Spring 2017

Coherent and non-coherent data detection algorithms in massive MIMO

Haider Ali Jasim Alshamary
University of Iowa

COHERENT AND NON-COHERENT DATA DETECTION ALGORITHMS IN MASSIVE MIMO

by

Haider Ali Jasim Alshamary

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Electrical and Computer Engineering
in the Graduate College of
The University of Iowa

May 2017

Thesis Supervisor: Assistant Professor Weiyu Xu

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Haider Ali Jasim Alshamary

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Electrical and Computer Engineering at the May 2017 graduation.

Thesis Committee: _____
Weiyu Xu,
Thesis Supervisor

_____
Soura Dasgupta

_____
Anton Kruger

_____
Raghu Mudumbai

_____
Jianfeng Cai

To whom is always with me.
To all my family.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Weiyu Xu for his limitless support and inspiration. His immense input brings this piece of work to be compatible and comprehensive. I am indebted to his insights and enthusiasm which carried my research to logical conclusions.

I would also like to extend my gratitude to my committee members, in particular, Professor Soura Dasgupta and Professor Anton Kruger for all the knowledge that I earned from their lectures and their comments toward my research. In addition, I would like to thank Cathy Kern and Dina Blanc who always go way beyond their duties to help me, whether it was inside or outside the ECE Department.

I extend my profound gratitude to all my friends and lab mates for their brotherly love, support, and company during my time in the university of Iowa. A special appreciation goes to my parents, my wife, and my big family in Iraq for believing in me and for all the values that they raised me with. I couldn't have come this far without you all.

ABSTRACT

In the past decade there has been a significant growth in the number of devices consuming data traffics. Billions of mobile data devices are now connected to the global wireless network. Real-time audio, video, and virtual reality applications require reliable wireless communications with high data throughput. One way to meet these requirements is increasing the number of transmit and/or receive antennas of the wireless communication systems. Massive multiple-input multiple-output (MIMO) has emerged as a promising candidate technology for the next generation (5G) wireless communications. Massive MIMO increases the spatial multiplexing gain and diversity gain by adding a large number of antennas to the base stations (BS) of wireless communication systems. However, designing efficient algorithms to decode transmitted signal with low complexity is a big challenge in massive MIMO. In this dissertation, we design and analyze novel algorithms to achieve near-optimal or optimal performance for coherent data detection, and joint channel estimation and signal detection in massive MIMO systems. The dissertation consists of three parts depending on the number of users at the transmitter side.

In the first part, we assume the channel state information is known at the receiver. We introduce a probabilistic approach to solve the problem of coherent signal detection using an optimized Markov Chain Monte Carlo (MCMC) algorithm. Two factors contribute to the speed of finding the optimal solution by the

MCMC detector: The probability of encountering the optimal solution when the Markov chain converges to the stationary distribution, and the mixing time of the MCMC detector. First, we compute the optimal value of the "temperature" parameter such that the MC encounters the optimal solution in a polynomially small probability. Second, we study the mixing time of the underlying Markov chain of the proposed MCMC detector.

In the second part, we consider optimal non-coherent signal detection for massive MIMO systems, when the channel state information is unknown at the receiver. We develop and analyze an optimal joint channel estimation and signal detection algorithm for massive (single-input multiple-output) SIMO wireless systems. We propose exact non-coherent data detection algorithms in the sense of generalized likelihood ratio test (GLRT). In addition to their optimality, these proposed tree search based algorithms provably have low expected complexity and work for general constellations. More specifically, despite the large number of the unknown channel coefficients for massive SIMO systems, we show that the expected computational complexity of these algorithms is linear in the number of receive antennas (N) and polynomial in channel coherence time (T). We prove that as $N \to \infty$, the number of tested hypotheses for each coherent block equals T times the cardinality of the constellation. Simulation results show that the optimal non-coherent data detection algorithms achieve significant performance gains (up to 5 dB improvement in energy efficiency) with low computational complexity.

In the third part, we consider non-coherent data detections for the uplink

transmissions of TDD massive MIMO systems. We propose an GLRT-optimal algorithm for joint channel estimation and data detection in massive MIMO systems. We show that the expected complexity of our algorithm grows polynomially in the channel coherence time (T). The proposed algorithm is novel in two aspects. First, the transmitted signal can be chosen from a general constellation, constant-modulus or nonconstant-modulus. Second, the algorithm offers the exact optimal solution with expected complexity polynomial in the coherent block interval. Simulation results demonstrate significant performance gains of our approach compared with suboptimal non-coherent detection schemes. To the best of our knowledge, this is the first algorithm which efficiently achieves GLRT-optimal non-coherent detections for massive MIMO systems with general constellations.

PUBLIC ABSTRACT

In the past decade there has been a significant growth in the number of devices consuming data traffics. Billions of mobile data devices are now connected to the global wireless network. Real-time audio, video, and virtual reality applications require reliable wireless communications with high data throughput. One way to meet these requirements is increasing the number of transmit and/or receive antennas of the wireless communication systems.

Massive multiple-input multiple-output (MIMO) has emerged as a promising candidate technology for the next generation (5G) wireless communications. Massive MIMO increases the spatial multiplexing gain and diversity gain by adding a large number of antennas to the base stations (BS) of wireless communication systems. However, designing efficient algorithms to decode transmitted signal with low complexity is a big challenge in massive MIMO.

In this dissertation, we design and analyze novel algorithms to achieve near-optimal or optimal performance for coherent data detection, and joint channel estimation and signal detection (JED) in massive MIMO systems. Our proposed algorithms decode the noisy received signal offering polynomial complexity in the coherent block interval. In addition, the transmitted signal can be chosen from any constellation including nonconstant-modulus constellations like 16-QAM. To the best of our knowledge, the proposed algorithms in this dissertation are the state-of-the-art to achieve JED for massive MIMO systems.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF ALGORITHMS

Algorithm

CHAPTER 1
INTRODUCTION

1.1 Motivation

The central aim of modern wireless communication systems is to provide better service quality, higher data rates and larger accessibility under any circumstance. The wide prevalence of modern wireless devices like smartphones, tablets, and laptops has fueled an extensive growth in wireless data traffic. According to Cisco's visual networking index (VNI) forecast, the global mobile data traffic between 2015 and 2020 is projected to increase 8-fold, and three-fourths of the world's mobile data traffic will be video [1]. Figure 1.1 illustrates the exponential increase in mobile data traffic. Wireless data traffic will grow at a compound annual growth rate (CAGR) of 53 percent from 2015 to 2020, reaching 30.6 exabytes per month by 2020.

In order to meet this explosive growth in data traffic and user density, developing new technologies for future wireless communication is required. These technologies need to improve the wireless systems throughput by increasing either the communication channel bandwidth or the spectral efficiency. Based on the United States' frequency allocation chart in 2016 [2], we can see that increasing the bandwidth is limited by the scarcity in the favourable communication frequency ranges and the paucity of the radio spectrum which is already over exploited. Adding antennas to the communication terminals is an effective way to increase the spectral

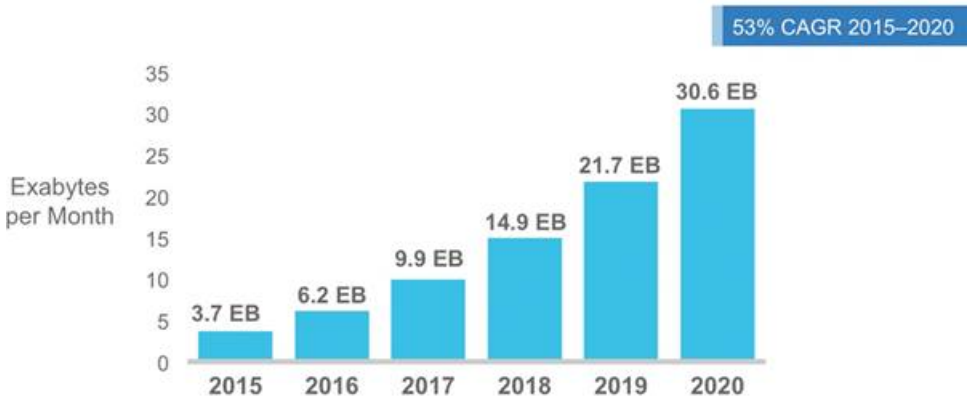


Figure 1.1: Global Mobile Data Traffic by 2020 [1].

efficiency. With the current fourth-generation (4G) mobile communication technology, improving the spectral efficiency is limited since 4G allows only up to 8 antennas at each base station (BS). The upcoming fifth-generation (5G) communication systems call for employing massive multiple-input multiple-output (MIMO) technologies. In fact, massive MIMO allows one to add hundreds or even thousands of antennas at a BS. This significant leap in the number of antennas at the wireless communication stations potentially increases the spectral efficiency. Thereby several autonomous users can simultaneously communicate with high throughput in the same time-frequency resources.

Massive MIMO is a promising candidate technology for future wireless communications. It promotes the attractive advantages of increasing the system capacity, and potentially reducing the transmitting power. However, in order to achieve the promises of massive MIMO, the receiver terminal needs to know the communication channel characteristics. Also, decoding the huge stream of the

transmitted data involves a tradeoff with the complexity of signal processing. Considering the unprecedented number of antennas at the BS, estimating the channel elements and detecting the transmitted signal will be a big challenge.

