Advanced Precoding and Detection Techniques for Large MIMO Systems

PAN, Jiaxian

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

The Chinese University of Hong Kong
May 2014

Abstract

Multiple-input multiple-output (MIMO) transmission has been at the core of wireless communication research for the past two decades. Driven by the explosive increase of data demand, the development of MIMO systems has entered a large-scale realm where there are dozens of or even more than a hundred antennas and users. The large number of antennas can significantly boost the system throughput and robustness against noise. However, the physical realization of such a large MIMO system can be very complicated and expensive. On the one hand, optimal signal processing algorithms usually have complexities that increase rapidly in the numbers of antennas and users. On the other hand, large number of antennas means increased hardware overheads, such as those of power amplifiers and D/A converters. This thesis considers efficient precoding and detection algorithms that can reduce implementation complexity and cost. Specifically, the thesis consists of the following three parts:

In the first part, we consider a fundamental problem in MIMO communication, namely MIMO detection. The traditional lattice decoding methods, as well as its efficient approximations by lattice reduction aided (LRA) methods, relax the symbol bounds in detection and thus suffer from performance loss. We propose a systematic adaptive regularization approach to lattice decoding to alleviate the adverse effect of symbol bound relaxation, which is based on the study of a Lagrangian dual relaxation (LDR) of the optimal maximum-likelihood (ML) detector. We find an intriguing relationship between lattice decoding and ML, which was not reported in the previous literature. Simulation results show that the proposed LDR approach can significantly outperform existing lattice decoding and LRA methods.

In the second part, we consider the vector perturbation approach which is a promising technique to achieve near-sum capacity and allows simple user processing in the multiuser multiple-input single-output (MISO) downlink scenario. However, the conventional vector perturbation designs can have very high per-antenna powers, which causes significant difficulty to power amplifier implemen-

tations. To tackle this problem, we propose a vector perturbation design with per-antenna power constraints (VP-PAPC). The resulting optimization problem is an integer program which requires a computationally demanding enumeration process. Lagrangian dual relaxation is used to transform the VP-PAPC problem into standard integer least square problems which may have efficient approximations. Simulation results show that the proposed method can effectively reduce the power back-off caused by high per-antenna power in conventional vector perturbation.

In the last part, we consider constant envelope (CE) precoding in the single-user MISO downlink scenario. CE precoding is recently proposed as a mean to utilize cheap but power-efficient power amplifiers in very large MIMO systems. We provide complete solutions to some fundamental signal processing issues in CE precoding which were only partially solved in the previous literature. In addition, we enhance CE precoding with antenna subset selection for transmit optimization and implementation cost reduction. Simulation results reveal that the proposed method only exhibits moderate power loss compared to non-CE beamforming but have the advantages of CE transmission and fewer active transmitting antennas.

摘 要

多輸入多輸出傳輸在過去二十多年來無線通信研究中一直處於中心地位。 人們對信息需求的爆炸性增長導致大規模多輸入多輸出系統的出現與發展。 在大規模多輸入多輸出系統中有幾十甚至上百的天線與用戶。 這種大規模天線能夠極大地提高系統容量及對噪聲的魯棒性。 然而，大規模天線系統的物理實現卻是十分困難的。 一方面，最優的信號處理算法通常需要指數增長的複雜度。 另一方面，數目繁多的天線意味大量包括功率放大器和數模轉換器在內的硬件開銷。 這篇論文的研究重點在於能夠降低信號處理複雜度和硬件開銷的信號檢測和預編碼算法。 具體而言，本論文的研究包括三部分：

在第一部分中，我們考慮多輸入多輸出系統中的一個基本問題—信號檢測。 格型解碼是信號檢測中的一種傳統方法。 但是格型解碼（以及其快速近似算法—格基規約輔助算法)放鬆了信號檢測中的符號邊界約束因而受到性能限制。 我們提出一種自適應的正則化方法來避免格型解碼中邊界約束鬆弛帶來的負面影響。 這種方法是基於最大似然解碼器的拉格朗日對偶鬆弛。 我們發現了格型解碼和最大似然解碼的一個十分有趣的關係，而這個關係在現有的文獻中並沒有被提及。 數值仿真結果顯示拉格朗日對偶鬆弛方法比現有的格型解碼更為優勝。

在第二部分中，我們考慮多用戶信號廣播中的矢量擾動方法。 矢量擾動是一種能夠接近信道總容量以及簡化用戶數據處理方法。 然而，傳統的矢量擾動會導致每根傳輸天線上都有相當大的功率， 導致天線模擬前端的硬件實現有相當大的難度。 我們提出一種每天線功率受限的矢量擾動方法來解決這個問題。 在這個方法中，我們需要解決一個整數規劃問題。 然而，求解這個整數規劃問題需要用到複雜度十分高的枚舉算法。 我們用拉格朗日對偶鬆弛方法把這個整數規劃轉化為標準的整數最小二乘問題，然後採用快速的近似算法來求解。 數值仿真顯示提出的方法能夠顯著地降低高每天線功率造成的功率回饋。

在最後一部分，我們考慮單用戶通信中的恆定包絡預編碼。 恆定包絡預編碼是一種最近被提出用於超大規模多輸入多輸出系統的方法。恆定包絡預編碼的優點在於能夠利用價格低廉但是功率效率高的功率放大器。 但是恆定

包絡預編碼中的一些信號處理問題在之前的文獻中只是得到了部分解答。 我們為這些信號處理問題提供了一個完整的解決方案。 更進一步地，我們用天線子集選擇來加强恆定包絡預編碼以優化天線傳輸信號及進一步降低天線成本。 數值仿真結果顯示包絡預編碼的性能只稍遜於傳統的波束成型方法，但是能恆定包絡傳輸和降低活動的天線數目。

Acknowledgements

I am much indebted to my supervisor, Prof. Wing-Kin Ma, for his invaluable guidance through my PhD study. I sincerely appreciate his patience, encourage and immense knowledge. I have learnt from him a lot on the art of research, teaching, writing and presentation. His persistence for perfection and rigorous scholarship have influenced me significantly. I would also like to express my thanks to Prof. Joakim Jaldén for his kind advices on my research. I would like to thank Prof. Tan Lee and Prof. Pak-Chung Ching for creating a friendly environment in the Digital Signal Processing Laboratory (DSP Lab) in the Chinese University of Hong Kong.

Many thanks go to my friends in DSP lab. I would like to thank Yujia Li, Meng Yuan, Feng Tian, Lan Shuai, Hongying Zheng, Ning Wang and Houwei Cao who gave me lots of suggestions on study and life in my early days at DSP lab. I am much grateful to Qiang Li, Xiaoxiao Wu, Hoi-To Wai, Ka-Kit Lee, Xiao Fu, Hing-Yin Tseng, and Haipeng Wang for all the insightful discussion on research.

Last but not least, I wish to express my immense gratitude to my parents, sisters, and brothers-in-law. In particular, I would like to thank Kexian and Xiaobing for encouraging me to start the PhD study, and Shuxian, Junfeng, Jiexian and Jianpeng for their constant supports during the past five years. Finally, I would like to express my deepest gratitude and appreciation to my beloved girlfriend Josephine for her love and encouragement.

This work is dedicated to my family.

Contents

List of Figures

List of Tables

Abbreviations

s.t.	subject to
SER	symbol error rate
SNR	signal to noise ratio
CSIT	channel state information at the transmitter
MIMO	multiple-input multiple-output
MISO	multiple-input single-output
SISO	single-input single-output
ML	maximum likelihood
SDR	semidefinite relaxation
LD	lattice decoder
NLD	naive lattice decoder
MMSE	minimum mean square error
LRA	lattice reduction-aided
SD	sphere decoding
LDR	Lagrangian dual relaxation
PS	projected subgradient
CE	constant envelope
MRT	maximum ratio transmission
PAPR	peak-to-average power ratio
AS	antenna-subset selection
UA	unequal amplitude
CVP	conventional vector perturbation
PAPC	per-antenna power constraint

Notations

$\mathbb{Z}$	- the set of integers
$\mathbb{R}$	- the set of real numbers
$\mathbb{C}$	- the set of complex numbers
$\mathbb{G}$	- the set of Gaussian numbers
$\mathbb{S}^N$	- the set of all N by N symmetric matrices
$\|\boldsymbol{x}\|_p$	- ℓ_p norm of a vector $\boldsymbol{x}$
$\boldsymbol{A}^H$	- the Hermitian (conjugate) transpose of $\boldsymbol{A}$
$\boldsymbol{A}^T$	- the transpose of $\boldsymbol{A}$
$\boldsymbol{A}^\dagger$	- the pseudo-inverse of $\boldsymbol{A}$
$\Pr\{a\}$	- the probability of event a
$\mathcal{CN}(\boldsymbol{u}, \boldsymbol{\Omega})$	- multivariate complex Gaussian distribution with mean $\boldsymbol{u}$ and covariance $\boldsymbol{\Omega}$
$\boldsymbol{A} \succeq \boldsymbol{B}$	- $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite
$\boldsymbol{a} \succeq \boldsymbol{b}$	- $\boldsymbol{a} - \boldsymbol{b}$ is element-wise nonnegative
$\|\boldsymbol{x}\|$	- the element-wise absolute value
$\mathrm{tr}(\boldsymbol{A})$	- the trace of the matrix $\boldsymbol{A}$
$\odot$	- the element-wise product
log	- the binary logarithm
$\mathbb{E}[\cdot]$	- the expectation operator
$\boldsymbol{D}(\cdot)$	- the diagonal operator
$\boldsymbol{j}$	- $\boldsymbol{j} = \sqrt{-1}$, the imaginary unit

Chapter 1

Introduction

1.1 MIMO Systems

The last two decades have seen dramatic demands for high speed wireless networks. Enormous endeavors from both academia and industry have focused on designing new communication systems to meet such demands. Multiple-input multiple-output (MIMO) systems, where both the transmitter and the receivers are equipped with multiple antennas, have been recognized as a major technique to boost system performance.

In the single-user (or point-to-point) scenario, MIMO systems are known to provide two types of gains — multiplexing gain and diversity gain. The multiplexing gain reflects the ability for an MIMO system to increase the data rate. It is shown in [1] that under an i.i.d. Rayleigh fading channel, an MIMO system with N transmit and M receive antennas can achieve a capacity asymptotically increasing as $\min\{N, M\}\log(\text{SNR})$, where SNR denotes the signal-to-noise ratio. Compared with a single-antenna system whose capacity is asymptotically close to $\log(\text{SNR})$, we can see that the MIMO system can increase the capacity by a factor of $\min\{M, N\}$ which is known as the multiplexing gain. Apart from the ability of increasing capacity, MIMO systems are also able to reduce the system error rate dramatically. The error rate of an MIMO system usually behaves approximately as $\frac{G}{\text{SNR}^d}$ for some constant G and d. The number d is called the diversity gain which describes the speed of the error rate approaching zero. The maximum diversity gain of an MIMO system is MN which is far superior to the

diversity of one in single-antenna systems [2]. In addition to single-user communication, MIMO techniques are also very useful in multiuser communications where a base station simultaneously serves multiple users. It is shown in [3, 4] that in the scenario of an N-antenna base station broadcasting informations to K single-antenna users, the sum capacity scales linearly in $\min\{K, N\}$.

Owing to these salient benefits of MIMO systems, the development trend of MIMO systems is to equip the base station with more antennas. The WLAN 802.11ac and the cellular network LTE-Advanced systems support 8 antennas at the base station. Systems operating in millimeter wave are expected to have 10 - 60 antennas [5–7]. Samsung announced in 2013 a 64-antenna adaptive array transceiver which is expected to be a core technology in the fifth generation of cellar networks [8]. Field tests of linear and circular arrays of 128 elements have been done in [9, 10]. Researchers envision that wireless communication systems would evolve into a massive scale with hundreds of antennas [11].

1.2 Motivation and Contribution of This Thesis

The realization of the performance gains of MIMO systems promised by theoretical results, however, is very complicated and expensive. The increasingly large number of antennas also means increasing difficulty in processing the transmit and receive signals. Typical operations of a transceiver include symbol synchronization, channel estimation, signal detection/precoding and channel encoding/decoding. Many signal processing problems for theses operations are not known to have fast algorithms. In fact, many of the problems, such as those in signal detection [12] and precoder designs [13, 14], are known to be NP-hard, which means that it is unlikely to find polynomial-time exact solutions for those problems. The high complexity of signal processing also complicates hardware implementations which in practice are highly constrained by space, weight, power and cost. On the other hand, each antenna of an MIMO system has its own analog frontend which includes impedance matching circuit, band-pass filter, digital-to-analog converter, and power amplifiers. Thus, an MIMO system requires lots of such expensive hardware. Finally, an MIMO system will

draw much power in daily operations which would put a financial burden on service providers and cause environmental impacts.

This thesis considers efficient high-performance detection and precoding algorithms for MIMO systems in various communication scenarios. The proposed algorithms also consider hardware-friendly schemes that can reduce hardware costs and improve power efficiency.

1.2.1 MIMO Detection

MIMO detection refers to the task of separating the transmitting signals from the observation of the receive signals. MIMO detection is a fundamental problem in communication as it is involved in a plethora of communication scenarios including single-user and multiuser uplink communications. If the noise follows a Gaussian distribution, the detection problem amounts to a least square problem with a finite-alphabet constraint. Generally, this detection problem is known to belong to the class of NP-hard problems. Attracted by the importance and hardness of MIMO detection, researchers have developed several classes of algorithms to tackle the MIMO detection problem. Notable detectors include sphere decoders [12, 15, 16], linear detectors [17, 18], semidefinite relaxation detectors [19–24], lattice decoders [15, 25–28], and lattice reduction-aided (LRA) detectors [29–35].

In Chapter 3 of this thesis, we investigate lattice decoding under PAM constellations. Our contributions lie in proposing a novel regularization approach in lattice decoding for tackling the out-of-bound symbol effects for lattice decoding. Specifically, our contributions are listed as follows.

- We propose a systematic approach for determining the regularization variable in lattice decoding by considering a Lagrangian dual relaxation (LDR) of the maximum-likelihood (ML) detection problem. We find that the LDR is to find the best diagonally regularized lattice decoding to approximate the ML detector, and all diagonal regularizations, including the traditional MMSE regularization, can be subsumed under the LDR formalism. We devise new detection algorithms which are based on the projected sub-

gradient method in finding the best regularization for lattice decoding. Simulation results show that the proposed LDR detectors can outperform the conventional lattice decoder and LRA detectors.

1.2.2 Transmit Precoding

Transmit precoding has been recognized as a major technique in MIMO systems to reduce the error rate and to approach the capacity promised by information theory. Some of the representative precoding strategies are space-time coding [2,36,37], beamforming [14,38], and vector perturbation [13,39]. The design of a precoding algorithm, however, must take into account hardware constraints. A notable one is the per-antenna power constraint (PAPC). The reason of considering PAPCs is that in practice each antenna has its own analog frontend. Each analog frontend, which includes a D/A converter and power amplifier, has its own operating region. If the signal input into the analog frontend exceeds the operating region, then low power efficiency, nonlinear amplification or even signal clipping may happen. Therefore, PAPC is an important design constraint in transmit precoder design. Beamforming with PAPCs has been considered in multiuser broadcast channels [40–43] and in the multicell scenario [44, 45]. Vector perturbation with PAPC is also considered in [46]. All these works focus on putting a power upper bounds on the average or instantaneous per-antenna power. A different form of PAPC is considered in [47–49] where the instantaneous power of the transmitting signal is further restricted to be constant. This means that the transmitting signal at each antenna is of constant envelope (CE). CE precoding, compared to beamforming and vector perturbation, can utilize cheap but highly power-efficient power amplifiers which are a must in large-scale MIMO systems. Following the convention in the literature, in the rest of thesis we will refer PAPCs as the kind of power constraints that upper bound the averaged or maximum per-antenna power, though literally PAPC means any form of constraints on the per-antenna power such as CE precoding.

In Chapter 4 we consider vector perturbation with PAPCs in multiuser multiple-input single-output (MISO) broadcast channels, and in Chapter 5 we

consider CE precoding in single-user MISO channels. Our contribution lies in proposing novel transmit strategies and devising fast algorithms. Specifically, our contributions are listed as follows.

- We propose a vector perturbation formulation with strict maximum per-antenna power constraint in the multiuser MISO broadcast channels. We show that the proposed formulation can achieve full transmit diversity in Gaussian fading channels. An efficient algorithm based on LDR and LRA methods is proposed. Simulation results show that the proposed methods can significantly reduce the power back-off due to high per-antenna power and have better error rate performance than the conventional vector perturbation method.

- We consider CE precoding in single-user MISO channels. We solve a fundamental problem in CE precoding — characterization of the noise-free receive signal region. We derive a simple, efficient and exact CE precoder algorithm whose complexity is linear in the number of antennas.

1.3 Organization of This Thesis

The organization of this thesis is as follows.

Chapter 2 introduces the MIMO detection problem and reviews several major MIMO detectors.

Chapter 3 proposes a novel MIMO detector based on the LDR of the ML detector. We also investigate the relationship between LDR and the regularized lattice decoder. The implementation of the LDR is elaborated and fast approximations are proposed.

Chapter 4 considers the multiuser broadcast scenario where we propose the VP-PAPC formulations. The feasibility and diversity of the VP-PAPC problems are investigated. We also derive the efficient algorithms based on the LDR and LRA techniques.

Chapter 5 focuses on CE precoding in the single-user MISO channels. This chapter is divided into two parts. First, we investigate the receive signal struc-

ture and propose a fast exact algorithm for CE precoding. We then develop two transmit strategies based on CE precoding.

Chapter 6 summarizes this thesis and presents our perspectives on future directions.

Chapter 2

MIMO Detection

This chapter reviews the MIMO detection problem.

2.1 System Model

We consider a single-user MIMO system where the transmitter and receiver are equipped with N and M antennas, respectively. Fig. 2.1 depicts the scenario.

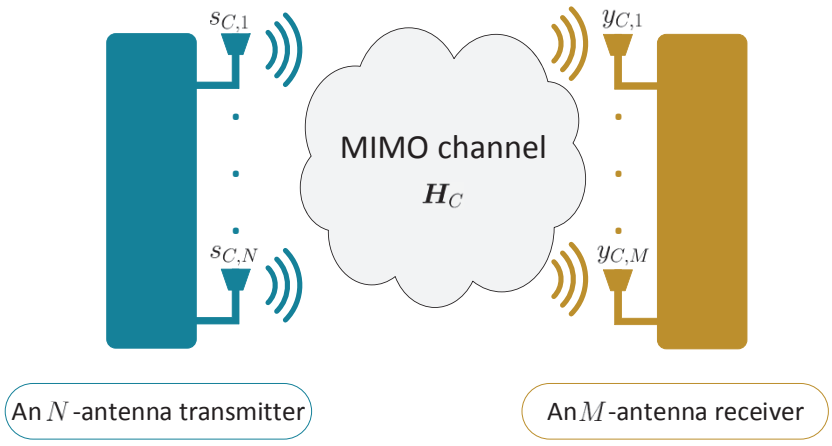


Figure 2.1: A single-user MIMO system with spatial multiplexing

The transmitter sends N complex-valued information symbols, denoted by $s_{C,i} \in \mathcal{S}_C$, $i = 1, \ldots, N$, where $\mathcal{S}_C \subset \mathbb{C}$ is the constellation set. Specifically, the ith symbol $s_{C,i}$ is transmitted by the ith antenna. The receive signal at the mth

receive antenna is given by

$$y_{C,j} = \sum_{i=1}^{N} h_{C,ji} s_{C,i} + \nu_{C,m} \tag{2.1}$$

where $h_{C,ji} \in \mathbb{C}$ is the channel coefficient from the ith transmit antenna to the jth receive antenna, and $\nu_{C,j} \in \mathbb{C}$ is the noise at the jth receive antenna. Collecting all receive signal $y_{C,j}$, we can write compactly the channel input-output relationship as

$$\boldsymbol{y}_C = \boldsymbol{H}_C \boldsymbol{s}_C + \boldsymbol{\nu}_C, \tag{2.2}$$

where $\boldsymbol{y}_C = [y_{C,1}, \ldots, y_{C,M}]^T$, $\boldsymbol{s}_C = [s_{C,1}, \ldots, s_{C,N}]^T$, $\boldsymbol{\nu}_C = [\nu_{C,1}, \ldots, \nu_{C,M}]^T$, and $\boldsymbol{H}_C \in \mathbb{C}^{M \times N}$ has $h_{C,ji}$ at its (j,i) position. We assume that the noise follows the distribution $\mathcal{CN}(\boldsymbol{0}, \sigma_{\nu_C}^2 \boldsymbol{I})$ and the constellation $\mathcal{S}$ is the standard $(u+1)^2$-QAM (u is a positive odd number), i.e. $\mathcal{S}_C = \mathcal{S} + \boldsymbol{j}\mathcal{S}$ with $\mathcal{S}$ being the $(u+1)$-PAM constellation $\mathcal{S} = \{\pm 1, \pm 3, \ldots, \pm u\}$. This model is known as the single-user spatial multiplexing system [50].

The signal model (2.2) can also describe the multi-user uplink scenario depicted in Fig. 2.2. In this case, N single-antenna users transmit simultaneously

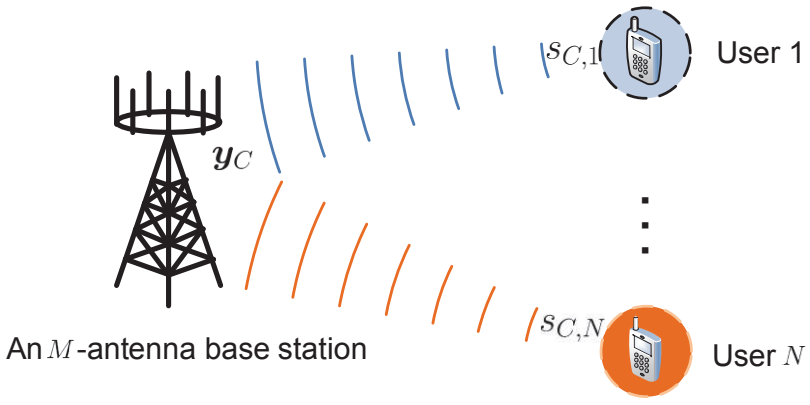


Figure 2.2: A multi-user MISO uplink system

to an M-antenna base station. The transmitting signal of the ith user is $s_{C,i}$. The receive signal at the jth antenna of the base station is also given by (2.1) with $h_{C,ji}$ being the channel coefficient between the ith user and the jth receive antenna. This results in exactly the same channel input-output relationship

as (2.2). In fact, model (2.2) is not new. It has been used in a plethora of communication scenarios, such as multiuser code division multiple access (CDMA) [19], space-time coding [36] in multi-antenna frequency-flat channels, space-frequency coding in multi-antenna orthogonal frequency division multiplexing (OFDM) [51], relay networks [52] and most recently, very large-scale antenna systems [11].

2.2 MIMO Detectors

The goal of MIMO detection is to detect the transmitted signal given the observation of the receive signal and the channel. For convenience in the subsequent development, let us convert the complex signal model (2.2) to a real one as follows

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{\nu} \tag{2.3}$$

where

$$\begin{aligned} \boldsymbol{y} &= \begin{bmatrix} \Re\{\boldsymbol{y}_C\} \\ \Im\{\boldsymbol{y}_C\} \end{bmatrix}, \quad \boldsymbol{H} = \begin{bmatrix} \Re\{\boldsymbol{H}_C\} & -\Im\{\boldsymbol{H}_C\} \\ \Im\{\boldsymbol{H}_C\} & \Re\{\boldsymbol{H}_C\} \end{bmatrix}, \\ \boldsymbol{s} &= \begin{bmatrix} \Re\{\boldsymbol{s}_C\} \\ \Im\{\boldsymbol{s}_C\} \end{bmatrix}, \quad \boldsymbol{\nu} = \begin{bmatrix} \Re\{\boldsymbol{\nu}_C\} \\ \Im\{\boldsymbol{\nu}_C\} \end{bmatrix}. \end{aligned} \tag{2.4}$$

The input and output problem sizes of the real model (2.3) are $N = 2N_C$ and $M = 2M_C$, respectively. The constellation of the information symbols $\boldsymbol{s}$ is the $(u+1)$-PAM constellation

$$\mathcal{S} = \{\pm 1, \pm 3, \ldots, \pm u\}.$$

The ML MIMO detector, which is optimal in minimizing the vector error probability of detecting $\boldsymbol{s}$ given knowledge of the channel $\boldsymbol{H}$, is the solution of the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{s}} \quad & \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 \\ \text{s.t.} \quad & s_i \in \{\pm 1, \pm 3, \ldots, \pm u\}, \ i = 1, \ldots, N. \end{aligned} \tag{2.5}$$

where $\|\cdot\|_2$ denotes the 2-norm. The ML problem (2.5) is known to be NP-hard for general $(\boldsymbol{y}, \boldsymbol{H})$. This implies that all the existing exact ML solvers, including

the well-known sphere decoders [12, 15], would be computationally prohibitive when N is large. In fact, it is shown in [53] that the sphere decoder exhibits exponential complexity with respect to the problem size N. The computational difficulty in solving the ML problem exactly has stimulated a number of works that aim to approximate the ML detector in an efficient manner. In the following subsections, we will review several major classes of MIMO detectors.

2.2.1 Linear Detector

Generally, a linear detector can be represented by

$$\begin{aligned} \tilde{\boldsymbol{s}} &= \boldsymbol{G}\boldsymbol{y} \\ \hat{\boldsymbol{s}} &= \mathcal{Q}(\tilde{\boldsymbol{s}}), \end{aligned} \tag{2.6}$$

where $\boldsymbol{G} \in \mathbb{R}^{N\times M}$, and $\mathcal{Q}$ is the element-wise quantization with respect to the constellation $\mathcal{S}$. It can be seen that linear detectors only involve a matrix-vector product and a quantization process, both of which can be computed very efficiently. Zero-forcing (ZF) and minimum-mean-square-error (MMSE) detectors are the most common linear detectors. In ZF, the matrix $\boldsymbol{G}$ is the channel pseudo-inverse $(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T$, which results in the following detector

$$\begin{aligned} \tilde{\boldsymbol{s}}_{\mathrm{ZF}} &= (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{y} \\ \hat{\boldsymbol{s}}_{\mathrm{ZF}} &= \mathcal{Q}(\tilde{\boldsymbol{s}}_{\mathrm{ZF}}). \end{aligned} \tag{2.7}$$

The name ZF follows from the fact that the channel pseudo-inverse $(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T$ totally eliminates intersymbol interference (ISI) in $\tilde{\boldsymbol{s}}_{\mathrm{ZF}}$. To see this, let us substitute (2.3) into (2.7) and obtain

$$\tilde{\boldsymbol{s}}_{\mathrm{ZF}} = \boldsymbol{s} + (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\nu}. \tag{2.8}$$

We can see that each transmitted symbol s_i only appears in $[\tilde{\boldsymbol{s}}_{\mathrm{ZF}}]_i$ and zero ISI is achieved. Though ZF avoids ISI, it also amplifies the effective noise $(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\nu}$ significantly, especially when $\boldsymbol{H}$ is ill-conditioned.

In order to strike a balance between ISI and effective noise, the MMSE detector minimizes the mean square error (MSE) between $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$. The matrix $\boldsymbol{G}$ is obtained by the minimizer of the following problem

$$\min_{\boldsymbol{G}\in\mathbb{C}^{N\times M}} \ \mathbb{E}_{\boldsymbol{s},\boldsymbol{\nu}}[\|\boldsymbol{s} - \boldsymbol{G}\boldsymbol{y}\|_2^2]. \tag{2.9}$$

This results in the matrix $\boldsymbol{G} = (\boldsymbol{H}^T\boldsymbol{H} + \frac{\sigma_\nu^2}{\sigma_s^2}\boldsymbol{I})^{-1}\boldsymbol{H}^T$, where σ_s^2 and σ_ν^2 denote the symbol and noise variance, respectively. Therefore, the MMSE detector is given by

$$\begin{aligned} \tilde{\boldsymbol{s}}_{\text{MMSE}} &= (\boldsymbol{H}^T\boldsymbol{H} + \tfrac{\sigma_\nu^2}{\sigma_s^2}\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y} \\ \hat{\boldsymbol{s}}_{\text{MMSE}} &= \mathcal{Q}(\tilde{\boldsymbol{s}}_{\text{MMSE}}). \end{aligned} \tag{2.10}$$

2.2.2 Decision Feedback Detector

Though the ZF and MMSE detectors are easy to implement, their performances are far from being optimal. The decision feedback (DF) technique can be used to enhance the performance of linear detectors. To describe the ZF-DF detector, let us denote the QR decomposition of $\boldsymbol{H}$ as $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} \in \mathbb{R}^{M\times N}$ is orthogonal and $\boldsymbol{R} \in \mathbb{R}^{N\times N}$ is upper triangular. Then the ZF detector can be equivalently rewritten as

$$\begin{aligned} \tilde{\boldsymbol{s}}_{\text{ZF}} &= \boldsymbol{R}^{-1}\tilde{\boldsymbol{y}} \\ \hat{\boldsymbol{s}}_{\text{ZF}} &= \mathcal{Q}(\tilde{\boldsymbol{s}}), \end{aligned} \tag{2.11}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{Q}^T\boldsymbol{y}$. By noting that $\boldsymbol{R}\tilde{\boldsymbol{s}}_{\text{ZF}} = \tilde{\boldsymbol{y}}$ and $\boldsymbol{R}$ is upper triangular, we can express $[\hat{\boldsymbol{s}}_{\text{ZF}}]_i$ as

$$[\hat{\boldsymbol{s}}_{\text{ZF}}]_i = \mathcal{Q}\left(\frac{\tilde{y}_i - \sum_{j=i+1}^{N} r_{ij}[\tilde{\boldsymbol{s}}_{\text{ZF}}]_j}{r_{ii}}\right) \tag{2.12}$$

for $i = N$ to $i = 1$. The idea of DF is to replace $[\tilde{\boldsymbol{s}}_{\text{ZF}}]_j$ by its decision $[\tilde{\boldsymbol{s}}_{\text{ZF}}]_j = \mathcal{Q}([\tilde{\boldsymbol{s}}_{\text{ZF}}]_j)$, which results in the following ZF-DF detector

$$[\hat{\boldsymbol{s}}_{\text{ZF-DF}}]_i = \mathcal{Q}\left(\frac{\tilde{y}_i - \sum_{j=i+1}^{N} r_{ij}[\hat{\boldsymbol{s}}_{\text{ZF-DF}}]_j}{r_{ii}}\right) \tag{2.13}$$

for $i = N$ to $i = 1$. The MMSE-DF detector can be obtained in a similar way by replacing the matrix $\boldsymbol{R}$ in (2.11)-(2.13) by the Cholesky factor of $\boldsymbol{H}^T\boldsymbol{H} + \frac{\sigma_\nu^2}{\sigma_s^2}\boldsymbol{I}$. It should be noted that though seemingly the ZF-DF and MMSE-DF detectors involve more operations than the ZF and MMSE detectors, their complexities are actually exactly the same.

2.2.3 Semidefinte Relaxation Detector

Several semidefinite relaxation (SDR) detectors have been proposed in [19–22, 24, 54] and analyzed in [55–57]. Here, we describe a representative SDR detector— the *bound-constrained semidefinite relaxation* (BC-SDR) [54]. A key reason why BC-SDR is representative is that BC-SDR is shown to be equivalent to two other relaxations, namely, the polynomial-inspired SDR (PI-SDR) [22] and virtually-antipodal SDR [21], which employ different ideas to relax and offer different insights in ML approximation; see [58] for details. For example, there are theoretically proven results on the approximation accuracy of VA-SDR [57, 59], and those results apply to BC-SDR (and also PI-SDR) by using the equivalence of the three SDRs.

To derive the BC-SDR, let us rewrite the ML detector as follows

$$\min_{\boldsymbol{s}\in\mathbb{R}^N} \ \mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{s}\boldsymbol{s}^T) - 2\boldsymbol{s}^T\boldsymbol{H}^T\boldsymbol{y} + \|\boldsymbol{y}\|_2^2 \tag{2.14a}$$

$$\text{s.t. } s_i^2 \in \{1^2, 3^2, \ldots, u^2\},\ i = 1, \ldots, N. \tag{2.14b}$$

Introducing a redundant constraint $\boldsymbol{S} = \boldsymbol{s}\boldsymbol{s}^T$ into (2.14), we turn (2.14) equivalently to

$$\min_{\boldsymbol{S}\in\mathbb{S}^N, \boldsymbol{s}\in\mathbb{R}^N} \ \mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{S}) - 2\boldsymbol{s}^T\boldsymbol{H}^T\boldsymbol{y} + \|\boldsymbol{y}\|_2^2 \tag{2.15a}$$

$$\text{s.t. } \boldsymbol{S} = \boldsymbol{s}\boldsymbol{s}^T, \tag{2.15b}$$

$$S_{ii} \in \{1, 3^2, \ldots, u^2\},\ i = 1, \ldots, N. \tag{2.15c}$$

where $\mathbb{S}^N$ denotes the set of all $N \times N$ real symmetric matrices. Both the constraints (2.15b) and (2.15c) are nonconvex and thus are not easy to deal with. The BC-SDR replaces these two difficult constraints with other easier ones. Eq. (2.15b) is replaced by $\boldsymbol{S} \succeq \boldsymbol{s}\boldsymbol{s}^T$ and (2.14b) is replaced by $1 \le S_{ii} \le u^2$ for $i = 1, \ldots, N$, where the notation $\boldsymbol{A} \succeq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite (PSD). This results in the following convex problem

$$\min_{\boldsymbol{S}\in\mathbb{S}^N, \boldsymbol{s}\in\mathbb{R}^N} \ \mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{S}) - 2\boldsymbol{s}^T\boldsymbol{H}^T\boldsymbol{y} + \|\boldsymbol{y}\|_2^2 \tag{2.16a}$$

$$\text{s.t. } \boldsymbol{S} \succeq \boldsymbol{s}\boldsymbol{s}^T, \tag{2.16b}$$

$$1 \le S_{ii} \le u^2,\ i = 1, \ldots, N. \tag{2.16c}$$

Problem (2.16) can be efficiently solved exactly by available general-purpose interior-point softwares [60, 61]. Specialized algorithm that runs much faster is also developed in [62]. After solving (2.16), one needs to convert the optimal solution $(\boldsymbol{S}^\star, \boldsymbol{s}^\star)$ of (2.16) to a decision $\hat{\boldsymbol{s}}_{\mathrm{SDR}}$. A simple approach is a direct quantization by

$$\hat{\boldsymbol{s}}_{\mathrm{SDR}} = \mathcal{Q}(\boldsymbol{s}^\star).$$

Better performance can be obtained by using the randomization technique [22]. The numerical results in [58] show that with a reasonable number of randomizations, the performance of BC-SDR could be closed to the ML detector.

2.2.4 Lattice Decoder

The study of lattice decoding for MIMO detection has received much attention [15, 25–28], owing to its good tradeoff between detection accuracy and complexity. A simple version of lattice decoding is naive lattice decoding (NLD) [25], where, instead of dealing with the ML problem (2.5), one considers a lattice decoding problem

$$\begin{aligned} \min_{\boldsymbol{s}} \quad & \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{s} \in 2\mathbb{Z}^N + \mathbf{1}, \end{aligned} \tag{2.17}$$

where $\mathbf{1}$ denotes an all-one vector of appropriate length, and $\mathbb{Z}$ is the set of all integers. The NLD problem is an unbounded relaxation of the ML problem—it ignores the symbol bound constraints $-u \le s_i \le u$ in the original ML problem, but keeps the symbols s_i in the discrete set $2\mathbb{Z}+1$. The reason for doing this is to facilitate the use of an efficient processing technique, namely, *lattice reduction*.

In lattice reduction, the channel matrix $\boldsymbol{H}$ is transformed to another channel matrix

$$\tilde{\boldsymbol{H}} = \boldsymbol{H}\boldsymbol{U}, \tag{2.18}$$

where $\tilde{\boldsymbol{H}}$ is called a lattice-reduced channel, and $\boldsymbol{U}$ is a unimodular matrix; i.e., $\boldsymbol{U} \in \mathbb{Z}^{N\times N}$ and $|\det(\boldsymbol{U})| = 1$. Loosely speaking, the transformation (2.18) is designed such that the transformed channel matrix $\tilde{\boldsymbol{H}}$ may be better conditioned with short and roughly orthogonal column vectors [63, 64]. A popular algorithm

for the lattice reduction process (2.18) is the Lenstra-Lenstra-Lovász (LLL) reduction [64,65]. Since a unimodular $\boldsymbol{U}$ satisfies $\boldsymbol{U}^{-1}\mathbb{Z}^N = \mathbb{Z}^N$, the symbol vector $\boldsymbol{s}$ can be transformed to $\tilde{\boldsymbol{s}} = \boldsymbol{U}^{-1}\boldsymbol{s}$ and the domain of $\tilde{\boldsymbol{s}}$ is $\tilde{\boldsymbol{s}} \in 2\mathbb{Z}^N + \boldsymbol{U}^{-1}\mathbf{1}$. Subsequently, we can equivalently recast the NLD problem (2.17) as

$$\begin{aligned} \min_{\tilde{\boldsymbol{s}}} \quad & \|\boldsymbol{y} - \tilde{\boldsymbol{H}}\tilde{\boldsymbol{s}}\|_2^2 \\ \text{s.t.} \quad & \tilde{\boldsymbol{s}} \in 2\mathbb{Z}^N + \boldsymbol{U}^{-1}\mathbf{1}, \end{aligned} \tag{2.19}$$

which is also a lattice decoding problem, but with a "better" channel $\tilde{\boldsymbol{H}}$. The equivalent lattice decoding problem (2.19) is then solved by applying a sphere decoder for unbounded integers. Empirical and theoretical results in [15] and [28, 66] show that lattice reduction can boost the speed for a sphere decoder to solve problem (2.19).

While lattice reduction provides an attractive way to handle the NLD problem (2.17), it also transforms the symbol bound set $\{\boldsymbol{s} \mid -u \leq s_i \leq u, \ \forall\, i\}$ to a complicated polyhedron $\{\tilde{\boldsymbol{s}} = \boldsymbol{U}^{-1}\boldsymbol{s} \mid -u \leq s_i \leq u, \ \forall\, i\}$. There is no known way for a sphere decoder to efficiently manage such a polyhedron bound constraint, and, for this reason, the latter is ignored in NLD. As a consequence, NLD may output an out-of-bound symbol decision. Unfortunately, the out-of-bound symbol events can be detrimental to the system error rate performance. It is shown that NLD fails to achieve the optimal diversity-multiplexing tradeoff (DMT) in general MIMO system models [25, 26]. This drawback has motivated endeavors that study a regularized version of the NLD problem [25, 27]

$$\begin{aligned} \min_{\boldsymbol{s}} \quad & \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T\boldsymbol{T}\boldsymbol{s} \\ \text{s.t.} \quad & \boldsymbol{s} \in 2\mathbb{Z}^N + \mathbf{1}, \end{aligned} \tag{2.20}$$

where $\boldsymbol{T}$ is a positive-semidefinite regularization matrix. The addition of the regularization term $\boldsymbol{s}^T\boldsymbol{T}\boldsymbol{s}$ penalizes symbol vectors $\boldsymbol{s}$ that are far away from the origin, thereby attempting to constrain the optimal solutions of the regularized lattice decoding problem (2.20) within the symbol bounds in an implicit manner. The analysis in [27] shows that regularization can alleviate the negative effect of having no explicit symbol bound constraints, and the regularized lattice decoding with a positive-definite $\boldsymbol{T}$ can achieve the same DMT as the true ML detector. A

well-known choice of $\boldsymbol{T}$ is the MMSE regularization matrix $\boldsymbol{T} = \sigma_\nu^2/\sigma_s^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. The corresponding lattice decoding is MMSE LD.

To solve the regularized lattice decoding problem (2.20), one first reformulate (2.20) as an integer least squares problem. Let $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{T}$ denote a square-root decomposition of $\boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{T}$, where $\boldsymbol{V}$ is the corresponding square-root factor. Consider a variable transformation $\boldsymbol{s} = 2\boldsymbol{z} + \boldsymbol{1}$ with $\boldsymbol{z} \in \mathbb{Z}^N$. Then, (2.20) is rewritten as

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & \|\boldsymbol{f} - \boldsymbol{V}\boldsymbol{z}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{z} \in \mathbb{Z}^N, \end{aligned} \tag{2.21}$$

where $\boldsymbol{f} = \frac{1}{2}\boldsymbol{V}^{-T}(\boldsymbol{H}^T\boldsymbol{y} - (\boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{T})\boldsymbol{1})$. Lattice reduction is then applied to the matrix $\boldsymbol{V}$ to obtain

$$\tilde{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{U}$$

where $\tilde{\boldsymbol{V}}$ is a lattice-reduced matrix and $\boldsymbol{U}$ is the corresponding unimodular matrix. With another variable transformation $\tilde{\boldsymbol{z}} = \boldsymbol{U}^{-1}\boldsymbol{z}$, we further rewrite (2.21) as

$$\begin{aligned} \min_{\tilde{\boldsymbol{z}}} \quad & \|\boldsymbol{f} - \tilde{\boldsymbol{V}}\tilde{\boldsymbol{z}}\|_2^2 \\ \text{s.t.} \quad & \tilde{\boldsymbol{z}} \in \mathbb{Z}^N. \end{aligned} \tag{2.22}$$

Now, we can apply a sphere decoder for unbounded integers to solve (2.22) exactly. The optimal solution of (2.20) can be obtained via the relationship

$$\boldsymbol{s} = 2\boldsymbol{U}\tilde{\boldsymbol{z}} + \boldsymbol{1}.$$

2.2.5 Lattice Reduction-Aided Detector

Lattice reduction-aided (LRA) detectors [29–35] are fast approximations of the lattice decoders. The most expensive operation in lattice decoding is the sphere decoding in solving (2.22). The idea of LRA detectors is to replace sphere decoding with other fast detectors. Two notable detectors are the linear and DF detectors.

When linear detector is adopted, an approximate solution of (2.22) is given by

$$\hat{\tilde{\boldsymbol{z}}}_{\text{LRA}-\text{linear}} = \lfloor \tilde{\boldsymbol{V}}^{-1}\boldsymbol{f} \rceil. \tag{2.23}$$

where $\lfloor \cdot \rceil$ denotes the element-wise integer rounding. The resulting detectors with zero regularization matrix $\boldsymbol{T} = \boldsymbol{0}$ and MMSE regularization matrix $\boldsymbol{T} = \frac{\sigma_v^2}{\sigma_s^2}\boldsymbol{I}$ are known as *ZF LRA-linear* and *MMSE LRA-linear* respectively in the literature.

To employ the DF detector, we first rewrite (2.22) as

$$\begin{aligned} \min_{\tilde{\boldsymbol{z}}} \quad & \|\tilde{\boldsymbol{f}} - \tilde{\boldsymbol{R}}\tilde{\boldsymbol{z}}\|_2^2 \\ \text{s.t.} \quad & \tilde{\boldsymbol{z}} \in \mathbb{Z}^N. \end{aligned} \tag{2.24}$$

where $\tilde{\boldsymbol{R}} \in \mathbb{R}^{N\times N}$ is the upper triangular square root of $\tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{V}}$ resulting from the Cholesky decomposition and $\tilde{\boldsymbol{f}} = \tilde{\boldsymbol{R}}^{-T}\tilde{\boldsymbol{V}}^T\boldsymbol{f}$. Due to the Cholesky factorization, $\tilde{\boldsymbol{R}}$ is upper triangular. Then, a DF operation follows to obtain a decision $\hat{\tilde{\boldsymbol{z}}}_{\text{LRA}-\text{DF}}$ by

$$[\hat{\tilde{\boldsymbol{z}}}_{\text{LRA}-\text{DF}}]_i = \left\lfloor \frac{\tilde{f}_i - \sum_{j=i+1}^{N} \tilde{r}_{ij}[\hat{\tilde{\boldsymbol{z}}}_{\text{LRA}-\text{DF}}]_j}{\tilde{r}_{ii}} \right\rceil$$

for $i = N$ to $i = 1$. The resulting detectors with zero regularization matrix $\boldsymbol{T} = \boldsymbol{0}$ and MMSE regularization matrix $\boldsymbol{T} = \frac{\sigma_v^2}{\sigma_s^2}\boldsymbol{I}$ are known as *ZF LRA-DF* and *MMSE LRA-DF* respectively in the literature.

Though LRA detectors are only approximation to the lattice decoders, it is shown in [27] that LRA detectors with LLL reduction and positive definite regularization $\boldsymbol{T}$ can preserve the optimal DMT in the exact lattice decoders.

2.3 Summary

In this chapter, we reviewed several classic MIMO detectors, namely the ML detector, the linear detectors, the DF detectors, the lattice decoder, and the LRA detectors.

Chapter 3

Lagrangian Dual Maximum-Likelihood Relaxation

3.1 Introduction

We have seen in the previous chapter that regularization plays a non-negligible role in lattice decoding and LRA methods. However, no existing works attempt to shed light on how the regularization should be optimally designed. Other than the well-known MMSE regularization, no other choice of regularization is offered in the literature.

In this chapter, we propose a systematic approach for determining the regularization variable by considering a Lagrangian dual relaxation (LDR) of the ML detection problem. As it turns out, the proposed LDR formulation is to find the best diagonally regularized lattice decoding solution to approximate the ML detector, and all diagonal regularizations, including the MMSE regularization, can be subsumed under the LDR formalism. Our analysis shows that for the 2-PAM case, strong duality holds between the LDR and ML problems. Also, for general PAM, we prove that the LDR problem yields a duality gap no worse than that of a representative relaxation method in MIMO detection, namely, semidefinite relaxation. To physically realize the proposed LDR, the projected subgradient method is employed to handle the LDR problem so that the best regularization can be found. The resultant method can physically be viewed as an adaptive symbol bound control wherein regularized lattice decoding is re-

cursively performed to correct the decision. Simulation results show that the proposed LDR approach can outperform the conventional MMSE-based lattice decoding and LRA approach.

The rest of this chapter is organized as follows. In Section 3.2, we propose our LDR formulation for the ML detection problem, analyze its relationship with the ML problem, and introduce the idea of the projected subgradient method for solving the LDR problem. This is followed by Section 3.3, where we explain in detail the practical implementations of the LDR methods. Simulation results are presented in Section 3.4 to demonstrate the performance of the proposed methods. Section 3.5 concludes this chapter.

