits used for quantization per user when channel quantization is performed. To cope with this variety of channel estimation errors, we propose a new feedback quantity referred to as CSIAI, which can track each user's error variance. Therefore, the components of each ith row of $\boldsymbol{E}$ have the error variance σ_i^2. The statistics of $\boldsymbol{E}$ can be estimated at the BS considering the CSIAI reporting to know the different error variances associated with users. Depending on the estimated channel matrix $\hat{\boldsymbol{H}}$, we need to compute the optimal precoder that minimizes the above expectation (5.54) taken over the distributions of $\boldsymbol{w}$ and $\boldsymbol{E}$.

5.4.2 MMSE based precoding

In this section, we consider the optimization problem proposed in (5.54), where we aim to derive the optimal precoding matrix which minimizes the MSE objective function given the estimated channel matrix $\hat{\boldsymbol{H}}$ at the BS. In order to compute the precoding matrix $\boldsymbol{F}_{\text{opt}}$, knowledge of the error variance σ_i^2 for each ith user is available at the BS thanks to the CSIAI reporting.

Given that the elements of the channel $\boldsymbol{H}$ have unit variance, the matrix $\hat{\boldsymbol{H}}$ has the same distribution as $\boldsymbol{H}$ with a reduced variance equal to $1-\sigma_i^2$ per each ith row. Let $\tilde{\boldsymbol{H}}$ be defined as

$$\tilde{\boldsymbol{H}} = \boldsymbol{D}\hat{\boldsymbol{H}} \tag{5.55}$$

where $\boldsymbol{D}$ is the diagonal matrix defined as

$$\boldsymbol{D} = \begin{bmatrix} \frac{1}{\sqrt{1-\sigma_1^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{1-\sigma_K^2}} \end{bmatrix} \tag{5.56}$$

It is evident to see that $\tilde{\boldsymbol{H}}$ has the same distribution as $\boldsymbol{H}$. Therefore, we introduce the system model of precoding with quantized feedback or channel mismatch, as shown in Figure 5.9. The received signal vector can be written as

$$\begin{aligned} \boldsymbol{y} &= \frac{1}{\sqrt{\hat{\gamma}}}\boldsymbol{D}\boldsymbol{H}\boldsymbol{F}\boldsymbol{s} + \boldsymbol{D}\boldsymbol{w} \\ &= \frac{1}{\sqrt{\hat{\gamma}}}\Big(\boldsymbol{s} + (\boldsymbol{D}\hat{\boldsymbol{H}}\boldsymbol{F} - \boldsymbol{I}_K)\boldsymbol{s} + \boldsymbol{D}\boldsymbol{E}\boldsymbol{F}\boldsymbol{s}\Big) + \boldsymbol{D}\boldsymbol{w} \end{aligned} \tag{5.57}$$

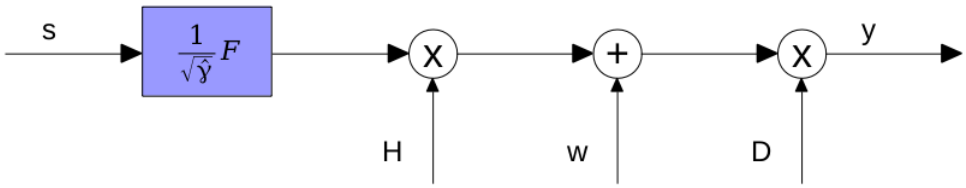


FIGURE 5.9: System model of precoding.

with an arbitrary precoding matrix $\boldsymbol{F}$, and $\hat{\gamma} = \mathrm{Tr}(\boldsymbol{F}^H\boldsymbol{F})/P$. The deviation vector is obtained as

$$\begin{aligned}\boldsymbol{d} &= \sqrt{\hat{\gamma}}\,\boldsymbol{y} - \boldsymbol{s} \\ &= (\boldsymbol{D}\hat{\boldsymbol{H}}\boldsymbol{F} - \boldsymbol{I}_K)\boldsymbol{s} + \boldsymbol{D}\boldsymbol{E}\boldsymbol{F}\boldsymbol{s} + \sqrt{\hat{\gamma}}\boldsymbol{D}\boldsymbol{w}\end{aligned} \tag{5.58}$$

Given the data vector $\boldsymbol{s}$ and the estimated channel matrix $\hat{\boldsymbol{H}}$, the MSE is obtained as a function of $\boldsymbol{F}$ by taking the expectation over $\boldsymbol{w}$ and $\boldsymbol{E}$

$$\begin{aligned}e(\boldsymbol{F}) &= \mathbb{E}_{\boldsymbol{w},\boldsymbol{E}}(\|d\|^2|\hat{\boldsymbol{H}}, \boldsymbol{s}) \\ &= \|(\boldsymbol{D}\hat{\boldsymbol{H}}\boldsymbol{F} - \boldsymbol{I}_K)\boldsymbol{s}\|^2 + \hat{\gamma}\sum_{i=1}^{K}\frac{\sigma_w^2 + \sigma_i^2}{1-\sigma_i^2}\end{aligned} \tag{5.59}$$

Let us denote $\sigma_{eq}^2 = \frac{1}{K}\sum_{i=1}^{K}\frac{\sigma_w^2+\sigma_i^2}{1-\sigma_i^2}$ to simplify equations. A comparison between (5.6) and (5.59) shows that the optimization problems of MSE share a similar form. It follows that the optimal precoding matrix is given by [105]

$$\boldsymbol{F}_{\text{opt}} = (\boldsymbol{D}\hat{\boldsymbol{H}})^H\left(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{\boldsymbol{H}})^H + K\sigma_{eq}^2\boldsymbol{I}_K\right)^{-1} \tag{5.60}$$

Now the optimal perturbation vector can be found as

$$\boldsymbol{v}_{\text{opt}} = \underset{v}{\text{argmin}}\ K\sigma_{eq}^2\tilde{\boldsymbol{s}}^H\left(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{\boldsymbol{H}})^H + K\sigma_{eq}^2\boldsymbol{I}_K\right)^{-1}\tilde{\boldsymbol{s}} \tag{5.61}$$

where $\tilde{\boldsymbol{s}} = \boldsymbol{s} + \boldsymbol{v}$ denotes the perturbed data vector. With Cholesky factorization of $\left(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{H})^H + K\sigma_{eq}^2\boldsymbol{I}_K\right)^{-1}$, (5.61) can be rewritten as

$$\boldsymbol{v}_{\text{opt}} = \underset{v}{\text{argmin}}\ \|\boldsymbol{L}^H(\boldsymbol{s}+\boldsymbol{v})\|^2 \tag{5.62}$$

where $\left(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{H})^H + K\sigma_{eq}^2\boldsymbol{I}_K\right)^{-1} = \boldsymbol{L}\boldsymbol{L}^H$ with $\boldsymbol{L}$ is a lower triangular matrix.

Special Case

We assume that the error matrix $\boldsymbol{E}$ has $K \times M$ independent elements with zero mean and error variance equal to σ_e^2, i.e., $\sigma_i^2 = \sigma_e^2 \; \forall \; i \in \{1, \ldots, K\}$. In this case, $\boldsymbol{D}$ is equal to the identity matrix scaled by some factor, that is

$$\boldsymbol{D} = \frac{1}{\sqrt{1-\sigma_e^2}} \boldsymbol{I} \tag{5.63}$$

Substituting (5.63) in (5.60), the optimal precoding matrix can be written as

$$\boldsymbol{F}_{\text{opt}} = \sqrt{1-\sigma_e^2} \times \hat{\boldsymbol{H}}^H \left(\hat{\boldsymbol{H}} \hat{\boldsymbol{H}}^H + K(\sigma_n^2 + \sigma_e^2) \boldsymbol{I}_K \right)^{-1} \tag{5.64}$$

Then, from the system model, the received signal vector is obtained as

$$\boldsymbol{y} = \boldsymbol{H} \hat{\boldsymbol{H}}^H \left(\hat{\boldsymbol{H}} \hat{\boldsymbol{H}}^H + K(\sigma_w^2 + \sigma_e^2) \boldsymbol{I}_K \right)^{-1} \boldsymbol{s} + \frac{1}{\sqrt{1-\sigma_e^2}} \boldsymbol{w} \tag{5.65}$$

Hence, we can say that our results share a similar form of the precoding matrix that appears in many studies where the error variance is the same for all users as in [106].

5.4.3 Performance analysis

This section provides equivalence relations between the quantized and perfect channel feedback in the MMSE criterion and SER. We analyze the performance of the system by considering the SINR value at each user receiver. Then we show how to scale the quantization level for each user in the MMSE-VP to achieve the full diversity order for VP precoding.

Equivalence relations

In the following, Lemmas 1 and 2 are applicable for both linear precoding and VP precoding. We use the term "MMSE based precoding" to refer to both linear MMSE and non-linear MMSE-VP.

Lemma 1: In a transmission system with imperfect CSI for K users, where each ith user has a channel estimation error of variance σ_i^2 and an additive noise of variance σ_w^2, the MMSE based precoding is equal to that in a transmission system with perfect CSI where the additive noise is of variance $\sigma_{eq}^2 = \frac{1}{K} \sum_{i=1}^{K} \frac{\sigma_w^2 + \sigma_i^2}{1-\sigma_i^2}$.

Proof: In the last section 5.4.2, we have shown that for a given $\hat{\mathbf{H}}$, there exists a corresponding $\tilde{\mathbf{H}}$ with an additive noise power $\sigma_{eq}^2 = \frac{1}{K}\sum_{i=1}^{K}\frac{\sigma_w^2+\sigma_i^2}{1-\sigma_i^2}$ which has the same distribution as $\mathbf{H}$. Based on equation (5.59), the resulting MMSE is averaged over the distribution of $\tilde{\mathbf{H}} = \mathbf{D}\hat{\mathbf{H}}$. So we can say that the MMSE based precoding is equivalent to that in a transmission system with perfect CSI, where the channel matrix is $\tilde{\mathbf{H}}$ and the noise variance is σ_{eq}^2.

Let $e(\sigma_w^2, (\sigma_1^2, \ldots, \sigma_K^2)^T)$ denote the SER of MMSE based precoding with imperfect CSI for K users, each ith user has a channel estimation error of variance σ_i^2. We note that $(\sigma_1^2, \ldots, \sigma_K^2)^T = 0^T$ corresponds to the SER with perfect CSI.

Lemma 2: In a transmission system with imperfect CSI for K users, where each ith user has a channel estimation error of variance σ_i^2 and an additive noise of variance σ_w^2, the SER of MMSE based precoding is equal to that in a transmission system with perfect CSI where the additive noise power is of variance $\sigma_{eq}^2 = \frac{1}{K}\sum_{i=1}^{K}\frac{\sigma_w^2+\sigma_i^2}{1-\sigma_i^2}$, namely

$$e(\sigma_w^2, (\sigma_1^2, \ldots, \sigma_K^2)^T) = e\left(\frac{1}{K}\sum_{i=1}^{K}\frac{\sigma_w^2+\sigma_i^2}{1-\sigma_i^2}, 0^T\right) \tag{5.66}$$

Proof: Since the deviation vector determines the number of incorrectly detected symbols, we can claim from (5.58) that both cases of perfect and imperfect CSI achieve the same SER under different noise variances.

Feedback load

It is proved in [34] that VP precoding with perfect channel feedback achieves full diversity of order M, namely

$$\lim_{\rho\to\infty}\frac{-\log(e(\sigma_w^2, 0^T))}{\log(\rho)} = \lim_{\sigma_w^2\to 0}\frac{\log(e(\sigma_w^2, 0^T))}{\log(\sigma_w^2)} = M \tag{5.67}$$

The following Lemma shows how much feedback load is required to obtain the same precoding diversity order under specific quantized channel feedback.

Lemma 3: To achieve the same diversity order as perfect channel feedback, MMSE-VP precoding with quantized channel feedback has to increase each user's feedback load by at least $3.32 \times M$ bits for every 10 dB increase in SNR.

Proof: see Appendix D.

SINR performance analysis

Assuming that the BS designs its transmitted signal using the proposed precoding technique, we derive a general expression for the SINR of each user and closed form expressions for the SINR value at high SNR. Let the ith row of $\boldsymbol{H}$ be given by $\boldsymbol{h}_i = \hat{\boldsymbol{h}}_i + \boldsymbol{e}_i$, where $\hat{\boldsymbol{h}}_i$ and $\boldsymbol{e}_i$ denote the ith rows of the matrices $\hat{\boldsymbol{H}}$ and $\boldsymbol{E}$, respectively. The received signal of the ith user can be written as

$$\boldsymbol{y}_i = \frac{1}{\sqrt{\hat{\gamma}}}(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\boldsymbol{f}_i\boldsymbol{s}_i + \frac{1}{\sqrt{\hat{\gamma}}}(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\sum_{j\neq i}\boldsymbol{f}_i\boldsymbol{s}_j + \boldsymbol{w}_i \tag{5.68}$$

where $\boldsymbol{f}_i$ denotes the ith column vector of $\boldsymbol{F}_{\text{opt}}$. From (5.68), we can notice that the signal and interference components, where $\boldsymbol{s}_i$ is the symbol intended for user i and the interfering signal is given by $(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\sum_{j\neq i}\boldsymbol{f}_i\boldsymbol{s}_j$. We can first compute the ith user SINR conditioned on $\hat{\boldsymbol{H}}$ and $\boldsymbol{e}_i$. In this case we let

$$\text{SINR}_i(\boldsymbol{e}_i) \equiv \frac{\mathbb{E}\left[\left|\frac{1}{\sqrt{\hat{\gamma}}}(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\boldsymbol{f}_i\boldsymbol{s}_i\right|^2 \middle| \hat{\boldsymbol{H}}, \boldsymbol{e}_i\right]}{\mathbb{E}\left[\left|\frac{1}{\sqrt{\hat{\gamma}}}(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\sum_{j\neq i}\boldsymbol{f}_i\boldsymbol{s}_i + \boldsymbol{w}_i\right|^2 \middle| \hat{\boldsymbol{H}}, \boldsymbol{e}_i\right]} \tag{5.69}$$

denote the ith user SINR value given the estimated channel matrix $\hat{\boldsymbol{H}}$ and the ith user channel estimation error $\boldsymbol{e}_i$. The expectation is taken over the distribution of the noise $\boldsymbol{w}_i$. From the above definition, a general expression for the ith user SINR value is given by

$$\text{SINR}_i = \mathbb{E}\left[\text{SINR}_i(\boldsymbol{e}_i)\middle|\hat{\boldsymbol{H}}\right] \tag{5.70}$$

where the expectation is taken over the distribution of the ith user channel estimation error $\boldsymbol{e}_i$ conditioned on the estimated channel matrix $\hat{\boldsymbol{H}}$. Given $\hat{\boldsymbol{H}}$ and $\boldsymbol{e}_i$, the expression in (5.69) can be easily computed and given by

$$\begin{aligned}\text{SINR}_i(\boldsymbol{e}_i) &= \frac{\frac{1}{\hat{\gamma}}|(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\boldsymbol{f}_i|^2}{\sigma_n^2 + \frac{1}{\hat{\gamma}}|\sum_{j\neq i}(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\boldsymbol{f}_i|^2} \\ &= \frac{\rho|(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\boldsymbol{f}_i|^2}{\text{Tr}(\boldsymbol{F}_{\text{opt}}^H\boldsymbol{F}_{\text{opt}}) + \rho\sum_{j\neq i}|(\hat{\boldsymbol{h}}_i + \boldsymbol{e}_i)\boldsymbol{f}_i|^2}\end{aligned} \tag{5.71}$$

At high enough SNR, the MMSE's performance converges to the ZF. This happens since the regularization coefficient in the MMSE precoding matrix is proportional to the inverse of SNR. Thus the MMSE precoder converges to the ZF precoder at asymptotically high SNR levels. Hence, at high SNR, the precoding matrix $\boldsymbol{F}_{\text{opt}}$ converges to the ZF solution with $\boldsymbol{f}_i$ given by the ith column vector of the pseudo-inverse of the channel matrix $\boldsymbol{H}$ noting that $\hat{\boldsymbol{h}}_i = \boldsymbol{h}_i$ and $\boldsymbol{e}_i = 0$. From (5.71), when the BS has perfect and full knowledge of the CSI, we can see that each user's average SINR grows linearly with the SNR. Hence the average SER goes to zero, and the achievable capacity of each user is unbounded.