In this dissertation, we investigate near-optimal or optimal channel estimation and signal detection in massive MIMO systems. Our focus is on developing and proposing new efficient algorithms massive MIMO signal detection. In fact, we developed an asymptotically optimal detection algorithm for large scale coherent MIMO systems. In addition, we proposed optimal noncoherent signal detection algorithms for massive SIMO and MIMO systems with low complexity.

1.2 Massive Multi-User (MU) MIMO Systems

In the foreseeable future there will always be a sustainable growth in the indispensable mobile communication commodity and users' density. Accordingly, the data traffic carried by global wireless networks continues to increase in an exponential trend [3], [4]. This ever increasing demand for wireless data traffic and better service quality obligates the operators to invest in new technologies including improving reliable and high capacity links. A great amount of research has focused on increasing the capacity of wireless communication channels through spacial multiplexing. Massive MIMO is conceived as a fundamental technology for the next generation of wireless communication [5]. Massive MIMO relies on adding numerous antennas at one or both sides of wireless communications.

Adding multiple antennas at the receiver side was traditionally known as

a viable technique to leverage spatial diversity and mitigate multiple path fading. However, Foschini and Telatar [6], [7] show that utilizing multiple antennas at the transmitter and the receiver enlarges the wireless system spectral efficiency. It has been shown, under ideal channel estimation, the capacity of the MIMO system increases linearly with the $\min\{M, N\}$ for a scattering environment, where M, and N are the number of receive and transmit antennas respectively. Marzetta and Hochwald in [8], [9] extended the previous work on MIMO systems using a block fading channel model for unknown channel state information (CSI) at the receiver. In addition, they show that channel capacity is limited by the coherence time T and it is pointless to increase N beyond T since the link capacity will not change. For non-coherent channel, [10] finds a general expression for the channel capacity using geometric approach. [10] shows that the capacity of non-coherent systems reaches that of systems with perfect CSI when $T \to \infty$.

A pioneering work of Marzetta [11] in noncooperative cellular wireless systems reveals the idea of massive (large scale) MIMO systems by equipping the communication terminals with a huge number of antennas. [11] mathematically showed that the effect of fast fading and non-correlated noise is eliminated as the number of receive antennas approaches infinity. Since then, extensive research has been invested in massive MIMO. For example, massive MIMO systems' information-theoretic and propagation aspects are discussed in [12], [13]. Research on massive MIMO has also focused on many other aspects, including transmit and receive schemes, the effect of pilot contamination, energy efficiency, and channel estima-

tion as reviewed in [14], [15]. Massive MIMO wireless systems are proposed as a emerging new technology which reaps the benefits of the traditional MIMO systems on a much larger scale.

Massive MIMO technology can be used in a multi-cell multi-user MIMO (MU-MIMO) system, where multiple cells exist with one BS in each cell. These BSs are equipped with a large number of antennas, hence they send massive data traffic to a large number of users. The more antennas are equipped in each BS, the higher degree of freedom is offered. Accordingly, MU-MIMO technology deals with all the users in the whole network at the same time. The users are differentiated from each other according to their spatial signature. Also, the users are simultaneously served in the same time-frequency resources. MU-MIMO improves system performance without increasing the transmitted power since it enables high data rate.

Massive MU-MIMO systems can be demonstrated in a cellular network architecture, where every 7 cells form a cluster. Each BS's cell provides service for multiple users each of whom is equipped with single or multiple antennas. Figure 1.2 illustrates massive MU-MIMO system of L cells with K single antenna user terminals (UTs). Each BS handles N number of antennas. Furthermore, we assume known CSI at the BS through specific training scheme depending on which system protocol [1]is using.

[1]Time-division duplex (TDD): Duplexing mode where both communication parties share the same frequency band for transmitting and receiving the information. Frequency-division duplex (FDD): Duplexing mode where the communication parties use different frequency bands for transmitting and receiving the information.

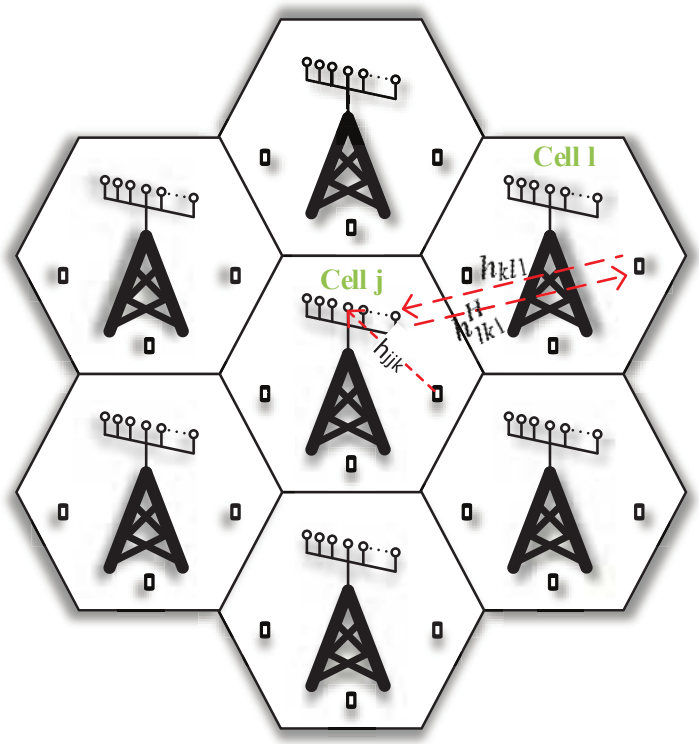


Figure 1.2: 7-cells cellular systems. Each BS equipped with N antennas and K users.

During the uplink (UL)[2], the channel model for MU-MIMO system can be represented as a linear combination of channel matrices from all the cells.

$$\mathbf{y}_j = \sum_{l=1}^{L} \mathbf{H}_{jl}\mathbf{x}_l + \mathbf{w}_j, \tag{1.1}$$

where $\mathbf{y}_j$ is the received vector at the j-th base station, $\mathbf{H}_{jl} = [\mathbf{h}_{jl1}\ \mathbf{h}_{jl2}\ \cdots\ \mathbf{h}_{jlK}]$ is the $N \times K$ channel matrix between the users of the l-th cell and the j-th BS, $\mathbf{x}_l$ is the transmitted vector from the users in cell l. $\mathbf{h}_{jl1}$ is usually modeled based on a

[2]It is the transmission mode in which the information data transmitted from the UTs to the base station, and it calls forward link as well.

deterministic correlation matrix and independent of fast-fading channel vector.

In downlink (DL)[3] phase, the received symbol by the m-th UT in the j-th cell is represented as:

$$y_{jm} = \sum_{l=1}^{L} \mathbf{h}_{jlm}^{H} \mathbf{s}_l + w_{jm}, \tag{1.2}$$

where $\mathbf{s}_l$ is the transmitted vector from the l-th BS, and w_{jm} is the noise at the m-th user of the j-th cell. In this setting, we assume TDD operation multiplexing mode which acquired by massive MIMO systems. Hence we assume channel reciprocity so that the downlink channel is the transpose of the uplink channel. The transmitted vector $\mathbf{s}_l$ is represented by the precoding matrix $A_l = [a_{l1}\ a_{l2} \cdots a_{lK}] \in \mathcal{C}^{N \times K}$ and the downlink transmitted data vector $\mathbf{x}_l$ which contains the information symbols for the K UTs in cell l,

$$\mathbf{s}_l = \sum_{k=1}^{K} a_{lk} x_{lk} = A_l \mathbf{x}_l. \tag{1.3}$$

The achievable rate of MU-MIMO network quantifies massive MIMO performance [13]. However, the ultimate achievable rate of the MU-MIMO systems is limited by interference, channel estimation error, and pilot contamination [16]. In addition, since most optimal performance detection algorithms obtain exponential complexity in system dimension, linear processing schemes are used instead in the down and up link signal detection like maximum ratio combining (MRC), Zero-Forcing (ZF), and minimum mean-square error (MMSE). These schemes for a large number of BS's antennas can demonstrate nearly-optimal performance [11].

[3]It is transmission mode in which the information data transmitted from the base station to the UTs, and it calls revers link as well.

1.3 Advantages of Massive MU-MIMO

Massive MU-MIMO is a potential technology for the future mobile communication since it increases wireless network throughput and reliability. We summarize the advantages of massive MU-MIMO as follows:

1. *Increasing channel capacity and system reliability*: For MU-MIMO systems with a large number of receive antennas at the BS and K number of users, we can increase the throughput without increasing the transmit power by simply increasing N and K. As $N \to \infty$, we can apply the law of large numbers to obtain a favorable propagation environment where the channel matrix columns pairwisely orthogonal. As a result of the favorable propagation, the channel capacity during the UL mode can be computed as follow,

$$C = \log_2 K(1 + N\rho_u),$$

   where ρ_u is the UL SNR, K is the multiplexing gain[4], and N is the array gain. So, as the system dimension increases, the spectral efficiency increases and equivalently system reliability.