3.2 Lagrangian Dual ML Relaxation

In this section, we consider a Lagrangian dual relaxation (LDR) formulation of the ML detection problem (2.5). An important motivation behind our endeavor is that LDR can provide us with the tightest approximation to the ML problem in a Lagrangian sense. Hence, by studying the LDR problem, we may be able to derive an approximate ML algorithm that yields good solution accuracy. Also, we will see that the LDR formulation shows relationship to regularized lattice decoding.

3.2.1 The LDR Formulation

Let us rewrite the ML problem (2.5) as

$$\begin{aligned} \min_{\boldsymbol{s}\in 2\mathbb{Z}^N+\mathbf{1}} \quad & \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2 \\ \text{s.t.} \quad & s_i^2 \le u^2,\ i=1,\ldots,N. \end{aligned} \tag{3.1}$$

Here, it is important to note that we define the problem domain of (3.1) as the discrete set $2\mathbb{Z}^N + \mathbf{1}$. Such an attempt is significantly different from that in many existing relaxed ML MIMO detection methods, such as semidefinite relaxation [19,21,22,54,58], where the problem domain is often the N-dimensional real space $\mathbb{R}^N$. Our goal is to derive the Lagrangian dual problem of the above

ML formulation. Let

$$\mathcal{L}(\boldsymbol{s},\boldsymbol{\lambda}) = \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{\lambda}^T\left(\boldsymbol{s}^2 - u^2\mathbf{1}\right) \tag{3.2}$$

denote the Lagrangian function of (3.1), where $\boldsymbol{\lambda} \succeq \mathbf{0}$ is the Lagrangian dual variable for the constraints $s_i^2 \leq u^2$ (note that the notation $\boldsymbol{\lambda} \succeq \mathbf{0}$ means that $\boldsymbol{\lambda}$ is elementwise non-negative), and $\boldsymbol{s}^2$ denotes the elementwise square of $\boldsymbol{s}$. The Lagrangian dual problem of the ML problem (3.1) is, by definition, given by

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & d(\boldsymbol{\lambda}) \\ \text{s.t.} \quad & \boldsymbol{\lambda} \succeq \mathbf{0}, \end{aligned} \tag{3.3}$$

where $d(\boldsymbol{\lambda}) = \min_{\boldsymbol{s}\in 2\mathbb{Z}^N+\mathbf{1}} \mathcal{L}(\boldsymbol{s},\boldsymbol{\lambda})$ is the dual function associated with (3.2), which can be expressed as

$$d(\boldsymbol{\lambda}) = \varphi(\boldsymbol{\lambda}) - u^2\boldsymbol{\lambda}^T\mathbf{1}, \tag{3.4}$$

$$(\Phi_{\boldsymbol{\lambda}}) \quad \varphi(\boldsymbol{\lambda}) = \min_{\boldsymbol{s}\in 2\mathbb{Z}^N+\mathbf{1}} \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T\boldsymbol{D}(\boldsymbol{\lambda})\boldsymbol{s}, \tag{3.5}$$

with $\boldsymbol{D}(\boldsymbol{\lambda})$ denoting a diagonal matrix whose ith diagonal is λ_i. For convenience, we will call problem (3.3) the *LDR problem*.

At this point, we see several interesting observations. First, problem $(\Phi_{\boldsymbol{\lambda}})$, as a constituent component of the dual function, is a diagonally regularized lattice decoding problem. Hence, the above formulated LDR problem exhibits relation to regularized lattice decoding. More connections between LDR and regularized lattice decoding will be revealed later. Second, the naive and MMSE lattice decoders can be seen as particular instances of problem $(\Phi_{\boldsymbol{\lambda}})$. Specifically, NLD chooses $\boldsymbol{\lambda} = \mathbf{0}$, while MMSE lattice decoding $\boldsymbol{\lambda} = \sigma_\nu^2/\sigma_s^2\mathbf{1}$. Third, the LDR problem (3.3) can be regarded as that of finding the best regularization vector $\boldsymbol{\lambda}$ among all the diagonally regularized lattice decoding instances to approximate the ML problem.

3.2.2 Optimality and Duality Gap Analysis

In this subsection, we analyze the optimality conditions of the LDR formulation in (3.3)-(3.5), that is, conditions under which the LDR problem exactly

solves the ML problem. Moreover, we also study the approximation quality of LDR by analyzing the duality gap $f^\star - d^\star$, where

$$f^\star = \min_{\boldsymbol{s}\in 2\mathbb{Z}^N+\mathbf{1},\ \boldsymbol{s}^2\preceq u^2\mathbf{1}} \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2 \tag{3.6}$$

$$d^\star = \max_{\boldsymbol{\lambda}\succeq\mathbf{0}} d(\boldsymbol{\lambda}) \tag{3.7}$$

denote the optimal objective values of the ML problem (3.1) and the LDR problem (3.3), respectively. In particular, a smaller $f^\star - d^\star$ would indicate better approximation accuracy, and zero $f^\star - d^\star$ means ML being achieved. Note that $f^\star - d^\star \geq 0$ by weak duality (see, e.g., [67]).

We first present a simple result based on a connection between NLD and LDR.

Fact 1 *Let*

$$\hat{\boldsymbol{s}}_{\mathrm{NLD}} \in \arg\min_{\boldsymbol{s}\in 2\mathbb{Z}^N+\mathbf{1}} \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2$$

be an optimal solution of the NLD problem (2.17)*. Consider instances where* $[\hat{\boldsymbol{s}}_{\mathrm{NLD}}]_i^2 \leq u^2$ *for all* $i = 1,\ldots,N$*. Then, the following statements hold:*

1. $\hat{\boldsymbol{s}}_{\mathrm{NLD}}$ *is an optimal solution of the ML problem* (3.1)*.*
2. *Strong duality, or* $f^\star - d^\star = 0$*, holds for the LDR problem. Also,* $\boldsymbol{\lambda} = \mathbf{0}$ *is an optimal solution of the LDR problem* (3.3)*.*
3. *For any optimal solution* $\boldsymbol{\lambda}^\star$ *of the LDR problem* (3.3)*,* $\hat{\boldsymbol{s}}_{\mathrm{NLD}}$ *is an optimal solution of problem* $(\Phi_{\boldsymbol{\lambda}})$ *for* $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\star$*.*

The proof of Fact 1 is given in Appendix 3.6.1. The idea behind is to exploit the fact that NLD is a special case of LDR. Fact 1 implies that for instances where NLD is ML-optimal, LDR is also ML-optimal. Hence, we should expect that LDR would perform better than NLD—this will be shown to be true by simulations later.

Next, we consider another optimality result. The following lemma will be needed.

Lemma 3.1 *Let $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ be an optimal solution of problem $(\Phi_{\boldsymbol{\lambda}})$. Suppose that $\boldsymbol{\lambda}$ satisfies*

$$\lambda_i > \gamma_m \tag{3.8}$$

for some $i \in \{1, \ldots, N\}$, where m is an odd positive integer,

$$\gamma_m = \frac{c(\boldsymbol{y}, \boldsymbol{H})}{(m+2)^2 - 1}, \tag{3.9}$$

$$c(\boldsymbol{y}, \boldsymbol{H}) = \min_{\boldsymbol{s} \in \{\pm 1\}^N} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 - \min_{\boldsymbol{s} \in 2\mathbb{Z}^N + 1} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2. \tag{3.10}$$

Then it must hold true that

$$[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 \leq m^2.$$

The proof of Lemma 3.1 is shown in Appendix 3.6.2. Lemma 3.1 is not only useful in establishing an optimality condition of LDR, as we will see, but it is also of independent interest as discussed in the following remark.

Remark 1: Intuitively, the idea of incorporating regularization in lattice decoding is based on the belief that regularization can pull the regularized lattice decoding solution within the symbol bounds. Lemma 3.1 provides a theoretical justification that this intuitive belief is indeed true. It quantifies, in a sufficient manner, how much regularization is needed to achieve a desired symbol bound constraint level in the regularized lattice decoding problem $(\Phi_{\boldsymbol{\lambda}})$.

Let us now turn our attention back to the optimality analysis. From Lemma 3.1, we have the following observation: For any $\boldsymbol{\lambda} \succ \gamma_u \mathbf{1}$, an optimal solution $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ of problem $(\Phi_{\boldsymbol{\lambda}})$ always satisfies $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}^2 \preceq u^2 \mathbf{1}$, which means that $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ is a feasible solution of the ML problem. Hence, we would hope that such a $\boldsymbol{\lambda}$ and the corresponding $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ are optimal to the LDR problem and the ML problem, respectively. Remarkably, we show that this is indeed true for the case of $u = 1$.

Theorem 3.1 *Consider $u = 1$, or the 2-PAM case. In this case, strong duality $f^\star - d^\star = 0$ always holds. In particular, any $\boldsymbol{\lambda} \succ \gamma_1 \mathbf{1}$ is an optimal solution of the LDR problem* (3.3)*, and an optimal solution $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ of problem $(\Phi_{\boldsymbol{\lambda}})$ for any $\boldsymbol{\lambda} \succ \gamma_1 \mathbf{1}$ is also an optimal solution of the ML problem* (3.1)*.*

The proof is relegated to Appendix 3.6.3. Theorem 3.1 indicates that in the 2-PAM constellation case, the LDR problem is ML-optimal for *any* given realization $(\boldsymbol{y}, \boldsymbol{H})$. This further implies that in principle, one can use regularized lattice decoding, which has no explicit symbol bound constraints, to solve the ML problem for the 2-PAM case.

An interesting question is whether the proof we use in Theorem 3.1 can be extended to the more general case of $u \geq 3$. We found that such extension is possible. Unfortunately, the result obtained becomes a loose bound on the duality gap, rather than strong duality. It also fails to cover the optimality condition in Fact 1. Herein, we give another analysis result that links to the BC-SDR detector

$$\begin{aligned} g^\star = \min_{\boldsymbol{S}\in\mathbb{S}^N, \boldsymbol{s}\in\mathbb{R}^N} \ & \mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{S}) - 2\boldsymbol{s}^T\boldsymbol{H}^T\boldsymbol{y} + \|\boldsymbol{y}\|_2^2 \\ \text{s.t. } & \boldsymbol{S} \succeq \boldsymbol{s}\boldsymbol{s}^T, \\ & 1 \leq S_{ii} \leq u^2,\ i = 1, \ldots, N. \end{aligned} \tag{3.11}$$

One can observe that LDR and BC-SDR are quite different from one another; the former and latter use discrete and continuous problem domains, respectively. In the following theorem, we provide a connection between LDR and BC-SDR.

Theorem 3.2 *For any realization* $(\boldsymbol{y}, \boldsymbol{H})$ *and any* $u \geq 1$*, the duality gap of LDR is better than or equal to that of BC-SDR:*

$$f^\star - d^\star \leq f^\star - g^\star. \tag{3.12}$$

The proof of Theorem 3.2 is relegated to Appendix 3.6.4. Theorem 3.2 is meaningful in establishing a relationship of the approximation qualities of LDR and BC-SDR. Specifically, it implies that LDR performs no worse than BC-SDR in terms of relaxation tightness. In addition, it is possible for LDR to yield strictly better duality gap than BC-SDR. For example, for the case of $u = 1$, we have shown in Theorem 3.1 that strong duality always holds for LDR. However, BC-SDR does not guarantee strong duality even for $u = 1$, as indicated in previous work [59, 68]. Also, owing to the equivalence of BC-SDR, PI-SDR and VA-SDR, the same conclusion applies to PI-SDR and VA-SDR.

3.2.3 Practical Realization of LDR via Projected Subgradient

We turn our attention to the realization of LDR. By Lagrangian duality theory [67], the LDR problem (3.3) is a convex optimization problem. In particular, its objective function $d(\boldsymbol{\lambda})$ is concave. However, this does not mean that $d(\boldsymbol{\lambda})$ is easy to maximize. From (3.4), we see that $d(\boldsymbol{\lambda})$ involves a minimization problem, namely, problem $(\Phi_{\boldsymbol{\lambda}})$ in (3.5). Thus, $d(\boldsymbol{\lambda})$ is in general a nondifferentiable function. Our optimization strategy is to employ the projected subgradient (PS) method [69], which is a convenient approach for solving nondifferentiable convex optimization problems.

The PS method for the LDR problem (3.3) is described as follows. Let $\boldsymbol{\lambda}^{(k)}$ denote the iterate generated by the PS method at the kth iteration. Given an initialization $\boldsymbol{\lambda}^{(1)}$, the iterates are recursively generated via

$$\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{P}_{\mathbb{R}_+^N}(\boldsymbol{\lambda}^{(k)} + \alpha_k \boldsymbol{g}^{(k)}), \quad k = 1, 2, \ldots \tag{3.13}$$

where $\boldsymbol{g}^{(k)}$ denotes a subgradient of $d(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^{(k)}$, $\{\alpha_k\}$ is a step-size sequence which is predetermined, and $\boldsymbol{P}_{\mathbb{R}_+^N}(\boldsymbol{\lambda})$ denotes the projection of its input $\boldsymbol{\lambda} \in \mathbb{R}^N$ onto the set of N-dimensional nonnegative vectors $\mathbb{R}_+^N$. The projection operator $\boldsymbol{P}_{\mathbb{R}_+^N}(\boldsymbol{\lambda})$ has a closed form; specifically, if we let $\boldsymbol{\mu} = \boldsymbol{P}_{\mathbb{R}_+^N}(\boldsymbol{\lambda})$, then $\mu_i = \max\{0, \lambda_i\}$ for all i. Using basic subgradient calculus results [69], the subgradient $\boldsymbol{g}^{(k)}$ is shown to be

$$\boldsymbol{g}^{(k)} = (\boldsymbol{s}^{(k)})^2 - u^2 \mathbf{1}, \tag{3.14}$$

where $\boldsymbol{s}^{(k)}$ is a solution of the regularized lattice decoding problem $(\Phi_{\boldsymbol{\lambda}})$ for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(k)}$; i.e.,

$$\boldsymbol{s}^{(k)} = \arg \min_{\boldsymbol{s} \in 2\mathbb{Z}^N + \mathbf{1}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T \boldsymbol{D}(\boldsymbol{\lambda}^{(k)}) \boldsymbol{s}. \tag{3.15}$$

We should discuss the convergence of the PS method. It is known in the optimization literature [69, 70] that under a few fairly mild assumptions, the PS method is guaranteed to converge to the optimal objective value, which is $d^\star$ here. For the LDR problem here, we can even pin down a simplified sufficient assumption for convergence—the PS method can achieve convergence to the

optimal dual value $d^\star$ for *any full column rank channel matrix* $\boldsymbol{H}$. A complete description for the PS convergence results mentioned above will be provided in Remark 3. Also, practical convergence issues will be discussed in Remark 4.

To summarize, we can solve the LDR problem by using the iterative PS procedure in (3.13)-(3.15). In particular, at each iteration, we need to solve the regularized lattice decoding problem in (3.15), and then use its solution $\boldsymbol{s}^{(k)}$ to update the regularization vector at the next iteration, $\boldsymbol{\lambda}^{(k+1)}$. At first, this may sound computationally more expensive than a one-shot regularized lattice decoding method such as MMSE LD. However, we will illustrate by simulations that with a careful initialization and implementation, a PS-based LDR detector can be computationally comparable to the MMSE LD detector. Also, as a practical alternative, we can consider efficient approximation schemes where we use inexact sphere decoding or LRA-DF detection in place of an exact solver for problem (3.15). The PS procedure provides interesting insight as described in the following remark.

Remark 2: Physically, the PS procedure above can be interpreted as a recursively regularized lattice decoding method wherein some form of adaptive symbol bound control is performed. To explain this, let $\lambda_i^{(k)}$ and $s_i^{(k)}$ denote the ith elements of $\boldsymbol{\lambda}^{(k)}$ and $\boldsymbol{s}^{(k)}$, respectively. Then, we can see from (3.13) and (3.14) that if $(s_i^{(k)})^2 > u^2$, then $\lambda_i^{(k+1)}$ will be increased. Likewise, if $(s_i^{(k)})^2 < u^2$, then $\lambda_i^{(k+1)}$ will be decreased. This means that if some symbols violate the symbol bound constraints, then the PS method at the next iteration will increase the regularization variables for those symbols. Subsequently, those larger regularization variables will tend to pull the associated symbols toward the origin at the next iteration, or set a more stringent upper bound on those symbols, as suggested in Lemma 3.1. Similarly, for symbols lying strictly within the symbol bounds, the corresponding regularization variables will be reduced at the next iteration.

We should also give some discussions on the convergence of the PS LDR procedure.

Remark 3: Theoretically, the optimal convergence of the PS LDR procedure is

usually guaranteed. To discuss this in more precise terms, we need to describe one available convergence result [69, 70]: Let $d_{\text{best}}^{(k)} = \max\{d_{\text{best}}^{(k-1)}, d(\boldsymbol{\lambda}^{(k)})\}$. For certain types of stepsize rules, such as the diminishing stepsize rule [1], we have $\lim_{k\to\infty} d_{\text{best}}^{(k)} = d^\star$ if every subgradient $\boldsymbol{g}^{(k)}$ is bounded; i.e., there exists a finite G such that

$$\|\boldsymbol{g}^{(k)}\|_2 \leq G, \quad \text{for } k = 1, 2, \ldots$$

In our LDR problem, the boundedness assumption above is the same as requiring every regularized lattice decoding solution $\boldsymbol{s}^{(k)}$ in (3.15) to be bounded. Intuitively, one would argue that problem (3.15), or $(\Phi_{\boldsymbol{\lambda}})$, should yield bounded solutions, except for pathological cases. In fact, this can be confirmed under a mild assumption:

Lemma 3.2 *Given a full column rank channel matrix $\boldsymbol{H}$, any optimal solution $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ of problem $(\Phi_{\boldsymbol{\lambda}})$ for any $\boldsymbol{\lambda} \succeq \mathbf{0}$ is bounded.*

The proof of Lemma 3.2 is given in Appendix 3.6.5. As a corollary of Lemma 3.2, the PS LDR procedure is theoretically guaranteed to converge to the optimal value for full column rank (and thus overdetermined) channels.

Remark 4: While the PS method provides an effective strategy to cope with certain difficult nondifferentiable convex optimization problems, such as the LDR problem here, it is also known to yield slow convergence in some applications. Despite this setback, the PS convergence speed may be improved if a good initialization can be found. Fortunately, such initialization seems to be available for LDR, as our extensive simulations have found. In Section 3.3.3, we will propose an initialization scheme that requires solving another (convex) optimization problem, but can significantly improve the convergence speed.

While the PS LDR realization procedure described above looks quite straightforward, there are fine details on how we can implement the method in a numerically efficient manner. This will be the focus of the next section.

[1] A typical example of the diminishing stepsize rule is $\alpha_k = \eta/\sqrt{k}$ for some fixed positive constant η.

3.3 Practical Implementation

In this section, we elaborate on how we implement the PS LDR method.

3.3.1 Lattice Decoding for Problem (3.15)

The core issue with implementing the PS LDR method lies in finding the solution of the regularized lattice decoding problem (3.15) at each iteration. One can directly handle this issue by seeing each problem (3.15) as a stand-alone regularized lattice decoding problem and solve it in the same as described in Chapter 2. However, we can make the implementation more efficient by utilizing the regularized lattice decoding solution in the previous iteration to improve the solution search process in the present iteration. To facilitate our description, we divide our development into three steps.

Step 1. Integer Least Squares Reformulation

Let

$$\boldsymbol{V}_k^T \boldsymbol{V}_k = \boldsymbol{H}^T \boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(k)}) \tag{3.16}$$

denote a square-root decomposition of $\boldsymbol{H}^T \boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(k)})$, where $\boldsymbol{V}_k \in \mathbb{R}^{N \times N}$ is a corresponding square-root factor. By the transformation

$$\boldsymbol{s} = 2\boldsymbol{z} + \mathbf{1}, \qquad \boldsymbol{z} \in \mathbb{Z}^N, \tag{3.17}$$

we can rewrite problem (3.15) as an integer least squares (LS) problem

$$\boldsymbol{z}^{(k)} = \arg \min_{\boldsymbol{z} \in \mathbb{Z}^N} \|\boldsymbol{f}_k - \boldsymbol{V}_k \boldsymbol{z}\|_2^2, \tag{3.18}$$

where $\boldsymbol{f}_k = \frac{1}{2} \boldsymbol{V}_k{}^{-T} \left(\boldsymbol{H}^T \boldsymbol{y} - \left(\boldsymbol{H}^T \boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(k)}) \right) \mathbf{1} \right)$. Note the relation $\boldsymbol{s}^{(k)} = 2\boldsymbol{z}^{(k)} + \mathbf{1}$.

Step 2. Lattice Reduction

We reformulate problem (3.18) to a lattice-reduced form. The popularized LLL algorithm is chosen for lattice reduction, and some of its operational details are concisely described as follows. Let $\boldsymbol{G} \in \mathbb{R}^{N \times N}$ be a matrix to be LLL-reduced. Denote its QR decomposition by $\boldsymbol{G} = \bar{\boldsymbol{Q}} \bar{\boldsymbol{R}}$ where $\bar{\boldsymbol{Q}} \in \mathbb{R}^{N \times N}$ is

unitary and $\bar{\boldsymbol{R}} \in \mathbb{R}^{N\times N}$ upper triangular. Given $(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{R}})$, an LLL algorithm finds a 3-tuple $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{U})$ such that

$$\boldsymbol{Q}\boldsymbol{R} = \bar{\boldsymbol{Q}}\bar{\boldsymbol{R}}\boldsymbol{U}, \tag{3.19}$$

$\boldsymbol{Q}\boldsymbol{R}$ is a lattice-reduced matrix of $\bar{\boldsymbol{Q}}\bar{\boldsymbol{R}} = \boldsymbol{G}$, and $\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{U} \in \mathbb{R}^{N\times N}$ are unitary, upper triangular and unimodular, respectively. For convenience, we will use the following notation

$$(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{U}) = \mathsf{LLL}(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{R}}) \tag{3.20}$$

to represent the LLL process. Returning to our problem, our task is to LLL-reduce the basis matrix $\boldsymbol{V}_k$ in problem (3.18). This can be done as follows:

$$(\bar{\boldsymbol{Q}}_k, \bar{\boldsymbol{R}}_k) = \mathsf{QR}(\boldsymbol{V}_k), \tag{3.21a}$$

$$(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{U}_k) = \mathsf{LLL}(\bar{\boldsymbol{Q}}_k, \bar{\boldsymbol{R}}_k), \tag{3.21b}$$

where QR is a shorthand notation for the QR decomposition process. We can see that $\boldsymbol{Q}_k\boldsymbol{R}_k$ forms an LLL-reduced basis matrix of $\boldsymbol{V}_k$, with $\boldsymbol{U}_k$ being the associated unimodular transformation matrix. By substituting $\boldsymbol{Q}_k\boldsymbol{R}_k = \boldsymbol{V}_k\boldsymbol{U}_k$, and introducing the transformation $\tilde{\boldsymbol{z}} = \boldsymbol{U}_k^{-1}\boldsymbol{z}$, we equivalently turn problem (3.18) to

$$\tilde{\boldsymbol{z}}^{(k)} = \arg\min_{\tilde{\boldsymbol{z}}\in\mathbb{Z}^N} \|\tilde{\boldsymbol{f}}_k - \boldsymbol{R}_k\tilde{\boldsymbol{z}}\|_2^2, \tag{3.22}$$

where $\tilde{\boldsymbol{f}}_k = {\boldsymbol{Q}_k}^T\boldsymbol{f}_k$.

An Alternative Option to Step 2 by Successive LLL Update

We offer an alternative option to Step 2 that can provide computational savings with the LLL reduction process. Essentially, if a given basis matrix $\boldsymbol{G}$ is almost LLL-reduced, then it would generally take fewer number of iterations for the LLL algorithm to complete the process. Now, rather than LLL-reducing $\boldsymbol{V}_k$ as in the previous Step 2, *we consider the LLL reduction of* $\boldsymbol{V}_k\boldsymbol{U}_{k-1}$*, where* $\boldsymbol{U}_{k-1}$ *is the LLL unimodular matrix of* $\boldsymbol{V}_{k-1}$. The idea is that if $\boldsymbol{V}_k$ and $\boldsymbol{V}_{k-1}$ are not too different, which is possible when there are only small changes with the PS update (cf. (3.13)), then $\boldsymbol{V}_k\boldsymbol{U}_{k-1}$ may already be well conditioned in

the LLL sense. Note that a similar idea has been considered in a different context, namely, LRA detection under time-correlated fading channels [71]. The proposed successive LLL update process is described as follows. We replace (3.21) by

$$(\bar{\boldsymbol{Q}}_k, \bar{\boldsymbol{R}}_k) = \mathsf{QR}(\boldsymbol{V}_k \boldsymbol{U}_{k-1}), \tag{3.23a}$$

$$(\boldsymbol{Q}_k, \boldsymbol{R}_k, \tilde{\boldsymbol{U}}_k) = \mathsf{LLL}(\bar{\boldsymbol{Q}}_k, \bar{\boldsymbol{R}}_k), \tag{3.23b}$$

$$\boldsymbol{U}_k = \boldsymbol{U}_{k-1} \tilde{\boldsymbol{U}}_k. \tag{3.23c}$$

It can be verified that $\boldsymbol{Q}_k \boldsymbol{R}_k$ is a lattice-reduced matrix of $\boldsymbol{V}_k$, and that $\boldsymbol{U}_k$ is the corresponding unimodular transformation matrix. Thus, the equivalent formulation in (3.22) applies. We should note that there is a more efficient way to compute (3.23). Let $\mathsf{chol}(\cdot)$ denote the Cholesky decomposition operation; i.e., $\boldsymbol{W} = \mathsf{chol}(\boldsymbol{A}) \Longleftrightarrow \boldsymbol{W}^T \boldsymbol{W} = \boldsymbol{A}, \boldsymbol{W}$ upper triangular. It can be shown that (3.23a)-(3.23b) can be equivalently implemented by

$$\boldsymbol{W}_k = \mathsf{chol}({\boldsymbol{U}_{k-1}}^T (\boldsymbol{H}^T \boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(k)})) \boldsymbol{U}_{k-1}), \tag{3.24a}$$

$$(\boldsymbol{Q}_k, \boldsymbol{R}_k, \tilde{\boldsymbol{U}}_k) = \mathsf{LLL}(\boldsymbol{I}, \bar{\boldsymbol{W}}_k). \tag{3.24b}$$

In particular, by using (3.24), we do not need to compute $\boldsymbol{V}_k$, which requires a square-root decomposition operation (cf. (3.16)).

Step 3. Applying a Sphere Decoder

As the last step, we solve problem (3.22) by a sphere decoding algorithm. Note that problem (3.22), which has its equivalent channel matrix $\boldsymbol{R}_k$ being upper triangular, already takes a standard problem form for a sphere decoding algorithm to run. Sphere decoding considers a within-sphere point search process; roughly speaking, the latter may be described by

$$\begin{aligned} \min_{\tilde{\boldsymbol{z}} \in \mathbb{Z}^N} \quad & \|\tilde{\boldsymbol{f}}_k - \boldsymbol{R}_k \tilde{\boldsymbol{z}}\|_2^2 \\ \text{s.t.} \quad & \|\tilde{\boldsymbol{f}}_k - \boldsymbol{R}_k \tilde{\boldsymbol{z}}\|_2^2 \leq C^{(k)} \end{aligned} \tag{3.25}$$

for some given squared sphere radius $C^{(k)}$. More complete descriptions of the within-sphere point processes used in various sphere decoding algorithms can be found in [12,15]. In practice, it is generally found that a well chosen sphere radius

may significantly narrow down the within-sphere search space, or the feasible set of (3.25), thereby making the optimal point search more efficient [15, 28]. Here, we make use of the result in the previous LDR iteration to set the sphere radius, namely, by

$$C^{(k)} = \|\boldsymbol{f}_k - \boldsymbol{V}_k \boldsymbol{z}^{(k-1)}\|_2^2. \tag{3.26}$$

The rationale is that if the previous iterate is LDR-optimal, or near LDR-optimal, then the sphere radius choice above may eliminate a substantial number of points that are not necessary to visit.

3.3.2 Putting Together the Algorithm

By plugging the above lattice decoding steps into the PS procedure in (3.13)-(3.15), we construct a complete LDR algorithm. A pseudo-code form description of the algorithm is given in Algorithm 3.1. In the algorithm, $\boldsymbol{\mathcal{Q}}$ denotes the elementwise quantization function to the symbol constellation set $\{\pm 1, \pm 3, \ldots \pm u\}$, and $\lfloor \cdot \rceil$ denotes the elementwise integer rounding function. Also, note that we adopt the successive LLL update method (Steps 5-8). In the sequel, we will call the resultant detector the *LDR lattice decoding (LD)* detector.

As previously mentioned, we can also consider variations of the LDR LD detector where efficient approximation schemes are used in place of the exact sphere decoder. One alternative is to force the sphere decoding algorithm to terminate when its number of nodes visited exceeds a prescribed limit— this leads to an inexact runtime-limited LDR LD detector. Another alternative is to apply the LRA-DF method [30, 72], or equivalently, the Babai's nearest plane algorithm [63], to problem (3.22). To be specific, we generate a point, denoted by $\tilde{\boldsymbol{z}}_{\mathsf{DF}}^{(k)} \in 2\mathbb{Z}^N + \mathbf{1}$, via

$$\tilde{z}_{\mathsf{DF},i}^{(k)} = \left\lfloor \frac{[\tilde{\boldsymbol{f}}_k]_i - \sum_{j=i+1}^{N} [\boldsymbol{R}_k]_{ij} \tilde{z}_{\mathsf{DF},j}^{(k)}}{[\boldsymbol{R}_k]_{ii}} \right\rceil, \quad \text{for } i = N, N-1, \ldots, 1. \tag{3.27}$$

We will name the subsequent detector the *LDR LRA-DF detector*.

Algorithm 3.1: The LDR LD Detector

input : a problem instance $(\boldsymbol{y}, \boldsymbol{H})$, a starting point $\boldsymbol{\lambda}^{(1)} \succeq \mathbf{0}$, a step-size sequence $\{\alpha_k\}$, and stopping parameters K_{max}, ϵ.

$k = 1$;

$\boldsymbol{U}^{(0)} = \boldsymbol{I}$;

$\tilde{\boldsymbol{z}}^{(0)} = \left\lfloor \frac{1}{2}((\boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(1)}))^{-1}\boldsymbol{H}^T\boldsymbol{y} + \mathbf{1}) \right\rceil$;

repeat

  $\boldsymbol{W}_k = \mathsf{chol}({\boldsymbol{U}_{k-1}}^T(\boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(k)}))\boldsymbol{U}_{k-1})$;

  $\boldsymbol{f}_k = \frac{1}{2}{\boldsymbol{W}_k}^{-T}{\boldsymbol{U}_{k-1}}^T(\boldsymbol{H}^T\boldsymbol{y} - (\boldsymbol{H}^T\boldsymbol{H} + \boldsymbol{D}(\boldsymbol{\lambda}^{(k)}))\mathbf{1})$;

  $(\boldsymbol{Q}_k, \boldsymbol{R}_k, \tilde{\boldsymbol{U}}_k) = \mathsf{LLL}(\boldsymbol{I}, \boldsymbol{W}_k)$;

  $\boldsymbol{U}_k = \boldsymbol{U}_{k-1}\tilde{\boldsymbol{U}}_k$;

  $\tilde{\boldsymbol{f}}_k = {\boldsymbol{Q}_k}^T\boldsymbol{f}_k$;

  $C^{(k)} = \|\boldsymbol{f}_k - \boldsymbol{W}_k\tilde{\boldsymbol{z}}^{(k-1)}\|_2^2$;

  run a sphere decoding algorithm, with $C^{(k)}$ as the squared sphere radius, to solve for

$$\tilde{\boldsymbol{z}}^{(k)} = \arg\min_{\tilde{\boldsymbol{z}} \in \mathbb{Z}^N} \|\tilde{\boldsymbol{f}}_k - \boldsymbol{R}_k\tilde{\boldsymbol{z}}\|_2^2$$

  $\boldsymbol{z}^{(k)} = \boldsymbol{U}_k\tilde{\boldsymbol{z}}^{(k)}$;

  $\boldsymbol{s}^{(k)} = 2\boldsymbol{z}^{(k)} + \mathbf{1}$;

  $\boldsymbol{g}^{(k)} = \left(\boldsymbol{s}^{(k)}\right)^2 - u^2\mathbf{1}$;

  $\boldsymbol{\lambda}^{(k+1)} = \boldsymbol{P}_{\mathbb{R}_+^N}\left(\boldsymbol{\lambda}^{(k)} + \alpha_k\boldsymbol{g}^{(k)}\right)$;

  $k = k + 1$;

until $k > K_{\mathsf{max}}$ or $\|\boldsymbol{\lambda}^{(k)} - \boldsymbol{\lambda}^{(k-1)}\|_2 \leq \epsilon$;

Find $\ell = \arg\min_{i=1,2,\ldots,k-1} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\mathcal{Q}}(\boldsymbol{s}^{(i)})\|_2$

output: $\hat{\boldsymbol{s}} = \boldsymbol{\mathcal{Q}}(\boldsymbol{s}^{(\ell)})$.

3.3.3 Box Relaxation as an Initialization

Like many other iterative optimization methods, the PS LDR method may exhibit fast convergence if a good starting point $\boldsymbol{\lambda}^{(1)}$ is given. For the initialization aspect, a simple and logical choice is to employ the MMSE regularization; i.e., $\boldsymbol{\lambda}^{(1)} = \sigma_\nu^2/\sigma_s^2\mathbf{1}$. By our empirical experience, the MMSE initialization scheme can indeed lead to reasonable improvement in PS convergence speed. But we also found a better initialization scheme based on our extensive numerical study. The idea is to consider a further relaxation of the LDR problem (3.3)

$$d^\star \geq \max_{\boldsymbol{\lambda}\succeq\mathbf{0}} \left\{ \min_{\boldsymbol{s}\in\mathbb{R}^N} \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{\lambda}^T(\boldsymbol{s}^2 - u^2\mathbf{1}) \right\}, \tag{3.28}$$

where we relax the original problem domain of $\boldsymbol{s}$ from $2\mathbb{Z}^N + \mathbf{1}$ to $\mathbb{R}^N$. To be specific, we aim at using the solution of the outer part of problem (3.28) to initialize the PS LDR method. It is interesting to note that problem (3.28) is related to a conventional MIMO detection method. Consider the following problem

$$\begin{aligned} &\min_{\boldsymbol{s}\in\mathbb{R}^N} \ \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2 \\ &\text{s.t. } s_i^2 \leq u^2, \qquad i=1,\ldots,N, \end{aligned} \tag{3.29}$$

which is a continuous *box relaxation* (BR) of the ML problem (3.1) and has previously been studied in the context of multiuser detection [20, 73]. It can be shown that problem (3.28) is equivalent to problem (3.29)—the equivalence is in the sense that the dual of the BR problem (3.29) takes exactly the same form as problem (3.28), and that strong duality holds owing to the fact that the BR problem (3.29) is convex and satisfies Slater's condition [67]. Hence, the initialization scheme based on (3.28) may be seen as using the BR method to warm-start the PS LDR method.

The BR initialization scheme proposed above does not have a closed form in general. In order to implement the BR initialization scheme efficiently, we need to build a low-complexity solver for problem (3.28). Our solution approach consists of two steps: First, we solve the BR problem (3.29), which has a simple structure and for which it is relatively easy to find an efficient optimization algorithm. Second, by utilizing the strong duality relationship of problems (3.28)

and (3.29), we construct a solution to problem (3.28) from the BR solution. For the first step, we custom-build an optimization algorithm for the BR problem (3.29); the algorithm is mainly based on the active set method in [74], with a modification to accelerate the LS procedure inside by the one-column updating method [75]. The details are heavily numerical, and are briefly explained in Appendix 3.6.6 and 3.6.7. For the second step, we consider the Karush-Kuhn-Tucker (KKT) conditions of the BR problem, which is shown to be

$$\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{s} + \boldsymbol{D}(\boldsymbol{\lambda})\boldsymbol{s} - \boldsymbol{H}^T\boldsymbol{y} = \boldsymbol{0}, \tag{3.30a}$$

$$\lambda_i(s_i^2 - u^2) = 0, \quad i = 1, \ldots, N, \tag{3.30b}$$

$$\lambda_i \geq 0, s_i^2 - u^2 \leq 0, \quad i = 1, \ldots, N. \tag{3.30c}$$

Since (3.30a)-(3.30c) is the necessary and sufficient conditions for $(\boldsymbol{s}, \boldsymbol{\lambda})$ to be optimal to problems (3.28) and (3.29), we plug the BR solution $\boldsymbol{s}$ obtained in the first step into (3.30a)-(3.30b) to find the corresponding optimal $\boldsymbol{\lambda}$. The resultant solution is shown to be

$$\lambda_i = \begin{cases} -\dfrac{[\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{s} - \boldsymbol{H}^T\boldsymbol{y}]_i}{s_i}, & \text{if } s_i \neq 0 \\ 0, & \text{if } s_i = 0 \end{cases} \tag{3.31}$$

for $i = 1, \ldots, N$. To summarize, the BR initialization scheme works by running a custom-built active set algorithm to find a solution $\boldsymbol{s}$ of the BR problem (3.29), substituting the obtained $\boldsymbol{s}$ into (3.31) to construct a solution $\boldsymbol{\lambda}$ of problem (3.28), and using the obtained $\boldsymbol{\lambda}$ as a starting point of the PS LDR method.

3.4 Simulations

Simulations were conducted to study the symbol error rate (SER) and complexity performance of the proposed LDR MIMO detection approach. We consider a standard complex-valued QAM MIMO scenario

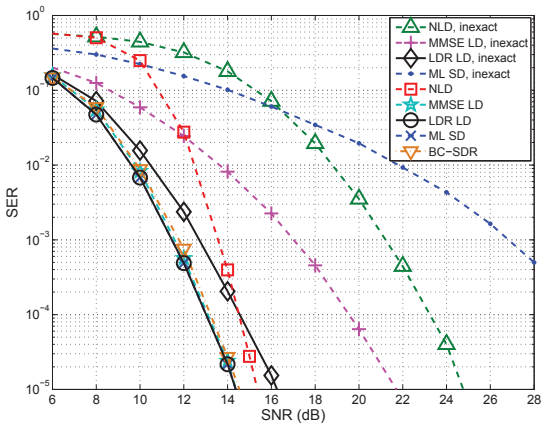
$$\boldsymbol{y}_C = \boldsymbol{H}_C\boldsymbol{s}_C + \boldsymbol{\nu}_C, \tag{3.32}$$

where the channel matrix $\boldsymbol{H}_C \in \mathbb{C}^{M_C \times N_C}$ follows an elementwise i.i.d. complex circular Gaussian distribution with zero mean and unit variance, the symbol

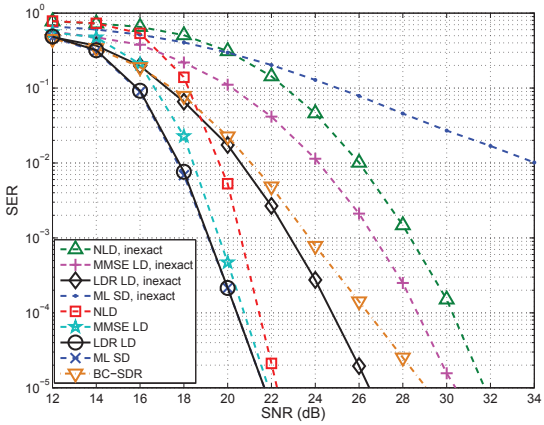
vector $\boldsymbol{s}_C \in \mathbb{C}^{N_C}$ is elementwise i.i.d. uniformly distributed with each element drawn from the standard $(u+1)^2$-QAM constellation set, $\sigma^2_{\nu_C} \in \mathbb{C}^{M_C}$ is additive white complex circular Gaussian noise with zero mean and variance $\sigma^2_{\nu_C}$, and M_C and N_C denotes the output and input problem sizes, respectively. The model (3.32) can be equivalently represented by the real-valued model via the transformation described in Chapter 2. The SNR is defined as $\text{SNR} = \mathbb{E}[\|\boldsymbol{H}_C\boldsymbol{s}_C\|^2]/\mathbb{E}[\|\boldsymbol{\nu}_C\|^2] = N_C\sigma^2_{s_C}/\sigma^2_{\nu_C}$, where $\sigma^2_{s_C}$ is the variance of the elements of $\boldsymbol{s}_C$.

The benchmarked algorithms are the ML sphere decoding (SD) detector, the MMSE LD detector, the NLD detector, runtime-limited inexact implementations of the aforementioned SD and LD detectors, and the BC-SDR detector (cf. problem (3.11) and [54]). The ML SD detector is implemented by the Schnorr-Euchner enumeration-based SD algorithm in [12, Algorithm 2]. The MMSE LD detector is also implemented by the same SD algorithm, and its lattice reduction process is LLL reduction. The LLL reduction algorithm we employed follows the pseudo-code description in [72]. The inexact MMSE LD detector is a variation of the MMSE LD detector where we force the SD algorithm to terminate when the number of nodes visited exceeds a prescribed worst-case limit, denoted by N_{node} here. We set $N_{\text{node}} = \min\{N^3, N \times \text{SNR}\}$ (SNR is in linear scale). The same applies to inexact ML SD and inexact NLD. For the BC-SDR detector, Gaussian randomization rounding [22] is employed and the number of randomizations is 64.

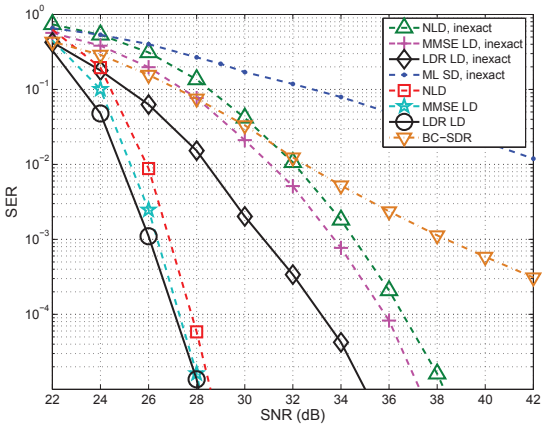
The settings of the proposed LDR LD detector, inexact LDR LD detector and LDR LRA-DF detector are as follows. Unless specified, the BR initialization scheme in Section 3.3.3 is used to generate the starting point. The stopping parameters of the PS iterations are set as $K_{\mathsf{max}} = 5$ and $\epsilon = 10^{-9}$; cf. Algorithm 3.1, line 17. We use the same SD implementations as in MMSE LD, both exact and inexact, to process each regularized lattice decoding problem in LDR LD.



(a) 4-QAM



(b) 16-QAM



(c) 64-QAM

Figure 3.1: Symbol error rate comparison for the SD-based detectors. $(M_C, N_C) = (16, 16)$.

3.4.1 Symbol Error Rate Performance of the LDR LD Detectors

In this subsection, we illustrate the SER performance of the LDR LD detector. The problem size is chosen as $(M_C, N_C) = (16, 16)$. Fig. 3.1 plots the SERs of the various detectors versus SNR under 4-QAM, 16-QAM and 64-QAM constellations. We have several key observations. First, the LDR LD detector achieves the same SER performance as the ML SD detector in the 4-QAM and 16-QAM cases. We are unable to verify whether the same desirable result holds for 64-QAM, since the ML SD detector is too slow to run in the 64-QAM case. Note that the identical performance of the ML SD and LDR LD detectors for the 4-QAM case is expected, since Theorem 3.1 shows that LDR is theoretically ML-optimal for 4-QAM. Second, the LDR LD detector gives SER performance no worse than the BC-SDR detector. In fact, the SER gaps between the LDR LD and BC-SDR detectors are significant for the 16-QAM and 64-QAM cases. This numerical observation is consistent with the duality gap theorem in Theorem 3.2, which suggests that LDR should provide approximation accuracies at least no worse than BC-SDR. Third, the LDR LD detector generally yields better SER performance than the NLD and MMSE LD detectors. Further comparing the LD detectors, we observe that NLD may suffer from more than 2dB performance loss relative to LDR LD, especially for 4-QAM and 16-QAM. The gap nevertheless reduces for 64-QAM. Another observation is that MMSE LD is quite close to LDR LD. This suggests that MMSE regularization is a good regularization under the exact SD implementation.

However, when we consider the runtime-limited inexact implementation, the SER gaps between the LDR method and the MMSE regularization-based method widen significantly. The inexact implementation results are also included in Fig. 3.1. We first notice that an inexact detector exhibits performance loss compared to its original exact counterpart, which is expected as a compromise for computational efficiency. The inexact LDR LD detector is seen to lose less, and perform best among all the inexact detectors. We observe that at SER= 10^{-5}, the SNR gains of the inexact LDR LD detector over the inexact MMSE LD detector (the second best) are 6dB, 4dB and 2dB for 4-QAM,

16-QAM and 64-QAM, respectively. The simulation results also show that the inexact ML SD detector is highly ineffective.

It is further observed that the inexact LDR LD detector has more performance advantages on small QAM sizes than large QAM sizes. This observation may be intuitively explained as follows: For small QAM sizes, constellation points on the symbol bounds constitute a larger portion of all constellation points, rendering out-of-bound symbol events more likely to occur. The LDR detector, which focuses on mitigating out-of-bound symbol effects, has more chances to kick in and improve the performance.

It is also interesting to examine the performance-complexity tradeoffs of the inexact LDR LD detector. Fig. 3.2 illustrates the SER performance of the inexact LDR LD detector and the inexact MMSE LD detector under various complexity limits. In the figure, the number c represents the power of the complexity limit. Specifically, for a given c, we set the worst-case complexity limit of the two inexact LD detectors to $N_{\text{node}} = \min\{N_C, N^{c-1} \times \text{SNR}\}$. We observe that the SER performance of the two detectors improves as c increases, and that the inexact LDR LD detector for $c = 4$ or above achieves a reasonably good performance. Also, for each given c, the inexact LDR LD detector is seen to outperform the inexact MMSE LD detector.

From all the observations above, we conclude that the LDR LD detector can yield near-ML performance, although the MMSE LD detector is also close. The inexact LDR LD detector has considerable SNR gains over inexact MMSE LD as well as the other inexact detectors; the gains are particularly significant for smaller QAM sizes.

