PSEUDO-RANDOM QUANTIZATION BASED DETECTION IN ONE-BIT MASSIVE MIMO SYSTEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKHAN YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JANUARY 2023

Approval of the thesis:

PSEUDO-RANDOM QUANTIZATION BASED DETECTION IN ONE-BIT MASSIVE MIMO SYSTEMS

submitted by **GÖKHAN YILMAZ** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. Ali Özgür Yılmaz
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Buyurman Baykal
Electrical and Electronics Engineering, METU _____

Prof. Dr. Ali Özgür Yılmaz
Electrical and Electronics Engineering, METU _____

Prof. Dr. Sinan Gezici
Electrical and Electronics Engineering, Bilkent University _____

Assoc. Prof. Dr. Ayşe Melda Yüksel Turgut
Electrical and Electronics Engineering, METU _____

Assist. Prof. Dr. Gökhan Muzaffer Güvensen
Electrical and Electronics Engineering, METU _____

Date: 18.01.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Gökhan Yılmaz

Signature :

ABSTRACT

PSEUDO-RANDOM QUANTIZATION BASED DETECTION IN ONE-BIT MASSIVE MIMO SYSTEMS

Yılmaz, Gökhan
M.S., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. Ali Özgür Yılmaz

January 2023, 110 pages

Analog-to-digital converter (ADC) units are one of the most power-hungry devices in the radio-frequency (RF) chains of massive multiple-input multiple-output (MIMO) systems. Therefore, utilizing low-resolution ADCs in uplink massive MIMO systems is a practical solution to decrease power consumption. However, when high modulation orders are employed for high-rate communication, the achievable rate saturates after a finite SNR value due to the stochastic resonance (SR) phenomenon. This thesis proposes a novel pseudo-random quantization (PRQ) scheme that can help compensate for the effects of SR and makes communication with high-order modulation schemes with one-bit quantization possible. The ADC thresholds at the receiver side of uplink one-bit massive MIMO systems are changed to work with the PRQ scheme. We modify linear detectors for one-bit non-zero threshold quantization and propose new detection methods for the frequency-flat and frequency-selective fading scenarios. For flat fading, we offer a two-stage detector that works with PRQ. The first stage is an iterative method called Boxed Newton Detector (BND) that utilizes Newton's method to maximize the log-likelihood. The second stage, Nearest Codeword Detector (NCD), exploits the first stage to create a small set of most likely candidates based on sign

constraints to increase performance. For frequency-selective fading, we design a new frequency-domain equalization (FDE) scheme, called the projected quasi-Newton detector (PQND), to optimize the log-likelihood using a quasi-Newton approach that works with PRQ in both orthogonal frequency division multiplexing (OFDM) and single carrier (SC) systems. The proposed methods outperform the existing detectors with comparable complexity.

Keywords: massive MIMO, detection, pseudo-random quantization, one-bit quantization, one-bit analog-to-digital converter (ADC)

ÖZ

BİR-BİT KİTLESEL MIMO SİSTEMLERDE SÖZDE RASTGELE NİCELEME BAZLI TESPİT

Yılmaz, Gökhan
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ali Özgür Yılmaz

Ocak 2023 , 110 sayfa

Analog-sayısal dönüştürücüler (ADC), kitlesel çok girdili çok çıktılı (MIMO) sistemlerin radyo-frekansı (RF) zincirlerinde yer alan en yüksek güç tüketimine sahip cihazlar arasında yer alır. Dolayısıyla, çıkış-yolu kitlesel MIMO sistemlerde düşük çözünürlüklü ADC kullanımı, güç tüketimini azaltmak adına pratik bir çözümdür. Buna rağmen, yüksek veri hızlarına ulaşmak için yüksek kipleme dereceleri ile çalışılırken, erişilebilir veri hızı belli bir işaret-gürültü oranının (SNR) üstüne çıkıldığında stokastik rezonans (SR) olgusundan dolayı doyuma ulaşır. Bu tezde SR olgusunun olumsuz etkilerini telafi ederek çıkış yolu bir-bit kitlesel MIMO sistemlerde yüksek kipleme dereceleri ile çalışılabilmesine olanak sağlayacak yeni bir sözde rastgele niceleme (PRQ) yöntemi, alıcı tarafta ADC ünitelerinin niceleme eşikleri değiştirilerek elde edilir. Geleneksel tespit yöntemleri, sıfırdan farklı eşikli bir-bit niceleme senaryosuna uyarlanır. Ayrıca, biri düz sönümlemeli ve diğeri frekans seçici sönümlemeli kanallar için olmak üzere iki yeni tespit yöntemi sunulur. Düz sönümlemeli kanallar için sunulan yöntem iki aşamalıdır. Kutulu Newton dedektörü (BND) adlı ilk aşamada Newton'un yöntemi baz alınarak olabilirlik fonksiyonu iteratif olarak optimize edilir.

En yakın kodlu-söz dedektörü (NCD) adlı ikinci aşamada ise performansı artırmak için ilk aşamadan gelen kestirim kullanılarak işaret kısıtlamalarına göre küçük bir en olasılıklı aday kümesi oluşturulur. Frekans seçici sönümlemeli kanallar için, olabilirlik fonksiyonunun bir yarı-Newton yöntemle optimize edildiği projeksiyonlu yarı-Newton dedektörü (PQND) isimli, PRQ ile hem dik-frekans-bölümlemeli çoğullama (OFDM) hem de tek taşıyıcılı (SC) sistemlerde kullanılabilecek yeni bir frekans bölgesi eko-layzırı tasarlanır. Önerilen yöntemler kullanılarak benzer hesaplama karmaşıklığı ile literatürde yer alan yöntemlerden daha yüksek tespit başarımı elde edilir.

Anahtar Kelimeler: kitlesel MIMO, tespit, sözde rastgele niceleme, bir-bit niceleme, bir-bit analog-sayısal dönüştürücü

To my dear family and loved ones

ACKNOWLEDGMENTS

I want to express my deepest gratitude to my advisor Prof. Ali Özgür Yılmaz. Beginning with the last year of my undergraduate studies, I have had the chance to learn many important aspects of research and engineering from him. It has been a privilege to have his guidance and mentorship. His support and confidence inspired me and helped me believe I could succeed in times of hesitation. This endeavor would not be possible from the beginning without his encouragement.

I am also grateful to Assist. Prof. Gökhan Muzaffer Güvensen for always being accommodating and understanding since the day I first met him and for helping me gain valuable insights into communication theory. I would like to extend my sincere thanks to Dr. Ali Bulut Üçüncü for always offering to help and share his ideas. His research has been one of the essential sources for my studies. Special thanks to Assist. Prof. Eren Balevi for helpful discussions and meetings.

I should thank Vodafone Turkey for funding my graduate studies as part of the 5G and Beyond Joint Graduate Support Program and the Information and Communication Technologies Authority of Turkey (BTK) for organizing this meaningful program to support research on communication technologies in our country. I also would like to offer my sincere thanks to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for the scholarship they granted me as part of the 2210-A program.

My family deserves important credit for supporting me in every part of my life. My mother Tülay, my father Ömer, and my sister Derya have always believed in me to succeed. I would be remiss in not mentioning my dear friends Buğra Esmer, Eren Ergünel, Fevzi Özgür, Furkan Akça, Hakan Hüdaverdi, and Halis Dönmez for always being there for me for more than a decade.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
ADC	Analog-to-Digital Converter
AWGN	Additive White Gaussian Noise
BER	Bit Error Rate
BMRC	Bussgang-based Maximum Ratio Combining
BND	Boxed Newton Detector
bpcu	bits per channel use
BPSK	Binary Phase Shift Keying
BS	Base Station
BZF	Bussgang-based Zero Forcing
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
CP	Cyclic Prefix
CSCG	Circularly Symmetric Complex Gaussian
CSI	Channel State Information
CT	Continuous Time
DAC	Digital-to-Analog Converter
DFT	Discrete Fourier Transform
DT	Discrete Time
FD	Frequency Domain
FDE	Frequency Domain Equalization
FFT	Fast Fourier Transform
ISI	Inter-Symbol Interference

I/Q	In-Phase/Quadrature
IID	Independent Identically Distributed
LDS	Large Delay Spread
LLN	Law of Large Numbers
LO	Local Oscillator
MC	Multi-Carrier
MIMO	Multiple-Input Multiple-Output
ML	Maximum Likelihood
MLSD	Maximum Likelihood Sequence Detection
MMSE	Minimum Mean Square Error
MRC	Maximum Ratio Combining
MUI	Multi-User Interference
NCD	Nearest Codeword Detector
OFDM	Orthogonal Frequency Division Multiplexing
PA	Power Amplifier
PAPR	Peak-to-Average Power Ratio
PDP	Power Delay Profile
PDF	Probability Distribution Function
PQND	Projected Quasi-Newton Detector
PRQ	Pseudo-Random Quantization
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase Shift Keying
RF	Radio Frequency
SIMO	Single-Input Multiple-Output
SISO	Single-Input Single-Output
SC	Single-Carrier
SDS	Small Delay Spread

SNR	Signal-to-Noise Ratio
SR	Stochastic Resonance
TD	Time Domain
ZF	Zero-Forcing
ZTQ	Zero-Threshold Quantization

CHAPTER 1

INTRODUCTION

Communication is an essential part of humans as social beings. Every individual desires to communicate ideas and feelings, search for help and wisdom or tell experiences. Throughout history, there have been various communication tools such as making signs and gestures, petroglyphs, pictograms, ideograms, all forms of art, and thousands of languages spread across the earth and time. As societies and technology advanced, communicating with people over long distances, in other cities, countries, and even continents became necessary. As the inventor of complex languages, humankind's need to communicate ideas over long distances inspired many scientific discoveries and inventions, such as the telegraph and the telephone, followed by the invention of more advanced technologies, such as the radio and television [5]. Then, Shannon revolutionized the field [6], which can be seen as the beginning of the Information Age. Digital communication systems were studied extensively during the 1960s. Networked communications systems were developed, starting with the first internet node in 1970. The transmission control protocol (TCP) and internet protocol (IP) were founded in 1980. A new age started with the invention of the world wide web (WWW) in 1993. The internet proliferated and provided connections between people, machines, devices, and processes that led to today's Internet of Everything (IoE).

During these developments, special attention must be given to cellular communication systems. The first generation (1G) of cellular networks was used during the 1980s with analog transmission using the frequency re-usage idea. Second-generation (2G) cellular communication was used in the 1990s, where both voice and text transmissions were applicable. It is the first standard to use digital communications, and the global system for mobile communications (GSM) standard first appeared as part of the 2G

standard. As the internet progressed, the general packet radio service (GPRS - 2.5G) and the enhanced data rates for GSM evolution (EDGE - 2.75G) were established for data communications. The third generation (3G) technology arrived with many improvements and innovations in 2001, such as code division multiple access (CDMA) and a significant increase in the bandwidth (from 200 kHz to 5 MHz) to support higher data rates [7].

After a decade, the fourth generation (4G) standard arrived. The demand for larger data rates was even higher, with 100 Mbps for high mobility and 1 Gbps for low mobility communication. The orthogonal frequency division multiple access (OFDMA) scheme replaced CDMA. Bandwidth usage increased up to 20 MHz. Quadrature phase shift keying (QPSK) to 64-quadrate amplitude modulation (QAM) schemes were utilized with turbo and low-density parity-check (LDPC) codes. The most important part of 4G concerning this thesis is the multiple-input multiple-output (MIMO) systems, where multiple antennas at the transmitter and the receiver units are utilized to multiply the data rate. The long-term evolution (LTE) became the dominant standard for 4G, and LTE Advanced Pro (4.5G) is the widely used standard today. Subsequently, the fifth generation (5G) standard has also started operating in many parts of the world. The 5G standard employs both sub-6 GHz and the mmWave band. It offers ultra-reliable low latency communications (URLLC), enhanced mobile broadband (eMBB), and massive machine-type communications (mMTC). The network is fully software-defined, and much larger bandwidths and frequencies are used. A significant development in the later versions of 4G and 5G is the usage of massive MIMO, where the base stations (BS) employ large antenna arrays for higher array gains and beamforming capabilities.

1.1 Problem Definition

The number of users of cellular networks, the need for more data rate and better quality of service (QoS) increase rapidly. Therefore, the physical layer of the communication systems becomes a critical factor in meeting the demands. Massive MIMO is one of the most important aspects of the physical layer design of the 5G and future cellular communication technologies [8]. The area throughput in a cellular network depends on the average cell density, bandwidth, and spectral efficiency. Cell density and

bandwidth can be limiting factors for cellular network designs. Hence, increasing spectral efficiency is critical for increasing area throughput. Massive MIMO systems are equipped with much larger numbers of antennas than the conventional MIMO setting, with the number of antennas at the BS vastly exceeding the number of users served by the BS. Therefore, massive MIMO can increase spectral efficiency by orders of magnitude due to much larger multiplexing and diversity gains. In addition to increased spectral efficiency, massive MIMO can provide hardware efficiency [9]. Distortion due to non-ideal transceiver hardware can be mitigated with a large number of BS antennas. The source of the hardware impairments could be quantization noise, sampling offsets or jitters of the analog-to-digital converters (ADCs), non-linearity of the power amplifiers (PAs), amplitude or phase imbalance in the in-phase/quadrature (I/Q) mixers, and phase noise in the local oscillators (LOs), among many others.

Despite many advantages, since each antenna at the BS requires separate radio-frequency (RF) chains, the power consumption and the hardware cost can become a burden as the number of antennas increases. ADC units are one of the main power-consuming components in the RF chains. Many studies deal with the modeling of the power consumption of ADCs [10, 11]. In general, the power consumption of an ADC increases linearly with the sampling rate and exponentially with the resolution. The resolution of an ADC is expressed in terms of bits, and the power consumption of an ADC unit is doubled for each added bit.

Hence, by taking advantage of the hardware efficiency of massive MIMO systems, low-resolution ADCs have become an attractive solution to deal with the power consumption issue [12]. Especially one-bit ADCs are very popular due to their additional benefits, such as having very simple circuitry and not needing automatic gain control (AGC) units since they work with only a single threshold level. There are different studies related to the channel capacity [13–16], and the achievable rate [17–20] of uplink MIMO systems where the BS is equipped with one-bit ADCs. The channel characteristics and whether the transmitter has access to the channel state information (CSI) are essential factors that affect the capacity, as in the case of infinite-resolution systems. One particular outcome for one-bit systems from these studies is that the channel capacity is finite, and the upper bound is related to the number of BS antennas. Also, the performance gap between the one-bit and infinite-resolution,

i.e., unquantized, systems is small at the low signal-to-noise ratio (SNR) regime. It gets more prominent as the SNR increases. Note that high-resolution systems can be approximated to an infinite-resolution system when the resolution is high enough, so the quantization distortion becomes negligible.

1.2 Related Work

An important issue related to nonlinear systems such as one-bit massive MIMO is the stochastic resonance (SR) phenomenon [21]. Unlike linear systems where a higher SNR leads to better performance, the nonlinear distortion induced by one-bit quantization leads to either performance degradation or performance saturation depending on the scenario [15]. Hence, the high-SNR performance is generally limited by quantization distortion. One-bit MIMO measurements are generally good at recovering the phase information compared to amplitude [22]. This limitation causes significant disadvantages in detecting symbols from high-order QAM constellations. Hence, previous one-bit massive MIMO literature generally focuses on modulation schemes such as binary phase shift keying (BPSK) or QPSK [23–25]. Different works reported performances for multi-user systems or in ISI channels for 16-QAM [2,26,27]. However, the high-SNR error floor or performance saturation was declared a significant limiting factor in these situations.

Randomization of quantization is an effective tool to mitigate the effects of quantization distortion [28], which is a version of dithering [29] widely used in audio and visual systems. Pseudo-random quantization (PRQ), where the digital processor perfectly knows the dither signal, can be even more beneficial due to additional information regarding the dither signal [30]. The dithering strategy can be helpful also in quantized MIMO systems. An antithetic dithering approach is used in [31], where negatively correlated dither signals are utilized by doubling the ADCs at each branch. Using uniformly distributed thresholds is also possible, as it is done in [32], where randomly generated thresholds that are updated for different channel realizations were utilized during the training of and estimation with the proposed neural network. [33] utilizes the dithering approach in downlink channels with one-bit quantized signal transmission at the BS, where a dither signal is added to the unquantized signal before quantization

and precoding operations.

Regarding detection methods for one-bit uplink massive MIMO systems, the literature is rich with works on the frequency-flat fading [2,3,23,24,27,34] and on the frequency-selective fading [3,4,35–39] scenarios. Linear processing based approaches [2,23,35] generally suffer from high SNR error floors. More sophisticated methods can result in better performance but at the cost of increased computational complexity. The techniques used for both frequency-flat and frequency-selective channels in one-bit massive MIMO systems are generally unsuitable to support high-order modulations such as 64-QAM, 256-QAM, or higher. In [20], it is shown that quantization results in circularly symmetric distortion with Gaussian distribution and radial distortion due to loss of amplitude information during detection using linear receivers in one-bit quantized orthogonal frequency division multiplexing (OFDM) systems. [20] also shows that frequency-selectivity of the wireless channel favors one-bit MIMO-OFDM systems, and a large number of channel taps helps lower the amplitude distortion. [35] focuses on linear and iterative block decision feedback equalizers for SC-frequency domain equalization (FDE). Bayes-optimal joint channel estimation and detection schemes are proposed for flat-fading systems in [26], and frequency-selective systems with OFDM in [36]. A reinforcement learning approach is proposed to compensate for the mismatch due to channel estimation errors while utilizing likelihood-based detection in [40]. [23] utilizes an ADMM-based algorithm for detection. A message-passing algorithm for frequency-selective channels is proposed in [37]. A recent study in [41] utilizes a hybrid scheme with analog processing and adaptive quantization thresholds to maximize the achievable rate.

The gradient-based optimization techniques have recently enjoyed popularity in one-bit MIMO detection [2, 4, 24, 34]. [34] is the first detector to rely on first-order optimization on the likelihood function. However, different system setups require fine-tuning the algorithm hyperparameters, such as step size. The works in [2, 24] utilize the deep unfolding technique [42], which unfolds each iteration of the gradient descent algorithm onto a neural network architecture. With this method, adaptive step sizes can be learned by determining a fixed number of iterations for the algorithm. However, different system setups require training for the given specific scenario. The works from [2, 3, 34] propose two-stage detection techniques for frequency-flat fading,

where the first stage is an equalizer and the second stage uses the first-stage estimate to create a set of most likely candidates to apply maximum likelihood (ML) detection with reduced complexity. [4] proposes the gradient descent algorithm for FDE under frequency-selective fading for MIMO-OFDM systems. However, it requires fine-tuning the algorithm step size for different setups and operating SNR values to obtain fast and favorable convergence.

1.3 Contributions

The contributions of this thesis can be summarized as follows:

- While many adopted system setups focus on one-bit quantization where all quantization thresholds are set to zero, we concentrate on non-zero threshold quantization. We derive linear filtering-based traditional detectors for this scenario. Specifically, we focus on the Bussgang-based and conventional quantization-unaware linear filters under both frequency-flat and frequency-selective fading.
- We propose a novel pseudo-random quantization (PRQ) scheme for one-bit uplink massive MIMO systems to reduce the adverse effects of the SR phenomenon on the detection performance. We show that the proposed method can increase the achievable rate in one-bit single-input multiple-output (SIMO) systems. Also, by following a coding theoretic approach and taking the one-bit quantized observations as codewords in space, we show that the minimum Hamming distance of the codebook for a given channel realization at infinite SNR is increased in one-bit massive MIMO systems with PRQ, compared to the scenario where conventional zero-threshold quantization (ZTQ) is used, especially for high-order QAM constellations.
- Similar to [2, 3, 34], we propose a new two-stage detection technique for one-bit massive MIMO systems operating under frequency-flat fading. This approach is based on the PRQ scheme, where quantization thresholds are modified to obtain a dithering effect. The first stage is called Boxed Newton Detector (BND), for which Newton's method, a second-order optimization technique, is adopted for fast convergence and to omit difficulties of selecting different step sizes since

Newton's method [43] selects the appropriate step size in all dimensions at each iteration. The second stage, called Nearest Codeword Detector (NCD), is used to refine the first-stage solution as in the existing methods but does so by taking the one-bit observations as codewords in space and creating a set of most likely candidates with respect to the minimum Hamming distance criterion. By using PRQ with BND-NCD methods, we show that massive MIMO systems can operate with high-order modulation schemes as 256-QAM and 1024-QAM, which were not attempted before and thus, their performances were not reported by any of the previous work in the literature to the best of our knowledge. The proposed detector outperforms the state-of-the-art detectors even when the conventional ZTQ scheme is employed.

- Then, we turn our focus to one-bit uplink massive MIMO-OFDM systems operating under frequency-selective fading. Due to severe nonlinear distortion, the conventional time-domain (TD) and frequency-domain (FD) conversions are not directly applicable in one-bit systems. Influenced by the BND approach, we propose a new FDE scheme named Projected Quasi-Newton Detector (PQND). This second-order optimization method utilizes the properties of massive MIMO systems with appropriate approximations to obtain low complexity. Due to the second-order derivative information, the algorithm does not require fine-tuning of the step size and can work with PRQ, unlike [4] to support high modulation orders such as 64-QAM and 256-QAM. The proposed method is also extended as a single carrier-frequency domain equalization (SC-FDE) scheme.

1.4 Notation

Lower-case letters represent scalars, lower-case bold letters represent column vectors, and upper-case bold letters represent matrices. $(.)^T$ represents the transpose and $(.)^H$ the Hermitian of a matrix. Scalars, vectors, and matrices with over-bar notation, e.g., $\bar{a}$, $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{A}}$ are complex-valued, and their real-valued counterparts are shown with no accent, i.e., a, $\boldsymbol{a}$, $\boldsymbol{A}$. The n^{th} element of a vector $\boldsymbol{a}$ is a_n. The n^{th} row and k^{th} column of a matrix $\boldsymbol{A}$ is denoted as $[\boldsymbol{A}]_{(n,k)}$. $\|.\|$ denotes the ℓ_2 norm of a vector. $\mathrm{diag}(.)$ operator creates a diagonal matrix by placing the elements of its vector arguments

on the main diagonal. It eliminates off-diagonal parts of its matrix arguments to create a diagonal matrix. $j = \sqrt{-1}$ is the imaginary unit. $\mathbb{R}$ and $\mathbb{C}$ are the set of real and complex numbers, respectively. $\mathbb{Z}$ is the set of integers, and $\mathbb{Z}^+$ denotes the set of positive integers. $\Re\{.\}$ and $\Im\{.\}$ give their arguments' real and imaginary parts, respectively. $|.|$ denotes the absolute value of its scalar arguments and the cardinality of its set arguments. $\langle . \rangle_V$ denotes the modulo V operator. $\bar{\boldsymbol{F}}$ is the unitary discrete Fourier transform (DFT) matrix of size $V \times V$, where $[\bar{\boldsymbol{F}}]_{n,k} = \mathrm{e}^{-j2\pi(n-1)(k-1)/V}/\sqrt{V}$. $\mathcal{F}_v\{.\} = \sum_{m=0}^{V-1}\{.\}\mathrm{e}^{-j2\pi mv/V}/\sqrt{V}$ is the unitary V-point DFT operator and $\mathcal{F}_m^{-1}\{.\} = \sum_{v=0}^{V-1}\{.\}\mathrm{e}^{+j2\pi mv/V}/\sqrt{V}$ is the unitary V-point inverse DFT operator. $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\phi(x) = \sqrt{\frac{1}{2\pi}}\exp(-\frac{x^2}{2})$ is the probability density function (PDF) of the standard Gaussian distribution. $\Phi(x) = \int_{-\infty}^{x}\phi(\tau)d\tau$ is the cumulative distribution function (CDF) of the standard Gaussian distribution. Each function is applied element-wise to its arguments. $\mathcal{H}(.)$ is the entropy, $\mathcal{H}_b(.)$ is the binary entropy, and $\mathcal{I}(.;.)$ is the mutual information operator. $\odot$ is the Hadamard, and $\otimes$ is the Kronecker product. The vectorization property is allowed so that the Hadamard product of a vector $\boldsymbol{a} \in \mathbb{R}^N$ and a matrix $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{b}_1 & \boldsymbol{b}_2 & \dots & \boldsymbol{b}_N\end{bmatrix}^T \in \mathbb{R}^{N\times K}$ results in $\boldsymbol{a} \odot \boldsymbol{B} = \begin{bmatrix}a_1\boldsymbol{b}_1 & \dots & a_2\boldsymbol{b}_2 & \dots a_n\boldsymbol{b}_N\end{bmatrix}^T = \mathrm{diag}(\boldsymbol{a})\boldsymbol{B} \in \mathbb{R}^{N\times K}$.

1.5 Outline

The remainder of this thesis is organized as follows:

- In Chapter 2, the basic system description to obtain the received signal models in both single-carrier (SC) and multi-carrier (MC) system setups is explained. Then, an ADC's operating principles and power consumption are analyzed. Finally, the benefits of using low-resolution ADCs to obtain power-efficient massive MIMO systems are discussed.
- In Chapter 3, detection methods under frequency-flat fading are examined. Traditional linear approaches and their modified forms for non-zero threshold quantization are obtained. Then, the proposed PRQ scheme is explained by analyzing the performance of PRQ using approaches from information theory

and coding theory. A new two-stage detection method, BND-NCD, is derived that works with the proposed PRQ scheme. The computational complexities of the proposed methods are compared with the existing techniques from the literature, and simulation results are shown to discuss the error performances.

- In Chapter 4, detection methods under frequency-selective fading are examined. The linear processing-based methods are again considered. The method used for the BND approach is modified for frequency-selective fading. The detector is simplified with appropriate approximations to obtain a quasi-Newton optimization technique, PQND, that can also work with PRQ. Finally, computational complexity analysis and simulation results are examined and discussed.

- We conclude our discussion in Chapter 5 with a summary of the topics covered in the thesis and comments on some possible research directions for the future.

CHAPTER 2

ONE-BIT UPLINK MASSIVE MIMO SYSTEMS AND ANALOG-TO-DIGITAL CONVERTERS

2.1 Motivation

This chapter aims to obtain the signal and system models in uplink one-bit massive MIMO systems. We obtain the mathematical models starting with the continuous-time (CT) signals up to the sampled and quantized discrete-time (DT) received signal. Then, we investigate the working principle of an ADC and analyze its power consumption using existing figure of merit (FoM) models to point out how crucial the power consumption issue can get in high-resolution massive MIMO systems and the advantage of utilizing low-resolution ADCs. While obtaining the signal model for one-bit quantized observations, we consider a non-zero threshold quantization scenario different than many other works in the literature [2, 3, 20, 26, 34, 36, 38–40, 44, 45], where ZTQ is adopted. We focus on the non-zero threshold quantization scenario to be able to use dithering by changing the quantization thresholds [30] instead of generating analog dither signals in Section 3.7.2.

Throughout the thesis, we focus on an $N \times K$ uplink massive MIMO system where K single-antenna users are served by a BS equipped with N receiver antennas, where $N \gg K$. A simple block diagram of the system is shown in Fig. 2.1, where RA stands for receiver antenna and UE stands for user equipment. Even though we assume each user has a single antenna, extension to multi-antenna users is straightforward by assuming each user transmits independent streams from their antennas. Furthermore, each user transmits bits using a discrete modulation alphabet of size M, such as M-QAM. The selected modulation alphabet is denoted as $\bar{\mathcal{M}}$. We adopt a block-

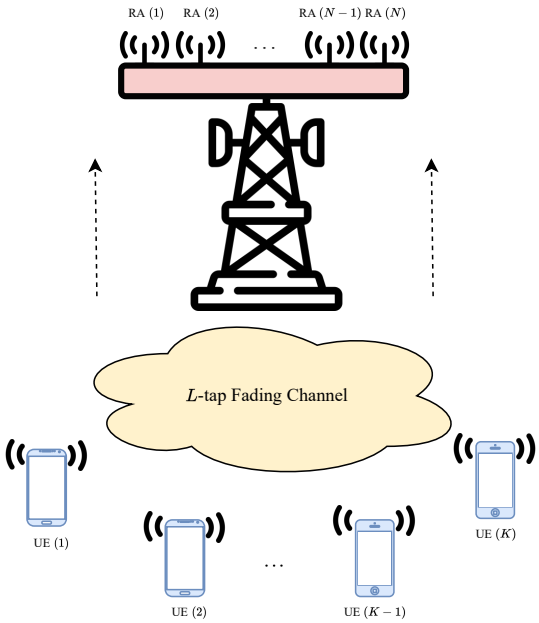


Figure 2.1: A basic illustration of the uplink massive MIMO system model, where RA stands for receiver antenna, and UE represents user equipment with a single antenna.

fading channel model, where the channel fading process remains constant for at least a data block duration of V symbols in the time domain (TD). The v^{th} symbol in the data block of the k^{th} user can be shown as $\bar{x}_k[v]$. Note that the average symbol power per user is $\mathbb{E}[|\bar{x}_k[v]|^2] = E_s = 1$. Since we deal with the detection task, we assume the BS has perfect knowledge of the channel state information (CSI), whose estimation in one-bit massive MIMO systems is extensively studied in the literature [3, 17, 24, 26, 31, 34, 36, 37, 46, 47]. Both SC and MC (OFDM) transmission schemes are considered in the thesis. We assume a frequency-selective fading channel with L taps in both cases. The special case of $L = 1$ corresponds to a frequency-flat fading scenario, and only SC transmission is considered for flat fading.

2.2 Single-Carrier (SC) System Description

There are minor differences between the SC and MC system models. Each user generates bits to be transmitted in the uplink. Throughout the thesis, we assume uncoded transmission of the users' bits. Hence, each sequence of $log_2(M)$ bits is mapped to one of the M symbols from the chosen modulation alphabet. After a

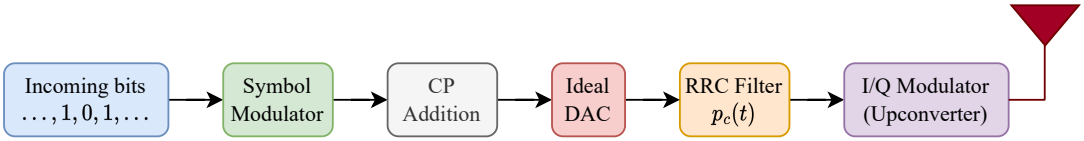


Figure 2.2: Block diagram of UE transmitter unit for uplink SC transmission.

data block of V constellation symbols is generated, each user transmits their symbols through an ideal digital-to-analog converter (DAC) and a Nyquist pulse shaping filter $p_c(t)$ such as root-raised cosine (RRC). We denote the symbol duration as T_s. Then, the baseband signal is upconverted in an I/Q modulator unit. Note that if the channel is frequency-selective, i.e., $L > 1$, a cyclic prefix (CP) of length $L_{\text{CP}} \geq L - 1$ is added at the beginning of each data block of length $V > L$. Inserting CP is a valuable technique to mitigate the effects of inter-symbol interference (ISI) at the receiver side at the expense of slightly decreasing the spectral efficiency. Even though CP is generally used with OFDM systems, it is also helpful with SC-FDE to obtain block circulant channel matrices, i.e., circular convolution operation. A basic block diagram of a transmitter unit, i.e., user equipment (UE), is shown in Fig. 2.2.