2. *Reducing transmitting power, increasing energy efficiency*: Due to coherent combining, the transmitted power proportions inversely with the number of transmitted antennas $P_t \propto 1/n_t$. Thus a massive reduction in the transmit power can be obtained as the number of transmit antennas increases. Since the ef-

[4]The multiplexing gain is $\min\{K, N\}$.

fective SNR in massive MIMO systems is ($\rho_u N$), transmitted power can be reduced by a factor of two as the N doubles maintaining the same quality of service [13]. During the DL, massive MIMO systems able to intensify the transmitted signal in one region since all the focused emitted signals can be collected accumulatively in a certain spot. Nevertheless, interference would occur due to distraction in the transmitted signal, which can be solved by using some precoding techniques. Also, increasing the array aperture potentially leads to increased system resolution and effectively mitigates environmental and health concerns which related to high transmitted power of mobile communications [12].

3. *Reducing component cost and improving robustness*: In contrast to conventional MIMO systems, massive MIMO systems use hundreds of cheap milli-Watt amplifiers instead of multiple expensive high power amplifiers. Further, it eliminates the need for the coaxial cables which used to connect the BS components, and hence reduces the system implementation cost [12]. Using large number of amplifiers makes massive systems unconstrained by the accuracy and linearity of individuals. Malfunctions in several antennas will not effectively reduce system performance [15].

4. *Improving system security*: Cyber-security threats is a serious growing concern since large number of international jamming makes substantially harmful interference to the communication system and costs a lot of money. A MU-

MIMO systems with a sustainable large number of antennas, constantly have a large number of degree of freedom which can be used to cancel jamming [17]. Also, using joint channel estimation and cleaver decoding for a massive MIMO systems could be effectively able to render jamming problem [15].

5. *Mitigating latency and simplifying signal processing*: Latency in the transmit signals is one of the problems that limits the performance of any wireless communication system. In particular, under fast fading scenarios, the transmitted signal is prone to get trapped in a fading dip due to multiple paths traveling. Fading renders the transmitted signal into weak small signals. The surplus number of antennas at each BS of MU-MIMO systems enlarges drastically the degree of freedom since channel matrix has a very low rank (large nullspace). The large number of antennas causes channel hardening [11], thereby the random channel matrices in massive MIMO become nearly determinist [12]. Also, the interference will depend mainly on the number of DoF per UT and not directly on N [13]. In essence, using additional antennas reduces potential interference effects and averages out fading and thermal noise.

1.4 Challenges in Massive MIMO Systems

Even though increasing the number of antennas has many advantages, there will be numerous challenges associated with massive MIMO techniques. We summarize the challenges as follows:

1. **Channel Estimation:** Knowledge of the channel state information (CSI) is es-

sential for the sake of achieving the substantial advantages of massive MIMO systems [14]. During the UL mode, the K active users send training sequences to the BS which uses these code sequences to estimate the channel. Training blocks length is usually scaled with the number of active users, and their columns are mutually orthogonal in order to mitigate inter symbol interference [18]. With a limited number of users, this estimation technique can be viable. In contrast, for highly densified areas, like cells associated with massive MU-MIMO, the excessive number of unknown channel parameters presents a big challenge on accurately estimating the channel gains [16]. First, the pilot block will share a part of the coherent block which could be used for sending information data. In fast fading environments the channel parameters change rapidly. A tradeoff between estimating the channel accurately by sending more pilot signals or using this fraction of pilot sequences to transmit real data will kick off. Second, constricting perfectly orthogonal code sequences is limited in practice, and replaced with quasi-orthogonal sequences which impose a substantial harmful interference. In the case of conventional MIMO systems, differential modulation techniques, blind and semi-blind, and pilot based algorithms are used to solve the problem of acquiring the CSI [19]–[22]. Although these algorithms have improved the performance of traditional non-coherent MIMO systems, they are not optimal for massive time-varying non-coherent channels.

Massive MIMO paradigm relies on TDD multiplexing operation protocol in

order to jump over the formidable task of estimating the channel during the DL mode. In TDD the channel reciprocity is exploited where both the UL and DL share the same frequency spectrum with different time slots. Hence, the DL channel is just some version of the estimated channel at the BS. Acquiring channel reciprocity is one of the central important features of TDD protocol. During UL transmitting mode, BSs with a large number of antennas will estimate the channel instead of single antenna users. As a result, the active users detect the desired signal by using the effective channel gain through utilizing some precoding techniques during the DL mode. Now if the UL channel is estimated inaccurately, a serious down performance will be imposed on the DL signal detection. Accordingly, acquiring the CSI at the BS posts a fundamental limit for communication systems not only during the UL mode, but rather during DL mode as well.

As opposed to TDD, FDD uses different frequencies for UL and DL. FDD needs generally at least twice the spectrum required by TDD[5]. In addition, FDD uses wasteful guard bands which allow for adequate spectrum separation between the transmit and receive channels. Although FDD is widely used in global systems of mobile (GSM) communication, cellular telephony systems, it is difficulty to apply in special antenna techniques like MIMO and beamforming. With a lager number of antennas at the BS, it is more difficult to design antennas with enough broad bandwidth to cover the UL and DL

[5]TDD require $2K < T$ while FDD requires $N + K < T$.

spectrum. Consequently, FDD is an intangible operating mode for massive MIMO systems; however, it can be possible in certain circumstances [15]. Noncoherent signal detection is worth investigating substitutional method rather estimating the channel and then detecting the signal.

2. **Signal Detection Complexity Cost:** The essential limitation of any wireless communication system is the ability to correctly decode the transmitted signal. This was pointed out earlier by Shannon: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point" [23]. In addition, ubiquity of the service coverage requires the ability to process and successfully detect the weak signals which appear in scattering environments or in fast fading trends. However, when it comes to signal detection, there is always a tradeoff between the performance and the complexity cost of the detection algorithms. For MIMO wireless systems, the transmitted signal is prone to fading, noise, interference and other attenuation sources which contaminate the received symbols. These obstacles make finding exact decoding algorithm with a scalable-complexity difficult to achieve [24], [25]. In fact, it has been shown that MIMO signal detection problem is an NP-hard optimization problem [26], and the complexity of the detection is exponential in the number of decision variable.

   A computationally efficient way of solving the ILS problem was introduced by using tree base search algorithms like sphere decoder (SD) [27]–[32]. For

example, [29], [30] used the SD algorithm to achieve ML non-coherent signal detection for SIMO systems and constant modulus constellation[6]. The sphere decoder reduces the computational complexity by restricting the detection search to a subset of the signal space. The SD searches over the lattice points which are inside a sphere radius r. In fact, it demonstrates, roughly, a cubic complexity over a wide range of signal to noise ratio (SNR) and moderate system dimension [31]. SD has two drawbacks: First, it introduced to achieve optimal detection only for constant modulus constellation. Second, for high dimension systems with low SNR, the complexity of the sphere decoder grows exponentially with the coherent time [33]. Attempts to prune the complexity of the sphere decoder were introduced using some search space realization in [34], [35] based on the combination of branch and bound tree search. However, these relaxation techniques increase the complexity per each node to be cubic.

Channel state partition approaches can achieve GLRT-optimal non-coherent detection with worst-case computational complexity polynomial in the coherent block length T [36]–[39]. In general, this line of work of polynomial complexity sequence detection applies to single-input single-output (SISO) systems, where the channel state can be represented by only a scaler variable. When extended to MIMO systems with many receive antennas, the

[6]By constant modulus constellation, like QPSK or BPSK, we mean that $\{\forall s_i \neq s_j \in \Omega\ ,\ |s_i|^2 = |s_j|^2\}$, while for nonconstant, like 16-QAM, $|s_i|^2$ may or may not equal $|s_j|^2$.

state partition approach is either inapplicable or will result in an extremely high computational complexity.

The auxiliary angle approach is used to solve low-rank convex quadratic maximization problem in order to achieve exact ML noncoherent detection with polynomial-complexity [40]–[42]. However, the polynomial complexity only holds for equal-energy signal constellation, PSK, and it does not apply to nonconstant-amplitude modulations, PAM, and QAM. Unlike the previous work which used auxiliary-angle approach for constant modulus constellation, [43] develops ML noncoherent sequence detection algorithm for PAM which extended to QAM modulations. Although this algorithm attains polynomial expected complexity in the entire coherence interval, it was introduced mainly to PAM modulation and for SISO systems.

It was shown joint detection offers better performance that symbol-by-symbol detection [44]. Joint Detection requires decoding each symbol considering the characteristics of the rest of the symbols in the coherent block. For massive MIMO systems there will be unprecedented flow of wireless data traffic and hence fast, efficient detection algorithms are required. Due to the lack of exact joint detection algorithms, it is believed that using suboptimal detection algorithms is preferable to use for massive MIMO systems [15]. As a consequence, optimal performance algorithms which involve large scale data detection is a big concern and challenge.

In summary, most of the existing non-coherent detection algorithms achieve

suboptimal performance with acceptable complexity cost. The existing GLRT-optimal non-coherent detection algorithms are limited to constant-modulus modulations or for SISO systems. The question of whether we can achieve GLRT-optimal non-coherent detection for massive SIMO and MIMO systems with general constellations while having a computational complexity polynomial in T has been open.

3. **Pilot contamination:** Due to the limitation in constricting orthogonal pilot sequences, different cells could share the same frequency-time resources, and different users in the network share the same orthogonal pilots. Assigning the same pilot for different users results the phenomena of pilot contamination which badly influences the channel estimation and system performance [45], [46].