In the second case when there are channel estimation errors and σ_i^2 is assumed to have a value that is greater than zero

$$\begin{aligned}\lim_{\rho\to\infty} \mathrm{SINR}_i(\boldsymbol{e}_i) &= \frac{\rho|(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\boldsymbol{f}_i|^2}{\mathrm{Tr}(\hat{\boldsymbol{F}}_{\mathrm{opt}}^H\hat{\boldsymbol{F}}_{\mathrm{opt}})+\rho\sum_{j\neq i}|(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\boldsymbol{f}_j|^2} \\ &= \frac{|(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\boldsymbol{f}_i^{\infty}|^2}{\sum_{j\neq i}|(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\boldsymbol{f}_j^{\infty}|^2}\end{aligned} \tag{5.72}$$

where $\boldsymbol{f}_i^{\infty}$ is the ith column vector of

$$\boldsymbol{F}_{\mathrm{opt}}^{\infty} = (\boldsymbol{D}\hat{\boldsymbol{H}})^H\left(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{\boldsymbol{H}})^H+\sum_{i=1}^{K}\frac{\sigma_i^2}{1-\sigma_i^2}\boldsymbol{I}\right)^{-1}$$

The SINR of the ith user is now given by

$$\mathrm{SINR}_i^{\infty} = \mathbb{E}\left[\left(\frac{|(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\boldsymbol{f}_i^{\infty}|^2}{\sum_{j\neq i}|(\hat{\boldsymbol{h}}_i+\boldsymbol{e}_i)\boldsymbol{f}_j^{\infty}|^2}\right)\middle|\hat{\boldsymbol{H}}\right] \tag{5.73}$$

which is obviously a function of the estimated channel matrix $\hat{\boldsymbol{H}}$, the channel estimation error variance σ_i^2 for each ith user, and is independent of the SNR value.

5.4.4 Simulation results

We consider an MU-MISO BC channel composed of a BS equipped with $M = 12$ transmit antennas serving $K = 12$ single antenna users at the same time. Performance is evaluated in terms of averaged SER overall users versus the SNR. The modulation scheme is 16-QAM. We average the performance through Monte Carlo simulations. We assume that the the channel estimation error $\boldsymbol{E}$ has i.i.d zero-mean complex Gaussian random variables with an error variance σ_i^2 per each ith row. Without loss of generality, we assume there are two different error variances $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$. Users are allocated into two groups, each ith group of six users has $\sigma_{e_i}^2$. We note that the proposed scheme can be more general and applied to any MU-MIMO configuration.

Table 5.1 presents some simulation results of the SER performance of the linear MMSE precoding under imperfect CSI and its equivalent perfect CSI whose SER appears within parentheses. It shows relative values between the two sets, which validate the equivalence relation given by (5.66). Table 5.2 also shows similar results for the non-linear MMSE-VP precoding which is in concordance with lemma 2.

TABLE 5.1: SER for different SNRs and settings with linear MMSE.

Setting\SNR	**18 dB**	**27 dB**	**36 dB**
$\sigma_{e_1}^2 = 2^{-4}$	5.007e-1	4.464e-1	4.379e-1
$\sigma_{e_2}^2 = 2^{-6}$	(5.005e-1)	(4.535e-1)	(4.470e-1)
$\sigma_{e_1}^2 = 2^{-6}$	3.674e-1	2.392-1	2.174e-1
$\sigma_{e_2}^2 = 2^{-8}$	(3.663e-1)	(2.417e-1)	(2.192e-1)
$\sigma_{e_1}^2 = 2^{-8}$	3.090e-1	1.224e-1	8.516e-2
$\sigma_{e_2}^2 = 2^{-10}$	(3.093e-1)	(1.223e-1)	(8.157e-2)
$\sigma_{e_1}^2 = 2^{-10}$	2.919e-1	7.952e-2	3.110e-2
$\sigma_{e_2}^2 = 2^{-12}$	(2.918e-1)	(7.922e-2)	(3.023e-2)

TABLE 5.2: SER for different SNRs and settings with MMSE-VP.

Setting\SNR	**14 dB**	**21 dB**
$\sigma_{e_1}^2 = 2^{-4}$	3.37e-1	2.00e-1
$\sigma_{e_2}^2 = 2^{-6}$	(3.42e-1)	(2.07e-1)
$\sigma_{e_1}^2 = 2^{-6}$	2.12e-1	3.47e-2
$\sigma_{e_2}^2 = 2^{-8}$	(2.09e-1)	(3.29e-2)
$\sigma_{e_1}^2 = 2^{-8}$	1.70e-1	4.36e-3
$\sigma_{e_2}^2 = 2^{-10}$	(1.69e-1)	(4.17e-3)

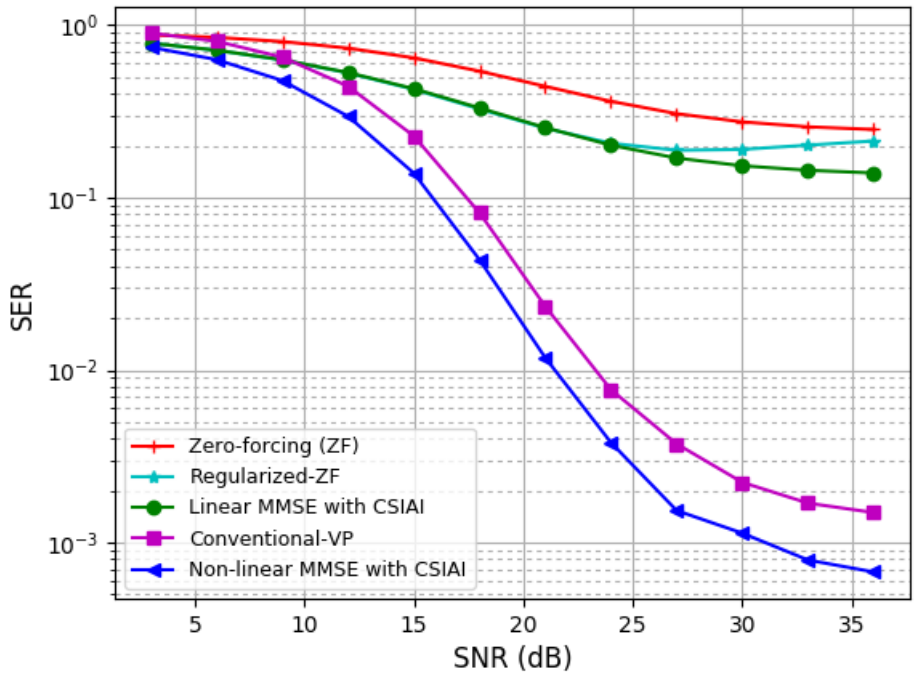


FIGURE 5.10: Performance of the new precoder with CSIAI compared to other existing precoding techniques for $M = 12$, $K = 12$, $\sigma^2_{e_1} = 2^{-7}$ and $\sigma^2_{e_2} = 2^{-9}$.

In Figure 5.10, we plot the average SER of the proposed precoding technique as a function of the SNR. For comparison, we plot the linear ZF's performance, the regularized-ZF, and the non-linear conventional-VP, with imperfect CSI on the exact figure. As we can see from Figure 5.10, the linear MMSE precoding technique exploiting the different values of channel estimation error variances outperforms the ZF and regularized-ZF. The proposed non-linear MMSE also performs better than the conventional-VP, which does not cope with CSI errors. We can see the ceiling effect where the average SER of all precoding techniques flattens for high SNR and does not improve by increasing the SNR. However, the proposed precoding technique can improve the ceiling and decrease the error floor level. We can say that the suggested MMSE precoding exploits well the channel estimation error statistics as long as CSIAI is reported. So the system performance is significantly enhanced.

Figure 5.11 shows the SER performance averaged over all MMSE-VP users under quantized and perfect channel feedback. The dashed line represents the perfect CSI where we have $\sigma^2_{e_1} = \sigma^2_{e_2} = 0$, whereas the other complete lines correspond to the imperfect CSI. In Figure 5.11, the channel estimation error variances are fixed and independent of SNR. It is well-observed that the proposed precoding matrix considering the different quantization error variances performs better than the one using the same variance. The error floor level due to imperfect CSI is decreased.

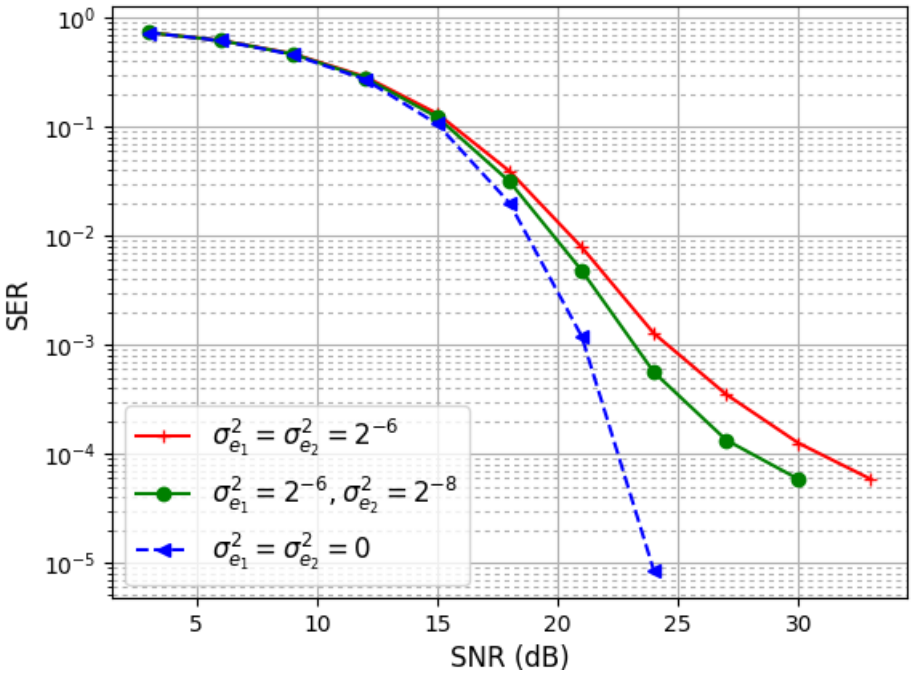


FIGURE 5.11: SER performance with fixed feedback load.

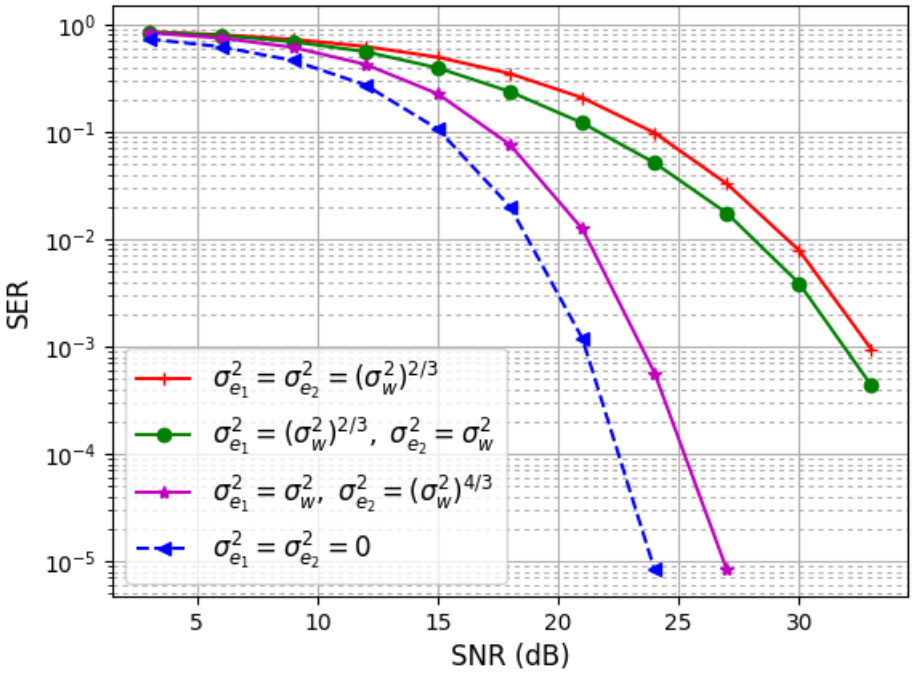


FIGURE 5.12: SER performance with varied feedback load.

With a varied feedback load per user such that $\sigma_i^2 = (\sigma_w^2)^{\alpha_i}$, the SER floor disappears as shown in Figure 5.12. When the feedback load is adjusted in such a way that $\alpha_i \geq 1 \; \forall \; i$, the SER is shown to decrease at the same rate as perfect channel feedback for high SNR region. On the other hand, the SER performance can not achieve the full diversity order when there exists at least one ith user with $\alpha_i < 1$, which confirms our theoretical results presented in section 5.4.3.

5.5 Summary

In this chapter, we investigate the downlink precoding in MU-MIMO BC systems. At first, we propose a combined VP precoding that takes into account different MCSs. The perturbation vector search is combined for all users as in conventional VP. The performance achieved by the proposed algorithm is optimal compared to existing solutions in the literature. We also propose the combined MMSE VP, which achieves better performance by minimizing the MSE criterion.

Moreover, we suggest an ordering of users according to their modulation sizes. Indeed, starting with the highest order in the search tree has the advantage of reducing complexity. Secondly, we develop the block VP algorithm in which the perturbation is applied separately for each block. We analytically derived the diversity order for the MIMO Gaussian BC. The block VP achieves the desired diversity order by fixing the block size. This would provide more flexibility to the BS to update its precoder while achieving the required diversity order. Finally, we investigate the downlink precoding when the CSI is erroneous. Thus, we introduce a new CSIAI reporting for massive MU-MIMO systems. The CSIAI is computed at the UE side and transmitted to the BS via uplink channels. It can be used to select the appropriate timing to trigger full CSI reporting, including PMI information. Based on this feedback quantity, we design an optimized precoding matrix under the MMSE criterion, which considers channel estimation errors. The proposed precoding technique outperforms previously existing techniques such as the ZF and the regularized-ZF. It was shown that an improvement in the average SER performance is achieved. The ceiling is improved, and the error floor level is decreased. We also model a channel vector quantization following the rate-distortion theory. We establish an equivalent relation between the SER of VP precoding with specific quantized and perfect channel feedback. Based on the equivalent relation, we show that the feedback load per user should scale at $M \log_2(\rho)$ to achieve full diversity order.

Chapter 6

Conclusions and Perspectives

This thesis is dedicated to analyzing, designing, and evaluating MIMO decoding techniques in the uplink reception and MIMO precoding approaches in the downlink transmission. Different MU-MIMO configurations are considered in the single-cell as well as the multi-cell environments.

We first address the problem of designing low-complexity decoding algorithms in the uplink MIMO reception. We take profit from recent advances in deep learning (DL) to propose the neural network (NN) assisted sphere decoder (SD), which significantly reduces the processing time for decoding compared to the original SD. Afterwards, we introduce the block recursive MIMO decoding, where the MIMO system is divided into small sub-blocks. This has the advantage to reduce the complexity of decoding, which increases exponentially with the size. Simulations results show the complexity reduction achieved by the proposed algorithms while preserving almost ML performance.

Second, MIMO decoding in the uplink reception for cloud-RAN (C-RAN) is investigated. Due to the prohibitive complexity of computations, the most efficient uplink C-RAN schemes are challenging to be implemented in practical systems. Therefore, we resort to recent advances in deep neural networks (DNNs) and propose QDNet architecture, representing a new and low complex method for uplink C-RAN subject to some quantization rules. Joint optimization of the quantization process done at the BSs and the decoding process completed at the central processor is performed. Simulation results show that QDNet achieves good performance for detection while requiring low complexity.

Third, we focus on the design of novel transmission schemes in the downlink MU-MIMO scenario. At first, we study the vector perturbation (VP) precoding, and we come up with the combined VP algorithm to serve multiple users applying different modulation coding schemes (MCSs). Secondly, we introduce the block VP precoding, which has the advantage of choosing the desired diversity order by fixing the

block size. Finally, we study the downlink precoding in MU-MIMO when several users are present in the network with different CSI accuracy, making the precoder more sensitive to CSI imperfections. So we propose a new quantity referred to as CSI accuracy indicator (CSIAI) reported by the UE. Accordingly, we design the downlink precoding based on CSIAI reporting. Simulation results show an improvement in the average BER performance.

This thesis's contributions can be implemented and developed to satisfy some objective functions subject to environmental constraints. In the following, we give insight into possible research directions.

In massive MU-MIMO systems, the number of BS antennas is enormous compared to the number of scheduled users. Non-linear precoding in these systems requires high computational complexity due to the high number of antennas at the transmitter side. This is a critical issue to be addressed. Indeed, someone should look forward to merging different types of precoding in such a way that further performance enhancements are achieved at reasonable computational complexity.

Nevertheless, when the number of users is high and much larger than the number of BS antennas, user scheduling in the downlink precoding must be performed to satisfy all network users. Thus, how to perform user scheduling is an exciting topic in the downlink MU-MIMO. New signalling between UEs and the BS could help make smart decisions and achieve users' fairness.

We have designed the downlink precoding in the single-cell environment taking into account the imperfect CSI at the BS. Also, it is essential to investigate the impact of CSI errors in the multi-cell environment. We should note that the CSI is available at different stages for the users, inside the cell or at the cell edge. How to explore the coordinated multi-point (CoMP) transmission in such scenarios could be the solution to overcome this problem of CSI signalling.

At last, it is interesting to investigate possible avenues of machine learning for communications. Applications of machine learning will be present in the next generation to restore some 5G architecture functionalities. Someone should think of how to profit from recent advances in DL and NNs to design reliable transmission schemes for beamforming, precoding and decoding in MU-MIMO systems.