Due to the advanced synchronization capabilities in 5G, we assume perfect synchronization between the users and no offset in sampling time or carrier frequency. The only source of hardware impairment in the system is the ADC units. Hence, transmitters, i.e., users, have ideal RF components. The adopted model in this thesis is very similar to the one used in [48], where a more detailed model for systems employing temporal oversampling can also be found. Thermal noise at the receiver side at each antenna $\bar{z}_n(t)$ for $n = 1, 2, \ldots, N$ is modeled as an independent circularly symmetric complex Gaussian (CSCG) process with power spectral density (PSD) N_0. Once the transmitted signal passes through the channel and the additive white Gaussian noise (AWGN) process corrupts the received signal, the bandpass signal is downconverted at

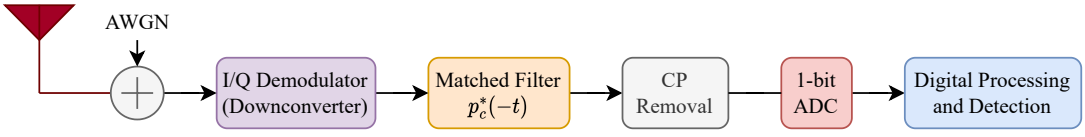


Figure 2.3: Block diagram of an RA unit RF chain for uplink SC transmission.

each antenna in an I/Q demodulator, and the CP part is discarded. A basic block diagram of an RA unit is shown in Fig. 2.3. As in [48], by denoting the channel impulse response (CIR) between the user k and receiver antenna n, at time t as $\bar{h}_{(n,k)}(t)$, the CT baseband received signal at receiver antenna n, at time t can be written as

$$\bar{d}_n(t) = \sum_{k=1}^{K} \sum_{\ell=0}^{L-1} \sum_{v=0}^{V-1} \bar{h}_{(n,k)}[\ell] \bar{x}_k[\langle v-\ell \rangle_V] p_c(t - vT_s) + \bar{z}_n(t), \tag{2.1}$$

for $n = 1, 2, \ldots, N$, where $\bar{h}_{n,k}[\ell] = \bar{h}_{(n,k)}(\ell T_s)$ and $\langle . \rangle_V$ denotes the modulo V operator to express the circular convolution operation due to CP. At each receiver antenna of the BS, the incoming CT signal is passed through a pulse-matched filter $p_c^*(-t)$, which is the complex conjugate time inverse version of the pulse shaping filter $p_c(t)$. The convolution of the pulse-shaping filter with its matched filter yields

$$p(t) = p_c(t) * p_c^*(-t), \tag{2.2}$$

where $*$ is the convolution operator, and the resulting filter is the raised cosine pulse whose symbol-rate samples can be found as

$$p[m] = p(mT_s) = \delta[m], \tag{2.3}$$

where $\delta[0] = 1$, $\delta[m] = 0$ for $m \neq 0$ is the DT unit impulse function. Convolution of the white noise process with the RRC filter results in

$$w_n(t) = \bar{z}_n(t) * p_c^*(-t), \tag{2.4}$$

for $n = 1, 2, \ldots, N$, where $w_n(t)$ is a bandlimited noise process. Once the received signal in (2.1) is pulse matched filtered, the output becomes

$$\begin{aligned} \bar{y}_n(t) &= \bar{d}_n(t) * p_c^*(-t) \\ &= \sum_{k=1}^{K} \sum_{\ell=0}^{L-1} \sum_{v=0}^{V-1} \bar{h}_{(n,k)}[\ell] \bar{x}_k[\langle v-\ell \rangle_V] p(t - vT_s) + \bar{w}_n(t). \end{aligned} \tag{2.5}$$

Now, the operations related to the RF chains are almost complete, and the remaining part is related to the analog-to-digital conversion, which includes sampling and quantization. ADCs and quantization operations are explained in detail in the following section. For now, we focus on the mathematical model of the sampled DT unquantized signal. Using (2.4), the DT received signal samples at the n^{th} antenna is obtained as

$$\bar{y}_n[m] = \bar{y}_n(mT_s) = \sum_{k=1}^{K} \sum_{\ell=0}^{L-1} \bar{h}_{(n,k)}[\ell] \bar{x}_k[\langle m-\ell \rangle_V] + \bar{w}_n[m], \tag{2.6}$$

for $n = 1, 2, \ldots, N$ and $m = 0, 1, \ldots, V-1$, where $w_n[m]$ is the $\mathcal{CN}(0, N_0)$ distributed noise sample at the n^{th} antenna, at time m. Noise samples are assumed to be independent both in time and space. Note that the roll-off factor of the RRC pulse does not play any role in this setup since the receiver employs Nyquist-rate ISI-free sampling.

When $L = 1$, the wireless channel is no longer frequency-selective and is modeled as frequency-flat. For this scenario, CP is no longer necessary and can be omitted. The resulting DT received signal samples at the n^{th} antenna under flat fading is found as

$$\begin{aligned} \bar{y}_n[m] &= \sum_{k=1}^{K} \bar{h}_{(n,k)}[0] \bar{x}_k[m] + \bar{w}_n[m] \\ \bar{y}_n &= \sum_{k=1}^{K} \bar{h}_{(n,k)} \bar{x}_k + \bar{w}_n. \end{aligned} \tag{2.7}$$

for $n = 1, 2, \ldots, N$. In (2.7), since there is no ISI, the time indices can be dropped. For a given channel realization, only the symbols sent at a particular time index affect the observations at the BS. Hence we obtain a memoryless system.

2.3 Multi-Carrier (MC) System Description

The MC system model is similar to the SC model with slight differences. For OFDM transmission, the overall bandwidth is divided into sub-bands, and each user assigns one of their constellation symbols from their data block to one of the subcarriers. For

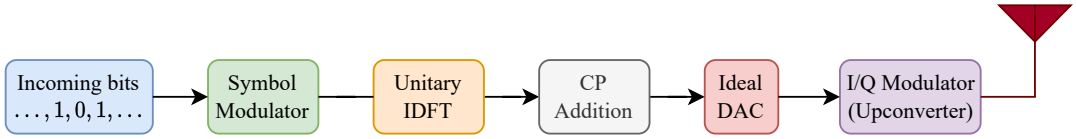


Figure 2.4: Block diagram of UE transmitter unit for uplink OFDM transmission.

this purpose, a unitary inverse DFT operation is conducted on the data block before transmission by each user. Note that the well-known fast Fourier transform (FFT) algorithm is applied in practice to perform the DFT operations. OFDM transmission does not explicitly utilize pulse-shaping filters. Hence, we utilize a rectangular pulse $p_r(t)$ instead of $p_c(t)$, which does not affect the mathematical model. UE and RA units for uplink OFDM transmission are shown in Fig. 2.4 and 2.5, respectively. By taking the inverse DFT operation into account, the received signal at the n^{th} antenna can be written as

$$\bar{d}_n(t) = \sum_{k=1}^{K} \sum_{\ell=0}^{L-1} \sum_{v=0}^{V-1} \bar{h}_{(n,k)}[\ell] \bar{s}_k[\langle v - \ell \rangle_V] p_r(t - vT_s) + \bar{z}_n(t), \tag{2.8}$$

for $n = 1, 2, \ldots, N$, where, $\bar{s}_k[m]$ is found as

$$\bar{s}_k[m] = \frac{1}{\sqrt{V}} \sum_{v=0}^{V-1} \bar{x}_k[v] e^{+j2\pi mv/V}. \tag{2.9}$$

for $k = 1, 2, \ldots, K$ and $m = 0, 1, \ldots, V-1$. Note that the pulse shaping filter is also a low-pass filter (LPF) for SC systems. The LPF at the analog front-end for OFDM serves this purpose. We assume the LPF is an infinitely sharp and perfectly flat filter in the passband. Hence, there is no interference from an adjacent band. Different than the SC system, the matched filter block is omitted. After CP is discarded, the sampled unquantized representation of the DT received signal can be obtained as

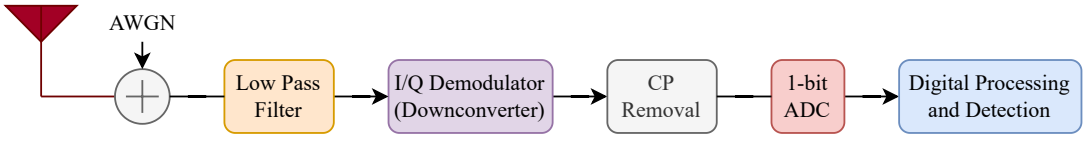


Figure 2.5: Block diagram of an RA unit RF chain for uplink OFDM transmission.

$$\bar{y}_n[m] = \sum_{k=1}^{K} \sum_{\ell=0}^{L-1} \bar{h}_{(n,k)}[\ell] \bar{s}_k[\langle m - \ell \rangle_V] + \bar{w}_n[m]. \tag{2.10}$$

for $n = 1, 2, \ldots, N$ and $m = 0, 1, \ldots, V-1$. The noise samples are IID with PDF $\mathcal{CN}(0, N_0)$ since the bandlimited noise process is sampled to obtain a flat PSD. The unquantized DT signal models in both SC and MC sections are just representations to obtain the quantized form as part of the ADC units. The following section covers the quantized signal model and ADC unit basics.

2.4 Analog-to-Digital Converters (ADCs)

2.4.1 Working Principle of an ADC

ADCs are one of the most crucial electronic components that bridge the real world and digital devices. The CT signals of real life, such as voltage, temperature, pressure, and sound need to be represented appropriately as DT signals to be processed digitally. Since digital devices all have finite memory, analog signals should be represented with limited samples. The rate at which a bandlimited CT signal should be sampled to be able to perfectly reconstruct it from its sampled DT version is determined by the Nyquist-Shannon sampling theorem [49]. The amplitude level of each sampled DT signal is also stored in digital devices. Hence, storing them in infinite precision is not an option. Therefore, samples should also be quantized and represented with the appropriate quantization labels. The basic block diagram of an ADC is shown in Fig. 2.6. For a fixed sampling period T_s, the analog signal to be digitized should first go through an anti-aliasing filter with a cutoff frequency chosen according to

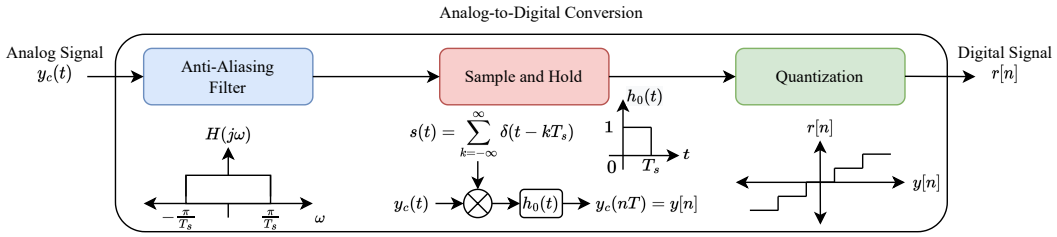


Figure 2.6: Basic configuration of an ADC.

the sampling theorem. In practice, a filter cannot be infinitely sharp, as illustrated in Fig. 2.6. However, this will not be of concern in this thesis since a filter can be sharp enough to be approximated to have a rectangular frequency response. Then comes a sample and hold block where the CT signal is multiplied with an impulse train, and its value is captured for one sampling period. Finally, quantization occurs where the signal amplitude is mapped to one of the quantization levels with finite precision to produce the digital output. A rigorous explanation and theory of CT to DT conversion and analog to digital conversion can be found in [50].

2.4.2 Quantized Signal Model

Infinite-resolution quantization would correspond to a quantizer with a linear transfer function and yield an error-free representation of the sample amplitudes. Unfortunately, finite resolution quantization, applicable in practice, results in quantization errors/quantization noise. Since binary representation is essential in digital circuits, ADC resolution is expressed by the number of bits that represent the amplitude information. For a b-bit quantizer, there can be 2^b quantization levels. In a basic implementation, b-bits would require $2^b - 1$ comparators. Hence, circuit complexity and power consumption increase as the ADC resolution increases. Throughout the thesis, the unquantized, i.e., infinite-resolution, samples of the analog signals will be used to obtain the quantized signals in the digital domain. The b-bit quantized version $\bar{r}[n]$ of a DT signal $\bar{y}[n]$ can be obtained as

$$\bar{r}[n] = Q_b(\Re\{\bar{y}[n]\}) + j\, Q_b(\Im\{\bar{y}[n]\}), \tag{2.11}$$

where $Q_b(.)$ represents the uniform scalar mid-rise quantization operator. In communication systems, we deal with complex signals with both in-phase (I) and quadrature (Q) parts. As a result, the quantization of a complex signal occurs in both parts separately. Assuming b-bit uniform quantization, a set of $2^b - 1$ quantization thresholds must be identified which can be shown as $\{\tau_1, \tau_2, \ldots, \tau_{2^b-1}\}$, where we can assume $-\infty = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{2^b-1} < \tau_{2^b} = \infty$. If we take $\Delta \in \mathbb{R}$ as the step size of the quantizer, each threshold value can be found as

$$\tau_n = (-2^{b-1} + n)\Delta \text{ for } n = \{1, 2, \ldots, 2^b - 1\}. \tag{2.12}$$

As a result, for an input $x \in \mathbb{R}$, the quantization operator can be defined as

$$Q_b(x) = \begin{cases} \tau_n - \frac{\Delta}{2}, & x \in (\tau_{n-1}, \tau_n] \text{ and } n \in \{1, 2, \ldots, 2^b - 1\} \\ (2^b - 1)\frac{\Delta}{2}, & x \in (\tau_{2^b-1}, \tau_{2^b}) \end{cases}. \tag{2.13}$$

For the special case of one-bit quantization, our main focus in this thesis, $b = 1$ results in $\tau = \tau_1 = 0$ according to (2.12). By taking $\Delta = 2$, the quantization operator becomes

$$Q_1(x) = \text{sign}(x). \tag{2.14}$$

Then we can re-write (2.11) as

$$\bar{r}[n] = \text{sign}(\Re\{\bar{y}[n]\}) + j\,\text{sign}(\Im\{\bar{y}[n]\}), \tag{2.15}$$

where $\text{sign}(x)$ for $x \in \mathbb{R}$ is the signum function which is defined as

$$\text{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}. \tag{2.16}$$

Consequently, one-bit quantization yields only the sign information of the input analog signal sample, and the amplitude information is completely lost. Note that for mid-rise uniform one-bit quantization, there is only one quantization threshold and $\tau = 0$. However, if the threshold is selected such that $\tau \neq 0$, (2.14) should be re-written as

$$Q_1(x) = \text{sign}(x - \tau). \tag{2.17}$$

If we assume a complex representation $\bar{\tau}$ of thresholds for I and Q ADCs, (2.15) can be updated as

$$\bar{r}[n] = \text{sign}(\Re\{\bar{y}[n] - \bar{\tau}\}) + j\,\text{sign}(\Im\{\bar{y}[n] - \bar{\tau}\}). \tag{2.18}$$

Note that the unquantized representations of the received signal (2.6), (2.7), and (2.10) can be directly utilized to obtain the one-bit quantized signals using either (2.15) or (2.18) depending on the scenario.

2.4.3 Power Consumption of an ADC

One-bit ADCs cause significant amplitude distortion on the digital output signal. What makes one-bit ADCs tolerable in communication systems is massive MIMO which may mitigate the effects of nonlinear quantization distortion thanks to the large spatial degrees of freedom provided by a large antenna array setup. Hence, large numbers of antennas in massive MIMO systems are significant to tolerate quantization errors. Moreover, what makes one-bit ADCs preferable in such a scenario is the previously mentioned simple circuitry and low power consumption compared to the high-resolution systems. The amount of comparator units required by an ADC increases exponentially as the resolution increases, directly affecting the system's overall power consumption.

Different models from the literature characterize the power consumption of an ADC [10]. Generally, a FoM is defined as a measure of efficiency that depends on several system properties. The first is the Walden FoM [11], which can be found as

$$\text{FoM}_{\text{Walden}} = \frac{P}{f_s 2^{\text{ENOB}}}, \tag{2.19}$$

where $f_s = 1/T_s$ is the sampling frequency, P is the power consumption, and ENOB is the effective number of bits of the ADC. ENOB is calculated for a target signal-to-noise-and-distortion ratio (SNDR) value as

$$\text{ENOB} = \frac{\text{SNDR}\,(\text{dB}) - 1.76}{6.02}. \tag{2.20}$$

ENOB is generally close to the actual resolution of the ADC. However, additional

distortion sources other than quantization results in a lower value for ENOB. As a result, the Walden FoM predicts that the power consumption of an ADC will double for each additional bit. A more recent FoM is the Schreier FoM from [51], which can be found as

$$\text{FoM}_{\text{Schreier}} = \text{SNDR}\,(\text{dB}) + 10\log_{10}\left(\frac{f_s}{2P}\right), \tag{2.21}$$

and it predicts that the power consumption will almost quadruple for each added bit. According to [10], for today's technology, the Walden FoM is a better measure for low-resolution ADCs, whereas the Schreier FoM is better suited for high-resolution systems.

The software radio implementation of ADCs is investigated in [1], where the authors discuss the design and implementation issues of ADCs. ENOB vs. f_s with constant power consumption curves from [1] are shown in Fig. 2.7. As seen in the figure, to work with a 100 MHz sampling rate, 10-bit resolution requires 1 W of power consumption, whereas a 20-bit ADC would require 1 kW, which is an outcome in accord with the Walden FoM. Power consumption increases almost linearly with the sampling rate and exponentially with the number of bits. Since the aim is to achieve higher data rates, larger bandwidths, hence larger sampling rates, cannot be avoided. As a result, utilizing low-resolution ADCs can be a viable solution to deal with these issues.

Moreover, a pair of ADCs must be deployed in a massive MIMO setup in each RF chain. For example, according to Fig. 2.7, the total power dissipation of ADCs in a 100-antenna system employing a 1 MHz sampling rate would be 0.2 W with 8-bit ADCs, whereas the ADCs would consume around 200 W with 18-bit resolution. Therefore, the overall system's power consumption may reach undesirable levels in massive MIMO systems when high-resolution ADCs are used.

Note that $1-4$ bits of resolution would still be considered low because commercial systems currently employ ADCs with resolutions higher than 10 bits. For this thesis, we focus on only one-bit ADCs since their implementation is straightforward with a single comparator unit. One operational amplifier (OPAMP) is sufficient to realize a

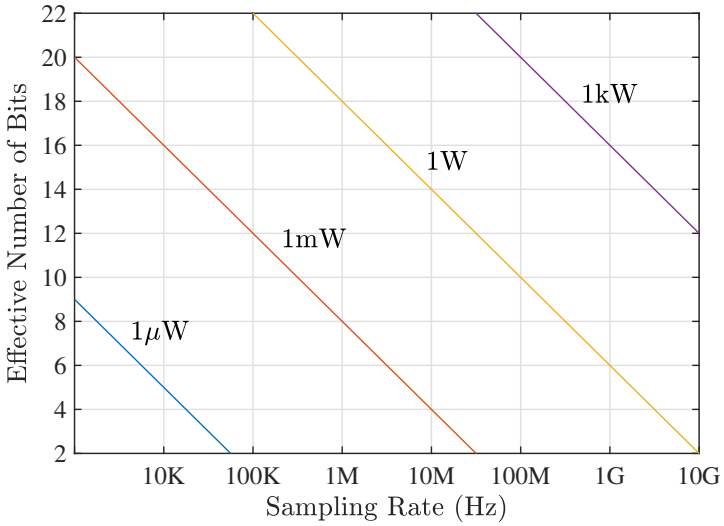


Figure 2.7: Resolution plotted against sampling rate of an ADC with constant power consumption curves from [1, Fig. 2], using the Walden FoM.

one-bit quantizer. Also, even though it is not visible in Fig. 2.6, an automatic gain control (AGC) unit must be deployed within a multi-bit ADC unit in practice since the total received power can vary in time which will require an adaptive selection of the quantizer step size Δ. On the other hand, one-bit ADCs do not require AGCs, with a single threshold located at the origin. Due to such additional benefits, our focus will be on one-bit quantization in this thesis.

CHAPTER 3

DETECTION UNDER FREQUENCY-FLAT FADING

3.1 Motivation

In this chapter, we work on detection under frequency-flat fading in uplink one-bit massive MIMO systems. After obtaining the detailed DT system model for which the fundamentals were given in Chapter 2, we focus on deriving the linear filters and the ML detector for the non-zero one-bit quantization scenario. To obtain accurate filtering methods, we utilize the Bussgang decomposition [52, 53] and obtain both Bussgang-based and conventional quantization-unaware filters. Such classification of filtering approaches was also used in [2, 23].

After briefly explaining the ML detector, we move on to our proposal for a new two-stage detection scheme for one-bit massive MIMO systems. The first stage is called Boxed Newton Detector (BND), which utilizes Newton's method with box constraints to optimize the log-likelihood and obtain an equalizer. The second stage is called Nearest Codeword Detector (NCD), which utilizes the output of a first-stage equalizer such as BND or linear filters to create a set of most likely candidates by taking the one-bit quantized observations as codewords in space. Then, the second-stage decisions are made by selecting the candidate in the reduced set that maximizes the likelihood. Such two-stage approaches are present in the literature. In [34], the first stage is based on the gradient descent algorithm, whereas [3] utilizes the support vector machine (SVM) approach to satisfy the sign constraints imposed by one-bit quantization, and [2] utilizes the deep unfolding method to optimize the log-likelihood for relatively faster convergence compared to [34]. For the second stage, [34] employs a nearest neighbor search algorithm whose complexity can be very large for high modulation orders and

large numbers of users. [2] tries to lower the second stage complexity by proposing a recursive algorithm to find a predefined number of candidates. [3] utilizes the work from [27] as the second stage, which proposes detection from a coding theoretic perspective by taking quantized observations as binary codewords in space.

In addition to a new detection scheme, we propose a new quantization scheme for uplink one-bit massive MIMO systems. The adverse effects of the SR phenomenon on the detection performance of uplink one-bit MIMO systems are encountered frequently in many studies [2, 3, 14, 19, 23, 24, 26]. Unlike unquantized systems, the performance is limited by quantization noise at high SNR. Influenced by pseudo-randomization of quantization for general signal processing applications in [30], we propose a new pseudo-random quantization scheme for one-bit uplink massive MIMO systems that can increase the achievable rate per user and support high order modulation schemes such as 256-QAM and 1024-QAM. We prefer to change the quantization thresholds instead of generating a dither signal with additional hardware to obtain an efficient system setup and thus pseudo-randomize the quantization operation in the spatial domain by relying on the large number of receiver antennas provided by the massive MIMO setup.

3.2 Contributions

The main contributions of this chapter can be summarized as:

- A novel PRQ scheme is proposed that can help mitigate the negative effects of SR. The proposed scheme relies on changing the quantization thresholds for dithering. The achievable rate in low-dimensional SIMO systems is shown to be increased with PRQ. Also, the minimum Hamming distance between the binary codewords at infinite SNR obtained via one-bit quantization in massive MIMO systems is shown to be increased with PRQ, which indicates how the proposed scheme can mitigate the effects of SR.
- As in [2, 23], Bussgang-based and conventional linear detectors are derived. However, different than [2, 23], linear detectors are modified for non-zero threshold one-bit quantization. With the appropriate scaling factor, conventional linear

receivers perform very closely to their Bussgang-based counterparts.

- A first-stage detector called the Boxed Newton Detector (BND) is proposed that relies on Newton's Method with box constraints to estimate the input. It outperforms the existing gradient-based detectors [2, 3, 34] in terms of error performance with comparable complexity. Also, it does not require hyper-parameter tuning like [2, 34] due to second-order derivative information.

- With a similar motivation as in [2, 3, 34], a second stage detector called the Nearest Codeword Detector (NCD) is proposed that creates a set of candidate vectors to refine the first stage solution with limited complexity. However, unlike [2], it does so by taking the one-bit observations as binary codewords to find a limited number of candidates based on the minimum Hamming distance criterion to lower the complexity and increase performance.

- By employing PRQ and BND-NCD, the proposed scheme can outperform the existing detectors in the literature with much lower error floors. Moreover, communication with high-order modulations such as 256-QAM, 1024-QAM, and 4096-QAM, whose performances were not reported by any of the previous works in the literature, is shown to be possible with PRQ. The proposed detector has better error performance compared to the state-of-the-art detectors from the literature, even when ZTQ is employed.

3.3 System Model

In this section, we build upon the system described in Section 2.2 and construct a vectorized notation to represent the DT baseband received signal. We consider an uplink massive MIMO system where K single-antenna users are served by a BS equipped with N antennas. A simple block diagram of the system model is shown in Fig. 3.1. Each user randomly selects an equally likely symbol from an M-QAM alphabet denoted by $\bar{\mathcal{M}}$. The vector of transmitted symbols from all users can be shown as $\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_K \end{bmatrix}^T$, where $\bar{x}_k \in \bar{\mathcal{M}}$, $\mathbb{E}[\bar{x}_k] = 0$, and $\mathbb{E}[|\bar{x}_k|^2] = E_s = 1$ for $k = 0, 1, \dots, K$. Users transmit their signals through a Nyquist pulse-shaping filter, an ideal DAC, and I/Q modulation.

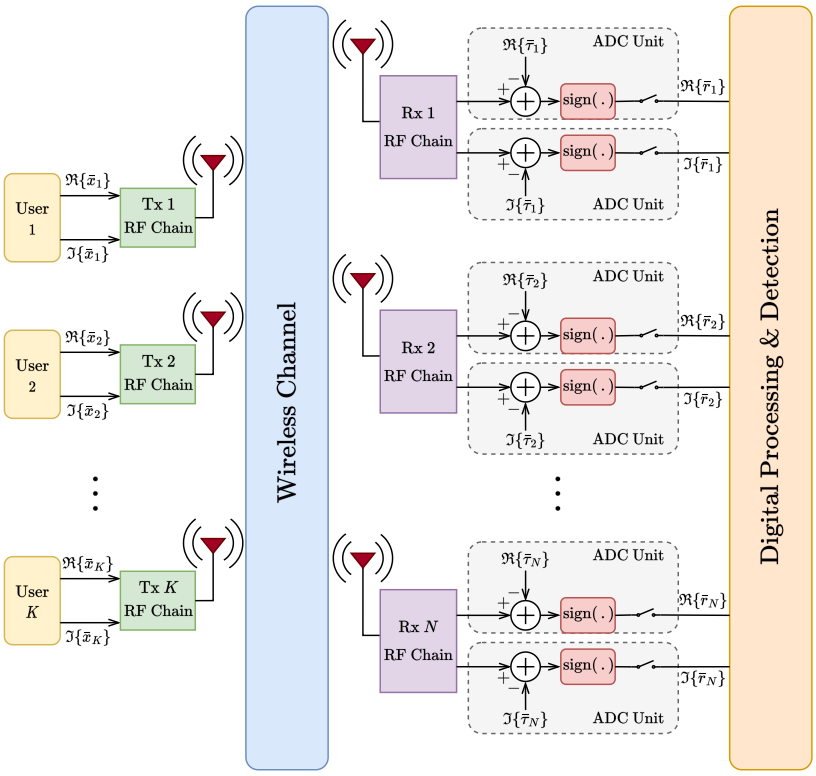


Figure 3.1: A block diagram that summarizes the single carrier (SC) system model for the flat-fading scenario.

We assume that the channel impulse response (CIR) between each user and receiver antenna is perfectly known by the BS and can be modeled with uncorrelated Rayleigh fading. Hence, CIR between receiver antenna n and user k has complex Gaussian distribution, i.e., $\bar{h}_{(n,k)} \sim \mathcal{CN}(0,1)$. The resultant $N \times K$ channel matrix is represented as $[\bar{\boldsymbol{H}}]_{(n,k)} = \bar{h}_{(n,k)}$. Assuming perfect synchronization, after the received signal is I/Q demodulated, pulse matched-filtered, and symbol-rate sampled, the unquantized discrete-time signal $\bar{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 & \bar{y}_2 & \dots & \bar{y}_N \end{bmatrix}^T$ is obtained as

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{H}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{w}}, \tag{3.1}$$

where $\bar{\boldsymbol{w}} = \begin{bmatrix} \bar{w}_1 & \bar{w}_2 & \dots & \bar{w}_N \end{bmatrix}^T$ is the zero-mean circularly symmetric Gaussian noise vector with PDF $\mathcal{CN}(\boldsymbol{0}_N, N_0\boldsymbol{I}_N)$. $\boldsymbol{I}_N$ is the identity matrix of size $N \times N$ and $\boldsymbol{0}_N$ is the all-zero vector of size N. The received signal is quantized by a pair of one-bit ADCs at each receiver antenna. The unquantized samples and the quantization

threshold vector $\bar{\boldsymbol{\tau}}$ can be used to obtain the quantized signal

$$\bar{\boldsymbol{r}} = \text{sign}(\Re\{\bar{\boldsymbol{y}} - \bar{\boldsymbol{\tau}}\}) + j \ \text{sign}(\Im\{\bar{\boldsymbol{y}} - \bar{\boldsymbol{\tau}}\}). \tag{3.2}$$

We define the SNR as the average SNR of each user at each receiver antenna:

$$\rho = \frac{\mathbb{E}[|\bar{x}_k|^2]}{\mathbb{E}[|\bar{w}_n|^2]} = \frac{E_s}{N_0} = \frac{1}{N_0}. \tag{3.3}$$

The relation between a complex vector $\bar{\boldsymbol{a}}$ and its real counterpart $\boldsymbol{a}$, and a complex matrix $\bar{\boldsymbol{A}}$ and its real counterpart $\boldsymbol{A}$ can be obtained as

$$\boldsymbol{a} = \begin{bmatrix} \Re\{\bar{\boldsymbol{a}}\} \\ \Im\{\bar{\boldsymbol{a}}\} \end{bmatrix} \text{ and } \boldsymbol{A} = \begin{bmatrix} \Re\{\bar{\boldsymbol{A}}\} & -\Im\{\bar{\boldsymbol{A}}\} \\ \Im\{\bar{\boldsymbol{A}}\} & \Re\{\bar{\boldsymbol{A}}\} \end{bmatrix}. \tag{3.4}$$

As a result, the overall input-output relation of the system can be expressed as

$$\boldsymbol{r} = \text{sign}(\boldsymbol{H}\boldsymbol{x} + \boldsymbol{w} - \boldsymbol{\tau}). \tag{3.5}$$

Note that each element of $\boldsymbol{x}$ belongs to the set of values a symbol from the modulation alphabet can take in one dimension. We denote this set as $\mathcal{M}$. For example, when $\bar{\mathcal{M}}$ is the 16-QAM alphabet, $\mathcal{M}$ can be expressed as the 4-PAM alphabet with an average power of $1/2$.