   Pilot contamination is particularly harmful to the performance and throughput of MU-MIMO systems since the number of required pilots scales with the number of cells, $N \times L$. In mobile environments the coherence time will be short, and it is impractical to use long training sequences. Consequently, using non-orthogonal uplink training sequences to estimate the channel causes a polluted estimated channel by other cells users' training signals. It has been shown by [11], [13] that pilot contamination in MU-MIMO is the only remaining impairment that cannot be eliminated as $N \to \infty$.

   We can illustrate the idea of pilot contamination using Figure 1.3 for 2 cells

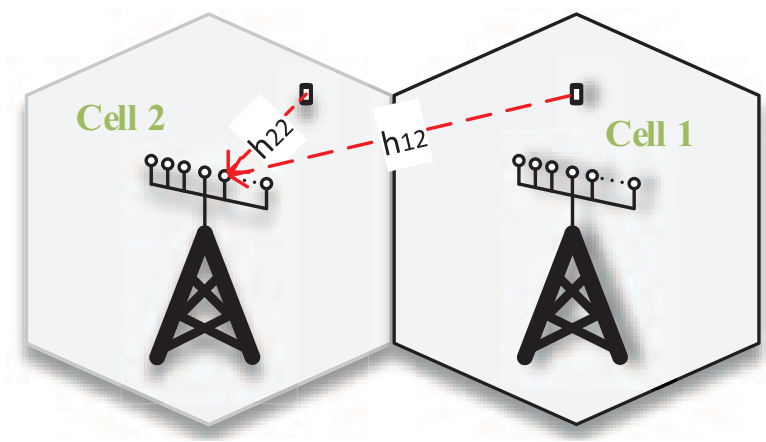


Figure 1.3: The estimated channel at BS2 $\hat{h}_{22} = c_1 h_{22} + c_2 h_{12} + c_3 w$.

system. Each cell is deployed with one BS and one user. Let us denote the channel between the i-th cell BS and the users in the j-th cell as $h_{i,j}$. Let assume that we assign the same pilot for both users. Then, the estimated channel at cell 2 is a linear combination of the channels between BS2 and the users in both cell 2 and 1, $\hat{h}_{22} = c_1 h_{22} + c_2 h_{12} + c_3 w$ where c_1, c_2 and c_3 are constant depend on the propagation factor.

For a cellular system with L cells, Let $\mathbf{p}_{jk}$ be the training vector transmitted by the k-th active user in the j-th cell. The received column vector by the m-th antenna of the l-th cell will be,

$$\mathbf{y}_{lm} = \sum_{j=1}^{L} \sum_{k=1}^{K} c h_{jlkm} \mathbf{p}_{jk} + \mathbf{w}_{lm},$$

where c is a constant including the transmitted power and the propagation factor, and $\mathbf{w}_{lm}$ is the additive noise. If we assume there is one UT at each BS,

the MMSE estimated channel at the $l-th$ cell is given by $\hat{\mathbf{h}}_{lj} = c_l \mathbf{p}^{\dagger} \mathbf{Y}_j$. Thus, In case of using the same training sequences, the estimated channel at the BS l will be some version of other cells estimated channel.

Based on [16], a simple version of the signal to interference and noise ratio (SINR) can be given as

$$\frac{1}{\underbrace{\frac{K\hat{L}}{cN}}_{Interference} + \underbrace{\frac{1}{\rho N}}_{Noise} + \underbrace{\alpha(\hat{L}-1)}_{Pilot\ contamination}},$$

where $\alpha \in (0,1)$, and $\hat{L} = 1 + \alpha(L-1)$. We can realize as $N \to \infty$, the first two terms in the denominator will vanish and the only performance limitation term that will stay is the pilot contamination.

1.5 Main Contributions

Our main contributions are summarized as follows:

Optimized Markov Chain Monte Carlo for Signal Detection in MIMO Systems: an Analysis of the Stationary Distribution and Mixing Time

Our main contributions in Chapter 2 are twofold: characterizing the stationary distribution, and bounding the mixing time of MCMC detectors. These results lead to an optimized MCMC detector for solving ILS problems. Firstly, we compute the optimal value of the "temperature" parameter, in the sense that the temperature has the desirable property that once the Markov chain has mixed to its stationary distribution, there is polynomially (and not exponentially) small

probability of encountering the optimal solution. This temperature is shown to be $O(\sqrt{\text{SNR}}/\ln(N))$, where SNR $> 2\ln(N)$, and N is the problem dimension. Secondly, we study the mixing time of the underlying Markov chain of the proposed MCMC detector. We find that, the mixing time of MCMC is closely related to whether there is a local minimum in the lattice structure of the ILS problems. Conventional wisdom proposes to use a temperature that is set to the noise standard deviation. On one hand, for some lattices without local minima, the mixing time of the Markov chain is independent of SNR, and grows polynomially in the problem dimension. On the other hand, for such a temperature choice, the mixing time of the Markov chain grows unbounded with SNR if the lattice has local minima.

We also study the probability of exist local minima in an ILS problem. For example, the probability of having local minima is $\frac{1}{3} - \frac{1}{\sqrt{5}} + \frac{2\arctan(\sqrt{\frac{5}{3}})}{\sqrt{5}\pi}$ for 2×2 Gaussian MIMO matrices. Simulation results indicate, when the system dimension $N \to \infty$, there seems to be at least one local minimum, but we do not have a rigorous proof of this phenomenon. Our theoretical and empirical results suggest that, to ensure fast mixing, for a fixed dimension N, very often the temperature for MCMC should be scaling at least as $\Omega(\sqrt{\text{SNR}})$. This is contrary to conventional wisdoms of using the standard deviation of channel noises [47], [48] as the temperature. Our simulation results show that the optimized MCMC detector achieves approximately ML detection in MIMO systems having a huge number of transmit and receive dimensions. We, however, have not been able to prove the scaling of the mixing time in terms of system dimension N for ILS problems.

Optimal Non-coherent Data Detection for Massive SIMO Wireless Systems with General Constellation: A Polynomial Complexity Solution

We can summarize our contribution in this chapter as follows:

- We propose GLRT-optimal joint channel estimation and data detection algorithms for SIMO systems with provably polynomial complexity. Our algorithms apply to general constellations, including nonconstant-modulus constellations. Our algorithms include a new breath-first tree-search algorithm which can find the GLRT-optimal non-coherent solution without requiring any predetermined search radius. To the best of our knowledge, these algorithms are the first set of efficient GLRT-optimal joint non-coherent data detection algorithms for massive SIMO systems using general constellations. We are thus able to provide the first set of error rate curve of the GLRT-optimal non-coherent detections for massive SIMO wireless systems with general constellations.

- Theoretically, we show that, under a large number of receive antennas in massive SIMO systems, both the sphere decoder algorithm and our own algorithms have expected computational complexity polynomial in the channel coherence time T and in the number of receive antennas. This is surprising, since we need to estimate a large number of unknown channel coefficients in massive SIMO systems. Moreover, we show that this is true as long as the number of recieve antennas grows polynomially in T.

- As a consequence of this work, we demonstrate the exact performance gap between the GLRT-optimal and suboptimal non-coherent data detection algorithms for massive SIMO systems. In fact, we demonstrate significant performance gains of our optimal non-coherent detection algorithms for nonconstant-modulus .

Efficient Optimal Joint Channel Estimation and Data Detection for Massive MIMO Systems

In this work, we demonstrate the exact performance gap between optimal and suboptimal non-coherent data detections in massive MIMO systems. We propose an efficient optimal joint channel estimation and data detection algorithm for massive MIMO systems with low expected complexity. Our algorithm is optimal in terms of the generalized likelihood ratio test (GLRT). We show that the expected complexity of our tree search based algorithm grows polynomially in the channel coherence time. In its essence, our approach is a branch-and-bound method on the residual energy of massive MIMO signals after projecting them onto certain subspaces. Moreover, our algorithm can provide benchmark performance against suboptimal low-complexity joint channel estimation and data detection algorithms. Simulation results demonstrate significant performance gains of our algorithm compared with suboptimal non-coherent detection schemes. To the best of our knowledge, this framework is the first GLRT-optimal non-coherent signal detection algorithm for massive MIMO systems with low computational complexity and optimal performance.

1.6 Organization

The rest of this thesis is organized as follows. Chapter 2 discusses the MCMC signal detection method for MIMO systems, and presents all the related analysis. Chapter 3 provides the joint channel estimation and data detection SIMO algorithms for massive SIMO systems, and shows the performance and the complexity with general constellation. Chapter 4 introduces our novel non-coherent signal detection massive MIMO algorithms for general constellation. Scope of future research is highlighted in 5.

CHAPTER 2
OPTIMIZED MARKOV CHAIN MONTE CARLO FOR SIGNAL DETECTION IN MIMO SYSTEMS: AN ANALYSIS OF THE STATIONARY DISTRIBUTION AND MIXING TIME

2.1 Introduction

In chapter 2 as a way to overcome the high complexity of SD's for ILS problems, we use approximate Markov Chain Monte Carlo (MCMC) detectors instead. MCMC method can provide the optimal solution asymptotically [49], [50] by performing the random walk according to the transition probability determined by the stationary distribution of a reversible Markov chain [50], [51]. MCMC detectors are proposed in [47], [48] for data detection in wireless communication; however, we optimize the performance of MCMC detector.