Chapter 7

Shortened French Version

7.1 Introduction

La technologie à entrées multiples et sorties multiples (MIMO) est considérée comme l'une des solutions les plus prometteuses et les plus efficaces de la 5G NR. Elle est développée pour répondre à la demande d'efficacité spectrale [9]. Les technologies MIMO exploitent les dimensions spatiales et temporelles pour coder et multiplexer davantage de symboles de données en utilisant une multiplicité d'antennes d'émission et de réception, sur une pluralité de tranches de temps. Ainsi, la capacité et la fiabilité des systèmes de communication MIMO peuvent être améliorées.

Dans ces systèmes, la détection multi-utilisateur (MU)-MIMO dans la réception de la liaison montante et le précodage dans la transmission de la liaison descendante permettent de séparer les flux de données utilisateur et de pré-annuler les interférences. Cependant, les performances du système se détériorent dans des conditions réalistes telles que la complexité raisonnable des processus de décodage et de précodage, la connaissance erronée des canaux et l'interférence des cellules adjacentes. Cette thèse se concentre sur les scénarios mentionnés ci-dessus pour la réception et la transmission dans les systèmes MU-MIMO.

7.2 Liaisons Montantes et Liaisons Descendantes dans les Systèmes MIMO

7.2.1 Aspects fondamentaux des systèmes MIMO

Les systèmes de communication MIMO traditionnels sont généralement appelés mono-utilisateurs MIMO (SU-MIMO) ou également point à point MIMO. Le point d'accès ou la station de base (BS), dans ce cas, ne communique qu'avec un seul terminal mobile (utilisateur). Le point d'accès et l'utilisateur sont équipés de plusieurs antennes, comme le montre la Figure 7.1. Les antennes d'émission envoient des signaux $(\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_{N_t})$ aux antennes de réception. Les signaux reçus sont désignés par

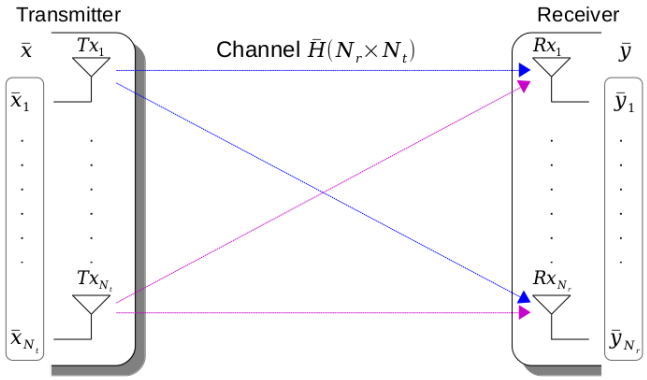


FIGURE 7.1: Modèle de système MIMO.

$(\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_{N_r})$. Nous exprimons le signal réçu à l'antenne $Rx_i \; \forall \, i \in \{1, \ldots, N_r\}$ comme

$$\bar{\boldsymbol{y}}_i = \sum_{j=1}^{N_t} \bar{\boldsymbol{h}}_{ij} \bar{\boldsymbol{x}}_j + \bar{\boldsymbol{w}}_i \tag{7.1}$$

Le modèle de canal MIMO peut être décrit par le système linéaire suivant

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{H}} \bar{\boldsymbol{x}} + \bar{\boldsymbol{w}} \tag{7.2}$$

où $\bar{\boldsymbol{y}}$ est le signal réçu, $\bar{\boldsymbol{H}}$ est le canal san fil dont les entrées sont supposées avoir une distribution de Rayleigh, $\bar{\boldsymbol{x}}$ est le vecteur des symboles transmis, et $\bar{\boldsymbol{w}}$ est le bruit Gaussian blan additif (AWGN), de moyenne nulle et variance $2\sigma_{\bar{w}}^2$. Tout d'abord, nous devons mentionner que tout au long de cette thèse, nous évitons de manipuler des variables à valeurs complexes, et convertissons (7.2) en sa représentation équivalente à valeurs réelles en utilisant la convention suivante

$$\boldsymbol{y} = \boldsymbol{H} \boldsymbol{x} + \boldsymbol{w} \tag{7.3}$$

où

$$\begin{aligned} \boldsymbol{y} = \begin{bmatrix} \Re(\bar{\boldsymbol{y}}) \\ \Im(\bar{\boldsymbol{y}}) \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} \Re(\bar{\boldsymbol{x}}) \\ \Im(\bar{\boldsymbol{x}}) \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} \Re(\bar{\boldsymbol{w}}) \\ \Im(\bar{\boldsymbol{w}}) \end{bmatrix}, \\ \boldsymbol{H} = \begin{bmatrix} \Re(\bar{\boldsymbol{H}}) & -\Im(\bar{\boldsymbol{H}}) \\ \Im(\bar{\boldsymbol{H}}) & \Re(\bar{\boldsymbol{H}}) \end{bmatrix} \end{aligned} \tag{7.4}$$

Contrairement au cas mono-utilisateur, la BS dans les systèmes MU-MIMO peut communiquer avec plusieurs terminaux mobiles. Nous avons la transmission de liaison montante où les multiples utilisateurs transmettent simultanément à la BS. Nous avons également la transmission descendante dans laquelle la BS transmet à de nombreux utilisateurs indépendants.

7.2.2 Dualité liaison montante - liaison Descendante

Plusieurs algorithmes de détection MIMO ont été proposés au fil des ans et ont fait l'objet d'une large couverture dans la littérature. D'une part, nous classons le décodage sous-optimal, qui comprend des techniques linéaires et non linéaires tels que les algorithmes ZF, MMSE et DFE. D'autre part, nous classons le décodage optimal, qui permet de détecter le maximum de vraisemblance (ML). Il comprend des techniques basées sur la représentation en treillis et des algorithmes séquentiels. Nous reconnaissons le décodeur par sphères (SD) qui utilise la recherche en profondeur (DFS), et l'algorithme de décodage en pile utilisant la recherche en largeur (BFS).

La capacité totale d'un canal de diffusion MU-MIMO est obtenue grâce à la technique du codage DPC. Cependant, la méthode DPC est très complexe. C'est pourquoi de nombreuses alternatives de précodage sont proposées, offrant une complexité raisonnable. Ces techniques de précodage peuvent être regroupées en deux catégories, selon qu'elles sont linéaires ou non linéaires. Les techniques de précodage linéaire exemplaires comprennent le ZF et le ZF régularisé (RZF). Avec le précodage ZF, le vecteur transmis est filtré en utilisant le pseudo-inverse de la matrice de canal, ce qui nécessite une puissance de transmission élevée. Des schémas de précodage non-linéaire ont été proposés dans la littérature pour améliorer les performances du précodage linéaire. Le précodage Tomlinson-Harashima (THP) et la perturbation vectorielle (VP) sont deux schémas non-linéaires bien connus.

Nous remarquons une dualité entre le traitement de la liaison montante et de la liaison descendante à la BS après réception ou avant transmission via plusieurs antennes. Les deux schémas de communication utilisent les informations de canal afin d'en supprimer l'effet. Les opérations de décodage sont similaires à celles du précodage, et la plupart des techniques exécutent un canal inverse pour la détection ou le précodage. Nous distinguons principalement les traitements linéaires et non-linéaires. On peut noter, par exemple, le ZF et le MMSE à la fois dans les opérations de décodage et de précodage. On retrouve également le THP dans le précodage MIMO, équivalant au décodage DFE, et le codeur sphérique (SE) qui est la même version du SD, exécuté du côté émetteur.

7.3 Décodage MIMO dans la Liaison Montante

7.3.1 Comptage des points du réseau dans la sphère

Parmi les nombreux problèmes liés aux réseaux de poins, beaucoup restent ouverts. Dans ce travail, nous nous concentrons sur le problème du comptage des points

du réseau dans la sphère. Désignons par $\mathcal{B}_r$ la boule Euclidienne de rayon r à dimension n centrée à l'origine. Le nombre de points de réseau à l'intérieur de la sphère dimensionnelle est proportionnel à son volume. Cependant, on ne connaît pas la densité des points de réseau à l'intérieur d'une sphère donnée. Le problème est également lié à la théorie de la complexité, en particulier au problème vectoriel le plus proche (CVP), et à la variante de Fincke et Pohst utilisée pour la détection MIMO [26]. L'algorithme SD est bien connu pour effectuer cette détection [24]. Ce type d'algorithmes nécessite une mise à l'échelle de grande complexité de manière exponentielle dans la dimension du réseau. Notre travail aborde ce problème en tirant parti des progrès récents des réseaux neuronaux profonds (DNNs) pour réduire la complexité de calcul. À cette fin, nous formons un DNN entièrement connecté pour prédire le nombre de points tombant dans une sphère donnée. Les données d'apprentissage sont obtenues par des implémentations de l'algorithme SD, donnant le nombre correct de points tombant dans la sphère. Celui-ci est, à son tour, obtenu à partir du rayon r, et de la matrice triangulaire supérieure $\boldsymbol{R}$ qui est dérivée de la décomposition "QR" de la matrice génératrice de réseau de points. Compte tenu de ces aspects, le DNN est formé en utilisant un ensemble de paires de vecteurs d'entrée-sortie $(\boldsymbol{x}, N)$ où $\boldsymbol{x}$ est le vecteur d'entrée, et N est le nombre réel de points du réseau à l'intérieur de la sphère. Nous définissons $\boldsymbol{x}$ sous la forme

$$\boldsymbol{x} = \frac{1}{r}\Big[\boldsymbol{R}_{11}, \ldots, \boldsymbol{R}_{nn}\Big]^T \tag{7.5}$$

où $\boldsymbol{R}_{ij}$ sont les coefficients de la matrice $\boldsymbol{R}$ $(1 \leq i \leq j \leq n)$. Le DNN prédit le nombre de points du réseau à sa couche de sortie comme

$$\hat{N}_p = f(\boldsymbol{x}; \boldsymbol{\theta}) \tag{7.6}$$

où $\boldsymbol{\theta}$ est le vecteur des paramètres DNN. Nous utilisons l'unité linéaire rectifiée bien connue (ReLU) comme fonction d'activation pour chaque couche du réseau neuronal (NN). Pour optimiser $\boldsymbol{\theta}$, nous utilisons l'erreur en pourcentage absolu moyen (MAPE) comme fonction de perte qui aboutit à la formule suivante

$$\tilde{L}(\theta) = \frac{1}{\#\{S_t\}} \sum_{i \in S_t} \left| \frac{N_p^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})}{N_p^{(i)}} \right| \tag{7.7}$$

où $N_p^{(i)}$ est la sortie souhaitée lorsque $\boldsymbol{x}^{(i)}$ est utilisé comme entrée. Comme méthode d'optimisation pour ajuster les paramètres, nous utilisons Adadelta [42]. Nous évaluons les performances du modèle NN proposé à travers plusieurs expériences de simulation. Nous considérons différentes tailles de systèmes, c.-à-d des réseaux de dimensions n, où n varie de 5 à 10. Le MAPE est couramment utilisé comme fonction de perte pour les problèmes de régression et l'évaluation du modèle grâce à son interprétation intuitive en erreur relative. La division par la valeur réelle $N_p^{(t)}$ au lieu de la valeur prédite $\hat{N}_p^{(t)}$ conduit à un résultat différent. Ce problème a été

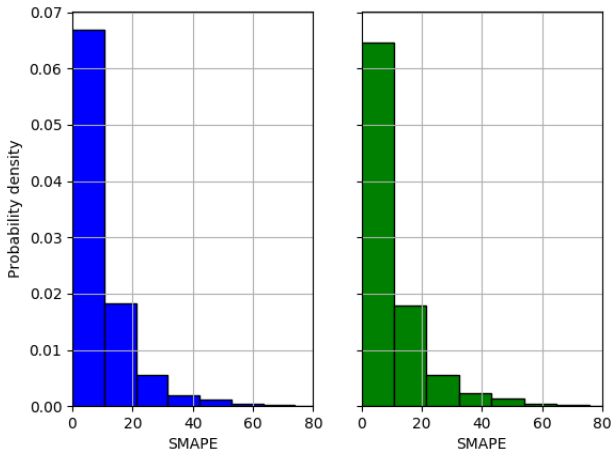


FIGURE 7.2: Tracés de l'histogramme SMAPE du modèle DNN pour la dimension $n = 10$.

soulevé dans [43] et [44]. Le MAPE symétrique (SMAPE) a été proposé pour fournir la symétrie et la robustesse contre les valeurs aberrantes en divisant la perte absolue par la moyenne arithmétique du $N_p^{(t)}$ réel et du $\hat{N}_p^{(t)}$ prédit

$$SMAPE = \frac{100\%}{T} \sum_{t=1}^{T} \frac{|N_p^{(t)} - \hat{N}_p^{(t)}|}{|N_p^{(t)} + \hat{N}_p^{(t)}|} \tag{7.8}$$

La Figure 7.2 montre l'histogramme d'erreur du DNN. Nous traçons la figure de gauche pour l'erreur d'ensemble d'apprentissage et celle de droite pour l'erreur d'ensemble de tests. Nous présentons des résultats pour des réseaux de points de dimensions 10, mais en général, nous observons des résultats similaires pour toutes les dimensions entre 5 et 10. Nous pouvons voir un pourcentage élevé de points dont le SMAPE est inférieur à 10%, ce qui indique que notre modèle s'adapte très bien au nombre de points. Dans le tableau 7.1, nous présentons le MAPE et le SMAPE pour chaque dimension de réseau de points sur l'ensemble d'apprentissage, et sur l'ensemble de tests dont les données se trouvent entre parenthèses. Nous observons la similitude entre l'erreur d'apprentissage et l'erreur de test, ce qui indique que notre modèle évite à la fois le sous-ajustement et le sur-ajustement grâce à l'utilisation des techniques de régularisation ℓ_1 et ℓ_2.

7.3.2 SD assisté par apprentissage

On peut montrer que, à la fois du pire des cas et du point de vue moyen, le SD requiert une complexité exponentielle [46]. Étant donné que le rayon de la sphère affecte directement la plage de recherche, il s'agit d'un paramètre important pour la conception. Par conséquent, nous profitons de l'approche d'apprentissage dans la

TABLE 7.1: Précision pour des réseaux arbitraires dans $\mathbb{R}^n$.

Dimension n	MAPE %	SMAPE %
6	14.575 (14.700)	7.055 (7.052)
7	15.958 (16.290)	8.700 (8.842)
8	17.021 (17.147)	9.028 (9.036)
9	16.377 (16.724)	8.463 (8.701)
10	17.078 (17.078)	9.247 (9.809)

section 7.3.1, qui prédit le nombre de points de réseau à l'intérieur de la sphère pour réduire la complexité de calcul de l'algorithme SD. Cependant, nous considérerons dans ce temps le point du signal reçu $\boldsymbol{y}$, comme une origine de la sphère. Le vecteur d'entrée du NN se présente maintenant sous la forme de

$$\boldsymbol{x} = \left[\boldsymbol{y}^T,\ \boldsymbol{R}_{11}, \ldots, \boldsymbol{R}_{nn},\ r\right]^T \tag{7.9}$$

Notre objectif principal est d'implémenter l'algorithme SD en utilisant un rayon initial amélioré, conduisant à un petit nombre de points de réseau à l'intérieur de la sphère. Dans ce cas, le nombre de points de réseau est prédit par le modèle NN en fonction du point de signal reçu, de la matrice génératrice et du rayon de la sphère. Nous commençons par prédire le nombre de points de réseau tombant à l'intérieur de la sphère avec un rayon initial égal à celui proposé dans [24]. Ensuite, si ce nombre attendu est grand, nous mettons à jour le rayon en utilisant une recherche dichotomique comme proposé dans [58]. En effet, nous divisons le rayon carré par deux, et nous prédisons à nouveau le nombre de points de réseau avec le nouveau rayon. Nous répétons la même procédure jusqu'à ce que nous atteignions un nombre prédit inférieur ou égal à un seuil donné. Enfin, nous commençons la phase de recherche de l'algorithme SD avec le rayon adapté.

Chaque fois que nous mettons à jour le rayon, des calculs NN sont nécessaires pour prédire le nombre de points du réseau et vérifier s'il est encore grand ou non. Cela conduit à une complexité moyenne supplémentaire liée aux calculs NN. Par conséquent, nous voulons évaluer le nombre moyen de calculs NN avant de démarrer l'algorithme SD. Avec un rayon initial $r_0^2 = 2n\sigma_w^2$, nous analysons théoriquement le nombre de mises à jour de rayon. En divisant successivement le rayon carré L fois jusqu'à atteindre un petit nombre attendu de points de réseau, nous démarrons l'algorithme SD avec le rayon $r_L^2 = r_0^2/2^L$. On arrive à déterminer le nombre d'itérations L en fonction du SNR (ρ exprimé en dB)

$$L = a\rho + b \tag{7.10}$$

où a et b sont les deux paramètres déterminés théoriquement. Sur la base de (7.10),

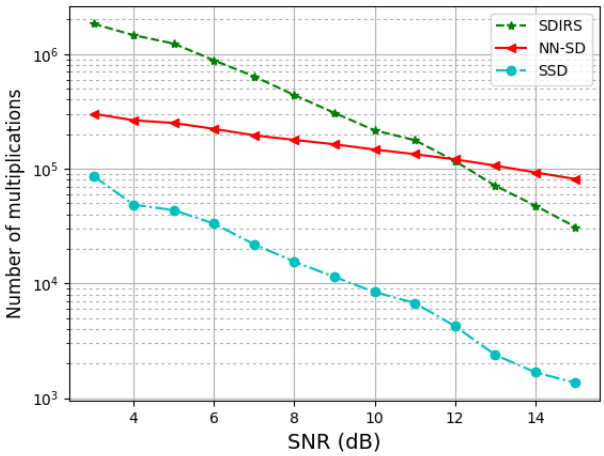


FIGURE 7.3: Nombre moyen de multiplications dans le processus de décodage.

nous proposons un nouvel algorithme, smart SD (SSD), sans calculs NN. Nous commençons la phase de recherche de SD avec un rayon amélioré égal à $r_0^2/2^L$.