3.4 Linear Detection Methods

Linear detectors can be very advantageous in communication systems due to their low complexities. As in [2, 23], we define two classes of linear filters: Bussgang-based and conventional. In the following subsections, the maximum ratio combining (MRC) and zero-forcing (ZF) filters are derived for both classes.

3.4.1 Bussgang-Based Linear Filters

One-bit quantization introduces significant nonlinear distortion on the received signal. Therefore, the overall system is no longer linear. Since we deal with a nonlinear system, the Bussgang theorem is a handy tool to characterize the received quantized signal's first and second-order statistics. Since the original work in [52], the Bussgang theorem has been widely used to design and analyze nonlinear systems. Naturally, it is also used in the design and analysis of one-bit MIMO systems as in [2, 23, 35, 48, 54], among many others.

For one-bit MIMO systems, the theorem takes the form of Bussgang decomposition, where nonlinear distortion, i.e., quantization noise, takes an additive form and is represented linearly on top of the received signal. This linear representation minimizes the variance of quantization noise, and quantization noise becomes uncorrelated with quantized and unquantized observations. The conventional Bussgang decomposition focuses on one-bit quantization when all thresholds are set to zero, i.e., $\bar{\boldsymbol{\tau}} = \mathbf{0}_N$. In [53], many practical derivations and a thorough analysis of the Bussgang theorem are included.

Note that the Bussgang decomposition is valid only when the input to the quantizer has Gaussian distribution. However, since the constellation symbols are chosen from a discrete alphabet, the unquantized observation vector does not have Gaussian distribution. Despite not having Gaussian distribution, especially at low SNR when the noise term is dominant or due to the central limit theorem (CLT) when the number of users is high, the distribution of $\bar{\boldsymbol{y}}$ is very close to Gaussian [17, 19]. Therefore, applying the Bussgang decomposition for designing and analyzing low-resolution systems is a common practice [2, 17, 19, 23, 41, 53, 54]. Since we focus on a more general scenario where thresholds can take arbitrary values, the generalized version of the Bussgang theorem must be used.

In [47], the Bussgang theorem is generalized for the non-zero threshold quantization scenario by taking the selection process of the Bussgang gain and bias terms as the problem of finding the minimum mean square error (MMSE) estimate of the quantized observation vector $\bar{\boldsymbol{r}}$ using the unquantized observation $\bar{\boldsymbol{y}}$ such that

$$\bar{\boldsymbol{r}} - \bar{\boldsymbol{b}} = \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{y}} + \bar{\boldsymbol{d}}, \tag{3.6}$$

where $\bar{\boldsymbol{b}} = \mathbb{E}[\bar{\boldsymbol{r}} \mid \bar{\boldsymbol{\tau}}]$ is the bias vector, $\bar{\boldsymbol{g}} \in \mathbb{C}^N$ is the Bussgang gain vector which can also be represented as a diagonal matrix composed of the entries of this vector, i.e., the MMSE filter, and $\bar{\boldsymbol{d}} \in \mathbb{C}^N$ is the quantization noise vector. Note that unlike [47], we adopt a notation with complex numbers by defining a complex extension to the Hadamard product such that

$$\bar{\boldsymbol{u}} \,\bar{\odot}\, \bar{\boldsymbol{v}} = \Re\{\bar{\boldsymbol{u}}\} \odot \Re\{\bar{\boldsymbol{v}}\} + j\Im\{\bar{\boldsymbol{u}}\} \odot \Im\{\bar{\boldsymbol{v}}\}. \tag{3.7}$$

Using the derivations from [47], the Bussgang gain vector and the bias vector can be calculated respectively as

$$\bar{g}_n = \sqrt{\frac{4}{\pi[\bar{\boldsymbol{C}}_y]_{(n,n)}}} \left(\exp\left(-\frac{\Re\{\bar{\tau}_n\}^2}{[\bar{\boldsymbol{C}}_y]_{(n,n)}} \right) + j \exp\left(-\frac{\Im\{\bar{\tau}_n\}^2}{[\bar{\boldsymbol{C}}_y]_{(n,n)}} \right) \right), \tag{3.8}$$

$$\bar{b}_n = 2\Phi\left(-\frac{\Re\{\bar{\tau}_n\}}{\sqrt{[\bar{\boldsymbol{C}}_y]_{(n,n)}/2}} \right) - 1 + j\left(2\Phi\left(-\frac{\Im\{\bar{\tau}_n\}}{\sqrt{[\bar{\boldsymbol{C}}_y]_{(n,n)}/2}} \right) - 1 \right), \tag{3.9}$$

where $\bar{\boldsymbol{C}}_y = \mathbb{E}[\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^H] = \bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^H + N_0\boldsymbol{I}_N$ is the covariance matrix of the unquantized observation vector.

For ease of notation, we define the conditionally zero-mean version of the quantized observation vector such that

$$\begin{aligned} \bar{\boldsymbol{r}}_e = \bar{\boldsymbol{r}} - \bar{\boldsymbol{b}} &= \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{H}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{w}} + \bar{\boldsymbol{d}} \\ &= \bar{\boldsymbol{H}}_e\bar{\boldsymbol{x}} + \bar{\boldsymbol{w}}_e, \end{aligned} \tag{3.10}$$

where $\bar{\boldsymbol{H}}_e \in \mathbb{C}^{N\times K}$ is the effective channel matrix which takes the Bussgang gains into account and $\bar{\boldsymbol{w}}_e \in \mathbb{C}^N$ is the effective noise vector which is the combination of quantized thermal noise and quantization distortion.

Now that we defined the linearized version of the input-output relation, we can move on to define the Bussgang-based MRC (BMRC) filter as in the well-known unquantized MIMO setup as

$$\bar{\boldsymbol{F}}_{\mathrm{BMRC}} = \bar{\boldsymbol{\lambda}}^{e} \odot \bar{\boldsymbol{H}}_{e}^{H}, \tag{3.11}$$

where $\bar{\boldsymbol{\lambda}}^{e}$ is the scaling constant to make sure we obtain a non-scaled estimate, and it is determined as

$$\bar{\lambda}_{k}^{e} = \frac{1}{\sum_{n=1}^{N} |[\bar{\boldsymbol{H}}_{e}]_{(n,k)}|^{2}}, \tag{3.12}$$

for $k = 1, 2, \ldots, K$.

Next, the Bussgang-based ZF (BZF) filter can be defined as

$$\bar{\boldsymbol{F}}_{\mathrm{BZF}} = (\bar{\boldsymbol{H}}_{e}^{H} \bar{\boldsymbol{H}}_{e})^{-1} \bar{\boldsymbol{H}}_{e}^{H}. \tag{3.13}$$

Note that the MRC filter tries to maximize the SNR, whereas the ZF filter focuses on maximizing the signal-to-interference ratio (SIR). However, unlike unquantized massive MIMO, an additional quantization distortion term limits the high SNR performance. A more advanced approach would be calculating the MMSE filter for this scenario. Unfortunately, a compact analytical expression for the Bussgang-based MMSE (BMMSE) filter [45] can not be found as indicated in [47]. The unavailability of a compact analytical form is due to Arcsine law [55] not being valid for non-zero threshold quantization.

Once a filter $\bar{\boldsymbol{F}}$ to be applied on the conditionally zero-mean observation vector $\bar{\boldsymbol{r}}_{e}$ is selected, the estimates can be obtained as

$$\tilde{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{F}} \bar{\boldsymbol{r}}_{e}. \tag{3.14}$$

Finally, if no further processing is to be applied, decisions can be obtained by symbol-by-symbol detection, i.e., element-wise minimum distance mapping for $k =$

$1, 2, \ldots, K$ such that

$$\hat{\tilde{x}}_k = \arg\min_{\bar{x} \in \bar{\mathcal{M}}} |\tilde{\bar{x}}_k - \bar{x}|. \tag{3.15}$$

3.4.2 Conventional Linear Filters

The conventional class of linear filters for unquantized MIMO detection are well-known and well-studied [56]. Like the Bussgang-based filters, they can be used for quantized MIMO detection as in many studies such as [2,19,23,46,54]. However, since one-bit measurement causes the loss of amplitude information, inserting appropriate labels to our quantized observations is a helpful way to work with variable-amplitude modulation schemes.

Note that the Bussgang decomposition tries to find the appropriate scaling each element of the unquantized observation vector goes through during one-bit quantization. Finding a quantization label is similar to this operation, except that all scaling must be the same to define a common quantization label. Hence, we continue with the Bussgang decomposition and look for approximations that make all entries of $\bar{\boldsymbol{g}}_n$ the same in both the I and Q parts. To do so, we first assume the threshold values are unknown and $\bar{\boldsymbol{\tau}} \sim \mathcal{CN}(\mathbf{0}_N, \sigma_\tau^2)$. This can be a useful assumption if the empirical distribution of the thresholds is close to Gaussian. When we take the threshold values as random with Gaussian distribution, they can be seen as an additional form of noise over AWGN. Therefore the effective noise variance becomes $\sigma_e^2 = N_0 + \sigma_\tau^2$ since the thermal noise and threshold selections are independent. With this selection, we can also act as if $\bar{\boldsymbol{\tau}} = \mathbf{0}_N$. Now, the effective unquantized observation covariance matrix becomes $\bar{\boldsymbol{C}}_{y,e} = \bar{\boldsymbol{H}}\bar{\boldsymbol{H}}^H + (N_0 + \sigma_\tau^2)\boldsymbol{I}_N$.

Finally, the wireless channel's effect must be considered to obtain a label that is the same for all branches. As stated in [45], at low SNR or when the number of users is large, the unquantized observation covariance matrix can be approximated by a diagonal matrix such that $\bar{\boldsymbol{C}}_{y,e} \cong (K + N_0 + \sigma_\tau^2)\boldsymbol{I}_N$. By using these approximations, the bias vector becomes $\bar{\boldsymbol{b}} = \mathbf{0}_N$ and each element of the Bussgang gain vector can be approximated as $\bar{g}_n \cong \sqrt{\frac{4}{\pi(K+N_0+\sigma_\tau^2)}}(1+j)$. Since the I and Q parts are now the

same, the input-output relation can be simplified as

$$\bar{\boldsymbol{r}}_e = \bar{\boldsymbol{r}} = \sqrt{\frac{4}{\pi(K + N_0 + \sigma_\tau^2)}}\bar{\boldsymbol{y}} + \bar{\boldsymbol{d}}. \tag{3.16}$$

By defining the quantization label ℓ_q as

$$\ell_q = \sqrt{\frac{\pi(K + N_0 + \sigma_\tau^2)}{4}}, \tag{3.17}$$

and multiplying both sides of (3.16) with ℓ_q, we can get

$$\begin{aligned} \ell_q \bar{\boldsymbol{r}} &= \bar{\boldsymbol{y}} + \ell_q \bar{\boldsymbol{d}} \\ &= \bar{\boldsymbol{H}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{w}} + \ell_q \bar{\boldsymbol{d}}. \end{aligned} \tag{3.18}$$

Using this quantization label, a non-scaled estimate of the $\bar{\boldsymbol{x}}$ vector can now be obtained when conventional linear filters are used. Finding the conventional MRC and ZF filters is now straightforward. The MRC filter can be calculated as

$$\bar{\boldsymbol{F}}_{\mathrm{MRC}} = \bar{\boldsymbol{\lambda}} \odot \bar{\boldsymbol{H}}^H, \tag{3.19}$$

where the elements of scaling vector $\bar{\boldsymbol{\lambda}}$ for $k = 1, 2, \ldots, K$ can be found as

$$\bar{\lambda}_k[v] = \frac{1}{\sum_{n=1}^{N} |[\bar{\boldsymbol{H}}_{(n,k)}|^2}. \tag{3.20}$$

Similarly, the ZF filter can be calculated as

$$\bar{\boldsymbol{F}}_{\mathrm{ZF}} = (\bar{\boldsymbol{H}}^H \bar{\boldsymbol{H}})^{-1} \bar{\boldsymbol{H}}^H. \tag{3.21}$$

Our motivation here is to introduce appropriate quantization labels to obtain a non-scaled estimate which can be critical while working with high-order modulation schemes such as 16-QAM or 64-QAM. BPSK or QPSK modulations do not require such a label since they have constant amplitude, and a scaled version of the estimate

would not cause any problems during detection. Also, we omit the conventional MMSE filter for this scenario since it reportedly has the same performance as that of the ZF filter [2].

Once a filter of choice $\bar{\boldsymbol{F}}$ is selected, it is applied on the quantized observation vector $\bar{\boldsymbol{r}}$ to obtain the estimates as

$$\tilde{\bar{\boldsymbol{x}}} = \ell_q \bar{\boldsymbol{F}} \bar{\boldsymbol{r}}. \tag{3.22}$$

Lastly, if no further processing is applied, decisions can be obtained as in (3.15).

3.5 Maximum Likelihood (ML) Detection

When the receiver perfectly knows CSI, the detection operation can be conducted using the likelihood function. Likelihood-based detectors have an important advantage against linear detectors in terms of bit error rate (BER) performance, especially at high SNR [2]. As explained in different works such as [2, 32, 34], the conditional probability mass function (PMF) of the quantized observations can be expressed as

$$\mathrm{p}(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\tau}, \boldsymbol{H}) = \prod_{n=1}^{2N} \Phi\left(\frac{r_n(\boldsymbol{h}_n^T \boldsymbol{x} - \tau_n)}{\sqrt{N_0/2}}\right), \tag{3.23}$$

where $\boldsymbol{h}_n^T$ is the n^{th} row of the channel matrix $\boldsymbol{H}$, i.e., $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \dots & \boldsymbol{h}_{2N} \end{bmatrix}^T$, and $\Phi(x)$ is the CDF of the standard Gaussian random variable. Then, by utilizing the log-likelihood, we can construct the ML detector as

$$\hat{\boldsymbol{x}}_{\text{ML}} = \underset{\boldsymbol{x} \in \mathcal{M}^{2K}}{\arg\max} \left\{ \mathbf{1}_N \ln\left(\Phi\left(\sqrt{\frac{2}{N_0}} \boldsymbol{r} \odot (\boldsymbol{H}\boldsymbol{x} - \boldsymbol{\tau})\right)\right)\right\}, \tag{3.24}$$

where the natural logarithm $\ln(.)$ and $\Phi(.)$ are applied element-wise to their arguments, and $\mathbf{1}_N$ is the column vector of size N which is composed of entries that are all equal to 1.

3.6 Proposed Detection Method: BND-NCD

3.6.1 First Stage: Boxed Newton Detector (BND)

When the modulation order M or the number of users K is large, complexity becomes infeasible for ML detection. Different works in the literature focus on a gradient-based approach to maximize the log-likelihood function [2, 24, 32, 34]. In this subsection, a similar procedure is followed with a new iterative receiver that uses a modified cost function and the Hessian information for faster convergence via Newton's method. For compactness, we denote the log-likelihood as

$$\mathcal{L}(\boldsymbol{x}) = \mathbf{1}_N \ln\left(\Phi\left(\sqrt{\frac{2}{N_0}}\boldsymbol{r} \odot (\boldsymbol{H}\boldsymbol{x} - \boldsymbol{\tau})\right)\right). \tag{3.25}$$

To begin with, we relax the constraint that $\boldsymbol{x}$ belongs to a finite input set and reformulate the ML detection problem as

$$\tilde{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathbb{R}^{2K}}{\arg\max}\left\{\mathcal{L}(\boldsymbol{x})\right\}. \tag{3.26}$$

Gradient-based optimization can be utilized for this problem due to the log-concavity of $\Phi(.)$. However, this approach is not helpful at high SNR since each entry of the input vector can take any value from the real numbers, which causes an increase in the saddle points of the log-likelihood function. To prevent diverging from the boundaries of the discrete constellation set, we insert a box constraint on each element of $\boldsymbol{x}$, as in [4]. For M-QAM, the in-phase or quadrature part of the symbol sent from any of the users must satisfy

$$|x_k| \leq M_b = \sqrt{\frac{3(\sqrt{M}-1)^2}{2(M-1)}}, \text{ for } k = 1, 2, \ldots, 2K, \tag{3.27}$$

where M_b denotes the constellation boundary on a single dimension. The new constrained optimization problem can be written as

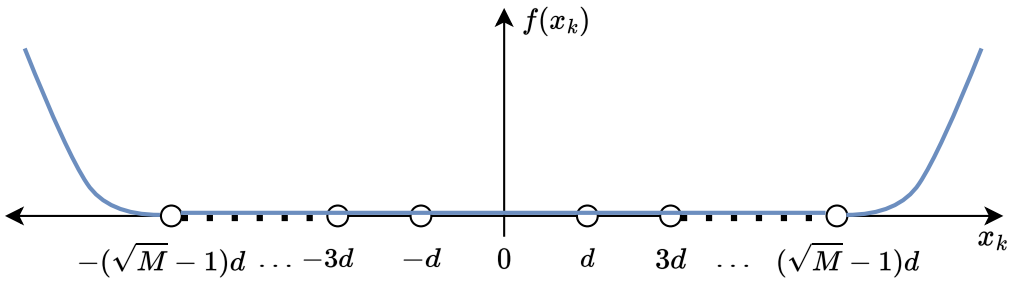


Figure 3.2: An illustration of the penalty function that enforces the box constraint on each element of the $\boldsymbol{x}$ vector during the iterative updates of BND for M-QAM constellation, where $d = \sqrt{\frac{3}{2(M-1)}}$.

$$\tilde{\boldsymbol{x}} = \underset{\substack{|x_k| \leq M_b \\ k=1,2,\ldots,2K}}{\arg\max}\ \mathcal{L}(\boldsymbol{x}). \tag{3.28}$$

Optimization with iterative approaches that rely on the gradient of the objective function requires the problem to be converted back to unconstrained optimization. Therefore, a penalty function is inserted into the objective function, which can be written as

$$\mathcal{P}(\boldsymbol{x}) = \frac{\theta}{2} \sum_{k=1}^{2K} \mathrm{R}(|x_k| - M_b)^2, \tag{3.29}$$

where $\theta \in \mathbb{R}$ is a constant that determines the strength of box constraints, and $\mathrm{R}(x) = \max\{0, x\}$ is the unit-ramp function. If we denote the penalty function applied on a single element in the set that reflects the box constraint as $f(x_k) = \mathrm{R}\{|x_k| - M_b\}^2$, we can visualize it as shown in Fig. 3.2. Hence, the final form of the unconstrained optimization problem to be solved can be obtained as

$$\tilde{\boldsymbol{x}} = \underset{\boldsymbol{x} \in \mathbb{R}^{2K}}{\arg\max}\{\mathcal{L}(\boldsymbol{x}) - \mathcal{P}(\boldsymbol{x})\}. \tag{3.30}$$

By denoting the argument of (3.30) as our cost function $\mathcal{J}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) - \mathcal{P}(\boldsymbol{x})$, the iterative update equation of Newton's Method becomes

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - (\nabla^2 \mathcal{J}(\boldsymbol{x}^k))^{-1} \nabla \mathcal{J}(\boldsymbol{x}^k), \tag{3.31}$$

where $\nabla = \begin{bmatrix} \frac{\partial}{\partial x_1} & \dots & \frac{\partial}{\partial x_{2K}} \end{bmatrix}^T$ can be expressed as a vector of size $2K$, and ∇^2 can be expressed as the outer product of the ∇ operator with itself: $\nabla^2 = \nabla \nabla^T$. Hence, the ∇^2 operator can be considered a $2K \times 2K$ matrix. $\nabla \mathcal{J}(\boldsymbol{x}^k)$ is the gradient of the cost function $\mathcal{J}$ with respect to $\boldsymbol{x}$ calculated at $\boldsymbol{x}^k$, i.e., the estimate of $\boldsymbol{x}$ at iteration k. Similarly, $\nabla^2 \mathcal{J}(\boldsymbol{x}^k)$ is the Hessian of the cost function $\mathcal{J}$ with respect to $\boldsymbol{x}$ calculated at $\boldsymbol{x}^k$.

We should go over the log-likelihood and penalty functions' first and second-order derivatives before expressing the gradient and Hessian information. The first-order derivatives of the log-likelihood function and the penalty function can be expressed as

$$\frac{d}{dx_k} \mathcal{L}(\boldsymbol{x}) = \sqrt{\frac{2}{N_0}} \sum_{n=1}^{2N} r_n [\boldsymbol{H}]_{(n,k)} \varphi\left(u_n\right), \tag{3.32}$$

$$\frac{d}{dx_k} \mathcal{P}(\boldsymbol{x}) = \theta \, \mathrm{sign}(x_k) \, \mathrm{R}(|x_k| - M_b), \tag{3.33}$$

respectively, for $k = 1, 2, \dots, 2K$, where $\boldsymbol{u} = \sqrt{\frac{2}{N_0}} \boldsymbol{r} \odot (\boldsymbol{H}\boldsymbol{x} - \boldsymbol{\tau})$, and $\varphi(x) = \frac{d}{dx} \ln(\Phi(x)) = \frac{\phi(x)}{\Phi(x)}$. Then the second-order derivatives can be obtained as

$$\frac{d^2}{dx_k dx_m} \mathcal{L}(\boldsymbol{x}) = \frac{2}{N_0} \sum_{n=1}^{2N} [\boldsymbol{H}]_{(n,k)} [\boldsymbol{H}]_{(n,m)} \psi\left(u_n\right), \tag{3.34}$$

$$\frac{d^2}{dx_k dx_m} \mathcal{P}(\boldsymbol{x}) = \theta \, \mathrm{U}(|x_k| - M_b) \delta[k - m], \tag{3.35}$$

for $k = 1, 2, \dots, 2K$ and $m = 1, 2, \dots, 2K$, where $\psi(x) = \frac{d^2}{dx^2} \ln(\Phi(x)) = -x\varphi(x) - \varphi^2(x)$ and $\mathrm{U}(x) = \max\{0, \mathrm{sign}(x)\}$ is the unit-step function. $psi(.)$ is the PDF of the standard Gaussian random variable.

Due to linearity of differentiation, we have $\nabla \mathcal{J}(\boldsymbol{x}) = \nabla \mathcal{L}(\boldsymbol{x}) - \nabla \mathcal{P}(\boldsymbol{x})$. As in [32], the gradient of the log-likelihood function can be found as

$$\nabla \mathcal{L}(\boldsymbol{x}) = \sqrt{\frac{2}{N_0}} \boldsymbol{H}^T \left(\boldsymbol{r} \odot \varphi \left(\boldsymbol{u} \right) \right), \tag{3.36}$$

where $\odot$ denotes the Hadamard product, and $\varphi(.)$ is applied element-wise to its arguments. Then, the gradient of the penalty function can be found as

$$\nabla \mathcal{P}(\boldsymbol{x}) = \theta \, \mathrm{sign}(\boldsymbol{x}) \odot \mathrm{R}(|\boldsymbol{x}| - M_b \mathbf{1}_{2K}), \tag{3.37}$$

where $\mathrm{R}(.)$ is applied element-wise to its arguments. Similarly, we have $\nabla^2 \mathcal{J}(x) = \nabla^2 \mathcal{L}(x) - \nabla^2 \mathcal{P}(x)$. The Hessian of the log-likelihood function or the Fisher information matrix can be calculated as

$$\nabla^2 \mathcal{L}(\boldsymbol{x}) = \frac{2}{N_0} \boldsymbol{H}^T \, \mathrm{diag} \left(\psi \left(\boldsymbol{u} \right) \right) \boldsymbol{H}, \tag{3.38}$$

where $\psi(.)$ is applied element-wise to its arguments. Since $\Phi(x)$ can approach zero exponentially fast, computations of $\ln(\Phi(x))$, $\varphi(x)$, and $\psi(x)$ in finite precision can cause problems such as divergent behavior or uncertainties. This problem is solved in Appendix A. Finally, the Hessian of the penalty function is

$$\nabla^2 \mathcal{P}(\boldsymbol{x}) = \theta \, \mathrm{diag}(\mathrm{U}(|\boldsymbol{x}| - M_b \mathbf{1}_{2K})), \tag{3.39}$$

where $\mathrm{U}(.)$ is applied element-wise to its arguments.

Now that the update rule is wholly defined, an initial solution $\boldsymbol{x}^0$ that is preferably not far away from the final solution should be found. The MRC estimate is a suitable selection for this purpose, and it can easily be found using (3.19) and (3.22). Note that at high SNR, when N_0 is very small, the Hessian matrix can become very close to singular since $\psi(x) \to 0$ as $x \to \infty$. To avoid such behavior, we define a damping factor ζ such that

$$\zeta = \max\{1, \rho / \rho_t\}, \tag{3.40}$$

Algorithm 1 Boxed Newton Detector (BND)

Input: $\boldsymbol{r}, \boldsymbol{H}, \boldsymbol{\tau}, \theta, \epsilon, M_b, T_{\text{max}}, N_0, \zeta$

Output: $\tilde{\boldsymbol{x}}$

1: Set the initial solution to the MRC estimate $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}^0$ using (3.19)
2: Apply the damping factor $N_0 \leftarrow \zeta N_0$
3: **for** $i = 1$ to T_{max} **do**
4: Calculate the gradient $\nabla \mathcal{J}(\tilde{\boldsymbol{x}})$ using (3.36) and (3.37)
5: Calculate the Hessian $\nabla^2 \mathcal{J}(\tilde{\boldsymbol{x}})$ using (3.38) and (3.39)
6: Calculate the step $\Delta \boldsymbol{x} \leftarrow (\nabla^2 J(\tilde{\boldsymbol{x}}))^{-1} \nabla J(\tilde{\boldsymbol{x}})$
7: Iterative update $\tilde{\boldsymbol{x}} \leftarrow \tilde{\boldsymbol{x}} - \Delta \boldsymbol{x}$ as in (3.31)
8: **if** $\|\Delta \boldsymbol{x}\|^2 < \epsilon \|\boldsymbol{x}\|^2$ **then**
9: **return** $\tilde{\boldsymbol{x}}$
10: **end if**
11: **end for**
12: **return** $\tilde{\boldsymbol{x}}$

where ρ_t is the threshold SNR. Before starting the iterative updates, N_0 is multiplied with this term to provide numeric stability. If the SNR is above the threshold, we process the signals as if the SNR is at the threshold level, and if the SNR is below the threshold, we operate with the actual SNR value. The complete procedure for BND is summarized in Algorithm 1. Once the iterative updates start, the algorithm is terminated if the maximum number of iterations T_{max} is reached or if further iterations do not cause significant changes, which is determined by the termination threshold ϵ. If a maximum exists, we expect each step's norm to get smaller at each iteration. Hence at each iteration, we measure how close we are to the maximum, as in line 8, and if we are close enough, the algorithm is terminated. The output of the algorithm $\tilde{\boldsymbol{x}}$ is the first-stage solution, and if a one-stage approach is to be followed, then symbol-by-symbol detection is applied on the estimate as in (3.15). If not, the estimate is supplied to the second stage for further processing.

3.6.2 Second Stage: Nearest Codeword Detector (NCD)

After finding an estimate using a first-stage method, NCD can be utilized to make more

Algorithm 2 Nearest Codeword Detector (NCD)

Input: $\tilde{\boldsymbol{x}}, \boldsymbol{r}, \boldsymbol{H}, \boldsymbol{\tau}, P, \gamma, U_{\max}, N_0$

Output: $\hat{\boldsymbol{x}}$

Find the nearest decision boundaries $\boldsymbol{t}$ as in (3.42)
Obtain the sets of reliable $\mathcal{R}$ (3.43) and unreliable $\mathcal{U}$ (3.44) indices
while $|\mathcal{U}| > U_{\max}$ **do**
    Decrease the size of the unreliable region as $\gamma \leftarrow 0.95\gamma$
    Reobtain the sets $\mathcal{R}$ and $\mathcal{U}$ according to the new γ
end while
Find the candidate element sets $(\mathcal{X}_k)_{k=1}^{2K}$ using (3.45)
Generate the candidate vector set $\mathcal{X}$ using (3.46)
if $|\mathcal{X}| > P$ **then**
    Apply symbol-by-symbol detection to get
    $\check{x}_k \leftarrow \arg\min_{x \in \mathcal{M}} |\tilde{x}_k - x|$ for $k = 1, 2, \ldots, 2K$
    Remove the symbol-by-symbol detected vector from the set $\mathcal{X} \leftarrow \mathcal{X} \setminus \{\check{\boldsymbol{x}}\}$
    if $|\mathcal{X}| > 1$ **then**
        Sort $\mathcal{X}$ according to the $\mathscr{H}(.)$ metric as in (3.50)
        Discard all elements of $\mathcal{X}$ except the first $P - 1$
    end if
    Add $\check{\boldsymbol{x}}$ back to the set to obtain the result from (3.51) $\mathcal{X} \leftarrow \mathcal{X} \cup \{\check{\boldsymbol{x}}\}$
end if
Apply ML detection on the set $\mathcal{X}$
return $\tilde{\boldsymbol{x}}$

accurate decisions compared to symbol-by-symbol detection. In this part, the first step is to decide on the reliability of each element of the first-stage estimate. Then, a set of candidate vectors is formed based on the reliability information of each element, very similar to the ideas from [2, 3, 34]. The candidate set is then narrowed down based on the minimum Hamming distance criterion in the codeword domain. Finally, ML detection is conducted on the smaller candidate set to make the final decisions. The complete summary of the proposed second stage detector is made in Algorithm 2, and the detailed explanation of the whole procedure is made in the following parts.