We introduce an optimized Markove Chain Monte Carlo (MCMC) technique for solving integer least-square (ILS) problems, which include Maximum Likelihood (ML) detection in lage scale multiple-input Multiple-output (MIMO) systems. ILS problem appears in many research areas, for example, communications, radar imaging, Monte Carlo second-moment estimation, bioinformatics, and lattice design [26], [52]. Sphere decoder (SD) is an efficient way for an exact solution of (ILS) problem. It is known that for a moderate problem size and a suitable range of Signal-to-Noise Ratios (SNR), SD has low computational complexity, which can be significantly smaller than an exhaustive search. However, for large scale MIMO system dimension, the average computational complexity of

traditional SD is exponential with system dimension [33].

Unlike the SD which perform well in high SNR regimes, MCMC detectors often suffer performance degradation in high SNR regimes. Moreover, the MCMC detectors in the literature were proposed mostly as practical heuristic detectors for digital communications. The theoretical understanding of MCMC detectors performance and complexity remains limited. For example, the mixing time (convergence rate) of the underlying Markov chains of these MCMC detectors, namely how fast these Markov chains mix to their stationary distributions, is not explicitly known. For the MCMC detectors in the literature [47], [48], the conditional transition probabilities of their underlying Markov chains were directly determined by the posterior likelihood of signal sequences [47], [48]. In other words, the standard deviation of channel noise was naturally applied as the "temperature" of these MCMC detectors [47], [48]. It was not clear either whether this choice of temperature is optimal, and what effect it will have on the performance and complexity of MCMC detectors.

To optimize MCMC detectors for ILS problems, we focus on two factors which contribute to the speed of finding the optimal solution by the MCMC detector: the probability of encountering the optimal solution when the Markov chain has converged to the stationary distribution, and the mixing time of the underlying Markov chain for the MCMC detector. In fact, if the optimal solution has a high probability in the stationary distribution, the MCMC detector will very likely encounter the optimal solution when its underlying Markov chain has mixed to

its stationary distribution. However, as we will see in this chapter, increasing the probability in the stationary distribution of the optimal solution often (even though not always) results in a slow mixing of the underlying Markov chain. Namely it takes a long time for the Markov chain to reach its stationary distribution. How to balance the mixing time and the stationary distribution for best performance of MCMC detectors is the main subject of this chapter.

The chapter is organized as follows. In Section 2.2 we present the system model that will be used throughout the chapter. The MCMC methods and background knowledge on Markov chain mixing time are described in Section 2.3. In Section 2.4 we analyze the probability of error for the ML detector. Section 2.5 treats the optimal selection of the temperature parameter α. Sections 2.6 and 2.7 derive bounds on the mixing time and discuss how to optimize MCMC parameters to ensure fast mixing. Simulation results are given in Section 2.8.

2.2 System Model

We consider a real-valued block-fading MIMO antenna system, with N transmit and N receive dimensions, with know channel coefficients. The received signal $\mathbf{y} \in \mathbb{R}^N$ can be expressed as

$$\mathbf{y} = \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\mathbf{x} + \boldsymbol{v} \ , \tag{2.1}$$

where $\mathbf{x} \in \Xi^N$ is the transmitted signal, and Ξ denotes the constellation set. To simplify the derivations we will assume that $\Xi = \{\pm 1\}$. $\boldsymbol{v} \in \mathbb{R}^N$ is the noise vec-

tor where each entry is Gaussian $\mathcal{N}(0,1)$ and independent identically distributed (i.i.d.), and $\mathbf{H} \in \mathbb{R}^{N\times N}$ denotes the channel matrix with i.i.d. $\mathcal{N}(0,1)$ entries. (In general, $\mathbf{H}$ can be any matrix, however, for analysis purposes we will focus on $\mathbf{H}$ with i.i.d. Gaussian elements.) SNR denotes the signal-to-noise ratio, namely,

$$\text{SNR} = \frac{\mathbb{E}\left\|\mathbf{y} - \boldsymbol{v}\right\|^2}{\mathbb{E}\|\boldsymbol{v}\|^2}. \tag{2.2}$$

For analysis purposes we will focus on the regime where SNR $> 2\ln(N)$, in order to get the probability of error of the ML detector to go to zero. Without loss of generality, we will assume that the all-minus-one vector was transmitted, $\mathbf{x} = -\mathbf{1}$. Therefore

$$\mathbf{y} = \boldsymbol{v} - \sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\mathbf{1}\ . \tag{2.3}$$

We are considering a minimization of the average error probability $P(\mathbf{e}) \triangleq P(\hat{\mathbf{x}} \neq \mathbf{x})$, which is obtained by performing Maximum Likelihood Sequence Detection (here simply referred to as ML detection) given by

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\Xi^N} \left\|\mathbf{y} - \sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\mathbf{x}\right\|^2 . \tag{2.4}$$

We emphasize that we focus on considering real-valued MIMO channels in this chapter. The results in this chapter can be adapted to complex-valued MIMO channels, constellations, and complex noises used in wireless systems in two ways. In one way, one can directly solve the same optimization problem provided in (2.4),

except for the fact that we allow $\mathbf{H}$ and $\mathbf{x}$ to take complex numbers. In another way, one can decompose the model in (2.1) into a real-valued model through the following decomposition:

$$\begin{pmatrix} \Re(\mathbf{y}) \\ \Im(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} \Re(\mathbf{H}) & -\Im(\mathbf{H}) \\ \Im(\mathbf{H}) & \Re(\mathbf{H}) \end{pmatrix} \begin{pmatrix} \Re(\mathbf{x}) \\ \Im(\mathbf{x}) \end{pmatrix} + \begin{pmatrix} \Re(\boldsymbol{v}) \\ \Im(\boldsymbol{v}) \end{pmatrix},$$

where $\Re(\cdot)$ and $\Im(\cdot)$ respectively denote the real and imaginary parts of a matrix.

2.3 MCMC Detector

One way of solving the optimization problem given in (2.4) is by using Markov Chain Monte Carlo (MCMC) detectors, which asymptotically converge to the optimal solution if the detector follows a reversible Markov chain [53]. We first describe our proposed MCMC detector based on reversible Markov chain, and then compare it with existing MCMC detectors in the literature.

2.3.1 Reversible MCMC Detector

In this chapter, we mainly focus on an MCMC detector which follows a reversible Markov chain and asymptotically converges to the stationary distribution [53]. Under the stationary distribution, the MCMC detector has a certain positive probability of visiting the optimal solution. This implies that if the MCMC detector is run for a sufficiently long time, it will be able to find the optimal solution to (2.4).

For this MIMO detection problem (2.4), the MCMC detector starts with a

certain N-dimensional feasible vector $\hat{\mathbf{x}}^{(0)}$ among the set $\{-1,+1\}^N$ of cardinality 2^N. Then the MCMC detector performs a random walk over $\{-1,+1\}^N$ based on the following reversible Markov chain. Assume that we are at time index l and the current state of the Markov chain is $\hat{\mathbf{x}}^{(l)} \in \{-1,+1\}^N$. In the next step, the Markov chain picks one random position index j uniformly out of $\{1,2,...,N\}$, and keeps the symbols of $\hat{\mathbf{x}}^{(l)}$ at other positions fixed. Then the MCMC detector computes the conditional probability of transferring to each constellation point at the j-th index. With the symbols at the $(N-1)$ other positions fixed, the probability that the j-th symbol adopts the value ω, is given by

$$p\left(\hat{\mathbf{x}}_j^{(l+1)} = \omega \,|\theta\right) = \frac{e^{-\frac{1}{2\alpha^2}\left\|\mathbf{y}-\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\hat{\mathbf{x}}_{j|\omega}\right\|^2}}{\sum\limits_{\hat{\mathbf{x}}_{j|\tilde{\omega}}\in\Xi} e^{-\frac{1}{2\alpha^2}\left\|\mathbf{y}-\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\hat{\mathbf{x}}_{j|\tilde{\omega}}\right\|^2}}, \tag{2.5}$$

where $\hat{\mathbf{x}}_{j|\omega}^T \triangleq \left[\hat{\mathbf{x}}_{1:j-1}^{(l)}, \omega, \hat{\mathbf{x}}_{j+1:N}^{(l)}\right]^T$, $\theta = \{\hat{\mathbf{x}}^{(l)}, j, \mathbf{y}, \mathbf{H}\}$ and $\alpha > 0$ is a tunable "temperature" parameter. So conditioned on the j-th position is chosen, the MCMC detector will with probability $p\left(\hat{\mathbf{x}}_j^{(l+1)} = \omega \,|\theta\right)$ transition to ω at the j-th position index. The initialization of the symbol vector $\hat{\mathbf{x}}^{(0)}$ can either be chosen randomly or as other heuristic solutions. Note that for a general constellation set Ξ, the procedure above, which summarized in Algorithm 1, still applies by replacing $\{-1,+1\}$ with Ξ.

For this type of MCMC detector, we care about the probability that such an algorithm encounters the true transmitted signal within a certain number of iterations. In general, determining this probability within a certain number of iterations

Algorithm 1: MCMC detector based on reversible Markov chain.