Maintenant, nous présentons des simulations informatiques du NN-SD par rapport à l'algorithme d'origine SDIRS qui commence la recherche avec le rayon r_o^2. On considère le canal 8×8 MIMO avec 16-QAM. Dans la Figure 7.3, nous traçons le nombre moyen de multiplications pour mesurer la complexité de calcul en fonction des rapports signal-sur-bruit (SNRs) faibles à modérés. On observe bien que le NN-SD réduit considérablement le nombre d'opérations par rapport à l'algorithme SDIRS. Cette réduction de complexité s'explique par le choix d'un rayon de sphère initial qui permet à un petit nombre de points de réseau de tomber à l'intérieur de la sphère, et ainsi la taille de l'arbre de recherche diminue dans le sens moyen. La Figure 7.4 affiche le nombre moyen de points de réseau tombant à l'intérieur de la sphère de décodage. Nous voyons que cette moyenne dans le NN-SD est presque constante en fonction du SNR, alors qu'elle est plus élevée dans l'algorithme SDIRS pour des SNRs faibles à modérés.

7.3.3 Décodage MIMO récursif par blocs

Dans cette section, nous proposons une nouvelle stratégie de décodage par blocs, qui est une méthode de généralisation du travail décrit dans [67], où le système MIMO est seulement divisé en deux blocs. L'idée est de résoudre les sous-systèmes issus de toute division en plus de deux blocs. Considérons la matrice triangulaire supérieure qui est divisée en k blocs comme le montre la Figure 7.5. Soit $(p_1, \ldots, p_k)$ les tailles de blocs satisfaisant l'égalité $\sum_{j=1}^{k} p_j = 2n$. $\boldsymbol{R}_i \in \mathbb{R}^{p_i \times p_i}$ et $\boldsymbol{B}_i \in \mathbb{R}^{p_{i+1} \times (\sum_{j=1}^{i} p_j)}$ sont les matrices triangulaires supérieures et de rétroaction, respectivement, où $i \in \{1, \ldots, k\}$. En conséquence, les vecteurs de signaux émis et

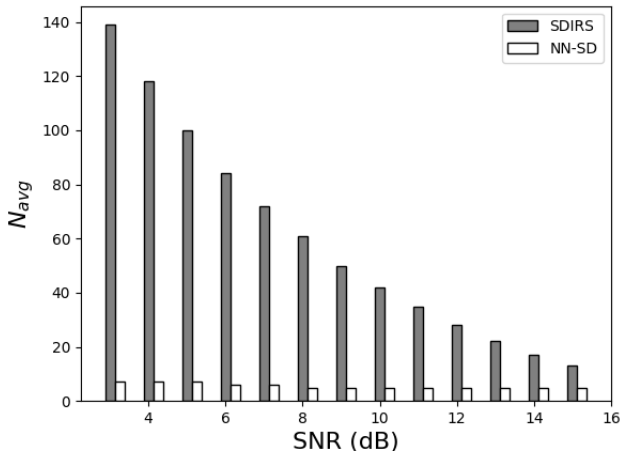


FIGURE 7.4: Nombre moyen de points de réseau (N_{avg}) tombant dans la sphère de recherche.

reçus sont divisés en $(\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(k)})$ et $(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(k)})$, respectivement. Notre méthode de décodage récursif proposée est implémentée en deux étapes principales. Tout d'abord, nous estimons les symboles d'information dans les premiers (k-1) blocs en utilisant le décodage par blocs pour obtenir un essai éventuel des symboles d'information incomplète. Deuxièmement, nous recherchons les données restantes qui minimisent la métrique globale de ML. Pour résumer, le processus de décodage comprend les étapes suivantes:

1. Choisir le nombre de blocs k et leurs tailles correspondantes $(p_1, \ldots, p_k)$;
2. Créer une liste de solutions pour le premier bloc en utilisant un décodage séquentiel. Cette liste est composée de la solution ML qui minimise la distance Euclidienne $\|\boldsymbol{y}^{(1)} - \boldsymbol{R}_1 \boldsymbol{s}^{(1)}\|^2$, et ses voisins;
3. Créer une nouvelle liste de solutions pour chaque candidat de la liste précédente afin de minimiser la métrique Euclidienne $\|\boldsymbol{y}^{(2)} - \boldsymbol{B}_1 s^{(1)} - \boldsymbol{R}_2 \boldsymbol{s}^{(2)}\|^2$;
4. **Répéter la dernière procédure jusqu'à le (k-1)ème bloc**;
5. Trier l'ensemble des candidats par ordre croissant de leur poids;
6. Rechercher le vecteur de symboles de données restant dans le $k$ème bloc en commençant par le haut des candidats classés.

La dernière étape se termine lorsque le poids le plus petit d'un candidat complet examiné est inférieur au poids partiel de celui à examiner. La deuxième, troisième et quatrième étapes nécessitent un ensemble de rayons $(r_1, \ldots, r_{k-1})$ représentant les seuils sur les poids pour créer les listes. Pour cela, ces rayons sont calculés sur la

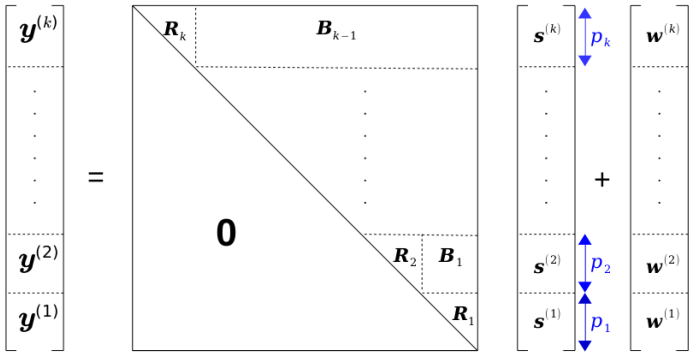


FIGURE 7.5: Division en blocs du système de décodage.

base de la dérivation d'une borne supérieure du taux d'erreur de trame P_{ef}:

$$P_{ef} \leq (\beta_1 + \beta_2)\rho^{-n} + \sum_{i=1}^{k-1} \frac{\Gamma(\frac{p_i}{2}, \frac{r_i^2}{\sigma_w^2})}{\Gamma(\frac{p_i}{2})} \tag{7.11}$$

L'ordre de diversité qui pourrait être atteint par ce schéma de décodage est contrôlé par le deuxième terme étant donné que le premier atteint une diversité complète. Pour garantir un ordre de diversité global d'au moins $d \in \{1, \ldots, n\}$, chaque terme de la somme doit décroître de l'ordre de ρ^{-d}. Cela revient à trouver pour chaque $i$ème bloc le seuil minimum r_i tel que

$$\frac{\Gamma(\frac{p_i}{2}, \frac{r_i^2}{2\sigma_w^2})}{\Gamma(\frac{p_i}{2})} \leq \delta\rho^{-d}, \; i \in \{1, \ldots, k-1\} \tag{7.12}$$

pour une constante positive δ qui contrôle le gain SNR. Dans notre travail, nous donnons le calcul analytique de r_i. En effet, l'inégalité sur r_i est résolue sur la base de l'inversion asymptotique des fonctions Gamma incomplètes [71]. Pour présenter des résultats numériques, nous considérons un canal MIMO 10×10 avec 16-QAM et une division en 2 et 3 blocs. La légende des figures indique les tailles de blocs $(p_1, \ldots, p_k)$. Dans la Figure 7.6, nous représentons le taux d'erreur binaire (BER) en fonction du SNR. Nous pouvons voir qu'en fixant la diversité cible à 10, le décodeur par blocs atteint des performances presque ML pour toutes les divisions considérées. Dans la Figure 7.7, nous traçons le temps de traitement moyen pour le décodage en fonction du SNR. Nous utilisons le même processeur informatique et le même langage de programmation C pour mesurer ce temps. On observe bien que l'algorithme SD a un temps de traitement plus important que le décodeur par blocs récursif.

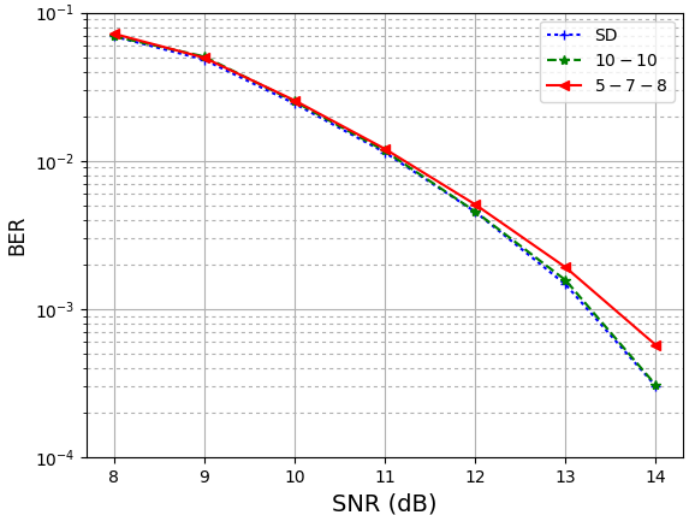


FIGURE 7.6: Performances BER du décodeur de bloc.

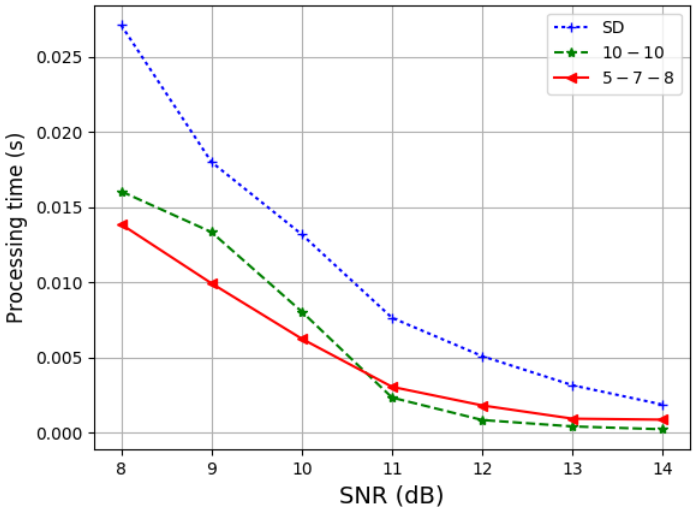


FIGURE 7.7: Temps de traitement moyen dans le décodeur de bloc par rapport à l'algorithme SD.

7.4 Compression Fronthaul dans les systèmes C-RAN assistée par Apprentissage

Dans cette section, le décodage MIMO dans la réception de la liaison montante reste à étudier. Cependant, cette fois, il sera examiné dans le cloud-RAN (C-RAN). Ce dernier est une architecture pour des réseaux cellulaires qui consistent en de nombreuses BSs afin de déplacer la charge de calcul vers le processeur central (CP). Cependant, en raison de la complexité des calculs, certains schémas C-RAN sont difficiles à mettre en œuvre dans des systèmes pratiques. En utilisant les DNNs, nous proposons un schéma pragmatique de C-RAN, appelé QDNet, soumis à certaines règles de quantification.

7.4.1 Modèle de système C-RAN et conception du problème

Nous considérons un modèle C-RAN de liaison montante comme le montre La Figure 7.8, où K utilisateurs distants avec N_t antennes de transmission émettent leurs messages indépendamment vers N BSs distantes. Chaque $n$ème BS est équipée de N_r antennes de réception et reliée au CP par un fronthaul de capacité limitée $C_n \ \forall \ n \in \{1, \dots, N\}$. Ainsi, les signaux reçus aux BSs doivent être comprimés et quantifiés de manière distribuée. Le message transmis par le $i$ème utilisateur est désigné comme $\bar{s}_i$ et appartient à une constellation finie $\bar{\mathcal{S}}$. En pratique, nous supposons que l'ensemble de la constellation $\bar{\mathcal{S}}$ est donné par une modulation QAM. Toutes les constellations sont normalisées à une puissance moyenne unitaire (par exemple, la constellation 4-QAM est représentée par $\{\pm\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}\}$). Le signal reçu $\bar{\boldsymbol{y}}_n$ au $n$ème BS peut être exprimé comme

$$\bar{\boldsymbol{y}}_n = \sum_{i=1}^{K} \bar{\boldsymbol{H}}_{ni} \bar{\boldsymbol{s}}_i + \bar{\boldsymbol{w}}_n \tag{7.13}$$

où $\bar{\boldsymbol{H}}_{in} \in \mathbb{C}^{N_r \times N_t}$ est la matrice de canal entre le $i$ème utilisateur et la $n$ème BS, et $\bar{\boldsymbol{w}}_n \sim \mathcal{CN}(0, 2\sigma_w^2 I_{N_r})$ est le bruit Gaussian reçu par la $n$ème BS. Le principal défi de la détection MIMO est l'utilisation de signaux à valeur complexe, moins courants dans l'apprentissage automatique. Ainsi, en utilisant la convention (7.4), nous convertissons (7.13) en sa représentation équivalente en valeur réelle

$$\boldsymbol{y}_n = \sum_{i=1}^{K} \boldsymbol{H}_{in} \boldsymbol{s}_i + \boldsymbol{w}_n \tag{7.14}$$

Désignons par $\boldsymbol{H} = (\boldsymbol{H}_1^T, \dots, \boldsymbol{H}_N^T)^T$ le canal global dans l'architecture C-RAN où $\boldsymbol{H}_n \ \forall \ n \in \{1, \dots, N\}$ est le canal vers la $n$ème BS en considérant tous les utilisateurs.

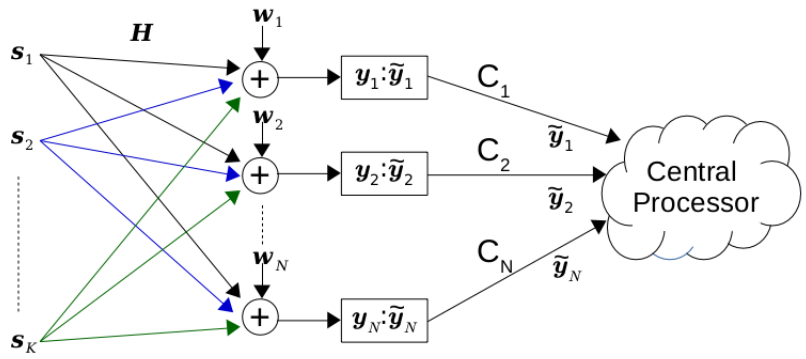


FIGURE 7.8: Un système C-RAN de liaison montante avec un fronthaul de capacité finie.

Ainsi, le signal reçu à toutes les BSs peut être exprimé comme

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{w} \tag{7.15}$$

où $\boldsymbol{s} = (\boldsymbol{s}_1^T, \ldots, \boldsymbol{s}_K^T)^T$ est le message transmis envoyé par tous les utilisateurs du système et $\boldsymbol{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_K^T)^T$ est le bruit Gaussian.