The set of decision boundaries utilized during symbol-by-symbol detection for M-

QAM constellations is found as

$$\mathcal{T} = \left\{ \pm n \sqrt{\frac{6}{M-1}} \,\middle|\, n \in \left\{ 0, 1, \ldots, \frac{\sqrt{M}-2}{2} \right\} \right\}. \tag{3.41}$$

For example, the decision boundary set for the QPSK constellation is $\{0\}$, and for 16-QAM constellation, it is $\left\{0, \pm 2\sqrt{\frac{1}{10}}\right\}$. Then, the closest decision boundary to the k^{th} element of the estimate $\tilde{\boldsymbol{x}}$ is defined as

$$t_k = \arg\min_{t \in \mathcal{T}} |\tilde{x}_k - t|, \tag{3.42}$$

for $k = 1, 2, \ldots, 2K$. The set of reliable and unreliable indices of the estimate can be found as

$$\mathcal{R} = \{k \mid k \in \{1, 2, \ldots, 2K\},\, |\tilde{x}_k - t_k| > \gamma\}, \tag{3.43}$$

$$\mathcal{U} = \{k \mid k \in \{1, 2, \ldots, 2K\} \setminus \mathcal{R}\}, \tag{3.44}$$

respectively. If the k^{th} element of the estimate is within a certain margin of its closest decision boundary defined by the hyperparameter $\gamma \in \mathbb{R}$, then it is not reliable to conduct symbol-by-symbol detection on $\tilde{x}_k$, for $k = 1, 2, \ldots, 2K$. Now, the set of possible assignments for each element of the estimate can be found as

$$\tilde{\mathcal{X}}_k = \begin{cases} \left\{ \arg\min_{x \in \mathcal{M}} |x_k - x| \right\}, & k \in \mathcal{R} \\ \left\{ t_k \pm \sqrt{\frac{3}{2(M-1)}} \right\}, & k \in \mathcal{U} \end{cases}, \tag{3.45}$$

for $k = 1, 2, \ldots, 2K$. If the k^{th} element of the estimate is reliable, then there is only one possible assignment to this element obtained by the minimum distance rule. If it is unreliable, the potential assignment is a set composed of two elements that are the neighbors to the closest decision boundary. Hence, each set's cardinality can be 1

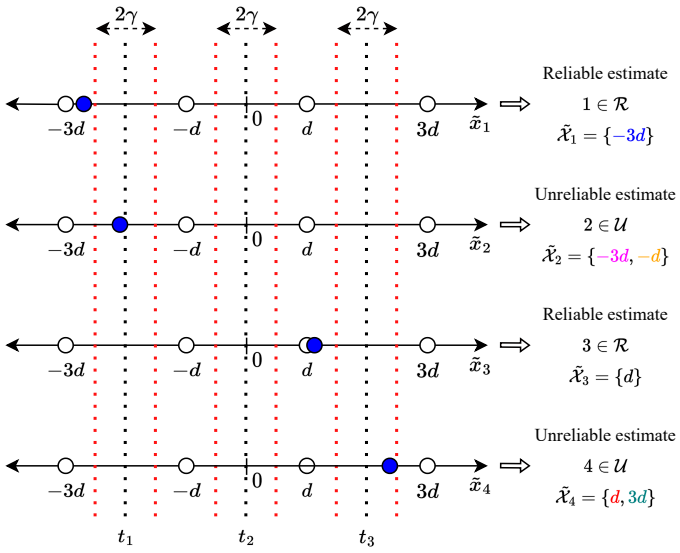


Figure 3.3: The candidate set formation example when $K = 2$ and $M = 16$, where $d = \sqrt{\frac{3}{2(M-1)}}$. The blue dots correspond to the estimates obtained from a first-stage detector. Each interval between the red dashed lines is mapped as unreliable, and outside of these intervals is mapped as reliable. The element sets $\tilde{\mathcal{X}}_k$ are formed as in (3.45), and the resultant example vector set obtained via (3.46) is shown in (3.47).

or 2. As in [2], the resultant set of candidate vectors from the combinations of each candidate element set can be obtained as

$$\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_1 \times \tilde{\mathcal{X}}_2 \times \ldots \times \tilde{\mathcal{X}}_{2K}, \tag{3.46}$$

where $\times$ denotes the Cartesian product. An example of the set formation is shown in Fig. 3.3, and the resultant color-coded set $\tilde{\mathcal{X}}$ can be shown as

$$\tilde{\mathcal{X}} = \left\{ \begin{bmatrix} -3d \\ -3d \\ d \\ d \end{bmatrix}, \begin{bmatrix} -3d \\ -3d \\ d \\ 3d \end{bmatrix}, \begin{bmatrix} -3d \\ -d \\ d \\ d \end{bmatrix}, \begin{bmatrix} -3d \\ -d \\ d \\ 3d \end{bmatrix} \right\}. \tag{3.47}$$

The cardinality of the candidate vector set $|\tilde{\mathcal{X}}|$ is at most 2^{2K}, which means the set can grow exponentially as the number of users increases, which can cause two problems. First, the size of the set may get too large when $|\mathcal{U}|$ is large, which can cause memory

problems. Also, a large candidate set would cause an undesirable complexity increase for ML detection at the final step. The first problem is dealt with adaptively changing γ, which can adjust the size of the unreliable region, and it is addressed in lines 3-6 of Algorithm 2. If the size of $\mathcal{U}$ is large, i.e., it is greater than $U_{\max}$, then decreasing γ can help us obtain a smaller unreliable region, hence a smaller set size for $\tilde{\mathcal{X}}$. For the second problem, similar to the idea from [2], the aim is to limit the search complexity by finding a subset $\mathcal{X}$ of $\tilde{\mathcal{X}}$ that includes the most likely candidates. In [34] and [3], the search complexity is not limited in the second stage, and in [2], a limited number of nearest neighbors to the estimate from the candidate set $\tilde{\mathcal{X}}$ are found. However, especially at high SNR, detection performance can benefit from the sign constraints imposed by one-bit quantization. Since quantized observations take binary values, $\boldsymbol{r}$ can easily be seen as a codeword in the spatial domain. In [27], this idea is exploited by a coding theoretic approach, and a new metric called weighted Hamming distance is proposed for detection. Since a first-stage solution already exists, a sophisticated metric such as the weighted Hamming distance is not required in this case. Hence, the size of the set of candidate vectors can be limited by finding their codeword representations, then ordering them according to their Hamming distance to the actual quantized observation vector. The spatial codeword representation $c(.)$ of a candidate vector $\acute{\boldsymbol{x}} \in \tilde{\mathcal{X}}$ is defined as

$$\begin{aligned} c(\acute{\boldsymbol{x}}) &= \operatorname{sign}\left(\mathbb{E}[\,\acute{\boldsymbol{y}} - \boldsymbol{\tau} \mid \acute{\boldsymbol{x}}, \boldsymbol{H}, \boldsymbol{\tau}\,]\right) \\ &= \operatorname{sign}\left(\boldsymbol{H}\acute{\boldsymbol{x}} - \boldsymbol{\tau}\right), \end{aligned} \tag{3.48}$$

where $\acute{\boldsymbol{y}}$ is the unquantized observation vector obtained when $\acute{\boldsymbol{x}}$ is sent. The Hamming distance between $\boldsymbol{r}$ and the spatial codeword representation of a selected vector $\acute{\boldsymbol{x}}$ can be expressed as

$$\mathscr{H}(\acute{\boldsymbol{x}}) = \mathrm{d_H}(\boldsymbol{r}, c(\acute{\boldsymbol{x}})), \tag{3.49}$$

where $\mathrm{d_H}\{\,.\,,\,.\,\}$ denotes the Hamming distance between its arguments. Now, the aim is to create a set $\mathcal{X} \subseteq \tilde{\mathcal{X}}$ whose cardinality is a hyperparameter denoted as P. It consists of the symbol-by-symbol detected vector $\check{\boldsymbol{x}}$ whose k^{th} element is

$\check{x}_k = \arg\min_{x \in \mathcal{M}} |\tilde{x}_k - x|$ for $k = 1, 2, \ldots, 2K$, and the $P-1$ candidate vectors from $\tilde{\mathcal{X}}$ such that the Hamming distance between their spatial codeword representations and $\boldsymbol{r}$ are the smallest. Including the symbol-by-symbol detection result as a candidate is enforced due to favorable performance, especially at low SNR. Next, to sort the candidate vectors according to the Hamming distance between their spatial codeword representations and the quantized observation vector, an ordered vector sequence is defined as

$$(\acute{\boldsymbol{x}}_i)_{i=1}^{|\mathcal{A}|} \in \tilde{\mathcal{X}} \setminus \{\check{\boldsymbol{x}}\} \text{ such that } \mathscr{H}(\acute{\boldsymbol{x}}_i) \leq \mathscr{H}(\acute{\boldsymbol{x}}_{i+1}). \tag{3.50}$$

Sequence indexing is a useful way to find the first $P-1$ candidates. Indexing may not be unique, and encountering a situation such as $\mathscr{H}(\acute{\boldsymbol{x}}_{P-1}) = \mathscr{H}(\acute{\boldsymbol{x}}_P)$ is possible. However, the performance gain obtained by the second stage increases as the SNR increases, and for large enough K, a situation such as $\mathscr{H}(\boldsymbol{x}) = \mathscr{H}(\acute{\boldsymbol{x}}_{P-1})$ is not likely. Hence, another approach to sort the candidates whose Hamming distances between their spatial codeword representations and $\boldsymbol{r}$ are equal to $\mathscr{H}(\acute{\boldsymbol{x}}_{P-1})$, is not necessary. The final set $\mathcal{X}$ of size P can be found as

$$\mathcal{X} = \{\acute{\boldsymbol{x}}_1, \acute{\boldsymbol{x}}_2, \ldots, \acute{\boldsymbol{x}}_{P-1}\} \cup \{\check{\boldsymbol{x}}\}. \tag{3.51}$$

Finally, ML detection (3.24) is applied on the reduced set $\mathcal{X}$ instead of $\mathcal{M}^{2K}$ to obtain the second-stage solution.

3.7 Proposed Quantization Method: Pseudo-Random Quantization (PRQ)

3.7.1 Stochastic Resonance and Dithering

SR is a counter-intuitive effect in nonlinear systems where the maximum performance is achieved at a finite SNR value in contrast to linear systems where higher SNR leads to better performance [21]. When the signal to be detected is not distinguishable by a sensor, the presence of white noise in a nonlinear system's input can help increase the detectability of the input signal [57]. Therefore, this phenomenon is

of multidisciplinary interest and can be encountered in electronics, biology, physics, and neuroscience. Electronic circuitry with thresholds [58], bistable and multistable dynamical systems [59], image processing and medical imaging [60], biological neurons [61], and artificial neural networks [62, 63] are examples of systems where SR is observed.

When the SNR of a nonlinear system is high, the intentional addition of noise into the system's input to reduce quantization errors is called dithering. SR and dithering are closely related, and dithering can be seen as a method for exploiting the SR phenomenon [29]. Generating artificial noise with a particular distribution is one way of dithering the input signal. However, the same dithering effect can be achieved by shifting the quantization thresholds. Thus, dithering can also be called randomized quantization, with both random [28] and pseudo-random [30] applications.

Even though there are not many studies regarding dithering in one-bit MIMO systems, there are some notable works related to non-zero threshold quantization. The idea of non-zero thresholds is also exploited in [47], where a channel estimation procedure with threshold design based on a set partitioning scheme is proposed, which requires computing and updating threshold values during channel estimation. Another channel estimation method is proposed in [41], where an adaptive threshold design procedure is utilized to minimize the estimation error. The thresholds are changed at consecutive symbol periods to increase estimation accuracy in both [47] and [41]. Recent work in [64] proposes a hybrid scheme that utilizes analog processing and adaptive quantization thresholds to maximize the achievable rate.

3.7.2 Pseudo-Random Quantization (PRQ) Scheme

Randomizing the quantization process can be a handy tool to mitigate quantization noise in low-resolution systems [28, 30]. Randomized quantization exploits dithering, where a randomly generated analog signal is added to/subtracted from the input to the quantizer. Suppose the randomly generated signal is kept in the memory of the digital processor. In that case, the procedure is pseudo-random (PRQ) since the operation is entirely deterministic. Otherwise, it is named random (RQ). Generating analog signals requires additional hardware, which may not be desirable. Therefore, a preferable

alternative is to change the quantization threshold of each ADC unit, which will result in the same dithering effect as generating an analog signal and subtracting it from the input signal. We change the domain of dithering from temporal (time) to spatial (antennas) with this approach.

In Fig. 3.4, the first version of dithering is illustrated where a randomly generated analog signal is subtracted from the incoming analog signal with the help of an additional DAC. PRNG is a pseudo-random number generator and helps obtain random sequences from a selected probability distribution. As seen in the figure, the advantage of PRQ is that we can utilize threshold values during the detection operation. However, employing an additional DAC in the RF chain can be undesirable since it increases the cost and power consumption of the system.

To obtain a more efficient system, we can rely on variable-threshold quantizers to help get the same effect as generating the dither signal and subtracting it from the received signal. The resulting setup is shown in Fig. 3.5. As seen in the figure, we rely on changing the quantization threshold instead of using an additional DAC. However, in practice, the response time of the ADC can become a burden if the threshold is varied at every symbol period. Hence, keeping the thresholds constant during channel coherence time would be preferable. Also, since we deal with massive MIMO systems,

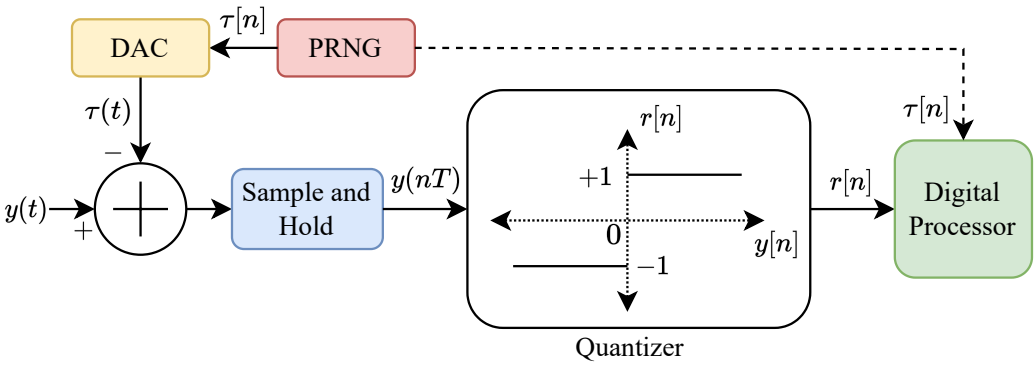


Figure 3.4: Illustrations of random (without the dashed line connection) and pseudo-random (with the dashed line connection) quantization schemes where the dither signal that is generated in the analog domain is subtracted from the incoming received signal, and the quantizer threshold is set to zero.

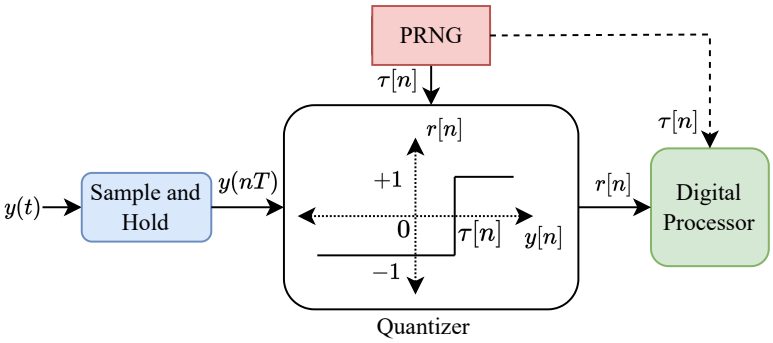


Figure 3.5: Illustrations of random (without the dashed line connection) and pseudo-random (with the dashed line connection) quantization schemes where the dithering effect is obtained by modifying the quantization thresholds. (It is assumed that $\tau[n] > 0$.)

randomization does not have to occur in time but can be utilized in space.

In the literature, different techniques for adaptive threshold optimization, such as [41,47], depend on the channel realization. However, these techniques require updating at different symbol periods or channel realizations. With the intuition we get from massive MIMO, we propose to randomize the spatial oversampling operation with the large antenna array equipped with one-bit ADCs. We propose a new quantization scheme for uplink one-bit massive MIMO systems to overcome the adverse effects of the SR phenomenon at high SNR, especially when the sum interference is small. From our previous observations, it is clear that the performance degradation does not occur at low SNR where thermal noise is the dominant distortion, which is reported by many other works from the literature [13, 54, 65]. However, at high SNR, either the performance peaks at a unique SNR value or a performance saturation occurs after a finite SNR.

When a dithering scheme is applied in the system, the distribution of the dither signal, the threshold values for our scenario, must be selected. By observing different scenarios, we saw that selecting the Gaussian distribution is a useful technique. Different selections, such as continuous uniform, and discrete uniform with binary or ternary alphabets, can also increase the high SNR performance. The selection of the appropriate distribution for the pseudo-random thresholds is also a noteworthy topic, which

is left as future work. In the context of this thesis, we focus on Gaussian distributed threshold values.

In general, one-bit massive MIMO systems are better at detecting phase than amplitude since the cardinality of the amplitude alphabet of the quantized observations is 1, but the cardinality of the phase alphabet is 4. Also, the fading channel introduces uniformly distributed phase shifts between all users and receiver antennas. Even though the channel also introduces amplitude variations, one-bit quantization causes a black/white or -1/+1 situation where the message to be decoded is a gray tone or a number between -1 and 1. Suppose that we have a real-valued AWGN channel scenario at a very high SNR to decode a message that can take values $\{\pm 1.0, \pm 0.5\}$, and we have four independent observations to decode the message. If the message 0.5 is sent, all we can deduce in a one-bit quantized scenario where the thresholds are set to zero is that the message is most likely greater than zero. However, suppose we change the thresholds to uniformly sample the region between -1 and $+1$. In that case, we can better understand the amplitude of the message rather than just its sign information. Thermal noise is a great help in understanding the signal's amplitude in such a coarsely quantized scenario, which relates to the SR phenomenon.

After deciding on the Gaussian distribution, the mean and the variance of the thresholds must also be selected. Since we are trying to detect signals with zero-mean and our unquantized observations are zero-mean, the intuitive selection is to generate the threshold with zero-mean. The selection of the variance parameter is more challenging than the mean since it is hard to analytically track the effects of the empirical distribution of a pseudo-random selection. From many trials and observations, we saw that the variance selection depends on system parameters such as the number of BS antennas, the number of users, and the SNR. If the variance is too small, the performance is not affected. If it is selected as too large, the performance gets worse since the threshold values become too large to differentiate relatively small amplitude differences.

Again, after many trials and observations, when all other parameters are fixed, we saw that increasing the number of BS antennas N requires the variance to be increased, whereas increasing the number of users K requires the SNR to be decreased. Also, the PRQ approach does not yield any benefits and causes performance degradation

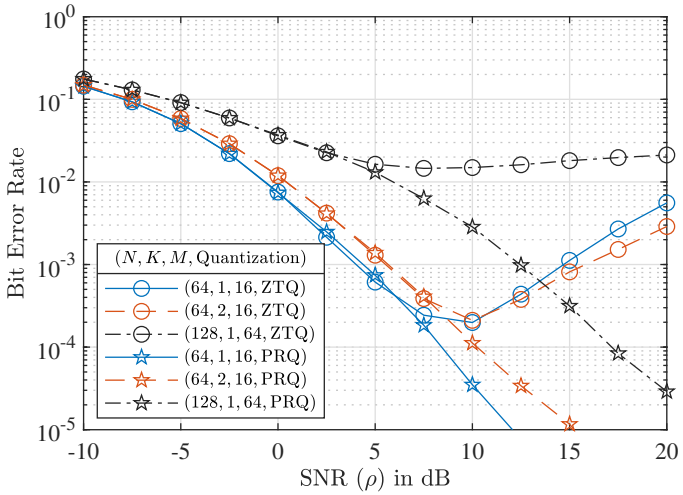


Figure 3.6: The BER plots obtained by the ML detector with ZTQ and PRQ in 64×1 and 64×2 systems with 16-QAM and 128×1 system with 64-QAM in the Rayleigh fading channel.

below a particular SNR value. Hence, gradually increasing the threshold variance after a starting SNR value ρ_{start} is an appropriate technique. In order to avoid selecting different variances for different scenarios, we parameterize the starting SNR value as

$$\rho_{\text{start}} = 5 \log_{10} \left(100 \frac{K}{N} \right) \text{ dB}. \tag{3.52}$$

This selection may not be optimal and relies only on empirical findings and observations. Now that the starting SNR value is selected, the threshold variance can be found accordingly using

$$\sigma_\tau^2 = \mathrm{R} \left(\frac{E_s}{\rho_{\text{start}}} - N_0 \right) = \mathrm{R} \left(\frac{1}{\rho_{\text{start}}} - N_0 \right). \tag{3.53}$$

The effective nonlinear channel between the users and the receiver antennas changes when the threshold information is used during baseband processing. If the information were not used, the effective nonlinear channel would have remained the same, and only a decrease in the SNR would occur since the thresholds would act as an additional form of noise over AWGN, as discussed while computing the conventional linear filters in Subsection 3.4.2.

Upon generation, rescaling the thresholds to strictly satisfy $\|\boldsymbol{\tau}\|^2 = N\sigma^2$ is helpful to obtain more stable results. The proposed PRQ scheme does not require updates for different channel realizations and depends only on the SNR. In Fig. 3.6, BER curves of 64×1 and 64×2 systems with 16-QAM, and a 128×1 system with 64-QAM are shown both with ZTQ and PRQ in the Rayleigh fading channel. The peak performance is achieved, i.e., SR occurs, between 5-10 dB of SNR for each system with ZTQ. The two-user performance with 16-QAM is better than that of the single-user at high SNR with ZTQ, which shows how multi-user interference (MUI) can act as a dither source. With PRQ, we not only stop the performance degradation but also obtain superior performance while approaching and at the SR point.

In addition to quantization, changing the sampling characteristics can help obtain better performance. In [16, 46, 54, 66], the merits of temporal oversampling in massive MIMO systems are shown. Instead of fixed-rate oversampling, dynamic oversampling is proposed as an efficient method to increase the achievable rate in [67]. Randomized sampling applications such as sampling with jitter can also be a valuable method for low-resolution systems, as discussed in [68]. However, due to correlation, changing the sampling characteristics complicates the noise and channel models. Symbol-rate sampling eases the baseband processing techniques and lowers the computational complexity of detection and estimation techniques.

3.7.3 Achievable Rate in One-Bit SIMO Systems

In this subsection, the achievable rate analysis in low-dimensional SIMO systems is made to better understand the merits of PRQ. Again, based on observations, we have seen that the thresholds should be optimized according to the given channel realization for the conventional MIMO systems, i.e., when the number of antennas is small. Thankfully, massive MIMO systems can benefit from a more robust approach due to a large number of antennas, and important performance gains can be obtained without optimizing the thresholds and utilizing PRQ. Threshold optimization is also tricky since the mutual information or the error probability should be optimized. There is no analytical formulation for the error probability, and the mutual information calculation is very costly and not even possible for large-scale systems due to memory

requirements. The problem of finding the capacity-achieving quantization thresholds in the case of availability of CSI only at the receiver can be written as

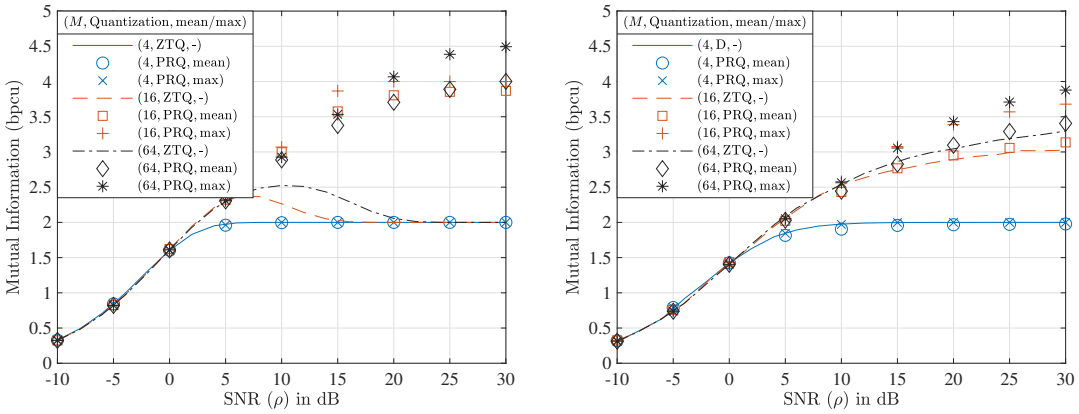
$$\hat{\boldsymbol{\tau}} = \underset{\boldsymbol{\tau}\in\mathbb{R}^{2N}}{\arg\max} \quad \mathcal{I}(\boldsymbol{r};\boldsymbol{x} \mid \boldsymbol{H},\boldsymbol{\tau}), \tag{3.54}$$

for fixed $\mathrm{p}(\boldsymbol{x})$, where $\mathcal{I}(\mathbf{r};\mathbf{x} \mid \boldsymbol{H},\boldsymbol{\tau})$ denotes the conditional mutual information between the quantized observation and the transmitted signal vectors given the channel realization $\boldsymbol{H}$ and the selected quantization threshold vector $\boldsymbol{\tau}$. For our setup, the analytical expression for the conditional mutual information can be written as

$$\begin{aligned}\mathcal{I}(\boldsymbol{r};\boldsymbol{x} \mid \boldsymbol{H},\boldsymbol{\tau}) = \\ -\frac{1}{M^K}\sum_{\boldsymbol{x}_1\in\mathcal{M}^{2K}}\sum_{\boldsymbol{r}\in\{\pm1\}^{2N}} \mathrm{p}(\boldsymbol{r} \mid \boldsymbol{x}_1,\boldsymbol{H},\boldsymbol{\tau})\log_2\left(\frac{1}{M^K}\sum_{\boldsymbol{x}_2\in\mathcal{M}^{2K}}\mathrm{p}(\boldsymbol{r} \mid \boldsymbol{x}_2,\boldsymbol{H},\boldsymbol{\tau})\right) \\ -\frac{1}{M^K}\sum_{\boldsymbol{x}_3\in\mathcal{M}^{2K}}\sum_{n=1}^{2N}\mathcal{H}_b\left(\Phi\left(\sqrt{2\rho}(\boldsymbol{h}_n^T\boldsymbol{x}_3-\tau_n)\right)\right),\end{aligned} \tag{3.55}$$

where $\mathrm{p}(\boldsymbol{r} \mid \boldsymbol{x},\boldsymbol{H},\boldsymbol{\tau}) = \prod_{n=1}^{2N}\Phi\left(\sqrt{2\rho}\, r_n(\boldsymbol{h}_n^T\boldsymbol{x}-\tau_n)\right)$ is the likelihood function, and $\mathcal{H}_b(.)$ is the binary entropy function. A detailed derivation of mutual information is given in Appendix B.

Finding the optimal thresholds is difficult since the solution is not necessarily unique, and the dimensionality is very high for the massive MIMO setup. To observe the effects of PRQ with the previously explained distribution of the thresholds for SIMO systems, we resort to random sampling by utilizing the Monte Carlo method. For the simulations, we aim to find the achievable rate with ZTQ and PRQ in a SIMO system where a BS equipped with 4 antennas is serving a single user in AWGN and Rayleigh fading channels. The conditional mutual information for any given channel realization is calculated for 500 random realizations of the thresholds for the PRQ scheme and 1 with zero thresholds for ZTQ. The average rate and the maximum rate obtained during trials for each channel are recorded with PRQ. Since the AWGN channel is deterministic with a unit CIR, only a single channel realization is utilized. For the Rayleigh channel, the average is calculated over 1000 channel realizations to reach the



(a) AWGN Channel

(b) Rayleigh Channel

Figure 3.7: Mutual information plotted against SNR for both AWGN (a) and Rayleigh (b) channels where $N = 4$, $K = 1$, and $M = 4, 16, 64$ with one-bit ZTQ (lines) and PRQ (markers) schemes. The average and the maximum rate obtained with the PRQ scheme are recorded.

final results.

The mutual information plots with respect to SNR are shown in Fig. 3.7a for the AWGN channel and in Fig. 3.7b for the Rayleigh fading channel. The performances in both channels exhibit no significant difference when the QPSK constellation is used. This is an intuitive result since each pair of ADCs from each antenna divides the 2D space into four regions, the cardinality of the QPSK alphabet is also four, and the modulation scheme is amplitude-invariant, i.e., the amplitude of each symbol is equal to $\sqrt{E_s}$. However, the behavior changes drastically for 16-QAM and 64-QAM. The SR phenomenon is visible in the AWGN channel with 16-QAM and 64-QAM constellations. We can no longer observe SR in this 4-antenna scenario for the Rayleigh fading channel. This seems to be due to the amplitude and phase-varying nature of the Rayleigh fading channel, which helps obtain different instantaneous SNR values at each branch. Nonetheless, the PRQ scheme outperforms the ZTQ scheme for both channel types, even on average. If the thresholds are optimized with respect to the given channel realization, the rate can be increased further in this low-dimensional setup.

3.7.4 Minimum Hamming Distance Analysis

Calculating mutual information for the massive MIMO setup is not feasible due to very large dimensionality. Therefore, another approach is followed to grasp how the PRQ scheme works for massive MIMO systems using an intuition from coding theory [27, 69]. Quantized observations can be seen as binary codewords in space for each given channel realization. Channel coding starts with the transmission of symbols from the uplink users. The signals pass through the channel and arrive at the receiver antennas. Additive noise corrupts the incoming signals, and the quantization operation occurs. The only quantization variable for our scenario is the vector of quantization thresholds. Hence, we can only intervene in the encoding process by changing the thresholds. The resultant spatial channel code will determine our error performance. A measure of the performance of a code is its minimum Hamming distance, which determines the diversity order of a code in fading channels [70]. In this part, we will investigate the average of the minimum Hamming distance of the space code obtained in the Rayleigh fading channel via Monte Carlo simulations at infinite SNR with both ZTQ and PRQ schemes. The infinite SNR scenario is chosen to understand the high SNR behavior, which is a critical part of our purposes. For a given channel and a given threshold realization, the minimum Hamming distance of the codebook is calculated for the infinite-SNR scenario as

$$d_{\mathrm{H}}^{\min} = \min_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}^{2K}} \mathrm{d}_{\mathrm{H}}(c(\boldsymbol{x}_1), c(\boldsymbol{x}_2)), \tag{3.56}$$

where $c(.)$ is the spatial codeword representation of its argument vector, which was defined in (3.48).