Input: $\mathbf{y}$, $\mathbf{H}$, initialization vector $\hat{\mathbf{x}}^{(0)}$, decision vector $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(0)}$ and the number of iterations n, set loop index $l = 0$
Output: The transmitted signal $\hat{\mathbf{x}}$

1. Pick a uniformly random position index j out of $\{1, 2, ..., N\}$.
2. Keep the symbols of $\hat{\mathbf{x}}^{(l)}$ at the $(N-1)$ other positions fixed, transition the j-th symbol of $\hat{\mathbf{x}}^{(l)}$ to ω with probability $p\left(\hat{\mathbf{x}}_j^{(l+1)} = \omega \,|\theta\right)$ specified in (2.5), for every $\omega \in \Xi$
3. Denote the new vector by $\hat{\mathbf{x}}^{(l+1)}$
4. If $\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}^{(l+1)}\|_2^2 < \|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|_2^2$, update $\hat{\mathbf{x}} := \hat{\mathbf{x}}^{(l+1)}$.
5. If $l \neq n-1$ go to (1), else stop the algorithm and present the output.

is difficult. However, things are relatively easy when we assume that the underlying Markov chain has mixed to the steady state distribution, which is easy to write down because it is easy to determine the probability P_{en} of encountering the true transmitted signal in steady state. Therefore, an upper bound on the expected time to find the optimal solution is determined by the mixing time (the time it takes to to reach the steady state) of the Markov chain, and the inverse of the probability P_{en} of encountering the true transmitted signal in the steady state.

We remark that α represents a tunable positive parameter which controls the mixing time of the Markov chain, and this parameter is also sometimes called the "temperature". If we let $\alpha \to \infty$, the MCMC detector is a just a uniform random walk in the signal space, namely in each iteration the detector chooses constellation points with equal probabilities, and the underlying Markov chain quickly mixes to

its steady state [51]. When α is close to 0, the MCMC detector will eventually "reside" at the optimal solution, but it may take a very long time to get there from an initial suboptimal signal vector.

Now the smaller α is, the larger the stationary probability for the optimal solution will be, and the easier it is for the MCMC detector to find the optimal solution in the stationary distribution. On the other hand, as α gets smaller, it often takes a long time for the Markov chain to converge to its stationary distribution. In fact, as we will show in this chapter, there is often a lower bound on α, in order to ensure the fast mixing of the Markov chain to its stationary distribution.

2.3.2 Comparisons with Conventional MCMC Detectors

Our proposed MCMC detector is different from conventional MCMC detectors [47], [48]. In [47], [48], the conditional transition probabilities of the underlying Markov chains were directly determined by the posterior likelihood of data sequences. In other words, the "temperature" α of these MCMC detectors is directly set as the standard deviation of channel noise [47], [48]. In this chapter, however, we have the freedom of optimizing this temperature parameter α.

Our proposed method is also very different from simulated annealing techniques where the temperature is slowly reduced until the detector converges to an acceptable solution. In our MCMC detector, the temperature is set as a *fixed* value, and we care about a fast mixing of the underlying Markov chain to a stationary probability distribution and a big enough probability of encountering the trans-

mitted signal in steady state. Unlike simulated annealing, we do not require the MCMC random walk to converge to the optimal solution in the end, but instead we would like the MCMC random walk to *encounter* the optimal solution as soon as possible. This is because in the MCMC detector, we always record the sequence which has the lowest distance metric.

2.3.3 Mixing Time

It is not hard to see that the Markov chain of MCMC detector is reversible and has 2^N states with the stationary distribution $e^{-\frac{1}{2\alpha^2}\left\|\mathbf{y}-\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\hat{\mathbf{x}}\right\|^2}$ (without normalization) for a state $\hat{\mathbf{x}}$. The $2^N \times 2^N$ transition matrix is denoted by P, and the element $P_{i,j}$ in the i-th ($1 \leq i \leq N$) row and j-th ($1 \leq j \leq N$) column is the conditional probability of transferring to state j given that the current state is i. So each row of P sums up to 1 and the transition matrix after t iterations is P^t. We denote the vector for the stationary distribution as π. Then for an $\epsilon > 0$, the mixing time $t(\epsilon)$ is a parameter describing how long it takes for the Markov chain to get close to the stationary distribution [51], namely,

$$t_{mix}(\epsilon) := \min\{t : \max_{\tilde{\mathbf{x}}} \|P^t(\tilde{\mathbf{x}}, \cdot) - \pi\|_{TV} \leq \epsilon\},$$

where $\|\mu - \nu\|_{TV}$ is the usual total variation distance between two distributions μ and ν over the state space $\{+1, -1\}^N$.

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{\mathbf{z} \in \{+1,-1\}^N} |\mu(\mathbf{z}) - \nu(\mathbf{z})|.$$

The mixing time is closely related to the spectrum of the transition matrix P. More precisely, for a reversible Markov chain, its mixing time is generally small when the gap between the largest and the second largest eigenvalue of P, namely $1 - \lambda_2$, is large. The inverse of this gap, $\frac{1}{1-\lambda_2}$, is called the relaxation time.

2.3.4 Sequential Markov Chain Monte Carlo Detectors

For simplicity of implementations, we also consider a sequential MCMC detector. The only difference between sequential MCMC detectors and reversible MCMC detectors is the way they choose the position index to update. Sequential MCMC detectors can have many *block iterations* which can be define for the sequential MCMC detector as an sequential update of all the N indices $\{1, \ldots, N\}$ in the estimated symbol vector $\hat{\mathbf{x}}$, starting from $j = 1$ to $j = N$. Namely, in one block iteration, we update N indices. For each index j, the updating rule for the sequential MCMC detector is the same as the reversible MCMC detector. We remark, however, that the mixing time results are only for reversible MCMC detectors.

2.3.5 Complexity of the MCMC Detector

The conditional probability for the j-th symbol in (2.5) can be computed efficiently by reusing the result obtained in earlier iterations, when we evaluate $\left\| \mathbf{y} - \sqrt{\mathrm{SNR}/N}\mathbf{H}\hat{\mathbf{x}}_{j|\omega} \right\|^2$. Since we are only changing the j-th symbol in the symbol vector, the difference $\mathbf{d}^{(l)} \triangleq \mathbf{y} - \sqrt{\mathrm{SNR}/N}\mathbf{H}\hat{\mathbf{x}}_{j|\omega}$ can be expressed as

$$\mathbf{d}^{(l)} = \mathbf{d}^{(l-1)} - \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{h}_j \Delta x_{j|\omega} \ , \tag{2.6}$$

where l is the index for the number of iterations, $\Delta x_{j|\omega} \triangleq x_{j|\omega}^{(l)} - x_{j|\tilde{\omega}}^{(l-1)}$, and $\mathbf{h}_j$ is the j-th column of $\mathbf{H}$. The computation of the conditional probability when changing the symbol in the j-th position costs $2N$ operations[1], where we define an operation as a Multiply and Accumulate (MAC) instruction. This leads to a complexity of $O\left(2N[|\Xi| - 1)]\right)$ operations per iteration, which grows linearly with $|\Xi|$.

2.3.6 MCMC Sampling Using QR- or QL-factorization

When the number of iterations in the MCMC detector is sufficiently larger than the system size, the complexity of MCMC detector can be reduced further using a QR- or QL-factorization, $\mathbf{H} = \tilde{\mathbf{Q}}\mathbf{R} = \mathbf{Q}\mathbf{L}$, thus (2.4) becomes

$$\min_{\mathbf{x}\in\Xi^N} \left\| \tilde{\mathbf{y}} - \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{L}\mathbf{x} \right\|^2 , \tag{2.7}$$

where $\hat{\mathbf{y}} \triangleq \mathbf{Q}^T\mathbf{y}$. Since $\mathbf{L}$ is a lower triangular matrix, the product $\mathbf{L}\mathbf{x}$ requires less operations compared to a full channel matrix. Suppose we need to update position index j at the current iteration and assume $\mathbf{d}^{(l-1)}$ is known, we only need to compute the indices from j to N in $\mathbf{d}^{(l)}$, since these are the only non-zero elements in $\mathbf{L}_{1:N,j}\Delta x_{j|\Xi}$. Thus, for a square channel matrix of size N the complexity of one iteration in the MCMC detector can roughly be reduced to half, namely $O\left(N[|\Xi| - 1]\right)$. This saving should be compared with the complexity of performing the QL-factorization, which requires $\mathcal{O}\left(\frac{2}{3}N^3 + 2N^2\right)$. Thus, we can achieve a complexity reduction when the number of iterations is $k > (\frac{3}{2}N^2 + N)/(|\Xi| - 1)$.

[1]We need to compute both the product $\mathbf{h}_j\Delta x_{j|\omega}$ and the inner product $(\mathbf{d}^{(l)})^T\mathbf{d}^{(l)}$.

2.4 Probability of Error

First, we derive the probability of error for ML detection in MIMO systems, then use the results to characterize the SNR regime of interest. The error probability is calculated by averaging over the random matrices $\mathbf{H}$ and random noises. We will state lemma 2.4.1 (the proof of which is provided in the appendix).