7.4.2 Conception du QDNet

Le schéma que nous proposons s'inspire de l'algorithme itératif de descente de gradient projeté. Pour une observation donnée $\boldsymbol{y}$, la probabilité $p(\boldsymbol{y}|\boldsymbol{s})$ est inversement proportionnelle à la distance $\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2$. En conséquence, un algorithme de descente de gradient projeté basé sur la détection ML peut être exprimé comme

$$\hat{\boldsymbol{s}}_{k+1} = \Pi\left[\hat{\boldsymbol{s}}_k - \frac{\partial\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|^2}{\partial \boldsymbol{s}}|_{\boldsymbol{s}=\boldsymbol{s}_k}\right] = \Pi\left[\hat{\boldsymbol{s}}_k + \delta_k\left(\boldsymbol{H}^T\boldsymbol{y} - \boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k\right)\right] \tag{7.16}$$

où $\boldsymbol{s}_k$ est l'estimation de $\boldsymbol{s}$ dans la $k$ème itération, $\Pi[.]$ est un opérateur de projection non linéaire, et δ_k est un pas. Intuitivement, chaque itération est une combinaison linéaire de $\hat{\boldsymbol{s}}_k$, $\boldsymbol{H}^T\boldsymbol{y}$, et $\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k$ suivie d'une projection non linéaire. Cela indique que les deux principaux ingrédients de l'architecture devraient être $\boldsymbol{H}^T\boldsymbol{y}$ et $\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k$. Notre construction est basée sur l'imitation de cette descente en gradient projetée comme une solution pour l'optimisation du ML. Il est clair de reconnaître le signal $\boldsymbol{H}^T\boldsymbol{y}$ à transmettre au CP au lieu du signal reçu $\boldsymbol{y}$ ayant une grande dimension. Tout d'abord, il est évident que les termes $\boldsymbol{H}^T\boldsymbol{y}$ et $\boldsymbol{H}^T\boldsymbol{H}$ peuvent être réécrits comme

$$\boldsymbol{H}^T\boldsymbol{y} = \sum_{n=1}^{N} \boldsymbol{H}_n^T\boldsymbol{y}_n \quad \& \quad \boldsymbol{H}^T\boldsymbol{H} = \sum_{n=1}^{N} \boldsymbol{H}_n^T\boldsymbol{H}_n \tag{7.17}$$

Chaque BS a ses processus de transformation et de quantification, tandis que le processus de décodage est effectué du côté CP avec un réseau partagé. Avant d'être

quantifié, le signal $\boldsymbol{H}_n^T \boldsymbol{y}_n$ est écrêté à la $n$ème BS en impliquant un opérateur de signe souple linéaire par morceaux $\psi_n(.)$ défini comme

$$\psi_n(x) = -v_n + \rho(x + v_n) - \rho(x - v_n) \tag{7.18}$$

où $\rho(x) = \max\{0, x\}$ et $v_n > 0$ est le paramètre de seuil d'écrêtage à optimiser pendant la phase d'apprentissage.

Désignons par $\tilde{\boldsymbol{r}}_n$ la sortie de la $n$ème BS qui va être transmise au CP. Une fois que tous les signaux sont reçus de toutes les BSs, le CP additionne ces signaux pour obtenir $\tilde{\boldsymbol{r}} = \sum_{n=1}^{N} \tilde{\boldsymbol{r}}_n$. Cela entraînerait la modification des itérations dans (7.16) comme suit

$$\hat{\boldsymbol{s}}_{k+1} = \Pi\Big[\hat{\boldsymbol{s}}_k + \delta_k\Big(\tilde{\boldsymbol{r}} - \boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k\Big)\Big] \tag{7.19}$$

Dans un premier temps, nous enrichissons ces itérations dans (7.19) en changeant la taille du pas δ_k pour chaque $k$ème itération par $\theta_k^{(1)}$ et $\theta_k^{(2)}$ correspondant respectivement à $\tilde{\boldsymbol{r}}$ et $\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k$, soit

$$\hat{\boldsymbol{s}}_{k+1} = \Pi\Big[\hat{\boldsymbol{s}}_k + \theta_k^{(1)}\tilde{\boldsymbol{r}} - \theta_k^{(2)}\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k\Big] \tag{7.20}$$

Dans la deuxième étape, pour imiter l'opérateur de projection non linéaire $\Pi[.]$, un débruiteur non linéaire $\zeta_k(.)$ est appliqué à $\boldsymbol{z}_k = \hat{\boldsymbol{s}}_k + \theta_k^{(1)}\tilde{\boldsymbol{r}} - \theta_k^{(2)}\boldsymbol{H}^T\boldsymbol{H}\hat{\boldsymbol{s}}_k$ pour produire $\hat{\boldsymbol{s}}_{k+1}$. Dans cette perspective, une fonction optimale de débruitage des éléments est donnée par

$$\zeta(\boldsymbol{z}; \sigma_k^2) = \frac{1}{Z}\sum_{\boldsymbol{s}_i \in \mathcal{S}} \boldsymbol{s}_i \exp\left(-\frac{\|\boldsymbol{s}_i - \boldsymbol{z}\|^2}{\sigma_k^2}\right) \tag{7.21}$$

où $Z = \sum_{\boldsymbol{s}_i \in \mathcal{S}} \exp\left(-\frac{\|\boldsymbol{s}_i - \boldsymbol{z}\|^2}{\sigma_k^2}\right)$. Comme le bruit d'écrêtage et de quantification est difficile à caractériser, nous prédisons la variance σ_k^2 en fonction du bruit du canal comme suit

$$\sigma_k^2 = \theta_k^{(4)} \times \left(\sigma_w^2 + \theta_k^{(3)}\right) \tag{7.22}$$

où $\theta_k^{(3)} > 0$ et $\theta_k^{(4)} > 0$ sont les paramètres à optimiser dans la $k$ème itération pendant la phase d'apprentissage. La Figure 7.9 montre un organigramme représentant une seule couche de QDNet qui correspond à la $k$ème itération du processus d'estimation. Le modèle n'a que quatre paramètres par couche: $\theta_k^{(1)}$, $\theta_k^{(2)}$, $\theta_k^{(3)}$ et $\theta_k^{(4)}$. Ces paramètres sont optimisés pendant la phase d'apprentissage qui se fait hors ligne sur des canaux Gaussiens i.i.d échantillonnés de manière aléatoire. Ensuite, les paramètres optimisés du NN sont utilisés pendant toute la phase de communication.

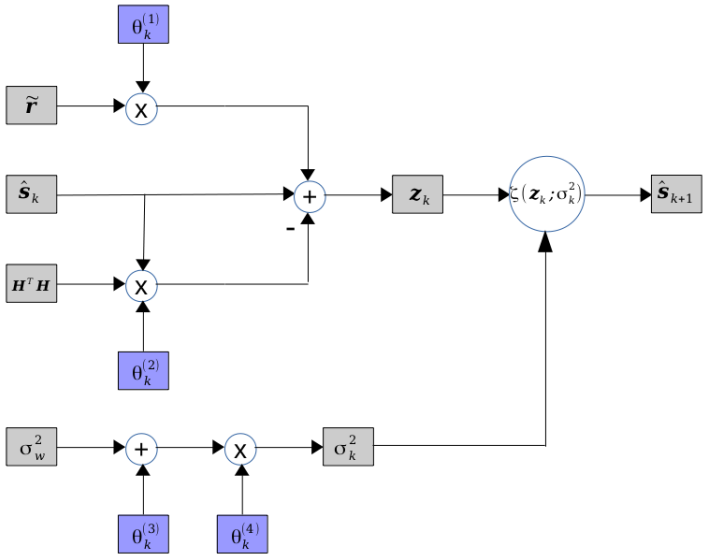


FIGURE 7.9: Organigramme représentant une seule couche de QDNet côté CP.

7.4.3 Résultats d'expérimentation

Les performances du système dépendent du détecteur MIMO considéré. Par conséquent, nous avons testé les performances des schémas de détection suivants:

- **QZF**: Détecteur ZF [86] avec des observations quantifiées de $\boldsymbol{H}^T\boldsymbol{y}$.
- **QSD**: Algorithme SD [86] avec des observations quantifiées de $\boldsymbol{U}^T\boldsymbol{y}$ où $\boldsymbol{U}$ est une matrice semi-unitaire $2N_r \times 2KN_t$. Ce dernier résulte de la décomposition en valeurs singulières (SVD) de la matrice de canal $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$
- **QDNet**: Notre proposition d'algorithme NN.

Le codage de quantification scalaire avec R_q bits est effectué dans tous les algorithmes de détection avant l'envoi des observations au CP par les BSs. Nous considérons $N = 3$ BSs dans le système C-RAN. Le canal MIMO correspondant à chaque BS a une entrée de taille $K = 4$ utilisateurs d'antenne unique et une sortie de $N_r = 6$ antennes de réception. Les Figures 7.10 et 7.11 montrent les performances BER en fonction du SNR des différents schémas pour les modulations 4-QAM et 16-QAM, respectivement. Nous pouvons voir que QDNet fonctionne bien pour la détection tant que les signaux quantifiés des BSs sont utilisés de manière constructive du côté CP dans l'architecture QDNet.

7.5 Précodage MU-MIMO dans la Liaison Descendante

Nous avons étudié le décodage MIMO en réception montante (plusieurs-à-un). Cette section se concentre sur le scénario de liaison descendante (un-à-plusieurs) et étudie

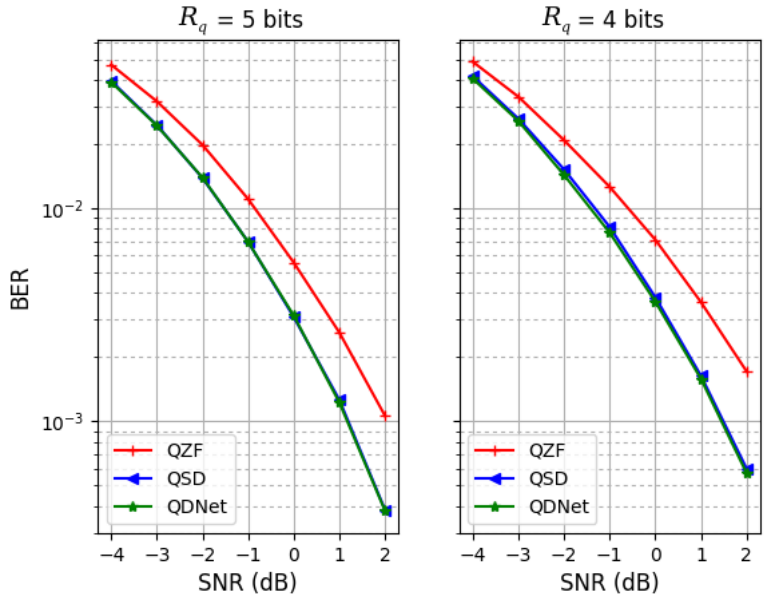


FIGURE 7.10: BER de différents schémas pour 4-QAM.

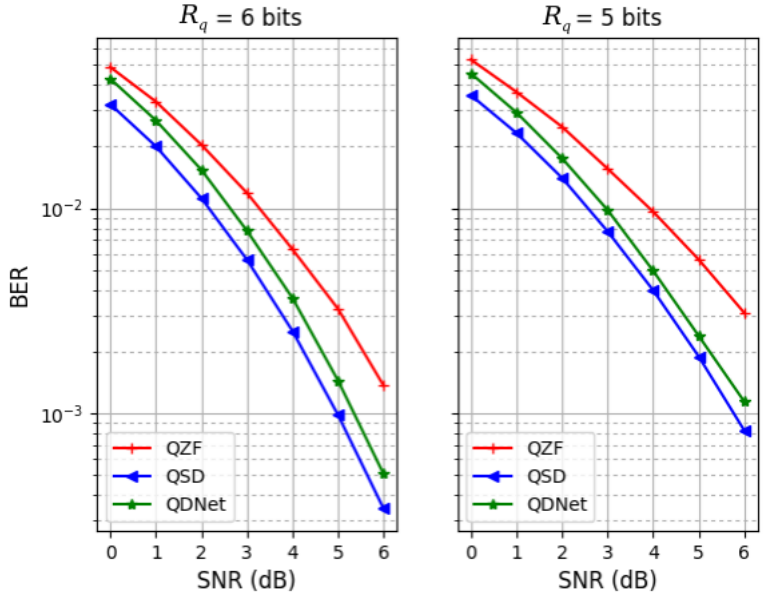


FIGURE 7.11: BER de différents schémas pour 16-QAM.

les techniques de précodage MU-MIMO. Nous développons des schémas de transmission fiables pour les communications multi-utilisateurs dans un environnement à cellule unique, basés sur des techniques de précodage linéaires et non-linéaires.

7.5.1 Précodage MU-MIMO pour la modulation adaptative

Le précodage VP conventionnel n'exploite pas le fait que les utilisateurs utilisent différents schémas de codage de modulation (MCSs), en fonction du rapport signal-sur-interférence-plus-bruit (SINR). Dans [87], la diagonalisation des blocs et le précodage VP sont combinés pour proposer le bloc VP diagonalisé (BD-VP). Ce dernier permet à différents utilisateurs d'appliquer différents schémas de modulation. En outre, les auteurs de [88] proposent le groupement d'utilisateurs VP (UG-VP), qui améliore les performances du précodage BD-VP. Ces solutions existantes ne sont pas optimales puisque le VP est appliqué pour chaque utilisateur ou groupe indépendamment. Pour conserver l'avantage de performance du VP conventionnel (Conv-VP), nous proposons dans nos travaux, le VP combiné (Comb-VP) pour atténuer l'interférence entre les utilisateurs appliquant différents MCSs. En effet, la nouvelle conception du précodage du VP combiné comprend une matrice diagonale $\boldsymbol{T}$ au lieu de la base modulo scalaire τ qui a une valeur constante

$$\boldsymbol{T} = \begin{bmatrix} \tau_1 & & \\ & \ddots & \\ & & \tau_{N_r} \end{bmatrix} \tag{7.23}$$

Le vecteur symbole de données $\boldsymbol{s}$ est perturbé par l'ajout d'un signal de perturbation $\boldsymbol{Tt}$, où $\boldsymbol{T}$ est une matrice diagonale d'éléments égaux aux bases modulos relatifs à chaque type de modulation, et $\boldsymbol{t}$ est le vecteur entier à dimension K. Ensuite, le signal d'émission peut être exprimé sous la forme

$$\boldsymbol{x} = \frac{1}{\sqrt{\beta}} \boldsymbol{F}(\boldsymbol{s} + \boldsymbol{Tt}) \tag{7.24}$$

où $\beta = \|\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{Tt})\|^2$ pour satisfaire la puissance d'émission de l'unité. Le VP combiné peut être représenté comme une recherche de réseau entier où, à l'émetteur, $\hat{\boldsymbol{t}}$ est choisi de telle sorte que β soit minimisé, c.-à-d.

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t} \in \mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}} \|\boldsymbol{F}(\boldsymbol{s} + \boldsymbol{Tt})\|^2 \tag{7.25}$$

Sur la base de l'idée du VP combiné, nous proposons le MMSE-VP combiné pour la modulation adaptative afin de minimiser le critère d'erreur quadratique moyenne (MSE). La matrice de précodage optimale qui minimise le MSE est obtenue en se

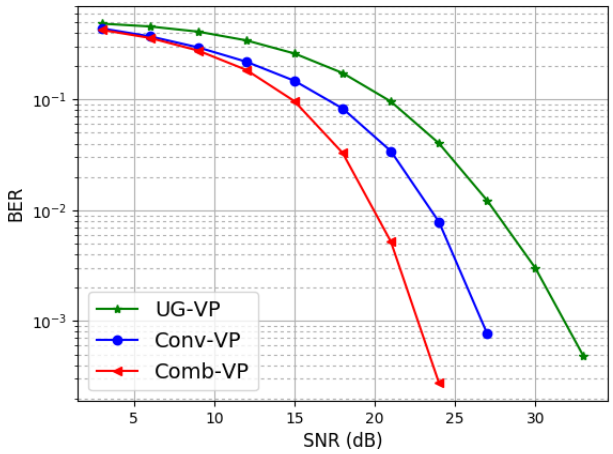


FIGURE 7.12: BER moyen de tous les utilisateurs avec 3 types de modulation différents.

référant à [35]

$$\boldsymbol{F}_o = \boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H + K\sigma_w^2\boldsymbol{I})^{-1} \tag{7.26}$$

Le vecteur de perturbation optimal peut être trouvé comme

$$\boldsymbol{t}_o = \underset{\boldsymbol{t}\in\mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}}\ K\sigma_w^2\tilde{\boldsymbol{s}}^H(\boldsymbol{H}\boldsymbol{H}^H + K\sigma_w^2\boldsymbol{I})^{-1}\tilde{\boldsymbol{s}} \tag{7.27}$$

où $\tilde{\boldsymbol{s}} = \boldsymbol{s} + \boldsymbol{T}\boldsymbol{t}$ désigne le vecteur de données perturbé. Avec la factorisation de Cholesky, (7.27) peut être réécrit comme

$$\boldsymbol{t}_o = \underset{\boldsymbol{t}\in\mathbb{C}\mathbb{Z}^K}{\operatorname{argmin}}\ \|\boldsymbol{L}^H(\boldsymbol{s} + \tau\boldsymbol{t})\|^2 \tag{7.28}$$

où $(\boldsymbol{H}\boldsymbol{H}^H + K\sigma_w^2\boldsymbol{I}_K)^{-1} = \boldsymbol{L}\boldsymbol{L}^H$ avec $\boldsymbol{L}$ une matrice triangulaire inférieure.

Maintenant, nous évaluons les performances du schéma de précodage VP combiné proposé avec le VP conventionnel [33], qui considère la base modulo la plus élevée pour tous les utilisateurs. Nous comparons également les résultats avec UG-VP [88]. On considère une BS équipée de $M = 8$ antennes d'émission desservantes $K = 8$ utilisateurs d'antenne unique en même temps. La Figure 7.12 montre les performances BER moyennes sur tous les utilisateurs, de UG-VP, Conv-VP et Comb-VP. On suppose que les utilisateurs $1, 2$ appliquent 4-QAM, les utilisateurs $3, 4, 5$ appliquent 16-QAM et les autres 64-QAM. Il est bien observé que notre algorithme proposé Comb-VP surpasse UG-VP et Conv-VP. La Figure 7.13 montre que le MMSE-VP combiné surpasse le ZF-VP combiné. En général, la diversité de ces deux précodeurs est la même. Cependant, nous avons observé un gain SNR obtenu par le MMSE-VP combiné dans toute la région SNR.