The average of the minimum Hamming distance in the Rayleigh fading channel is obtained for both ZTQ and PRQ scenarios where 10^4 randomly generated channel realizations are matched with randomly generated quantization thresholds. Table 3.1 shows the results for different scenarios. An interesting result obtained from these simulations is that when ZTQ is employed with $M \geq 16$, the minimum Hamming distance of the space code is 0, which means that the code is not uniquely decodable. Errors will indeed occur during the decoding operation due to the SR phenomenon; however, if we

Table 3.1: Average Minimum Hamming Distance of the Space Code in the Rayleigh Fading Channel Calculated Using (3.56)

N	K	M	with ZTQ	with PRQ
64	1	4	64.00	47.59
64	2	4	35.41	28.98
64	1	16	0.00	11.92
64	2	16	0.00	5.03
64	1	64	0.00	1.85
128	1	16	0.00	25.63
128	1	64	0.00	6.39

utilize the PRQ scheme, which gives us a chance to intervene in the codebook design, the minimum Hamming distance of the code increases. Hence, a superior performance compared to the ZTQ scheme will be obtained. QPSK modulation does not benefit from PRQ, as was the case during the mutual information calculations. Also, a larger alphabet size leads to a smaller minimum Hamming distance, whereas a larger number of antennas, i.e., larger codeword length, leads to an increased minimum Hamming distance for the space code with PRQ. Since the rate of the code is $R_c = \frac{K \log_2(M)}{2N}$, increasing K or M results in a lower, and increasing N leads to a larger minimum Hamming distance.

3.8 Computational Complexity Analysis

In this part, we compare the computational complexities of the proposed methods with that of the existing detectors. A downside of using Newton's method in the proposed first-stage detector BND is that it requires matrix inversion. Even though the gradient descent algorithm may not require matrix inversion, Newton's method requires much fewer iterations to reach the optimum [43, 71]. The part that dominates the complexity of BND is the calculation of the step towards the optimum, shown in line 7 of Algorithm 1. Multiplication of the inverse of the Hessian and the gradient does not require calculating the inverse of the matrix directly. However, calculating the

Table 3.2: Computational Complexity Comparison of the Proposed Detectors with the Existing Detectors from the Literature

Method	**Preprocessing**	**Stage 1**	**Stage 2**
BND-NCD	-	$\mathcal{O}(NK^2T)$	$\mathcal{O}(NK^2)$
MRC/BMRC	$\mathcal{O}(NK)$	$\mathcal{O}(NK)$	-
ZF/BZF	$\mathcal{O}(NK^2)$	$\mathcal{O}(NK)$	-
ML	$\mathcal{O}(NK\lvert\mathcal{M}\rvert^{2K})$	$\mathcal{O}(N\lvert\mathcal{M}\rvert^{2K})$	-
SVM-based	-	$\mathcal{O}(NK\kappa(N))$	-
OBMNet-NNS	Offline Training	$\mathcal{O}(NKL)$	$\mathcal{O}(\max(M,N)KM)$

Hessian matrix itself is a more computationally complex procedure since it involves a complexity of $\mathcal{O}(NK^2)$ where $N \gg K$ for massive MIMO systems. Since this procedure is repeated until the termination of the algorithm, the resulting complexity is $\mathcal{O}(NK^2T)$, where T denotes the number of iterations of the algorithm. The average number of iterations of BND is around $3-5$, and the number of unreliable indices for NCD is generally small at high SNR. Therefore, the complexity-dominant part of NCD is the last part where ML detection is applied on the reduced set with cardinality P. Even though we let P get as large as $2K$ during simulations, the average cardinality of the candidate set is small at high SNR, resulting in a complexity of $\mathcal{O}(NK^2)$ also for NCD. Note that computation of the nonlinear function $\psi(x)$ can be made using only $\varphi(x)$ since they have the same arguments.

The complexities of the proposed BND-NCD and the existing detectors from the literature are shown in Table 3.2. The MRC-based detectors offer the lowest complexity with poor error performance. Then comes ZF-based detectors offering better but inadequate performance and with higher computational complexity due to matrix inversion. The ML detector has the highest complexity, which is not feasible to operate in commercial systems. SVM-based detector from [3] reportedly has complexity $\mathcal{O}(NK\kappa(N))$, where $\kappa(.)$ is a super-linear function, and large values of N may lead to unaffordable complexity. OBMNet from [2], reportedly has complexity $\mathcal{O}(NKL)$ in its first stage, where L is the number of layers of the proposed deep neural network (DNN) architecture and a complexity of $\mathcal{O}(\max(M,N)KM)$ in its second stage NNS

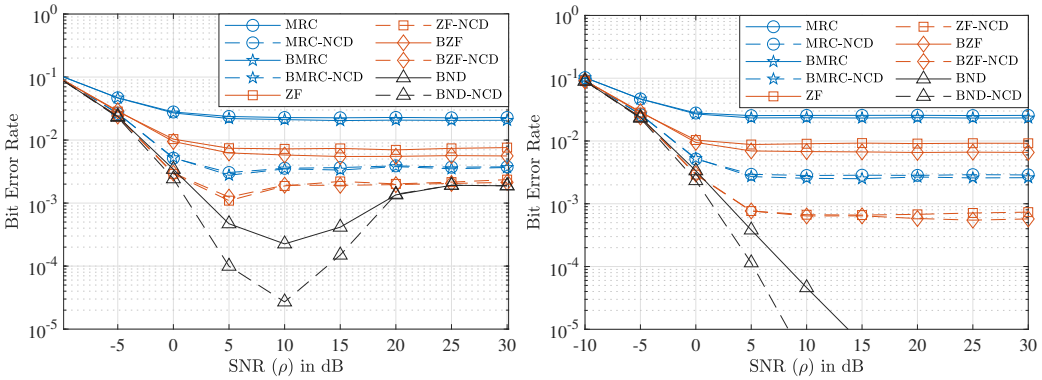
algorithm. Even though OBMNet does not require matrix inversion, the neural network needs an offline training stage. A much fewer number of iterations is necessary by BND, which makes the complexities of the detectors comparable. Also, the algorithm can be terminated early when BND's stopping criterion is satisfied. In contrast, the number of layers defined as part of the deep unfolded network architecture is a constant for OBMNet. Finally, the second stage NNS algorithm of [2] is a recursive algorithm with many for-loops. NCD omits for-loops and calculates the candidate set efficiently. MMSE-based filters are not included in the discussions since their complexity-performance trade-off is poor compared to the ZF-based filters. The MMSE filter performs the same as the ZF filter, and the BMMSE filter for non-zero threshold quantization does not have a compact analytical expression.

3.9 Simulation Results

In this section, the error performances of the proposed methods are investigated. For BND, the maximum number of iterations $T_{\max} = 15$, the damping factor ζ is calculated by taking $\rho_t = 20 - 160\frac{K}{N}$ dB, the penalty control parameter $\theta = 20$, and the termination threshold $\epsilon = \frac{10^{-4}}{\log_2(M)}$. For NCD, the maximum size of the set of nearest spatial codewords $P = 2K$, to adjust the size of the unreliable region, $\gamma = d/2$ with $d = \sqrt{\frac{3}{2(M-1)}}$, and the maximum size of the set of unreliable indices $U_{\max} = K$. While using PRQ, the starting SNR ρ_{start} is chosen according to (3.52), and the generated threshold values are chosen using (3.53). All simulation results are obtained with the uncorrelated Rayleigh fading channel model as explained in Section 3.3.

3.9.1 Performance Comparison of Detection Methods with ZTQ and PRQ

We now compare the performances of the linear detectors with that of the proposed methods. The linear detectors are also matched with the proposed NCD in the second stage. The resultant BER curves obtained by ZTQ and the proposed PRQ schemes are shown in Fig. 3.8. We have a 128×4 uplink MIMO system with the 16-QAM constellation. BND outperforms the linear detectors, ZF-type receivers top their MRC-



(a) ZTQ (b) PRQ

Figure 3.8: One and two-stage BER performances of the linear and the proposed detectors in a 128×4 system with 16-QAM using ZTQ (a) and PRQ (b) schemes, where linear detectors are also matched with the proposed NCD for the second stage.

type counterparts, and the Bussgang-based receivers perform slightly better than the conventional linear receivers. The performance gap between the conventional linear and Bussgang-based linear detectors is smaller than the previously reported works from the literature [2, 23] due to the quantization label ℓ_q. Substantial gains can be obtained by utilizing a two-stage approach for all scenarios. Linear detectors do not benefit much from PRQ, and only the second stage performance of BZF differs where a slightly lower error floor is obtained. While the linear detectors utilize only the first and second-order statistics of the unquantized input signal with Gaussian assumption for the Bussgang decomposition, the likelihood-based detectors can perfectly characterize the posterior probability. SR is visible for all detectors with ZTQ, whereas it is no longer observed at least down to BER $= 10^{-5}$ for BND and BND-NCD with PRQ.

3.9.2 Effect of Changing the Number of Users and the Number of BS Antennas on the High SNR Performance

Next, we investigate the effect of the number of users on the system's performance with the proposed BND-NCD by utilizing both ZTQ and PRQ. BER performances for $N = 64$ with $M = 16$, $N = 128$ with $M = 64$, and $N = 256$ with $M = 256$ with respect to the number of users at $\rho = 30$ dB are plotted in Fig. 3.9. Starting with

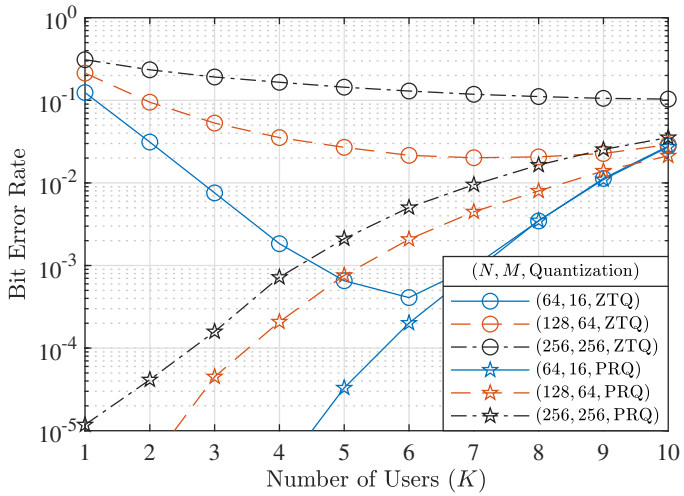


Figure 3.9: The BER performance of BND-NCD obtained by ZTQ and PRQ at $\rho = 30$ dB with respect to the number of users when $N = 64$ with 16-QAM, $N = 128$ with 64-QAM, and $N = 256$ with 256-QAM.

$K = 1$, increasing the number of users results in better performance up to a certain point with ZTQ, which also shows how MUI acts as a source of dither. As the number of users increases, the performance gap between the ZTQ and PRQ schemes decreases. They perform the same after some point, which happens at $K = 8$ for $N = 64$ and larger values of K for both $N = 128$ and $N = 256$. As N increases, the starting point where ZTQ and PRQ schemes perform the same is delayed to a larger value of K, which means that the merits of the PRQ scheme get more evident as the number of BS antennas increases.

The performance in SIMO systems that employ ML and the proposed BND-NCD with various high-order modulations, specifically, $M = 256, 1024, 4096$, are now examined. The BER performances obtained with ZTQ and PRQ schemes against the logarithm of the number of BS antennas are shown in Fig. 3.10. Based on this and the previous results, we can state that when K/N is small, the spatial degrees of freedom of the system is not adequately utilized with the ZTQ scheme. Moreover, we can also see that modulation orders such as 256, 1024, and 4096 can be used in one-bit quantized systems with PRQ if a sufficient number of BS antennas are utilized. Note that a different hyperparameter selection for NCD is made for this part, as indicated in the caption, to increase the performance of the proposed detector for the single-user

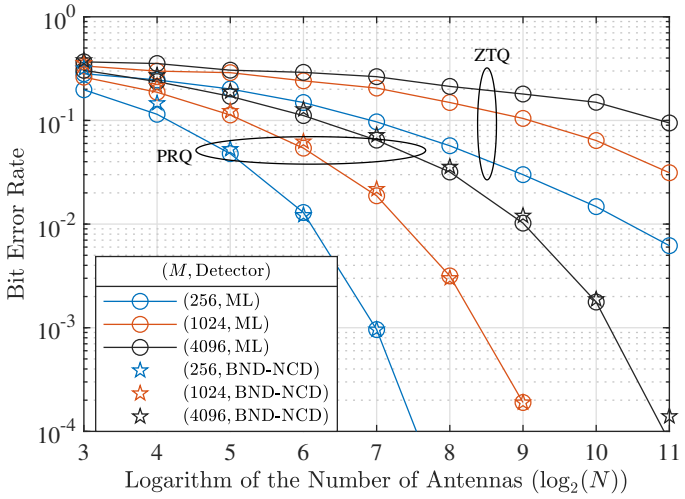


Figure 3.10: BER performance with respect to the number of antennas at $\rho = 30$ dB obtained for 256-QAM, 1024-QAM and 4096-QAM when $K = 1$ using ML with ZTQ and PRQ, and BND-NCD with PRQ. For this simulation only, $U_{\max} = 2$ and $P = \log_2(M)/2$.

system. The proposed two-stage detector can perform very close to ML with much lower computational complexity. For example, for 4096-QAM, ML needs to search for all the 4096 symbols to maximize the log-likelihood function. The proposed scheme requires $3 - 4$ iterations of BND and a search within a much smaller set $\mathcal{X}$ with cardinality at most $\log_2(M)/2 = 6$.

3.9.3 Performance Comparison of the Proposed and Existing Methods

The performances of the proposed detector and the existing detectors from the literature, namely, SVM-based from [3], and OBMNet from [2] are compared in Fig. 3.11, where the BER performances of one and two-stage detectors for a 128×8 system are obtained with 16-QAM. The second stage for OBMNet is the nearest neighbor search (NNS) algorithm from [2], and the SVM-based detector is utilized as a single-stage detector. The cardinality of the set of nearest neighbors to conduct ML detection for the second stage detector of OBMNet is also chosen as $2K = 16$. The plots show that the high-SNR error floor of the one-stage BND coincides with that of the two-stage OBMNet-NNS, the most recently proposed high-order modulation supporting detection scheme from the literature. Even when ZTQ is employed, BND-

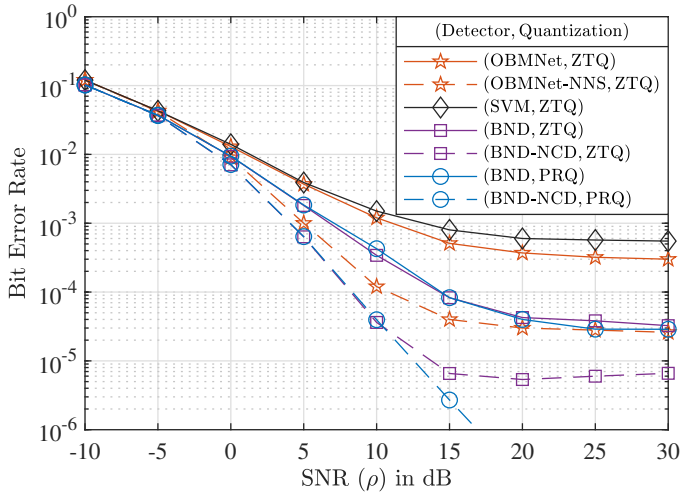
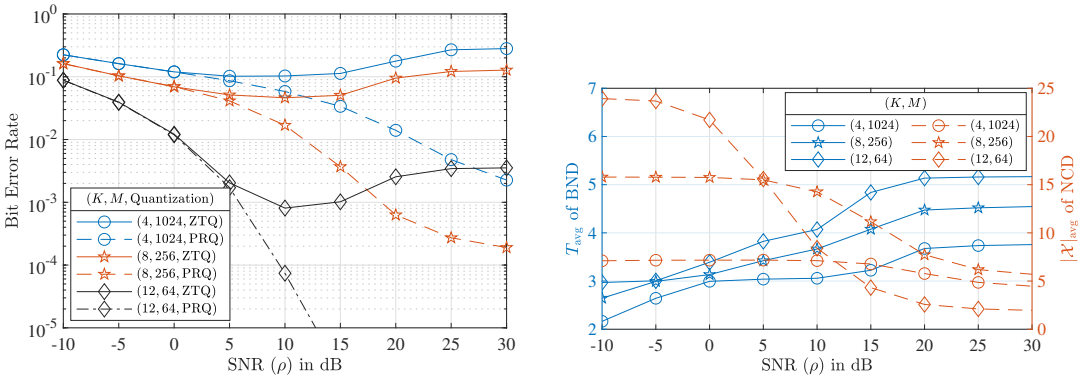


Figure 3.11: The BER performance comparison of the proposed detector with OBMNet [2] and SVM-based [3] detectors from the literature in a 128×8 system that employs 16-QAM. NNS is the nearest neighbor search algorithm, the second-stage detector from [2]. The maximum cardinality of the set of nearest neighbors for NNS is chosen as $2K = 16$, the same as NCD.

NCD outperforms both machine learning based approaches. Also, unlike any of the detectors included for comparison, the proposed two-stage BND-NCD with PRQ reduces the error floor below 10^{-6}. BND requires 4.5 iterations on average, whereas the number of layers of OBMNet, which is a DNN-based detector, is 15. Note that the SVM-based detector and OBMNet are selected among other candidates since they support 16-QAM, and their complexities are comparable to that of the BND-NCD method.

3.9.4 Performance and Complexity with Multi-User and High-Order Modulations

The error performances of several high-order modulation schemes in multi-user settings are examined as a final comparison. Additionally, complexity-related measurements regarding the average number of iterations of BND and the average size of the reduced set $\mathcal{X}$ are also made. BER performances of systems employing 512 BS antennas are obtained for $K = 4$ with 1024-QAM, $K = 8$ with 256-QAM, and $K = 12$ with 64-QAM in Fig. 3.12a with ZTQ and PRQ, and complexity-related measurements are



(a) BER Performance

(b) Computational Complexity

Figure 3.12: BER performance (a) and complexity-related measurements (b) against SNR obtained with BND-NCD when $N = 512$ for $K = 4$ with 1024-QAM, $K = 8$ with 256-QAM, $K = 12$ with 64-QAM. BER plots are obtained with ZTQ and PRQ, whereas complexity-related measurements are recorded with PRQ. T_{avg} is the average number of iterations of BND, and $|\mathcal{X}|_{\text{avg}}$ is the average size of the reduced set in NCD.

plotted in Fig. 3.12b with PRQ. The proposed PRQ scheme can mitigate the effects of SR and lower the error floors. Hence, the sources of spatial degrees of freedom are better exploited by PRQ for such a large value of N. For complexity, we can see that $3 - 5$ iterations are required for BND, and the number of iterations increases as K increases. $|\mathcal{X}|_{\text{avg}}$ decreases as SNR increases, which means that employing only the first-stage BND is sufficient at low SNR since the performance gaps between the one and two-stage approaches are very small at low SNR, which was shown in Fig. 3.8. At high SNR, even though the number of users is the largest for 64-QAM, it also has the smallest $|\mathcal{X}|_{\text{avg}}$. Therefore, as M increases, a more accurate estimation of the amplitudes is required. Note that the termination threshold ϵ of BND and the maximum size of the set of candidates P of NCD can be changed to obtain a trade-off between error performance and computational complexity.

3.10 Discussion

This chapter proposes a new two-stage detector for uplink one-bit massive MIMO systems that operate with pseudo-random quantization (PRQ). In the first stage, BND

estimates the transmitted signal using the log-likelihood function via Newton's Method. A penalty method to realize the box constraints related to the chosen constellation is also utilized to avoid saddle points during optimization. The second-stage detector, NCD, is proposed to refine the estimate from the first stage to enhance detection performance. The second stage creates a small set to select the candidate that maximizes the likelihood using the first-stage estimate. The set formation is based on the sign constraints imposed by one-bit quantization. The proposed PRQ scheme helps mitigate the detrimental effects of SR on detection performance. With the proposed two-stage detector and PRQ, one-bit multi-user massive MIMO systems can operate with high-order modulation schemes such as 256-QAM or 1024-QAM, and superior error performance to the existing detectors can be obtained with comparable complexity.

CHAPTER 4

DETECTION UNDER FREQUENCY-SELECTIVE FADING

4.1 Motivation

The BND-NCD approach is advantageous for SC-MIMO systems that employ PRQ under frequency-flat fading. However, frequency-selective fading is a more frequently encountered scenario where the wireless channel introduces ISI on the received signal at the BS. Time-domain equalization (TDE) can be very costly since both the channel length (L) and the data block length (V) can be very large in practice. Therefore, resorting to frequency-domain equalization (FDE) tools is beneficial. However, since it is not easy to statistically model the quantization distortion in the FD in one-bit MIMO systems, FDE based on the likelihood function is not straightforward. The motivation to utilize a detection scheme based on the likelihood function comes from our observations in Chapter 3, where we saw that the linear methods do not benefit much from PRQ compared to the likelihood-based BND-NCD. Therefore, just like the idea from the BND-NCD approach in Chapter 3, a detection scheme that would be compatible with the non-zero threshold idea to employ PRQ and optimize the log-likelihood with affordable computational complexity is the motivation in this chapter. We develop the system model by building upon Sections 2.2 and 2.3. Then, we again focus on linear detectors by utilizing the Bussgang decomposition. Next, we develop a low-complexity FDE scheme called PQND. With a similar motivation as in Chapter 3, we adopt the PRQ scheme also for the frequency-selective fading scenario and explain its usage. In the end, we discuss the computational complexity and obtain the error performances of the proposed methods.

Different solutions for detection have been proposed for low-resolution systems op-

erating under frequency-selective fading. Many of these methods focus on an SC transmission scheme to benefit from the relatively low peak-to-average power ratio (PAPR) and amplitude variation compared to OFDM. Various types of message-passing algorithms are present in the literature. An algorithm based on Bussgang decomposition and Ungerboeck factorization for CP-free SC systems is presented in [37]. In [22], phase-only measurements are utilized for detection, and in [72], massive MIMO with spatial modulation is considered. Even though SC transmission can be favorable, OFDM is a widely used scheme with many advantages in resource allocation and has variants such as OFDMA. In [4], a first-order projected gradient descent algorithm, 1BOX, is utilized in one-bit OFDM systems, which will be used as the benchmark algorithm in this chapter. Exact and inexact expectation maximization (EM) algorithms are deployed in [39] and [38], respectively. Note that 1BOX is chosen as the benchmark since exact EM requires high complexity, and inexact EM performs very close to 1BOX in terms of error rate. A significant setback of the 1BOX method is that it utilizes a first-order optimization technique, i.e., the gradient descent algorithm, and different system setups require the algorithm step size to be adjusted for fast convergence.

4.2 Contributions

The main contributions of this chapter can be summarized as follows:

- As in the case of SC systems in Chapter 3, we utilize PRQ in our system to exploit dithering by modifying the quantization thresholds to mitigate the negative effects of SR on detection performance in frequency-selective fading channels with OFDM and SC-FDE.
- With a similar motivation as in Section 3.4, we obtain the appropriate modifications for linear detectors under frequency-selective fading, this time by utilizing FD relations.
- We formulate equalization in one-bit MIMO-OFDM systems as a constrained optimization problem, and propose to solve it using a second-order optimization technique, projected Newton's method. Unlike the 1BOX method from [4],

this approach does not require selecting suitable step sizes for different system setups due to the additional second-order derivative information.

- After deriving the necessary relations using Newton's method, we utilize two approximations to decouple equalization among different subcarriers and to avoid matrix inversion to reach the proposed Projected Quasi-Newton Detector (PQND).
- By employing the PRQ scheme and the proposed PQND, we show that one-bit uplink massive MIMO-OFDM systems can support high-order modulation schemes such as 64-QAM and 256-QAM in multi-user settings, and the proposed detector outperforms the benchmark detector from [4] in terms of error performance with very similar computational complexity.

4.3 System Model

The fundamentals of the signal and system models were discussed in Chapter 2. Building upon the multi-carrier system description in Section 2.3 to obtain compact and vectorized forms, we consider an $N \times K$ uplink massive MIMO-OFDM system with V subcarriers, where K single-antenna users are served by a BS equipped with N antennas. Extension to CP-SC systems is straightforward and will be discussed in the

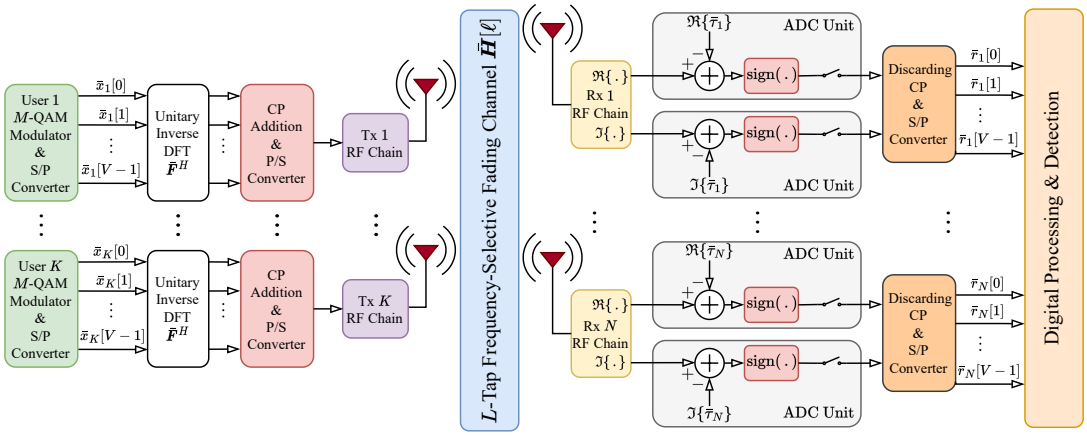


Figure 4.1: A block diagram that summarizes the OFDM system model in the frequency-selective fading scenario. S/P is for serial-to-parallel, and P/S is for parallel-to-serial.

following sections. Constellation symbols to be transmitted are selected independently from an M-QAM alphabet, denoted as $\bar{\mathcal{M}}$, with equal likelihood at all subcarriers and for all users. A block diagram that summarizes the adopted system model is shown in Fig. 4.1. The FD symbol of the k^{th} user at the v^{th} subcarrier is shown as $\bar{x}_k[v]$ for $k = 1, 2, \ldots, K$ and $v = 0, 1, \ldots, V-1$, where $\mathbb{E}[|\bar{x}_k[v]|^2] = E_s = 1$. We denote the vector containing the FD symbols of all users at the v^{th} subcarrier as $\bar{\boldsymbol{x}}[v] = \begin{bmatrix} \bar{x}_1[v] & \bar{x}_2[v] & \ldots & \bar{x}_K[v] \end{bmatrix}^T$ for $v = 0, 1, \ldots, V-1$. The concatenated version of an FD vector for all subcarriers or a TD vector for all time indices can be obtained as

$$\underline{\bar{\boldsymbol{x}}} = \begin{bmatrix} \bar{\boldsymbol{x}}[0]^T & \bar{\boldsymbol{x}}[1]^T & \ldots & \bar{\boldsymbol{x}}[V-1]^T \end{bmatrix}^T. \tag{4.1}$$

Each user takes their FD symbols' inverse unitary DFT before transmission to obtain their TD signals, which can be shown as

$$\bar{\boldsymbol{s}}[m] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{x}}[v] e^{+j2\pi mv/V}, \tag{4.2}$$

for $m = 0, 1, \ldots, V-1$. Equivalently,

$$\underline{\bar{\boldsymbol{s}}} = \bar{\boldsymbol{Q}}_K^H \underline{\bar{\boldsymbol{x}}}, \tag{4.3}$$

where $\bar{\boldsymbol{Q}}_i = \bar{\boldsymbol{F}} \otimes \boldsymbol{I}_i$ for $i \in \mathbb{Z}^+$ and $\bar{\boldsymbol{F}}$ is the unitary DFT matrix of size $V \times V$, whose n^{th} row and k^{th} column can be expressed as

$$[\bar{\boldsymbol{F}}]_{(n,k)} = \frac{1}{\sqrt{V}} \exp\left(-j2\pi \frac{nk}{V}\right), \tag{4.4}$$

for $n = 0, 1, \ldots, V-1$ and $k = 0, 1, \ldots, V-1$. Hence, $\bar{\boldsymbol{Q}}_K^H$ corresponds to each user's unitary inverse DFT operation separately.