Lemma 2.4.1 (Gaussian Integral)**.** *Let* $\mathbf{v}$ *and* $\mathbf{x}$ *be i.i.d* $\mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ *each. Let* a *and* $\eta \leq 0$ *be constant numbers. We further assume that* $1 - 2a^2\eta(1+2\eta) > 00$*. Then*

$$\mathbb{E}\left\{e^{\eta\left(\|\mathbf{v}+a\mathbf{x}\|^2 - \|\mathbf{v}\|^2\right)}\right\} = \left(\frac{1}{1-2a^2\eta(1+2\eta)}\right)^{N/2}. \tag{2.8}$$

Let us first look at the probability of error using maximum likelihood detection. We will make an error if there exists a vector $\mathbf{x} \neq -\mathbf{1}$ such that

$$\left\|\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\mathbf{x}\right\|^2 \leq \left\|\mathbf{y} + \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\mathbf{1}\right\|^2 = \|\boldsymbol{v}\|^2 .$$

In other words,

$$\begin{aligned} P_e &= \mathrm{Prob}\left(\left\|\mathbf{y} - \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\mathbf{x}\right\|^2 \leq \|\boldsymbol{v}\|^2\right) \\ &= \mathrm{Prob}\left(\left\|\boldsymbol{v} + \sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}(-\mathbf{1}-\mathbf{x})\right\|^2 \leq \|\boldsymbol{v}\|^2\right), \end{aligned}$$

for some $\mathbf{x} \neq -\mathbf{1}$, which can be formulated as

$$P_e \;=\; \text{Prob}\left(\left\|\boldsymbol{v} + 2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2 \leq \|\boldsymbol{v}\|^2\right),$$

for some $\boldsymbol{\delta} \neq 0$, where $\boldsymbol{\delta} \triangleq \frac{1}{2}(-\mathbf{1} - \mathbf{x})$. Note that in the above equation $\boldsymbol{\delta}$ is a vector of zeros and -1's. Now using the union bound

$$P_e \leq \sum_{\boldsymbol{\delta}\neq 0} \text{Prob}\left(\left\|\boldsymbol{v} + 2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2 \leq \|\boldsymbol{v}\|^2\right). \tag{2.9}$$

We will use the Chernoff bound to bound the quantity inside the summation. Thus,

$$\text{Prob}\left(\left\|\boldsymbol{v} + 2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2 \leq \|\boldsymbol{v}\|^2\right) \tag{2.10a}$$

$$\leq \mathbb{E}\left\{e^{-\beta\left(\left\|\boldsymbol{v}+2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2 - \|\boldsymbol{v}\|^2\right)}\right\} \tag{2.10b}$$

$$= \left(\frac{1}{1 + 8\frac{\text{SNR}\|\boldsymbol{\delta}\|^2}{N}\beta(1-2\beta)}\right)^{N/2}, \tag{2.10c}$$

where $\beta \geq 0$ is the Chernoff parameter, and where we have used Lemma 2.8 with $\eta = -\beta$ and $a = 2\sqrt{\frac{\text{SNR}\|\boldsymbol{\delta}\|^2}{N}}$, since

$$\mathbb{E}\left\{\left(2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right)\left(2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right)^T\right\} = 4\frac{\text{SNR}\|\boldsymbol{\delta}\|^2}{N}\mathbf{I}_N.$$

The optimal value for β is $\frac{1}{4}$, which yields the tightest bound

$$\text{Prob}\left(\left\|\boldsymbol{v} + 2\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2 \leq \|\boldsymbol{v}\|^2\right) \leq \left(\frac{1}{1 + \frac{\text{SNR}\|\boldsymbol{\delta}\|^2}{N}}\right)^{N/2}. \tag{2.11}$$

Note that this depends only on $\|\boldsymbol{\delta}\|^2$, the number of nonzero entries in $\boldsymbol{\delta}$. Plugging this into the union bound yields

$$P_e \leq \sum_{i=1}^{N} \binom{N}{i} \left(\frac{1}{1 + \frac{\text{SNR}i}{N}}\right)^{N/2}. \tag{2.12}$$

Let us first look at the linear (i.e., i proportional to N) terms in the above sum,

$$\binom{N}{i} \left(\frac{1}{1 + \frac{\text{SNR}i}{N}}\right)^{N/2} \approx e^{NH(\frac{i}{N}) - \frac{N}{2}\ln\left(1 + \frac{\text{SNR}i}{N}\right)},$$

where $H(\cdot)$ is entropy in "nats". Clearly, if

$$\lim_{N\to\infty} \text{SNR} = \infty,$$

then the linear terms go to zero (superexponentially fast).

Let us now look at the sublinear terms. In particular, let us look at $i = 1$:

$$N\left(\frac{1}{1 + \frac{\text{SNR}}{N}}\right)^{N/2} \approx Ne^{-\text{SNR}/2}.$$

Clearly, to have this term go to zero, we require that SNR $> 2\ln N$. A similar argument shows that all other sublinear terms also go to zero, and so we have:

Lemma 2.4.2 (SNR scaling). *If SNR* $> 2\ln N + f(N)$*, where* $f(N)$ *is an arbitrary function that goes to* ∞ *as* $N \to \infty$*, then* $P_e \to 0$ *as* $N \to \infty$*.*

We remark that the requirement on SNR $> 2\ln N + f(N)$ to guarantee small sequence error probability is near optimal. In fact, one can show that, for a channel matrix with orthogonal columns, if SNR $< (1-\epsilon)\ln N$, where $\epsilon > 0$ is an arbitrary constant, the sequence error probability is then lower bounded by a positive constant as $N \to \infty$.

2.5 Computing the Optimal α

In this section, we derive the optimal value of the "temperature" parameter which controls the mixing time of the underlying Markov chain. The temperature has the desirable property that once the Markov chain has mixed to steady state, there is only polynomially (and not exponentially) small probability of encountering the optimal solution.

2.5.1 Mean of $\pi_{-\mathbf{1}}$

In the following section we compute the expected value of the stationary probabilities of the states, where the expectation is taken over random Gaussian $\mathbf{H}$ and noises. More specifically, we are examining the probability of state $\mathbf{x} = -\mathbf{1}$, denoted by $\pi_{-\mathbf{1}}$ (recall that we assumed that $-\mathbf{1}$ is transmitted symbol vector). This calculation has a lot in common with the one in Section 2.4. Let $\boldsymbol{\delta}$ be a vector of zeros and ones, then:

$$\pi_{-\mathbf{1}} = \frac{e^{-\frac{1}{2\alpha^2}\left\|\mathbf{y}+\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H1}\right\|^2}}{\sum_{\mathbf{x}} e^{-\frac{1}{2\alpha^2}\left\|\mathbf{y}+\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{Hx}\right\|^2}} \tag{2.13a}$$

$$= \frac{e^{-\frac{1}{2\alpha^2}\|\boldsymbol{v}\|^2}}{\sum_{\mathbf{x}} e^{-\frac{1}{2\alpha^2}\left\|\boldsymbol{v}+\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}(\mathbf{x}-\mathbf{1})\right\|^2}} \tag{2.13b}$$

$$= \frac{e^{-\frac{1}{2\alpha^2}\|\boldsymbol{v}\|^2}}{\sum_{\boldsymbol{\delta}} e^{-\frac{1}{2\alpha^2}\left\|\boldsymbol{v}+2\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2}} \tag{2.13c}$$

$$= \frac{1}{\sum_{\boldsymbol{\delta}} e^{-\frac{1}{2\alpha^2}\left(\left\|\boldsymbol{v}+2\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2-\|\boldsymbol{v}\|^2\right)}}, \tag{2.13d}$$

Now, using Jensen's inequality and the convexity of $\frac{1}{t}$ when $t > 0$,

$$\mathbb{E}\left\{\pi_{-\mathbf{1}}\right\} \geq \frac{1}{\mathbb{E}\left\{\frac{1}{\pi_{-\mathbf{1}}}\right\}}$$

$$= \frac{1}{\mathbb{E}\left\{\sum_{\boldsymbol{\delta}} e^{-\frac{1}{2\alpha^2}\left(\left\|\boldsymbol{v}+2\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2-\|\boldsymbol{v}\|^2\right)}\right\}}$$

$$= \frac{1}{\sum_{\boldsymbol{\delta}} \mathbb{E}\left\{e^{-\frac{1}{2\alpha^2}\left(\left\|\boldsymbol{v}+2\sqrt{\frac{\mathrm{SNR}}{N}}\mathbf{H}\boldsymbol{\delta}\right\|^2-\|\boldsymbol{v}\|^2\right)}\right\}}$$

$$= \frac{1}{1+\sum_{\boldsymbol{\delta}\neq 0}\left(\frac{1}{1+4\frac{\mathrm{SNR}\|\boldsymbol{\delta}\|^2}{N}\frac{1}{\alpha^2}\left(1-\frac{1}{\alpha^2}\right)}\right)^{N/2}} \tag{2.14a}$$

$$= \frac{1}{1+\sum_{i=1}^{N}\binom{N}{i}\left(\frac{1}{1+\frac{\beta i}{N}}\right)^{N/2}}. \tag{2.14b}$$