3.4.2 Convergence of the LDR LD Detector

This subsection aims at studying the convergence behaviors of the PS iterations of the LDR LD detector. We start with considering the objective value convergence for a single problem realization. Fig. 3.3 shows the result, where the dual objective values $d_{\text{best}}^{(k)}$ of the LDR LD detector are plotted against the PS iteration number k (see Remark 3 for the precise definition of $d_{\text{best}}^{(k)}$). We tested

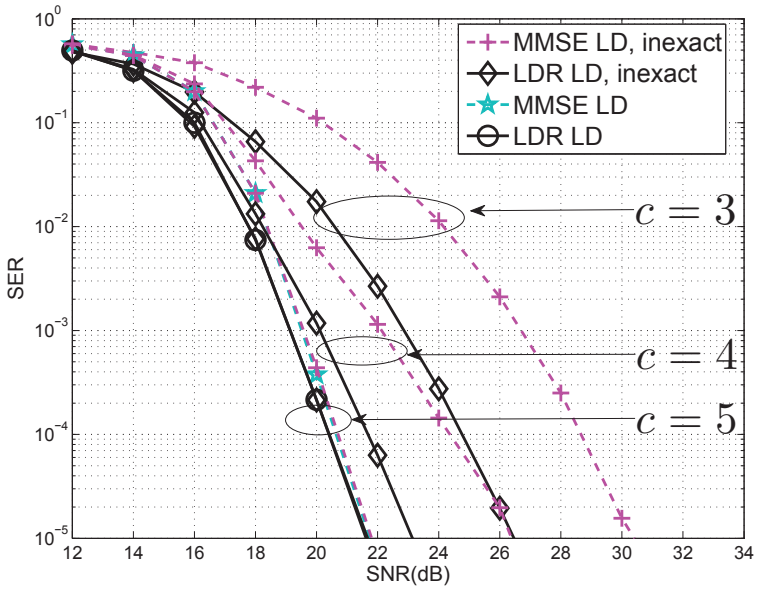


Figure 3.2: Symbol error rate comparison of the inexact LD detectors under various complexity limits. $(M_C, N_C) = (16, 16)$, 16-QAM.

the three initialization schemes, namely, random initialization, MMSE initialization and BR initialization. The problem size is $(M_C, N_C) = (16, 16)$, and the QAM size is 16. For reference, we also plot the optimal ML objective value, obtained via the ML SD detector. From Fig. 3.3, we observe that the three differently initialized LDR LD detectors converge to the same objective value. Also, for this problem realization, they converge to the optimal ML objective value. However, the convergence speed of the three initialization schemes has significant differences. It is seen that random initialization is the worse, taking more than 35 iterations to converge. MMSE initialization is better than random initialization, but the best is BR initialization: BR initialization takes only one iteration to converge at SNR=22dB, and three iterations at SNR=16dB.

We further illustrate the convergence speed of the LDR LD detectors by examining the average numbers of PS iterations to terminate. The stopping parameters of the PS iterations is set to be $K_{\max} = 50$ and $\epsilon = 10^{-9}$. A number of $10,000$ independent problem realizations was run. Again, we consider 16-QAM and $M_C = N_C$. Table 3.1 shows the average numbers of iterations with respect to the problem size N_C; the SNR is fixed at SNR= 22dB. It is observed

that the BR-initialized LDR LD detector takes about one iteration on average to complete the task, while the MMSE-initialized LDR LD detector takes $2-4$ iterations. Table 3.2 shows the average number of iterations with respect to the SNR, with $N_C = 16$. We can see that the average numbers of iterations of the LDR LD detectors tend to increase when the SNR decreases. Nevertheless, the BR-initialized LDR LD detector is very efficient in terms of the number of iterations used.

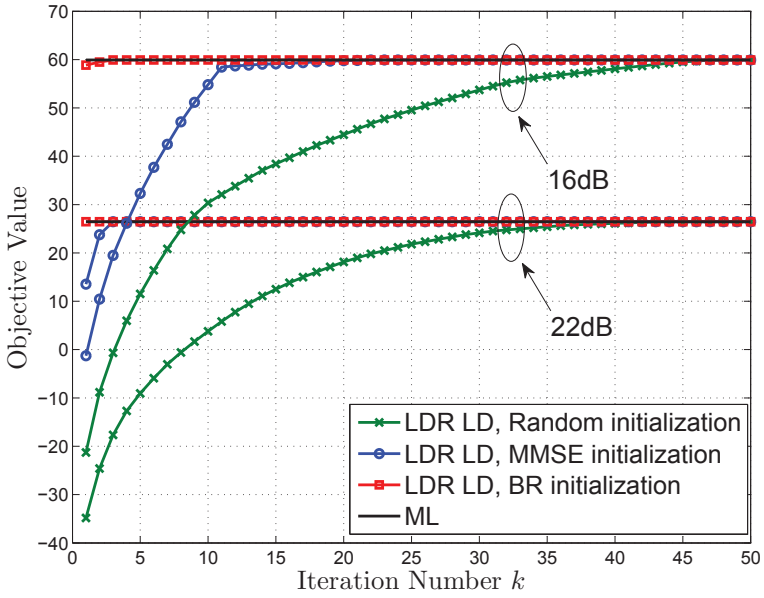


Figure 3.3: Convergence of the LDR LD detector in one realization. 16-QAM, $(M_C, N_C) = (16, 16)$.

	Problem size N_C					
	6	10	14	18	22	26
BR init.	1.083	1.016	1.009	1.010	1.008	1.006
MMSE init.	2.040	2.000	3.000	3.000	4.000	4.000

Table 3.1: Average number of PS iterations of the LDR LD detector. 16-QAM, SNR=22dB, $M_C = N_C$, $K_{\max} = 50$, $\epsilon = 10^{-9}$.

	SNR (dB)				
	16	17.5	19	20.5	22
BR init.	5.746	2.026	1.098	1.025	1.009
MMSE init.	23.76	9.402	5.115	4.001	3.000

Table 3.2: Average number of PS iterations of the LDR LD detector. 16-QAM, $(M_C, N_C) = (16, 16)$, $K_{\mathsf{max}} = 50$, $\epsilon = 10^{-9}$.

3.4.3 Complexity Performance of the LDR LD Detectors

In this subsection, we examine the complexities of the proposed LDR detectors. Fig. 3.4 presents the average number of floating point operations (FLOPs) of the various detectors with respect to the problem sizes, where we set SNR=22dB, $M_C = N_C$ and the QAM size to be 16. Note that the LDR LD detector is initialized by BR, and the overhead of computing the initialization, i.e., BR optimization, has been included in evaluating the FLOPs of the LDR detectors. For problem sizes $N_C \leq 6$, the complexity of the ML SD detector is very low, and is even faster than the suboptimal detectors. But its complexity becomes unacceptably large as the problem size increases, rendering the ML SD detector impractical for large problem sizes. All the other detectors have much lower complexities than the ML SD detector for problem sizes $N_C \geq 10$. We can see that the complexities of LDR LD and MMSE LD increase at a rate much lower than that of the ML SD detector, though they still exhibit exponential complexity behaviors eventually. Moreover, as expected, the complexities of the inexact detectors are much lower than those of their exact counterparts when the problem size is large.

The complexity comparison in Fig. 3.4 also reveals that for problem sizes $N_C \geq 18$, the LDR LD detector can be more efficient than the MMSE LD. This result seems counter-intuitive, since the LDR LD detector is an iterative LD method, rather than one-shot LD as in MMSE LD. The reason for this actually lies in the chosen initialization scheme for LDR LD, i.e., BR initialization. As illustrated previously, the BR-initialized LDR LD detector takes very few number of iterations to converge. For the problem setting here, the average number

of iterations is about one (cf. Table 3.1). Moreover, by empirical experience, we found that regularization via BR is helpful in improving the SD computational speed. To explain, in Table 3.3 we give a breakdown of the complexities of MMSE LD and LDR LD. In the table, "SD" represents the FLOPs consumed by the SD algorithm, and "others" the FLOPs of other operations, which include LLL reduction, BR optimization (for LDR LD only), and other matrix operations. We can see that LDR LD is always more expensive than MMSE LD on "others"; this makes sense because LDR LD requires solving the BR problem for initialization. However, LDR LD is much cheaper than MMSE LD on "SD" when $N_C \geq 14$. In fact, the SD complexities of both LDR LD and MMSE LD dominates the total complexities for $N_C \geq 14$. As a result, the LDR LD detector can be faster than MMSE LD for moderate to large problem sizes. As a future direction, it will be interesting to further investigate why BR regularization can accelerate SD so significantly.

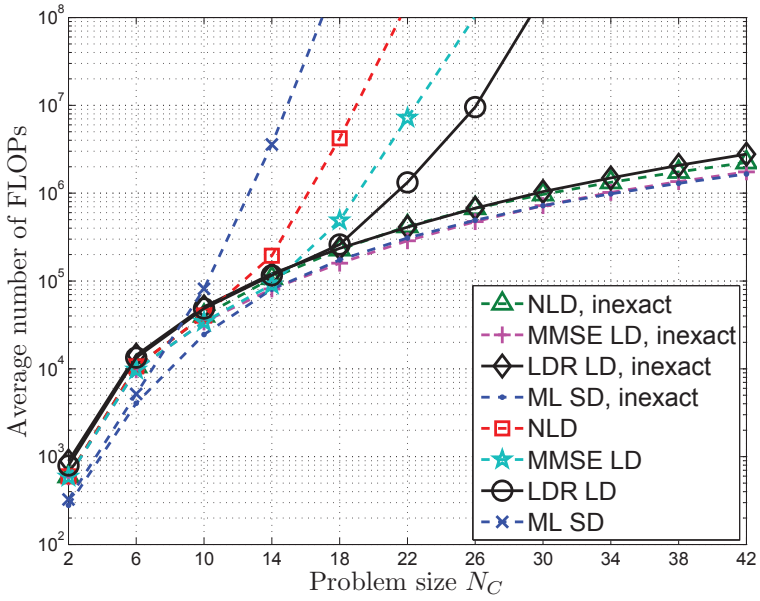


Figure 3.4: Average number of FLOPs of various detectors. 16-QAM, SNR=22dB, $M_C = N_C$.

		Problem Size N_C			
		6	14	22	26
SD	MMSE LD	**4.0e2**	1.9e4	7.0e6	1.0e8
	LDR LD	4.1e2	**5.6e3**	**9.7e5**	**9.0e6**
Others	MMSE LD	**9.4e3**	**7.2e4**	**2.1e5**	**3.1e5**
	LDR LD	1.1e4	8.0e4	3.5e5	5.4e5
Total	MMSE LD	**9.8e3**	9.1e4	7.2e6	1.0e8
	LDR LD	1.1e4	**8.5e4**	**1.3e6**	**9.5e6**

Table 3.3: Average number of FLOPs of the operations of the LDR LD and MMSE LD detectors. 16-QAM, SNR=22 dB, $M_C = N_C$.

3.4.4 Symbol Error Rate Performance of the LDR LRA-DF Detector

In this subsection, we test the SER performance of the LDR LRA-DF detector. The 16-QAM case is considered. Fig. 3.5 shows the SERs of the LDR LRA-DF detector and the MMSE LRA-DF detector under various problem sizes. It can be observed that the LDR LRA-DF detector exhibits dramatic performance gains compared to MMSE LRA-DF as the problem size increases. While LDR LRA-DF needs about 27dB to achieve SER=10^{-5} for the three problem sizes tested, MMSE LRA-DF requires a much higher SNR to achieve the same SER level—and this is particularly true for larger problem sizes. This demonstrates that the LDR method is also very effective in boosting the performance of the LRA-DF receiver approach.

3.5 Summary

This chapter addressed a regularization optimization problem in lattice decoding, by considering the LDR of the ML detection problem. It was found that the LDR problem can be regarded as that of finding the best diagonally regularized lattice decoding to approximate the ML detector, and that the well-known NLD and MMSE LD detectors can be seen as particular instances of LDR. We proved that in the 2-PAM constellation case, lattice decoding with a proper regularization is optimal. We also established a connection between

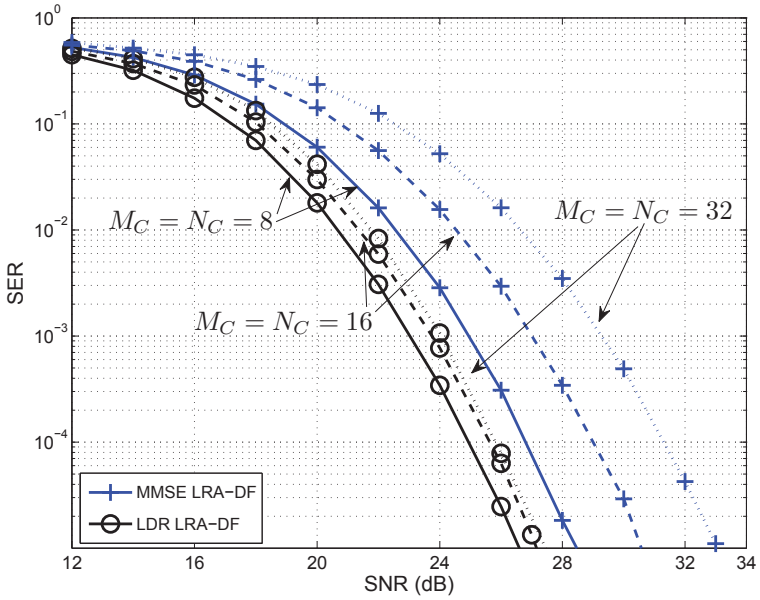


Figure 3.5: Symbol error rates of the LDR and MMSE LRA-DF detectors. 16-QAM.

LDR and a previously developed semidefinite relaxation-based method, showing that the former yields relaxation tightness at least no worse than the latter. The projected subgradient method was derived to solve the LDR problem, thereby obtaining the best regularization. Based on the idea of projected subgradient, we developed the LDR LD detector, and its approximations using conventional suboptimal lattice decoding methods. Simulation results showed that the LDR LD approach yields promising symbol error probability and complexity performance.

3.6 Appendix

3.6.1 Proof of Fact 1

Statement 1 is straightforward: Since the NLD problem (2.17) is a relaxation of the ML problem (3.1), NLD is automatically ML-optimal under instances where the NLD solution $\hat{\boldsymbol{s}}_{\mathrm{NLD}}$ is a feasible point of the ML problem. For Statement 2, we first note that $\|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\mathrm{NLD}}\|_2^2 = \varphi(\boldsymbol{0}) = d(\boldsymbol{0}) \leq d^\star \leq f^\star$. Since we have $f^\star = \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\mathrm{NLD}}\|_2^2$ by Statement 1, we obtain $f^\star = d^\star$. This subsequently implies that $\boldsymbol{\lambda} = \boldsymbol{0}$ is an optimal solution of the LDR problem. For

Statement 3, we prove the statement using contradiction. Suppose that $\hat{\boldsymbol{s}}_{\mathrm{NLD}}$ is not an optimal solution of problem $(\Phi_{\boldsymbol{\lambda}})$ for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\star$. This is equivalent to saying

$$\varphi(\boldsymbol{\lambda}^\star) < \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\mathrm{NLD}}\|_2^2 + \hat{\boldsymbol{s}}_{\mathrm{NLD}}^T \boldsymbol{D}(\boldsymbol{\lambda}^\star)\hat{\boldsymbol{s}}_{\mathrm{NLD}}. \tag{3.33}$$

Applying (3.33) to the strong duality result in Statement 2, we have

$$\begin{aligned} 0 &= d(\boldsymbol{\lambda}^\star) - f^\star && \text{(3.34a)} \\ &< \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\mathrm{NLD}}\|_2^2 + \hat{\boldsymbol{s}}_{\mathrm{NLD}}^T \boldsymbol{D}(\boldsymbol{\lambda}^\star)\hat{\boldsymbol{s}}_{\mathrm{NLD}} - u^2 \mathbf{1}^T \boldsymbol{\lambda}^\star - f^\star && \text{(3.34b)} \\ &= \left(\hat{\boldsymbol{s}}_{\mathrm{NLD}}^2 - u^2\mathbf{1}\right)^T \boldsymbol{\lambda}^\star && \text{(3.34c)} \end{aligned}$$

where (3.34c) is owing to $f^\star = \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\mathrm{NLD}}\|_2^2$. However, since $\boldsymbol{\lambda}^\star \succeq \mathbf{0}$ and $[\hat{\boldsymbol{s}}_{\mathrm{NLD}}]_i^2 \leq u^2$ for all $i = 1, \ldots, N$, we always obtain $\left(\hat{\boldsymbol{s}}_{\mathrm{NLD}}^2 - u^2\mathbf{1}\right)^T \boldsymbol{\lambda}^\star \leq 0$, which is a contradiction to (3.34c).

3.6.2 Proof of Lemma 3.1

The proof is by contradiction. Assume that (3.8) holds, and yet $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 > m^2$. Then we have

$$\begin{aligned} \varphi(\boldsymbol{\lambda}) &= \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2^2 + \hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}^T \boldsymbol{D}(\boldsymbol{\lambda})\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}} \\ &\geq \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2^2 + \lambda_i \left((m+2)^2 - 1\right) + \boldsymbol{\lambda}^T \mathbf{1}, \end{aligned} \tag{3.35}$$

where we have used the fact that $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_j^2 \geq 1$ for $j \neq i$ and $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 \geq (m+2)^2$ (recall $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}} \in 2\mathbb{Z}^N + \mathbf{1}$). By applying (3.8) to the second term of (3.35), we get

$$\begin{aligned} \varphi(\boldsymbol{\lambda}) &> \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2^2 + c(\boldsymbol{y}, \boldsymbol{H}) + \boldsymbol{\lambda}^T \mathbf{1} \\ &\geq \min_{\boldsymbol{s} \in \{\pm 1\}^N} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{\lambda}^T \mathbf{1}, \end{aligned} \tag{3.36}$$

where (3.36) is due to $\|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2^2 \geq \min_{\boldsymbol{s} \in 2\mathbb{Z}+\mathbf{1}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2$. From (3.36), we can equivalently write

$$\varphi(\boldsymbol{\lambda}) > \min_{\boldsymbol{s} \in \{\pm 1\}^N} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T \boldsymbol{D}(\boldsymbol{\lambda})\boldsymbol{s} \geq \varphi(\boldsymbol{\lambda}), \tag{3.37}$$

where the second inequality is by the definition of $\varphi(\boldsymbol{\lambda})$, cf. (3.5). Eq. (3.37) is clearly a contradiction. We therefore conclude that under the condition in (3.8), $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 \leq m^2$ must hold.

3.6.3 Proof of Theorem 3.1

Our proof is based on a strong duality result for general integer programming problems [76]. The result, for the ML problem, is stated as follows.

Lemma 3.3 *[Strong Lagrangian duality [76]] Let $\boldsymbol{\lambda} \succeq \mathbf{0}$, and $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}} \in 2\mathbb{Z}^N + \mathbf{1}$ be an optimal solution of problem $(\Phi_{\boldsymbol{\lambda}})$. If the following conditions are satisfied:*

$$\begin{aligned} \hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}^2 &\preceq u^2\mathbf{1}, \\ \boldsymbol{\lambda}^T\left(\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}^2 - u^2\mathbf{1}\right) &= 0, \end{aligned} \tag{3.38}$$

then $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ are optimal solutions of the the LDR problem (3.3) and ML problem (3.1), respectively. Moreover, strong duality $f^\star - d^\star$ holds.

Suppose that $\boldsymbol{\lambda} \succ \gamma_1\mathbf{1}$. By Lemma 3.1, we know that $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 \leq 1$ for all i. Since $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}} \in 2\mathbb{Z}^N + \mathbf{1}$, we must have $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 = 1$ for all i. Thus, the conditions in Lemma 3.3 are satisfied, and strong duality holds. It also follows from Lemma 3.3 that any $\boldsymbol{\lambda} \succ \gamma_1\mathbf{1}$ is LDR-optimal, and the corresponding $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ ML-optimal.

3.6.4 Proof of Theorem 3.2

The statement in Theorem 3.2 is the same as saying that $d^\star \geq g^\star$. To prove this result, we first use the implication

$$\boldsymbol{s} \in 2\mathbb{Z}^N + \mathbf{1} \Longrightarrow \boldsymbol{s}^2 \succeq \mathbf{1}$$

to obtain a lower bound on $d(\boldsymbol{\lambda})$ in (3.4)-(3.5)

$$d(\boldsymbol{\lambda}) \geq \min_{\boldsymbol{s}\in\mathbb{R}^N,\ \boldsymbol{s}^2\succeq\mathbf{1}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{\lambda}^T(\boldsymbol{s}^2 - u^2\mathbf{1}). \tag{3.39}$$

By weak duality, the right-hand side (RHS) of (3.39) is lower-bounded by its Lagrangian dual, which is given by

$$d(\boldsymbol{\lambda}) \geq \max_{\boldsymbol{\mu}\succeq\mathbf{0}} \min_{\boldsymbol{s}\in\mathbb{R}^N} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{\lambda}^T(\boldsymbol{s}^2 - u^2\mathbf{1}) - \boldsymbol{\mu}^T(\boldsymbol{s}^2 - \mathbf{1}), \tag{3.40}$$

where $\boldsymbol{\mu} \succeq \mathbf{0}$ is the dual variable for the constraint $\boldsymbol{s}^2 \succeq \mathbf{1}$. By recalling that $d^\star = \max_{\boldsymbol{\lambda}\succeq\mathbf{0}} d(\boldsymbol{\lambda})$, we arrive at

$$d^\star \geq \max_{\substack{\boldsymbol{\lambda}\succeq\mathbf{0},\\ \boldsymbol{\mu}\succeq\mathbf{0}}} \min_{\boldsymbol{s}\in\mathbb{R}^N} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{\lambda}^T(\boldsymbol{s}^2 - u^2\mathbf{1}) - \boldsymbol{\mu}^T(\boldsymbol{s}^2 - \mathbf{1}). \tag{3.41}$$

Our next step is to show that the RHS of (3.41) is equivalent to $g^\star$ in (3.11). The RHS of (3.41) can be rewritten as

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\mu},r}\ & r \\ \text{s.t. } & \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{s}\|_2^2+\boldsymbol{\lambda}^T(\boldsymbol{s}^2-u^2\mathbf{1})-\boldsymbol{\mu}^T(\boldsymbol{s}^2-\mathbf{1})\geq r,\ \forall \boldsymbol{s}\in\mathbb{R}^N \\ & \boldsymbol{\lambda}\succeq\mathbf{0},\ \boldsymbol{\mu}\succeq\mathbf{0}. \end{aligned} \tag{3.42}$$

By the lemma given in [77, p.163], we can equivalently turn problem (3.42) to

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\mu},r}\ & r \\ \text{s.t. } & \begin{bmatrix} \boldsymbol{H}^T\boldsymbol{H}+\boldsymbol{D}(\boldsymbol{\lambda}-\boldsymbol{\mu}) & -\boldsymbol{H}^T\boldsymbol{y} \\ -\boldsymbol{y}^T\boldsymbol{H} & \boldsymbol{\mu}^T\mathbf{1}-u^2\boldsymbol{\lambda}^T\mathbf{1}-r+\|\boldsymbol{y}\|_2^2 \end{bmatrix}\succeq\mathbf{0}, \\ & \boldsymbol{\lambda}\succeq\mathbf{0},\ \boldsymbol{\mu}\succeq\mathbf{0}. \end{aligned} \tag{3.43}$$

Let us take a look at the dual of problem (3.43). The Lagrangian function of problem (3.43) is written as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\mu},r,\boldsymbol{p},\boldsymbol{q},\boldsymbol{X}) \\ =&r+\boldsymbol{p}^T\boldsymbol{\mu}+\boldsymbol{q}^T\boldsymbol{\lambda} \\ &+\mathrm{tr}\boldsymbol{X}\begin{bmatrix} \boldsymbol{H}^T\boldsymbol{H}+\boldsymbol{D}(\boldsymbol{\lambda}-\boldsymbol{\mu}) & -\boldsymbol{H}^T\boldsymbol{y} \\ -\boldsymbol{y}^T\boldsymbol{H} & \boldsymbol{\mu}^T\mathbf{1}-u^2\boldsymbol{\lambda}^T\mathbf{1}-r+\|\boldsymbol{y}\|_2^2 \end{bmatrix}, \end{aligned}$$

where $\boldsymbol{p}\succeq\mathbf{0},\boldsymbol{q}\succeq\mathbf{0}$ and $\boldsymbol{X}\succeq\mathbf{0}$ are the dual variables for $\boldsymbol{\mu}\succeq\mathbf{0}$, $\boldsymbol{\lambda}\succeq\mathbf{0}$ and the first constraint in (3.43), respectively. By partitioning $\boldsymbol{X}$ as

$$\boldsymbol{X}=\begin{bmatrix} \boldsymbol{S} & \boldsymbol{s} \\ \boldsymbol{s}^T & t \end{bmatrix}$$

where $\boldsymbol{S}\in\mathbb{S}^N$, $\boldsymbol{s}\in\mathbb{R}^N$, $t\in\mathbb{R}$, the function $\mathcal{L}$ can be reorganized as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\mu},r,\boldsymbol{p},\boldsymbol{q},\boldsymbol{X})=&(1-t)r+(\boldsymbol{q}+\boldsymbol{d}(\boldsymbol{S})-tu^2\mathbf{1})^T\boldsymbol{\lambda} \\ &+(\boldsymbol{p}-\boldsymbol{d}(\boldsymbol{S})+t\mathbf{1})^T\boldsymbol{\mu} \\ &+\mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{S})-2\boldsymbol{s}^T\boldsymbol{H}^T\boldsymbol{y}+t\|\boldsymbol{y}\|_2^2. \end{aligned}$$

We note that the dual function respective to the above $\mathcal{L}$ is bounded only if the first three terms above are zero; i.e., $1-t=0$, $\boldsymbol{q}+\boldsymbol{d}(\boldsymbol{S})-tu^2\mathbf{1}=\mathbf{0}$ and

$\boldsymbol{p} - \boldsymbol{d}(\boldsymbol{S}) + t\mathbf{1} = \mathbf{0}$. As a result, the dual of problem (3.43) is

$$\min_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{S},\boldsymbol{s}} \quad \operatorname{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{S}) - 2\boldsymbol{s}^T\boldsymbol{H}^T\boldsymbol{y} + \|\boldsymbol{y}\|_2^2 \tag{3.44a}$$

$$\text{s.t.} \quad \boldsymbol{q} + \boldsymbol{d}(\boldsymbol{S}) - u^2\mathbf{1} = \mathbf{0}, \boldsymbol{p} - \boldsymbol{d}(\boldsymbol{S}) + \mathbf{1} = \mathbf{0}, \tag{3.44b}$$

$$\boldsymbol{p} \succeq \mathbf{0},\ \boldsymbol{q} \succeq \mathbf{0}, \tag{3.44c}$$

$$\begin{bmatrix} \boldsymbol{S} & \boldsymbol{s} \\ \boldsymbol{s}^T & 1 \end{bmatrix} \succeq \mathbf{0}. \tag{3.44d}$$

Since problem (3.44) is strictly feasible and bounded from below, by the conic duality theorem [77, Theorem 2.4.1], problems (3.43) and (3.44) attain the same optimal objectives; i.e., strong duality holds. Furthermore, by substituting (3.44b) into (3.44c) and applying Schur's complement [67] to (3.44d), we show that problem (3.44) is same as the BC-SDR problem in (3.11). The desired result $d^\star \geq g^\star$ is therefore concluded.

3.6.5 Proof of Lemma 3.2

First, we note that

$$\|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2 \geq \|\boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2 - \|\boldsymbol{y}\|_2 \geq \sigma_{\min}\|\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2 - \|\boldsymbol{y}\|_2, \tag{3.45}$$

where $\sigma_{\min}$ is the smallest singular value of $\boldsymbol{H}$, which is strictly positive for a full column rank $\boldsymbol{H}$. From (3.45), we obtain

$$\|\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2 \leq \frac{1}{\sigma_{\min}} \left(\|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2 + \|\boldsymbol{y}\|_2\right). \tag{3.46}$$

Second, recall from $(\Phi_{\boldsymbol{\lambda}})$ that $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ is an optimal solution of

$$\min_{\boldsymbol{s} \in 2\mathbb{Z}^N + \mathbf{1}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T\boldsymbol{D}(\boldsymbol{\lambda})\boldsymbol{s}. \tag{3.47}$$

Let

$$\mathcal{I} = \{\ i \in \{1, \ldots, N\} \mid \lambda_i > \gamma_1\ \},$$
$$\mathcal{J} = \{\ i \in \{1, \ldots, N\} \mid \lambda_i \leq \gamma_1\ \},$$

where γ_1 has been defined in (3.9). By Lemma 3.1, $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ must satisfy $[\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}]_i^2 \leq 1$ for all $i \in \mathcal{I}$. Hence, problem (3.47) can be equivalently expressed as

$$\begin{aligned} \min_{\boldsymbol{s}} \ & \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}_{\mathcal{J}}^T \boldsymbol{D}(\boldsymbol{\lambda}_{\mathcal{J}}) \boldsymbol{s}_{\mathcal{J}} \\ \text{s.t. } & \boldsymbol{s}_{\mathcal{I}}^2 = \mathbf{1}, \ \boldsymbol{s}_{\mathcal{J}} \in 2\mathbb{Z}^{|\mathcal{J}|} + \mathbf{1}, \end{aligned} \tag{3.48}$$

where $\boldsymbol{s}_{\mathcal{I}}$ denotes a subvector of $\boldsymbol{s}$ whose elements are $\{s_i\}_{i\in\mathcal{I}}$, and $\boldsymbol{s}_{\mathcal{J}}$ and $\boldsymbol{\lambda}_{\mathcal{J}}$ are defined in the same way. By noting that $\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}$ is optimal to (3.48), we have

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2^2 &\leq \|\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2^2 + \hat{\boldsymbol{s}}_{\boldsymbol{\lambda},\mathcal{J}}^T \boldsymbol{D}(\boldsymbol{\lambda}_{\mathcal{J}}) \hat{\boldsymbol{s}}_{\boldsymbol{\lambda},\mathcal{J}} \\ &= \min_{\boldsymbol{s}_{\mathcal{I}}^2 = \mathbf{1}, \ \boldsymbol{s}_{\mathcal{J}} \in 2\mathbb{Z}^{|\mathcal{J}|} + \mathbf{1}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 + \boldsymbol{s}_{\mathcal{J}}^T \boldsymbol{D}(\boldsymbol{\lambda}_{\mathcal{J}}) \boldsymbol{s}_{\mathcal{J}} \\ &\leq \|\boldsymbol{y} - \boldsymbol{H}\mathbf{1}\|_2^2 + \mathbf{1}^T \boldsymbol{D}(\boldsymbol{\lambda}_{\mathcal{J}}) \mathbf{1} \\ &\leq \|\boldsymbol{y} - \boldsymbol{H}\mathbf{1}\|_2^2 + N\gamma_1. \end{aligned} \tag{3.49}$$

Moreover, it can be shown from (3.9)-(3.10) that

$$\gamma_1 \leq \frac{1}{8}\|\boldsymbol{y} - \boldsymbol{H}\mathbf{1}\|_2^2. \tag{3.50}$$

Finally, by plugging (3.49)-(3.50) into (3.46), we obtain a finite upper bound on $\|\hat{\boldsymbol{s}}_{\boldsymbol{\lambda}}\|_2$, as desired.

3.6.6 Active Set Method for the BR problem

We consider solving the BR problem (3.29) using the active set method [74, 78]. The BR problem is

$$\begin{aligned} \min_{\boldsymbol{s}} \quad & f(\boldsymbol{s}) \triangleq \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s}\|_2^2 \\ \text{s.t.} \quad & -u \leq s_i \leq u, \ i = 1, \ldots, N, \end{aligned} \tag{3.51}$$

where $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{s} \in \mathbb{R}^N$, and $\boldsymbol{H} \in \mathbb{R}^{M\times N}$. We assume that $\boldsymbol{H}$ is of full column-rank.

The general idea of active set method is to identify those $s_i^\star$ that are active on the upper or lower bounds, i.e. $s_i^\star = \pm u$, where $\boldsymbol{s}^\star$ is the optimal solution of the BR problem. Let $\mathcal{L} = \{i \mid s_i^\star = -u\}$ and $\mathcal{U} = \{i \mid s_i^\star = u\}$ collect the indexex of $\boldsymbol{s}_i^\star$ that are $-u$ and u, respectively. If we know $\mathcal{L}$ and $\mathcal{U}$ in advance,

then the BR problem can be simplified as.

$$\min_{s_i, i \notin \mathcal{L} \cup \mathcal{U}} \quad \left\| \tilde{\boldsymbol{y}} - \sum_{i \notin \mathcal{L} \cup \mathcal{U}} \boldsymbol{h}_i s_i \right\|_2^2 \tag{3.52}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \sum_{i \in \mathcal{L}} \boldsymbol{h}_i(-u) - \sum_{i \in \mathcal{U}} \boldsymbol{h}_i u$. The optimization variables of (3.52) are $\{s_i\}_{i \notin \mathcal{L} \cup \mathcal{U}}$ which are unconstrained. Hence, problem (3.52) is a simple least square problem which can be easily solved. However, generally $\mathcal{L}$ and $\mathcal{U}$ are not known. The active set method is an iterative algorithm to identify $\mathcal{L}$ and $\mathcal{U}$.

We will briefly describe how the active set method works. Suppose at the current iteration, we have an iterate $\tilde{\boldsymbol{s}}$ and its bounding information

$$\mathcal{L} = \{i \mid \tilde{s}_i = -u\} \quad \text{and} \quad \mathcal{U} = \{i \mid \tilde{s}_i = u\}.$$

We want to find a step $\boldsymbol{p} \in \mathbb{R}^N$ such that after this step, the new iterate $\boldsymbol{s} = \tilde{\boldsymbol{s}} + \boldsymbol{p}$ can have a smaller objective value, but those s_i with $i \in \mathcal{L} \cup \mathcal{U}$ are still fixed at the corresponding $-u$ and u. This problem can be formulated as

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \|\boldsymbol{y} - \boldsymbol{H}(\tilde{\boldsymbol{s}} + \boldsymbol{p})\|_2^2 \\ \text{s.t.} \quad & p_i = 0, \quad \forall i \in \mathcal{L} \cup \mathcal{U}, \\ & p_i \in \mathbb{R}, \quad \forall i \notin \mathcal{L} \cup \mathcal{U}. \end{aligned} \tag{3.53}$$

Note that we do not constrain s_i with $i \notin \mathcal{L} \cup \mathcal{U}$ to satisfy $-u \leq s_i \leq u$. Problem (3.53) is a least square problem, and the solution can be easily obtained. Let us denote the optimal solution by $\boldsymbol{p}^\star$. Depending on whether $\boldsymbol{p}^\star$ is zero or not, we will take different strategies.

Case 1: $\boldsymbol{p}^\star \neq \boldsymbol{0}$. The case of $\boldsymbol{p}^\star \neq \boldsymbol{0}$ means that we can strictly decrease the objective value from $f(\tilde{\boldsymbol{s}})$ to $f(\tilde{\boldsymbol{s}} + \boldsymbol{p}^\star)$, since we assume that $\boldsymbol{H}$ is of full rank and the solution of problem (3.53) is unique. But if we take a full step $\boldsymbol{p}^\star$, the constraint $-u \leq \tilde{s}_i + p_i^\star \leq u$ may be violated for some $i \notin \mathcal{L} \cup \mathcal{U}$. (The constraints for those $i \in \mathcal{L} \cup \mathcal{U}$ will not be violated as $\tilde{s}_i$ is feasible and $p_i^\star = 0$.)

If $\boldsymbol{s} = \tilde{\boldsymbol{s}} + \boldsymbol{p}^\star$ is feasible, then we take a full step

$$\boldsymbol{s} = \tilde{\boldsymbol{s}} + \boldsymbol{p}^\star.$$

Update $\mathcal{L}$ and $\mathcal{U}$ according to $\boldsymbol{s}$, and start a new iteration with problem (3.53) with the new iterate $\boldsymbol{s} = \tilde{\boldsymbol{s}} + \boldsymbol{p}^\star$ and the corresponding bounding information $\mathcal{L}$ and $\mathcal{U}$.

If $\boldsymbol{s} = \tilde{\boldsymbol{s}} + \boldsymbol{p}^\star$ is infeasible, some s_i violates the constraint $-u \leq s_i \leq u$. Then we want to take the largest step-size $\alpha \geq 0$ such that

$$\boldsymbol{s} = \tilde{\boldsymbol{s}} + \alpha \boldsymbol{p}^\star$$

is still feasible . The step-size α can be easily determined as

$$\alpha = \min_{i \notin \mathcal{L} \cup \mathcal{U}} \alpha_i,$$

where $\alpha_i = (u - \tilde{s}_i)/p_i^\star$ when $p_i^\star > 0$ and $\alpha_i = -(u + \tilde{s}_i)/p_i^\star$ when $p_i^\star < 0$. Here, α_i is the largest step-size that the ith constraint will not be violated. Then, we take the step

$$\boldsymbol{s} = \tilde{\boldsymbol{s}} + \alpha \boldsymbol{p}^\star.$$

Now, there must be one or more indexes $i \notin \mathcal{L} \cup \mathcal{U}$ such that s_i is $-u$ or u. Choose such an index and move it to $\mathcal{L}$ if $s_i = -u$ and to $\mathcal{U}$ if $s_i = u$. A new iteration begins with the new iterate $\tilde{\boldsymbol{s}}$ with $\mathcal{L}$ and $\mathcal{U}$.

Case 2: $\boldsymbol{p}^\star = \boldsymbol{0}$. If it happens that $\boldsymbol{p}^\star = \boldsymbol{0}$, then $\tilde{\boldsymbol{s}}$ is already a minimizer of (3.53). This suggests that $\tilde{\boldsymbol{s}}$ could be the solution of the original BR problem (3.51). To verify, we check the KKT conditions of (3.51) which can be written as

$$\begin{cases} w_i \leq 0, & i \in \mathcal{L} \\ w_i \geq 0, & i \in \mathcal{U} \end{cases} \tag{3.54}$$

where $\boldsymbol{w}$ is the gradient of $f(\boldsymbol{s})$ at $\tilde{\boldsymbol{s}}$ and is given by

$$\boldsymbol{w} = \boldsymbol{H}^T(\boldsymbol{y} - \boldsymbol{H}\tilde{\boldsymbol{s}}). \tag{3.55}$$

If (3.54) holds true, then $\tilde{\boldsymbol{s}}$ is the solution of (3.51).

If the KKT condition does not hold, we need to remove one index t from $\mathcal{L} \cup \mathcal{U}$. The reason is that $\tilde{\boldsymbol{s}}$ has already minimized $f(\boldsymbol{s})$ subject to $\tilde{s}_i = -u,\ \forall\ i \in \mathcal{L}$ and $\tilde{s}_i = u,\ \forall\ i \in \mathcal{U}$. Progress can not be made if we do not drop one index from $\mathcal{L} \cup \mathcal{U}$. To choose an index t from $\mathcal{L} \cup \mathcal{U}$, let $\mathcal{V}$ collect those indexes with w_i violating the KKT conditions, i.e. $\mathcal{V} = \{i \mid w_i > 0, i \in \mathcal{L}\} \cup \{i \mid w_i < 0, i \in \mathcal{U}\}$. Recall that in gradient descent method, the descent strategy is to walk along

the negative gradient direction with some step-size β, i.e. $\tilde{\boldsymbol{s}} + \beta \boldsymbol{w}$. So $|w_i|$ can be viewed as the aggressiveness of $\tilde{s}_i$ to move. Naturally, we will choose t as

$$t = \max_{i \in \mathcal{V}} |w_i|. \tag{3.56}$$

Then we remove t from $\mathcal{L}$ or $\mathcal{U}$. A new iteration begins with the same iterate $\tilde{\boldsymbol{s}}$, but different $\mathcal{L}$ and $\mathcal{U}$.

We briefly discuss the convergence here. On the one hand, we have seen that if a full step can be taken, a strict decrease in $f(\boldsymbol{s})$ can be assured. It can be shown that as long as the step-size α is not zero, a strict decrease in $f(\boldsymbol{s})$ can be still achieved. This means that if $\boldsymbol{p}^\star \neq \boldsymbol{0}$, a strict decrease in $f(\boldsymbol{s})$ is achieved. On the other hand, it can be shown that t in (3.56) will lead to a strict decrease in the function value $f(\boldsymbol{s})$. This is because it can be shown that at the next iteration, the optimal solution $\boldsymbol{p}^\star$ of the problem (3.53) is not zero, and we can take a positive step length. Thus, a strict decrease of function value $f(\boldsymbol{s})$ is achieved at every iteration, and every iteration corresponds to a particular pair of $\mathcal{L}$ and $\mathcal{U}$. Because there is only finitely pairs of $\mathcal{L}$ and $\mathcal{U}$, the active set method can find the exact optimal solution in a finite time.

Note that we can properly combine case 1 and case 2 for better efficiency. It can be observed that in case 1 when $\boldsymbol{p}^\star \neq \boldsymbol{0}$ and a full step $\boldsymbol{s} = \tilde{\boldsymbol{s}} + \boldsymbol{p}^\star$ is taken, at the next iteration $\boldsymbol{p}^\star$ becomes zero, and case 2 will follow. Hence, we can directly execute the operations in case 2 whenever $\boldsymbol{p}^\star \neq \boldsymbol{0}$ and a full step is taken. The description of the general idea of active set method is completed. We provide the pseudo-code in Algorithm 3.2. In Section 3.6.7, we introduce a rank-one pseudo-inverse updating method that can significantly reduce the complexity of the active set method.

3.6.7 One-column Pseudo-inverse Update for the Active Set Method

We describe a one-column pseudo-inverse update method [75] to speed up the active set method. A close look at Algorithm 3.2 would reveal that the most heavy computational cost of Algorithm 3.2 lies in Line 1, i.e. the computation

of the least square problem (3.53) which is rewritten as follows

$$\begin{aligned} \min_{\bar{\boldsymbol{p}}} \quad & \|\bar{\boldsymbol{y}} - \bar{\boldsymbol{H}}\bar{\boldsymbol{p}}\|_2 \\ \text{s.t.} \quad & \bar{\boldsymbol{p}} \in \mathbb{R}^n \end{aligned} \tag{3.57}$$

where $\bar{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{s} \in \mathbb{R}^m$, $\bar{\boldsymbol{H}} = [h_i]_{i \notin \mathcal{L} \cup \mathcal{U}}$, and $\bar{\boldsymbol{p}} = [p_i]_{i \notin \mathcal{L} \cup U}$, $m = M$, and $n = N - |\mathcal{L} \cup \mathcal{U}|$. The optimal solution $\bar{\boldsymbol{p}}$ is given by

$$\bar{\boldsymbol{p}}^\star = (\bar{\boldsymbol{H}}^T \bar{\boldsymbol{H}})^{-1} \bar{\boldsymbol{H}}^T \bar{\boldsymbol{y}},$$

which would cost a computational complexity $\mathcal{O}(\frac{1}{3}n^3 + n^2 m)$. Among all the operation in computing $\bar{\boldsymbol{p}}^\star$, the computation of $(\bar{\boldsymbol{H}}^T \bar{\boldsymbol{H}})^{-1}$ is the most expensive and have complexity $\mathcal{O}(\frac{1}{3}n^3 + n^2 m)$. Thus, if we can speed up the computation of $(\bar{\boldsymbol{H}}^T \bar{\boldsymbol{H}})^{-1}$, the active set method can be much faster.

The introduced method exploits the fact that in two consecutive iterations of Algorithm 3.2 the optimization problem (3.57) would not be too different. To describe, let us denote by $\bar{\boldsymbol{H}}_0$ the matrix in the previous iteration and $\bar{\boldsymbol{H}}$ in the current iteration. Then the relationship of $\bar{\boldsymbol{H}}_0$ and $\bar{\boldsymbol{H}}$ can be summarized as follows.

- **Removing a column.** If Line 17 of Algorithm 3.2 is executed in the previous iteration, then $\bar{\boldsymbol{H}}$ is a submatrix of $\bar{\boldsymbol{H}}_0$ with one column removed. Specifically, $\bar{\boldsymbol{H}}$ and $\bar{\boldsymbol{H}}_0$ satisfy

$$\bar{\boldsymbol{H}}_0 = [\bar{\boldsymbol{H}}, \bar{\boldsymbol{h}}]. \tag{3.58}$$

- **Inserting a column.** If Line 11 of Algorithm 3.2 is executed in the previous iteration, then $\bar{\boldsymbol{H}}$ contains $\bar{\boldsymbol{H}}_0$ and an additional column. Let $\bar{\boldsymbol{h}}$ denote the column to be inserting to $\bar{\boldsymbol{H}}_0$ Then, $\bar{\boldsymbol{H}}$ is

$$\bar{\boldsymbol{H}} = [\bar{\boldsymbol{H}}_0, \bar{\boldsymbol{h}}]. \tag{3.59}$$

Note that we have assumed for simplicity that the column to be removed or inserted is the last column. The case of other column can be done via the same operations with an additional permutation step. We can see that the difference between $\bar{\boldsymbol{H}}_0$ and $\bar{\boldsymbol{H}}$ is only an column. If we have computed $\bar{\boldsymbol{H}}_0$ and $(\bar{\boldsymbol{H}}_0^T \bar{\boldsymbol{H}}_0)^{-1}$

in the previous active set method iteration, then the computation of $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1}$ at the current iteration can accelerated.

We will use the following identity for an invertible matrix $\boldsymbol{A}$ [79].

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}_{12}^T & a_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b}_{12} \\ \boldsymbol{b}_{12}^T & b_{22} \end{bmatrix}, \tag{3.60}$$

where

$$\begin{aligned} \boldsymbol{B}_{11} &= \boldsymbol{F}_{11}^{-1} \\ \boldsymbol{b}_{12} &= -\boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}f_{22}^{-1} \\ b_{22} &= f_{22}^{-1} \end{aligned}$$

and

$$\boldsymbol{F}_{11} = \boldsymbol{A}_{11} - \boldsymbol{a}_{12}\boldsymbol{a}_{12}^T a_{22}^{-1} \tag{3.61}$$

$$f_{22} = a_{22} - \boldsymbol{a}_{12}^T\boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}. \tag{3.62}$$

By the matrix inversion lemma, $\boldsymbol{F}_{11}$ and f_{22} also satisfy

$$\boldsymbol{F}_{11}^{-1} = \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}\boldsymbol{a}_{12}^T\boldsymbol{A}_{11}^{-1}f_{22}^{-1}. \tag{3.63}$$

Using these identities, we can compute $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1}$ via $(\bar{\boldsymbol{H}}_0^T\bar{\boldsymbol{H}}_0)^{-1}$.