We assume the BS has perfect knowledge of the CIR and the multipath channel has L taps. Hence, each user adds a CP of length $L_{\text{CP}} \geq L-1$ to the beginning of their TD signals to help mitigate the effects of ISI at the receiver side. Then, the

CP-added version of the TD signal vector is passed through an ideal DAC and I/Q modulated. The CIR at the ℓ^{th} tap, between the k^{th} user and n^{th} receiver antenna is denoted as $\bar{h}_{(n,k)}[\ell]$ for $k = 1, 2, \ldots, K$, $n = 1, 2, \ldots, N$, and $\ell = 0, 1, \ldots, L-1$. The ℓ^{th} tap of the CIR between all users and BS antennas can be represented as $\bar{\boldsymbol{H}}[\ell]$, where $[\bar{\boldsymbol{H}}[\ell]]_{(n,k)} = \bar{h}_{(n,k)}[\ell]$ with $\sum_{\ell=0}^{L-1} \mathbb{E}[|\bar{h}_{(n,k)}[\ell]|^2] = 1$. Assuming perfect synchronization, the received signal is I/Q demodulated, symbol-rate sampled, and CP is discarded. The unquantized received TD signal at time m can be expressed as

$$\bar{\boldsymbol{y}}[m] = \sum_{\ell=0}^{L-1} \bar{\boldsymbol{H}}[\ell]\bar{\boldsymbol{s}}[\langle m-\ell\rangle_V] + \bar{\boldsymbol{w}}[m], \tag{4.5}$$

for $m = 0, 1, \ldots, V-1$, where $\bar{y}_n[m]$ and $\bar{w}_n[m]$ for $n = 1, 2, \ldots, N$ represent the unquantized version of the received signal and the thermal noise sample at antenna n and time m, respectively. $\bar{\boldsymbol{w}}[m] \sim \mathcal{CN}(\boldsymbol{0}_N, N_0\boldsymbol{I}_N)$ for $m = 0, 1, \ldots, V-1$, and there is no temporal correlation among noise samples. An equivalent fully vectorized form of (4.5) can also be obtained. To do so, we define the block circulant MIMO channel circular convolution matrix of size $NV \times KV$ as

$$\underline{\bar{\boldsymbol{H}}}_b = \begin{bmatrix} \bar{\boldsymbol{H}}[0] & \boldsymbol{0}_{N\times K} & \ldots & \boldsymbol{0}_{N\times K} & \bar{\boldsymbol{H}}[L-1] & \ldots & \bar{\boldsymbol{H}}[1] \\ \bar{\boldsymbol{H}}[1] & \bar{\boldsymbol{H}}[0] & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \bar{\boldsymbol{H}}[L-1] \\ \bar{\boldsymbol{H}}[L-1] & \ldots & \bar{\boldsymbol{H}}[1] & \bar{\boldsymbol{H}}[0] & \boldsymbol{0}_{N\times K} & \ldots & \boldsymbol{0}_{N\times K} \\ \boldsymbol{0}_{N\times K} & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \boldsymbol{0}_{N\times K} \\ \boldsymbol{0}_{N\times K} & \ldots & \boldsymbol{0}_{N\times K} & \bar{\boldsymbol{H}}[L-1] & \ldots & \bar{\boldsymbol{H}}[1] & \bar{\boldsymbol{H}}[0] \end{bmatrix}, \tag{4.6}$$

which, due to CP, can be expressed as

$$\underline{\bar{\boldsymbol{H}}}_b = \bar{\boldsymbol{Q}}_N^H \underline{\bar{\boldsymbol{\Lambda}}}_b \bar{\boldsymbol{Q}}_K, \tag{4.7}$$

where $\underline{\bar{\boldsymbol{\Lambda}}}_b$ is the block diagonal FD MIMO channel matrix, which can be shown as

$$\underline{\bar{\boldsymbol{\Lambda}}}_b = \begin{bmatrix} \bar{\boldsymbol{\Lambda}}[0] & \mathbf{0}_{N\times K} & \dots & \mathbf{0}_{N\times K} \\ \mathbf{0}_{N\times K} & \bar{\boldsymbol{\Lambda}}[1] & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{N\times K} \\ \mathbf{0}_{N\times K} & \dots & \mathbf{0}_{N\times K} & \bar{\boldsymbol{\Lambda}}[V-1] \end{bmatrix}. \tag{4.8}$$

The decomposition shown in (4.7) utilizes the fact that the columns of the DFT matrix are eigenvectors of any circulant matrix [73] and applies this to the MIMO scenario as in [46]. Hence, the unquantized representation of the received signal can be found using

$$\begin{aligned} \underline{\bar{\boldsymbol{y}}} &= \underline{\bar{\boldsymbol{H}}}_b \bar{\boldsymbol{Q}}_K^H \underline{\bar{\boldsymbol{x}}} + \underline{\bar{\boldsymbol{w}}} \\ &= \bar{\boldsymbol{Q}}_N^H \underline{\bar{\boldsymbol{\Lambda}}}_b \underline{\bar{\boldsymbol{x}}} + \underline{\bar{\boldsymbol{w}}} \\ &= \underline{\bar{\boldsymbol{G}}}\, \underline{\bar{\boldsymbol{x}}} + \underline{\bar{\boldsymbol{w}}}, \end{aligned} \tag{4.9}$$

where $\underline{\bar{\boldsymbol{G}}} = \bar{\boldsymbol{Q}}_N^H \underline{\bar{\boldsymbol{\Lambda}}}_b$ represents the effective channel matrix, $\underline{\bar{\boldsymbol{w}}}$ is the spatially and temporally white thermal noise vector. Then, the unquantized representation of the observations can be used to find the one-bit quantized observations as

$$\underline{\bar{\boldsymbol{r}}} = \operatorname{sign}(\Re\{\underline{\bar{\boldsymbol{y}}} - \underline{\bar{\boldsymbol{\tau}}}\}) + j \operatorname{sign}(\Im\{\underline{\bar{\boldsymbol{y}}} - \underline{\bar{\boldsymbol{\tau}}}\}), \tag{4.10}$$

where $\underline{\bar{\boldsymbol{\tau}}}$ is the time-invariant quantization threshold vector with components $\bar{\boldsymbol{\tau}}[m_1] = \bar{\boldsymbol{\tau}}[m_2] = \bar{\boldsymbol{\tau}}$ for $m_1, m_2 \in \{0, 1, \dots, V-1\}$. These thresholds need not be updated for different channel realizations. The detailed procedure regarding pseudo-random threshold assignment is explained in Section 4.7. The quantized observation from the n^{th} antenna at time m can be shown as $\bar{r}_n[m]$ for $n = 1, 2, \dots, N$ and $m = 0, 1, \dots, V-1$.

Due to the structure of the proposed detector, we need to be able to work with real numbers only. Using the notation with real numbers as defined in (3.4), (4.10) can be written as

$$\underline{\boldsymbol{r}} = \operatorname{sign}\left(\underline{\boldsymbol{G}}\, \underline{\boldsymbol{x}} - \underline{\boldsymbol{\tau}} + \underline{\boldsymbol{w}}\right). \tag{4.11}$$

We define the SNR as the ratio of the average received signal power per user to the average noise power at each antenna such that

$$\rho = \frac{\mathbb{E}\left[\sum_{\ell=0}^{L-1} |\bar{h}_{(n,k)}[\ell]\bar{s}_k[\langle m-\ell\rangle_V]|^2\right]}{\mathbb{E}[|\bar{w}_n[m]|^2]} = \frac{E_s}{N_0} = \frac{1}{N_0}. \tag{4.12}$$

4.4 Linear Detection Methods

Due to their low complexities, linear methods are also popular for the frequency-selective fading scenario [35, 48]. However, ISI introduced by the channel increases the complexity and analysis of one-bit MIMO systems under frequency-selective fading. Conversion to the FD from TD is not straightforward as in unquantized systems due to the nonlinearity of one-bit quantization that occurs in the TD. In this section, just like in Section 3.4, we derive the Bussgang-based and conventional linear filters for one-bit massive MIMO systems, this time for the frequency-selective fading scenario. Before starting, defining the FD counterparts of the relations shown in the system model would be helpful. We define the FD representations of the TD quantized and unquantized observations, the TD thermal noise samples, and the TD CIR matrices as

$$\bar{\boldsymbol{\kappa}}[v] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{r}}[m] e^{-j2\pi mv/V}, \tag{4.13}$$

$$\bar{\boldsymbol{\eta}}[v] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{y}}[m] e^{-j2\pi mv/V}, \tag{4.14}$$

$$\bar{\boldsymbol{q}}[v] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{w}}[m] e^{-j2\pi mv/V}, \tag{4.15}$$

$$\bar{\boldsymbol{\Lambda}}[v] = \sum_{\ell=0}^{L-1} \bar{\boldsymbol{H}}[\ell] e^{-j2\pi \ell v/V}, \tag{4.16}$$

for $v = 0, 1, \ldots, V-1$, respectively. As a result, the unquantized input-output relation of the system can be expressed in the FD as

$$\bar{\boldsymbol{\eta}}[v] = \bar{\boldsymbol{\Lambda}}[v]\bar{\boldsymbol{x}}[v] + \bar{\boldsymbol{q}}[v], \tag{4.17}$$

for $v = 0, 1, \ldots, V-1$. Since the TD noise samples $\bar{\boldsymbol{w}}[m]$ are Gaussian, and DFT is a linear and unitary transform, the distribution of the FD noise samples $\bar{\boldsymbol{q}}[v]$ are the same as the TD samples, i.e., $\mathcal{CN}(\mathbf{0}_N, N_0\boldsymbol{I}_N)$.

4.4.1 Bussgang-Based Linear Filters

We need to linearize the system's input-output relation in the TD as we did for the frequency-flat systems. Note that derivations are made for a system employing OFDM transmission, but they can easily be extended to an SC system. As in Section 3.4, by adopting the generalized Bussgang decomposition for non-zero threshold quantization from [47], the linearized input-output relation using the fully vectorized notation can be expressed as

$$\underline{\bar{\boldsymbol{r}}}_e = \underline{\bar{\boldsymbol{r}}} - \underline{\bar{\boldsymbol{b}}} = \underline{\bar{\boldsymbol{g}}} \,\bar{\odot}\, \underline{\bar{\boldsymbol{y}}} + \underline{\bar{\boldsymbol{d}}}, \tag{4.18}$$

where $\underline{\bar{\boldsymbol{r}}}_e \in \mathbb{C}^{NV}$ is the conditionally zero-mean version of the quantized observations, $\underline{\bar{\boldsymbol{d}}} \in \mathbb{C}^{NV}$ is the quantization noise vector, $\underline{\bar{\boldsymbol{g}}} \in \mathbb{C}^{NV}$ is the vector of Bussgang gains, and $\underline{\bar{\boldsymbol{b}}} \in \mathbb{C}^{NV}$ is the bias vector over all time indices. Using the relations (3.8) and (3.9) from the frequency-flat scenario, the stationarity of the observations due to CP [46], and the time-invariance of the threshold vector we can state that

$$\bar{\boldsymbol{g}} = \bar{\boldsymbol{g}}[0] = \ldots = \bar{\boldsymbol{g}}[V-1] \in \mathbb{C}^N, \tag{4.19}$$

$$\bar{\boldsymbol{b}} = \bar{\boldsymbol{b}}[0] = \ldots = \bar{\boldsymbol{b}}[V-1] \in \mathbb{C}^N. \tag{4.20}$$

Hence, the Bussgang gain and bias vectors are time-invariant over a data block. We can write an alternative form of the linearized input-output as

$$\begin{aligned}\bar{\boldsymbol{r}}_e[m] = \bar{\boldsymbol{r}}[m] - \bar{\boldsymbol{b}} &= \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{y}}[m] + \bar{\boldsymbol{d}}[m] \\ &= \sum_{\ell=0}^{L-1} \left(\bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{H}}[\ell]\right) \bar{\boldsymbol{s}}[\langle m-\ell\rangle_V] + \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{w}}[m] + \bar{\boldsymbol{d}}[m] \\ &= \sum_{\ell=0}^{L-1} \bar{\boldsymbol{H}}_e[\ell] \bar{\boldsymbol{s}}[\langle m-\ell\rangle_V] + \bar{\boldsymbol{w}}_e[m],\end{aligned} \tag{4.21}$$

for $m = 0, 1, \ldots, V-1$, where $\bar{\boldsymbol{r}}_e[m] = \bar{\boldsymbol{r}}[m] - \bar{\boldsymbol{b}}$ is the bias compensated quantized observation vector at time m, $\bar{\boldsymbol{H}}_e[\ell] = \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{H}}[\ell]$ for $\ell = 0, 1, \ldots, L-1$ is the effective channel matrix, and $\bar{\boldsymbol{w}}_e[m] = \bar{\boldsymbol{g}} \,\bar{\odot}\, \bar{\boldsymbol{w}}[m] + \bar{\boldsymbol{d}}[m]$ for $m = 0, 1, \ldots, V-1$ is the effective noise vector. Similar to the frequency-flat scenario, the Bussgang gain vector and the bias vector can be calculated as

$$\bar{g}_n = \sqrt{\frac{4}{\pi \beta_n}} \left(\exp\left(-\frac{\Re\{\bar{\tau}_n\}}{\beta_n} \right) + j \exp\left(-\frac{\Im\{\bar{\tau}_n\}}{\beta_n} \right) \right), \tag{4.22}$$

$$\bar{b}_n = 2\Phi\left(-\frac{\Re\{\bar{\tau}_n\}}{\sqrt{\beta_n/2}} \right) - 1 + j \left(2\Phi\left(-\frac{\Im\{\bar{\tau}_n\}}{\sqrt{\beta_n/2}} \right) - 1 \right), \tag{4.23}$$

where the vector $\boldsymbol{\beta} \in \mathbb{R}^N$ consisting of the diagonal part of the TD cross-covariance matrix $\bar{\boldsymbol{C}}_y[m] = \mathbb{E}[\bar{\boldsymbol{y}}[n+m]\bar{\boldsymbol{y}}[n]^H]$ calculated at time $m = 0$ can be found as

$$\beta_n = [\bar{\boldsymbol{C}}_y[0]]_{(n,n)} = \frac{1}{V} \sum_{v=0}^{V-1} [\bar{\boldsymbol{C}}_\eta^{(v)}]_{(n,n)}, \tag{4.24}$$

and $\bar{\boldsymbol{C}}_\eta^{(v)} = \mathbb{E}[\bar{\boldsymbol{\eta}}[v]\bar{\boldsymbol{\eta}}[v]^H] = \bar{\boldsymbol{\Lambda}}[v]\bar{\boldsymbol{\Lambda}}[v]^H + N_0 \boldsymbol{I}_N$. This approach is taken from [46] to efficiently calculate the TD covariance matrices using the FD covariance matrices since calculations in the TD require taking the ISI introduced by the channel into account.

Now, the linearized TD relation can be transferred to the FD by taking the unitary DFT of both sides of (4.21) to obtain

$$\bar{\boldsymbol{\kappa}}_e[v] = \bar{\boldsymbol{\Lambda}}_e[v]\bar{\boldsymbol{x}}[v] + \bar{\boldsymbol{q}}_e[v], \tag{4.25}$$

for $v = 0, 1, \ldots, V-1$, where the FD representations of the bias compensated quantized observation vector $\bar{\boldsymbol{r}}_e[m]$, the effective channel matrix $\bar{\boldsymbol{H}}_e[\ell]$, and effective noise vector $\bar{\boldsymbol{w}}_e[v]$ for $v = 0, 1, \ldots, V-1$ can be calculated respectively as

$$\bar{\boldsymbol{\kappa}}_e[v] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{r}}_e[m] e^{-j2\pi mv/V}, \tag{4.26}$$

$$\bar{\boldsymbol{q}}_e[v] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{w}}_e[m] e^{-j2\pi mv/V}, \tag{4.27}$$

$$\bar{\boldsymbol{\Lambda}}_e[v] = \sum_{\ell=0}^{L-1} \bar{\boldsymbol{H}}_e[\ell] e^{-j2\pi \ell v/V}. \tag{4.28}$$

For SC transmission, we can replace $\bar{\boldsymbol{x}}[v]$ with

$$\bar{\boldsymbol{a}}[v] = \frac{1}{\sqrt{V}} \sum_{m=0}^{V-1} \bar{\boldsymbol{x}}[m] e^{-j2\pi mv/V}, \tag{4.29}$$

to obtain the FD relations since the constellation symbols are now transmitted in the TD.

According to (4.25), subcarrier-level processing is applicable in the FD, and it is straightforward to calculate the BMRC and BZF filters. The BMRC and BZF can be calculated, respectively as

$$\bar{\boldsymbol{F}}_{\mathrm{BMRC}}[v] = \bar{\boldsymbol{\lambda}}^e[v] \odot \bar{\boldsymbol{\Lambda}}_e[v]^H, \tag{4.30}$$

$$\bar{\boldsymbol{F}}_{\mathrm{BZF}}[v] = (\bar{\boldsymbol{\Lambda}}_e[v]^H \bar{\boldsymbol{\Lambda}}_e[v])^{-1} \bar{\boldsymbol{\Lambda}}_e[v]^H, \tag{4.31}$$

where the MRC scaling vector is obtained as

$$\bar{\lambda}_k^e[v] = \frac{1}{\sum_{n=1}^{N} |[\bar{\boldsymbol{\Lambda}}_e[v]]_{(n,k)}|^2}. \tag{4.32}$$

As in the frequency-flat scenario, the BMRC filter maximizes the SNR, whereas the BZF filter maximizes the SIR for the given scenario. The BMMSE filter is again not included in our discussion due to the unavailability of a compact representation for the covariance matrix of the quantized observations for the non-zero threshold quantization scenario since the Arcsine law cannot be used.

For an OFDM system, a Bussgang-based linear filter of choice $\bar{\boldsymbol{F}}[v]$ is applied to the FD representation of the bias-compensated quantized observations to obtain estimates as

$$\tilde{\bar{\boldsymbol{x}}}[v] = \bar{\boldsymbol{F}}[v]\bar{\boldsymbol{\kappa}}_e[v], \tag{4.33}$$

for $v = 0, 1, \ldots, V-1$. For SC systems, $\bar{\boldsymbol{x}}[v]$ and $\tilde{\bar{\boldsymbol{x}}}[v]$ are replaced with $\bar{\boldsymbol{a}}[v]$ and $\tilde{\bar{\boldsymbol{a}}}[v]$, respectively. For SC systems, after the estimates $\tilde{\bar{\boldsymbol{a}}}[v]$ are found, they are converted back to the TD to obtain the estimates of the constellation symbols as

$$\tilde{\bar{\boldsymbol{x}}}[m] = \frac{1}{\sqrt{V}} \sum_{v=0}^{L-1} \tilde{\bar{\boldsymbol{a}}}[v] e^{+j2\pi mv/V}. \tag{4.34}$$

If further processing is not applied to the estimates, decisions are obtained by element-wise minimum distance mapping, i.e., symbol-by-symbol detection, as

$$\hat{\bar{x}}_k[m] = \underset{\bar{x} \in \bar{\mathcal{M}}}{\arg\min} \, |\tilde{\bar{x}}_k[m] - \bar{x}|, \tag{4.35}$$

for $k = 1, 2, \ldots, K$ and $m = 0, 1, \ldots, V-1$.

4.4.2 Conventional Linear Filters

As in Subsection 3.4.2, we try to find the appropriate quantization label to utilize the quantization-unaware conventional linear filters. As discussed before, due to the stationarity of the observations as a result of using CP, by utilizing the approximations and assumptions from Subsection 3.4.2, the quantization label can be found as

$$\ell_q = \sqrt{\frac{\pi(K + N_0 + \sigma_\tau^2)}{4}}, \tag{4.36}$$

where we assume the effective covariance matrix of the unquantized observations at $m = 0$ is $\bar{\boldsymbol{C}}_{y,e}[0] = (K + N_0 + \sigma_\tau^2)\boldsymbol{I}_N$ which can be valid when the number of channel taps L and the number of users K is large or for the low SNR regime.

The MRC and ZF filters [56] can be found as

$$\bar{\boldsymbol{F}}_{\mathrm{MRC}}[v] = \bar{\boldsymbol{\lambda}} \odot \bar{\boldsymbol{\Lambda}}[v]^H, \tag{4.37}$$

$$\bar{\boldsymbol{F}}_{\mathrm{ZF}}[v] = (\bar{\boldsymbol{\Lambda}}[v]^H \bar{\boldsymbol{\Lambda}}[v])^{-1} \bar{\boldsymbol{\Lambda}}[v]^H, \tag{4.38}$$

where the elements of the MRC scaling vector for $k = 1, 2, \ldots, 2K$ are defined as

$$\bar{\lambda}_k[v] = \frac{1}{\sum_{n=1}^{N} |[\bar{\boldsymbol{\Lambda}}[v]]_{(n,k)}|^2}. \tag{4.39}$$

Similar to (4.33), a conventional linear filter of choice $\bar{\boldsymbol{F}}[v]$ is applied to the FD representation of the quantized observations $\bar{\boldsymbol{\kappa}}[v]$ as

$$\tilde{\bar{\boldsymbol{x}}}[v] = \ell_q \bar{\boldsymbol{F}}[v] \bar{\boldsymbol{\kappa}}[v], \tag{4.40}$$

for $v = 0, 1, \ldots, V-1$. (4.34) is again valid for the SC transmission and (4.35) is used for symbol-by-symbol detection. Note that the MMSE filter is again not included in our discussions since it yields the same performance as the ZF filter [2, 23].

4.5 Maximum Likelihood Sequence Detection (MLSD)

In the presence of independent AWGN on each branch of a one-bit quantized channel, as in [2, 4, 32, 34], the log-likelihood of a quantized observation vector $\underline{\boldsymbol{r}}$, given a transmitted signal $\underline{\boldsymbol{x}}$, a MIMO channel $\underline{\boldsymbol{G}}$, a vector of quantization thresholds $\underline{\boldsymbol{\tau}}$, can be expressed as

$$\mathcal{L}(\underline{\boldsymbol{x}}) = \mathbf{1}_{2NV}^{T} \ln\left(\Phi\left(\sqrt{\frac{2}{N_0}}\underline{\boldsymbol{r}} \odot (\underline{\boldsymbol{G}}\,\underline{\boldsymbol{x}} - \underline{\boldsymbol{\tau}})\right)\right). \tag{4.41}$$

Then, we construct the maximum likelihood sequence detection (MLSD) for this problem as

$$\hat{\underline{\boldsymbol{x}}}_{\text{MLSD}} = \underset{\underline{\boldsymbol{x}} \in \mathcal{M}^{2KV}}{\arg\max} \quad \mathcal{L}(\underline{\boldsymbol{x}}) \tag{4.42}$$

where $\mathcal{M}$ is the set of possible values an FD symbol can take in either the in-phase or quadrature part.

Even though the MLSD approach can provide high performance in terms of achievable rate, it is not feasible for practical systems where V is generally very large, and the search space consists of M^{KV} candidates. Equivalents of MLSD with lower complexity can be derived. For example, in [74], the Viterbi equalizer is constructed for a one-bit CP-free SC system. However, the complexity is still large in multi-user systems or when utilizing high modulation orders. Note that subcarrier-wise ML detection in the FD is not applicable since conversion to the FD causes an unknown conditional distribution for the quantized observations due to quantization noise.

4.6 Proposed Detection Method: Projected Quasi-Newton Detector (PQND)

To derive a low-complexity FDE scheme that can work with non-zero thresholds, we start with MLSD. Then, we construct an equalizer using Newton's method with additional constraints. Since Newton's method prevents subcarrier-wise processing and requires matrix inversion, we move one step further with approximations to construct a quasi-Newton method for equalization.

4.6.1 Equalization with Newton's Method

To utilize gradient-based optimization techniques, we relax the discrete input set constraint in (4.42) and reformulate the problem as

$$\begin{aligned} \underline{\tilde{\boldsymbol{x}}} = \underset{\underline{\boldsymbol{x}} \in \mathbb{R}^{2KV}}{\arg\max} \quad & \mathcal{L}(\underline{\boldsymbol{x}}) \\ \text{subject to} \quad & |\underline{x}_i| \leq M_b, \quad i = 1, 2, \ldots, 2KV, \\ & \|\underline{\boldsymbol{x}}\|^2 = KV, \end{aligned} \tag{4.43}$$

where M_b is the boundary of the chosen constellation in one of the I/Q parts. For M-QAM, M_b can be found as in (3.27).

While relaxing the discrete input set constraint, we utilize two restrictions. Each element of $\underline{\boldsymbol{x}}$ is inside the boundaries of the constellation, which is named as box constraint [4], and the norm of $\underline{\boldsymbol{x}}$ is set to KV, which is named as norm constraint. The box constraint prevents diverging from the constellation's boundaries during equalization, using a priori knowledge of the chosen constellation scheme. The norm constraint is related to the law of large numbers (LLN), for which we utilize practical values of V being large in commercial systems. We assume the total number of constellation symbols sent from all users in one OFDM symbol duration KV is large. Since DFT operations are all unitary, we have $\mathbb{E}[\|\underline{\boldsymbol{x}}\|^2] \rightarrow KVE_s = KV$ as $KV \rightarrow \infty$.

Different strategies, such as the penalty method in Chapter 3 or the projection method [2,4], can be followed to apply constraints with gradient-based optimization. We resort to the projection method for both constraints, which does not require modifying the cost function. To solve (4.43), we start with Newton's method with projection. The update equation of Newton's method can be written as

$$\underline{\boldsymbol{x}}^{(t)} = \mathfrak{P}^{(t)} \left(\underline{\boldsymbol{x}}^{(t-1)} - \alpha (\nabla^2 \mathcal{L}^{(t-1)})^{-1} \nabla \mathcal{L}^{(t-1)} \right), \tag{4.44}$$

where $\alpha \in \mathbb{R}$ is the step size, and $\mathfrak{P}^{(t)}(.)$ is the projection function at iteration t which is formally addressed in the next subsection. For now, we deal with only gradient and Hessian calculations.

Similar to Section 3.6.1, in (4.44), $\nabla \triangleq \left[\frac{\partial}{\partial \underline{x}_1} \quad \ldots \quad \frac{\partial}{\partial \underline{x}_{2KV}} \right]^T$ can be expressed as a vector of size $2KV$, and ∇^2 can be expressed as the outer product of ∇ operator with itself: $\nabla^2 = \nabla \, \nabla^T$. Hence, the ∇^2 operator can be considered a $2KV \times 2KV$ matrix.

$\nabla\mathcal{L}^{(t)}$ is the gradient of the log-likelihood function $\mathcal{L}$ with respect to $\underline{\boldsymbol{x}}$ calculated at iteration t. Similarly, $\nabla^2\mathcal{L}^{(t)}$ is the Hessian of the log-likelihood function $\mathcal{L}$ with respect to $\underline{\boldsymbol{x}}$, i.e., the Fisher information matrix, at iteration t. The gradient and Hessian are functions of $\underline{\boldsymbol{x}}$, and we drop the argument for ease of notation. Also, we define a new vector

$$\underline{\boldsymbol{u}} = \sqrt{\frac{2}{N_0}}\underline{\boldsymbol{r}} \odot (\underline{\boldsymbol{G}}\,\underline{\boldsymbol{x}} - \underline{\boldsymbol{\tau}}), \tag{4.45}$$

for compactness. Then, the gradient of the log-likelihood function can be found as

$$\nabla\mathcal{L} = \sqrt{\frac{2}{N_0}}\underline{\boldsymbol{G}}^T \left(\underline{\boldsymbol{r}} \odot \varphi\left(\underline{\boldsymbol{u}}\right)\right), \tag{4.46}$$

where $\varphi(x) \triangleq \frac{d}{dx}\ln(\Phi(x)) = \phi(x)/\Phi(x)$. The Hessian matrix can be found as

$$\nabla^2\mathcal{L} = \frac{2}{N_0}\underline{\boldsymbol{G}}^T \operatorname{diag}\left(\psi\left(\underline{\boldsymbol{u}}\right)\right)\underline{\boldsymbol{G}}, \tag{4.47}$$

where $\psi(x) \triangleq \frac{d^2}{dx^2}\ln(\Phi(x)) = -x\varphi(x) - \varphi^2(x)$, and it is applied element-wise to its arguments. Since $\Phi(x)$ can approach zero exponentially fast, the computations of $\ln(\Phi(x))$, $\varphi(x)$, and $\psi(x)$ in finite precision can cause problems such as divergent behavior and uncertainties. The solution in Appendix A is also used herein.

4.6.2 Equalization with the Proposed Quasi-Newton Method

Newton's method can provide complexity reduction compared to the MLSD approach. However, it does not allow subcarrier-level equalization and requires a large matrix inversion. In this subsection, we attempt to solve these problems using approximations to obtain an equalization method with lower complexity. We define the step $\Delta\underline{\boldsymbol{x}} = (\nabla^2\mathcal{L})^{-1}\nabla\mathcal{L}$, which can be written as

$$\Delta\underline{\boldsymbol{x}} = \left(\frac{2}{N_0}\underline{\boldsymbol{G}}^T \operatorname{diag}\left(\psi\left(\underline{\boldsymbol{u}}\right)\right)\underline{\boldsymbol{G}}\right)^{-1}\left(\sqrt{\frac{2}{N_0}}\underline{\boldsymbol{G}}^T \left(\underline{\boldsymbol{r}} \odot \varphi\left(\underline{\boldsymbol{u}}\right)\right)\right). \tag{4.48}$$

Since $\underline{\boldsymbol{G}} = \boldsymbol{Q}_N^T\underline{\boldsymbol{\Lambda}}_b$, a more explicit form can be obtained as

$$\Delta\underline{\boldsymbol{x}} = \sqrt{\frac{N_0}{2}} \left(\underline{\boldsymbol{\Lambda}}_b^T \boldsymbol{Q}_N \operatorname{diag}\left(\psi\left(\underline{\boldsymbol{u}}\right)\right) \boldsymbol{Q}_N^T \underline{\boldsymbol{\Lambda}}_b\right)^{-1} \left(\underline{\boldsymbol{\Lambda}}_b^T \boldsymbol{Q}_N \left(\underline{\boldsymbol{r}} \odot \varphi\left(\underline{\boldsymbol{u}}\right)\right)\right). \tag{4.49}$$

The matrix to be inverted in (4.49) is a $2KV \times 2KV$ matrix without any additional property for simplification. To simplify the relations, we first aim to approximate $\operatorname{diag}(\psi(\underline{\boldsymbol{u}}))$ to a multiple of the identity matrix $\gamma \boldsymbol{I}_{2NV}$ such that

$$\gamma = \frac{1}{2NV} \sum_{n=1}^{2NV} \psi(\underline{u}_n). \tag{4.50}$$

This assumption is valid for the low SNR regime when N_0 is large so that the elements of $\bar{\underline{\boldsymbol{u}}}$ are very close to zero. For the high SNR regime, it is very likely that all elements of $\bar{\underline{\boldsymbol{u}}}$ are positive and large values. Since the rate of change of $\psi(.)$ is very low for large positive arguments, this assumption is also helpful for high SNR.