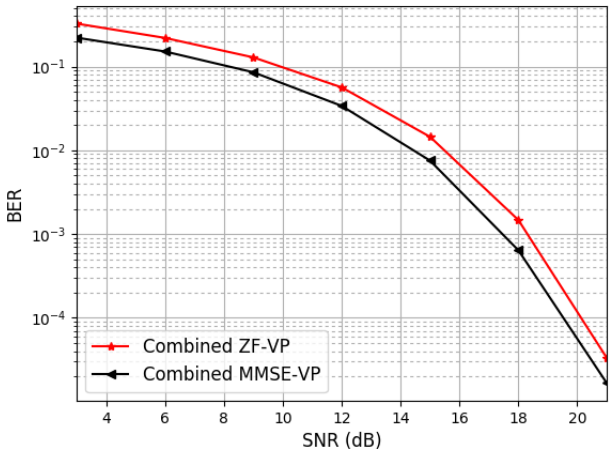


FIGURE 7.13: Performances du précodeur combiné MMSE-VP.

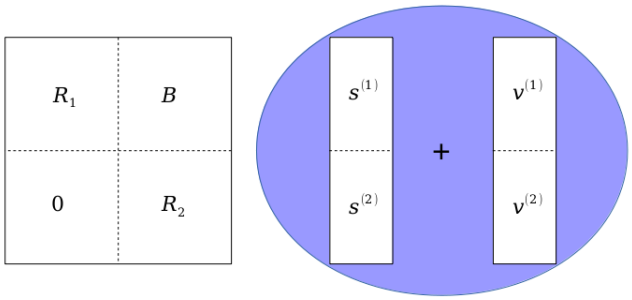


FIGURE 7.14: Block division of the precoding system.

7.5.2 Précodage MU-MIMO récursif par blocs

Cette section présente une technique de précodage de faible complexité appelée algorithme de VP par blocs (Block-VP) basé sur la décomposition "QR" de la matrice de précodage. Dans Block-VP, le VP est appliqué pour chaque bloc en tenant en compte les informations de retour des blocs précédemment perturbés. Ainsi, la perturbation ne sera pas appliquée pour chaque groupe indépendamment comme dans le cas de BD-VP et UG-VP. Le schéma proposé permet d'obtenir l'ordre de diversité souhaité en fixant la taille des blocs. Nous considérons la division en blocs de la matrice triangulaire supérieure $\boldsymbol{R}$ qui est développée à partir de la décomposition "QR" de la matrice de précodage $\boldsymbol{F} = \boldsymbol{QR}$. Nous divisons le système MU-MIMO en deux blocs, comme le montre la Figure 7.14. Le premier bloc en haut est de taille l_1 et le second bloc en bas est de taille l_2. En conséquence, le symbole de données et les vecteurs de perturbation sont divisés en $(\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)})$ et $(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)})$, respectivement.

Sur la base d'une analyse mathématique, nous concluons que dans un système MIMO Gaussien avec M antennes d'émission et K $(\leq M)$ utilisateurs à antenne unique,

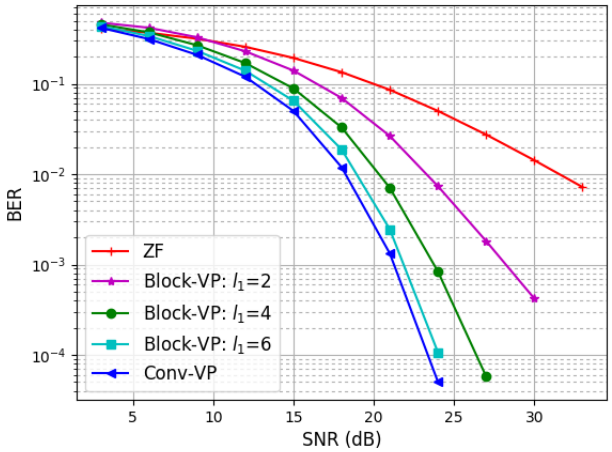


FIGURE 7.15: Résultats de Block VP pour un système 8×8 avec des blocs de taille variable.

l'ordre de diversité pour chaque utilisateur obtenu par le Block-VP est

$$d = M - l_2 \tag{7.29}$$

où $0 \leq l_2 \leq K-1$ est la taille du bloc en bas. En conséquence, on peut dire qu'en termes de diversité, la taille du premier bloc en haut $l_1 = K - l_2$, définit l'ordre de diversité cible $d = M - K + l_1$. Par conséquent, le précodage VP par blosc permet d'obtenir une réduction de complexité couplée à l'ordre de diversité souhaité.

Pour valider notre analyse pour le Block-VP proposé, nous calculons l'exposant d'erreur numériquement à partir des courbes BER en fonction du SNR. La Figure 7.15 montre les résultats du Block-VP, pour un système carré avec $M = K = 8$ et une division en deux blocs de taille variable $l_1 = 2, 4$ et 6. Le schéma de modulation utilisé est le 16-QAM. Les résultats de la simulation confirment que l'ordre de diversité est égal à $M - K + l_1$ dans tous les cas.

7.5.3 Précodage avec des précisions CSI différentes

Pour faire face aux imperfections du canal, nous proposons un signalement d'un nouvel indicateur, appelé CSIAI, qui mesure l'erreur de canal de chaque utilisateur. Cet indicateur représente la précision d'informations sur l'état du canal (CSI) et peut-être calculé en fonction des conditions du canal. Dans diverses études comme dans [99, 104] et leurs références, les coefficients de canal connus de la BS s'écartent du vrai canal par une erreur Gaussienne. Ce modèle capture divers scénarios tels que les erreurs dues à l'estimation de canal, le retard de rétroaction, la quantification de canal dans les systèmes FDD et la non-concordance de réciprocité dans les systèmes TDD. Quand la BS n'a qu'une estimation $\hat{\boldsymbol{H}}$ du vrai canal $\boldsymbol{H}$, alors la relation

entre $\hat{\boldsymbol{H}}$ et $\boldsymbol{H}$ est donnée par

$$\boldsymbol{H} = \hat{\boldsymbol{H}} + \boldsymbol{E} \tag{7.30}$$

où nous supposons que la matrice d'erreur $\boldsymbol{E}$ a des composantes aléatoires Gaussiennes de moyenne nulle. Dans la littérature, les composantes aléatoires de $\boldsymbol{E}$ sont supposées avoir généralement la même variance d'erreur σ_e^2. Cependant, ce n'est pas un scénario réaliste, car la variance d'erreur est associée à chaque utilisateur. En effet, plusieurs utilisateurs sont présents dans le réseau avec une précision CSI différente. Par conséquent, les composantes de chaque $i$ème ligne de $\boldsymbol{E}$ ont la variance d'erreur σ_i^2. Nous notons que les statistiques de $\boldsymbol{E}$ peuvent être estimées à la BS en considérant le signalement de CSIAI qui inclut les différentes variances d'erreur associées aux utilisateurs. Etant donné que les éléments du vrai canal $\boldsymbol{H}$ ont une variance unitaire, la matrice $\hat{\boldsymbol{H}}$ a la même distribution que $\boldsymbol{H}$ avec une variance réduite égale à $1-\sigma_i^2$ pour chaque $i$ème ligne. Soit $\tilde{\boldsymbol{H}}$ défini comme suit

$$\tilde{\boldsymbol{H}} = \boldsymbol{D}\hat{\boldsymbol{H}} \tag{7.31}$$

où $\boldsymbol{D}$ est la matrice diagonale ayant les éléments $\left\{\left(1-\sigma_1^2\right)^{-0.5}, \ldots, \left(1-\sigma_K^2\right)^{-0.5}\right\}$.

Il est évident de voir que $\tilde{\boldsymbol{H}}$ a la même distribution que $\boldsymbol{H}$. Par conséquent, nous introduisons le modèle système de précodage avec rétroaction quantifiée ou discordance de canal, comme le montre la Figure 7.16. Le vecteur de signal reçu peut être écrit comme

$$\boldsymbol{y} = \frac{1}{\sqrt{\hat{\gamma}}}\boldsymbol{DHFs} + \boldsymbol{Dw} = \frac{1}{\sqrt{\hat{\gamma}}}(\boldsymbol{s} + (\boldsymbol{D}\hat{\boldsymbol{H}}\boldsymbol{F} - \boldsymbol{I}_K)\boldsymbol{s} + \boldsymbol{DEFs}) + \boldsymbol{Dw} \tag{7.32}$$

avec une matrice de précodage arbitraire $\boldsymbol{F}$, et $\hat{\gamma} = \mathrm{Tr}(\boldsymbol{F}^H\boldsymbol{F})/P$. Il s'ensuit que la matrice de précodage optimale est donnée par [105]

$$\boldsymbol{F}_{\text{opt}} = (\boldsymbol{D}\hat{\boldsymbol{H}})^H(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{\boldsymbol{H}})^H + K\sigma_{eq}^2\boldsymbol{I}_K)^{-1} \tag{7.33}$$

où $\sigma_{eq}^2 = \frac{1}{K}\sum_{i=1}^{K}\frac{\sigma_w^2+\sigma_i^2}{1-\sigma_i^2}$. Par conséquent, le vecteur de perturbation optimal peut être trouvé comme

$$\boldsymbol{v}_{\text{opt}} = \underset{v}{\text{argmin}}\ K\sigma_{eq}^2\tilde{\boldsymbol{s}}^H(\boldsymbol{D}\hat{\boldsymbol{H}}(\boldsymbol{D}\hat{\boldsymbol{H}})^H + K\sigma_{eq}^2\boldsymbol{I}_K)^{-1}\tilde{\boldsymbol{s}} \tag{7.34}$$

où $\tilde{\boldsymbol{s}} = \boldsymbol{s} + \boldsymbol{v}$ désigne le vecteur de données perturbé.

Dans la Figure 7.17, nous traçons le BER moyen de la technique de précodage proposée en fonction du SNR. Sur la même figure, nous traçons les performances des techniques ZF linéaire, ZF régularisé, et VP conventionnel non-linéaire, avec CSI imparfait. Comme nous pouvons le voir sur la Figure 7.17, la technique de précodage MMSE linéaire exploitant les différentes valeurs de variances d'erreur surpasse le

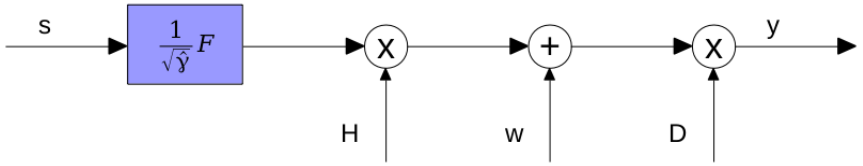


FIGURE 7.16: Modèle système de précodage.

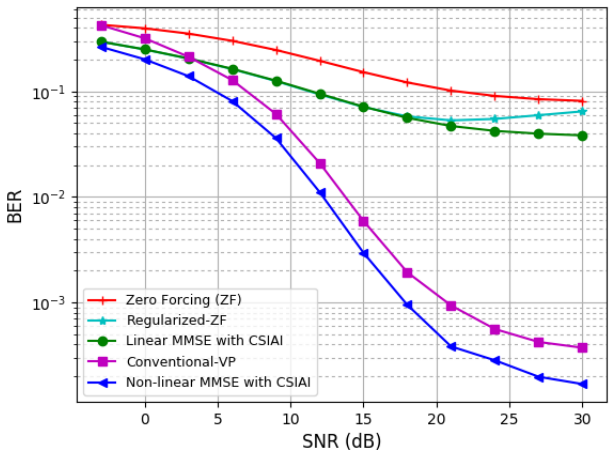


FIGURE 7.17: Performances du nouveau précodeur avec CSIAI avec $M = 12, K = 12$

ZF et le ZF régularisé. On peut voir aussi que le MMSE non-linéaire proposé a de meilleures performances que le VP conventionnel qui ne supporte pas les erreurs CSI. Nous pouvons voir l'effet-plafond où le BER moyen de toutes les techniques de précodage s'aplatit pour un SNR élevé et ne s'améliore pas en augmentant le SNR. Cependant, la technique de précodage proposée peut améliorer le plafond et diminuer le niveau du plancher d'erreur. À ce stade, nous pouvons dire que le précodage MMSE suggéré exploite bien les statistiques d'erreur de canal tant que le CSIAI est signalé. Ainsi, les performances du système sont considérablement améliorées.

7.6 Conclusion

Cette thèse a été dédiée à l'analyse, la conception et l'évaluation des techniques de décodage MIMO dans la réception montante et de précodage MIMO dans la transmission descendante. Différentes configurations MU-MIMO sont envisagées dans les environnements à cellule unique et à cellules multiples.

Les contributions de cette thèse peuvent être mises en œuvre et développées pour satisfaire diverses fonctions objectives avec des contraintes environnementales.

Appendix A

Chapter 3: Number of Radius Updates in the NN-SD

To develop equation (3.17), we have used the following function, which approximates the volume of a sphere B_r with some radius r containing N_p lattice points to the volume of N_p fundamental parallelotopes:

$$N_p = \#\{\boldsymbol{x} \in \boldsymbol{\Lambda} : \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2 \leq r\} \approx \frac{\text{vol}\,(\mathcal{B}_r)}{\det\,(\boldsymbol{\Lambda})} \tag{A.1}$$

where $\det\,(\Lambda)$ is the determinant of Λ, and $\text{vol}(B_r) = r^n \pi^{n/2}/\Gamma(n/2+1) = r^n V_n$.

Let the SNR be defined as $\rho = 10\log_{10}(P/\sigma_w^2)$, and let $r_0^2 = 2n\sigma_w^2$ be the initial square radius. By successively dividing the square radius r_0^2 L times till expecting a small number of lattice points N_p less or equal to a fixed threshold N_{th}, we start the SD search phase with the square radius $r_L^2 = r_0^2/2^L$. We obtain

$$N_p \;\approxeq\; \frac{r_L^n V_n}{\det\,(\Lambda)} \approxeq \frac{\left(2n\sigma_w^2/2^L\right)^{n/2}}{\det\,(\Lambda)} \tag{A.2}$$

By applying $\log_{10}(.)$, we obtain

$$\log_{10}(N_p) = \frac{n}{2}\Big(\log_{10}(2n\sigma_w^2) - L\log_{10}(2)\Big) + \log_{10}(V_n) - \log_{10}(\det\,(\Lambda)) \tag{A.3}$$

Now, L can be written as

$$L = \frac{1}{\log_{10}(2)}\log_{10}(2n\sigma_w^2) + \frac{2}{n\log_{10}(2)}\Big(\log_{10}(V_n) - \log_{10}(\det\,(\Lambda) - \log_{10}(N_p)\Big) \tag{A.4}$$

By using the SNR definition, and taking the expectation over the lattice Λ, we get equation (3.17), where a and b are defined as in (3.18).