- **Removing a column.** In this case, we are given $(\bar{\boldsymbol{H}}_0^T\bar{\boldsymbol{H}}_0)^{-1}$ with $\bar{\boldsymbol{H}}_0 = [\bar{\boldsymbol{H}}, \bar{\boldsymbol{h}}]$ and want to find $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1}$. Set $\boldsymbol{A} = \bar{\boldsymbol{H}}_0^T\bar{\boldsymbol{H}}_0$ and $\boldsymbol{B} = \boldsymbol{A}^{-1}$, we have

  $$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}_{12}^T & a_{22} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}} & \bar{\boldsymbol{H}}^T\bar{\boldsymbol{h}} \\ \bar{\boldsymbol{h}}^T\bar{\boldsymbol{H}} & \bar{\boldsymbol{h}}^T\bar{\boldsymbol{h}} \end{bmatrix}. \tag{3.64}$$

  By (3.63), we have

  $$\begin{aligned} (\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1} &= \boldsymbol{A}_{11}^{-1} \\ &= \boldsymbol{F}_{11}^{-1} - \boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}\boldsymbol{a}_{12}^T\boldsymbol{A}_{11}^{-1}f_{22}^{-1} \\ &= \boldsymbol{B}_{11} - \boldsymbol{b}_{12}\boldsymbol{b}_{12}^T b_{22}^{-1}. \end{aligned}$$

  As $\boldsymbol{B}$ is known, $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1}$ can be easily computed by (3.64) via a rank-one update, which costs a complexity $\mathcal{O}(2n^2)$.

- **Inserting a column.** In this case, we are given $(\bar{\boldsymbol{H}}_0^T \bar{\boldsymbol{H}}_0)^{-1}$ and $\bar{\boldsymbol{H}}_0$. We would like to compute $(\bar{\boldsymbol{H}}^T \bar{\boldsymbol{H}})^{-1}$ with $\bar{\boldsymbol{H}} = [\bar{\boldsymbol{H}}_0, \bar{\boldsymbol{h}}]$. Setting $\boldsymbol{A} = \bar{\boldsymbol{H}}^T \bar{\boldsymbol{H}}$ and $\boldsymbol{B} = \boldsymbol{A}^{-1}$, we have

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{a}_{12} \\ \boldsymbol{a}_{12}^T & a_{22} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{H}}_0^T \bar{\boldsymbol{H}}_0 & \bar{\boldsymbol{H}}_0^T \bar{\boldsymbol{h}} \\ \bar{\boldsymbol{h}}^T \bar{\boldsymbol{H}}_0 & \bar{\boldsymbol{h}}^T \bar{\boldsymbol{h}} \end{bmatrix}$$

and

$$\begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b}_{12} \\ \boldsymbol{b}_{12}^T & b_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1} \boldsymbol{a}_{12} f_{22}^{-1} \\ -\boldsymbol{a}_{12}^T \boldsymbol{A}_{11}^{-1} f_{22}^{-1} & f_{22}^{-1} \end{bmatrix}, \tag{3.65}$$

where

$$\begin{aligned} \boldsymbol{F}_{11}^{-1} &= \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1} \boldsymbol{a}_{12} \boldsymbol{a}_{12}^T \boldsymbol{A}_{11}^{-1} f_{22}^{-1} \\ f_{22} &= a_{22} - \boldsymbol{a}_{12}^T \boldsymbol{A}_{11}^{-1} \boldsymbol{a}_{12}. \end{aligned} \tag{3.66}$$

As $\boldsymbol{A}_{11}^{-1} = (\bar{\boldsymbol{H}}_0^T \bar{\boldsymbol{H}}_0)^{-1}$ is known and $\boldsymbol{a}_{12} = \bar{\boldsymbol{H}}_0^T \bar{\boldsymbol{h}}$ can be computed easily, $\boldsymbol{B}$ can be computed by (3.65) and (3.66) very efficiently. The complexity is again $\mathcal{O}(2n^2)$.

Algorithm 3.2: Active Set Method for the BR Broblem

input : $\boldsymbol{y}$, $\boldsymbol{H}$, u, and an initialization $\boldsymbol{s}^0$

Set $\boldsymbol{s} = \boldsymbol{s}^0$ and compute the bounding information $\mathcal{L}$ and $\mathcal{U}$.

repeat

  Obtain the optimal solution $\boldsymbol{p}^\star$ of the following least square problem

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \|\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{s}+\boldsymbol{p})\|_2^2 \\ \text{s.t.} \quad & p_i = 0, \quad \forall i \in \mathcal{L} \cup \mathcal{U}, \\ & p_i \in \mathbb{R}, \quad \forall i \notin \mathcal{L} \cup \mathcal{U}. \end{aligned}$$

  if $-u\mathbf{1} \preceq \boldsymbol{s} + \boldsymbol{p}^\star \preceq u\mathbf{1}$ **then**

    Update $\boldsymbol{s} = \boldsymbol{s} + \boldsymbol{p}^\star$;

    Set $\boldsymbol{w} = \boldsymbol{H}^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{s})$;

    if $w_i \le 0$ *for all* $i \in \mathcal{L}$ *and* $w_i \ge 0$ *for all* $i \in \mathcal{U}$ **then**

      Optimal solution is found and stop;

    else

      Compute $t = \arg\max_{i \in \mathcal{L}\cup\mathcal{U}} |w_i|$;

      Remove t from $\mathcal{L}$ or $\mathcal{U}$;

    end

  else

    Compute α_i for all $i \notin \mathcal{L} \cup \mathcal{U}$,

$$\alpha_i = \begin{cases} (u - s_i)/p_i^\star, & \text{if } p_i^\star > 0 \\ -(u + s_i)/p_i^\star, & \text{if } p_i^\star < 0 \\ \infty, & \text{if } p_i^\star = 0. \end{cases}$$

    Solve $\alpha = \min_{i \notin \mathcal{L}\cup\mathcal{U}} \alpha_i$, and set α as the objective and $i^\star$ as the minimizer.

    Update $\boldsymbol{s} = \boldsymbol{s} + \alpha \boldsymbol{p}^\star$;

    Update $\mathcal{L} = \mathcal{L} \cup \{i^\star\}$ if $s_{i^\star} = -u$ and $\mathcal{U} = \mathcal{U} \cup \{i^\star\}$ if $s_{i^\star} = u$;

  end

until *some stopping criteria are satisfied*;

output: $\boldsymbol{s}$

Algorithm 3.3: Removing the jth column

1 Permute column j and row j of $\boldsymbol{B}$ to the last column and row, receptively;

2 Partition $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b}_{12} \\ \boldsymbol{b}_{12}^T & b_{22} \end{bmatrix}$;

3 Set $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1} = \boldsymbol{B}_{11} - \boldsymbol{b}_{12}\boldsymbol{b}_{12}^T/b_{22}$;

Algorithm 3.4: Inserting a column to position j

1 $\boldsymbol{a}_{12} = \bar{\boldsymbol{H}}_0^T\bar{\boldsymbol{h}}$;

2 $\boldsymbol{u} = \boldsymbol{A}_{11}^{-1}\boldsymbol{a}_{12}$;

3 $b_{22} = 1/(\bar{\boldsymbol{h}}^T\bar{\boldsymbol{h}} - \boldsymbol{u}^T\boldsymbol{a}_{12})$;

4 $\boldsymbol{b}_{12} = -b_{22}\boldsymbol{u}$;

5 $\boldsymbol{B}_{11} = \boldsymbol{A}_{11}^{-1} + b_{22}\boldsymbol{u}^T\boldsymbol{u}$;

6 $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b}_{12} \\ \boldsymbol{b}_{12}^T & b_{22} \end{bmatrix}$;

7 Permute column j and row j of $(\bar{\boldsymbol{H}}^T\bar{\boldsymbol{H}})^{-1}$ to the last column and row, respectively.

Chapter 4

Vector Perturbation with Per-antenna Power Constraint

4.1 Introduction

In this chapter, we consider the downlink scenario where the base station broadcasts different information to different users simultaneously. The non-cooperative nature among users in multiuser communication means that the base station must carefully design its transmitting signal to cancel multiuser interference. Vector perturbation [13, 39] is a promising technique to achieve the sum capacity of the broadcast channel. One salient feature of vector perturbation is that the burden of signal processing goes with the base station and the users only require simple processing such as modulo operation and scalar quantization. Such an advantage of vector perturbation is very desirable in mobile device communications where the battery life is a great concern. From the signal processing point of view, vector perturbation absorbs two ingredients from the channel inversion method [80] and Tomlinson-Harashima precoding (THP) [81,82]. The first one is that vector perturbation inverses the channel effect at the base station so that multiuser interference is totally eliminated. The second ingredient is the perturbation technique originated from THP. A perturbation vector is deliberately added to the information vector for minimizing the power of the unnormalized transmitting signal, which results in significantly reduced effective noise at the users. In this thesis, we will name the vector

perturbation scheme of [13, 39] as conventional vector perturbation (CVP) [83].

However, the practical implementation of CVP is hindered by its very high per-antenna power. The CVP design only considers a total power constraint and thus may allocate a significant portion of the total power to a single antenna. It is difficult for the analog frontend of each antenna to meet such a high per-antenna power requirement. When the per-antenna power exceeds the operating region of the analog frontend, signal distortion or even signal clipping may occur which significantly degrades the system performance. Such a problem can be handled by introducing power back-off, but this in turn will degrade the error rate performance.

In this chapter, we consider vector perturbation with per-antenna power constraint (VP-PAPC). We add explicitly a PAPC into the CVP problem so that a strict upper bound on the maximum per-antenna power is guaranteed. We show that the resulting VP-PAPC problems, which have more stringent constraints than CVP, are always feasible and the corresponding transmission schemes achieve the same diversity as CVP. We develop fast algorithm to handle the VP-PAPC problem by using the LDR and LRA-DF techniques. Simulation results show that VP-PAPC schemes can effectively limit the per-antenna power, thereby avoiding signal clipping and reducing power back-off in CVP.

The rest of this chapter is organized as follows. We introduce the system model and vector perturbation in Section 4.2. Then, in Section 4.3, we propose the VP-PAPC formulation under the instantaneous power normalization assumption and investigate its property. This is followed by Section 4.4 where we consider VP-PAPC with short-term power normalization and propose LDR fast approximations. We then use simulation results to demonstrate the performance of the proposed methods in Section 4.5. Section 4.6 summarizes this chapter.

4.2 Background

4.2.1 System Model

We consider a multiuser MISO broadcast system model. A base station that is equipped with N antennas serves $M \leq N$ single-antenna users. The base

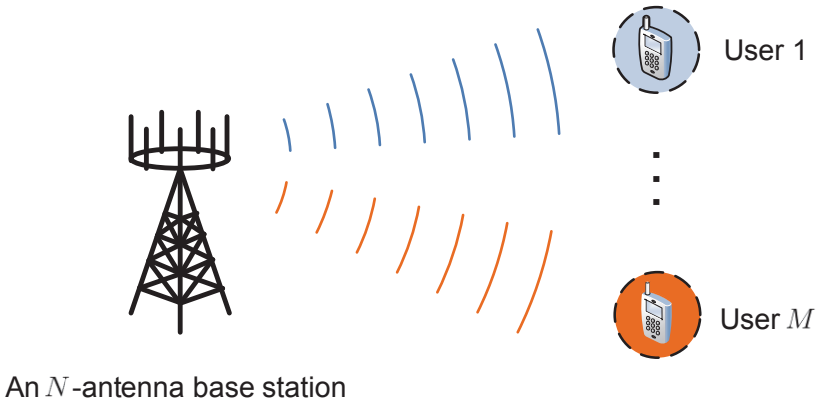


Figure 4.1: An N-antenna base station transmits to M single-antenna users.

station sends the information vector $\boldsymbol{s} \in \mathcal{S}^M$ simultaneously to all the users with s_m intended to the mth user. The receive signal of the mth user is

$$y_m = \boldsymbol{h}_m^H \boldsymbol{x} + \nu_m, \tag{4.1}$$

where $\boldsymbol{h}_m \in \mathbb{C}^N$ is the channel of the mth user, ν_m the noise at the mth user, and $\boldsymbol{x}$ the transmitting signal of the base station. By stacking $\boldsymbol{y} = [y_1, \ldots, y_M]^T$, $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_M]^T$, and $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots \boldsymbol{h}_M]^H$, we have a compact representation of the signal model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\nu}, \tag{4.2}$$

where the noise $\boldsymbol{\nu}$ is assumed to follow $\mathcal{CN}(\boldsymbol{0}, \sigma_\nu^2 \boldsymbol{I})$. The transmission signal $\boldsymbol{x}$ is subject to the total power constraint

$$\|\boldsymbol{x}\|_2^2 \leq P_{\mathrm{T}}, \tag{4.3}$$

where P_{T} is the available total power budget. The goal of the precoder is to transmit reliably the information vector $\boldsymbol{s}$ to the users by designing the transmitting signal $\boldsymbol{x}$ according to $\boldsymbol{s}$, the channel $\boldsymbol{H}$, and the total power budget P_{T}.

4.2.2 Channel Inversion and Vector Perturbation

In the multiuser scenario, the lack of cooperative processing of the received signals makes it impossible to cancel multiuser interference without a proper precoding at the base station. One possible precoding strategy that can completely eliminate multiuser interference is the channel inversion method [80]. In channel inversion method, the transmitting signal takes the following form

$$\boldsymbol{x} = \sqrt{\frac{P_{\mathrm{T}}}{\gamma}}\boldsymbol{d}, \tag{4.4a}$$

$$\boldsymbol{d} = \boldsymbol{H}^{\dagger}\boldsymbol{s}, \tag{4.4b}$$

where $\boldsymbol{H}^{\dagger} = \boldsymbol{H}^{H}(\boldsymbol{H}\boldsymbol{H}^{H})^{-1} \in \mathbb{C}^{N\times M}$ is the Moore-Penrose pseudo-inverse of the channel $\boldsymbol{H}$, and γ is a power normalization factor that makes the transmitting signal $\boldsymbol{x}$ satisfy the total power constraint. Here, $\boldsymbol{x}$ and $\boldsymbol{d}$ are named as normalized and unnormalized transmitting signals respectively, as $\boldsymbol{x}$ alway satisfies the total power constraint while $\boldsymbol{d}$ may not. Due to channel inversion, the receive signal of the mth user can be written as

$$y_m = \boldsymbol{h}_m^H\boldsymbol{x} + \nu_m = \sqrt{\frac{P_{\mathrm{T}}}{\gamma}}s_m + \nu_m. \tag{4.5}$$

The mth user, only based on its own received signal y_m, can detect the intended symbol s_m without multiuser interference. The detection can be done by quantizing

$$\tilde{s}_m = \sqrt{\frac{\gamma}{P_{\mathrm{T}}}}y_m = s_m + \sqrt{\frac{\gamma}{P_{\mathrm{T}}}}\nu_m \tag{4.6}$$

with respect to the constellation $\mathcal{S}$. The problem of channel inversion is that when $\boldsymbol{H}$ is close to being ill-conditioned, the unnormalized transmitting signal $\boldsymbol{d}$ has a high power, i.e. $\|\boldsymbol{d}\|_2^2$ is large. In this case, in order to ensure the power constraint (4.3) on the normalized transmitting signal $\boldsymbol{x}$ is met, the power normalization factor γ must be large. However, as seen in (4.6), a large γ would significantly increase the effective noise power at the user, rendering the detection more vulnerable to noise.

The conventional vector perturbation (CVP) technique proposed by Peel, Hochwald and Swindlehurst in [13, 39] can significantly alleviate the problem

encountered in the channel inversion method. Similar to the channel inversion method, CVP uses channel inversion as a means to cancel the multiuser interference. The salient feature of CVP is the use of the perturbation technique which originates from Tomlinson-Harashima precoding (THP) [81,82]. In CVP, the transmitting signal $\boldsymbol{x}$ and $\boldsymbol{d}$ take the following form

$$\begin{aligned} \boldsymbol{x} &= \sqrt{\frac{P_{\mathrm{T}}}{\gamma}}\boldsymbol{d}, \\ \boldsymbol{d} &= \boldsymbol{H}^{\dagger}(\boldsymbol{s}+\delta\boldsymbol{l}), \end{aligned} \tag{4.7}$$

where δ is a constant and $\boldsymbol{l} \in \mathbb{G}^M$ is a perturbation vector to be determined. Here, the perturbation vector $\boldsymbol{l}$ belongs to the set of Gaussian numbers $\mathbb{G}^N = (\mathbb{Z}+\boldsymbol{j}\mathbb{Z})^N$, which means that the real and imaginary parts of $\boldsymbol{l}$ are integers. The constant δ is usually chosen such that $\mathrm{conv}(\mathcal{S}+\delta\boldsymbol{l})$ is non-overlapping for all $\boldsymbol{l} \in \mathbb{G}^N$, where $\mathrm{conv}(\cdot)$ denotes the convex hull. For example, for the $(u+1)^2$-QAM constellation (u is a positive old number), δ is chosen as $\delta = 2(u+1)$.

From the discussion of the channel inversion method, we have seen that the performance bottleneck arises from the large power of the unnormalized transmitting signal. To deal with this problem, CVP finds an optimal perturbation vector that minimizes the power of the unnormalized transmitting power, i.e.,

$$\begin{aligned} (\mathrm{CVP}) \quad \min_{\boldsymbol{d},\boldsymbol{l}} \quad & \|\boldsymbol{d}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{d} = \boldsymbol{H}^{\dagger}(\boldsymbol{s}+\delta\boldsymbol{l}), \\ & \boldsymbol{d} \in \mathbb{C}^N,\ \boldsymbol{l} \in \mathbb{G}^M. \end{aligned} \tag{4.8}$$

The CVP problem is an integer least square problem which can be solved optimally by the sphere decoding algorithms [12,16] or approximately by the lattice reduction aided methods [29,63].

At the mth user, the receive signal after scaled by $\sqrt{\gamma/P_{\mathrm{T}}}$ is given by

$$\sqrt{\frac{\gamma}{P_{\mathrm{T}}}}y_m = s_m + \delta l_m + \sqrt{\frac{\gamma}{P_{\mathrm{T}}}}\nu_m. \tag{4.9}$$

The effect of the perturbation l_m can be canceled by applying the modulo operation

$$\tilde{s}_m = \sqrt{\frac{\gamma}{P_{\mathrm{T}}}}y_m \bmod \delta = \left(s_m + \sqrt{\frac{\gamma}{P_{\mathrm{T}}}}\nu_m\right) \bmod \delta. \tag{4.10}$$

A decision operation then follows to quantize $\tilde{s}_m$ to the constellation $\mathcal{S}$. The effective noise power $\gamma(\sigma_\nu^2/P_{\mathrm{T}})$ is much smaller than that of the channel inversion method, as power normalization factor γ is small due to the minimization of the power of the unnormalized transmitting signal. It is shown in [83, 84] that CVP can achieve the full transmit diversity N over the i.i.d. Gaussian fading channel.

4.2.3 Per-antenna Power Constraint and p-Sphere Encoder

In practical multi-antenna system implementations, each antenna has its own analog frontend including D/A conversion and power amplifier. The analog frontend has its own linear operation region. If the transmitting signal input into the analog frontend has a very high power, the output signal will suffer from nonlinear amplification or even signal clipping. Thus, it would be desirable to constrain the per-antenna power in practice. Per-antenna power constraint (PAPC) is also motivated by distributed antenna systems where the antennas of a system are geographically separated and are powered by different power supplies. However, as we can see from (4.7) and (4.8), CVP does not have any control on the per-antenna power. As a result, CVP occasionally puts much of its total available power to a single antenna. In [46] the p-sphere encoder is proposed to reduce the per-antenna power. The p-sphere encoder, instead of minimizing the ℓ_2-norm of the unnormalized signal as done in CVP, minimizes the ℓ_p-norm ($p \geq 2$).

$$\begin{aligned} (p\text{-sphere encoder}) \quad \min_{\boldsymbol{d},\boldsymbol{l}} \quad & \|\boldsymbol{d}\|_p^2 \\ \text{s.t.} \quad & \boldsymbol{d} = \boldsymbol{H}^\dagger(\boldsymbol{s} + \delta\boldsymbol{l}), \\ & \boldsymbol{d} \in \mathbb{C}^N,\ \boldsymbol{l} \in \mathbb{G}^M. \end{aligned} \tag{4.11}$$

As a numerical result in [46], the p-sphere encoder can reduce the probability of a large per-antenna power. However, the p-sphere encoder does not guarantee a worst-case per-antenna power consumption, as it does not have an explicit control on the per-antenna power.

In this thesis, we consider the following explicit per-antenna power constraint

$$|\boldsymbol{x}|^2 \preceq \boldsymbol{\alpha} P_{\mathrm{T}}, \tag{4.12}$$

where $|\cdot|^2$ denote the element-wise square of absolute value, and $\boldsymbol{\alpha} \succeq \mathbf{0}$ is a coefficient that determines the fraction of maximum allowable per-antenna power to the total power P_{T}. This explicit PAPC can make sure that the per-antenna power budget is strictly satisfied. We assume that

$$\|\boldsymbol{\alpha}\|_1 \geq 1 \tag{4.13a}$$

$$\|\boldsymbol{\alpha}\|_\infty \leq 1 \tag{4.13b}$$

where (4.13a) ensures that the total power constraint (4.3) is not redundant and (4.13b) guarantees that not every per-antenna power constraints in (4.12) are redundant.

4.2.4 Power Normalization

Since the invention of CVP, there has been an undesirable assumption on CVP. In many existing works such as [13,85–89], the power normalization factor γ is chosen as

$$\gamma = \|\boldsymbol{d}\|_2^2. \tag{4.14}$$

This is known as instantaneous power normalization. With this choice of γ, the total power constraint is always met. However, the power normalization factor γ, which is essential for the receive signal detection, depends on the information vector $\boldsymbol{s}$. This means that the base station needs to broadcast γ to all users via other communication link other than CVP itself. This assumption may not hold true in practice.

A more practical assumption is the short-term power normalization wherein a block fading channel is considered. Suppose that the channel is static for a block of T time slots and in each time slot the unnormalized transmitting signal is given by $\boldsymbol{d}_t$. One possible choise of the power normalization factor γ is

$$\gamma = \max_{t=1,\ldots,T} \|\boldsymbol{d}_t\|_2^2. \tag{4.15}$$

The factor γ is then used to normalize all unnormalized transmitting signal $\boldsymbol{d}_t$ for $t = 1, \ldots, T$. As the whole block of transmission, only one normalization factor γ is required, it can be sent to all users at the beginning of each transmission block.

In the following section, we will propose vector perturbation with per-antenna power constraint (VP-PAPC) formulations based on the instantaneous power normalization. Then, in Section 4.4, we extend the VP-PAPC formulation to the short-term power normalization assumption.

4.3 Vector Perturbation with Per-antenna Power Constraint with Instantaneous Power Normalization

4.3.1 Problem Formulation

In this subsection, we will propose the vector perturbation formulation with per-antenna power constraint.

We consider the following form of transmitting signal

$$\boldsymbol{x} = \sqrt{\frac{P_{\mathrm{T}}}{\gamma}}\boldsymbol{d}, \tag{4.16a}$$

$$\boldsymbol{d} = \boldsymbol{H}^{\dagger}(\boldsymbol{s} + \delta\boldsymbol{l}) + \sigma\boldsymbol{H}_{\perp}\boldsymbol{u}, \tag{4.16b}$$

where $\boldsymbol{H}_{\perp} \in \mathbb{C}^{M\times(N-M)}$ is an orthogonal basis of the orthogonal complement of $\mathcal{R}(\boldsymbol{H}^{\dagger})$, $\sigma > 0$ a parameter, and $\boldsymbol{u} \in \mathbb{G}^{N-M}$ a variable to be determined. Compared with the CVP signal (4.7), we can see that the augmented CVP signal (4.16b) has an additional term $\sigma\boldsymbol{H}_{\perp}\boldsymbol{u}$. As we will see, this term is an important augmentation of the CVP signal and plays an important role in the VP-PAPC formulation.

As the power normalization factor γ is proportional to the effective noise power at the user, we try to minimize γ in the VP-PAPC formulation, specifically

$$\min_{\boldsymbol{d},\boldsymbol{l},\boldsymbol{u}} \quad \gamma \tag{4.17a}$$

$$\text{s.t.} \quad \gamma = \|\boldsymbol{d}\|_2^2, \tag{4.17b}$$

$$|\boldsymbol{x}|^2 \preceq \boldsymbol{\alpha} P_{\mathrm{T}}, \tag{4.17c}$$

$$\boldsymbol{x} = \sqrt{\frac{P_{\mathrm{T}}}{\gamma}}\boldsymbol{d}, \tag{4.17d}$$

$$\boldsymbol{d} = \boldsymbol{H}^{\dagger}(\boldsymbol{s} + \delta\boldsymbol{l}) + \sigma\boldsymbol{H}_{\perp}\boldsymbol{u}, \tag{4.17e}$$

$$\boldsymbol{d} \in \mathbb{C}^N,\ \boldsymbol{l} \in \mathbb{G}^M, \boldsymbol{u} \in \mathbb{G}^{N-M}, \tag{4.17f}$$

where (4.17b) is due to the instantaneous power normalization (4.14), (4.17c) is the per-antenna power constraint (4.12), and (4.17e) is the augmented CVP signal (4.16b).

Substituting (4.17b) into (4.17a) and (4.17d) into (4.17c), we can eliminate the variables γ and $\boldsymbol{x}$ in (4.17) and obtain the following equivalent problem is obtained.

$$\min_{\boldsymbol{d},\boldsymbol{l},\boldsymbol{u}} \quad \|\boldsymbol{d}\|_2^2 \tag{4.18a}$$

$$\text{s.t.} \quad \boldsymbol{d} = \boldsymbol{H}^{\dagger}(\boldsymbol{s} + \delta\boldsymbol{l}) + \sigma\boldsymbol{H}_{\perp}\boldsymbol{u}, \tag{4.18b}$$

$$|\boldsymbol{d}|^2 \preceq \boldsymbol{\alpha}\|\boldsymbol{d}\|_2^2, \tag{4.18c}$$

$$\boldsymbol{d} \in \mathbb{C}^N,\ \boldsymbol{l} \in \mathbb{G}^M, \boldsymbol{u} \in \mathbb{G}^{N-M}. \tag{4.18d}$$

Problem (4.18), which has an additional PAPC (4.18c) compared to CVP (4.8), is not a standard integer least square problem. The traditional sphere decoding (SD) algorithm [12, 16] could not be used to solve (4.18). Inspired by [46], we tailor-design an SD algorithm to solve (4.18); details can be found in Appendix 4.7.6.

The additional PAPC (4.18c) also raises an important question on the feasibility of (4.18). While CVP (4.8) must be feasible, it is not clear whether the VP-PAPC (4.18) is feasible or not. As we will see in the next section, the VP-PAPC problem is indeed feasible under some mild condition. We also show in the next subsection that VP-PAPC can achieve the same diversity order as CVP.

4.3.2 Feasibility and Diversity

In this subsection, we investigate the feasibility condition and the diversity order of VP-PAPC.

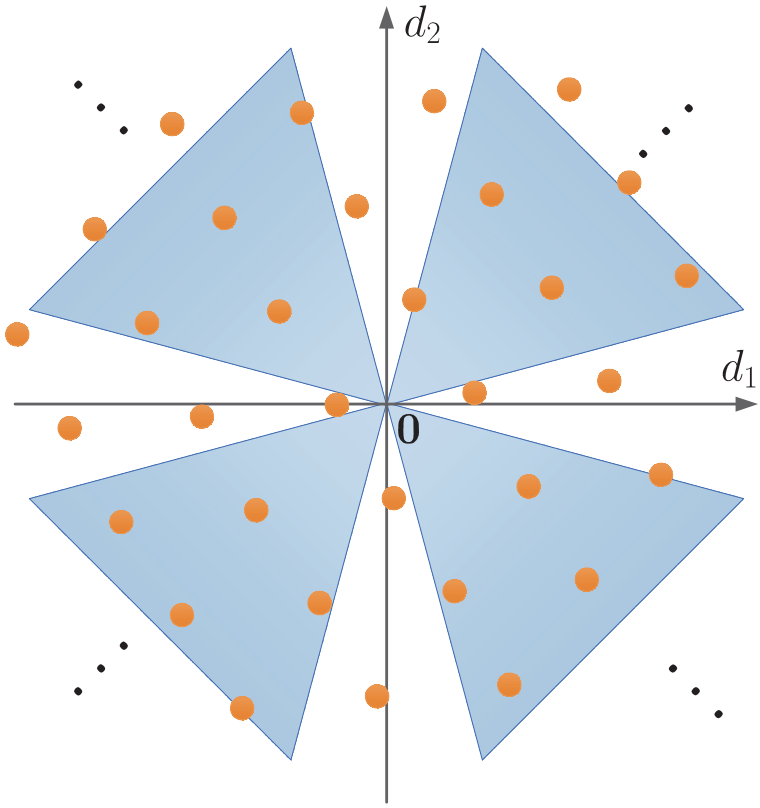


Figure 4.2: A geometric interpretation of the VP-PAPC problem

In order to provide intuition on the structure of the VP-PAPC problem (4.18), we illustrate in Fig. 4.2 an example of the feasible set of (4.18). The matrices $\boldsymbol{H}^\dagger$ and $\boldsymbol{H}_\perp$ belong to $\mathbb{C}^{2\times 1}$. For convenience of the following description, let us denote the translated lattice corresponding to (4.18b) by

$$\mathcal{A} = \left\{ \boldsymbol{d} = \boldsymbol{H}^\dagger \boldsymbol{s} + \delta \boldsymbol{H}^\dagger \boldsymbol{l} + \sigma \boldsymbol{H}_\perp \boldsymbol{u} \mid \boldsymbol{l} \in \mathbb{G}^M, \boldsymbol{u} \in \mathbb{G}^{N-M} \right\}. \tag{4.19}$$

and the PAPC set corresponding to (4.18c) by

$$\mathcal{V} = \{ \boldsymbol{d} \mid |\boldsymbol{d}|^2 \preceq \boldsymbol{\alpha} \|\boldsymbol{d}\|_2^2 \}. \tag{4.20}$$

Thus, the feasible solutions of (4.18) are given the by the intersection $\mathcal{A} \cap \mathcal{V}$. In Fig. 4.2, the blue region represents (part of) the set $\mathcal{V}$ and the orange dots

represents (part of) the translated lattice $\mathcal{A}$. Therefore, the feasible set is the set of orange points inside the blue region, and the one that is closet to the origin is the optimal solution of (4.18).

It can be seen from Fig. 4.2 that the translated lattice $\mathcal{A}$ lies in the whole space $\mathbb{C}^2$ and cannot be contained in any proper subspace, as the generator matrix $[\delta \boldsymbol{H}^\dagger, \sigma \boldsymbol{H}_\perp]$ of $\mathcal{A}$ is always of full rank. The PAPC set $\mathcal{V}$, which occupies a large part of the whole space $\mathbb{C}^M$, must contain some points of $\mathcal{A}$. This means that (4.18) is feasible. This intuition can be made rigorous and leads to the following proposition regarding the feasibility of the VP-PAPC problem (4.18).

Proposition 4.1 *Assume that* $\|\boldsymbol{\alpha}\|_1 > 1$. *Then, the VP-PAPC problem* (4.18) *is feasible for any realization of channel* $\boldsymbol{H}$ *and information symbol* $\boldsymbol{s}$.

The proof is relegated to Appendix 4.7.2. Proposition 4.1 shows that the VP-PAPC problem, which uses the augmented CVP signal (4.16b), is always feasible under the mild condition $\|\boldsymbol{\alpha}\|_1 > 1$. However, if the CVP signal (4.7) is used instead, the resulting formulation could be infeasible. To be precise, let us consider the following problem

$$\min_{\boldsymbol{d},\boldsymbol{l}} \quad \|\boldsymbol{d}\|_2^2 \tag{4.21a}$$

$$\text{s.t.} \quad \boldsymbol{d} = \boldsymbol{H}^\dagger(\boldsymbol{s} + \delta \boldsymbol{l}), \tag{4.21b}$$

$$|\boldsymbol{d}|^2 \preceq \boldsymbol{\alpha}\|\boldsymbol{d}\|_2^2, \tag{4.21c}$$

$$\boldsymbol{d} \in \mathbb{C}^N,\ \boldsymbol{l} \in \mathbb{G}^M, \tag{4.21d}$$

where the augmented CVP signal is replaced with CVP signal in (4.21b). We have the following proposition.

Proposition 4.2 *Suppose that the constellation* $\mathcal{S}$ *does not contain the origin. If there exists an index* $\mathcal{N} = \{n_1, \ldots, n_M\}$ *such that*

$$\mathcal{V}_\mathcal{N} = \{\boldsymbol{d} \mid |\boldsymbol{d}|^2 \preceq \boldsymbol{\alpha}\|\boldsymbol{d}\|_2^2,\ d_n = 0 \text{ for } n \notin \mathcal{N}\} \tag{4.22}$$

contains only the origin, then there exists an $\boldsymbol{H}^\dagger \in \mathbb{C}^{N\times M}$ *such that problem* (4.21) *is infeasible for all* $\boldsymbol{s} \in \mathbb{C}^M$.

The proof is relegated to Appendix 4.7.3. To provide some intuition of Proposition 4.2, we illustrate in Fig. 4.3 an example that (4.21) is infeasible. The channel inversion $\boldsymbol{H}^\dagger \in \mathbb{C}^{2\times 1}$ is tall matrix. As $\boldsymbol{H}^\dagger$ is tall, the translated lattice $\mathcal{A}_{\text{CVP}} = \left\{\boldsymbol{d} = \boldsymbol{H}^\dagger \boldsymbol{s} + \delta \boldsymbol{H}^\dagger \boldsymbol{l} \mid \boldsymbol{l} \in \mathbb{G}^M\right\}$ corresponding to (4.21b) lies in a line but not the whole space $\mathbb{C}^2$. Therefore, $\mathcal{A}_{\text{CVP}}$ does not intersect $\mathcal{V}$, which means that (4.21) is actually infeasible.

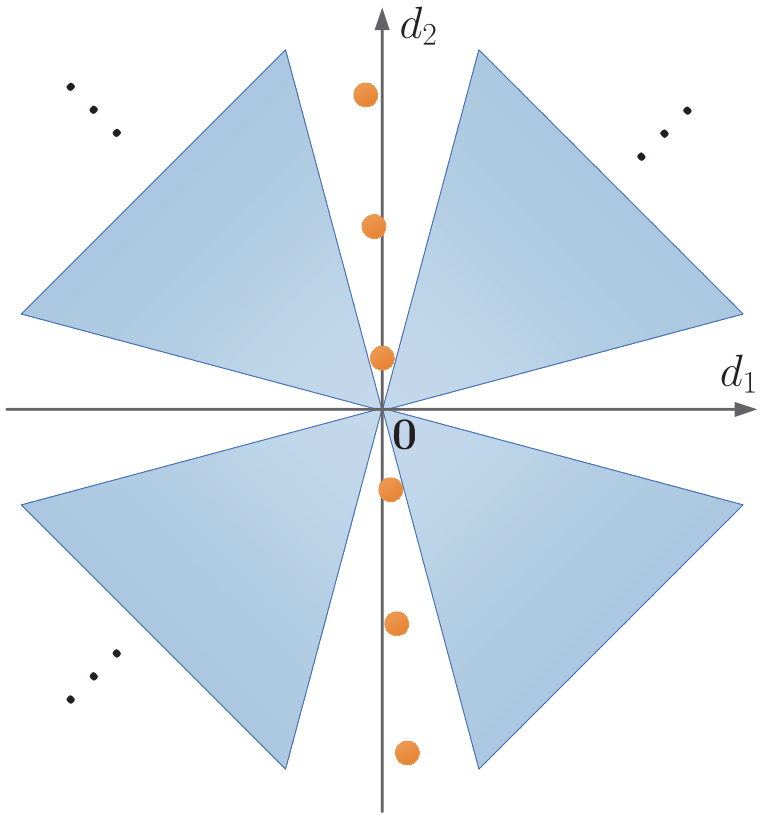


Figure 4.3: A geometric interpretation of CVP problem with an additional PAPC.

One implication of Proposition 4.2 is that in the case of equal PAPC coefficients $\alpha_1 = \ldots = \alpha_M = \alpha$, if

$$\alpha < 1/(N - M) \tag{4.23}$$

holds true, then the premise of Proposition 4.2 is satisfied and thus (4.21) can be infeasible. Therefore, for the CVP signal working with an additional PAPC, a stringent PAPC requirement (α is small) and the idea of using a large number of antenna to serve a few users ($N - M$ is small) are conflicting. In contrast, the augmented CVP signal always work with any numbers of antennas and users.

Next, we turn to the diversity aspect of VP-PAPC. The diversity of a precoding method is defined as

$$d = -\lim_{\mathrm{SNR}\to\infty} \frac{\log \Pr\{\hat{\boldsymbol{s}} \neq \boldsymbol{s}\}}{\log \mathrm{SNR}}, \tag{4.24}$$

where $\mathrm{SNR} = P_{\mathrm{T}}/\sigma_\nu^2$ and $\hat{\boldsymbol{s}}$ is the decision at the users. The diversity order d is the asymptotic slope of the SER curve in a log-log scale and describes how fast the SER decays. It is shown in [84] that under i.i.d. Gaussian fading channels CVP achieves a diversity order of $d_{\mathrm{CVP}} = N$, which is the highest possible diversity for an N-antenna base station. As VP-PAPC has a higher unnormalized transmitting signal and thus a higher power normalization factor γ than CVP due to the additional PAPC (4.18c), the VP-PAPC formulation could have a smaller diversity than CVP. Surprisingly, it can be shown that VP-PAPC achieves the same diversity N as CVP.

Proposition 4.3 *Suppose that* $\|\boldsymbol{\alpha}\|_1 > 1$ *and that each element of* $\boldsymbol{H}$ *follows an i.i.d.circular complex Gaussian distribution with unit variance and zero mean. Then, VP-PAPC under the instantaneous power normalization assumption achieves a diversity of* N*.*

The proof is inspired by [83, 84] is relegated to Appendix 4.7.4. As VP-PAPC and CVP have the same diversity, it would be expected that the SER of VP-PAPC would not degraded seriously compared to that of CVP. This will be demonstrated by numerical results in Section 4.5.

4.4 Vector Perturbation with Per-antenna Power Constraint with Short-term Power Normalization

4.4.1 Problem Formulation

In this subsection, we consider VP-PAPC under the short-term power normalization assumption. The channel $\boldsymbol{H}$ is assumed to be fixed within a block of T time slots. In the tth time slot, the base station transmits information

symbols $\boldsymbol{s}_t$ by the augmented CVP signal

$$\begin{aligned} \boldsymbol{d}_t &= \boldsymbol{H}^\dagger \boldsymbol{s}_t + \delta \boldsymbol{H}^\dagger \boldsymbol{l}_t + \sigma \boldsymbol{H}_\perp \boldsymbol{u}_t \\ \boldsymbol{x}_t &= \sqrt{\frac{P_{\mathrm{T}}}{\gamma}} \boldsymbol{d}_t \end{aligned} \tag{4.25}$$

for $t = 1, \ldots, T$. The total and per-antenna power constraint at each time slots are given by

$$\begin{aligned} \|\boldsymbol{x}_t\|_2^2 &\leq P_{\mathrm{T}}, \\ |\boldsymbol{x}_t|^2 &\preceq \boldsymbol{\alpha} P_{\mathrm{T}} \end{aligned} \tag{4.26}$$

for $t = 1, \ldots, T$. Then, the VP-PAPC problem under the short-term power normalization assumption can be formulated as

$$\begin{aligned} \min_{\{\boldsymbol{l}_t, \boldsymbol{d}_t, \boldsymbol{u}_t\}_{t=1}^T, \gamma} \quad & \gamma \\ \text{s.t.} \quad & \|\boldsymbol{x}_t\|_2^2 \leq P_{\mathrm{T}}, \\ & |\boldsymbol{x}_t|^2 \preceq \boldsymbol{\alpha} P_{\mathrm{T}}, \\ & \boldsymbol{x}_t = \sqrt{\frac{P_{\mathrm{T}}}{\gamma}} \boldsymbol{d}_t, \\ & \boldsymbol{d}_t = \boldsymbol{H}^\dagger \boldsymbol{s}_t + \delta \boldsymbol{H}^\dagger \boldsymbol{l}_t + \sigma \boldsymbol{H}_\perp \boldsymbol{u}_t, \\ & \boldsymbol{d}_t \in \mathbb{C}^N,\ \boldsymbol{l}_t \in \mathbb{G}^M,\ \boldsymbol{u}_t \in \mathbb{G}^{N-M},\ \ t = 1, \ldots, T. \end{aligned} \tag{4.27}$$

which can be simplified as

$$\begin{aligned} \min_{\{\boldsymbol{l}_t, \boldsymbol{d}_t, \boldsymbol{u}_t\}_{t=1}^T, \gamma} \quad & \gamma \\ \text{s.t.} \quad & \|\boldsymbol{d}_t\|_2^2 \leq \gamma, \\ & |\boldsymbol{d}_t|^2 \preceq \gamma \boldsymbol{\alpha}, \\ & \boldsymbol{d}_t = \boldsymbol{H}^\dagger \boldsymbol{s}_t + \delta \boldsymbol{H}^\dagger \boldsymbol{l}_t + \sigma \boldsymbol{H}_\perp \boldsymbol{u}_t, \\ & \boldsymbol{d}_t \in \mathbb{C}^N,\ \boldsymbol{l}_t \in \mathbb{G}^M,\ \boldsymbol{u}_t \in \mathbb{G}^{N-M},\ \ t = 1, \ldots, T. \end{aligned} \tag{4.28}$$

We can see that (4.28) is always feasible. Regarding the diversity, we have the following proposition.

Proposition 4.4 *Suppose that each element of* $\boldsymbol{H}$ *follows an i.i.d.circular complex Gaussian distribution with unit variance and zero mean. VP-PAPC under the short-term power normalization assumption achieves a diversity of* N*.*

The proof is relegated to Appendix 4.4.

4.4.2 Lagrangian Dual Relaxation Approximation

In this subsection, we develop efficient approximation to (4.28) by using the Lagrangian dual relaxation (LDR) proposed in Chapter 3.