While $\operatorname{diag}(\psi(\underline{\boldsymbol{u}}))$ helps obtain the second-order derivative information in each branch, we replace it with the sample mean to obtain a form that helps simplify the matrix multiplications and not deviate much from Newton's method to obtain fast convergence. A similar approximation was made in [75] to calculate feed-forward and feedback filters in an iterative equalizer for SC systems to obtain block diagonal form in the FD. The resultant expression for $\Delta\underline{\boldsymbol{x}}$ can be obtained as

$$\Delta\underline{\boldsymbol{x}} \cong \frac{1}{\gamma} \sqrt{\frac{N_0}{2}} \left(\underline{\boldsymbol{\Lambda}}_b^T \underline{\boldsymbol{\Lambda}}_b\right)^{-1} \underline{\boldsymbol{\Lambda}}_b^T \boldsymbol{Q}_N \left(\underline{\boldsymbol{r}} \odot \varphi\left(\underline{\boldsymbol{u}}\right)\right). \tag{4.51}$$

Note that with this approximation, decoupling between the real and imaginary parts due to second-order derivative calculation can also be avoided. We can now safely go back to the notation with complex numbers. The complex counterpart of (4.44) for the quasi-Newton approach can be obtained as

$$\bar{\underline{\boldsymbol{x}}}^{(t)} = \mathfrak{P}^{(t)} \left(\bar{\underline{\boldsymbol{x}}}^{(t-1)} - \alpha \Delta \bar{\underline{\boldsymbol{x}}}^{(t-1)}\right), \tag{4.52}$$

where, using (4.51), $\Delta\bar{\underline{\boldsymbol{x}}}$ can be written as

$$\Delta\underline{\bar{\boldsymbol{x}}} = \frac{1}{\gamma}\sqrt{\frac{N_0}{2}}\left(\underline{\bar{\boldsymbol{\Lambda}}}_b^H\underline{\bar{\boldsymbol{\Lambda}}}_b\right)^{-1}\underline{\bar{\boldsymbol{\Lambda}}}_b^H\bar{\boldsymbol{Q}}_N\left(\underline{\bar{\boldsymbol{r}}}\,\bar{\odot}\,\bar{\varphi}\left(\underline{\bar{\boldsymbol{u}}}\right)\right), \tag{4.53}$$

by defining $\bar{\varphi}(\bar{\boldsymbol{a}})$ as

$$\bar{\varphi}(\bar{\boldsymbol{a}}) = \varphi(\Re\{\bar{\boldsymbol{a}}\}) + j\,\varphi(\Im\{\bar{\boldsymbol{a}}\}). \tag{4.54}$$

Note that $\underline{\bar{\boldsymbol{u}}}$ from its each component $\{\bar{\boldsymbol{u}}[m]\}_{m=0}^{V-1}$, and γ can be found with the complex notation as

$$\bar{\boldsymbol{u}}[m] = \sqrt{\frac{2}{N_0}}\bar{\boldsymbol{r}}[m]\,\bar{\odot}\,\left(\mathcal{F}_m^{-1}\{\bar{\boldsymbol{\Lambda}}[v]\bar{\boldsymbol{x}}[v]\} - \bar{\boldsymbol{\tau}}\right), \tag{4.55}$$

$$\gamma = \frac{1}{2NV}\sum_{n=1}^{N}\sum_{m=0}^{V-1}\psi\left(\Re\{\bar{u}_n[m]\}\right) + \psi\left(\Im\{\bar{u}_n[m]\}\right), \tag{4.56}$$

respectively, where $\mathcal{F}_m^{-1} = \frac{1}{\sqrt{V}}\sum_{v=0}^{V-1}(.)\mathrm{e}^{+j2\pi mv/V}$ denotes the unitary inverse DFT operation. Since $\underline{\bar{\boldsymbol{\Lambda}}}_b$ is a block diagonal matrix, subcarrier-level processing is now applicable. (4.53) can be written for each subcarrier $v = 0, 1, \ldots, V-1$ separately as

$$\Delta\bar{\boldsymbol{x}}[v] = \frac{1}{\gamma}\sqrt{\frac{N_0}{2}}\left(\bar{\boldsymbol{\Lambda}}[v]^H\bar{\boldsymbol{\Lambda}}[v]\right)^{-1}\bar{\boldsymbol{\Lambda}}[v]^H\mathcal{F}_v\left\{\bar{\boldsymbol{r}}[m]\,\bar{\odot}\,\bar{\varphi}\left(\bar{\boldsymbol{u}}[m]\right)\right\}, \tag{4.57}$$

where $\mathcal{F}_v = \frac{1}{\sqrt{V}}\sum_{m=0}^{V-1}(.)\mathrm{e}^{-j2\pi mv/V}$ denotes the unitary DFT operation. To avoid matrix inversion, we introduce one last approximation using the properties of massive MIMO systems. Notice that $\left(\bar{\boldsymbol{\Lambda}}[v]^H\bar{\boldsymbol{\Lambda}}[v]\right)^{-1}\bar{\boldsymbol{\Lambda}}[v]^H$ from (4.57) is the ZF filter. By exploiting large numbers of BS antennas, we assume $\bar{\boldsymbol{\Lambda}}[v]^H\bar{\boldsymbol{\Lambda}}[v]$ is diagonally-dominant, and $\left(\bar{\boldsymbol{\Lambda}}[v]^H\bar{\boldsymbol{\Lambda}}[v]\right)^{-1}$ can be approximated by the diagonal matrix $\mathrm{diag}(\boldsymbol{\lambda}[v])$, where $\boldsymbol{\lambda}[v]$ corresponds to the MRC scaling vector from (4.39).

Hence, we replace the ZF filter expression during the step calculations with the MRC filter by relying on massive MIMO, i.e., $N \gg K$, to obtain the final form

$$\Delta\bar{\boldsymbol{x}}[v] \cong \frac{1}{\gamma}\sqrt{\frac{N_0}{2}}\,\boldsymbol{\lambda}[v] \odot \left(\bar{\boldsymbol{\Lambda}}[v]^H\mathcal{F}_v\left\{\bar{\boldsymbol{r}}[m]\,\bar{\odot}\,\bar{\varphi}\left(\bar{\boldsymbol{u}}[m]\right)\right\}\right). \tag{4.58}$$

The projection function at iteration t, $\mathfrak{P}^{(t)}(.)$ can be expressed as

$$\mathfrak{P}^{(t)}\left(\underline{\bar{\boldsymbol{x}}}\right) = \begin{cases} \mathfrak{P}_{\text{box}}\left(\underline{\bar{\boldsymbol{x}}}\right), & 1 \leq t < T \\ \mathfrak{P}_{\text{norm}}\left(\underline{\bar{\boldsymbol{x}}}\right), & t = T \end{cases}, \tag{4.59}$$

where T denotes the total number of iterations. The box projection function $\mathfrak{P}_{\text{box}}(.)$ can be found as

$$\mathfrak{P}_{\text{box}}(\bar{x}) = \mathfrak{P}^{I}_{\text{box}}\left(\Re\{\bar{x}\}\right) + j\mathfrak{P}^{Q}_{\text{box}}\left(\Im\{\bar{x}\}\right), \tag{4.60}$$

where $\mathfrak{P}^{I}_{\text{box}}(.) = \mathfrak{P}^{Q}_{\text{box}}(.)$ which is defined as

$$\mathfrak{P}^{I}_{\text{box}}(x) = \text{sign}(x)\min\{|x|, M_b\}, \tag{4.61}$$

and each function is applied element-wise to its arguments. We define the norm projection function $\mathfrak{P}_{\text{norm}}(.)$ as

$$\mathfrak{P}_{\text{norm}}(\underline{\bar{\boldsymbol{x}}}) = \frac{\sqrt{KV}}{\|\underline{\bar{\boldsymbol{x}}}\|}\underline{\bar{\boldsymbol{x}}}. \tag{4.62}$$

Note that utilizing box projection is redundant at the last iteration T since minimum distance mapping is conducted on each element of the estimate after equalization is complete. Also, utilizing the norm projection function at each iteration may cause problems during convergence. As in [2], norm projection is applied only at the last iteration.

Before starting the iterative updates, an initial solution should be found, preferably not far from the optimum. MRC estimate is a suitable choice with its low complexity, which can be found using (4.37) and (4.40).

At high SNR, singular Hessian matrices can be encountered during the iterative updates since N_0 takes very small values and $\psi(x) \to 0$ as $x \to \infty$. To avoid singularities in such circumstances, we again utilize a damping factor ζ as in (3.40) with ρ_t as the threshold SNR. The actual N_0 is multiplied with ζ before starting the iterative updates to operate as if we are at the threshold SNR if the actual SNR exceeds the threshold.

Algorithm 3 Projected Quasi-Newton Detector (PQND)

Input: $\{\bar{\boldsymbol{r}}[m]\}_{m=0}^{V-1}, \{\bar{\boldsymbol{\Lambda}}[v]\}_{v=0}^{V-1}, \bar{\boldsymbol{\tau}}, \alpha, T, N_0, \zeta$

Output: $\{\hat{\tilde{\boldsymbol{x}}}[v]\}_{v=0}^{V-1}$

Calculate $\{\boldsymbol{\lambda}[v]\}_{v=0}^{V-1}$ as in (4.39)
Set the initial solution $\{\tilde{\bar{\boldsymbol{x}}}[v] \leftarrow \bar{\boldsymbol{x}}^0[v]\}_{v=0}^{V-1}$ by using (4.37)
Apply the damping factor $N_0 \leftarrow \zeta N_0$ (3.40)
for $t = 1$ to T **do**
    Calculate $\{\bar{\boldsymbol{u}}[m]\}_{m=0}^{V-1}$ using (4.55)
    Calculate γ as in (4.56)
    Calculate the step $\{\Delta\bar{\boldsymbol{x}}[v]\}_{v=0}^{V-1}$ using (4.58)
    Update $\{\tilde{\bar{\boldsymbol{x}}}[v] \leftarrow \tilde{\bar{\boldsymbol{x}}}[v] - \alpha\Delta\bar{\boldsymbol{x}}[v]\}_{v=0}^{V-1}$ as part of (4.52)
    if $t < T$ **then**
        Apply box projection as part of (4.52) using (4.60)
        $\{\tilde{\bar{\boldsymbol{x}}}[v]\}_{v=0}^{V-1} \leftarrow \mathfrak{P}_{\text{box}}\left(\{\tilde{\bar{\boldsymbol{x}}}[v]\}_{v=0}^{V-1}\right)$
    else
        Apply norm projection as part of (4.52) using (4.62)
        $\{\tilde{\bar{\boldsymbol{x}}}[v]\}_{v=0}^{V-1} \leftarrow \mathfrak{P}_{\text{norm}}\left(\{\tilde{\bar{\boldsymbol{x}}}[v]\}_{v=0}^{V-1}\right)$
    end if
end for
Apply minimum distance mapping as in (4.35)
$\{\{\hat{\tilde{x}}_k[v] \leftarrow \arg\min_{\bar{x}\in\bar{\mathcal{M}}} |\tilde{\bar{x}}_k[v] - \bar{x}|\}_{k=1}^{K}\}_{v=0}^{V-1}$
return $\{\hat{\tilde{\boldsymbol{x}}}[v]\}_{v=0}^{V-1}$

After the iterative updates are complete, symbol-by-symbol minimum distance mapping is applied on the estimate for $k = 1, \ldots, K$ and $v = 0, \ldots, V-1$ to obtain the final decisions as in (4.35).

The complete procedure for the proposed PQND is summarized in Algorithm 3. We started with the update equation of Newton's method (4.44) and utilized two approximations to obtain a low-complexity equalization scheme. The first is related to the identity approximation of the diagonal matrix $\text{diag}(\psi(\underline{\boldsymbol{u}})) \cong \gamma \boldsymbol{I}_{NV}$ as in (4.51). The second is the diagonally dominant approximation for $\left(\bar{\boldsymbol{\Lambda}}[v]^H \bar{\boldsymbol{\Lambda}}[v]\right)^{-1} \cong \text{diag}(\boldsymbol{\lambda}[v])$ as in (4.58). Therefore, a projected quasi-Newton method is constructed

that does not require matrix inversion. Note that the proposed method can easily be modified for channel estimation purposes, which is left as future work. Finally, notice that a second-stage approach such as NCD was not proposed in addition to PQND. This issue is related to the complexity of the nearest neighbor search being unfeasible under frequency-selective fading and subcarrier-wise ML detection not being an option.

4.6.3 Extension to SC-FDE

As discussed in Section 4.4, both OFDM and SC systems can benefit from the circular convolution properties in situations where the system employs CP. Hence, the OFDM model can easily be extended for SC-FDE. For OFDM transmission, the transmitted TD signal is the unitary inverse DFT of the constellation symbols as shown in (4.9). For SC-FDE, the transmitted TD signal directly corresponds to the constellation symbols. Hence, for SC-FDE, (4.9) can be re-written as

$$\begin{aligned} \underline{\bar{\boldsymbol{y}}} &= \underline{\bar{\boldsymbol{H}}}_b \, \underline{\bar{\boldsymbol{x}}} + \underline{\bar{\boldsymbol{w}}} \\ &= \bar{\boldsymbol{Q}}_N^H \underline{\bar{\boldsymbol{\Lambda}}}_b \, \bar{\boldsymbol{Q}}_K \, \underline{\bar{\boldsymbol{x}}} + \underline{\bar{\boldsymbol{w}}} \\ &= \underline{\bar{\boldsymbol{G}}} \, \underline{\bar{\boldsymbol{a}}} + \underline{\bar{\boldsymbol{w}}}, \end{aligned} \tag{4.63}$$

where $\underline{\bar{\boldsymbol{G}}} = \bar{\boldsymbol{Q}}_N^H \underline{\bar{\boldsymbol{\Lambda}}}_b$ again, and $\underline{\bar{\boldsymbol{a}}} = \bar{\boldsymbol{Q}}_K \, \underline{\bar{\boldsymbol{x}}}$. Hence, the signal to be equalized is the unitary DFT of the transmitted constellation symbols. The whole procedure is the same as that of the OFDM system, except before applying the box constraints, the FD signal must be converted to the TD, and after the projection, it must be converted back to the FD to continue the iterative updates. After equalization is complete, the equalized signal is converted back to the TD to obtain the estimates of the constellation symbols.

4.7 Proposed Quantization Method: Pseudo-Random Quantization (PRQ)

We utilize the PRQ scheme to mitigate the effects of the SR phenomenon under frequency-selective fading, which is discussed thoroughly for the flat-fading scenario

in Section 3.7. Each pair of ADCs in the BS RF chains have a complex-valued threshold that separately represents the thresholds in I/Q parts. The thresholds are generated from the Gaussian distribution, i.e., $\bar{\boldsymbol{\tau}} \sim \mathcal{CN}(\mathbf{0}_N, \sigma_\tau^2 \boldsymbol{I}_N)$. Multipath diversity or frequency-selectivity is an important factor in one-bit massive MIMO systems that helps deal with SR [20]. Hence, different from the frequency-flat fading scenario, the number of channel taps (L), in addition to the number of BS antennas (N) and the number of users (K), must be used to determine the threshold variance. By relying on many observations, we parameterize the starting SNR ρ_{start}, which determines the point of SNR above which pseudo-random thresholds are utilized, as

$$\rho_{\text{start}} = \begin{cases} 0.15K^2 + L_s - 2.50\log_2(N) + 14 \text{ dB}, & \text{for OFDM} \\ 0.15K^2 + L_s - 2.50\log_2(N) + 16 \text{ dB}, & \text{for SC-FDE} \end{cases} \tag{4.64}$$

where L_s is the number of strong, i.e., high-power channel taps, which is determined as

$$L_s = \sum_{\ell=0}^{L-1} \frac{\text{sign}(p_d[\ell] - 1/L) + 1}{2}, \tag{4.65}$$

and $p_d[\ell] = \mathbb{E}[|\bar{h}_{n,k}[\ell]|^2]$ is the power delay profile (PDP) of the channel. Hence, the number of strong taps is equal to the number of taps whose power is greater than the power distributed between the taps of a uniform PDP channel. The proposed PRQ scheme does not require updates for different channel realizations and depends only on the SNR. The threshold variance is selected accordingly as in (3.53).

The starting SNR for SC-FDE is chosen as 2 dB higher than OFDM. According to our trials, this selection seemed more appropriate. This can be related to the higher amplitude variation in OFDM signals, which requires a larger variance in the dither signal, i.e., the quantization thresholds. We can think of the OFDM or SC transmission process under frequency-selective fading as $VK\log_2(M)$ message bits being transmitted in one data block, and a codeword of length $2VN$ is getting received. The channel state and quantization thresholds are essential factors affecting coding performance. Utilizing PRQ in frequency-selective channels is a way to intervene in the codebook design to obtain better performance thanks to the more accurate

amplitude recovery of the one-bit quantized signals. A detailed analysis for frequency-selective fading is not included since the analysis becomes too complicated due to the nonlinear distortion introduced in the TD.

4.8 Computational Complexity Analysis

In the proposed PRQ scheme, the ADC thresholds are kept the same for different channel realizations as long as the average SNR remains the same. If the SNR changes, appropriate scaling of the thresholds is sufficient, and there is no need for regenerating the pseudo-random numbers. The proposed PQND algorithm has many advantages. PQND can be adapted to different non-zero quantization threshold scenarios, making likelihood function-based FDE possible. Hence, a large number of channel taps is not a significant source of complexity. The computational complexity analysis of the proposed PQND is given in Table 4.1, where the complexity-dominant parts are examined with respect to their positions in Algorithm 3. Line 2 is the computation of the MRC estimate as an initial solution. Then, the remaining parts are repeated for T iterations as part of PQND. The arguments $\{\bar{\boldsymbol{u}}[m]\}_{m=0}^{N-1}$ of the nonlinear functions are calculated in Line 5. The sample mean of the average second-order derivative information γ is computed in Line 6. Note that computation of the nonlinear function $\psi(x)$ can be made using only $\varphi(x)$. Therefore, computing $\varphi(.)$ is sufficient, which can easily be handled using a look-up table. The step $\{\Delta\bar{\boldsymbol{x}}[v]\}_{v=0}^{V-1}$ towards the optimum for a given iteration is calculated in Line 7.

Since $N \gg K$ for massive MIMO, the complexity-dominant parts of the algorithm are Lines 5 and 7 both with complexity $\mathcal{O}(TVN\log_2(V)) + \mathcal{O}(TVNK)$. Hence,

Table 4.1: Computational Complexity Analysis per Data Block of the Proposed Projected Quasi-Newton Detector (PQND)

Line in Algorithm 3	**Number of Flops**
2	$\mathcal{O}(VNK)$
5	$\mathcal{O}(TVN\log_2(V)) + \mathcal{O}(TVNK)$
7	$\mathcal{O}(TVN\log_2(V)) + \mathcal{O}(TVNK)$

we can state that the complexity of PQND increases linearly with the number of BS antennas and users and super-linearly with the number of subcarriers due to the usage of the Fast Fourier Transform (FFT) algorithm. The complexity does not depend on the modulation order. However, increasing the number of iterations when the modulation order is high can be preferable to obtain a more accurate equalizer. The benchmark algorithm to compare with the proposed PQND is the 1BOX detector from [4], which also utilizes FDE tools based on the likelihood function. However, it does so by using first-order optimization. PQND and 1BOX have comparable complexity with similar computation steps. However, since PQND involves a second-order optimization technique, it requires fewer iterations to achieve a target error rate compared to 1BOX, as can be seen in Section 4.9. The complexity of MRC is $\mathcal{O}(VNK)$, and it is $\mathcal{O}(VNK^2)$ for ZF with poor performance at high SNR. Other nonlinear methods in the literature generally have high complexities [3, 36, 39]. The work in [38] offers a comparable complexity to 1BOX and PQND. However, the error performance of the detector is reportedly very similar to that of 1BOX.

4.9 Simulation Results

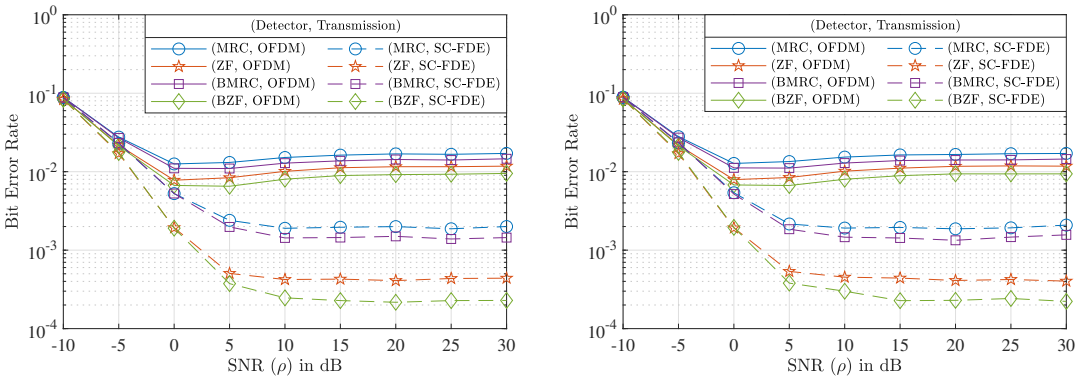
In this section, the error performance of the proposed detector is investigated for different scenarios. For PQND, there are three user-defined parameters. The step size of the proposed quasi-Newton optimization technique is chosen as $\alpha = 0.7$ for all scenarios. Newton's method can generally work with $\alpha = 1$. Due to approximations used while obtaining PQND, relatively smaller step size is used to avoid divergence. $T = 6$ iterations are considered for all scenarios. The threshold SNR for PQND is determined as $\rho_t = 20 - 150\frac{K}{N}$ dB, and the damping factor ζ is found accordingly as in (3.40). The variance of quantization thresholds is determined as in (4.64) and (3.53) when PRQ is utilized. The number of subcarriers is chosen as $V = 1024$. The frequency-selective wireless channel is modeled with Rayleigh fading, i.e., $\bar{h}_{(n,k)}[\ell] \sim \mathcal{CN}(0, p_d[\ell])$, and each coefficient is statistically independent both in space and time. BER performance is investigated with two PDPs throughout the simulations. The power of each tap in an exponential PDP channel can be calculated as

$$p_d[\ell] = \frac{\exp(-\mu\ell)}{\sum_{m=0}^{L-1}\exp(-\mu m)}, \tag{4.66}$$

where μ is the power decay rate, the sum in the denominator is added to satisfy unit average power. To model a small delay spread (SDS) channel, we use exponential PDP where $L = 8$ and $\mu = 1$, which results in $L_s = 3$. To model a large delay spread (LDS) channel, we utilize TDL-A delay profile from [76], which is a tapped delay line (TDL) channel model introduced by 3GPP for link level simulations of 5G systems, for which $L = 23$ and $L_s = 7$.

4.9.1 Performance Comparison of Linear Detectors with ZTQ and PRQ

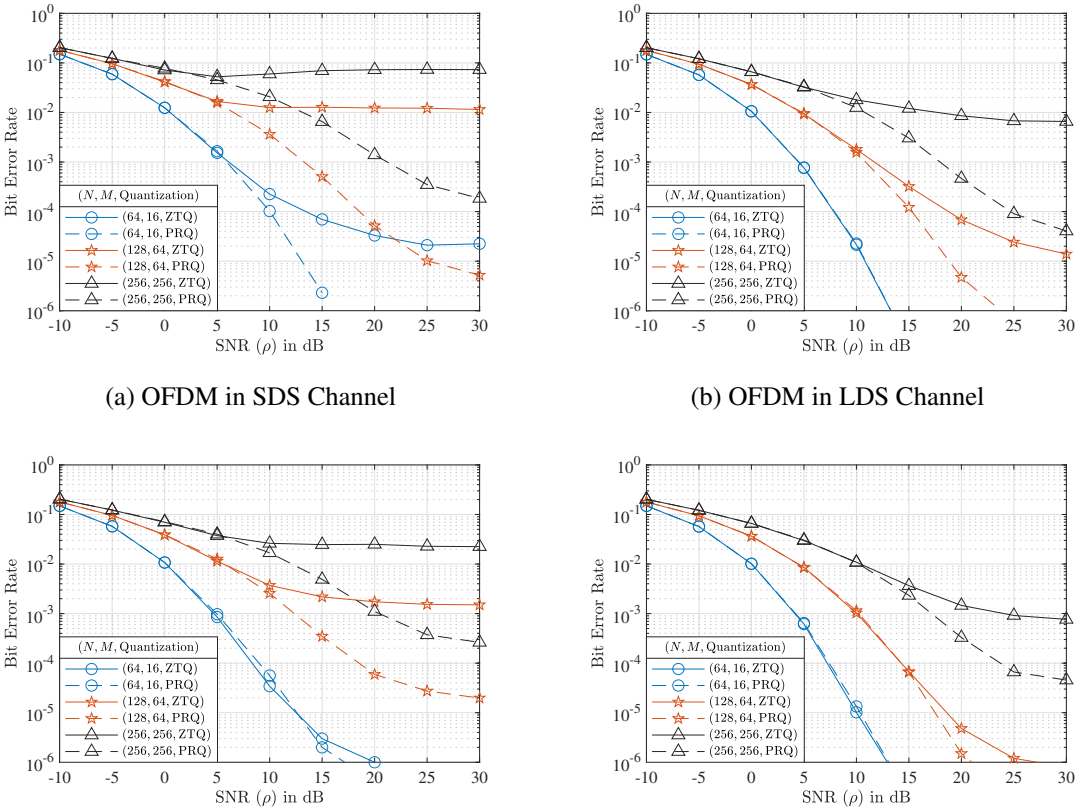
We start by comparing the ZTQ and PRQ performances of the linear detectors in the SDS channel. As shown in Fig. 4.2, the linear detectors perform the same for ZTQ and PRQ. We observed similar findings for the frequency-flat fading scenario in Section 3.9. Hence, we can again state that PRQ does not increase the performance of linear methods. Also, for this setup where a small number of users is being served, the SC-FDE scheme significantly outperforms the OFDM scheme, which suggests that the amplitude variation in OFDM systems causes a disadvantage for one-bit



(a) ZTQ

(b) PRQ

Figure 4.2: The BER performance comparison of the linear detectors using OFDM and SC-FDE with 16-QAM constellation in a 128×2 MIMO system in the SDS channel using ZTQ (a) and PRQ (b).



(a) OFDM in SDS Channel

(b) OFDM in LDS Channel

(c) SC-FDE in SDS Channel

(d) SC-FDE in LDS Channel

Figure 4.3: The BER performance of PQND in the SDS (a-c) and LDS (b-d) channel models using OFDM (a-b) and SC-FDE (c-d) schemes with respect to SNR (ρ) when $K = 2$ for $N = 64$ with 16-QAM, $N = 128$ with 64-QAM, and $N = 256$ and 256-QAM obtained with ZTQ and PRQ schemes.

MIMO systems when linear detectors are used. Similar findings were also reported in [77], where it is shown that the SC-FDE scheme shows more tolerance to hardware impairments compared to OFDM.

4.9.2 OFDM and SC-FDE Performance in the SDS and LDS Channels with ZTQ and PRQ

In Fig. 4.3, the performances of systems with two users are investigated when ZTQ and PRQ schemes are utilized with high-order modulations in both SDS and LDS channels and using PQND with both OFDM and SC-FDE schemes. The results are

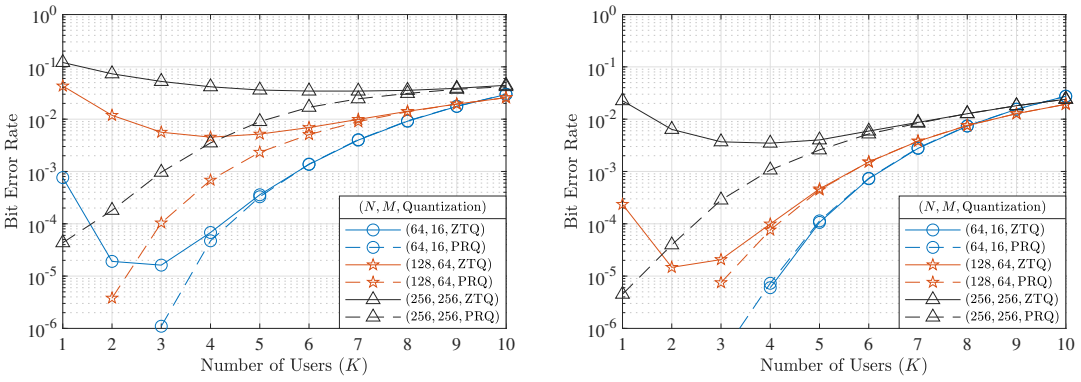
obtained for $N = 64$ with $M = 16$, $N = 128$ with $M = 64$, and $N = 256$ and $M = 256$. In general, the high SNR performance starts to saturate with ZTQ after some point. The spatial degrees of freedom are exploited much better with PRQ, and the high SNR error floors are decreased to significantly lower levels. Most of the proposed methods for frequency-flat and frequency-selective channel scenarios focus on QPSK and 16-QAM constellations. According to Fig. 4.3, higher-order modulation schemes can also be applicable in one-bit massive MIMO systems under frequency-selective fading with the PRQ scheme. In general, the error performance of SC-FDE seems to be superior to that of OFDM with ZTQ, and similar findings were reported in [77]. However, an interesting outcome of using PRQ is that the performances obtained with SC-FDE and OFDM become very similar. Hence, from these observations, in a scenario with a small number of users, the performances of OFDM and SC-FDE schemes can be increased to similar levels using PRQ. It can also be seen that the influence of PRQ is more critical in the SDS channel. Hence, it can be observed that increased frequency-selectivity helps obtain larger multipath diversity to reduce amplitude distortions. Similar findings were also reported in [78], where the error performance of an unquantized system is studied. As in Chapter 3, a larger number of BS antennas increases the performance gains obtained with PRQ. While all setups benefit from PRQ in the SDS channel, ZTQ and PRQ result in the same error performance for the 64×2 system in the LDS channel due to the higher ISI induced by the LDS channel compared to the SDS channel.

4.9.3 Effect of Changing the Number of Users and the Number of BS Antennas on the High SNR Performance

So far, the merits of PRQ are more evident for the SDS channel for a fixed number of users. In Fig. 4.4, the error performances in the SDS and LDS channels with respect to the number of users are plotted at $\rho = 30$ dB of SNR again for $N = 64$ with 16-QAM, $N = 128$ with 64-QAM, and $N = 256$ with 256-QAM both using ZTQ and PRQ with the PQND method for OFDM transmission. Since the high SNR performance is the limiting factor for high-order modulations, $\rho = 30$ dB is selected to better understand the high SNR behavior of ZTQ and PRQ. Similar to the findings from Chapter 3, starting with the single-user scenario, increasing the number of users results in better

performance at the beginning, which are implications of the SR phenomenon and MUI serving as a source of dither. The performance gain obtained with PRQ is more prominent when the number of users is small. With the increased ISI and multipath diversity, performance in the LDS channel is again better compared to the SDS channel. Starting with $K = 1$, the number of users for which ZTQ and PRQ begin to perform the same is smaller for the LDS channel. These results suggest that at high SNR, up to a certain point of sum interference composed of MUI and ISI, massive MIMO-OFDM systems can benefit from PRQ to achieve higher rates per user by employing higher modulation orders as opposed to the conventional ZTQ scheme.

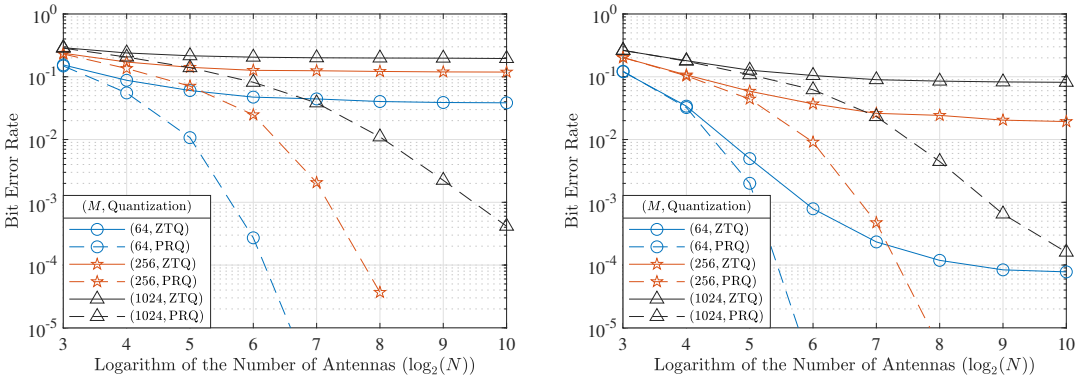
In Fig. 4.5, we check the 64-QAM, 256-QAM, and 1024-QAM error performances in a SIMO system with respect to the number of antennas at $\rho = 30$ dB both in the SDS and LDS channels using PQND. As can be seen, the SDS channel does not allow the usage of these high-order modulation schemes with the conventional ZTQ, even when the number of antennas is very large. Among the chosen modulation schemes, only 64-QAM transmission seems viable in the LDS channel. Hence, in the SIMO scenario where there is no MUI, ISI helps lower the amplitude distortion as suggested in [20] to support high-order modulation schemes with variable amplitudes. However, even in the LDS channel, employing PRQ is necessary to work with 256-QAM and



(a) SDS Channel

(b) LDS Channel

Figure 4.4: BER performances in the SDS (a) and LDS (b) channel models with respect to the number of users (K) at $\rho = 30$ dB of SNR for $N = 64$ with 16-QAM, $N = 128$ with 64-QAM, and $N = 256$ with 256-QAM obtained with ZTQ and PRQ schemes using PQND and OFDM.