Appendix B

Chapter 5: Upper Bound Proof

By definition of the Lovász-reduced basis, the following conditions hold [107]

$$|\mu_{j,i}| \leq 1/2,\ j < i \tag{B.1}$$

$$\|\overline{\boldsymbol{f}}^*_{1,i} + \mu_{i-1,i}\overline{\boldsymbol{f}}^*_{1,i-1}\|^2 \geq \frac{3}{4}\|\overline{\boldsymbol{f}}^*_{1,i-1}\|^2,\ i = 2,\ldots,l_1 \tag{B.2}$$

In particular, by using (B.1) in (B.2), it follows

$$\|\overline{\boldsymbol{f}}^*_{1,i}\|^2 \geq \frac{1}{2}\|\overline{\boldsymbol{f}}^*_{1,i-1}\|^2,\ i = 2,\ldots,l_1 \tag{B.3}$$

We attempt to find an upper bound to the second term of the sample error power in (5.21) for lattice reduction with rounding-off approximation. We use the decomposition of $\boldsymbol{z}_1 = (\boldsymbol{F}_2 - \tilde{\boldsymbol{F}}_2)(\boldsymbol{s}_2 + \boldsymbol{v}_2) + \boldsymbol{F}_1\boldsymbol{s}_1 \in \overline{\boldsymbol{F}}_1\mathbb{C}^{l_1}$

$$\boldsymbol{z}_1 = \sum_{i=1}^{l_1} \beta_i \overline{\boldsymbol{f}}_{1,i} \tag{B.4}$$

In the rounding-off procedure, the lattice translation is given by

$$\boldsymbol{F}_1\boldsymbol{v}_1 = \sum_{i=1}^{l_1} \lambda_i \overline{\boldsymbol{f}}_{1,i},\ \text{with}\ \lambda_i = \mathcal{Q}_\Lambda(\beta_i) \tag{B.5}$$

Hence, we can write the following identities:

$$\begin{aligned}
\boldsymbol{z}_1 + \boldsymbol{F}_1\boldsymbol{v}_1 &= \sum_{i=1}^{l_1}(\beta_i + \lambda_i)\overline{\boldsymbol{f}}_{1,i} \\
&= \sum_{i=1}^{l_1}(\beta_i + \lambda_i)\sum_{j=1}^{i}\mu_{j,i}\overline{\boldsymbol{f}}^*_{1,j} \\
&= \sum_{j=1}^{l_1}\overline{\boldsymbol{f}}^*_{1,j}\sum_{i=j}^{l_1}(\beta_i + \lambda_i)\mu_{j,i}
\end{aligned} \tag{B.6}$$

Therefore, we obtain

$$\|\boldsymbol{z}_1 + \boldsymbol{F}_1\boldsymbol{v}_1\|^2 \leq \sum_{j=1}^{l_1} \|\overline{\boldsymbol{f}}^*_{2,j}\|^2 R^2 \Big(1 + \frac{l_1 - j}{4}\Big) \tag{B.7}$$

$$\leq \|\overline{\boldsymbol{f}}^*_{2,l_1}\|^2 \sum_{i=0}^{l_1-1} 2^i \Big(1 + \frac{i}{4}\Big) \tag{B.8}$$

$$= \|\overline{\boldsymbol{f}}^*_{2,l_1}\|^2 R^2 \Big(2^{l_1} - 2^{l_1-1} + l_1 2^{l_1-2} - \frac{1}{2}\Big) \tag{B.9}$$

$$\leq \frac{3}{4} l_1 2^{l_1} R^2 \|\overline{\boldsymbol{f}}^*_{2,l_1}\|^2 \tag{B.10}$$

where in (B.7) we used the covering radius definition and (B.1), in (B.8) we used (B.3), in (B.9) we used the identities: $\sum_{i=0}^{n} x^n = \frac{1-x^{n+1}}{1-x}$ and $\sum_{i=0}^{n} i x^i = x\frac{1-x^n}{(1-x)^2} - \frac{nx^{n+1}}{1-x}$, and finally in (B.10) we used the fact that $l_1 \geq 1$.

Appendix C

Chapter 5: Proposition Proof

Firstly, we recognize that $\overline{\boldsymbol{H}}_1$ and $\tilde{\boldsymbol{H}}_1$ are two generator matrices of the same lattice. By using (5.19), we obtain

$$(\tilde{\boldsymbol{H}}_1^H)^\dagger = \boldsymbol{F}_1 = \overline{\boldsymbol{F}}_1 \boldsymbol{U}_1 = (\overline{\boldsymbol{H}}_1^H)^\dagger \boldsymbol{U}_1 \tag{C.1}$$

which yields

$$\overline{\boldsymbol{H}}_1 = \tilde{\boldsymbol{H}}_1 \boldsymbol{U}_1^H \tag{C.2}$$

where $\boldsymbol{U}_2^H$ is an unimodular matrix. By taking the QR-decomposition of $\boldsymbol{H}^H$

$$\boldsymbol{H}^H = (\boldsymbol{H}_1 \;\; \boldsymbol{H}_2) = (\boldsymbol{Q}_1 \;\; \boldsymbol{Q}_2) \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{B} \\ 0 & \boldsymbol{R}_2 \end{pmatrix} \tag{C.3}$$

we can write

$$\tilde{\boldsymbol{H}}_1 = (\boldsymbol{Q}_1 \;\; \boldsymbol{Q}_{11}) \triangleq \overline{\boldsymbol{Q}}_1 \begin{pmatrix} \boldsymbol{R}_1 \\ 0 \end{pmatrix} \tag{C.4}$$

for some $M \times (M-K)$ orthogonal matrix $\boldsymbol{Q}_{11}$ such that $\overline{\boldsymbol{Q}}_1$ is $M \times (M-l_2)$ orthonormal. Let $\overline{\boldsymbol{Q}}_u$ be randomly chosen and uniformly distributed over the manifold of complex unitary $(M-l_2) \times (M-l_2)$ matrices and independent of $\boldsymbol{R}_1$

$$\overline{\boldsymbol{Q}}_u = (\boldsymbol{Q}_u | \boldsymbol{Q}_{uu}) \tag{C.5}$$

where $\boldsymbol{Q}_u$ is $(n-l_2) \times l_1$ and $\boldsymbol{Q}_{uu}$ is $(M-l_2) \times (M-K)$. Let $\boldsymbol{H}_c$ be defined as

$$\boldsymbol{H}_c = \boldsymbol{Q}_u \boldsymbol{R}_1 \tag{C.6}$$

and the $M \times (M-l_2)$ orthonormal matrix $\boldsymbol{Q}_c$ as

$$\boldsymbol{Q}_c = \overline{\boldsymbol{Q}}_1 \overline{\boldsymbol{Q}}_u^H \tag{C.7}$$

From (C.4), we obtain

$$\tilde{\boldsymbol{H}}_1 = \boldsymbol{Q}_c \boldsymbol{H}_c \tag{C.8}$$

By combining (C.8) and (C.2), we can write

$$\overline{\boldsymbol{H}}_1 = \boldsymbol{Q}_c \boldsymbol{H}_c \boldsymbol{U}_1^H \tag{C.9}$$

Hence, by definition 1, we conclude that $\overline{\boldsymbol{H}}_1$ is congruent to $\boldsymbol{H}_c$. Moreover, by applying a known result on the QR decomposition of Gaussian matrices [96], we note that the $l_1 \times l_1$ upper triangular matrix $\boldsymbol{R}_1$ in (C.3) is such that its diagonal entries are χ^2 random variables with $2(M-i+1)$ degrees of freedom, the off-diagonal elements are complex Gaussian and all entries are independently distributed. Therefore, if we multiply $\boldsymbol{R}_1$ to the left by an independently distributed random matrix $\boldsymbol{Q}_u$, uniformly distributed over the manifold of complex $(M-l_2) \times l_1$ matrices such that $\boldsymbol{Q}_u^H \boldsymbol{Q}_u = \boldsymbol{I}$, we obtain by construction that (C.6) is an $(M-l_2) \times l_1$ standard complex Gaussian matrix. The assert of the proposition then follows.

Appendix D

Chapter 5: Feedback Load

Let the feedback load for each ith user increases with the SNR such that

$$\sigma_i^2 = k_i(\sigma_w^2)^{\frac{z_i}{M}} + o(\sigma_w^2) \tag{D.1}$$

where $z_i \ (\geqslant 1) \in \mathbb{N}$. Let $\alpha_i = \frac{z_i}{M}$ for the ith user, be the lowest power with the polynomial, namely

$$\lim_{\sigma_w^2 \to 0} \frac{o(\sigma_w^2)}{(\sigma_w^2)^{\alpha_i}} = 0 \tag{D.2}$$

Accordingly, the sum $\sum_{i=1}^{K} \frac{\sigma_w^2 + \sigma_i^2}{1 - \sigma_i^2}$ that appears in Lemmas 1 and 2 in section 5.4.3 can be written as

$$\sum_{i=1}^{K} \frac{\sigma_w^2 + \sigma_i^2}{1 - \sigma_i^2} = \sum_{i=1}^{K} \frac{\sigma_w^2 + k_i(\sigma_w^2)^{\alpha_i} + o(\sigma_w^2)}{1 - k_i(\sigma_w^2)^{\alpha_i} + o(\sigma_w^2)} \tag{D.3}$$

From (5.67), we can express the SER with perfect channel feedback as

$$\log\Big(e(\sigma_w^2, 0^T)\Big) = M\log(\sigma_w^2) + o\Big(\log(\sigma_w^2)\Big) \tag{D.4}$$

where

$$\lim_{\sigma_w^2 \to 0} \frac{o(\log(\sigma_w^2))}{\log(\sigma_w^2)} = 0 \tag{D.5}$$

Using (5.66), (D.3), and (D.4), we obtain

$$\begin{aligned} &\log(e(\sigma_w^2, (\sigma_1^2, \ldots, \sigma_K^2)^T)) \\ &= M\log\left(\sum_{i=1}^{K} \frac{\sigma_w^2 + k_i(\sigma_w^2)^{\alpha_i} + o(\sigma_w^2)}{1 - k_i(\sigma_w^2)^{\alpha_i} + o(\sigma_w^2)}\right) + o\left(\log\left(\sum_{i=1}^{K} \frac{\sigma_w^2 + k_i(\sigma_w^2)^{\alpha_i} + o(\sigma_w^2)}{1 - k_i(\sigma_w^2)^{\alpha_i} + o(\sigma_w^2)}\right)\right) \end{aligned} \tag{D.6}$$

The diversity order can then be determined as

$$\lim_{\sigma_w^2 \to 0} \frac{\log\left(e\left(\sigma_w^2, (\sigma_1^2, \ldots, \sigma_K^2)^T\right)\right)}{\log(\sigma_w^2)} = \begin{cases} M & \min\limits_i \{z_i\} \geq M \\ \min\limits_i \{z_i\} & 1 \leq \min\limits_i \{z_i\} < M \end{cases} \tag{D.7}$$

Thus we have to keep $\sigma_i^2 = k_i(\sigma_w^2)^{1+\varepsilon_i}$ to achieve full diversity order where $\varepsilon_i \geq 0$ for $i \in \{1, \ldots, K\}$. Following the rate distortion theory in [108], the lower bound of mutual information between $\boldsymbol{h}_i$ and $\hat{\boldsymbol{h}}_i$ is given by

$$R(D_i) = M \log_2 \frac{M}{D_i}, \; D_i \leq M \tag{D.8}$$

where D_i is the average squared-error distortion between $\boldsymbol{h}_i$ and $\hat{\boldsymbol{h}}_i$. For convenience, we let $\sigma_i^2 = D_i/M$ and we obtain the following relation between the feedback load and SNR

$$R(D_i) = M\Big(\log_2(\rho) - \log_2(k_i)\Big) \tag{D.9}$$

which reveals that we need an extra $3.32 \times M$ bits per user for every 10 dB SNR to obtain the full diversity. In the general case, each user should get $3.32 \times d$ bits for every 10 dB SNR to experience the diversity order d ($\leq M$) of the system.

Bibliography

[1] E. Dahlman, S. Parkvall, and J. Skold. *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang. What will 5G be? *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, et al. Scenarios for 5G mobile and wireless communications: the vision of the METIS project. *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, 2014.

[4] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder. 5G: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, 2017.

[5] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli. 5G evolution: A view on 5G cellular technology beyond 3GPP release 15. *IEEE Access*, vol. 7, pp. 127639–127651, 2019.

[6] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi. 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[7] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh. 5G URLLC: Design challenges and system concepts. In *IEEE Int. Symp. Wireless Commun. Syst.*, pages 1–6, 2018.

[8] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. D. Carvalho, Y. Ji, Č. Stefanović, et al. Towards massive connectivity support for scalable mMTC communications in 5G networks. *IEEE Access*, vol. 6, pp. 28969–28992, 2018.

[9] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson. The role of small cells, coordinated multipoint, and massive MIMO in 5G. *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 44–51, 2014.

[10] S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, 1998.

[11] J. C. Belfiore, G. Rekaya, and E. Viterbo. The golden code: a 2/spl times/2 full-rate space-time code with nonvanishing determinants. *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1432–1436, 2005.

[12] E. G. Larsson and L. Van der Perre. Massive MIMO for 5G. 2017.

[13] B. Panzner, W. Zirwas, S. Dierks, M. Lauridsen, P. Mogensen, K. Pajukoski, and D. Miao. Deployment and implementation strategies for massive MIMO in 5G. In *IEEE Globecom Workshops*, pages 346–351, 2014.

[14] F. W. Vook, A. Ghosh, and T. A. Thomas. MIMO and beamforming solutions for 5G technology. In *IEEE MTT-S Int. Microw. Symp.*, pages 1–4, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] L. Deng, J. Li, J. T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, et al. Recent advances in deep learning for speech research at Microsoft. In *IEEE Int. Conf. Acoust., Speech, Signal Process.*, volume 26, page 64, 2013.

[17] M. Ibnkahla. Applications of neural networks to digital communications–a survey. *Elsevier Signal Process.*, vol. 80, no. 7, pp. 1185–1215, 2000.

[18] M. Bkassiny, Y. Li, and S. K. Jayaweera. A survey on machine-learning techniques in cognitive radios. *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1136–1159, 2013.

[19] H. Ye, G. Y. Li, and B. H. Juang. Power of deep learning for channel estimation and signal detection in OFDM systems. *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, 2018.

[20] E. Nachmani et al. Deep learning methods for improved decoding of linear codes. *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 119–131, 2018.

[21] N. Farsad and A. Goldsmith. Detection algorithms for communication systems using deep learning. *arXiv preprint arXiv:1705.08044*, 2017.

[22] T. O'Shea and J. Hoydis. An introduction to deep learning for the physical layer. *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, 2017.

[23] S. Dörner et al. Deep learning based communication over the air. *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132–143, 2018.

[24] E. Viterbo and J. Boutros. A universal lattice code decoder for fading channels. *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1639–1642, 1999.

[25] K. S. E. Zigangirov. Some sequential decoding procedures. *Problemy Peredachi Informatsii*, vol. 2, no. 4, pp. 13–25, 1966.

[26] 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.

[27] E. Viterbo and E. Biglieri. A universal decoding algorithm for lattice codes. In *14 Colloque sur le traitement du signal et des images, FRA, 1993*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1993.

[28] G. R. B. Othman. *Nouvelles Constructions algébriques de codes spatio-temporels atteignant le compromis" Multiplexage-Diversité"*. PhD thesis, 2004.

[29] G. R. B. Othman, R. Ouertani, and A. Salah. The spherical bound stack decoder. In *IEEE Int. Conf. Wireless, Mobile Comput., Netw. Commun.*, pages 322–327, 2008.

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

[31] C. B. Peel, B.M. Hochwald, and A.L. Swindlehurst. A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization. *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[32] C. Windpassinger, R. F. H. Fischer, T. Vencel, and J.B. Huber. Precoding in multiantenna and multiuser communications. *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1305–1316, 2004.

[33] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst. A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: Perturbation. *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, 2005.

[34] M. Taherzadeh, A. Mobasher, and A.K. Khandani. Communication over MIMO broadcast channels using lattice-basis reduction. *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4567–4582, 2007.

[35] D. A. Schmidt, M. Joham, and W. Utschick. Minimum mean square error vector precoding. *European Trans. Telecommun.*, vol. 19, no. 3, pp. 219–231, 2008.

[36] M. Mazrouei-Sebdani and W. A. Krzymień. On MMSE vector-perturbation precoding for MIMO broadcast channels with per-antenna-group power constraints. *IEEE Trans. Signal Process.*, vol. 61, no. 15, pp. 3745–3751, 2013.

[37] H. W. Lenstra Jr. Lattices. 2008.

[38] K. M. Tsang. Counting lattice points in the sphere. *Bulletin of the London Mathematical Society*, vol. 32, no. 6, pp. 679–688, 2000.

[39] E. Krätzel. *Lattice points*, volume 33. Springer Science & Business Media, 1989.

[40] J. H. Conway and N. J. A. Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media, 2013.

[41] M. Mohammadkarimi, M. Mehrabi, M. Ardakani, and Y. Jing. Deep learning-based sphere decoding. *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4368–4378, 2019.

[42] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[43] J. S. Armstrong and F. Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *Int. J. Forecast.*, vol. 26, no. 1, pp. 150, 1993.

[44] S. Makridakis. Accuracy measures: theoretical and practical concerns. *Int. J. Forecast.*, vol. 9, no. 4, pp. 527–529, 1993.

[45] A. Meyer. On the number of lattice points in a small sphere. In *Workshop on coding and cryptography*, pages 463–472, 2011.

[46] B. Hassibi and H. Vikalo. On the sphere-decoding algorithm I. Expected complexity. *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, 2005.

[47] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger. Closest point search in lattices. vol. 48, no. 8, pp. 2201–2214, 2002.

[48] A. M. Chan and I. Lee. A new reduced-complexity sphere decoder for multiple antenna systems. In *IEEE Int. Conf. Commun.*, volume 1, pages 460–464, 2002.

[49] L. G. Barbero and J. S. Thompson. Fixing the complexity of the sphere decoder for MIMO detection. *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, 2008.

[50] B. Shim and I. Kang. Sphere decoding with a probabilistic tree pruning. *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4867–4878, 2008.

[51] R. Gowaikar and B. Hassibi. Statistical pruning for near-maximum likelihood decoding. *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2661–2675, 2007.

[52] B. Shim and I. Kang. Radius-adaptive sphere decoding via probabilistic tree pruning. In *IEEE Workshop Signal Process. Adv. Wireless Commun.*, pages 1–5, 2007.

[53] X. W. Chang, J. Wen, and X. Xie. Effects of the LLL reduction on the success probability of the Babai point and on the complexity of sphere decoding. *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4915–4926, 2013.

[54] W. Zhao and G. B. Giannakis. Sphere decoding algorithms with improved radius search. *IEEE Trans. Commun.*, vol. 53, no. 7, pp. 1104–1109, 2005.