Problem (4.28) involves seemingly a joint optimization of all perturbation variables within the whole block. However, it is actually separable. We can solve (4.28) by separately solving

$$\min_{\boldsymbol{l}_t,\boldsymbol{d}_t,\boldsymbol{u}_t,\gamma_t} \quad \gamma_t \tag{4.29a}$$

$$\text{s.t.} \quad \|\boldsymbol{d}_t\|_2^2 \leq \gamma_t, \tag{4.29b}$$

$$|\boldsymbol{d}_t|^2 \preceq \gamma_t \boldsymbol{\alpha}, \tag{4.29c}$$

$$\boldsymbol{d}_t = \boldsymbol{H}^\dagger \boldsymbol{s}_t + \delta \boldsymbol{H}^\dagger \boldsymbol{l}_t + \sigma \boldsymbol{H}_\perp \boldsymbol{u}_t, \tag{4.29d}$$

$$\boldsymbol{d}_t \in \mathbb{C}^N,\ \boldsymbol{l}_t \in \mathbb{G}^M,\ \boldsymbol{u}_t \in \mathbb{G}^{N-M}, \tag{4.29e}$$

for $t = 1, \ldots, T$. Let $\boldsymbol{l}_t^\star$, $\boldsymbol{d}_t^\star$, $\boldsymbol{u}_t^\star$, and $\gamma_t^\star$ denote an optimal solution of (4.29). Then, an optimal solution of (4.28) is given by $\{\boldsymbol{l}_t^\star, \boldsymbol{d}_t^\star, \boldsymbol{u}_t^\star\}_{t=1}^T$ and

$$\gamma^\star = \max_{t=1,\ldots,T} \gamma_t^\star.$$

Problem (4.29) can be solved exactly by the modified sphere encoder described in Appendix 4.7.6. However, as we will see in the simulations, the modified sphere encoder has a very high complexity due to the integer programming nature of (4.29). Here, we focus on deriving efficient approximate algorithm by applying the LDR approach. To proceed, let us denote the dual variables corresponding to (4.29b) and (4.29c) by $\omega_t \in \mathbb{R}_+$ and $\boldsymbol{\lambda}_t \in \mathbb{R}_+^N$. Then, a partial Lagrangian function of (4.29) can be written as

$$\begin{aligned} \mathcal{L} &= \gamma_t + \omega_t(\|\boldsymbol{d}_t\|_2^2 - \gamma_t) + \boldsymbol{\lambda}_t^T(|\boldsymbol{d}_t|^2 - \gamma_t \boldsymbol{\alpha}) \\ &= (1 - \omega_t - \boldsymbol{\lambda}^T \boldsymbol{\alpha})\gamma_t + \|\boldsymbol{D}(\sqrt{\omega_t \mathbf{1} + \boldsymbol{\lambda}_t})\boldsymbol{d}_t\|_2^2 \end{aligned} \tag{4.30}$$

The dual problem by definition is given by

$$\max_{\omega_t\geq 0,\boldsymbol{\lambda}_t\succeq\mathbf{0}}\left\{\begin{array}{ll}\min\limits_{\boldsymbol{l}_t,\boldsymbol{d}_t,\boldsymbol{u}_t} & (1-\omega_t-\boldsymbol{\lambda}_t^T\boldsymbol{\alpha})\gamma_t+\|\boldsymbol{D}(\sqrt{\omega_t\mathbf{1}+\boldsymbol{\lambda}_t})\boldsymbol{d}_t\|_2^2\\ \text{s.t.} & \boldsymbol{d}_t=\boldsymbol{H}^\dagger\boldsymbol{s}_t+\delta\boldsymbol{H}^\dagger\boldsymbol{l}_t+\sigma\boldsymbol{H}_\perp\boldsymbol{u}_t,\\ & \boldsymbol{d}_t\in\mathbb{C}^N,\ \boldsymbol{l}_t\in\mathbb{G}^M,\ \boldsymbol{u}_t\in\mathbb{G}^{N-M}.\end{array}\right\} \tag{4.31}$$

To prevent the inner minimization from being unbounded below, it is required that

$$1-\omega_t-\boldsymbol{\lambda}_t^T\boldsymbol{\alpha}=0.$$

Hence, (4.31) is equivalently written as

$$\begin{array}{ll}\max\limits_{\omega_t,\boldsymbol{\lambda}_t} & \varphi(\omega_t,\boldsymbol{\lambda}_t)\\ \text{s.t.} & \omega_t\geq 0,\boldsymbol{\lambda}_t\succeq\mathbf{0},\\ & 1-\omega_t-\boldsymbol{\lambda}_t^T\boldsymbol{\alpha}=0\end{array} \tag{4.32}$$

with the dual function

$$\begin{array}{ll}\varphi(\omega_t,\boldsymbol{\lambda}_t)=\min\limits_{\boldsymbol{l}_t,\boldsymbol{d}_t,\boldsymbol{u}_t} & \|\boldsymbol{D}(\sqrt{\omega_t\mathbf{1}+\boldsymbol{\lambda}_t})\boldsymbol{d}_t\|_2^2\\ \text{s.t.} & \boldsymbol{d}_t=\boldsymbol{H}^\dagger\boldsymbol{s}_t+\boldsymbol{H}^\dagger\boldsymbol{l}_t+\sigma\boldsymbol{H}_\perp\boldsymbol{u}_t,\\ & \boldsymbol{d}_t\in\mathbb{C}^N,\ \boldsymbol{l}_t\in\mathbb{G}^M,\ \boldsymbol{u}_t\in\mathbb{G}^{N-M}.\end{array} \tag{4.33}$$

The projected subgradient (PS) method is used to handle (4.32). For the iterate $(w_t^{(k)},\boldsymbol{\lambda}_t^{(k)})$ at the kth iteration of the PS method, the subgradient is given by

$$\boldsymbol{g}_t^{(k)}=\begin{bmatrix}g_{\omega_t}^{(k)}\\ \boldsymbol{g}_{\boldsymbol{\lambda}_t}^{(k)}\end{bmatrix}=\begin{bmatrix}\|\boldsymbol{d}_t^{(k)}\|_2^2\\ |\boldsymbol{d}_t^{(k)}|_2^2.\end{bmatrix} \tag{4.34}$$

where $\boldsymbol{d}_t^{(k)}$ denotes an optimal solution of (4.33) at $(w_t^{(k)},\boldsymbol{\lambda}_t^{(k)})$. The next iterate $(w_t^{(k+1)},\boldsymbol{\lambda}_t^{(k+1)})$ is updated according to

$$\begin{bmatrix}\omega_t^{(k+1)}\\ \boldsymbol{\lambda}_t^{(k+1)}\end{bmatrix}=\boldsymbol{P}_{\mathcal{F}}\left(\begin{bmatrix}\omega_t^{(k)}\\ \boldsymbol{\lambda}_t^{(k)}\end{bmatrix}+\beta_k\begin{bmatrix}g_{\omega_t}^{(k)}\\ \boldsymbol{g}_{\boldsymbol{\lambda}_t}^{(k)}\end{bmatrix}\right), \tag{4.35}$$

where β_k is the predefined step-size and $\boldsymbol{P}_{\mathcal{F}}(\cdot)$ denotes the projection onto the feasible set

$$\mathcal{F}=\{(\omega_t,\boldsymbol{\lambda}_t)\mid\omega_t\geq 0,\boldsymbol{\lambda}_t\succeq\mathbf{0},1-\omega_t-\boldsymbol{\lambda}_t^T\boldsymbol{\alpha}=0\}.$$

The computation of this projection operator $\boldsymbol{P}_{\mathcal{F}}$ is a water-filling-type problem and has a semi-closed-form solution; details can be found in Appendix 4.7.7. The dual function (4.33) is a standard integer least square problem which can be solved by a sphere decoder optimally or by the LRA-DF [29,30] approximately and efficiently. In this thesis, we adopt the LRA-DF solver for complexity reduction. The resulting precoder is named as *VP-PAPC LDR LRA-DF*.

4.5 Simulations

In this section, we use simulations to demonstrate the performance of the proposed VP-PAPC methods. The settings are described as follows. The channel matrix $\boldsymbol{H}$ follows an element-wise i.i.d. complex circular Gaussian distribution with zero mean and unit variance. The information symbols $\boldsymbol{s}$ are drawn in an i.i.d. fashion from the standard $(u+1)^2$-QAM constellation set, where u is an odd integer. Unless otherwise specified, we use 64-QAM constellation. The SNR is defined as $\text{SNR} = P_{\text{T}}/\sigma_\nu^2$.

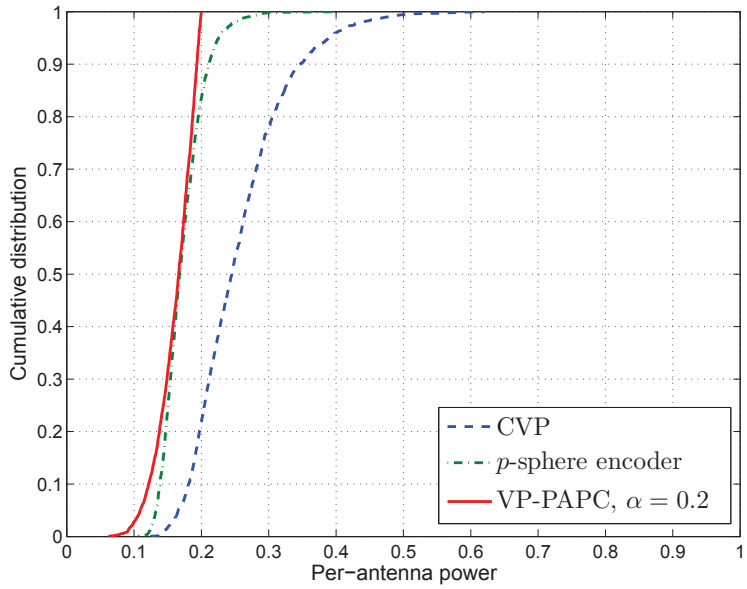
The benchmarked algorithms are the p-sphere encoder, CVP and its LRA-DF approximation [30]. For the $(u+1)^2$-QAM constellation, the CVP parameter are $\delta = 2(u+1)$ and $\sigma = \delta/10$. We choose $p = 10$ for the p-sphere encoder. In the short-term power normalization, the block length is $T = 100$. For the PS method for the VP-PAPC LDR LRA-DF method, we set the number of maximum iterations as $K_{\max} = 5$ and the step size as $\beta_k = 0.01/k$. In the case of instantaneous power normalization, the power normalization factors γ of all precoding methods are chosen according to (4.14). In the case of short-term power normalization, for all precoding methods γ is chosen as

$$\gamma = \max_{t=1,\ldots,T} \{\|\boldsymbol{d}_t\|_2^2, \|\boldsymbol{D}(\sqrt{\boldsymbol{\alpha}})^{-1}\boldsymbol{d}_t\|_\infty^2\} \tag{4.36}$$

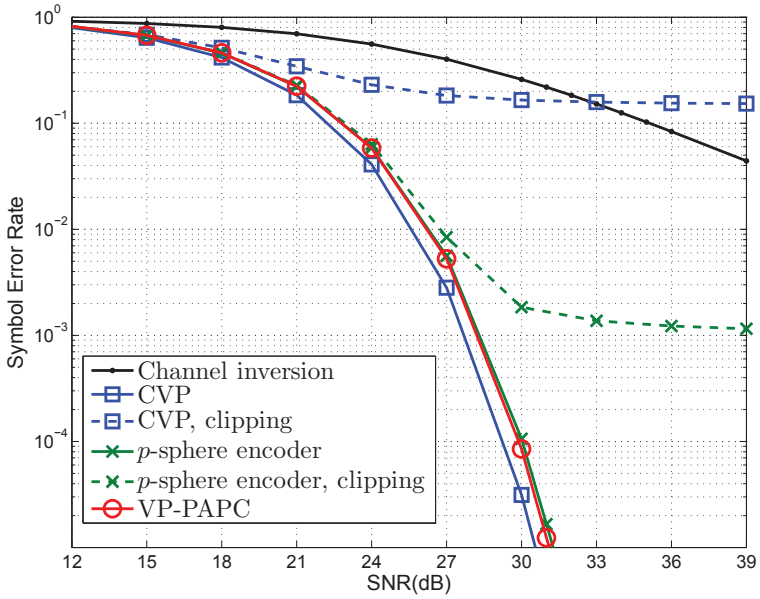
so that both the total and per-antenna power constraints are met. Comparing to the choice of the power normalization factor γ in (4.15), the one in (4.36) introduces additional power back-off to avoid signal clipping. We will consider VP-PAPC under the instantaneous power normalization assumption in the first subsection and then short-term normalization assumption in the second subsec-

tion.

4.5.1 Instantaneous Power Normalization



(a) Per-antenna power distribution



(b) Symbol error rate

Figure 4.4: Performance comparison under the instantaneous power normalization assumption. $(M, N) = (12, 12)$. $\alpha_n = 0.2$ for $n = 1, \ldots, N$.

In Fig. 4.4, we present the cumulative distribution of the per-antenna power

and SER performance of various precoding methods. The number of users and transmitting antennas are $(M, N) = (12, 12)$. The PAPC coefficients α_n are the same and are equal to 0.2. Fig. 4.4(a) shows the distribution of the power at the first transmitting antenna normalized by the total power, i.e. $|x_1|^2/P_{\mathrm{T}}$. Note that the results in this figure also represent the distribution of all other antennas, as the distributions of transmitting signals at all antennas are the same. It can be seen that CVP occasionally puts much power into a single antenna. In extreme cases, more than 50% of the total power is consumed by one antenna. In addition, CVP violates the PAPC $\alpha = 0.2$ with probability around 80%. The p-sphere encoder has a lower per-antenna power than CVP. However, it still violates the PAPC with probability around 20%. The VP-PAPC method, which has an explicit PAPC, never puts more than 20% of the total power to one antenna. This results demonstrates the effectiveness of VP-PAPC in reducing the per-antenna power of CVP.

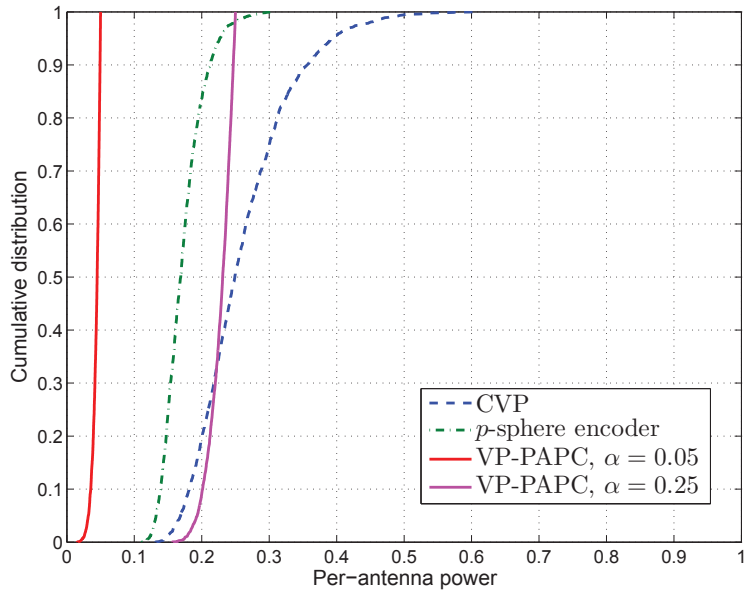
In order to investigate the impact of signal clipping to the SER, we introduce the following clipping function

$$\mathrm{clipping}_{\eta}(ae^{\boldsymbol{\jmath}\theta}) = \begin{cases} ae^{\boldsymbol{\jmath}\theta}, & \text{if } |a| \leq \eta, \\ \eta e^{\boldsymbol{\jmath}\theta}, & \text{if } |a| > \eta. \end{cases}$$

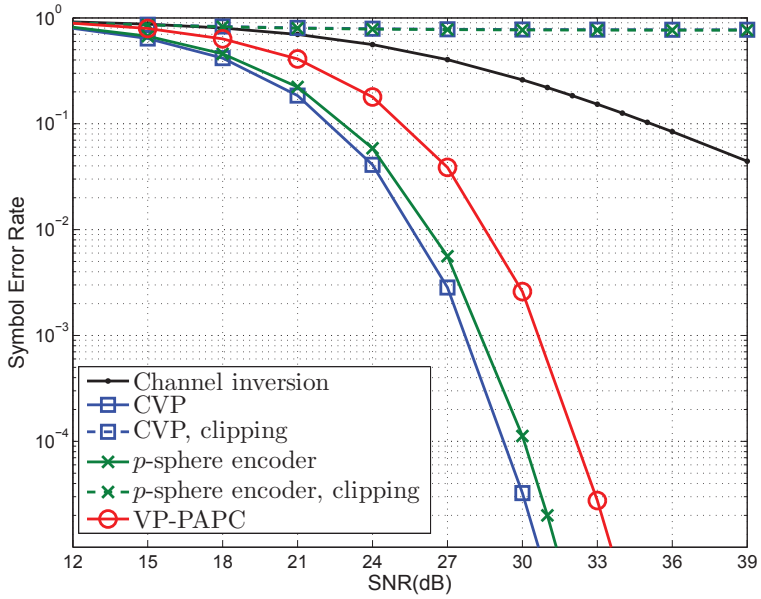
The transmitting signal $\tilde{x}_i$ is given by $\tilde{x}_n = \mathrm{clipping}_{\sqrt{\alpha_n P_{\mathrm{T}}}}(x_n)$ for $n = 1, \ldots, N$, where x_n is original transmitting signal of CVP, p-sphere or VP-PAPC. This means that when the transmitting signal has a large power, the signal is clipped to a maximum power $\alpha_n P_{\mathrm{T}}$ without affecting the signal phase.

Fig. 4.4(b) shows the SER result of VP-PAPC. It can be seen that when signal clipping is not applied, CVP, p-sphere encoder and VP-PAPC have very similar SER performances; CVP is 0.5dB better than p-sphere encoder and VP-PAPC. However, when signal clipping is present, both CVP and p-sphere encoder exhibits error floors. The error floors of CVP and p-sphere encoder are 10^{-1} and 10^{-3}, respectively. As VP-PAPC never violates the strict PAPC, signal clipping has no impact on VP-PAPC. The results in this figure also confirm that VP-PAPC achieves the same diversity as CVP.

Fig. 4.5 shows the results of the unequal PAPC coefficients case. The settings



(a) Per-antenna power distribution



(b) Symbol error rate

Figure 4.5: Performance comparison under the instantaneous power normalization assumption. $(M, N) = (12, 12)$. $\alpha_n = 0.05$ for $n = 1, \ldots, 6$ and $\alpha_n = 0.25$ for $n = 7, \ldots, 12$.

except the PAPC coefficient $\boldsymbol{\alpha}$ are the same as those of Fig. 4.4; the coefficient $\boldsymbol{\alpha}$ is chosen as $\alpha_n = 0.05$ for $n = 1, \ldots, 6$ and $\alpha_n = 0.25$ for $n = 7, \ldots, 12$, i.e. the first half of the transmitting antennas has a PAPC coefficient 0.05 and

the second half 0.25. Fig. 4.5(a) shows the distributions of the powers at the first and seventh transmitting antennas of VP-PAPC which have PAPC of α_n equal to 0.05 and 0.25, respectively. It can be seen that VP-PAPC can strictly meet this PAPC requirement. Again, both CVP and p-sphere encoder frequently violate the PAPC requirement. In fact, for the first six antennas which has a very stringent PAPC requirement of $\alpha_n = 0.05$, CVP and p-sphere encoder never satisfy the PAPC in all simulated trials. This serious violation of PAPC would lead to heavy signal clipping, as confirmed by the SER result in Fig. 4.5. It can be seen from Fig. 4.5 that when clipping is present, CVP and p-sphere encoder do not work at all. In contrast, VP-PAPC only has a 3dB loss compared to CVP without clipping.

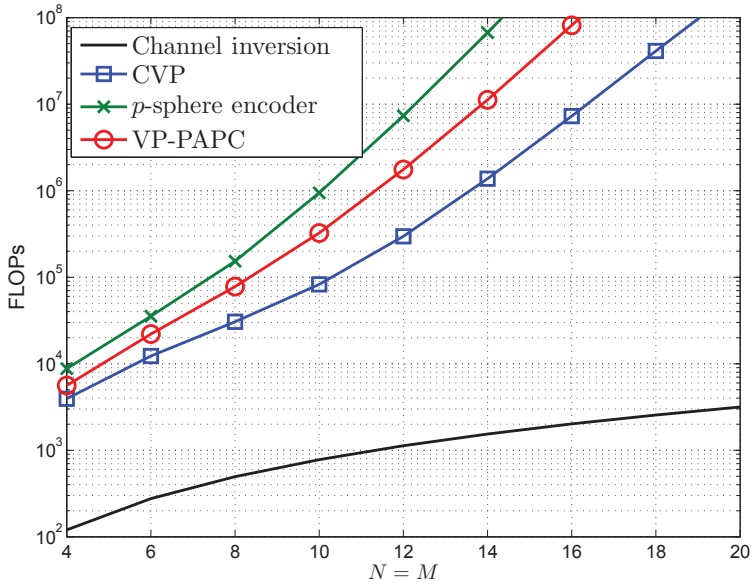


Figure 4.6: Average number of floating point operations (FLOPs) under instantaneous power normalization. $\alpha_n = 2/N$ for $n = 1, \ldots, N$.

We investigate the complexity of solving the VP-PAPC problem in Fig. 4.6 where the settings are $M = N$ and all PAPC coefficients are the same and are equal to $2/N$. It can be seen that VP-PAPC can be much faster than the p-sphere encoder. The computational advantage of VP-PAPC is more significant at larger problem sizes. It can be 10 times faster than the p-sphere encoder at problem size $(N, M) = (14, 14)$. CVP is less computationally demanding than

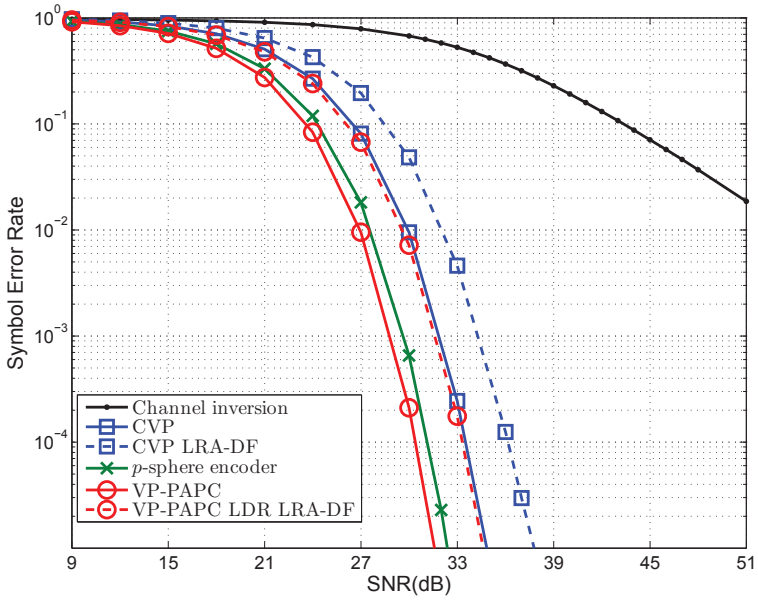
both p-sphere encoder and VP-PAPC.

From all the observations above, we conclude that under the instantaneous power normalization assumption, VP-PAPC can strictly satisfy the PAPC and thus avoid signal clipping. The SER performance of VP-PAPC can be closed to CVP without clipping. The complexity of VP-PAPC can be significantly lower than that of p-sphere encoder.

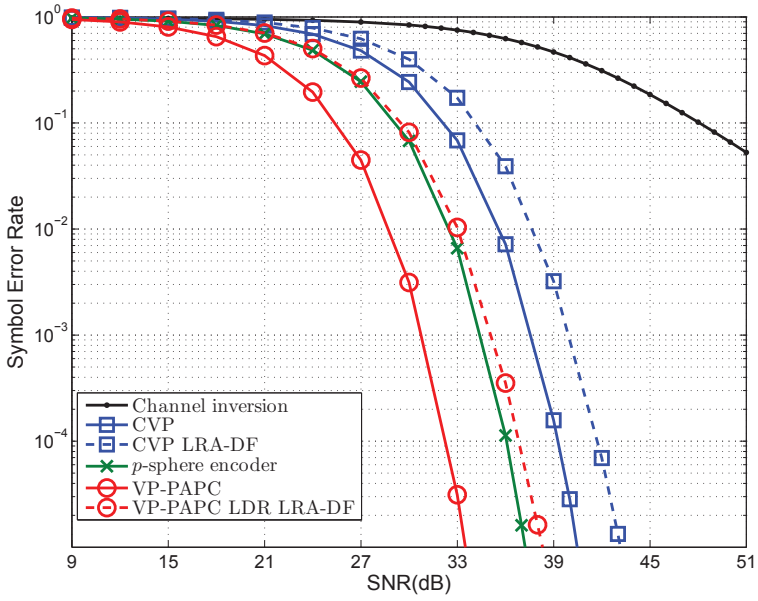
4.5.2 Short-term Power Normalization

Fig. 4.7 shows the SER performance of VP-PAPC and its LDR LRA-DF approximation under the short-term power normalization assumption. Fig. 4.7(a) shows the case that all PAPC coefficients α_n are equal to 0.2. The settings are the same as those in Fig. 4.4. Note that CVP and p-sphere encoder will not suffer from signal clipping due to the choice of power normalization factor γ according to (4.36). But the power back-off due to (4.36) leads to some performance degradation of p-sphere encoder and CVP. One can observe that p-sphere encoder and CVP are worse than VP-PAPC by 1dB and 3dB, respectively. The LRA-DF approximation versions of CVP and VP-PAPC have some performance loss compared to their exact counterparts; the losses are 3dB. VP-PAPC LDR LRA-DF is 3dB better than CVP LRA-DF. Note that there is no efficient approximation of the p-sphere encoder in the literature. In Fig. 4.7(b) we change the PAPC coefficients to $\alpha_n = 0.05$ for $n = 1, \ldots, 6$ and $\alpha_n = 0.25$ for $n = 7, \ldots, 12$. It can be observed that the performance advantages of VP-PAPC over p-sphere encoder and CVP are 5dB and 7dB respectively, which are greater than those in the equal PAPC coefficients case. Moreover, VP-PAPC LDR LRA-DF has a 5dB gain compared to CVP LRA-DF.

In Fig. 4.8, we show the result in a large problem size $(M, N) = (40, 40)$, where exact CVP, p-sphere encoder, and VP-PAPC are too computationally demanding. In Fig. 4.8(a) the all PAPC coefficients α_n are equal to 0.1 and in Fig. 4.8(b) the first half of PAPC coefficients are 0.01 and the remaining half are 0.05. The observations are similar to those in Fig. 4.8. The performance gap between VP-PAPC LDR LRA-DF and CVP LRA-DF is 5dB in unequal PAPC



(a)

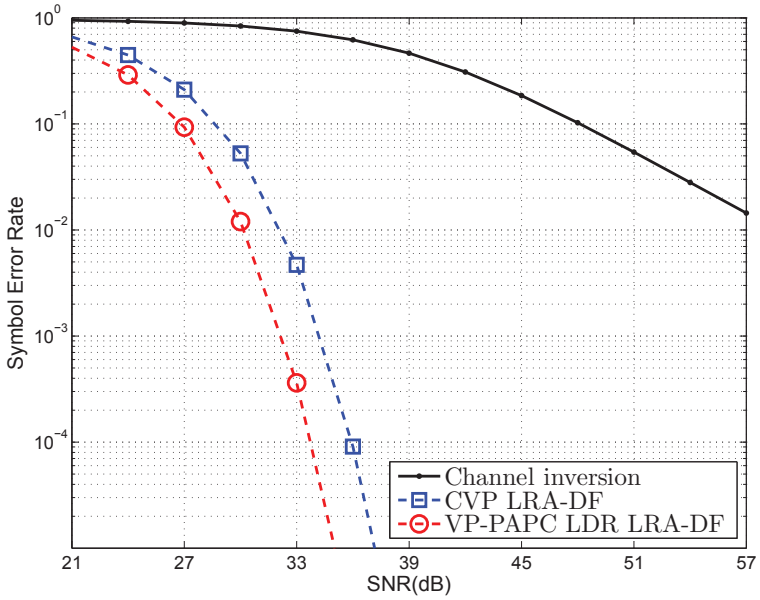


(b)

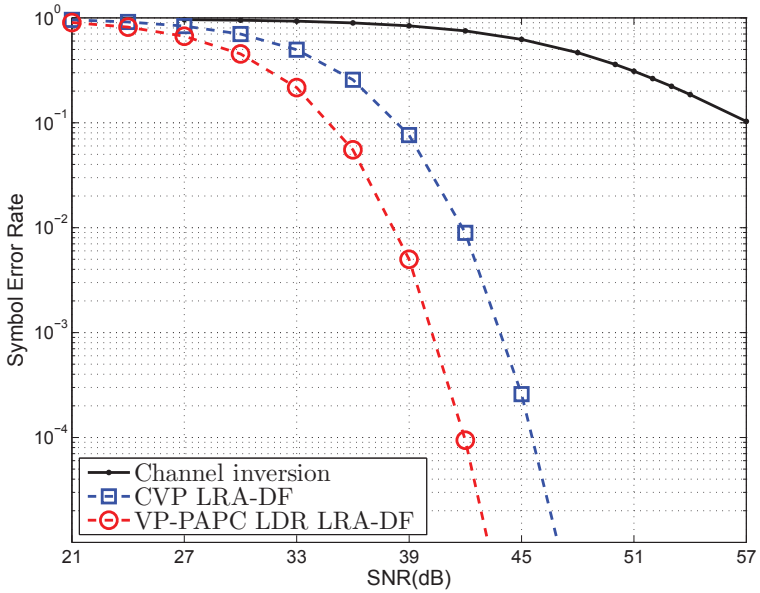
Figure 4.7: Symbol error rate performance under the short-term power normalization assumption. $(M, N) = (12, 12)$. (a) $\alpha_n = 0.2$ for $n = 1, \ldots, 12$. (b) $\alpha_n = 0.05$ for $n = 1, \ldots, 6$ and $\alpha_n = 0.25$ for $n = 7, \ldots, 12$.

coefficients case, which is greater than the 3dB gap in equal PAPC coefficients case.

Fig. 4.9 presents the the average number of FLOPs of various precoding



(a)



(b)

Figure 4.8: Symbol error rate performance under the short-term power normalization assumption. $(M, N) = (40, 40)$. (a) $\alpha_n = 0.1$ for $n = 1, \ldots, 40$. (b) $\alpha_n = 0.01$ for $n = 1, \ldots, 20$ and $\alpha_n = 0.05$ for $n = 21, \ldots, 40$.

methods. It can be seen that compared with the exact CVP and VP-PAPC, CVP LRA-DF and VP-PAPC LDR LRA-DF have a very low complexity and thus can be used in large antenna array systems. VP-PAPC LDR LRA-DF is

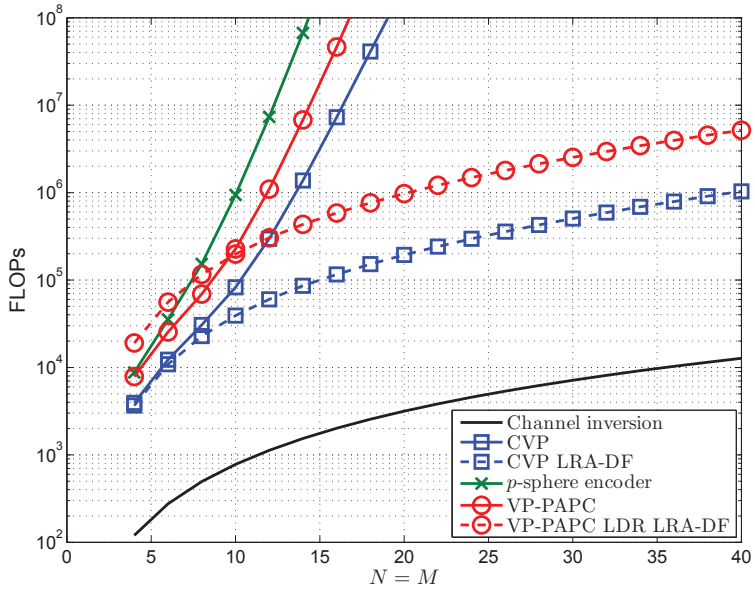


Figure 4.9: Average number of (FLOPs) under short-term power normalization. $\alpha_n = 2/N$ for $n = 1, \ldots, N$.

slower than CVP LRA-DF due to the iteration nature of the LDR approach.

From the observations above, we conclude that under the short-term power normalization assumption, VP-PAPC outperforms both CVP and p-sphere encoder. CVP LRA-DF and VP-PAPC LDR LRA-DF are much more efficient than the exact CVP, p-sphere encoder and VP-PAPC. VP-PAPC LDR LRA-DF can be better than CVP LRA-DF by more than 3dBs.

4.6 Summary

In this chapter, we considered the per-antenna power constrained vector perturbation under the instantaneous and short-term power normalization assumptions. We show that the formulated VP-PAPC can achieve full transmit diversity under both assumptions. A modified sphere encoder is used to solve the VP-PAPC problems. We developed a fast approximation algorithm for the short-term VP-PAPC by using the LDR and LRA-DF techniques. Simulation results reveal that the performances of VP-PAPC are very promising.

4.7 Appendix

4.7.1 Lemma 4.1

Let $\mathcal{B}_r(\boldsymbol{d}_o)$ denotes a closed ball with center $\boldsymbol{d}_o$ and radius r, i.e. $\mathcal{B}_r(\boldsymbol{d}_o) = \{\boldsymbol{d} \in \mathbb{C}^N \mid \|\boldsymbol{d} - \boldsymbol{d}_o\|_2 \leq r\}$. Then, we have the following lemma.

Lemma 4.1 *Suppose that $\|\boldsymbol{\alpha}\|_1 > 1$ and $r > 0$. There exists a vector $\boldsymbol{d}_o$ such that $\mathcal{B}_r(\boldsymbol{d}_o) \subset \mathcal{V} \cap \mathcal{B}_{cr}(\mathbf{0})$, where c is a constant depending on $\boldsymbol{\alpha}$ only.*

Let us consider the vector $\boldsymbol{d}_o$ in the form of $\boldsymbol{d}_o = t\sqrt{\boldsymbol{\alpha}}$ with $t \geq 0$, where $\sqrt{\boldsymbol{\alpha}}$ denotes the element-wise square root. Consider the following feasibility problem

$$\begin{aligned} \text{find} \quad & t \\ \text{s.t.} \quad & \mathcal{B}_r(t\sqrt{\boldsymbol{\alpha}}) \subset \mathcal{V}, \\ & t \geq 0. \end{aligned} \tag{4.37}$$

If we can find a feasible $t^\star$ in the form of $t^\star = c'r$ for some constant c' depending on $\boldsymbol{\alpha}$ only, then

$$\mathcal{B}_r(\boldsymbol{d}_0) \subset \mathcal{V}$$

with $\boldsymbol{d}_0 = c'r\sqrt{\boldsymbol{\alpha}}$. In addition, it is always true that

$$\mathcal{B}_r(\boldsymbol{d}_0) = \mathcal{B}_r(c'r\sqrt{\boldsymbol{\alpha}}) \subset \mathcal{B}_{cr}(\mathbf{0})$$

with $c = c'\|\sqrt{\boldsymbol{\alpha}}\|_2 + 1$. Hence, we have

$$\mathcal{B}_r(\boldsymbol{d}_o) \subset \mathcal{V} \cap \mathcal{B}_{cr}(\mathbf{0})$$

with c depending on N and $\boldsymbol{\alpha}$ only, and thus complete the proof.

In the remaining of the proof, our goal is to find a $t^\star = c'r$ that is feasible to (4.37). Note that we can arbitrarily tighten (4.37) as long as the resulting problem still has a feasible solution $t^\star = c'r$. We first tighten (4.37) as follows

$$\begin{aligned} \text{find} \quad & t \\ \text{s.t.} \quad & \mathcal{B}_r(t\sqrt{\boldsymbol{\alpha}}) \subset \mathcal{V} \\ & t \geq r/\|\sqrt{\boldsymbol{\alpha}}\|_2. \end{aligned} \tag{4.38}$$

To further tighten (4.38), let us rewrite rewrite $\mathcal{V}$ as

$$\begin{aligned}\mathcal{V} &= \{\boldsymbol{d} \in \mathbb{C}^N \mid |\boldsymbol{d}|^2 \preceq \boldsymbol{\alpha}\|\boldsymbol{d}\|_2^2\} \\ &= \{\boldsymbol{d} \in \mathbb{C}^N \mid 0 \geq \|\boldsymbol{D}^{-1}(\sqrt{\boldsymbol{\alpha}})\boldsymbol{d}\|_\infty^2 - \|\boldsymbol{d}\|_2^2\}.\end{aligned}$$

Then the first constraint of (4.38) is the same as

$$0 \geq \left(\begin{array}{ll} \max\limits_{\boldsymbol{d}\in\mathbb{C}^N} & \|\boldsymbol{D}^{-1}(\sqrt{\boldsymbol{\alpha}})\boldsymbol{d}\|_\infty^2 - \|\boldsymbol{d}\|_2^2 \\ \text{s.t.} & \|\boldsymbol{d} - t\sqrt{\boldsymbol{\alpha}}\|_2^2 \leq r^2. \end{array} \right),$$

which can be tightened by

$$0 \geq \left(\begin{array}{ll} \max\limits_{\boldsymbol{d}\in\mathbb{C}^N} & \|\boldsymbol{D}^{-1}(\sqrt{\boldsymbol{\alpha}})\boldsymbol{d}\|_\infty^2 \\ \text{s.t.} & \|\boldsymbol{d} - t\sqrt{\boldsymbol{\alpha}}\|_2^2 \leq r^2 \end{array} \right) - \left(\begin{array}{ll} \min\limits_{\boldsymbol{d}\in\mathbb{C}^N} & \|\boldsymbol{d}\|_2^2 \\ \text{s.t.} & \|\boldsymbol{d} - t\sqrt{\boldsymbol{\alpha}}\|_2^2 \leq r^2 \end{array} \right). \tag{4.39}$$

Let us derive a closed-form solution of this constraint. After changing the variable $\boldsymbol{z} = \boldsymbol{d} - t\sqrt{\boldsymbol{\alpha}}$, the first optimization problem of (4.39) becomes

$$\begin{array}{ll} \max\limits_{\boldsymbol{z}\in\mathbb{C}^N} & \|t\boldsymbol{1} + \boldsymbol{D}^{-1}(\sqrt{\boldsymbol{\alpha}})\boldsymbol{z}\|_\infty^2 \\ \text{s.t.} & \|\boldsymbol{z}\|_2^2 \leq r^2. \end{array}$$

By noting that t is positive, the optimal objective value is given by

$$\left(t + \frac{r}{\sqrt{\alpha_{\min}}}\right)^2$$

where $\alpha_{\min} = \min_{n=1,\ldots,N} \alpha_n$. Similarly, the second optimization problem of (4.39) is rewritten as

$$\begin{array}{ll} \min\limits_{\boldsymbol{z}\in\mathbb{C}^N} & \|\boldsymbol{z} + t\sqrt{\boldsymbol{\alpha}}\|_2^2 \\ \text{s.t} & \|\boldsymbol{z}\|_2^2 \leq r^2. \end{array}$$

By noting that $t \geq r/\|\sqrt{\boldsymbol{\alpha}}\|_2$, we obtain the optimal objective value

$$(t\|\sqrt{\boldsymbol{\alpha}}\|_2 - r)^2.$$

Thus, (4.39) is the same as

$$\begin{aligned}0 \geq& \left(t + \frac{r}{\sqrt{\alpha_{\min}}}\right)^2 - (t\|\sqrt{\boldsymbol{\alpha}}\|_2 - r)^2 \\ =&(1 - \|\sqrt{\boldsymbol{\alpha}}\|_2^2)t^2 + 2(1/\sqrt{\alpha_{\min}} + \|\sqrt{\boldsymbol{\alpha}}\|_2)rt + (1/\alpha_{\min} - 1)r^2.\end{aligned}$$

For convenience, let

$$\begin{aligned} a &= 1 - \|\sqrt{\boldsymbol{\alpha}}\|_2^2 \\ b &= 2(1/\sqrt{\alpha_{\min}} + \|\sqrt{\boldsymbol{\alpha}}\|_2)r \\ d &= (1/\alpha_{\min} - 1)r^2 \end{aligned}$$

denote the coefficients of the quadratic, linear and constant terms, respectively. Then, problem (4.38) is tightened by

$$\begin{aligned} \text{find} \quad & t \\ \text{s.t.} \quad & at^2 + bt + d \leq 0, \\ & r/\|\sqrt{\boldsymbol{\alpha}}\|_2 \leq t. \end{aligned} \tag{4.40}$$

By the assumption that $\|\boldsymbol{\alpha}\|_1 > 1$, we have $a < 0$. By noting that $a < 0$, $b > 0$ and $d > 0$, the smallest feasible solution is

$$t^\star = \max\left\{ r/\|\sqrt{\boldsymbol{\alpha}}\|_2, \sqrt{\frac{b^2 - 4ad}{4a^2}} - \frac{b}{2a} \right\}. \tag{4.41}$$

Substituting a, b and d into (4.41), $t^\star$ can be written as $t^\star = c'r$, where c' is given by $c' = \max\left\{1/\|\sqrt{\boldsymbol{\alpha}}\|_2, \sqrt{\frac{(1/\alpha_{\min}+\|\sqrt{\boldsymbol{\alpha}}\|_2)^2-(1-\|\sqrt{\boldsymbol{\alpha}}\|_2^2)(1/\alpha_{\min}-1)}{(1-\|\sqrt{\boldsymbol{\alpha}}\|_2^2)^2}} - \left(\frac{1/\sqrt{\alpha_{\min}}+\|\sqrt{\boldsymbol{\alpha}}\|_2}{1-\|\sqrt{\boldsymbol{\alpha}}\|_2^2}\right)\right\}$. It can be seen that c' depends on $\boldsymbol{\alpha}$ only. Thus, we complete the proof.

4.7.2 Proof of Proposition 4.1

Let us denote $\boldsymbol{G} = [\delta\boldsymbol{H}^\dagger, \sigma\boldsymbol{H}_\perp]$ and $\boldsymbol{z} = [\boldsymbol{l}^T, \boldsymbol{u}^T]^T$. The covering radius of a lattice $\mathcal{L}(\boldsymbol{G}) = \{\boldsymbol{G}\boldsymbol{z} \mid \boldsymbol{z} \in \mathbb{G}^N\}$ is defined as

$$\zeta(\boldsymbol{G}) = \max_{\boldsymbol{a}\in\mathbb{C}^N} \min_{z\in\mathbb{G}^N} \|\boldsymbol{G}(\boldsymbol{a} - \boldsymbol{z})\|_2. \tag{4.42}$$

From the definition of $\zeta(\boldsymbol{G})$, it can be seen that $\zeta(\boldsymbol{G})$ is the maximum distance from any point in the subspace $\mathcal{R}(\boldsymbol{G})$ to the lattice $\mathcal{L}(\boldsymbol{G})$. As $\boldsymbol{G}$ is square and of full rank, $\mathcal{R}(\boldsymbol{G})$ is the same as the entire space $\mathbb{R}^N$. Thus, $\zeta(\boldsymbol{G})$ is simply the distance from any point in the space to $\mathcal{L}(\boldsymbol{G})$. Then, for any point $\boldsymbol{d}_0' \in \mathbb{R}^N$, the ball $\mathcal{B}_{\zeta(\boldsymbol{G})}(\boldsymbol{d}_0')$ must contain at least one point of the lattice $\mathcal{L}(\boldsymbol{G})$.

On the other hand, by Lemma 4.1 there exists a point $\boldsymbol{d}_o$ such that

$$\mathcal{B}_{\zeta(\boldsymbol{G})}(\boldsymbol{d}_o) \subset \mathcal{V},$$

because the premise $\|\boldsymbol{\alpha}\|_1 > 1$ is satisfied. Then, by setting $\boldsymbol{d}_o' = \boldsymbol{d}_0 - \boldsymbol{H}^\dagger \boldsymbol{s}$, we known that $\mathcal{B}_{\zeta(\boldsymbol{G})}(\boldsymbol{d}_0 - \boldsymbol{H}^\dagger \boldsymbol{s})$ contains at least one point of $\mathcal{L}(\boldsymbol{G})$. Let us denote this point by $\boldsymbol{G}\boldsymbol{z}_0$ with $\boldsymbol{z}_0 = [\boldsymbol{s}_0^T, \boldsymbol{u}_0^T]^T \in \mathbb{G}^N$. Then we have

$$\boldsymbol{G}\boldsymbol{z}_0 + \boldsymbol{H}^\dagger \boldsymbol{s} \in \mathcal{B}_{\zeta(\boldsymbol{G})}(\boldsymbol{d}_0) \subset \mathcal{V},$$

or equivalently

$$\boldsymbol{H}^\dagger \boldsymbol{s} + \delta \boldsymbol{H}^\dagger \boldsymbol{l}_0 + \sigma \boldsymbol{H}_\perp \boldsymbol{u}_0 \in \mathcal{V}.$$

Thereby, we complete the proof.

4.7.3 Proof of Proposition 4.2

Set $\boldsymbol{H}^\dagger$ as $\boldsymbol{H}^\dagger = [\boldsymbol{e}_{n_1}, \ldots, \boldsymbol{e}_{n_M}]$, where $\boldsymbol{e}_{n_m}$ denotes the n_mth unit vector. Then it follows that

$$\begin{aligned} &\mathcal{R}(\boldsymbol{H}^\dagger) \cap \mathcal{V} \\ =&\{\boldsymbol{d} \mid |\boldsymbol{d}|^2 \preceq \boldsymbol{\alpha}\|\boldsymbol{d}\|_2^2,\ \boldsymbol{d} = \boldsymbol{H}^\dagger \boldsymbol{z}, \boldsymbol{z} \in \mathbb{C}^M\} \\ \subset&\mathcal{V}_{\mathcal{N}} \\ =&\{\boldsymbol{0}\}. \end{aligned} \tag{4.43}$$

Noting that $\mathcal{A}_{\mathrm{CVP}}$ belongs to $\mathcal{R}(\boldsymbol{H}^\dagger)$, we have

$$\mathcal{A}_{\mathrm{CVP}} \cap \mathcal{V} \subset \{\boldsymbol{0}\}. \tag{4.44}$$

As the parameter δ of CVP is chosen such that $\mathrm{conv}(\mathcal{S} + \delta \boldsymbol{l})$ is non-overlapping for every $\boldsymbol{l} \in \mathbb{G}^N$, we have that $\boldsymbol{s} + \delta \boldsymbol{l}$ is nonzero for any $\boldsymbol{s} \in \mathcal{S}^M$ and $\boldsymbol{l} \in \mathbb{G}^M$. Noting that $\boldsymbol{H}^\dagger$ is of full column-rank, it follows that the origin does not belong to $\mathcal{A}_{\mathrm{CVP}}$. Thus, we have the desired result that

$$\mathcal{A}_{\mathrm{CVP}} \cap \mathcal{V} = \emptyset. \tag{4.45}$$

4.7.4 Proof of Proposition 4.3

The proof is based on [84, 90]. By Lemma 4.1, it can be seen that the point $\boldsymbol{z}_0 = [\boldsymbol{l}_0, \boldsymbol{u}_0]$ in (4.12) in the proof of Proposition 4.1 can be chosen in a way such that $\boldsymbol{d}_0 = \boldsymbol{H}^\dagger \boldsymbol{s} + \boldsymbol{G}\boldsymbol{z}_0$ satisfies

$$\begin{aligned} &|\boldsymbol{d}_0|^2 \preceq \boldsymbol{\alpha}\|\boldsymbol{d}_0\|_2^2 \\ &\|\boldsymbol{d}_0\|_2^2 \leq c^2 \zeta(\boldsymbol{G}). \end{aligned} \tag{4.46}$$

where $\boldsymbol{G} = [\delta \boldsymbol{H}^\dagger, \sigma \boldsymbol{H}_\perp]$ and $\zeta(\boldsymbol{G})$ is the covering radius of the lattice $\mathcal{L}(\boldsymbol{G})$. Let Γ denote the objective value of the VP-PAPC problem (4.18). Then we have

$$\Gamma \leq \|\boldsymbol{d}_0\|_2^2 \leq c^2 \zeta^2(\boldsymbol{G}). \tag{4.47}$$

Let us compute $\zeta^2(\boldsymbol{G})$ as follows.

$$\begin{aligned}
\zeta^2(\boldsymbol{G}) &= \max_{\boldsymbol{a}_1 \in \mathbb{C}^M, \boldsymbol{a}_2 \in \mathbb{C}^{N-M}} \min_{\boldsymbol{l} \in \mathbb{G}^M, \boldsymbol{u} \in \mathbb{G}^{N-M}} \left\| \left[\delta \boldsymbol{H}^\dagger,\ \sigma \boldsymbol{H}_\perp\right] \left(\begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix} - \begin{bmatrix} \boldsymbol{l} \\ \boldsymbol{u} \end{bmatrix} \right) \right\|^2 \\
&= \max_{\boldsymbol{a}_1 \in \mathbb{G}^M, \boldsymbol{a}_2 \in \mathbb{C}^{N-M}} \min_{\boldsymbol{l} \in \mathbb{G}^M, \boldsymbol{u} \in \mathbb{G}^{N-M}} \|\delta \boldsymbol{H}^\dagger(\boldsymbol{a}_1 - \boldsymbol{l}) + \sigma \boldsymbol{H}_\perp(\boldsymbol{a}_2 - \boldsymbol{u})\|_2^2 \\
&\overset{(a)}{=} \max_{\boldsymbol{a}_1 \in \mathbb{G}^M, \boldsymbol{a}_2 \in \mathbb{C}^{N-M}} \min_{\boldsymbol{l} \in \mathbb{G}^M, \boldsymbol{u} \in \mathbb{G}^{N-M}} \delta^2 \|\boldsymbol{H}^\dagger(\boldsymbol{a}_1 - \boldsymbol{l})\|_2^2 + \sigma^2 \|\boldsymbol{H}_\perp(\boldsymbol{a}_2 - \boldsymbol{u})\|_2^2 \\
&= \max_{\boldsymbol{a}_1 \in \mathbb{G}^M} \min_{\boldsymbol{l} \in \mathbb{G}^M} \delta^2 \|\boldsymbol{H}^\dagger(\boldsymbol{a}_1 - \boldsymbol{l})\|_2^2 + \max_{\boldsymbol{a}_2 \in \mathbb{C}^{N-M}} \min_{\boldsymbol{u} \in \mathbb{G}^{N-M}} \sigma^2 \|\boldsymbol{H}_\perp(\boldsymbol{a}_2 - \boldsymbol{u})\|_2^2 \\
&\overset{(b)}{=} \delta^2 \zeta^2(\boldsymbol{H}^\dagger) + \sigma^2 (N-M)/2
\end{aligned} \tag{4.48}$$

where (a) is because $\boldsymbol{H}^\dagger$ and $\boldsymbol{H}_\perp$ are orthogonal; (b) is due to the fact that $\boldsymbol{H}_\perp$ is semi-unitary.