(a) SDS Channel

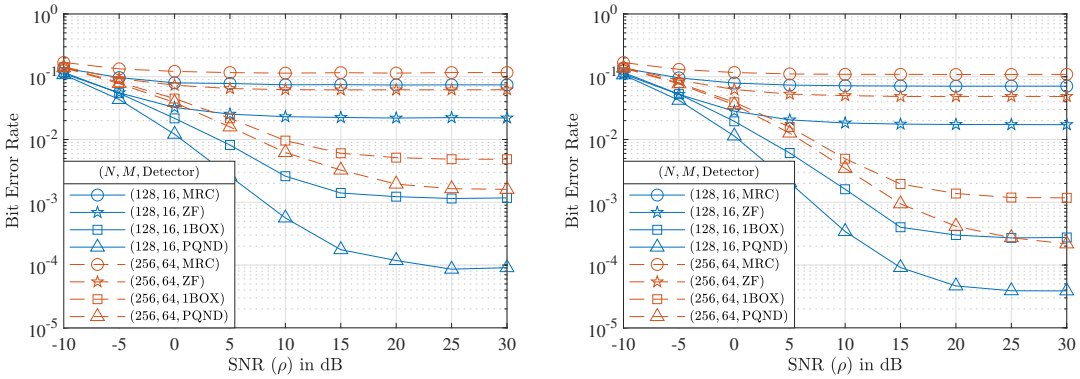
(b) LDS Channel

Figure 4.5: The BER performance of a SIMO system in the SDS (a) and LDS (b) channel models with respect to the logarithm of the number of antennas ($\log_2(N)$) at $\rho = 30$ dB of SNR with 64-QAM, 256-QAM, and 1024-QAM obtained with ZTQ and PRQ schemes using PQND and OFDM.

1024-QAM.

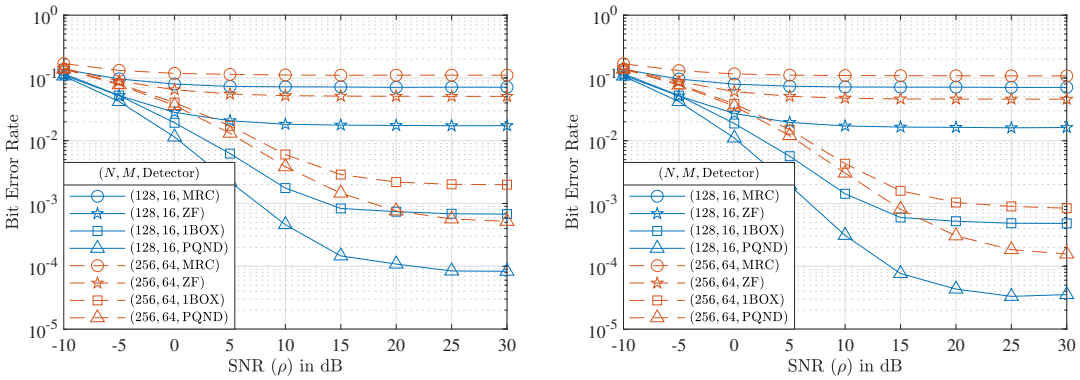
4.9.4 Performance with Multi-User and High-Order Modulations

Finally, the BER performances of different detectors from the literature are compared in Fig. 4.6 in both SDS and LDS channels. The results are obtained for $K = 10$ users for two setups where $N = 128$ with 16-QAM for the first and $N = 256$ with 64-QAM for the second. Linear detectors MRC and ZF are used for comparison along with 1BOX from [4] and the proposed PQND. The BMRC and BZF filters are not included in the comparison since they perform very similarly to their conventional quantization-unaware counterparts, as shown in Fig. 4.2. 1BOX is a first-order optimization-based equalization method from [4], for which derivations are similar to PQND except for the Hessian information. 1BOX involves only box projections at each iteration, though for fairness during comparison, we utilize the same projection function for 1BOX as PQND as described in (4.59). Hence, different than [4], norm projection is used at the final iteration of 1BOX, which leads to better performance. Regarding the damping constant, a similar discussion about acting as if the SNR is lower than its actual value is made in [4]. From several trials, we saw that choosing $\rho_t = 15$ dB for 1BOX is



(a) OFDM in SDS Channel

(b) OFDM in LDS Channel



(c) SC-FDE in SDS Channel

(d) SC-FDE in LDS Channel

Figure 4.6: Comparison of the BER performances of MRC, ZF, 1BOX [4], and PQND methods in the SDS (a-c) and LDS (b-d) channel models using OFDM (a-b) and SC-FDE (c-d) with respect to SNR for a 128×10 system with 16-QAM and a 256×10 system with 64-QAM.

suitable for both system setups. $T = 6$ iterations of 1BOX are utilized, the same as PQND. The step size is an issue for first-order optimization methods, especially for varying SNR. Again, by observing the performance for different setups, we select the step size of 1BOX as 0.007 (defined as κ in [4]). Note that even though the gradient expression involves the $\sqrt{N_0/2}$ constant at the beginning, as can be seen in (4.46), first-order optimization-based methods generally discard the term and take it as part of the step size, examples of which can be found in [2, 34]. However, for the proposed PQND, the SNR-dependent scaling is not discarded and becomes helpful with the second-order derivative information to select a fixed step size for all scenarios.

By comparing the plots, we can see that the LDS channel helps obtain better error performance than the SDS channel. Linear detectors MRC and ZF perform much poorly compared to the more sophisticated likelihood-based methods 1BOX and PQND. Note that all plots from Fig. 4.6 are obtained with ZTQ since ZTQ and PRQ perform the same when $K = 10$ according to Fig. 4.4. Hence the results with PRQ can be considered the same. For a fixed iteration number, the proposed PQND outperforms 1BOX in all scenarios, which is expected since second-order techniques converge faster than the first-order methods [71]. The performance of both OFDM and SC-FDE are very similar in this scenario as opposed to Fig. 4.3, where 2-user performances are shown. In this case, for the 10-user performance, OFDM is no longer inferior to SC-FDE, and both schemes perform almost the same. Even though SC-FDE may have lower amplitude variation than OFDM, in multi-user settings where the number of users is large, the received unquantized signals of both SC-FDE and OFDM schemes should share very similar statistical properties as a result of the central limit theorem (CLT). In [19, 45], it is also stated that the validity of assuming Gaussian distribution for the received signal increases at low SNR or when the number of users is large. The frequency-selective ISI channel is also important for this assumption since the unquantized received signal is a mixture of K users' last L transmitted signals.

4.10 Discussion

This chapter proposes a new detection method, PQND, that operates with PRQ for one-bit massive MIMO-OFDM and CP-SC massive MIMO systems. PQND is derived based on Newton's method with additional approximations to obtain a quasi-Newton method to optimize the log-likelihood function by operating at subcarrier level in the FD. By utilizing PQND and PRQ, one-bit massive MIMO systems can support high-order modulation schemes in low-user regimes and benefit from higher rates per user at high SNR with both OFDM and SC transmission schemes. The proposed detector outperforms the first-order optimization-based benchmark method 1BOX from [4] with comparable complexity.

CHAPTER 5

CONCLUSION

5.1 Summary

In this thesis study, detection methods in one-bit pseudo-randomly quantized uplink massive MIMO systems under both frequency-flat and frequency-selective fading scenarios are studied. We start with introductory chapters explaining how employing low-resolution ADCs can be an important solution to possible power consumption burdens in massive MIMO systems and provide the system description.

The main part of the work begins with detection under frequency-flat fading. We derive the Bussgang-based and conventional linear filters modified for non-zero threshold quantization. Moreover, we propose a new two-stage detection scheme that relies on Newton's method with box constraints and a nearest neighbor search algorithm based on the sign constraints imposed by one-bit quantization in the second stage. Then, we discuss a new pseudo-random quantization scheme for one-bit massive MIMO systems that modifies quantization thresholds to obtain a dithering effect. The proposed scheme does not require updates for different channel realizations, and it is not affected by different realizations of the thresholds in massive MIMO setups thanks to a large number of BS antennas. By combining the proposed BND-NCD and PRQ, uplink one-bit massive MIMO systems can operate with high modulation orders such as 256-QAM and 1024-QAM. Also, the proposed scheme outperforms existing high modulation order supporting detection schemes used for frequency-flat fading with comparable complexity.

Then, we turn our attention to frequency-selective fading. Using FDE tools, we also derive the linear filtering approaches for non-zero threshold quantization un-

der frequency-selective fading. Influenced by the BND method, we construct an equalizer for the frequency-selective fading channels using Newton's method. However, since the complexity of Newton's method becomes a significant burden under frequency-selective fading, we utilize two approximations to decouple equalization among subcarriers and to avoid matrix inversion. The proposed second-order PQND method can outperform the benchmark detector 1BOX that utilizes first-order optimization with similar complexity. Also, communication with high modulation orders such as 64-QAM and 256-QAM are shown to be possible using the proposed PQND and PRQ schemes under frequency-selective fading in one-bit massive MIMO-OFDM and one-bit CP-SC massive MIMO systems.

5.2 Future Research Directions

Changing quantization characteristics as shown in this thesis or sampling characteristics as in [16, 46, 48, 66, 67] can be very beneficial to increase the achievable rate in low-resolution massive MIMO systems. Conventional approaches such as zero-threshold quantization and Nyquist-rate sampling can sometimes be replaced with more sophisticated techniques, as in this study and the previously mentioned work. Since the demand for more data-rate will surely increase, low-resolution systems may require unconventional approaches that may result in significant performance improvements, as in this thesis.

Even though this thesis covers detection with PRQ under perfect CSI at the receiver assumption, channel estimation is a crucial task in practice. Many studies in the literature focus on the channel estimation task in one-bit massive MIMO systems such as [3, 4, 17, 31, 34, 47]. However, investigating the performance of PRQ for the channel estimation task can also be worth attention since PRQ provides better amplitude recovery of the signal to be estimated.

Coarse quantization induces significant nonlinearity in the input-output relation. Therefore artificial intelligence (AI) based methods have enjoyed popularity for the past few years. Many recent works in the literature use machine learning or deep learning algorithms for various tasks in communication systems. In [44], a survey on AI techniques

for the physical layer design of one-bit MIMO systems is discussed. Examples of AI applications for one-bit MIMO systems can be found in [2, 3, 24, 32, 40, 79]. AI tools can be a great solution for nonlinear and complex system architectures. For future studies, it can be important to search for AI tools that can optimize the system parameters, such as the thresholds, or use them with detection and estimation algorithms in more complex system models.

Multi-bit ADCs and oversampling effects are also interesting topics to be considered with PRQ for future research. It would be easy to modify the proposed BND and PQND methods for multi-bit ADCs since only the likelihood function needs to be updated to consider the probability of an interval between the quantization thresholds. However, oversampling would complicate the structure greatly due to the temporal correlation among noise samples. In such a scenario, resorting to machine learning tools would be preferable. It would also be very interesting to consider PRQ within such systems.

There are many works from the literature that focus only on in-band interference [2, 3, 23, 26, 32, 35, 36, 38–40, 67]. Considering the adjacent channel interference as in [46] would yield a more accurate characterization of the system setup when there is considerable leakage in the front-end filters. This work can be considered a starting point to build upon more complex scenarios.

REFERENCES

[1] F. Rivet, Y. Deval, J.-B. Begueret, D. Dallet, P. Cathelin, and D. Belot, "The experimental demonstration of a SASP-based full software radio receiver," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 5, pp. 979–988, 2010.

[2] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "Linear and deep neural network-based receivers for massive MIMO systems with one-bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7333–7345, 2021.

[3] L. V. Nguyen, A. L. Swindlehurst, and D. H. N. Nguyen, "SVM-based channel estimation and data detection for one-bit massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2086–2099, 2021.

[4] S. H. Mirfarshbafan, M. Shabany, S. A. Nezamalhosseini, and C. Studer, "Algorithm and VLSI design for 1-bit data detection in massive MIMO-OFDM," *IEEE Open Journal of Circuits and Systems*, vol. 1, pp. 170–184, 2020.

[5] E. Uysal and A. Ö. Yılmaz, *EE 435 - Communications I Lecture Notes*. Middle East Technical Universit Department of Electrical and Electronics Engineering, Ankara, Turkey, Fall 2019, [Online].

[6] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[7] A. Ö. Yılmaz, *EE 728 - Wireless Communications Lecture Notes*. Middle East Technical University Department of Electrical and Electronics Engineering, Ankara, Turkey, Fall 2021, [Online].

[8] F. Jameel, Faisal, M. A. A. Haider, and A. A. Butt, "Massive MIMO: A survey of recent advances, research numbers and future directions," in *2017 International Symposium on Recent Advances in Electrical Engineering (RAEE)*, pp. 1–6, Oct. 2017.

[9] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[10] B. Murmann, "ADC performance survey 1997-2021." Available: http://web.stanford.edu/~murmann/adcsurvey.html. Online.

[11] R. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 539–550, Apr. 1999.

[12] J. Liu, Z. Luo, and X. Xiong, "Low-resolution ADCs for wireless communication: A comprehensive survey," *IEEE Access*, vol. 7, pp. 91291–91324, 2019.

[13] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Transactions on Communications*, vol. 57, pp. 3629–3639, Dec. 2009.

[14] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5498–5512, 2015.

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

[16] S. Krone and G. Fettweis, "Capacity of communications channels with 1-bit quantization and oversampling at the receiver," in *2012 35th IEEE Sarnoff Symposium*, pp. 1–7, 2012.

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

[18] T.-K. Kim, M. Min, and Y.-S. Jeon, "Performance bound for MIMO systems using one-bit ADCs over Rayleigh fading channels," *IEEE Transactions on Vehicular Technology*, vol. 71, pp. 9067–9072, Aug. 2022.

[19] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput

analysis of massive MIMO uplink with low-resolution ADCs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 4038–4051, 2017.

[20] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 87–100, 2017.

[21] M. D. McDonnell and D. Abbott, "What is stochastic resonance? definitions, misconceptions, debates, and its relevance to biology," *PLoS Computational Biology*, vol. 5, no. 5, p. e1000348, 2009.

[22] S. Wang, L. Zhang, Y. Li, J. Wang, and E. Oki, "Multiuser MIMO communication under quantized phase-only measurements," *IEEE Transactions on Communications*, vol. 64, pp. 1083–1099, Mar. 2016.

[23] Ö. T. Demir and E. Björnson, "ADMM-based one-bit quantized signal detection for massive MIMO systems with hardware impairments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9120–9124, May 2020.

[24] S. Khobahi, N. Shlezinger, M. Soltanalian, and Y. C. Eldar, "LoRD-Net: Unfolded deep detection network with low-resolution receivers," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5651–5664, 2021.

[25] Y.-S. Jeon, N. Lee, S.-N. Hong, and R. W. Heath, "One-bit sphere decoding for uplink massive MIMO systems with one-bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 4509–4521, July 2018.

[26] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2541–2556, 2016.

[27] S.-N. Hong, S. Kim, and N. Lee, "A weighted minimum distance decoding for uplink multiuser MIMO systems with low-resolution ADCs," *IEEE Transactions on Communications*, vol. 66, pp. 1912–1924, May 2018.

[28] I. Bilinskis, "Randomized quantization," in *Digital Alias-Free Signal Processing*, pp. 87–106, John Wiley & Sons, Inc., 2007.

[29] R. A. Wannamaker, S. P. Lipshitz, and J. Vanderkooy, "Stochastic resonance as dithering," *Physical Review E*, vol. 61, no. 1, pp. 233–236, 2000.

[30] I. Bilinskis, "Pseudo-randomized quantizing," in *Digital Alias-Free Signal Processing*, pp. 107–126, John Wiley & Sons, Inc., 2007.

[31] D. K. W. Ho and B. D. Rao, "Antithetic dithered 1-bit massive MIMO architecture: Efficient channel estimation via parameter expansion and PML," *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2291–2303, 2019.

[32] S. Khobahi, N. Naimipour, M. Soltanalian, and Y. C. Eldar, "Deep signal recovery with one-bit quantization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2987–2991, May 2019.

[33] A. K. Saxena, A. Mezghani, and R. W. Heath, "Linear CE and 1-bit quantized precoding with optimized dithering," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 310–325, 2020.

[34] J. Choi, J. Mo, and R. W. Heath, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2005–2018, 2016.

[35] J. Guerreiro, R. Dinis, and P. Montezuma, "Low-complexity SC-FDE techniques for massive MIMO schemes with low-resolution ADCs," *IEEE Transactions on Communications*, vol. 67, pp. 2368–2380, Mar. 2019.

[36] H. He, C.-K. Wen, and S. Jin, "Bayesian optimal data detector for hybrid mmWave MIMO-OFDM systems with low-resolution ADCs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 469–483, 2018.

[37] A. B. Üçüncü, G. M. Güvensen, and A. Ö. Yılmaz, "A reduced complexity Ungerboeck receiver for quantized wideband massive SC-MIMO," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4921–4936, 2021.

[38] M. Shao and W.-K. Ma, "Divide and conquer: One-bit MIMO-OFDM detection by inexact expectation maximization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4890–4894, 2021.

[39] J. García, J. Munir, K. Roth, and J. A. Nossek, "Channel estimation and data equalization in frequency-selective MIMO systems with one-bit quantization," 2016.

[40] Y.-S. Jeon, N. Lee, and H. V. Poor, "Robust data detection for MIMO systems with one-bit ADCs: A reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1663–1676, 2020.

[41] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "MIMO networks with one-bit ADCs: Receiver design and communication strategies," *IEEE Transactions on Communications*, vol. 70, no. 3, pp. 1580–1594, 2022.

[42] A. Balatsoukas-Stimming and C. Studer, "Deep unfolding for communications systems: A survey and some new directions," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 266–271, 2019.

[43] A. Galántai, "The theory of Newton's method," *Journal of Computational and Applied Mathematics*, vol. 124, pp. 25–44, Dec. 2000.

[44] Y.-S. Jeon, D. Kim, S.-N. Hong, N. Lee, and R. W. Heath, "Artificial intelligence for physical-layer design of MIMO communications with one-bit ADCs," *IEEE Communications Magazine*, vol. 60, pp. 76–81, July 2022.

[45] A. Abdallah, M. M. Mansour, A. Chehab, and L. M. Jalloul, "MMSE detection for 1-bit quantized massive MIMO with imperfect channel estimation," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2018.

[46] A. B. Üçüncü, E. Björnson, H. Johansson, A. Ö. Yılmaz, and E. G. Larsson, "Performance analysis of quantized uplink massive MIMO-OFDM with oversampling under adjacent channel interference," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 871–886, 2020.

[47] Q. Wan, J. Fang, H. Duan, Z. Chen, and H. Li, "Generalized Bussgang LMMSE channel estimation for one-bit massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 19, pp. 4234–4246, June 2020.

[48] A. B. Üçüncü, *Massive Multiple-Input Multiple-Output Communication Systems with Low-Resolution Quantizers*. PhD thesis, Middle East Technical University, Dec. 2021.

[49] C. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, pp. 10–21, Jan. 1949.

[50] A. V. Oppenheim and R. W. Schafer, "Sampling of continuous-time signals," in *Discrete-Time Signal Processing*, pp. 140–239, Prentice Hall Press, 2nd ed., 2009.

[51] A. M. A. Ali, A. Morgan, C. Dillon, G. Patterson, S. Puckett, P. Bhoraskar, H. Dinc, M. Hensley, R. Stop, S. Bardsley, D. Lattimore, J. Bray, C. Speir, and R. Sneed, "A 16-bit 250-MS/s IF sampling pipelined ADC with background calibration," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 12, pp. 2602–2612, 2010.

[52] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," technical report, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1952.

[53] Ö. T. Demir and E. Björnson, "The Bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions [lecture notes]," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 131–136, 2021.

[54] A. B. Üçüncü and A. Ö. Yılmaz, "Oversampling in one-bit quantized massive MIMO systems and performance analysis," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 7952–7964, 2018.

[55] J. Van Vleck and D. Middleton, "The spectrum of clipped noise," *Proceedings of the IEEE*, vol. 54, no. 1, pp. 2–19, 1966.

[56] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3109–3132, 2019.

[57] M. Chen, N. Q. Hu, G. J. Qin, and Y. M. Yang, "A study on additional-signal-enhanced stochastic resonance in detecting weak signals," in *2008 IEEE Inter-*

national Conference on Networking, Sensing and Control, pp. 1636–1640, Apr. 2008.

[58] G. Harmer, B. Davis, and D. Abbott, "A review of stochastic resonance: circuits and measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, pp. 299–309, Apr. 2002.

[59] Q. Ye, H. Huang, X. He, and C. Zhang, "A study on the parameters of bistable stochastic resonance systems and adaptive stochastic resonance," in *IEEE International Conference on Robotics, Intelligent Systems and Signal Processing*, pp. 484–488, 2003.

[60] L. Yue, P. Ganesan, B. S. Sathish, C. Manikandan, A. Niranjan, V. Elamaran, and A. F. Hussein, "The importance of dithering technique revisited with biomedical images—a survey," *IEEE Access*, vol. 7, pp. 3627–3634, 2019.

[61] F. Moss, L. M. Ward, and W. G. Sannita, "Stochastic resonance and sensory information processing: a tutorial and review of application," *Clinical Neurophysiology*, vol. 115, no. 2, pp. 267–281, 2004.

[62] S. Mitaim and B. Kosko, "Adaptive stochastic resonance in noisy neurons based on mutual information," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1526–1540, Nov. 2004.

[63] A. Patel and B. Kosko, "Stochastic resonance in continuous and spiking neuron models with Levy noise," *IEEE Transactions on Neural Networks*, vol. 19, pp. 1993–2008, Dec. 2008.

[64] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "MIMO networks with one-bit ADCs: Receiver design and communication strategies," *IEEE Transactions on Communications*, vol. 70, no. 3, pp. 1580–1594, 2022.

[65] J. Liu, Z. Luo, and X. Xiong, "Low-resolution ADCs for wireless communication: A comprehensive survey," *IEEE Access*, vol. 7, pp. 91291–91324, 2019.

[66] A. B. Üçüncü and A. Ö. Yılmaz, "Performance analysis of faster than symbol rate sampling in 1-bit massive MIMO systems," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2017.

[67] Z. Shao, L. T. N. Landau, and R. C. de Lamare, "Dynamic oversampling for 1-bit ADCs in large-scale multiple-antenna systems," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3423–3435, 2021.

[68] I. Bilinskis, "Direct randomization of sampling," in *Digital Alias-Free Signal Processing*, pp. 127–138, John Wiley & Sons, Inc., 2007.

[69] A. Goldsmith, "Coding in wireless channels," in *Wireless Communications*, pp. 228–282, Cambridge University Press, 2005.

[70] Y. Shang, D. Wang, and X.-G. Xia, "Signal space diversity techniques with fast decoding based on MDS codes," *IEEE Transactions on Communications*, vol. 58, pp. 2525–2536, Sep. 2010.

[71] J. Liang, "Gradient descent and Newton's method with backtracking line search in linear regression," in *2021 2nd International Conference on Computing and Data Science (CDS)*, pp. 394–397, Jan. 2021.

[72] S. Wang, Y. Li, and J. Wang, "Multiuser detection in massive spatial modulation MIMO with low-resolution ADCs," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2156–2168, 2015.

[73] A. Goldsmith, "Multicarrier modulation," in *Wireless Communications*, pp. 374–402, Cambridge University Press, 2005.

[74] H. Lee, Y.-S. Jeon, and N. Lee, "Quantized Viterbi algorithm: Maximum likelihood sequence detection for SIMO ISI channels with low-precision ADCs," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, June 2018.

[75] G. M. Güvensen and A. Ö. Yılmaz, "A general framework for optimum iterative blockwise equalization of single carrier MIMO systems and asymptotic performance analysis," *IEEE Transactions on Communications*, vol. 61, no. 2, pp. 609–619, 2013.

[76] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," Technical Report (TR) 38.901, 3rd Generation Partnership Project, 2020. Version 16.1.0.

[77] M. Wu, D. Wuebben, A. Dekorsy, P. Baracca, V. Braun, and H. Halbauer, "Hardware impairments in millimeter wave communications using OFDM and SC-FDE," in *WSA 2016; 20th International ITG Workshop on Smart Antennas*, pp. 1–8, 2016.

[78] B.-G. Kang, S. Han, and S. Park, "Impacts of frequency selectivity on the error performance of time-domain equalizers in MIMO systems," in *2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pp. 491–494, 2016.

[79] Y.-S. Jeon, S.-N. Hong, and N. Lee, "Supervised-learning-aided communication framework for MIMO systems with low-resolution ADCs," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7299–7313, 2018.

APPENDIX A

COMPUTATIONS OF THE NONLINEAR FUNCTIONS RELATED TO THE LOG-LIKELIHOOD

Computation of $\ln(\Phi(x))$ can be a problem when x is small, and divergence towards $-\infty$ for small arguments in finite precision can be encountered. We can approximate $\ln(\Phi(x))$ using its first-order derivative to avoid possible divergent behavior as

$$\ln(\Phi(x)) \cong \ln(\Phi(c)) + \varphi(c)(x - c), \tag{A.1}$$

for $x < c \in \mathbb{R}$, where $\varphi(x) = \frac{d}{dx}\ln(\Phi(x)) = \frac{\phi(x)}{\Phi(x)}$ is the first order derivative of $\ln(\Phi(x))$. $c = -38.2$ is a suitable choice for MATLAB's arithmetic calculations since divergent behavior is observed for smaller values. However, the first and second-order derivatives $\varphi(x)$ and $\psi(x)$ also show divergent behavior around the same point c. The asymptotic behavior of $\varphi(x)$ towards $-\infty$ can be found as

$$\lim_{x\to-\infty} \frac{\phi(x)}{\Phi(x)} = \lim_{x\to-\infty} -\frac{x\phi(x)}{\phi(x)} = \lim_{x\to-\infty} -x, \tag{A.2}$$

which means that for $x < c$, $\varphi(x) = -x$ and $\frac{d^2}{dx^2}\ln(\Phi(x)) = \psi(x) = -1$ can be utilized in practice. Note that a look-up table can be generated to calculate nonlinear functions to decrease the complexity.

APPENDIX B

DERIVATION OF THE CONDITIONAL MUTUAL INFORMATION BETWEEN THE QUANTIZED OBSERVATION AND TRANSMIT SIGNAL VECTORS

Starting with the definition of conditional mutual information:

$$\mathcal{I}(\boldsymbol{r};\boldsymbol{x} \mid \boldsymbol{H},\boldsymbol{\tau}) = \mathcal{H}(\boldsymbol{r} \mid \boldsymbol{H},\boldsymbol{\tau}) - \mathcal{H}(\boldsymbol{r} \mid \boldsymbol{x},\boldsymbol{H},\boldsymbol{\tau}), \tag{B.1}$$

where $\mathcal{H}(.)$ is the entropy function. The first term in (B.1) can be calculated by the definition of conditional entropy as

$$\mathcal{H}(\boldsymbol{r} \mid \boldsymbol{H},\boldsymbol{\tau}) = -\sum_{\boldsymbol{r}\in\{\pm 1\}^{2N}} \mathrm{p}(\boldsymbol{r} \mid \boldsymbol{H},\boldsymbol{\tau}) \log_2(\mathrm{p}(\boldsymbol{r} \mid \boldsymbol{H},\boldsymbol{\tau})), \tag{B.2}$$

where $p(\boldsymbol{r} \mid \boldsymbol{H},\boldsymbol{\tau})$ can be found by averaging the likelihood function $p(\boldsymbol{r} \mid \boldsymbol{x},\boldsymbol{H},\boldsymbol{\tau}) = \prod_{n=1}^{2N} \Phi\left(\frac{r_n(\boldsymbol{h}_n^T\boldsymbol{x}-\tau_n)}{\sqrt{N_0/2}}\right)$ over all possible $\boldsymbol{x}$ vectors such that

$$\mathrm{p}(\boldsymbol{r} \mid \boldsymbol{H},\boldsymbol{\tau}) = \frac{1}{M^K} \sum_{\boldsymbol{x}'\in\mathcal{M}^{2K}} p(\boldsymbol{r} \mid \boldsymbol{x}=\boldsymbol{x}',\boldsymbol{H},\boldsymbol{\tau}). \tag{B.3}$$

Now that the first term in (B.1) is found, we can move on to the second term. Due to the conditional independence of the quantized observations given the input vector, the channel matrix, and the quantization thresholds, we can write

$$\mathcal{H}(\boldsymbol{r} \mid \boldsymbol{x},\boldsymbol{H},\boldsymbol{\tau}) = \sum_{n=1}^{2N} \mathcal{H}(r_n \mid \boldsymbol{x},\boldsymbol{H},\tau_n). \tag{B.4}$$

Then, we can find the conditional entropy as

$$\mathcal{H}(r_n \mid \boldsymbol{x}, \boldsymbol{H}, \tau_n) = \frac{1}{M^K} \sum_{\boldsymbol{x}' \in \mathcal{M}^{2K}} \mathcal{H}(r_n \mid \boldsymbol{x} = \boldsymbol{x}', \boldsymbol{H}, \tau_n), \tag{B.5}$$

for $n = 1, ..., 2N$. Again, by definition, and since each r_n is a binary random variable,

$$\begin{aligned} \mathcal{H}(r_n \mid \boldsymbol{x}, \boldsymbol{H}, \tau_n) &= \sum_{r_n \in \{\pm 1\}} p(\boldsymbol{r} \mid \boldsymbol{x}, \boldsymbol{H}, \boldsymbol{\tau}) \log_2(p(\boldsymbol{r} \mid \boldsymbol{x}, \boldsymbol{H}, \boldsymbol{\tau})) \\ &= \mathcal{H}_b \left(\Phi \left(\frac{\boldsymbol{h}_n^T \boldsymbol{x} - \tau_n}{\sqrt{N_0/2}} \right) \right). \end{aligned} \tag{B.6}$$