[55] H. Vikalo, B. Hassibi, and T. Kailath. Iterative decoding for MIMO channels via modified sphere decoding. *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2299–2311, 2004.

[56] Z. Yang, C. Liu, and J. He. A new approach for fast generalized sphere decoding in MIMO systems. *IEEE Signal Process. Lett.*, vol. 12, no. 1, pp. 41–44, 2005.

[57] F. Zhao and S. Qiao. Radius selection algorithms for sphere decoding. In *Proc. 2nd Canadian Conf. Computer Sci. Softw. Eng.*, pages 169–174, 2009.

[58] M. A. Khsiba and G. R. B. Othman. Sphere decoder with dichotomic search. In *IEEE Annu. Int. Symp. Pers. Indoor, Mobile Radio Commun.*, pages 1–7, 2017.

[59] M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 324–335, 2010.

[60] W. J. Choi, R. Negi, and J. M. Cioffi. Combined ML and DFE decoding for the V-BLAST system. In *IEEE Int. Conf. Commun.*, 2000.

[61] S. D. Howard. Low complexity essentially maximum likelihood decoding of perfect space-time block codes. In *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009.

[62] S. Sirinaunpiboon, A. R. Calderbank, and S. D. Howard. Fast essentially maximum likelihood decoding of the Golden code. *IEEE Trans. Inf. Theory*, 2011.

[63] A. Barreal, C. Hollanti, and D. Karpuk. Reduced complexity decoding of nxn algebraic space-time codes. *arXiv preprint arXiv:1501.06686*, 2015.

[64] L. P. Natarajan and B. S. Rajan. An adaptive conditional zero-forcing decoder with full-diversity, least complexity and essentially ML performance for STBCs. *IEEE Trans. Signal Process.*, 2013.

[65] M. Liu, J. F. Hélard, M. Crussiere, and M. Hélard. Reduced-complexity maximum-likelihood decoding for 3D MIMO code. In *IEEE Wireless Commun., Netw. Conf.*, 2013.

[66] T. Xu and X. G. Xia. On space-time code design with a conditional PIC group decoding. *IEEE Trans. Inf. Theory*, 2011.

[67] M. A. Khsiba and G. R. B. Othman. Semi-exhaustive reduced-complexity recursive block decoding for MIMO systems. In *IEEE Int. Conf. Telecommun.*, pages 1–6, 2016.

[68] A. Askri, G. R. B. Othman, and M. A. Khsiba. Block recursive MIMO decoding. In *IEEE Int. Conf. Telecommun.*, pages 116–120, 2018.

[69] T. W. Anderson (1958). *An introduction to multivariate statistical analysis*.

[70] T. H. Liu. Comparisons of two real-valued MIMO signal models and their associated ZF-SIC detectors over the rayleigh fading channel. *IEEE Trans. Wireless Commun*, 2013.

[71] N. M. Temme. Asymptotic inversion of incomplete gamma functions. *Mathematics of Computation*, 1992.

[72] T. Q. Quek, M. Peng, O. Simeone, and W. Yu. *Cloud radio access networks: Principles, technologies, and applications*. Cambridge University Press, 2017.

[73] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-constrained cloud radio access networks: Insights and challenges. *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, 2015.

[74] Y. Zhou and W. Yu Wei. Optimized backhaul compression for uplink cloud radio access network. *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, 2014.

[75] Y. Zhou and W. Yu. Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN. *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4138–4151, 2016.

[76] N. Samuel, T. Diskin, and A. Wiesel. Learning to detect. *IEEE Trans. Signal Process. 2554–2564*, vol. 67, no. 10, pp. 2554–2564, 2019.

[77] G. Gao, C. Dong, and K. Niu. Sparsely connected neural network for massive MIMO detection. In *IEEE 4th Int. Conf. Computer Commun.*, pages 397–402, 2018.

[78] H. He, C. K. Wen, S. Jin, and Geoffrey G. Y. Li. A model-driven deep learning network for MIMO detection. In *IEEE Global Conf. Signal Inf. Process.*, pages 584–588, 2018.

[79] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming. Adaptive neural signal detection for massive MIMO. *IEEE Trans. Wireless Commun.*, 2020.

[80] S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[81] D. J. Love, R. W. Heath, and T. Strohmer. Grassmannian beamforming for multiple-input multiple-output wireless systems. *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, 2003.

[82] E. Björnson, J. Hoydis, and L. Sanguinetti. Massive MIMO networks: Spectral, energy, and hardware efficiency. *Found. Trends Signal Process.*, vol. 11, no. 3-4, pp. 154–655, 2017.

[83] J. Ma and L. Ping. Orthogonal AMP. *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[84] C. Jeon, R. Ghods, A. Maleki, and C. Studer. Optimality of large MIMO detection via approximate message passing. In *IEEE Int. Symp. Inf. Theory*, pages 1227–1231. IEEE, 2015.

[85] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[86] M. A. Albreem and N. A. H. B. Ismail. A review: detection techniques for LTE system. *Springer Telecommun. Syst.*, vol. 63, no. 2, pp. 153–168, 2016.

[87] C. B. Chae, S. Shim, and R. W. Heath. Block diagonalized vector perturbation for multiuser MIMO systems. *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp.4051–4057, 2008.

[88] R. Chen, C. Li, J. Li, and Y. Zhang. Low complexity user grouping vector perturbation. *IEEE Wireless Commun. Lett.*, vol. 1, no. 3, pp. 189–192, 2012.

[89] A. Li and C. Masouros. A constellation scaling approach to vector perturbation for adaptive modulation in MU-MIMO. *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 289–292, 2015.

[90] F. Tosato and M. Sandell. Diversity analysis of group vector perturbation precoding. *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6426–6432, 2012.

[91] M. Salehi and J. Proakis. Digital communications. *McGraw-Hill Education*, vol. 31, pp. 32, 2007.

[92] W. Banaszczyk. New bounds in some transference theorems in the geometry of numbers. *Mathematische Annalen*, vol. 296, no. 1, pp. 625–635, 1993.

[93] L. Babai. On Lovász' lattice reduction and the nearest lattice point problem. *Combinatorica*, vol. 6, no. 1, pp. 1–13.

[94] A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, vol. 261, no. 4, pp. 515–534, 1982.

[95] M. S. Bartlett. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, vol. 53, pp. 260–283, 1934.

[96] A. M. Tulino, S. Verdú, and S. Verdu. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.

[97] T. Yoo, N. Jindal, and A. Goldsmith. Multi-antenna downlink channels with limited feedback and user selection. *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491.

[98] N. Jindal. Mimo broadcast channels with finite-rate feedback. *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, 2006.

[99] D. J. Ryan, I. B. Collings, I. V. L. Clarkson, and R. W. Heath. Performance of vector perturbation multiuser MIMO systems with limited feedback. *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2633, 2009.

[100] L. Sun and M. Lei. Quantized CSI-based Tomlinson-Harashima precoding in multiuser MIMO systems. *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1118–1126, 2013.

[101] M. Payaró, A. P. Iserte, A. I. P. Neira, and M. A. Lagunas. Robust design of spatial Tomlinson-Harashima precoding in the presence of errors in the CSI. *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2396–2401, 2007.

[102] P. Lu and H. C. Yang. Vector perturbation precoding for MIMO broadcast channel with quantized channel feedback. In *IEEE Global Telecommun. Conf.*, pages 1–5, 2009.

[103] J. Maurer, J. Jalden, D. Seethaler, and G. Matz. Vector perturbation precoding revisited. *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 315–328, 2010.

[104] M. Mazrouei-Sebdani and W. A. Krzymien. Vector perturbation precoding for network MIMO: Sum rate, fair user scheduling, and impact of backhaul delay. *IEEE Trans. Veh. Technol.*, vol. 61, no. 9, pp. 3946–3957, 2012.

[105] F. Dietrich. *Robust signal processing for wireless communications*, volume 2. Springer Science & Business Media, 2007.

[106] A. D. Dabbagh and D. J. Love. Multiple antenna MMSE based downlink precoding with quantized feedback or channel mismatch. *IEEE Trans. Commun.*, vol. 56, no. 11, pp. 1859–1868, 2008.

[107] H. Napias. A generalization of the LLL-algorithm over euclidean rings or orders. *Journal de théorie des nombres de Bordeaux*, vol. 8, no. 2, pp. 387–396, 1996.

[108] T. M Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Publications

Journal Articles

- A. Askri, C. Zhang, and G. R. B. Othman: **Learning assisted Fronthaul Compression for Multi-Antenna Uplink C-RAN**. – *Submitted to IEEE Access*, May 2021.
- A. Askri and G. R. B. Othman: **Vector Perturbation and User Grouping for the Downlink in Multi-User MIMO Systems**. – *Submitted to IEEE Open J. Signal Process.*, March 2021.
- A. Askri and G. R. B. Othman: **Multi-User MIMO Downlink Precoding for Users with Different CSI Accuracy**. – *Submitted to IEEE Trans. Commun.*, March 2021.

Conference Papers

- C. Zhang, A. Askri, and G. R. B. Othman: **Distributed DNN based Processing for Uplink Could-RAN**. – *IEEE Int. Symp. Inf. Theory*, July 2021, Melbourne, Victoria, Australia.
- A. Askri, and G. R. B. Othman: **Precoding in Massive MU-MIMO Systems Based on New CSI Accuracy Indicator Reporting**. – *IEEE Int. Conf. Telecommun.*, October 2020, Bali, Indonesia.
- A. Askri, and G. R. B. Othman: **Combined Vector Perturbation for Adaptive Modulation in MU-MIMO**. – *IEEE Int. Conf. Telecommun.*, October 2020, Bali, Indonesia.
- A. Askri, G. R. B. Othman, and H. Ghauch: **Counting Lattice Points in the Sphere using Deep Neural Networks**. – *IEEE Asilomar Conf. Sig., Syst., Comput.*, November 2019, Pacific Grove, CA, USA.
- A. Askri, and G. R. B. Othman: **DNN assisted Sphere Decoder**. – *IEEE Int. Symp. Inf. Theory*, July 2019, Paris, France.
- A. Askri, G. R. B. Othman, and M. A. Khsiba: **Block Recursive MIMO Decoding**. – *IEEE Int. Conf. Telecommun.*, June 2018, Saint-Malo, France.

Patents

- G. R. B. Othman, A. Askri, and C. Zhang: **METHOD AND SYSTEM FOR COORDINATED MULTI POINT TRANSMISSION COORDINATION**. – *Eur. App.*, EP21305203.8, February 2021.
- G. R. B. Othman, C. Zhang, and A. Askri: **SYSTEM AND METHOD FOR UPLINK COORDINATED TRANSMISSION IN AN OPEN RADIO ACCESS NETWORK**. – *Eur. App.*, EP21305203.8, February 2021.
- C. Zhang, G. R. B. Othman, A. Askri, and S. E. Hajri: **PRE-PROCESSING IN UPLIKNK RAN USING NEURAL NETWORK**. – *Eur. App.*, PCT/FR2020/051895, October 2020.
- G. R. B. Othman and A. Askri: **DEVICES AND METHODS FOR RECURSIVE BLOCK PRECODING**. – *Eur. App.*, EP20306114.8, September 2020.
- G. R. B. Othman, A. Askri, and S. E. Hajri: **Combined vector perturbation for adaptive modulation coding schemes**. – *US App.*, US62/968,515, February 2020.
- G. R. B. Othman, A. Askri, and F. Tosato: **APPARATUS, METHOD AND COMPUTER PROGRAM**. – *Eur. App.*, EP2019/079368, October 2019.
- G. R. B. Othman, and A. Askri: **DEVICES AND METHODS FOR MACHINE LEARNING ASSISTED SPHERE DECODING**. – *Eur. App.*, EP19305886.4, July 2019.
- G. R. B. Othman, and A. Askri: **DEVICES AND METHODS FOR MACHINE LEARNING ASSISTED PRECODING**. – *Eur. App.*, EP193-05887.2, July 2019.
- G. R. B. Othman, and A. Askri: **DEVICES AND METHODS FOR LATTICE POINTS ENUMERATION**. – *Eur. App.*, EP19305888.0, July 2019.

ECOLE DOCTORALE

Titre: La réception de Liaison Montante et la Transmission de Liaison Descendante en MU-MIMO pour la 5G

Mots clés: MU-MIMO, Réseaux de Points, Algèbre, Réseaux de Neurones Profonds

Résumé:
Les technologies à entrées multiples et sorties multiples (MIMO) ont été développées pour augmenter la capacité du système et offrir une meilleure fiabilité de la liaison. Ils permettent une architecture réseau dense qui permettra à de nombreux utilisateurs de se connecter dans la même zone sans subir de ralentissements. Les réseaux 5G et au-delà utiliseront ces technologies MIMO avec de nombreuses petites antennes permettant au faisceau de se concentrer sur une zone donnée. Couplées à des bandes haute fréquence, l'utilisation de ces antennes augmentera considérablement le débit.

Dans ces systèmes, la détection multi-utilisateurs (MU)-MIMO dans la réception de la liaison montante et le précodage dans la transmission de la liaison descendante permettent de séparer les flux de données utilisateur et de pré-annuler les interférences. Cependant, certains défis doivent être relevés dans des conditions réalistes telles que dans des conditions réalistes telles que la complexité raisonnable des processus de décodage et de précodage, la connaissance erronée des canaux et l'interférence des cellules adjacentes. Cette thèse aborde toutes ces limitations ci-dessus pour la réception en liaison montante et la transmission en liaison descendante dans les systèmes MU-MIMO.

Pour la réception sur la liaison montante, nous étudions l'algorithme bien connu de décodage par sphères (SD) pour la détection MIMO. Nous cherchons à réduire sa complexité qui augmente de manière exponentielle avec le nombre d'antennes et la taille de la constellation. Ainsi, nous profitons des récentes avancées dans le domaine des réseaux de neurones (NNs) pour développer le SD assisté par les NNs de faible complexité. Nous proposons également le décodage MIMO récursif par blocs, qui atteint presque la performance de maximum de vraisemblance (ML). En utilisant les réseaux neuronaux profonds (DNNs), nous suggérons un nouveau schéma peu complexe pour le traitement et la détection du signal dans la liaison montante du cloud-RAN (C-RAN). Ce schéma DNN vise à imiter toute la transmission en liaison montante C-RAN, qui prend en compte les contraintes de quantification au niveau des unités radio distantes (RRUs) et les observations corrompues au niveau du processeur central (CP).

Dans la transmission en liaison descendante, nous étudions le précodage de la perturbation vectorielle (VP) non-linéaire. Nous concevons le VP combiné pour servir plusieurs utilisateurs avec différents schémas de codage de modulation (MCSs). Nous introduisons également l'algorithme VP par blocs, qui fusionne le précodage linéaire et non-linéaire pour offrir un compromis accordable entre complexité et performance. Pour traiter les informations erronées sur l'état du canal (CSI) dans le précodage de la liaison descendante, nous développons le nouvel indicateur de précision CSI pour concevoir un nouveau précodeur moins sensible aux erreurs CSI.

Title: The Uplink Reception and Downlink Transmission in MU-MIMO for 5G

Keywords: MU-MIMO, Lattices (groups), Algebra, Deep Neural Networks

Abstract:

Multiple-input multiple-output (MIMO) technologies were developed to increase system capacity and offer better link reliability. They allow a dense network architecture that will allow many users to connect in the same area without experiencing slowdowns. 5G networks and beyond will use these MIMO technologies with many small antennas allowing the beam to be focused on a given area. Coupled with high-frequency bands, the use of these antennas will significantly increase throughput.

In such systems, multi-user (MU)-MIMO detection in the uplink reception and MU-MIMO precoding in the downlink transmission enable separating user data streams and pre-cancelling interference. However, some challenges have to be met under realistic conditions, such as the reasonable complexity of the decoding and precoding processes, the erroneous channel knowledge, and the adjacent cell interference. This thesis addresses all these limitations above for the uplink reception and the downlink transmission in MU-MIMO systems.

In the uplink reception, we study the well-known sphere decoding (SD) algorithm for MIMO detection. We seek to reduce its complexity which increases exponentially with the number of antennas and the constellation size. Thus, we profit from recent advances in neural networks (NNs) to develop the low-complexity NN assisted SD. We also propose the block recursive MIMO decoding, achieving almost the maximum likelihood (ML) performance. Using deep neural networks (DNNs), we suggest a new and low complex scheme for signal processing and cloud-RAN (C-RAN) detection. This DNN scheme aims to mimic the whole transmission in uplink C-RAN, which considers the quantization constraints at the radio remote units (RRUs) and the corrupted observations at the central processor (CP).

In the downlink transmission, we study the non-linear vector perturbation (VP) precoding. We design the combined VP to serve multiple users with different modulation coding schemes (MCSs). We also introduce the block VP algorithm, which merges linear and non-linear precoding to offer a tunable tradeoff between complexity and performance. To deal with the erroneous channel state information (CSI) in the downlink precoding, we develop the new CSI accuracy indicator reporting to design a novel precoder that is less sensitive to CSI errors.

Institut Polytechnique de Paris

91120 Palaiseau, France