Using Theorem 2.2 in [91], we have

$$\lambda_1(\boldsymbol{H}^H)\zeta(\boldsymbol{H}^\dagger) \leq N, \tag{4.49}$$

where $\lambda_1(\boldsymbol{H}^H)$ is the shortest nonzero point of the lattice $\mathcal{L}(\boldsymbol{H}^H)$ defined as

$$\lambda_1(\boldsymbol{H}^H) = \min_{\boldsymbol{z} \in \mathbb{G}^N \setminus \{\boldsymbol{0}\}} \|\boldsymbol{H}^H \boldsymbol{z}\|_2.$$

Substituting (4.49) and (4.48) into (4.47), we obtain

$$\Gamma \leq c^2 \delta^2 N^2 \lambda_1^{-2}(\boldsymbol{H}^H) + c^2 \sigma^2 (N-M)/2. \tag{4.50}$$

When SNR is large ($\mathrm{SNR}/2 \geq c^2\sigma^2(N-M)$), we have

$$\begin{aligned}
&\Pr\{\Gamma \geq \mathrm{SNR}\} \\
\leq &\Pr\{c^2\delta^2 N^2 \lambda_1^{-2}(\boldsymbol{H}^H) + c^2\sigma^2(N-M)/2 \geq \mathrm{SNR}\} \\
= &\Pr\{c^2\delta^2 N^2 \lambda_1^{-2}(\boldsymbol{H}^H) + (c^2\sigma^2(N-M)/2 - \mathrm{SNR}/2) \geq \mathrm{SNR}/2\} \\
\leq &\Pr\{c^2\delta^2 N^2 \lambda_1^{-2}(\boldsymbol{H}^H) \geq \mathrm{SNR}/2\}.
\end{aligned} \tag{4.51}$$

By [84, Lemma 3], there exists a constant β such that for any $\epsilon > 0$ it holds that

$$\Pr\{\lambda_1(\boldsymbol{H}^H) \leq \epsilon\} \leq \beta\epsilon^{2N} \cdot \max\{-(\ln\epsilon)^{N+1}, 1\}. \tag{4.52}$$

Using (4.52) with $\epsilon = \sqrt{2}c\delta N/\sqrt{\mathrm{SNR}}$, (4.51) is bounded by

$$\begin{aligned} &\Pr\{\Gamma \geq \mathrm{SNR}\} \\ \leq &\beta(2c^2\delta^2N^2)^N\mathrm{SNR}^{-N} \cdot \max\{-(\ln\sqrt{2}c\delta N/\sqrt{\mathrm{SNR}})^{N+1}, 1\}. \end{aligned} \tag{4.53}$$

It follows that

$$\begin{aligned} &-\lim_{\mathrm{SNR}\to\infty} \frac{\Pr\{\Gamma \geq \mathrm{SNR}\}}{\log \mathrm{SNR}} \\ \geq &-\lim_{\mathrm{SNR}\to\infty} \left(\frac{\log\beta(2c^2\delta^2N^2)^N}{\log\mathrm{SNR}} + \frac{\log\mathrm{SNR}^{-N}}{\log\mathrm{SNR}} + \frac{\log(\ln\mathrm{SNR})^{N+1}}{\log\mathrm{SNR}} \right) \\ = &N \end{aligned} \tag{4.54}$$

By [90, Lemma 1], the diversity d is equal to

$$d = -\lim_{\mathrm{SNR}\to\infty} \frac{\log\Pr\{\Gamma \geq \mathrm{SNR}\}}{\log\mathrm{SNR}} \geq N. \tag{4.55}$$

Thus, we obtain a lower bound of the diversity order.

For the upper bound, let us denote Γ_{CVP} by the objective value of the CVP problem (4.8). As every optimal solution $\boldsymbol{l}^\star$ of the VP-PAPC problem (4.18) is feasible to the CVP problem (4.8), we have $\Gamma_{\mathrm{CVP}} \leq \Gamma$. This implies that the instantaneous noise of VP-PAPC has higher power than that of CVP. It follows that VP-PAPC has a higher SER and thus lower diversity than CVP. It has been shown in [84, 90] that CVP achieves a diversity of N. It follows that

$$d \leq N. \tag{4.56}$$

We conclude by (4.55) and (4.56) that VP-PAPC achieves a diversity of N.

4.7.5 Proof of Proposition 4.4

Let use rewrite (4.29) as follows

$$\begin{aligned} \gamma_t^\star = \min_{\boldsymbol{l}_t, \boldsymbol{d}_t, \boldsymbol{u}_t, \gamma_t} \quad & \max\{\|\boldsymbol{d}_t\|_2^2, \|\boldsymbol{D}(\sqrt{\boldsymbol{\alpha}})^{-1}\boldsymbol{d}_t\|_\infty^2\} \\ \text{s.t.} \quad & \boldsymbol{d}_t = \boldsymbol{H}^\dagger\boldsymbol{s}_t + \delta\boldsymbol{H}^\dagger\boldsymbol{l}_t + \sigma\boldsymbol{H}_\perp\boldsymbol{u}_t, \\ & \boldsymbol{d}_t \in \mathbb{C}^N,\ \boldsymbol{l}_t \in \mathbb{G}^M,\ \boldsymbol{u}_t \in \mathbb{G}^{N-M}. \end{aligned} \tag{4.57}$$

Consider the following problem

$$\begin{aligned} \min_{\boldsymbol{l}_t,\boldsymbol{d}_t,\boldsymbol{u}_t,\gamma_t} \quad & \|\boldsymbol{d}_t\|^2 \\ \text{s.t.} \quad & \boldsymbol{d}_t = \boldsymbol{H}^\dagger \boldsymbol{s}_t + \delta \boldsymbol{H}^\dagger \boldsymbol{l}_t + \sigma \boldsymbol{H}_\perp \boldsymbol{u}_t \\ & \boldsymbol{d}_t \in \mathbb{C}^N,\ \boldsymbol{l}_t \in \mathbb{G}^M,\ \boldsymbol{u}_t \in \mathbb{G}^{N-M}. \end{aligned} \tag{4.58}$$

Let us denote the optimal solution by $\tilde{\boldsymbol{d}}_t^\star$. As $\tilde{\boldsymbol{d}}_t^\star$ is feasible to (4.57), we have

$$\gamma_t^\star \leq \max\{\|\tilde{\boldsymbol{d}}_t\|_2^2, \|\boldsymbol{D}(\sqrt{\boldsymbol{\alpha}})^{-1}\tilde{\boldsymbol{d}}_t\|_\infty^2\}.$$

On the other hand, it is easy to verify that the function $\max\{\|\boldsymbol{d}\|_2, \|\boldsymbol{D}(\sqrt{\boldsymbol{\alpha}})^{-1}\boldsymbol{d}\|_\infty\}$ is a norm. Therefore, by norm equivalence in finite dimensional vector space, it follows that there exists some constant $c > 0$ such that

$$\gamma_t^\star \leq \max\{\|\tilde{\boldsymbol{d}}_t\|_2^2, \|\boldsymbol{D}(\sqrt{\boldsymbol{\alpha}})^{-1}\tilde{\boldsymbol{d}}_t\|_\infty^2\} \leq c^2\|\tilde{\boldsymbol{d}}_t\|_2^2.$$

Then, we have

$$\gamma^\star = \max_{t=1,\ldots,T} \gamma_t^\star \leq c^2 \max_{t=1,\ldots,T} \|\tilde{\boldsymbol{d}}_t\|_2^2 \leq c^2\zeta^2([\delta\boldsymbol{H}^\dagger, \sigma\boldsymbol{H}_\perp]).$$

The rest of the proof is similar to that of Proposition 4.3 and is omitted for brevity.

4.7.6 Modified Sphere Encoder

We consider the following form of integer program problem

$$\min_{\boldsymbol{d},\boldsymbol{z}} \quad f(\boldsymbol{d}) \tag{4.59a}$$

$$\text{s.t.} \quad \boldsymbol{\phi}(\boldsymbol{d}) \preceq \boldsymbol{0}, \tag{4.59b}$$

$$\boldsymbol{d} = \boldsymbol{r} - \boldsymbol{G}\boldsymbol{z}, \tag{4.59c}$$

$$\boldsymbol{z} \in \mathbb{G}^N, \boldsymbol{d} \in \mathbb{C}^N. \tag{4.59d}$$

where $\boldsymbol{\phi}(\boldsymbol{d})$ is real vector-valued function. We assume that there are positive constants c_1 and c_2 such that the objective function satisfies

$$c_1\|\boldsymbol{d}\|_2 \leq f(\boldsymbol{d}) \leq c_2\|\boldsymbol{d}\|_2. \tag{4.60}$$

Both (4.18) and (4.29) can be written in the form of (4.59); see (4.29b).

We use the Schnorr-Euchner (SE) [16] to enumerate candidate solution of (4.59) that satisfy (4.59c) and

$$\|\boldsymbol{d}\|_2^2 \leq C^2, \tag{4.61}$$

where $C >$ is a radius initially set as infinity. The rationale of using the SE strategy is that SE strategy is a very efficient way of searching for candidate solution that minimizes $\|\boldsymbol{d}\|_2$. By (4.60), solutions that minimize $\|\boldsymbol{d}\|_2$ also tend to minimize $f(\boldsymbol{d})$. Once the SE enumeration find a candidate solution $\boldsymbol{z}$, we first check the feasibility of $\boldsymbol{z}$ by checking (4.59b). If $\boldsymbol{z}$ is feasible and yields a better objective value than the best candidate solution previously found, $\boldsymbol{z}$ is kept as the best solution found. The radius C is updated according to

$$C = \frac{1}{c_1} f(\boldsymbol{d})$$

with $\boldsymbol{d} = \boldsymbol{r} - \boldsymbol{H}\boldsymbol{z}$, as the optimal solution $\boldsymbol{d}^\star$ of (4.59) satisfies $\|\boldsymbol{d}^\star\|_2 \leq \frac{1}{c_1} f(\boldsymbol{d}^\star) \leq \frac{1}{c_1} f(\boldsymbol{d})$. The SE enumeration continues with the updated radius C. If $\boldsymbol{z}$ is not feasible or its objective value is worse than that of the best candidate solution previously found, $\boldsymbol{z}$ is discarded and SE enumeration continue with the same radius C. The SE enumeration stops when all lattice points within the radius (4.61) has been enumerated.

4.7.7 Projector Operator

The projection of a given vector $(\omega_t^0, \boldsymbol{\lambda}_t^0)$ onto the set $\mathcal{F} = \{(\omega_t, \boldsymbol{\lambda}_t) \mid \omega_t \geq 0, \boldsymbol{\lambda}_t \succeq \boldsymbol{0}, \omega_t + \boldsymbol{\lambda}_t^T \boldsymbol{\alpha} = 1\}$ is given by the optimal solution of the following optimization problem

$$\begin{aligned} \min_{\omega_t, \boldsymbol{\lambda}_t} \quad & \frac{1}{2} \left\| \begin{bmatrix} \omega_t^0 - \omega_t \\ \boldsymbol{\lambda}_t^0 - \boldsymbol{\lambda}_t \end{bmatrix} \right\|_2^2 \\ \text{s.t.} \quad & \omega_t \geq 0,\ \boldsymbol{\lambda}_t \succeq \boldsymbol{0},\ \omega_t + \boldsymbol{\lambda}_t^T \boldsymbol{\alpha} = 1. \end{aligned} \tag{4.62}$$

For convenience, let us denote by $\boldsymbol{c} = (\omega_t, \boldsymbol{\lambda}_t)$, $\boldsymbol{c}^0 = (\omega_t^0, \boldsymbol{\lambda}_t^0)$, and $\boldsymbol{w} = (1, \boldsymbol{\alpha})$. Then (4.62) is recast as

$$\begin{aligned} \min_{\boldsymbol{c}} \quad & \frac{1}{2} \left\| \boldsymbol{c}^0 - \boldsymbol{c} \right\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{c} \succeq \boldsymbol{0}, \boldsymbol{c}^T \boldsymbol{w} = 1. \end{aligned} \tag{4.63}$$

This problem can be solved via solving the KKT condtions

$$\boldsymbol{c} - \boldsymbol{c}^0 + \theta \boldsymbol{w} - \boldsymbol{u} = \boldsymbol{0} \tag{4.64a}$$

$$\boldsymbol{u} \succeq \boldsymbol{0} \tag{4.64b}$$

$$\boldsymbol{x} \succeq \boldsymbol{0} \tag{4.64c}$$

$$\boldsymbol{c}^T \boldsymbol{\mu} = 0 \tag{4.64d}$$

$$\boldsymbol{w}^T \boldsymbol{c} = 1, \tag{4.64e}$$

where θ and $\boldsymbol{u}$ are the dual variables associated with the constraints of (4.63) It can be seen from (4.64a) - (4.64d) that for a given θ, $\boldsymbol{c}$ is given by

$$\boldsymbol{c} = [\boldsymbol{c}^0 - \theta \boldsymbol{w}]^+.$$

By (4.64e), the variable θ can be determined by solving

$$f(\theta) \triangleq \boldsymbol{w}^T [\boldsymbol{c}^0 - \theta \boldsymbol{w}]^+ = 1. \tag{4.65}$$

To solve this equation, let us assume that $c_1^0/w_1 \leq \ldots \leq c_N^0/w_N$. As $f(\theta)$ is a continuous nonincreasing function and $f(c_N^0/w_N) = 0$, there exists an index n' such that

$$f\left(\frac{c_{n'}^0}{w_{n'}}\right) \leq 1 < f\left(\frac{c_{n'-1}^0}{w_{n'-1}}\right).$$

This means that the solution $\theta^\star$ of (4.65) belongs to the region

$$\frac{c_{n'-1}^0}{w_{n'-1}} < \theta^\star \leq \frac{c_{n'}^0}{w_{n'}}.$$

Thus, we have

$$\begin{aligned} f(\theta^\star) &= \sum_{n=n'}^{N} [c_n^0 - \theta^\star w_n]^+ w_n + \sum_{n=1}^{n'-1} [c_n^0 - \theta^\star w_n]^+ w_n \\ &= \sum_{n=n'}^{N} (c_n^0 - \theta^\star w_n) w_n \\ &= 1. \end{aligned}$$

It follows that

$$\theta^\star = \frac{\sum_{n=n'}^{N} c_n^0 w_n}{\sum_{n=n'}^{N} w_n^2}.$$

Chapter 5

Constant Envelope Precoding

5.1 Introduction

In the previous chapter, we have considered per-antenna power constraint in vector perturbation where the maximum per-antenna power is bounded. Such a form of maximum per-antenna power constraint can help prevent signal clipping as well as increase the power efficiency of the power amplifier. However, channel inversion and vector perturbation methods, even when the maximum per-antenna power constraint is imposed, still have transmitting signals whose instantaneous power varies according to the channel realization and information vector. These transmitting signals can only be amplified by linear amplifiers that can accommodate a large signal power variation. These highly linear power amplifiers are expensive to implement and have lower power efficiency. On the other hand, the base station should be cost-effective and power efficient, as the large power consumption by the communication industry has become a global concern. The cost and power efficiency are particularly important issues in the emerging massive MIMO system [11, 92, 93] where the base station could have more than a hundred antennas.

In order to overcome the cost and power efficiency issues in power amplifiers, some very recent works [47–49] advocate the concept of *constant envelope (CE) precoding* in massive MIMO systems. In CE precoding, the transmitting signal at each antenna is restricted to have a constant amplitude irrespective of the channel and information symbol realization, and only the phases of the per-antenna

transmitting signals are used to convey information to the receiver. Since the CE signal has a constant amplitude, the instantaneous power is fixed as a constant as well. Therefore, the transmitter can use nonlinear but highly power-efficient switched-mode power amplifiers that can be cheaply implemented. While CE precoding provides an attractive signal processing way to manage power efficiency and reduce the implementation costs of power amplifiers, it also brings new signal processing challenges.

This chapter concentrates on CE precoding for single-user MISO channels. There are two fundamental challenges in this context. The first challenge is the *characterization of the region of all possible noise-free receive signals* generated by CE precoding. This problem is crucial in determining whether a given constellation can be supported by CE precoding. Mohammed and Larsson in their pioneering work [47] prove that the noise-free receive signal region is a doughnut region, i.e., a region between two circles centered at the origin of the complex plane. Moreover, the radius of the outer circle is shown to be the sum of all channel amplitudes. However, the inner radius is not known in that reference. The second challenge is the *phase recovery problem*, which is a precoder problem at the transmitter side and plays an indispensable role in CE precoding implementations. The problem is to find the phases of the CE signals such that the CE signals, after coherently combined by the channels, form a desired information signal. Mathematically, the phase recovery problem amounts to solving a highly nonlinear equation. The work in [47] handles the phase recovery problem by formulating the problem as an optimization problem, and applying gradient descent.

In this chapter, we develop an alternative approach to provide a complete characterization of the noise-free received signal region. Our approach not only provides a simple proof of the characterization results in [47], but also gives a simple expression of the inner radius of the doughnut region. Thus, with this new result, one can easily check whether CE precoding can support a given constellation. More importantly, the inductive and constructive nature of our approach leads to a direct solution to the phase recovery problem. We derive an

efficient phase recovery algorithm that solves the phase recovery problem exactly with a complexity linear in the number of antennas. In addition, we consider two novel CE precoding scenarios, wherein the system has the flexibility to perform either antenna-subset selection (AS) or unequal amplitude (UA) transmission. We formulate SER minimization problems where imperfect CSIT, constellation supportability and total power constraints are also taken into account. As it turns out, the UA strategy results in optimization problems that can be transformed to second-order cone programs (SOCPs) and thus have efficient exact solutions by available algorithms [60, 61]. For the AS strategy, though the resulting problems are combinatorial and nonconvex, we devise polynomial-time exact searching algorithms. Simulation results will show that CE precoding via an optimal AS and UA design can achieve SER performance comparable to the representative non-CE maximum ratio transmission (MRT) method.

The rest of this chapter is organized as follows. We introduce the system model and CE precoding in Section 5.2.1. Then, in Section 5.3, we provide the signal region characterization and propose the exact phase recovery algorithm. This is followed by Section 5.4, where we formulate the design optimization of CE precoding under the AS and UA strategies, and propose optimal and efficient algorithms for the formulated problems. Simulation results are presented in Section 5.5 to demonstrate the performance of the proposed methods. Section 5.6 summarizes this chapter.

5.2 Background

5.2.1 System Model and CE Precoding Problems

We consider a standard single-user MISO channel model

$$y = \boldsymbol{h}^T \boldsymbol{x} + \nu, \tag{5.1}$$

where $y \in \mathbb{C}$ is the receive signal; $\boldsymbol{x} = [x_1, \dots x_N]^T \in \mathbb{C}^N$ is the transmitting signal, with x_i being the transmitting signal at ith antenna and N being the number of antennas; $\boldsymbol{h} = [h_1, \dots, h_N]^T \in \mathbb{C}^N$ is the channel vector; $\nu \in \mathbb{C}$ is complex circular additive white Gaussian noise whose mean and variance are

0 and σ_ν^2, respectively. The problem is to transmit information symbols drawn from a symbol constellation, given channel state information at the transmitter (CSIT). To describe, let $\mathcal{S}$ be the symbol constellation (e.g., QAM). The set $\mathcal{S}$ is assumed to have unit average power with its symbols, that is, $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}|s|^2 = 1$. The task is to design the transmitting signal $\boldsymbol{x}$ such that the noise-free received signal equals

$$d \triangleq \boldsymbol{h}^T\boldsymbol{x} = \alpha \cdot s, \tag{5.2}$$

where $s \in \mathcal{S}$ is the information symbol to be transmitted, and $\alpha > 0$ is a constant. Note that the constant α describes the effective channel gain at the receiver.

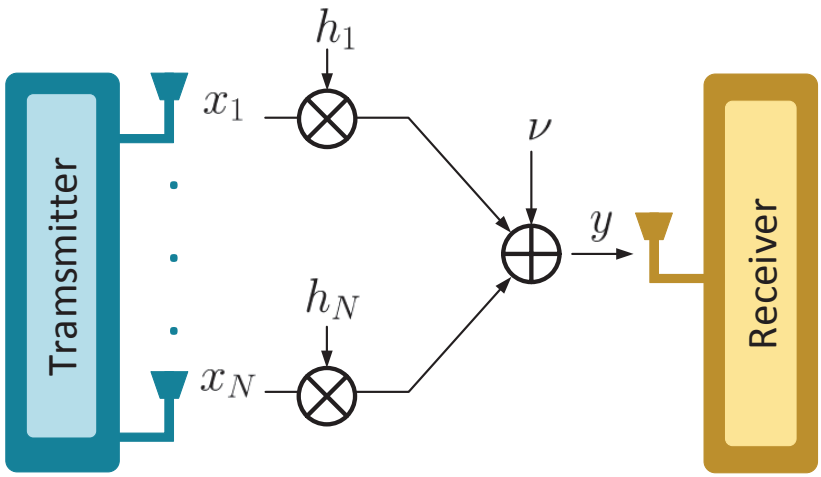


Figure 5.1: A single-user MISO model.

A simple, convenient way to carry out the precoding task mentioned above is channel inversion method which in the single-user scenario is normally called maximum ratio transmission (MRT). The MRT transmitting signal takes the form of takes the form

$$\boldsymbol{x}_{\mathrm{MRT}} = \sqrt{P_{\mathrm{T}}}\frac{\boldsymbol{h}^*}{\|\boldsymbol{h}\|_2}s, \tag{5.3}$$

where P_{T} is the average total transmission power. Note that the resulting effective channel gain is $\alpha_{\mathrm{MRT}} = \sqrt{P_{\mathrm{T}}}\|\boldsymbol{h}\|_2$. It can be easily shown that the average per-antenna power of MRT equals a constant $\mathbb{E}[|x_{\mathrm{MRT},i}|^2] = \frac{P_{\mathrm{T}}}{N}$ for an i.i.d fading

channel. However, the *instantaneous per-antenna powers*, $|x_{\mathrm{MRT},i}|^2$, are dependent on the realization of $\boldsymbol{h}$ and s, and may vary dramatically from zero to $\max_{s\in\mathcal{S}} P_{\mathrm{T}}|s|^2$. In order to accommodate the large variations of the instantaneous per-antenna power, the RF amplifier built for MRT signals must have a wide linear region, which inevitably leads to a low power efficiency. The power efficiency for such highly linear RF amplifiers is typically about $0.15-0.25$ [47,94].

The difficulty in using highly power-efficient RF amplifiers for large antenna array systems has recently motivated the use of constant envelope (CE) signals for transmission [47]. CE precoding is a nonlinear scheme with respect to the information symbols. In essence, we constrain the transmitting signal x_i of each antenna to take the form of

$$x_i = \sqrt{\frac{P_{\mathrm{T}}}{N}} e^{\boldsymbol{j}\theta_i}, \text{ for } i = 1, \ldots, N, \tag{5.4}$$

where $\theta_i \in [0, 2\pi)$ is the phase of x_i. In contrast to MRT, the *instantaneous power* of CE signal x_i is fixed at $|x_i|^2 = \frac{P_{\mathrm{T}}}{N}$, which is independent of the channel and information symbols. Hence, the RF amplifiers for CE signals can have a high power efficiency ranging from 0.75 to 0.85 [47, 94].

While the CE signal (5.4) enables the use of highly power-efficient RF amplifiers, it also presents new challenges. The first challenge is the characterization of the set of all possible noise-free receive signal, which is defined as

$$\mathcal{D} \triangleq \left\{ \sqrt{\frac{P_{\mathrm{T}}}{N}} \sum_{i=1}^{N} h_i e^{\boldsymbol{j}\theta_i} \;\middle|\; \theta_i \in [0, 2\pi), i = 1, \ldots, N \right\}. \tag{5.5}$$

The motivation for characterizing $\mathcal{D}$ is that it underpins the feasibility for CE precoding to transmit, fixing a constellation $\mathcal{S}$. Specifically, we must ensure that $\alpha\mathcal{S} \subset \mathcal{D}$ for some $\alpha > 0$, for otherwise the CE precoding scheme is unable to generate all information symbols. To get some intuitive insight, in Fig. 5.2 we use pictures to illustrate two cases where CE precoding is able and unable to support a given symbol constellation, respectively. In the figure, the blue region represents the noise-free receive signal region $\mathcal{D}$ and the dots are the constellation points of $\mathcal{S}$. The 16-QAM constellation is used in the illustrations. In Fig. 5.2(a), all the constellation points lie in $\mathcal{D}$, which means that CE precoding is able to

support the constellation. An unsupportable counterpart is shown in Fig. 5.2(b); we see that part of the constellation points is outside $\mathcal{D}$. Also, no matter how we scale the 16-QAM constellation points, there are always some constellation points not covered by $\mathcal{D}$ in Fig. 5.2(b).

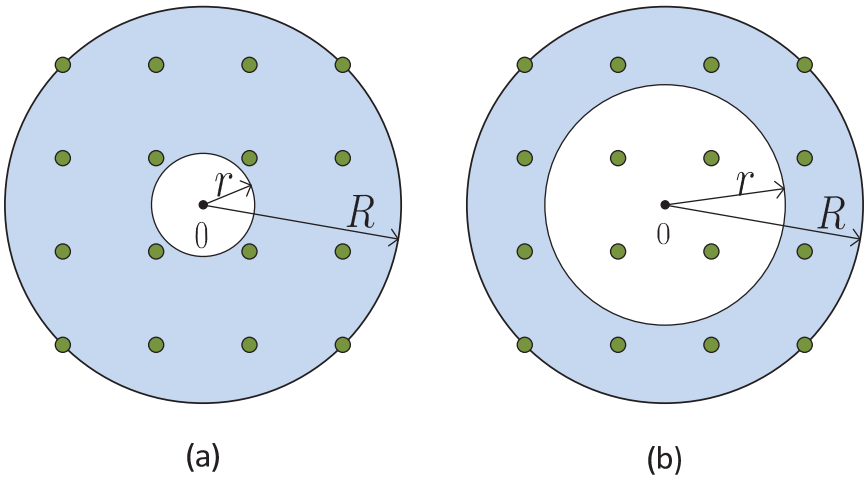


Figure 5.2: The noise-free receive signal region $\mathcal{D}$.

The second challenge is the phase recovery problem. For each $d = \alpha s$, $s \in \mathcal{S}$, the CE precoder needs to find a phase vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]^T$ that solves

$$\text{find } \boldsymbol{\theta} \in [0, 2\pi)^N \tag{5.6a}$$

$$\text{s.t. } d = \sqrt{\frac{P_{\mathrm{T}}}{N}} \sum_{i=1}^{N} h_i e^{\boldsymbol{j}\theta_i}, \tag{5.6b}$$

that is, we wish to shape a desired noise-free receive signal value d by recovering the corresponding phase vector $\boldsymbol{\theta}$ at the transmitter side. Unlike MRT which is a linear precoding scheme, CE precoding has a highly nonlinear relationship between the noise-free receive signal d and the phase vector $\boldsymbol{\theta}$. This nonlinear phase recovery problem introduces a challenge in efficient CE precoding in practice.

5.2.2 Prior Work

The pioneering work [47] by Mohammed and Larsson shows that $\mathcal{D}$ is a doughnut region given by

$$\mathcal{D} = \{d \in \mathbb{C} \mid r \leq |d| \leq R\}, \tag{5.7}$$

where r and R are scalars depending on $\boldsymbol{h}$. Moreover, r and R are shown in the same reference to satisfy

$$r \leq \sqrt{\frac{P_{\mathrm{T}}}{N}}\|\boldsymbol{h}\|_{\infty}, \quad R = \sqrt{\frac{P_{\mathrm{T}}}{N}}\|\boldsymbol{h}\|_{1}. \tag{5.8}$$

However, the exact value of r is not known. The work in [47] also considered the phase recovery problem, where the phase recovery problem in (5.6) is formulated as a minimization problem

$$\min_{\boldsymbol{\theta}} \left| d - \sqrt{\frac{P_{\mathrm{T}}}{N}} \sum_{i=1}^{N} h_i e^{\boldsymbol{j}\theta_i} \right|. \tag{5.9}$$

To solve problem (5.9), the gradient descent method was proposed for large N; a two-step algorithm combining depth-first-search (DFS) and gradient descent was also proposed for small N ($N \leq 10$).

5.3 Signal Region Characterization and Exact Phase Recovery

In this section we characterize and analyze the noise-free receive signal region $\mathcal{D}$ through a proof different from that by Mohammed and Larsson [47]. In particular, it is shown that the inner radius r of $\mathcal{D}$ has a simple closed-form expression. We also propose an exact and closed-form solution for the phase recovery problem (5.9), which is derived by taking insight from our alternative signal region characterization proof.

For notational convenience in the subsequent development, denote for $i = 1, \ldots, N$,

$$g_i = \sqrt{\frac{P_{\mathrm{T}}}{N}}|h_i|, \quad \phi_i = \theta_i + \varphi_i,$$

where φ_i is the argument of h_i. Then, (5.6b) and (5.5) can be equivalently expressed as

$$d = \sum_{i=1}^{N} g_i e^{\boldsymbol{j}\phi_i}, \tag{5.10}$$

$$\mathcal{D} = \left\{ \sum_{i=1}^{N} g_i e^{\boldsymbol{j}\phi_i} \middle| \phi_i \in [0, 2\pi), i = 1, \ldots, N \right\}. \tag{5.11}$$

Without loss of generality, we assume that g_1 and g_2 are respectively the first and the second largest elements in $\{g_i\}_{i=1}^N$, i.e. $g_1 \geq g_2 \geq g_i \geq 0$ for $i = 3, \ldots, N$. We also define, for $i = 1, \ldots, N$,

$$\mathcal{D}_i \triangleq \left\{ d_i = \sum_{j=1}^{i} g_j e^{\boldsymbol{j}\phi_j} \;\middle|\; \phi_j \in [0, 2\pi), j = 1, \ldots, i \right\}. \tag{5.12}$$

Physically, $\mathcal{D}_i$ can be interpreted as the noise-free receive signal region when only the first i antennas are used. Note that $\mathcal{D}_N = \mathcal{D}$.

5.3.1 Characterization of $\mathcal{D}$

The results for our CE receive signal region characterization are summarized in the following theorem.

Theorem 5.1 *For every $i \in \{1, \ldots, N\}$, the set $\mathcal{D}_i$ is a doughnut region*

$$\mathcal{D}_i = \{d_i \in \mathbb{C} \;\mid\; r_i \leq |d_i| \leq R_i\}, \tag{5.13}$$

where $R_i = \sum_{j=1}^{i} g_j$ and $r_i = \max\{g_1 - \sum_{j=2}^{i} g_j,\ 0\}$. In particular, the noise-free receive signal region in (5.5) *or* (5.11) *is given by*

$$\mathcal{D} = \{d \in \mathbb{C} \;\mid\; r \leq |d| \leq R\},$$

where $R = \sum_{j=1}^{N} g_j$ and $r = \max\{g_1 - \sum_{j=2}^{N} g_j,\ 0\}$, and the two radii can be alternatively expressed as

$$R = \sqrt{\frac{P_{\mathrm{T}}}{N}} \|\boldsymbol{h}\|_1, \tag{5.14}$$

$$r = \sqrt{\frac{P_{\mathrm{T}}}{N}} \max\{2\|\boldsymbol{h}\|_\infty - \|\boldsymbol{h}\|_1, 0\}. \tag{5.15}$$

The proof of Theorem 5.1 will be described in the next subsection. Theorem 5.1 completes the previous noise-free receive signal region characterization by Mohammed and Larsson [47], where we provide an explicit expression for the inner radius r in (5.15).

Theorem 5.1 provides several important implications. First, the inner doughnut radius expression in (5.15) suggests that $\mathcal{D}$ should be a disk region in many

practical cases in large antenna array systems—in order for r to be strictly positive, it must hold true that there exists a channel element h_i whose amplitude $|h_i|$ is greater than $\sum_{j\neq i}|h_j|$, or, simply speaking, much greater than the other channel elements' amplitudes. In practice, one would expect that it is not too likely to encounter an instance in which one channel element amplitude $|h_i|$ is so dominant over the others. In fact, for the i.i.d. Gaussian channel, the following result can be proven.

Proposition 5.1 *Suppose that each element of* $\boldsymbol{h}$ *follows an i.i.d. circular complex Gaussian distribution with zero mean. Then,*

$$\frac{1}{N^{N-2}} \leq \Pr\{r>0\} \leq \frac{1}{(N-1)!}.$$

The proof is based on direct integration of the distribution of the ordered statistics of $\{|h_i|\}_{i=1}^N$ [95]; the details are relegated to Appendix 5.7.2. The pioneering work [47] has provided a similar result that $\Pr\{r \geq c(\log N)/\sqrt{N}\}$ converges to zero as N goes to infinity for all $c>0$. We can see that the result in Proposition 5.1 provides a better guarantee of r being zero; this is owing to the fact that Proposition 5.1 is proven from the explicit expression of r in (5.15), while the previous result was not. Proposition 1 states that $\Pr\{r>0\}$ decays factorially fast in N. For example, for $N=10$, we have $\Pr\{r>0\} \leq 3\times 10^{-7}$. For very large array systems, where N could be more than 100, it is expected that $\Pr\{r>0\}$ is virtually zero. This indicates that with high probability, the doughnut region is essentially a disk region (for i.i.d. Gaussian channels).

Second, the proof of Theorem 5.1 provides insights into solving the phase recovery problem (5.6) in an exact and polynomial-time manner; this will be elaborated upon in the subsequent subsections. In this regard, we should note that our proof of the noise-free receive signal region characterization is constructive, based on induction and satisfiability of some nonlinear equations. It is the inductive and constructive nature of the proof that allows us to transfer the idea to the phase recovery problem. Also, we should point out that our proof is different from the previous pioneering proof of the noise-free receive signal region characterization [47]; the latter, simply speaking, is based on an existence

argument. Third, the simple closed-form nature of the inner doughnut radius r in (5.15), as well as that of the outer doughnut radius R in (5.14), allow one to easily check whether a given symbol constellation is supportable by the CE precoding scheme. In fact, the simple doughnut radii characterization in (5.14)–(5.15) will allow us to perform CE precoder optimization in a tractable fashion. The latter will be considered in Section 5.4.

5.3.2 Proof of Theorem 5.1

The proof of Theorem 1 is as follows. The main idea is to show by induction from $i = 1$ to $i = N$ that $\mathcal{D}_i = \{d_i \in \mathbb{C} \mid r_i \leq |d_i| \leq R_i\}$. For $i = 1$, this holds true obviously. For $i = 2$, from the definition of $\mathcal{D}_2$ in (5.12), we can see that if $d_2 \in \mathcal{D}_2$, then $r_2 = g_1 - g_2 \leq |d_2| \leq g_1 + g_2 = R_2$ by triangular inequality. Conversely, if d_2 satisfies $g_1 - g_2 \leq |d_2| \leq g_1 + g_2$, then we need to find (ϕ_1, ϕ_2) satisfying

$$d_2 = g_1 e^{\boldsymbol{j}\phi_1} + g_2 e^{\boldsymbol{j}\phi_2}. \tag{5.16}$$

It can be easily verified that the (ϕ_1, ϕ_2) given below is a solution,

$$\begin{aligned} \phi_1 &= \arccos\left(\frac{g_1^2 + |d_2|^2 - g_2^2}{2g_1|d_2|}\right) + \omega_2 \\ \phi_2 &= \arccos\left(\frac{g_1^2 + g_2^2 - |d_2|^2}{2g_1 g_2}\right) + \phi_1 + \pi \end{aligned} \tag{5.17}$$

where ω_2 is the argument of d_2. Hence, (5.13) is true for $i = 2$.

For $i \geq 3$, we need to invoke the following lemma which reveals the relationship between $\mathcal{D}_i$ and $\mathcal{D}_{i-1}$.

Lemma 5.1 *Let*

$$\begin{aligned} \mathcal{A} &= \{\ x \in \mathbb{C} \mid r_a \leq |x| \leq R_a\ \}, \\ \mathcal{B} &= \{\ y \in \mathbb{C} \mid |y| = r_b\ \}, \\ \mathcal{C} &= \{\ z \in \mathbb{C} \mid z = x + y, x \in \mathcal{A}, y \in \mathcal{B}\ \}, \end{aligned}$$

and suppose that

$$R_a - r_a \geq 2r_b. \tag{5.18}$$

Then, $\mathcal{C}$ is a doughnut region

$$\mathcal{C} = \{\ z \in \mathbb{C} \mid r_c \leq |z| \leq R_c\ \},$$

with

$$r_c = \max\{r_a - r_b, 0\}, \quad R_c = R_a + r_b.$$

Moreover, for any $z \in \mathcal{C}$, we can construct $x \in \mathcal{A}, y \in \mathcal{B}$ such that $z = x + y$ holds. Specifically, such (x, y) is obtained by setting

$$y = \begin{cases} r_b e^{\boldsymbol{j}\phi_z}, & |z| \geq R_a - r_b \\ r_b e^{\boldsymbol{j}(\phi_z + \pi)}, & |z| < R_a - r_b \end{cases} \tag{5.19}$$

and $x = z - y$, where ϕ_z denotes the argument of z.

By the definitions of $\mathcal{D}_i$ and $\mathcal{D}_{i-1}$, $\mathcal{D}_i$ can be written as

$$\mathcal{D}_i = \{d_i \in \mathbb{C} \ \mid \ d_i = d_{i-1} + \tilde{d}_i, d_{i-1} \in \mathcal{D}_{i-1}, |\tilde{d}_i| = g_i\}. \tag{5.20}$$

Suppose that $\mathcal{D}_{i-1}$ is a doughnut region with radii $r_{i-1} = \max\{g_1 - \sum_{j=2}^{i-1} g_j, 0\}$ and $R_{i-1} = \sum_{j=1}^{i-1} g_j$. Then,

$$R_{i-1} - r_{i-1} \geq R_2 - r_2 = 2g_2 \geq 2g_i,$$

which satisfies the premise (5.18) of Lemma 1. Applying Lemma 1 to (5.20), we have that $\mathcal{D}_i$ is a doughnut region with radii $r_i = \max\{r_{i-1} - g_i, 0\} = \max\{g_1 - \sum_{j=2}^{i} g_j, 0\}$ and $R_i = R_{i-1} + g_i = \sum_{j=1}^{i} g_j$.

5.3.3 Exact Phase Recovery

In this subsection, we propose an exact phase recovery algorithm for (5.10) which has a linear complexity in the problem size N.

The main idea of the proposed algorithm is derived from the proof of Theorem 5.1. Assume that $r \leq |d| \leq R$, for otherwise (5.10) has no solution by Theorem 1. Let $d_N \triangleq d$. Observe from (5.20) that if d_N belongs to $\mathcal{D}_N$, then a ϕ_N exists such that $d_N - g_N e^{\boldsymbol{j}\phi_N}$ belongs to $\mathcal{D}_{N-1}$; again, a ϕ_{N-1} exists such that $(d_N - g_N e^{\boldsymbol{j}\phi_N}) - g_{N-1} e^{\boldsymbol{j}\phi_{N-1}}$ belongs to $\mathcal{D}_{N-2}$. Repeating this argument, it

can be seen that d_N can be decomposed in the form of $d_N = \sum_{i=1}^{N} g_i e^{\boldsymbol{j}\phi_i}$. Hence, it suffices to choose ϕ_i such that

$$d_{i-1} \triangleq d_i - g_i e^{\boldsymbol{j}\phi_i} \in \mathcal{D}_{i-1}. \tag{5.21}$$

from $i = N$ down to $i = 2$. At the end of the process, the resultant $\boldsymbol{\phi}$ is a solution[1] of (5.10).

The proof of Theorem 1 already offers a way to determine a ϕ_i for (5.21). By (5.19) in Lemma 1, we can see that for $i \geq 3$, ϕ_i can be chosen as

$$\phi_i = \begin{cases} \omega_i, & \text{if } |d_i| \geq R_{i-1} - g_i, \\ \omega_i + \pi, & \text{if } |d_i| < R_{i-1} - g_i, \end{cases} \tag{5.22}$$

where ω_i is the argument of d_i. For $i = 2$, by noting $r_1 = R_1 = g_1$, it can be seen that (5.21) is equivalent to the equation in (5.16). Then ϕ_2 and ϕ_1 can be chosen as (5.17).

We can see that the proposed algorithm only involves N steps of operations. Hence, the complexity of the proposed algorithm is $\mathcal{O}(N)$. The description of the proposed algorithm is complete, and we provide the pseudo code in Algorithm 5.1.

5.4 Robust Transmit Optimization of CE Precoding with Channel Uncertainty

This section turns the attention to the design optimization of CE precoding, where we are allowed to either select a subset of antennas, or allocate power unequally for each antenna, to maximize system performance.

In the previous system model in Section 5.2.1, we assume perfect CSIT. Here, we consider imperfect CSIT. Under such scenarios, the single-user MISO channel model in (5.1) should be replaced by

$$y = (\bar{\boldsymbol{h}} + \Delta\boldsymbol{h})^T \boldsymbol{x} + \nu, \tag{5.23}$$

[1]Note that ϕ_1 is automatically obtained when choosing ϕ_2 such that $d_2 - g_2 e^{\boldsymbol{j}\phi_2} = d_1$, since d_1 is of the form of $d_1 = g_1 e^{\boldsymbol{j}\phi_1}$.

Algorithm 5.1: Exact phase recovery for problem (5.6).

input : $d = |d|e^{\boldsymbol{j}\omega}$, $h_i = |h_i|e^{\boldsymbol{j}\varphi_i}$, $i = 1, \ldots, N$, with

$$|h_1| \geq |h_2| \geq |h_i| \geq 0, \ \forall\ i \geq 3.$$

$g_i = \sqrt{\frac{P_{\mathrm{T}}}{N}}|h_i|$, $i = 1, \ldots, N$;

$R_N = \sum_{j=1}^{N} g_j$; $\quad r_N = \max\{g_1 - \sum_{j=2}^{N} g_j,\, 0\}$;

if $|d| > R_N$ **or** $|d| < r_N$ **then**

  return. (There is no solution);

end

$d_N = |d|$;

for $i \leftarrow N$ **to** 3 **do**

  $R_{i-1} = R_i - g_i$;

  if $d_i \geq R_{i-1} - g_i$ **then**

    $\phi_i = \omega$; $\ d_{i-1} = d_i - g_i$;

  else

    $\phi_i = \omega + \pi$; $\ d_{i-1} = d_i + g_i$;

  end

end

$\phi_1 = \omega + \arccos \frac{g_1^2 + d_2^2 - g_2^2}{2g_1 d_2}$;

$\phi_2 = \phi_1 + \pi + \arccos \frac{g_1^2 + g_2^2 - d_2^2}{2g_1 g_2}$;

output: $\{\theta_i\}_{i=1}^{N} = \{\phi_i - \varphi_i\}_{i=1}^{N}$

where $\bar{\boldsymbol{h}}$ is the channel estimate, of which the transmitter has full information; and $\Delta\boldsymbol{h}$ is the channel uncertainty. We consider two classic channel uncertainty models, namely the stochastic model [96–98] and the deterministic model [99–101]. In the stochastic uncertainty model, the channel uncertainty $\Delta\boldsymbol{h}$ is modeled as a zero-mean Gaussian random vector

$$\Delta\boldsymbol{h} \sim \mathcal{CN}(\boldsymbol{0}, \delta^2\boldsymbol{I}),$$

where δ^2 represents the uncertainty level. In the deterministic uncertainty model, the channel uncertainty $\Delta\boldsymbol{h}$ is assumed to be deterministic unknown and lie in the box region

$$\|\Delta\boldsymbol{h}\|_\infty \leq \epsilon,$$

where $\epsilon \geq 0$ is a parameter controlling the level of uncertainty.

In the previously considered CE precoding scheme, we have each antenna's transmitting signal taking the plain form $x_i = \sqrt{\frac{P_{\mathrm{T}}}{N}} e^{\boldsymbol{j}\theta_i}, i = 1, \ldots, N$, with $\{\theta_i\}_{i=1}^N$ varying in accordance with the information symbol $s \in \mathcal{S}$ for a given channel. For convenience, this previous CE precoding strategy will be named *plain CE precoding* in the sequel. Now, we are interested in extending the plain CE precoding strategy by adopting *antenna selection (AS)*. The rationale is to allow the system to adapt to the channel conditions, so that power can be more efficiently utilized over a good subset of channels. We also wish to perform the latter in a manner that is robust against channel uncertainties. Moreover, AS is meaningful in that the hardware overheads associated with the number of RF chains may be reduced. The AS strategy is formulated as follows. We write

$$x_i = a_i e^{\boldsymbol{j}\theta_i}, \quad i = 1, \ldots, N, \tag{5.24}$$

where $a_i \in \mathbb{R}_+$ is the signal amplitude at the ith antenna. Specifically, if the ith antenna is not selected, then we shut down the ith antenna by setting $a_i = 0$. On the other hand, if the ith antenna is selected, then we set the corresponding a_i to a common amplitude value $b \in \mathbb{R}_+$. More concisely, the feasible set of the amplitude vector $\boldsymbol{a} = [a_1, \ldots, a_N]^T$ can be expressed as

$$\mathcal{A}_{\mathrm{AS}} = \{\boldsymbol{a} \in \{0, b\}^N \mid b \in \mathbb{R}_+\}. \tag{5.25}$$

It should be noted that $\boldsymbol{a}$ is fixed for all transmitted information symbols $s \in \mathcal{S}$, or, in practice, fixed over the whole transmission period. The design task is to choose a good $\boldsymbol{a}$ given the channel estimate $\bar{\boldsymbol{h}}$; the precise problem formulation will be given in the next two subsections.

This chapter also explores an alternative transmit strategy where the amplitudes a_i's can be freely selected in a soft manner; that is, the feasible set of $\boldsymbol{a}$ is

$$\mathcal{A}_{\mathrm{UA}} = \mathbb{R}_+^N. \tag{5.26}$$

We will call the above strategy the *unequal amplitude (UA) strategy*. The UA strategy requires that the system is able to allocate unequal power for each

antenna, the hardware implementation of which would be more demanding than the AS strategy. However, UA is a relaxed version of AS and hence will deliver performance at least no worse than that of AS.

Our interest is in designing the AS and UA strategies by formulating the design as a power-constrained quality of service (QoS) maximization problem. In the design to be formulated, the supportability of the symbol constellation $\mathcal{S}$ is also taken into consideration.

We will consider the stochastic uncertainty model in the next subsection where we develop the problem formulations and fast algorithms. Then, in the second subsection we extend the problem formulations and fast algorithms to the deterministic uncertainty model.

5.4.1 Robust Design with Stochastic Channel Uncertainty

Problem Formulation

Our desired performance measure is the SER averaged over the random channel uncertainty. To derive it, we first re-examine the transmitter side's operation. Given an amplitude design $\boldsymbol{a}$ and a channel estimate $\bar{\boldsymbol{h}}$, the transmitter performs CE precoding by satisfying the equation

$$\alpha s = \sum_{i=1}^{N} \bar{h}_i a_i e^{\boldsymbol{j}\theta_i} \tag{5.27}$$

for any information symbol $s \in \mathcal{S}$, where $\alpha > 0$ is some constant [cf. (5.2) and (5.24), as well as the counterpart for the plain strategy in (5.6b)]. From (5.23) and (5.27), the receive signal can be written as

$$y = \alpha s + \left(\sum_{i=1}^{N} \Delta h_i a_i e^{\boldsymbol{j}\theta_i} + \nu \right). \tag{5.28}$$

The receiver obtains a symbol decision, denoted by $\hat{s}$, by applying threshold decision on y/α. By the previously laid assumptions of $\Delta\boldsymbol{h} \sim \mathcal{CN}(\boldsymbol{0}, \delta^2 \boldsymbol{I})$ and $\nu \sim \mathcal{CN}(0, \sigma_\nu^2)$, and by assuming that s is i.i.d. uniform, it can be shown that the SER averaged over the channel uncertainty $\Delta\boldsymbol{h}$ is upper-bounded by

$$\mathbb{E}_{\Delta\boldsymbol{h}} \left[\Pr\{\hat{s} \neq s | \Delta\boldsymbol{h}\} \right] \leq \frac{1}{|\mathcal{S}|} \sum_{\substack{s,s' \in \mathcal{S} \\ s \neq s'}} Q\left(\frac{\alpha |s - s'|}{\sqrt{2}\sqrt{\delta^2 \|\boldsymbol{a}\|_2^2 + \sigma_\nu^2}} \right), \tag{5.29}$$

where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$, and $Q(\cdot)$ is the Q-function. The upper bound in (5.29) is obtained by following the spirit as in the SER derivations in some other work, e.g., [83, 102]; and note that the former is well known to be an accurate approximation of the SER. It can be easily shown that the upper bound in (5.29) is a decreasing function of the term $\alpha^2/(\delta^2\|\boldsymbol{a}\|_2^2 + \sigma_\nu^2)$—which will be called the *effective receive SNR* here. Hence, we may minimize the SER in an accurate manner by maximizing the effective receive SNR.

Based on the aforedescribed problem setup, we formulate our desired design (for both the AS and UA strategies) as an optimization problem

$$\max_{\boldsymbol{a}\in\mathbb{R}_+^N, \alpha\in\mathbb{R}_+} \quad \frac{\alpha}{\sqrt{\delta^2\|\boldsymbol{a}\|_2^2 + \sigma_\nu^2}} \tag{5.30a}$$

$$\text{s.t. } \alpha\mathcal{S} \subset \mathcal{D}(\bar{\boldsymbol{h}} \odot \boldsymbol{a}) \tag{5.30b}$$

$$\boldsymbol{a}^T\boldsymbol{a} \leq P_{\mathrm{T}} \tag{5.30c}$$

$$\boldsymbol{a} \leq \sqrt{P_{\mathrm{PA}}}\mathbf{1} \tag{5.30d}$$

$$\boldsymbol{a} \in \mathcal{A}, \tag{5.30e}$$

where $\mathcal{D}(\bar{\boldsymbol{h}} \odot \boldsymbol{a})$ denotes the noise-free receive signal region defined by the equivalent channel $\bar{\boldsymbol{h}} \odot \boldsymbol{a}$ (whose expression will be considered soon), $\mathcal{A}$ is either the AS feasible set $\mathcal{A}_{\mathrm{AS}}$ or the UA feasible set $\mathcal{A}_{\mathrm{UA}}$, and P_{T} and P_{PA} are the maximum allowable total and per-antenna transmission powers, respectively. As can seen in problem (5.30), we aim to maximize the effective receive SNR, subject to the total transmission power constraint, per-antenna transmission power constraints, and supportability of the given symbol constellation $\mathcal{S}$. In particular, (5.30b) means that the doughnut region $\mathcal{D}(\bar{\boldsymbol{h}} \odot \boldsymbol{a})$ can support the constellation $\alpha\mathcal{S}$.

We proceed to reformulate problem (5.30) to a more convenient form. From the result of Theorem 1, the constraint (5.30b) can be explicitly expressed as

$$2\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_\infty - \|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1 \leq \alpha|s| \leq \|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1, \ \forall\ s \in \mathcal{S}. \tag{5.31}$$

Let $|s|_{\max} = \max_{s\in\mathcal{S}}|s|$ and $|s|_{\min} = \min_{s\in\mathcal{S}}|s|$ denote the maximum and minimum amplitudes of the symbols in $\mathcal{S}$, respectively. Eq. (5.31) is equivalent

to

$$\begin{aligned} &2\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_\infty - \|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1 \leq \alpha |s|_{\min}, \\ &\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1 \geq \alpha |s|_{\max}. \end{aligned} \tag{5.32}$$

Substituting (5.32) into problem (5.30) yields the following equivalent design optimization problem

$$\max_{\boldsymbol{a}\in\mathbb{R}_+^N, \alpha\in\mathbb{R}_+} \quad \frac{\alpha}{\sqrt{\delta^2\|\boldsymbol{a}\|_2^2 + \sigma_\nu^2}} \tag{5.33a}$$

$$\text{s.t.} \quad \frac{1}{|s|_{\min}}(2\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_\infty - \|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1) \leq \alpha \tag{5.33b}$$

$$\frac{1}{|s|_{\max}}\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1 \geq \alpha \tag{5.33c}$$

$$(5.30c) \text{ – } (5.30e). \tag{5.33d}$$

Observe that at the optimum of problem (5.33), α must take the value

$$\alpha = \frac{1}{|s|_{\max}}\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1. \tag{5.34}$$

Hence, we can directly substitute (5.34) into problem (5.33) and rewrite the design optimization problem as

$$\max_{\boldsymbol{a}\in\mathbb{R}_+^N} \quad \frac{\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1}{|s|_{\max}\sqrt{\delta^2\|\boldsymbol{a}\|_2^2 + \sigma_\nu^2}} \tag{5.35a}$$

$$\text{s.t.} \quad \frac{1}{|s|_{\min}}(2\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_\infty - \|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1) \leq \frac{1}{|s|_{\max}}\|\bar{\boldsymbol{h}} \odot \boldsymbol{a}\|_1 \tag{5.35b}$$

$$(5.30c) \text{ – } (5.30e). \tag{5.35c}$$

Notice that once we solve problem (5.35), we can obtain the corresponding optimal α by the relation in (5.34). Problem (5.35) can be further simplified. With a slight abuse of notations, let us denote

$$\boldsymbol{g} = [|\bar{h}_1|, \ldots, |\bar{h}_N|]^T$$

to be the amplitude vector of the channel. Since both $\boldsymbol{g}$ and $\boldsymbol{a}$ are nonnegative,

problem (5.35) can be equivalently written as

$$\max_{\boldsymbol{a}\in\mathbb{R}_+^N} \quad \frac{\boldsymbol{g}^T\boldsymbol{a}}{|s|_{\max}\sqrt{\delta^2\|\boldsymbol{a}\|_2^2+\sigma_\nu^2}} \tag{5.36a}$$

$$\text{s.t.} \quad \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} \max_{i=1,\ldots,N} g_i a_i \le \boldsymbol{g}^T\boldsymbol{a} \tag{5.36b}$$

$$\boldsymbol{a}^T\boldsymbol{a} \le P_{\mathrm{T}}, \quad \boldsymbol{a} \le \sqrt{P_{\mathrm{PA}}}\mathbf{1} \tag{5.36c}$$

$$\boldsymbol{a} \in \mathcal{A}. \tag{5.36d}$$

At this point, let us investigate the structure of the equivalent design optimization problem in (5.36). The objective function is quasi-concave. Hence, if $\mathcal{A}$ is a convex set, problem (5.36) is a quasi-convex problem which can be solved by a bisection search methodology. In the case of the convex UA region $\mathcal{A}_{\mathrm{UA}}$, we can solve problem (5.36) in a smarter way via the Charnes-Cooper transformation, which converts problem (5.36) into a convex problem; this will be demonstrated in the next subsection. However, in the AS case problem (5.36) is combinatorial and nonconvex. We will propose a polynomial-time algorithm that solves problem (5.36) exactly.

Optimization for the Unequal Amplitude Case

To describe the Charnes-Cooper transformation [103] for the design optimization problem in the UA case, let us define the following transformation

$$\boldsymbol{a} = \boldsymbol{z}/\xi \tag{5.37}$$

for some $\boldsymbol{z} \ge 0$ and $\xi > 0$. Then we can turn problem (5.36), with $\mathcal{A} = \mathcal{A}_{\mathrm{UA}}$, to

$$\min_{\boldsymbol{z}\in\mathbb{R}_+^N,\xi\in\mathbb{R}_+} \quad \frac{|s|_{\max}\sqrt{\delta^2\|\boldsymbol{z}\|_2^2+\sigma_\nu^2\xi^2}}{\boldsymbol{g}^T\boldsymbol{z}} \tag{5.38a}$$

$$\text{s.t.} \quad \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} \max_{i=1,\ldots,N} g_i z_i \le \boldsymbol{g}^T\boldsymbol{z}, \tag{5.38b}$$

$$\|\boldsymbol{z}\|_2 \le \sqrt{P_{\mathrm{T}}}\xi, \quad \boldsymbol{z} \le \sqrt{P_{\mathrm{PA}}}\mathbf{1}\xi, \tag{5.38c}$$

$$\xi > 0, \tag{5.38d}$$

where the objective function (5.38a) is the inverse of (5.36a). Since any feasible solution of problem (5.38) can be scaled by any positive number without affecting

the feasibility and objective value, we can impose without loss of generality an additional constraint that $\boldsymbol{g}^T\boldsymbol{z} = 1$. The resultant problem is

$$(\text{CE}-\text{UA}) \min_{\boldsymbol{z}\in\mathbb{R}^{\text{N}}_+,\xi\in\mathbb{R}_+} \quad |s|_{\max}\sqrt{\delta^2\|\boldsymbol{z}\|_2^2+\sigma_\nu^2\xi^2} \tag{5.39a}$$

$$\text{s.t. } \boldsymbol{g}^T\boldsymbol{z} = 1 \tag{5.39b}$$

$$\frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} \max_{i=1,\ldots,N} g_i z_i \le 1, \tag{5.39c}$$

$$\|\boldsymbol{z}\|_2 \le \sqrt{P_{\text{T}}}\xi, \quad \boldsymbol{z} \le \sqrt{P_{\text{PA}}}\mathbf{1}\xi, \tag{5.39d}$$

$$\xi \ge 0 \tag{5.39e}$$

where we relax the constraint $\xi > 0$ in (5.38d) to $\xi \ge 0$ in (5.39e). This relaxation is actually tight; if $\xi = 0$, then (5.39d) implies that $\boldsymbol{z} = \mathbf{0}$ which violates (5.39b). Problem (5.39) is a second-order cone program (SOCP), which can be efficiently solved by available algorithms [60, 61]. Once problem (5.39) is solved, an optimal solution of problem (5.36) can be recovered via (5.37).

Optimization for the Antenna Selection Case

The focus here is developing efficient algorithm for solving the AS design optimization problem. By substituting $\mathcal{A} = \mathcal{A}_{\text{AS}}$, the design optimization problem in (5.36) can be expressed as

$$\max_{\boldsymbol{a}\in\mathbb{R}^N_+,b\in\mathbb{R}_+} \quad \frac{\boldsymbol{g}^T\boldsymbol{a}}{|s|_{\max}\sqrt{\delta^2\|\boldsymbol{a}\|_2^2+\sigma_\nu^2}} \tag{5.40a}$$

$$\text{s.t. } \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} \max_{i=1,\ldots,N} g_i a_i \le \boldsymbol{g}^T\boldsymbol{a} \tag{5.40b}$$

$$\boldsymbol{a}^T\boldsymbol{a} \le P_{\text{T}} \tag{5.40c}$$

$$\boldsymbol{a} \in \{0, b\}^N \tag{5.40d}$$

$$0 \le b \le \sqrt{P_{\text{PA}}}. \tag{5.40e}$$

Problem (5.40) is a combinatorial optimization problem, which in the worst case involves enumerating 2^N possible solutions. However, as shown in the following proposition, problem (5.40) has a salient property that can be used to reduce the search space to a very manageable size.

Proposition 5.2 *Suppose that* $\boldsymbol{g}$ *is arranged in non-increasing order* $g_1 \geq g_2 \geq \ldots \geq g_N$. *Then problem* (5.40) *has an optimal solution in the form of*

$$\boldsymbol{a} = [\mathbf{0}_{i-1}^T,\ b\mathbf{1}_{k-i+1}^T,\ \mathbf{0}_{N-k}^T]^T \tag{5.41}$$

for some $0 \leq b \leq \sqrt{P_{\mathrm{PA}}}$ *and indices* $i, k \in \{1, \ldots, N\}$, $i \leq k$.

Proof: Let $\tilde{\boldsymbol{a}} \in \{0, \tilde{b}\}^N$ be a feasible solution of problem (5.40), where $0 \leq \tilde{b} \leq \sqrt{P_{\mathrm{PA}}}$. For this $\tilde{\boldsymbol{a}}$, let i be the index that indicates the first active antenna; i.e., to find the smallest i such that $\tilde{a}_j = 0$ for all $j = 1, \ldots, i-1$. Also, let n be the number of nonzero elements of $\tilde{\boldsymbol{a}}$ (or the number of active antennas specified by $\tilde{\boldsymbol{a}}$). From $\tilde{\boldsymbol{a}}$ we construct another point

$$\boldsymbol{a} = [\mathbf{0}_{i-1}^T, \tilde{b}\mathbf{1}_{k-i+1}^T, \mathbf{0}_{N-k}^T]^T, \tag{5.42}$$

and we argue that $\boldsymbol{a}$ is feasible and attains an objective value at least no worse than that of $\tilde{\boldsymbol{a}}$. First, by the nondecreasing order of $\boldsymbol{g}$, it can be verified that

$$\boldsymbol{g}^T\boldsymbol{a} \geq \boldsymbol{g}^T\tilde{\boldsymbol{a}} \tag{5.43}$$

$$\max_{i=1,\ldots,N} g_i a_i = \max_{i=1,\ldots,N} g_i \tilde{a}_i. \tag{5.44}$$

By noting that $\tilde{\boldsymbol{a}}$ satisfies (5.40b), and by substituting (5.43)-(5.44) into (5.40b), we show that $\boldsymbol{a}$ also satisfies (5.40b). Second, since $\boldsymbol{a}$ is just a permutation of $\tilde{\boldsymbol{a}}$, we have $\|\boldsymbol{a}\|_2 = \|\tilde{\boldsymbol{a}}\|_2 \leq P_{\mathrm{T}}$. Hence, $\boldsymbol{a}$ satisfies (5.40c). Third, the structure in (5.42) already implies that $\boldsymbol{a}$ satisfies (5.40d)–(5.40e). Consequently, we have proven that $\boldsymbol{a}$ is a feasible solution of problem (5.40). Moreover, by (5.43) and the result $\|\boldsymbol{a}\|_2 = \|\tilde{\boldsymbol{a}}\|_2$, we see from the objective function (5.40a) that the objective value achieved by $\boldsymbol{a}$ is greater than or equal to that by $\tilde{\boldsymbol{a}}$.

The derivations above further implies that if $\tilde{\boldsymbol{a}}$ is an optimal solution, then we can always construct a feasible solution $\boldsymbol{a}$ which takes the structure in (5.42) and yields an objective value no less than the optimal—which means that $\boldsymbol{a}$ must be optimal. Proposition 5.2 is therefore obtained, as desired. □

Now, by assuming that $\boldsymbol{g}$ is ordered in nonincreasing order and by using the

result in Proposition 5.2, we recast problem (5.40) as

$$\max_{i,k,b\in\mathbb{R}_+} \quad \frac{b\sum_{j=i}^{k} g_j}{|s|_{\max}\sqrt{\delta^2 b^2(k-i+1)+\sigma_\nu^2}} \tag{5.45a}$$

$$\text{s.t.} \quad \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} g_i \le \sum_{j=i}^{k} g_j \tag{5.45b}$$

$$b^2(k-i+1) \le P_{\mathrm{T}} \tag{5.45c}$$

$$0 \le b \le \sqrt{P_{\mathrm{PA}}} \tag{5.45d}$$

$$i,k \in \{1,\ldots,N\}, \quad i \le k. \tag{5.45e}$$

For a given pair of indices (i,k), problem (5.45) is feasible if and only if the constraint (5.45b) is satisfied. Supposing that (5.45b) holds, the corresponding optimal solution of b, denoted by $b_{i,k}^\star$, is easily shown to be

$$b_{i,k}^\star = \min\{\sqrt{P_{\mathrm{T}}/(k-i+1)}, \sqrt{P_{\mathrm{PA}}}\}. \tag{5.46}$$

Therefore, solving problem (5.45) amounts to checking the feasibility condition (5.45b), computing the optimal solution (5.46) for all pairs of indexes (i,k), and choosing the one that has the largest objective value. The resulting algorithm is shown in Algorithm 5.2. The complexity of the algorithm can be computed by noting that there are $N(N+1)/2$ candidate solutions to search; and that for each candidate, checking the feasibility and computing the objective value takes a complexity of $\mathcal{O}(N)$. Thus, the total complexity is $\mathcal{O}(N^3)$. The complexity can be reduced to $\mathcal{O}(N^2)$ by exploiting the similarity in computing $b_{i,k}^\star$ and $b_{i,k+1}^\star$. Details can be found in Algorithm 5.2.

As a final remark, the proposed algorithm can be easily modified to account for an additional constraint on the maximum number of active antennas. We omit the technical details here, as it is straightforward.

5.4.2 Robust Design with Deterministic Channel Uncertainty

In the deterministic channel uncertainty model, we are also interested in minimize the SER. In contrast with the stochastic model, here we are interested

Algorithm 5.2: An optimal search for problem (5.40).

input : P_{T}, P_{PA}, $|s|_{\max}$, $|s|_{\min}$, $\boldsymbol{g} = [|h_1|, \ldots, |h_N|]^T$, with $|h_1| \geq |h_2| \geq \ldots \geq |h_N|$.

$\rho^\star = -\infty$;
for $i \leftarrow 1$ **to** N **do**
  $g_{\mathrm{sum}} = 0$;
  for $k \leftarrow i$ **to** N **do**
    $g_{\mathrm{sum}} = g_{\mathrm{sum}} + g_k$;
    if $\frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} g_i \leq g_{\mathrm{sum}}$ **then**
      $b^\star_{i,k} = \min\{\sqrt{P_{\mathrm{T}}/(k-i+1)}, \sqrt{P_{\mathrm{PA}}}\}$;
      $\rho^\star_{i,k} = \frac{b^\star_{i,k} g_{\mathrm{sum}}}{|s|_{\max}\sqrt{\delta^2 (b^\star_{i,k})^2 (k-i+1)+\sigma_\nu^2}}$;
      if $\rho^\star_{i,k} \geq \rho$ **then**
        $b^\star = b^\star_{i,k}$; $\rho^\star = \rho^\star_{i,k}$;
        $i^\star = i$; $k^\star = k$;
      end
    end
  end
end
$\boldsymbol{a}^\star = [\mathbf{0}^T_{i^\star-1},\ b^\star \mathbf{1}^T_{k^\star-i^\star+1},\ \mathbf{0}^T_{N-k^\star}]^T$;

output: $\boldsymbol{a}^\star$, $\rho^\star$.

in the SER under the worst channel uncertainty:

$$\max_{\|\Delta \boldsymbol{h}\|_\infty \leq \epsilon} \Pr\{\hat{s} \neq s; \Delta \boldsymbol{h}\}. \tag{5.47}$$

Using the classical result in [102] we can upper bound the SER (5.47) by

$$\begin{aligned} &\max_{\|\Delta \boldsymbol{h}\|_\infty \leq \epsilon} \Pr\{\hat{s} \neq s; \Delta \boldsymbol{h}\} \\ \leq &(|\mathcal{S}|-1) \max_{\substack{s \neq s', \\ \|\Delta \boldsymbol{h}\| \leq \epsilon}} \Pr\{|y - \alpha s'| < |y - \alpha s|; \Delta \boldsymbol{h}\} \\ = &(|\mathcal{S}|-1) Q(\alpha \eta - 2\epsilon \|\boldsymbol{a}\|_1). \end{aligned} \tag{5.48}$$

Therefore, we can formulate the worst-case SER minimization problem as follows

$$\begin{aligned} \max_{\boldsymbol{a}\in\mathbb{R}_+^N,\alpha\in\mathbb{R}_+} \quad & \alpha\eta - 2\epsilon\|\boldsymbol{a}\|_1 \\ \text{s.t.} \quad & \alpha\mathcal{S} \subset \mathcal{D}(\bar{\boldsymbol{h}}\odot\boldsymbol{a}) \\ & \boldsymbol{a}^T\boldsymbol{a} \le P_{\mathrm{T}} \\ & \boldsymbol{a} \le \sqrt{P_{\mathrm{PA}}}\mathbf{1} \\ & \boldsymbol{a} \in \mathcal{A}. \end{aligned} \tag{5.49}$$

Following the same reformulation as (5.30)-(5.36), we rewrite (5.49) as

$$\begin{aligned} \max_{\boldsymbol{a}\in\mathbb{R}_+^N} \quad & \left(\frac{\eta}{|s|_{\max}}\boldsymbol{g} - 2\epsilon\mathbf{1}\right)^T\boldsymbol{a} \\ \text{s.t.} \quad & \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} \max_{i=1,\ldots,N} g_i a_i \le \boldsymbol{g}^T\boldsymbol{a} \\ & \boldsymbol{a}^T\boldsymbol{a} \le P_{\mathrm{T}} \\ & \boldsymbol{a} \le \sqrt{P_{\mathrm{PA}}}\mathbf{1} \\ & \boldsymbol{a} \in \mathcal{A}, \end{aligned} \tag{5.50}$$

where $\boldsymbol{g} = [|h_1|,\ldots,|h_N|]$, and the optimal solution $\alpha^\star$ of (5.49) is given by

$$\alpha^\star = \frac{1}{|s|_{\max}}\|\boldsymbol{h}\odot\boldsymbol{a}^\star\|_1 \tag{5.51}$$

with $\boldsymbol{a}^\star$ the optimal solution of (5.50).

It can be observed that (5.50) is a convex optimization problem if the set $\mathcal{A}$ is convex. In particular, for the UA set $\mathcal{A}_{\mathrm{UA}} = \mathbb{R}_+^N$, (5.50) is an SOCP which can be efficiently and optimally solved by available algorithms [60, 61].

Next, let us consider the AS case shown below

$$\begin{aligned} \max_{\boldsymbol{a}\in\mathbb{R}_+^N} \quad & \left(\frac{\eta}{|s|_{\max}}\boldsymbol{g} - 2\epsilon\mathbf{1}\right)^T\boldsymbol{a} \\ \text{s.t.} \quad & \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}} \max_{i=1,\ldots,N} g_i a_i \le \boldsymbol{g}^T\boldsymbol{a} \\ & \boldsymbol{a}^T\boldsymbol{a} \le P_{\mathrm{T}} \\ & \boldsymbol{a} \in \{0,b\}^N \\ & 0 \le b \le \sqrt{P_{\mathrm{PA}}}. \end{aligned} \tag{5.52}$$

It is straightforward to see that Proposition 5.2 can be applied to (5.52) as well. Therefore, we can turn problem (5.52) equivalently to

$$\max_{i,k,b\in\mathbb{R}_+} \quad b\sum_{j=i}^{k}\left(\frac{\eta}{|s|_{\max}}g_j - 2\epsilon\right) \tag{5.53a}$$

$$\text{s.t.} \quad \frac{2|s|_{\max}}{|s|_{\min}+|s|_{\max}}g_i \le \sum_{j=i}^{k} g_j \tag{5.53b}$$

$$b^2(k-i+1) \le P_{\mathrm{T}} \tag{5.53c}$$

$$0 \le b \le \sqrt{P_{\mathrm{PA}}} \tag{5.53d}$$

$$i,k \in \{1,\ldots,N\}, \quad i \le k. \tag{5.53e}$$

This problem can be solved optimally in the same way as that of (5.45). The algorithm can be adopted from Algorithm 5.2 by changing Line 8 of Algorithm 5.2 to

$$\rho^{\star}_{i,k} = b^{\star}_{i,k}\left(\frac{\eta}{|s|_{\max}}g_{\mathrm{sum}} - 2(k-i+1)\epsilon\right).$$

The complexity is also given by $\mathcal{O}(N^2)$.

5.5 Simulations

In this section, we use simulations to demonstrate the performance of CE precoding. Unless otherwise specified, we use the following simulation settings. The number of transmitting antennas is $N = 128$. The per-antenna power is set as $P_{\mathrm{PA}} = P_{\mathrm{T}}/16$, and the noise variance σ_ν^2 unity. We consider two types of channel models. In the first model, each element of the estimated channel vector $\bar{\boldsymbol{h}}$ follows a circular Gaussian distribution $\mathcal{CN}(0,1)$ in an i.i.d. manner. In the second model, ten elements of the channel have a line of sight (LOS) component and follow a distribution $\mathcal{CN}(10,1)$, while the rest of the channel elements follow $\mathcal{CN}(0,1)$. The channel elements are independently distributed. The second model is a simplified version of the experimental measurement results in [9, 10] which show that a large linear antenna array can experience large disparity in channel strength across antenna elements. For convenience of explaining simulation results to be shown, the two models above will be called channel

model one and channel model two, respectively. The simulation results are averages of 1000 independent channel realizations.

We compare the performance of the plain CE precoding scheme in (5.4), the proposed AS CE precoding scheme whose optimal design is obtained from Algorithm 5.2, and the proposed UA CE precoding scheme whose optimal design is achieved by solving the SOCP in problem (5.38) and (5.50). Also, as a minor technical remark, the constant α of each CE precoding scheme is computed by (5.34) and (5.51). We also benchmark the above constant envelope schemes against the MRT scheme [cf. (5.3)], which has non-constant envelope and does not take into account per-antenna power constraints. Their SER performances in the perfect CSI, stochastic channel uncertainty, and deterministic channel uncertainty cases will be considered in the first three subsections. Then, in the fourth subsection, we will examine the performance of the proposed exact phase recovery algorithm (Algorithm 5.1).

5.5.1 Performance in the Perfect CSIT Case

In Fig. 5.3, we present the SER performance of the various precoding schemes. The symbol constellation $\mathcal{S}$ is 16-QAM. The dash and solid lines represent the performances under channel model one and channel model two, respectively. We see that in both channel models, MRT outperforms all the CE precoding schemes; for example, the gap between MRT and UA CE precoding is about 2.5dB at the SER level 10^{-6}. This is not surprising, since CE precoding is a more stringent way of transmission than MRT. However, it should be recalled that the performance advantage of MRT comes at a price of higher hardware implementation costs and lower power efficiency with the RF amplifiers. Among the three CE precoding schemes, the UA strategy gives the best performance, and the plain strategy the worst. In channel model one, we observe that the performance of plain CE precoding can be quite close to that of UA CE precoding; the performance difference is just 1dB. Moreover, AS CE precoding shows a small performance gain compared to plain CE precoding (0.4dB at the SER level 10^{-6}). This observation suggests that uniform amplitudes in CE precoding may

provide near-optimal performance in scenarios where all the channel elements have similar magnitudes on average. However, in channel model two where some channel elements have stronger magnitudes, plain CE precoding does not perform well. Both the AS and UA CE precoding schemes are better than the plain CE precoding scheme, specifically, by about 4dB and 5dB respectively. Also, the performance difference between the AS and UA strategies is less than 1dB, which is small.

In Fig. 5.4 we show another set of SER results, where the symbol constellation is changed to 64-QAM and the other simulation settings are the same as those in Fig. 5.3. We observe similar results as in the previous 16-QAM case: In channel model one, the plain CE precoding scheme yields SER performance close to that of the AS and UA CE precoding schemes. In channel model two, the AS and UA schemes have similar SER performance and outperform the plain scheme. Moreover, in this 64-QAM case, we notice that MRT outperforms UA CE precoding by 3.6dB; in comparison, the gap in the 16-QAM case is 2.5dB (see Fig. 5.3). This suggests that for larger QAM sizes, MRT exhibits higher SNR gains than CE precoding. But please be noted that increasing the QAM size in MRT also incurs higher PAPR, and lower power efficiency with the RF amplifier.

We can explain why the performance gap between the MRT scheme and the CE precoding schemes widens with the QAM size, specifically, by analysis. As described previously, the MRT scheme in (5.3) has an effective channel gain

$$\alpha_{\mathrm{MRT}} = \sqrt{P_{\mathrm{T}}}\|\boldsymbol{h}\|_2. \tag{5.54}$$

An upper bound on the effective channel gain of the UA CE precoding scheme is given in the following proposition.

Proposition 5.3 *Assume the perfect CSIT case. The effective channel gain of the UA CE precoding scheme is upper-bounded by*

$$\alpha_{\mathrm{CE-UA}} \leq \frac{1}{|s|_{\max}}\sqrt{P_{\mathrm{T}}}\|\boldsymbol{h}\|_2. \tag{5.55}$$

Also, in a setting where there are no per-antenna power constraints and $\boldsymbol{h}$ *is i.i.d. zero-mean circular Gaussian, equality in* (5.55) *is attained with probability*

at least

$$1 - \frac{N}{2^{N-1}}. \tag{5.56}$$

The proof of Proposition 5.3 is relegated to Appendix 5.7.3. Comparing (5.54) and (5.55), we see that MRT is at least $|s|_{\max}$ times better than UA CE precoding in terms of the effective channel gain. Also, note that the argument above applies to the AS and plain CE precoding schemes as well, since the latters are more restrictive versions of the UA CE precoding scheme. Since a larger QAM size has a larger $|s|_{\max}$, we analytically confirm that the performance difference between MRT and CE precoding increases with the QAM size.

We then investigate how CE precoding performs in other problem sizes N in Fig. 5.5, where Fig. 5.5 (a), (b) and (c) show the performance of $N = 128$, 64, and 32, respectively. It can be seen that generally in channel model one, all CE methods perform similarly; in channel model two, both CE AS and CE UA are close and the plain CE shows some performance degradation compared to CE UA and CE AS. In addition, one can observe that all precoding methods have better performances in larger problem sizes than smaller problem sizes. For example, for all precoding methods, the powers to achieve SER level of 10^{-6} at problem size $N = 128$ are around 4dB less than those at $N = 64$. This observation confirms the necessity of using large antenna array and developing fast algorithms for it. On the other hand, it is also shown that CE precoding also works well when the problem size is not particular large.

We then investigate the number of active antennas in the AS CE scheme. Fig. 5.6 shows the distribution of the active number of antennas, where the settings are $N = 128$, 16-QAM and $P_{\mathrm{T}} = -4$dB. In the figure, the vertical solid line marks the average numbers of active antennas. For channel model one and two, the average number of antennas are 96 and 16, respectively; this means that the AS strategy has, on average, used 75% and 12.5% of the total 128 transmitting antennas in channel model one and two, respectively. We also see that the number of active antennas spreads from 85 to 110 in the first channel model, and is always 16 in the second channel model. This observation, together the previous SER result, indicates that the AS strategy is effective in

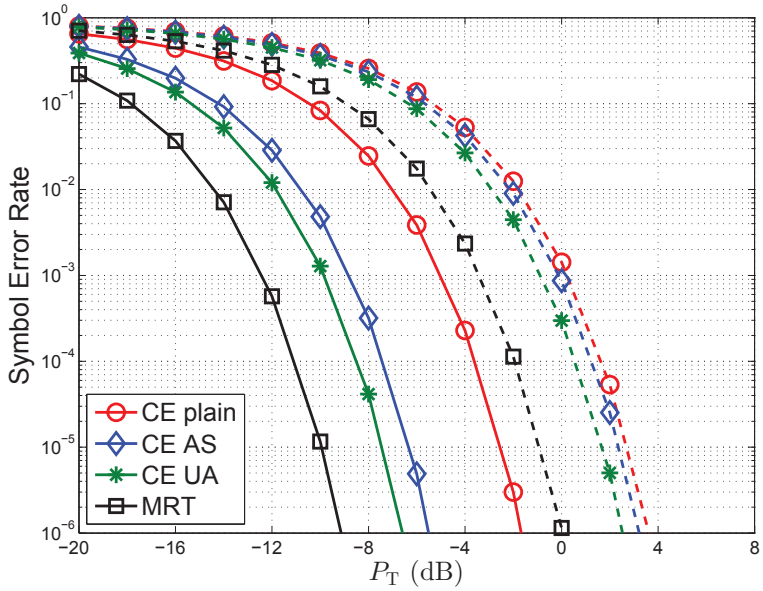


Figure 5.3: Symbol error rate in the perfect CSIT case. $N = 128$, 16-QAM. Dash line: channel model one; Solid line: channel model two.

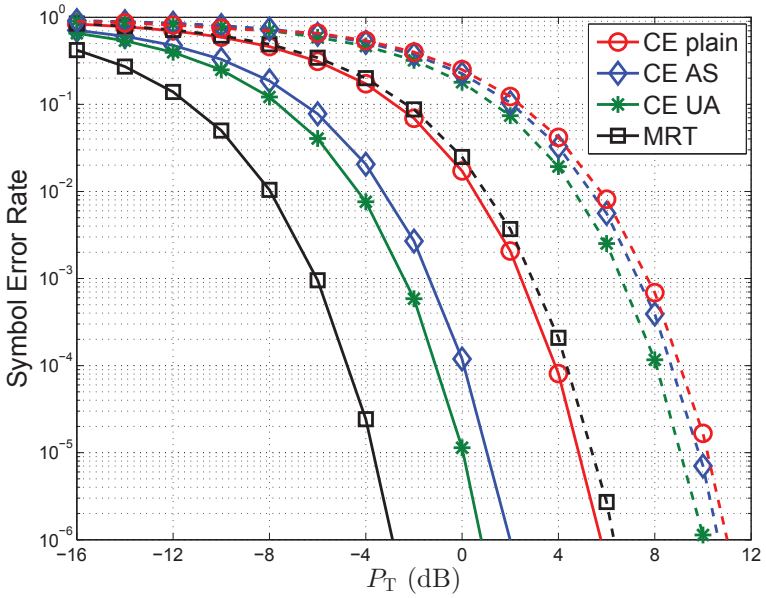
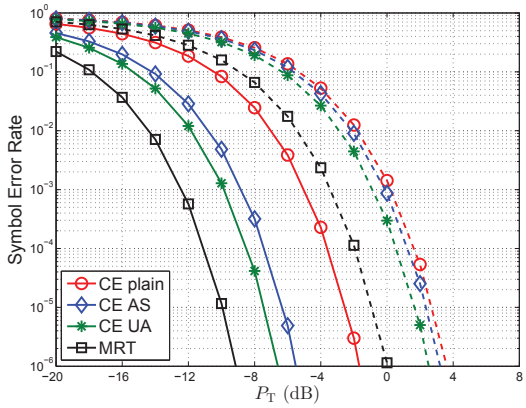
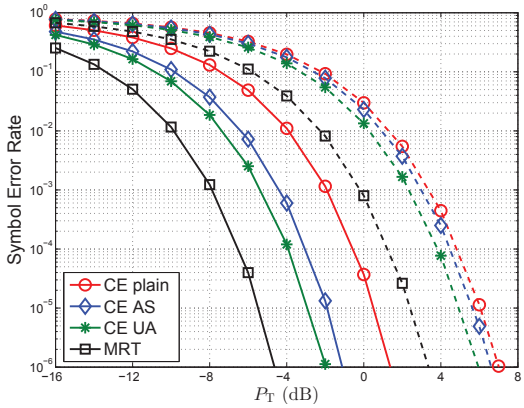


Figure 5.4: Symbol error rate in the perfect CSIT case. $N = 128$, 64-QAM. Dash line: channel model one; Solid line: channel model two.

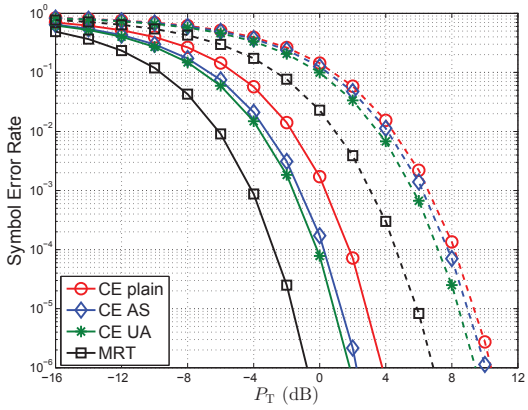
using a smaller number of the number of antennas to achieve comparable SER performance, especially in channels where some channel elements have stronger contributions over the others.



(a) $N = 128$



(b) $N = 64$



(c) $N = 32$

Figure 5.5: Symbol error rate in the perfect CSIT case. $N = 128$, 16-QAM. Dash line: channel model one; Solid line: channel model two. (a) is a replica of Fig. 5.3. In channel model two of (b) and (c), five and three channel elements are distributed as $\mathcal{CN}(10, 1)$ respectively, while the rest follow $\mathcal{CN}(0, 1)$.

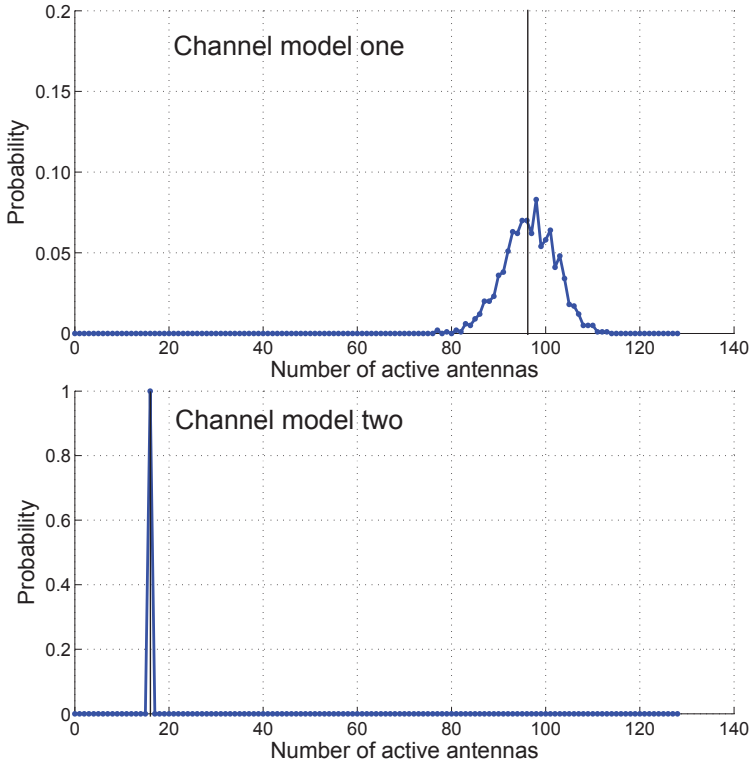


Figure 5.6: Distribution of the active antennas in the AS CE precoding scheme. $N = 128$, 16-QAM, and $P_{\mathrm{T}} = -4$dB.

5.5.2 Performance in the Stochastic Channel Uncertainty Case

We show the SER performance in the stochastic channel uncertainty case in Fig. 5.7. The simulation settings are the same as those in Fig. 5.3, and in addition the channel uncertainty level is $\delta^2 = 0.2$. Generally, the performances of all the precoding schemes in the stochastic channel uncertainty case are similar to those in the perfect CSIT case. Comparing Fig. 5.7 and its perfect CSIT counterpart in Fig. 5.3, we further observe that in channel model two, all the precoding schemes show almost the same performance as in the perfect CSIT case. In contrast, in channel model one, all the precoding schemes show some performance degradation compared to the perfect CSIT case. Furthermore, in the stochastic channel uncertainty case, the performance gaps between the plain and UA CE precoding schemes are larger: 2dB and 5.5dB for channel model one and channel model two, respectively. The AS CE precoding scheme is still within 1dB in comparison to the UA CE precoding scheme.

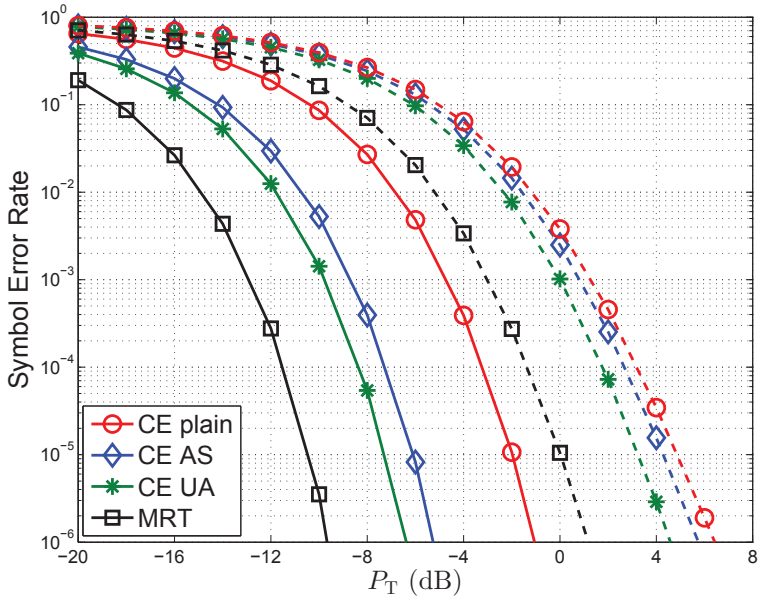


Figure 5.7: Symbol error rate in the stochastic channel uncertainty case. $N = 128$, 16-QAM, and $\delta^2 = 0.2$. Dash line: channel model one; Solid line: channel model two.

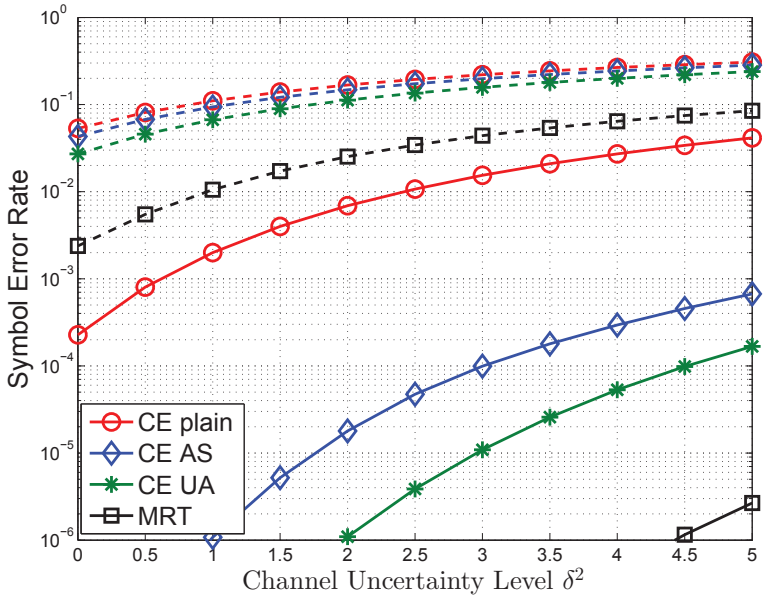


Figure 5.8: Symbol error rate versus the channel uncertainty level δ^2. $N = 128$, 16-QAM, and $P_T = -4$dB. Dash line: channel model one; Solid line: channel model two.

In Fig. 5.8 we show the SER performance with respect to the channel uncertainty level δ^2. It can be seen that in channel model one, the performances of all the CE precoding schemes are quite close to each other, irrespective of the

channel uncertainty level. In the channel model two, we observe that the UA and AS CE precoding scheme can withstand a larger channel uncertainty level than the plain CE precoding scheme under the same SER specification.

5.5.3 Performance in the Deterministic Channel Uncertainty Case

In this subsection, we investigate the performance of CE precoding in the deterministic uncertainty model. The simulation settings are the same as those in Fig. 5.3 with channel uncertainty level $\epsilon = 0.1$. The observations are similar to those in Fig. 5.7 of the stochastic uncertainty model. CE UA and CE AS are closed in both two channel models and are much better than plain CE in channel model two. We investigate how the SER scales with the channel uncertainty ϵ in Fig. 5.10. We can see that in channel model one, all CE methods are quite close to each other; in channel model two, the plain CE shows a much worse SER than CE AS and CE UA. CE AS and CE UA exhibit almost the same SER.

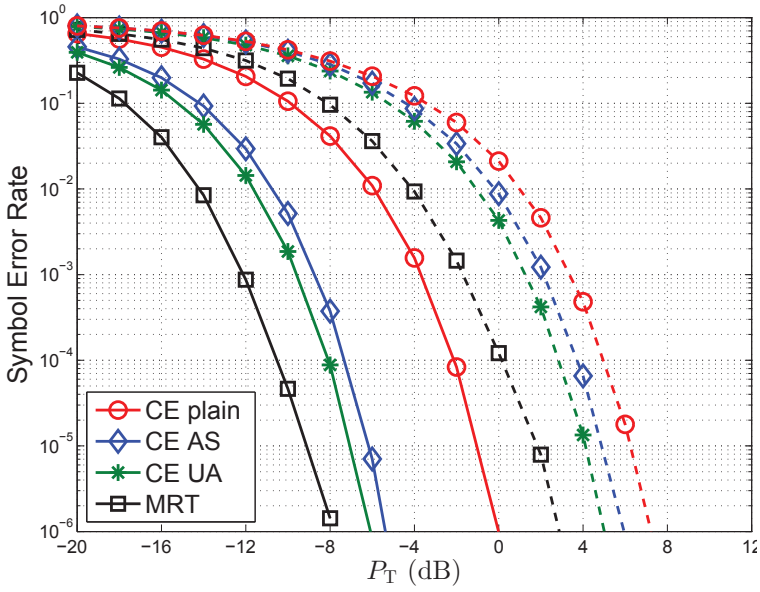


Figure 5.9: Symbol error rate in the deterministic channel uncertainty case. $N = 128$, 16-QAM, and $\epsilon = 0.1$. Dash line: channel model one; Solid line: channel model two.

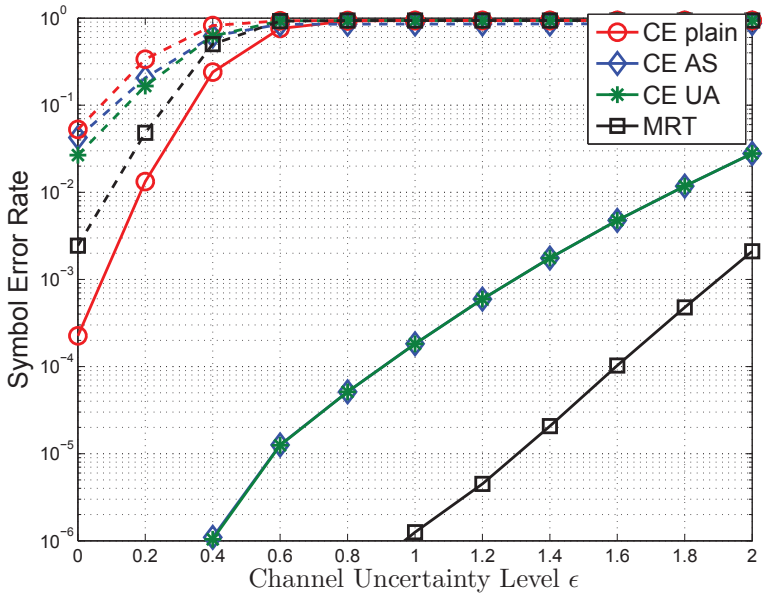


Figure 5.10: Symbol error rate versus the channel uncertainty level ϵ. $N = 128$, 16-QAM, and $P_{\mathrm{T}} = -4$dB. Dash line: channel model one; Solid line: channel model two.

5.5.4 Comparison between Exact Phase Recovery and Gradient Descent

In this subsection, we compare the proposed exact phase recovery algorithm (Algorithm 5.1) and the previous gradient descent method [47] in terms of accuracy and complexity. Channel model one with perfect CSIT is assumed, and the plain CE precoding scheme is adopted. We use the Armijo rule [104] for the step-size selection in the gradient descent method, and we stop the algorithm when the objective value of (5.9) is smaller than $\epsilon = 0.01$.

Fig. 5.11 compare the accuracies of the two algorithms by showing their SER performance. We can see that the SER performance of the gradient descent method is almost the same as that of the proposed exact phase recovery algorithm. This observation suggests that the gradient descent method should approach the exact solution. However, the gradient descent method is computationally more demanding than the exact phase recovery algorithm, as shown in Fig. 5.12 where the average numbers of floating point operations (FLOPs) of the two algorithms are compared. It can be seen that the proposed method is

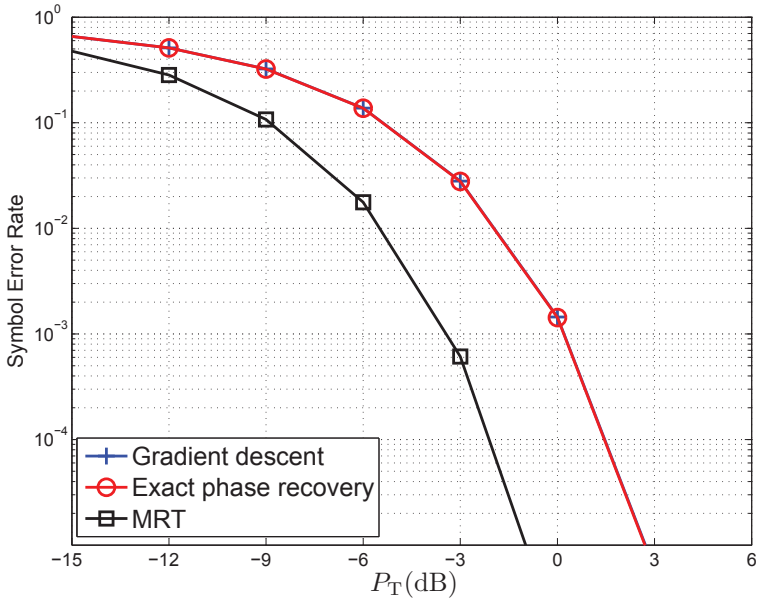


Figure 5.11: Symbol error rate comparison between two phase recovery algorithms. $N = 128$ and 16-QAM.

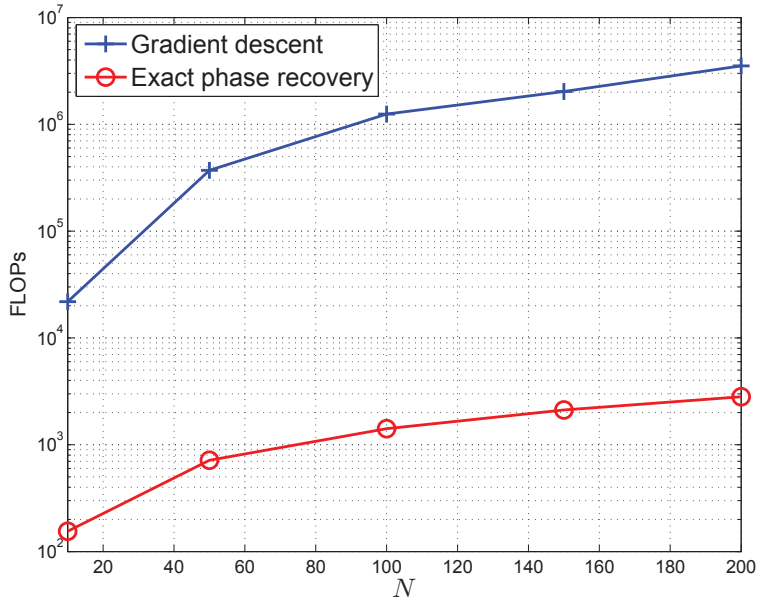


Figure 5.12: Complexity comparison between two phase recovery algorithms. $N = 128$ and 16-QAM.

much faster than the gradient descent method, and their FLOPs gap is much more significant at larger numbers of antennas.

5.6 Summary

In this chapter we investigated transceiver design problems in CE precoding for single-user MISO channels. A simple and efficient precoder algorithm for achieving exact phase recovery was devised, and optimal designs for CE precoding with antenna selection and unequal per-antenna power allocation were developed. Simulation results demonstrated that the proposed CE precoding designs can outperform the previous plain CE precoding scheme, especially when there are significant disparities among channel coefficients. The fundamental problem of characterizing the noise-free receive signal region was also solved in this chapter, where we complete the characterization results previously studied by the pioneering work by Mohammed and Larsson [47].

5.7 Appendix

5.7.1 Proof of Lemma 5.1

First, we show that any $z \in \mathcal{C}$ must satisfy $r_c \leq |z| \leq R_c$. For any $x \in \mathcal{A}$, $y \in \mathcal{B}$, we have that

$$|x+y| \leq |x| + |y| \leq R_a + r_b = R_c$$

and that

$$|x+y| \geq \max\{0, |x| - |y|\} \geq \max\{0, r_a - r_b\}.$$

This means that $r_c \leq |z| \leq R_c$ must hold.

Next, we show that any $z \in \mathbb{C}$, $r_c \leq |z| \leq R_c$ must lie in $\mathcal{C}$. The proof is by construction. We consider two cases, namely $|z| \geq R_a - r_b$, and $|z| < R_a - r_b$. For the case of $|z| \geq R_a - r_b$, set

$$y = r_b e^{\boldsymbol{\jmath}\phi_z}, \quad x = (|z| - r_b)e^{\boldsymbol{\jmath}\phi_z}.$$

It holds true that $z = x + y$, and that $y \in \mathcal{B}$. The question left is whether $x \in \mathcal{A}$. We first observe that $|x| \geq |z| - r_b \geq R_a - 2r_b \geq r_a$, where the last inequality is due to (5.18). Moreover, we have $|x| = |z| - r_b \leq R_c - r_b = R_a$. Hence, x lies in

$\mathcal{A}$. For the case of $|z| < R_a - r_b$, set

$$y = r_b e^{\boldsymbol{j}(\phi_z+\pi)}, \quad x = (|z| + r_b)e^{\boldsymbol{j}\phi_z}.$$

Again, since $z = x + y$ and $y \in \mathcal{B}$, we seek to show $x \in \mathcal{A}$. One can easily verify that $|x| = |z| + r_b \geq r_c + r_b \geq r_a$ and $|x| = |z| + r_b < R_a - r_b + r_b = R_b$. Hence, $x \in \mathcal{A}$ is true. We therefore conclude that any $z \in \mathbb{C}$, $r_c \leq |z| \leq R_c$, satisfies $z = x + y$ for some $x \in \mathcal{A}, y \in \mathcal{B}$, or equivalently $z \in \mathcal{C}$. It is also clear from the above proof that such (x, y) can be constructed via (5.19) and $x = y - z$.

5.7.2 Proof of Proposition 5.1

Assume without loss of generality that the variance of each element of $\boldsymbol{h}$ is one. Let $g_i = |h_i|$, $i = 1, \ldots, N$ and $0 \leq g_{(N)} < g_{(N-1)} \ldots < g_{(1)}$ denote the ordered statistics of $\{g_i\}_{i=1}^N$. As h_i is circular complex Gaussian distributed with zero mean and unit variance, g_i follows a Rayleigh distribution whose PDF is given by

$$f_g(g; \sigma) = \frac{g}{\sigma^2} e^{-g^2/2\sigma^2}, \quad g \geq 0 \tag{5.57}$$

with $\sigma^2 = 1/2$.

As $\{g_i\}_{i=1}^N$ are absolutely continuous i.i.d. random variables, it follows from [95] that the joint PDF of $\{g_{(i)}\}_{i=1}^N$ is given by

$$\begin{aligned} & f_{g_{(1)},g_{(2)},\ldots,g_{(N)}}(g_1, g_2, \ldots, g_N) \\ =& N! \prod_{i=1}^{N} f_g(g_i; 1/\sqrt{2}) \\ =& N! \prod_{i=1}^{N} 2g_i e^{-g_i^2} \end{aligned} \tag{5.58}$$

with domain $\mathcal{T} \triangleq \{(g_1, \ldots, g_N) \mid 0 \leq g_N < g_{N-1} \ldots < g_1\}$. Hence, we can express $\Pr\{r > 0\}$ as

$$\begin{aligned} & \Pr\{r > 0\} \\ =& \Pr\left\{ g_{(1)} - \sum_{i=2}^{N} g_{(i)} > 0 \right\} \\ =& \int_0^\infty \int_{g_N}^\infty \cdots \int_{g_3}^\infty \int_{\sum_{i=2}^N g_i}^\infty f_{g_{(1)},g_{(2)},\ldots,g_{(N)}}(g_1, g_2, \ldots, g_N) \mathrm{d}g_1 \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N. \end{aligned} \tag{5.59}$$

The integral in (5.59) with respect to g_1 can be computed as

$$\begin{aligned}&\int_{\sum_{i=2}^N g_i}^{\infty} f_{g_{(1)},g_{(2)},\ldots,g_{(N)}}(g_1,g_2,\ldots,g_N)\mathrm{d}g_1\\ =&N!\left(\prod_{i=2}^{N}2g_ie^{-g_i^2}\right)\int_{\sum_{i=2}^N g_i}^{\infty} f_g(g_1;1/\sqrt{2})\mathrm{d}g_1\\ =&N!\left(\prod_{i=2}^{N}2g_ie^{-g_i^2}\right)e^{-(\sum_{i=2}^N g_i)^2}.\end{aligned}\tag{5.60}$$

To prove the upper bound, let us define for $2\le j\le N$,

$$\begin{aligned}&\alpha_j(g_j;g_{j+1},\ldots,g_N)\\ =&\frac{N!}{(j-1)!(j-2)!}\left(\prod_{i=j}^{N}2g_ie^{-g_i^2}\right)e^{-(\sum_{i=j}^N g_i)^2}e^{-((j^2-j-2)g_j^2+2(j-2)(\sum_{i=j+1}^N g_i)g_j)},\end{aligned}$$

where $0\le g_N<g_{N-1}<\ldots<g_j$. We first show that

$$\int_{g_{j+1}}^{\infty}\alpha_j(g_j;g_{j+1},\ldots,g_N)\mathrm{d}g_j\le\alpha_{j+1}(g_{j+1};g_{j+2},\ldots,g_N)\tag{5.61}$$

is true for $0<g_N<g_{N-1}\ldots<g_{j+1}$. To see this, we need the following inequality

$$\int_c^{\infty}2xe^{-(ax^2+bx)}\mathrm{d}x\le\frac{1}{a}e^{-(ac^2+bc)},\tag{5.62}$$

where $a>0$, $b\ge0$ and $c\ge0$ are constant. This can be easily verified by noting that

$$2xe^{-(ax^2+bx)}\le2xe^{-(ax^2+bc)}$$

for $x\ge c$. The left hand side of (5.61) can be computed as

$$\begin{aligned}&\int_{g_{j+1}}^{\infty}\alpha_j(g_j;g_{j+1},\ldots,g_N)\mathrm{d}g_j\\ =&\frac{N!}{(j-1)!(j-2)!}\left(\prod_{i=j+1}^{N}2g_ie^{-g_i^2}\right)e^{-(\sum_{i=j+1}^N g_i)^2}\\ &\times\int_{g_{j+1}}^{\infty}2g_je^{-((j^2-j)g_j^2+2(j-1)(\sum_{i=j+1}^N g_i)g_j)}\mathrm{d}g_j.\end{aligned}\tag{5.63}$$

Applying (5.62) to the above equation via setting $a=j(j-1)$, $b=2(j-1)(\sum_{i=j+1}^N g_i)$, and $c=g_{j+1}$, we obtain the desired inequality in (5.63).

Substituting (5.60) into (5.59) and applying (5.61) for $j = 2, \ldots, N$, (5.59) can be upper bounded by

$$\begin{aligned}
&\int_0^\infty \int_{g_N}^\infty \ldots \int_{g_3}^\infty N! \left(\prod_{i=2}^{N} 2g_i e^{-g_i^2} \right) e^{-(\sum_{i=2}^{N} g_i)^2} \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N \\
=&\int_0^\infty \int_{g_N}^\infty \ldots \int_{g_3}^\infty \alpha_2(g_2; g_3, \ldots, g_N) \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N \\
\leq&\int_0^\infty \int_{g_N}^\infty \ldots \int_{g_4}^\infty \alpha_3(g_3; g_4, \ldots, g_N) \mathrm{d}g_3 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N \\
\leq&\int_0^\infty \alpha_N(g_N) \mathrm{d}g_N \\
=&\frac{1}{(N-1)!}.
\end{aligned}$$

To show the lower bound, we use the inequality $\left(\sum_{i=2}^{N} g_i\right)^2 \leq (N-1) \sum_{i=2}^{N} g_i^2$. Then, (5.59) is lower bounded by

$$\int_0^\infty \int_{g_N}^\infty \ldots \int_{g_3}^\infty N! \left(\prod_{i=2}^{N} 2g_i e^{-Ng_i^2} \right) \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N. \tag{5.64}$$

To compute the above integral, let us define for $2 \leq j \leq N-1$,

$$\beta_j(g_j; g_{j+1}, \ldots, g_N) = \frac{N!}{N^{j-2}(j-2)!} \left(\prod_{i=j}^{N} 2g_i e^{-Ng_i^2} \right) e^{-N(j-2)g_j^2}, \tag{5.65}$$

with domain $0 < g_N < g_{N-1} < \ldots < g_j$. It can be shown that

$$\int_{g_{j+1}}^\infty \beta_j(g_j; g_{j+1}, \ldots, g_N) \mathrm{d}g_j = \beta_{j+1}(g_{j+1}; g_{j+2}, \ldots, g_N) \tag{5.66}$$

is true for $0 < g_N < g_{N-1} < \ldots < g_{j+1}$. Indeed, we have

$$\begin{aligned}
&\int_{g_{j+1}}^\infty \beta_j(g_j; g_{j+1}, \ldots, g_N) \mathrm{d}g_j \\
=&\frac{N!}{N^{j-1}(j-1)!} \left(\prod_{i=j+1}^{N} 2g_i e^{-Ng_i^2} \right) \int_{g_{j+1}}^\infty f_g(g_j; 1/\sqrt{2N(j-1)}) \mathrm{d}g_j \\
=&\frac{N!}{N^{j-1}(j-1)!} \left(\prod_{i=j+1}^{N} 2g_i e^{-Ng_i^2} \right) e^{-N(j-1)g_{j+1}^2} \\
=&\beta_{j+1}(g_{j+1}; g_{j+2}, \ldots, g_N).
\end{aligned}$$

Now, applying (5.66) to (5.64) for $j = 2, \ldots, N-1$, (5.64) is equal to

$$\begin{aligned}
&\int_0^\infty \int_{g_N}^\infty \cdots \int_{g_3}^\infty \beta_j(g_2; g_3, \ldots, g_N) \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N \\
=&\int_0^\infty \beta_N(g_N) \mathrm{d}g_N \\
=&\frac{1}{N^{N-2}} \int_0^\infty f_g(g_N; 1/\sqrt{2N(N-1)}) \mathrm{d}g_N \\
=&\frac{1}{N^{N-2}}.
\end{aligned}$$

5.7.3 Proof of Proposition 5.3

In the perfect CSIT case, the design optimization problem under the UA strategy can be written as

$$\alpha_{\mathrm{CE-UA}} = \max_{\boldsymbol{a} \in \mathbb{R}_+^N} \ \frac{1}{|s|_{\max}} \|\boldsymbol{h} \odot \boldsymbol{a}\|_1 \tag{5.67a}$$

$$\text{s.t.} \ \frac{2|s|_{\max}}{|s|_{\min} + |s|_{\max}} \|\boldsymbol{h} \odot \boldsymbol{a}\|_\infty \leq \|\boldsymbol{h} \odot \boldsymbol{a}\|_1 \tag{5.67b}$$

$$\boldsymbol{a}^T \boldsymbol{a} \leq P_{\mathrm{T}} \tag{5.67c}$$

$$\boldsymbol{a} \leq \sqrt{P_{\mathrm{PA}}} \mathbf{1}; \tag{5.67d}$$

cf. the design optimization formulation in (5.35). Let us relax problem (5.67) by removing the constraints (5.67b) and (5.67d). By the Cauchy-Schwartz inequality, the optimal solution of the corresponding relaxed problem is shown to be

$$\boldsymbol{a}^\star = \frac{\sqrt{P_{\mathrm{T}}}}{\|\boldsymbol{h}\|_2} \boldsymbol{h}^*. \tag{5.68}$$

Since (5.68) achieves an objective value $\frac{\sqrt{P_{\mathrm{T}}}}{|s|_{\max}} \|\boldsymbol{h}\|_2$, we obtain the desired result in (5.55).

To prove the result in (5.56), consider the following inequality

$$2\|\boldsymbol{h} \odot \boldsymbol{a}^\star\|_\infty \leq \|\boldsymbol{h} \odot \boldsymbol{a}^\star\|_1. \tag{5.69}$$

As $|s|_{\max} \geq |s|_{\min}$, (5.69) implies that (5.67b) holds for $\boldsymbol{a}^\star$. Consequently, it can be easily shown that the occurrence of the event (5.69) implies that $\boldsymbol{a}^\star$ is an optimal solution to problem (5.67) without constraints (5.67d). Thus, if the probability of (5.69) being violated is $N/2^{N-1}$ under i.i.d. zero-mean Gaussian

channels, we obtain the desired result in (5.56). The problem can be boiled down to that of proving that

$$\Pr\{2\|\boldsymbol{g}\|_\infty - \|\boldsymbol{g}\|_1 > 0\} = \frac{N}{2^{N-1}}, \tag{5.70}$$

where $\boldsymbol{g} = [|h_1|^2, \ldots, |h_N|^2]^T$. The proof follows the same spirit as in that of Proposition 5.1. Here, we only show the key steps due to space limit. Without loss of generality, let us assume that the variance of each h_i is two. Then, all $g_i = |h_i|^2$ are exponentially distributed with the PDF given by

$$f_g(g) = \frac{1}{2}e^{-g/2}, \quad g \geq 0.$$

Let $0 \leq g_{(N)} < g_{(N-1)} < \ldots < g_{(1)}$ be the ordered statistics of $\{g_i\}_{i=1}^N$. Following the result in [95], the joint PDF of $\{g_{(i)}\}_{i=1}^N$ can be expressed as

$$\begin{aligned} & f_{g_{(1)},g_{(2)},\ldots,g_{(N)}}(g_1, g_2, \ldots, g_N) \\ =& N! \prod_{i=1}^{N} f_g(g_i) \\ =& \frac{N!}{2^N} e^{-\sum_{i=1}^N g_i/2} \end{aligned} \tag{5.71}$$

with domain $\mathcal{T} \triangleq \{(g_1, \ldots, g_N) \mid 0 \leq g_N < g_{N-1} < \ldots < g_1 < \infty\}$. Therefore,

$$\begin{aligned} & \Pr\{2\|\boldsymbol{g}\|_\infty - \|\boldsymbol{g}\|_1 > 0\} \\ =& \Pr\left\{ g_{(1)} - \sum_{i=2}^{N} g_{(i)} > 0 \right\} \\ =& \int_0^\infty \int_{g_N}^\infty \cdots \int_{g_3}^\infty \int_{\sum_{i=2}^N g_i}^\infty f_{g_{(1)},g_{(2)},\ldots,g_{(N)}}(g_1, g_2, \ldots, g_N) \mathrm{d}g_1 \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N. \end{aligned} \tag{5.72}$$

First, let us compute the integral in (5.72) with respect to g_1 as follows

$$\begin{aligned} & \int_{\sum_{i=2}^N g_i}^\infty f_{g_{(1)},g_{(2)},\ldots,g_{(N)}}(g_1, g_2, \ldots, g_N) \mathrm{d}g_1 \\ =& \frac{N!}{2^{N-1}} e^{-\sum_{i=2}^N g_i/2} \int_{\sum_{i=2}^N g_i}^\infty f_g(g_1) \mathrm{d}g_1 \\ =& \frac{N!}{2^{N-1}} e^{-\sum_{i=2}^N g_i}. \end{aligned} \tag{5.73}$$

Next, we define

$$\gamma_j(g_j; g_{j+1}, \ldots, g_N) = \frac{N!}{2^{N-1}(j-2)!} e^{-\sum_{i=j+1}^N g_i} e^{-(j-1)g_j}. \tag{5.74}$$

Then, it can be shown that

$$\int_{g_{j+1}}^{\infty} \gamma_j(g_j; g_{j+1}, \ldots, g_N) \mathrm{d}g_j = \gamma_{j+1}(g_{j+1}; g_{j+2}, \ldots, g_N). \tag{5.75}$$

This can be proved by

$$\begin{aligned} &\int_{g_{j+1}}^{\infty} \gamma_j(g_j; g_{j+1}, \ldots, g_N) \mathrm{d}g_j \\ =&\frac{N!}{2^{N-1}(j-2)!} e^{-\sum_{i=j+1}^{N} g_i} \int_{g_{j+1}}^{\infty} e^{-(j-1)g_j} \mathrm{d}g_j \\ =&\gamma_{j+1}(g_{j+1}; g_{j+2}, \ldots, g_N). \end{aligned} \tag{5.76}$$

Finally, by applying (5.73) and (5.75) from $j = 2$ to $N - 1$, we have

$$\begin{aligned} &\Pr\{2\|\boldsymbol{g}\|_\infty - \|\boldsymbol{g}\|_1 > 0\} \\ =&\int_0^{\infty} \int_{g_N}^{\infty} \cdots \int_{g_3}^{\infty} \gamma_2(g_2; g_3, \ldots, g_N) \mathrm{d}g_2 \ldots \mathrm{d}g_{N-1} \mathrm{d}g_N \\ =&\int_0^{\infty} \gamma_N(g_N) \mathrm{d}g_N \\ =&\int_0^{\infty} \frac{N!}{2^{N-1}(N-2)!} e^{-(N-1)g_N} \mathrm{d}g_N \\ =&\frac{N}{2^{N-1}}. \end{aligned}$$

We therefore have shown (5.70) and complete the proof.

 □

Chapter 6

Conclusion

6.1 Summary

With the increasing demand for faster wireless connectivity, wireless communication systems are evolving into MIMO systems equipped with tens or hundreds of antennas. Developing detection and precoding algorithms with low complexity but high performance in such systems is of paramount importance and is a tremendous challenge. The detection and precoding algorithms should also take into account hardware constraints for cheaper and more power-efficient implementations. In this thesis, we proposed several algorithms to meet this challenge in various communication scenarios, including

- MIMO detection: We developed a regularization optimization approach to tackle the out-of-bound symbol relaxation problem in lattice decoding. The proposed approach is based on the study of the LDR of the ML detector. We showed that the LDR approach, which is lattice decoding-based, is also connected to the semidefinite relaxation-based methods. We developed practical MIMO detectors by using the projected subgradient method for solving the LDR problem and the LRA-DF approximation method for reducing the complexity. Simulation results showed that the proposed methods have promising performance-complexity tradeoffs.

- Multi-user MISO broadcasting: We considered per-antenna power constraints in vector perturbation to reduce signal clipping and power back-off due to limited dynamic range of power amplifiers. We showed that the

proposed VP-PAPC schemes achieve the full transmit diversity in Gaussian fading channels. We developed an efficient approximation algorithm for VP-PAPC by using the LDR and the LRA-DF techniques. Simulation results demonstrated that VP-PAPC can avoid signal clipping in instantaneous power normalization and reduce the power back-off in short-term power normalization.

- Single-user MISO channels: We studied CE precoding which enables cheap but highly power-efficient power amplifiers. We provided a complete characterization of the noise-free receive signal region and developed an efficient and exact CE precoding algorithm. We formulated and solved efficiently robust QoS maximization problems in CE precoding where antenna selection or unequal per-antenna power allocation is allowed in the transmitter. Simulation results demonstrated that the performances of the proposed CE precoding designs are better than that of the previous plain CE precoding scheme.

6.2 Future Directions

While this thesis has proposed and analyzed several detection and precoding schemes, there are a few directions worth further investigations:

MIMO Detection

- This thesis focuses on hard decision-based MIMO detection. Since soft decision-based MIMO detection is important to practical systems (see the literature, such as [105] and the references therein), it would be meaningful to further study how the proposed LDR solutions may be used to provide soft decisions. Such a future direction seems feasible, since some existing soft decision generation tricks for LRA methods, such as bit flipping [106] and K-best SD [107], are also applicable to the LDR approach. It would also be interesting to investigate whether the structures of LDR may be exploited to provide efficient methods for soft decision generations.

- The present research endeavor demonstrated that the BR initialization scheme developed for the LDR LD approach plays a non-negligible role. While the LDR LD detector can converge to the optimum for any given feasible initialization, as suggested by convergence results in optimization theory, our extensive simulation results have shown that BR initialization can improve both the convergence speeds and sphere decoding complexities quite significantly. Thus, an interesting future direction would be to further understand the merits of BR initialization via analysis.

Vector Perturbation with Per-antenna Power Constraint

- This thesis considered exact CSIT in VP-PAPC. However, the CSIT in practice is not accurate due to noise corruption, feedback delay, and quantization. Thus, the study of the effect of inaccurate CSIT on VP-PAPC is very meaningful. Stochastic channel uncertainty in vector perturbation has been considered in [83] which shows that channel uncertainty can reduce the system diversity. It remains interesting to see how deterministic channel uncertainty affects the system performance.

- Long-term power normalization in vector perturbation is also of great importance. In long-term power normalization, the power normalization factor is taken to be the power of the unnormalized transmitting signal averaged over the channel and information vector. However, it is difficult to calculate this long term power normalization factor because for a given instance of the channel and information vector the computation of the power of the unnormalized transmitting signal is hard already. The introduction of PAPC will further complicate the computation. It would be interesting to develop efficient ways to compute the power normalization factor.

Constant Envelope Precoding

- This thesis considered CE precoding in single-user MISO channels. A very interesting extension would be to consider the scenarios of single-user MIMO channels and multi-user MISO downlink channels [48]. In these two scenarios, the noise-free receive signal region characterization and phase

recovery for CE precoding seem to be hard. In fact, the phase recovery problem can be written as a constant-modulus least squares problem, which has been reported to be NP-hard in general [108]. Developing efficient approximate solutions for the two scenarios mentioned above appears to be a meaningful and appealing future direction.

- Another interesting direction is robust design with unequal channel uncertainties across transmit antennas. In both stochastic and deterministic models, unequal channel uncertainties pose no challenge to the UA case; the resulting optimization problems are still convex problem. But for AS, the optimization problems are again likely to be NP-hard. In fact, our recent work [109] shows that in the deterministic uncertainty model the power minimization formulation, which is a variant of the QoS maximization formulation in this thesis, is NP-hard. It would be interesting to confirm the NP-hardness of the AS QoS maximization formulation and develop efficient approximate solutions.

Bibliography

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

[2] V. Tarokh, N. Seshadri, and A.R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

[3] P. Viswanath and D.N.C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[4] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.

[5] C.H. Doan, S. Emami, D.A. Sobel, A.M. Niknejad, and R.W. Brodersen, "Design considerations for 60GHz CMOS radios," *IEEE Commun. Mag.*, vol. 42, no. 12, pp. 132–140, Dec. 2004.

[6] S. Ranvier, J. Kivinen, and P. Vainikainen, "Millimeter-wave MIMO radio channel sounder," *IEEE Trans. Instrum. Meas.*, vol. 56, no. 3, pp. 1018–1024, Jun. 2007.

[7] S. Ranvier, M. Kyro, K. Haneda, T. Mustonen, C. Icheln, and P. Vainikainen, "VNA-based wideband 60GHz MIMO channel sounder with 3-D arrays," in *IEEE Radio and Wireless Symposium*, Jan. 2009, pp. 308–311.

[8] Samsung, "Samsung announces world's first 5G mmWave mobile technology," May 2013, available: `http://global.samsungtomorrow.com/?p=24093`.

[9] X. Gao, F. Tufvesson, O. Edfors, and F. Rusek, "Measured propagation characteristics for very-large MIMO at 2.6GHz," in *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2012, pp. 295–299.

[10] S. Payami and F. Tufvesson, "Channel measurements and analysis for very large array systems at 2.6GHz," in *European Conference on Antennas and Propagation (EUCAP)*, 2012, pp. 433–437.

[11] F. Rusek, D. Persson, B.K. Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[12] M.O. Damen, H. El Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2389–2402, Oct. 2003.

[13] B.M. Hochwald, C.B. Peel, and A.L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, 2005.

[14] N.D. Sidiropoulos, T.N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[15] A.D. Murugan, H. El Gamal, M.O. Damen, and G. Caire, "A unified framework for tree search decoding: rediscovering the sequential decoder," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 933–953, Mar. 2006.

[16] C.P. Schnoor and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Math. Programming*, vol. 66, pp. 181–191, 1994.

[17] Z. Xie, R.T. Short, and C.K. Rushforth, "A family of suboptimum detectors for coherent multiuser communications," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 4, pp. 683–690, May 1990.

[18] R. Lupas and S. Verdu, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 123–136, Jan. 1989.

[19] W.-K. Ma, T.N. Davidson, K.M. Wong, Z.-Q. Luo, and P.-C. Ching, "Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE Trans. Signal Process*, vol. 50, no. 4, pp. 912–922, Apr. 2002.

[20] P.H. Tan, L.K. Rasmussen, and T.J. Lim, "Constrained maximum-likelihood detection in CDMA," *IEEE Trans. Commun.*, vol. 49, no. 1, pp. 142 –153, Jan. 2001.

[21] Z. Mao, X. Wang, and X. Wang, "Semidefinite programming relaxation approach for multiuser detection of QAM signals," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4275 –4279, Dec. 2007.

[22] A. Wiesel, Y.C. Eldar, and S.S. Shitz, "Semidefinite relaxation for detection of 16-QAM signaling in MIMO channels," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 653 – 656, Sept. 2005.

[23] Y. Yang, C. Zhao, P. Zhou, and W. Xu, "MIMO detection of 16-QAM signaling based on semidefinite relaxation," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 797–800, Nov. 2007.

[24] A. Mobasher, M. Taherzadeh, R. Sotirov, and A.K. Khandani, "A near-maximum-likelihood decoding algorithm for MIMO systems based on semi-

definite programming," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3869–3886, Nov. 2007.

[25] H. El Gamal, G. Caire, and M.O. Damen, "Lattice coding and decoding achieve the optimal diversity-multiplexing tradeoff of MIMO channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 968–985, Jun. 2004.

[26] M. Taherzadeh and A. K. Khandani, "On the limitations of the naive lattice decoding," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4820–4826, Oct. 2010.

[27] J. Jaldén and P. Elia, "DMT optimality of LR-aided linear decoders for a general class of channels, lattice designs, and system models," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4765 –4780, Oct. 2010.

[28] A.K. Singh, P. Elia, and J. Jaldén, "Achieving a vanishing SNR gap to exact lattice decoding at a subexponential complexity," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3692–3707, Jun. 2012.

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

[30] C. Windpassinger and R.F.H. Fischer, "Low-complexity near-maximum-likelihood detection and precoding for MIMO systems using lattice reduction," in *Proc. Information Theory Workshop*, Munich, Germany, 31 March-4 April 2003, pp. 345–348.

[31] X. Ma and W. Zhang, "Performance analysis for MIMO systems with lattice-reduction aided linear equalization," *IEEE Trans. Commun.*, vol. 56, no. 2, pp. 309–318, Feb. 2008.

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

[33] L. Luzzi, G.R. Othman, and J. Belfiore, "Augmented lattice reduction for MIMO decoding," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2853–2859, Sept. 2010.

[34] S. Liu, C. Ling, and D. Stehlé, "Decoding by sampling: A randomized lattice algorithm for bounded-distance decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5933–5945, Sept. 2011.

[35] L. Luzzi, D. Stehlé, and C. Ling, "Decoding by embedding: Correct decoding radius and DMT optimality," *IEEE Trans. Inf. Theory*, vol. 59, pp. 2960 – 2973, 2013.

[36] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[37] V. Tarokh, H. Jafarkhani, and A.R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[38] A.B. Gershman, N.D. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten, "Convex optimization-based beamforming," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 62–75, May 2010.

[39] C.B. Peel, B.M. Hochwald, and A.L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[40] A. Tolli, M. Codreanu, and M. Juntti, "Linear multiuser MIMO transceiver design with quality of service and per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3049–3055, 2008.

[41] W. Yu and T. Lan, "Transmitter optimization for the multi-antenna downlink with per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.

[42] S.-R. Lee, J.-S. Kim, S.-H. Moon, H.-B. Kong, and I. Lee, "Zero-forcing beamforming in multiuser MISO downlink systems under per-antenna power constraint and equal-rate metric," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 228–236, Jan. 2013.

[43] M. Ding, M. Zhang, H. Luo, and W. Chen, "Leakage-based robust beamforming for multi-antenna broadcast system with per-antenna power constraints and quantized CDI," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5181–5192, Nov. 2013.

[44] T.M. Kim, F. Sun, and A.J. Paulraj, "Low-complexity MMSE precoding for coordinated multipoint with per-antenna power constraint," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 395–398, Apr. 2013.

[45] G. Dartmann, X. Gong, W. Afzal, and G. Ascheid, "On the duality of the max-min beamforming problem with per-antenna and per-antenna-array power constraints," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 606–619, Feb. 2013.

[46] F. Boccardi and G. Caire, "The p-sphere encoder: Peak-power reduction by lattice precoding for the MIMO Gaussian broadcast channel," *IEEE Trans. Commun.*, vol. 54, no. 11, pp. 2085 –2091, Nov. 2006.

[47] S.K. Mohammed and E.G. Larsson, "Single-user beamforming in large-scale MISO systems with per-antenna constant-envelope constraints: The doughnut channel," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3992–4005, Jun. 2012.

[48] S.K. Mohammed and E.G. Larsson, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 1059–1071, 2013.

[49] S.K. Mohammed and E.G. Larsson, "Constant-envelope multi-user precoding for frequency-selective massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 2, no. 5, pp. 547 – 550, Jul. 2013.

[50] P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *URSI International Symposium on Signals, Systems, and Electronics*, Sept. 1998, pp. 295–300.

[51] W. Su, Z. Safar, and K.J.R. Liu, "Full-rate full-diversity space-frequency codes with optimum coding advantage," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 229 –249, Jan. 2005.

[52] J.N. Laneman and G.W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.

[53] J. Jaldén and B. Ottersten, "On the complexity of sphere decoding in digital communications," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1474–1484, Apr. 2005.

[54] N.D. Sidiropoulos and Z.-Q. Luo, "A semidefinite relaxation approach to MIMO detection for high-order QAM constellations," *IEEE Signal Process. Lett.*, vol. 13, no. 9, pp. 525–528, Sept. 2006.

[55] J. Jaldén and B. Ottersten, "The diversity order of the semidefinite relaxation detector," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1406–1422, Apr. 2008.

[56] M. Kisialiou and Z.-Q. Luo, "Probabilistic analysis of semidefinite relaxation for binary quadratic minimization," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1906–1922, 2010.

[57] A.M.-C. So, "Non-asymptotic performance analysis of the semidefinite relaxation detector in digital communications," *Preprint*,

2010, available: http://www1.se.cuhk.edu.hk/~manchoso/papers/mimo_sdp_mpsk.pdf.

[58] W.-K. Ma, C.-C. Su, J. Jaldén, T.-H. Chang, and C.-Y. Chi, "The equivalence of semidefinite relaxation MIMO detectors for higher-order QAM," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 6, pp. 1038–1052, Dec. 2009.

[59] M. Kisialiou and Z.-Q. Luo, "Probabilistic analysis of semidefinite relaxation for binary quadratic minimization," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1906–1922, Mar. 2010.

[60] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, pp. 625–635, 1999.

[61] M.C. Grant and S.P. Boyd, *The CVX Users' Guide*, CVX Research, Inc., Sept. 2013.

[62] W.-K. Ma, C.-C. Su, J. Jaldén, and C.-Y. Chi, "Some results on 16-QAM MIMO detection using semidefinite relaxation," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, Mar. 2008, pp. 2673–2676.

[63] L. Babai, "On Lovasz' lattice reduction and the nearest lattice point problem," *Combinactorica*, vol. 6, no. 1, pp. 1–13, 1986.

[64] D. Wübben, D. Seethaler, J. Jaldén, and G. Matz, "Lattice reduction," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 70 –91, May 2011.

[65] A.K. Lenstra, H.W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261, pp. 515–534, 1982.

[66] D. Seethaler, J. Jaldén, C. Studer, and H. Bölcskei, "On the complexity distribution of sphere decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5754 –5768, Sept. 2011.

[67] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[68] J. Jaldén, C. Martin, and B. Ottersten, "Semidefinite programming for detection in linear systems-optimality conditions and space-time decoding," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, 2003, vol. 4, pp. IV – 9–12.

[69] D.P. Bertsekas, *Convex Analysis and Optimization*, Athena Scientific, 2003, with A. Nedić and A. E. Ozdaglar.

[70] S. Boyd and A. Mutapcic, "Subgradient methods," *Notes of EE364b, Standford University*, 2008.

[71] H. Najafi, M.E.D. Jafari, and M.O. Damen, "On adaptive lattice reduction over correlated fading channels," *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1224 –1227, May. 2011.

[72] D. Wübben, R. Böhnke, V. Kühn, and K.-D. Kammeyer, "Near-maximum-likelihood detection of MIMO systems using MMSE-based lattice reduction," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, Jun. 2004, vol. 2, pp. 798–802.

[73] A. Yener, R.D. Yates, and S. Ulukus, "CDMA multiuser detection: a nonlinear programming approach," *IEEE Trans. Commun.*, vol. 50, no. 6, pp. 1016 –1024, Jun. 2002.

[74] P.B. Stark and R.L. Parker, "Bounded-variable least-squares: an algorithm and applications," *Computational Statistics*, vol. 10, pp. 129–141, 1995.

[75] M. Emtiyaz, "Updating inverse of a matrix when a column is added/removed," 2008, available: `http://www.cs.ubc.ca/~emtiyaz/Writings/OneColInv.pdf`.

[76] D. Li and X. Sun, *Nonlinear Integer Programming*, Springer, 2006.

[77] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization*, MPS-SIAM Series on Optimization, 2001.

[78] J. Nocedal and Stephen J.W., *Numerical optimization*, Springer, 2006.

[79] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[80] T. Haustein, C. von Helmolt, E. Jorswieck, V. Jungnickel, and V. Pohl, "Performance of MIMO systems with channel inversion," in *IEEE VTC - Spring*, 2002, vol. 1, pp. 35 – 39.

[81] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electronics Letters*, vol. 7, no. 5, pp. 138–139, 1971.

[82] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Trans. Commun.*, vol. 20, no. 4, pp. 774–780, 1972.

[83] J. Maurer, J. Jaldén, D. Seethaler, and G. Matz, "Vector perturbation precoding revisited," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 315 –328, Jan. 2011.

[84] M. Taherzadeh, A. Mobasher, and A.K. Khandani, "Communication over MIMO broadcast channels using lattice-basis reduction," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4567 –4582, Dec. 2007.

[85] W.S. Chua, C. Yuen, and F. Chin, "A continuous vector-perturbation for multi-antenna multi-user communication," in *IEEE VTC - Spring*, Apr. 2007, pp. 1806 –1810.

[86] J. Park and B. Shim, "A vector perturbation based transmit diversity scheme for multiuser MIMO systems," in *Int. Symp. Personal Indoor and Mobile Radio Comm.*, Sept. 2009, pp. 3238 –3242.

[87] E.Y. Kim and J. Chun, "Optimum vector perturbation minimizing total mse in multiuser MIMO downlink," in *Proc. IEEE ICC*, Jun. 2006, vol. 9, pp. 4242 –4247.

[88] P. Lu and H.-C. Yang, "Vector perturbation precoding for MIMO broadcast channel with quantized channel feedback," in *Proc. IEEE Global Conf. Commun.*, Dec. 2009, pp. 1 –5.

[89] B. Lee and B. Shim, "A vector perturbation with virtual users for multiuser MIMO downlink," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, Mar. 2010, pp. 3406 –3409.

[90] J. Jaldén, J. Maurer, and G. Matz, "On the diversity order of vector perturbation precoding with imperfect channel state information," in *Proc. IEEE SPAWC*, Jul. 2008, pp. 211 –215.

[91] W. Banaszczyk, "New bounds in some transference theorems in the geometry of numbers," *Mathematische Annalen*, vol. 296, no. 1, pp. 625–635, 1993.

[92] C. Shepard, H. Yu, N. Anand, L.E. Li, T. L. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proc. ACM Int. Conf. Mobile Computing and Networking (MobiCom)*, Aug. 2012.

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

[94] S.C. Cripps, *RF Power Amplifiers for Wireless Communications*, Artech Publishing House, 1999.

[95] B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja, *A First Course in Order Statistics*, New York: Wiley - Interscience, 1992.

[96] T. Weber, A. Sklavos, and M. Meurer, "Imperfect channel-state information in MIMO transmission," *IEEE Trans. Commun.*, vol. 54, no. 3, pp. 543–552, Mar. 2006.

[97] G. Jöngren, M. Skoglund, and B. Ottersten, "Combining beamforming and orthogonal space-time block coding," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 611–627, Mar. 2002.

[98] M.B. Shenouda and T.N. Davidson, "On the design of linear transceivers for multiuser systems with channel uncertainty," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 1015–1024, Aug. 2008.

[99] N. Vučić, H. Boche, and S. Shi, "Robust transceiver optimization in downlink multiuser MIMO systems," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3576–3587, 2009.

[100] J. Wang and D.P. Palomar, "Worst-case robust MIMO transmission with imperfect channel knowledge," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3086–3100, 2009.

[101] Y. Guo and B.C. Levy, "Robust MSE equalizer design for MIMO communication systems in the presence of model uncertainties," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1840–1852, May 2006.

[102] J. Proakis, *Digital Communication*, McGraw-Hill Higher Education, fourth edition, 2001.

[103] A. Charnes and W. W. Cooper, "Programming with linear fractional functionals," *Naval Res. Logist. Quarter*, vol. 9, pp. 181–186, 1962.

[104] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[105] P. Fertl, J. Jaldén, and G. Matz, "Performance assessment of MIMO-BICM demodulators based on mutual information," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1366–1382, 2012.

[106] R. Wang and G.B. Giannakis, "Approaching MIMO channel capacity with soft detection based on hard sphere decoding," *IEEE Trans. Commun.*, vol. 54, no. 4, pp. 587–590, 2006.

[107] X.-F. Qi and K. Holt, "A lattice-reduction-aided soft demapper for high-rate coded MIMO-OFDM systems," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 305–308, 2007.

[108] A.M.-C. So, J. Zhang, and Y. Ye, "On approximating complex quadratic optimization problems via semidefinite programming relaxations," *Mathematical Programming*, vol. 110, no. 1, pp. 93–110, 2007.

[109] J. Pan and W.-K. Ma, "Antenna subset selection optimization for large-scale MISO constant envelope precoding," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, 2014, accepted.