In (2.14a) we have used Lemma 2.8 and in (2.14b) we have defined $\beta \triangleq 4\text{SNR}\frac{1}{\alpha^2}(1-\frac{1}{\alpha^2})$. While it is possible to focus on the linear and sublinear terms in the above summation separately, to give conditions for $\mathbb{E}\{\pi_{-\mathbf{1}}\}$ to have the form of $1/\text{poly}(N)$, we will be interested in the exact exponent of the poly and so we need a more accurate estimate. To do this we shall use saddle point integration. To this end, note that

$$\binom{N}{i}\left(\frac{1}{1+\frac{\beta i}{N}}\right)^{N/2} \leq e^{NH(\frac{i}{N})-\frac{N}{2}\ln\left(1+\frac{\beta i}{N}\right)},$$

where $H(\cdot)$ represents the entropy in "nats", and the inequality is through using the following inequality from [54]:

$$\frac{e^{NH(\frac{i}{N})}}{N+1} \leq \binom{N}{i} \leq e^{NH(\frac{i}{N})}.$$

And so the summation in the denominator of (2.14b) can be approximated as:

$$\sum_{i=1}^{N}\binom{N}{i}\left(\frac{1}{1+\frac{\beta i}{N}}\right)^{N/2} \leq \sum_{i=1}^{N} e^{NH(\frac{i}{N})-\frac{N}{2}\ln\left(1+\frac{\beta i}{N}\right)} \tag{2.15a}$$

$$\leq N\max_{i} e^{NH(\frac{i}{N})-\frac{N}{2}\ln\left(1+\frac{\beta i}{N}\right)} \tag{2.15b}$$

$$\leq N\max_{x\in[0,1]} e^{NH(x)-\frac{N}{2}\ln(1+\beta x)} \tag{2.15c}$$

$$= Ne^{Nf(x_0)}, \tag{2.15d}$$

where x_0 is the saddle point of $f(\cdot)$, i.e.,$f'(x_0) = 0$. In our case,

$$f(x) = -x\ln x - (1-x)\ln(1-x) - \frac{1}{2}\ln(1+\beta x)\ ,$$

and so

$$f'(x) = \ln\frac{1-x}{x} - \frac{1}{2}\frac{\beta}{1+\beta x}\ .$$

In general, it is not possible to solve for $f'(x_0) = 0$ in closed form. However, in our case, we assume that $\beta = 4\text{SNR}\frac{1}{\alpha^2}(1-\frac{1}{\alpha^2}) \gg 1$ (In fact, we must have $\beta \to \infty$ as $N \to \infty$. Otherwise, (2.15) will be exponential in N). In this case, it is not too hard to verify that the saddle point is given by $x_0 \approx e^{-\frac{\beta}{2}}$. And hence,

$$\begin{aligned} f(x_0) &= -e^{-\frac{\beta}{2}}\ln e^{-\frac{\beta}{2}} - (1-e^{-\frac{\beta}{2}})\ln(1-e^{-\frac{\beta}{2}}) \\ &\quad - \frac{1}{2}\ln(1+\beta e^{-\frac{\beta}{2}}) \\ &\approx \frac{\beta}{2}e^{-\frac{\beta}{2}} + e^{-\frac{\beta}{2}} - \frac{1}{2}\beta e^{-\frac{\beta}{2}} \\ &= e^{-\frac{\beta}{2}}\ . \end{aligned}$$

Replacing these into the saddle point expression in (2.15) shows that

$$\sum_{i=1}^{N}\binom{N}{i}\left(\frac{1}{1+\frac{\beta i}{N}}\right)^{N/2} \le N\exp\left(Ne^{-\frac{\beta}{2}}\right). \tag{2.16}$$

We want $\mathbb{E}\{\pi_{-1}\}$ to behave as $\frac{1}{N^\zeta}$ and according to (2.13) this means that we want the expression in (2.16) to behave as N^ζ, where $\zeta > 1$ is a positive number. Let

us take

$$Ne^{Ne^{-\frac{\beta}{2}}} = N^{\zeta} .$$

Solving for β yields

$$\beta = 2(\ln N - \ln(\zeta - 1)). \tag{2.17}$$

Finally, this choice of β means that we have

$$4\text{SNR}\frac{1}{\alpha^2}\left(1 - \frac{1}{\alpha^2}\right) = 2(\ln N - \ln(\zeta - 1)) ,$$

and so we have the following result.

Lemma 2.5.1 (Mean of π_{-1})**.** *Let $\zeta > 1$ be a positive constant. As $N \to \infty$, if α is chosen such that*

$$\frac{\alpha^2}{1 - \frac{1}{\alpha^2}} = \frac{2SNR}{\ln N - \ln(\zeta - 1)} , \tag{2.18}$$

then

$$\mathbb{E}\left\{\frac{1}{\pi_{-1}}\right\} \leq 1 + N^{\zeta} , \tag{2.19}$$

and

$$\mathbb{E}\left\{\pi_{-1}\right\} \geq \frac{1}{1 + N^{\zeta}} . \tag{2.20}$$

When we have an upper bound on $\mathbb{E}\left\{\frac{1}{\pi_{-1}}\right\}$, we can then use the Markov inequality to give upper bounds on the probability that $\frac{1}{\pi_{-1}}$ exceeds a certain threshold. More precisely, we have $P(\frac{1}{\pi_{-1}} > N^{\gamma'}) \leq \mathbb{E}\left\{\frac{1}{\pi_{-1}}\right\}/N^{\gamma'} \leq N^{-(\gamma'-\zeta)}$ for any γ' (here we omit the '1' in $1+N^{\zeta}$ when N is large). This means that with probability close to

1 as $N \to \infty$, the expected time to encounter the transmitted signal in steady state is no bigger than $N^{\gamma'}$, for every $\gamma' > \zeta$.

2.5.2 Value of α

In this subsection we investigate how α behaves as a function of the SNR and the system dimension, if α is chosen according to (2.18). In general, the larger α is, the faster the Markov chain mixes. However, choosing α any larger than this means that the probability of finding the optimal solution in stationary distribution is exponentially small. Thus, when choosing the value of α, there is a trade-off between faster mixing time of the Markov chain (due to an increase of α), and faster encountering the optimal solution in stationary distribution. In the following, we evaluate (2.18) with $\zeta = i$, denoted as $\alpha_{\zeta=i}$ and we also approximate α in (2.18) by neglecting the terms $\ln\ln(N)$ and $\ln(\zeta - 1)$, leading to

$$\frac{\alpha^4}{\alpha^2 - 1} = \frac{2\text{SNR}}{\ln(N)} \,. \tag{2.21}$$

From (2.21) we see that

$$\alpha^2 = \frac{\text{SNR}}{\ln(N)} \pm \sqrt{\left(\frac{\text{SNR}}{\ln(N)}\right)^2 - 2\frac{\text{SNR}}{\ln(N)}} \,, \tag{2.22}$$

which implies that $\tilde{\alpha}$ becomes complex when SNR $< 2\ln(N)$. However, as stated in Section 2.2 we focus on SNR $> 2\ln(N)$. Since we are solving a quadratic equation we get two values of α^2, representing the region in which (2.20) is satisfied. Based

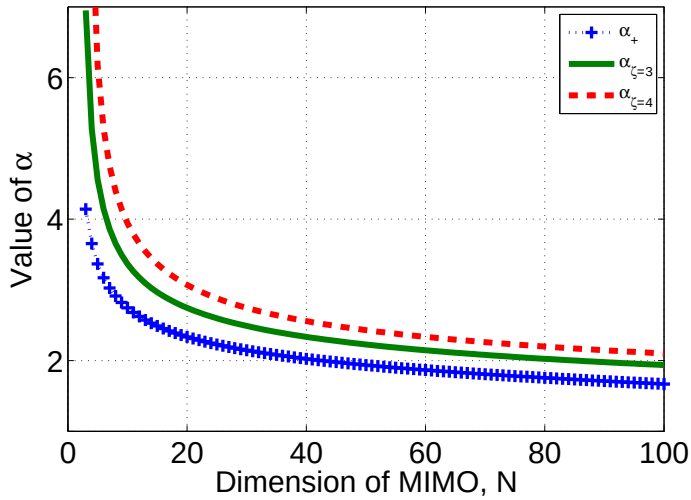


Figure 2.1: Value of α vs. system size N, for SNR = 10 dB.

on the considerations given above, we prefer the value of α^2 obtained by the plus sign in (2.22), denoted α_+^2, in order to achieve the fastest mixing time. In Figure 2.1 the values of $\alpha_{\zeta=3}$, $\alpha_{\zeta=4}$, and α_+ have been plotted as a function of the system dimension for SNR = 10 dB. Our simulations suggest that the computed values of α_+ give better MIMO detection performance when SNR is large, compared with using channel noise variance for α^2. Our simulations also suggest that the computed value of α is very close to the optimal choice, even in the case where the condition SNR $> 2\ln(N)$ is not satisfied.

2.5.3 Mixing Time of Markov Chain

So far, we have examined the largest possible α such that the optimal sequence has a reasonable stationary probability. However, all this was based on assuming that we have reached the stationary distribution. As α also affects the

speed of getting to the stationary distribution, it is interesting to quantify the mixing time of MCMC detectors. In the next sections, we will discuss how the mixing time is related to α and the underlying lattice structures in ILS problems.

2.6 Mixing Time

Starting from this section, we consider the mixing time for MCMC for ILS problems and study how the mixing time for ILS problem depends on the linear matrix structure and SNR. For simplicity, we use $\acute{\mathbf{H}}$ to represent $\sqrt{\frac{\text{SNR}}{N}}\mathbf{H}$, and the model we are currently considering is

$$\mathbf{y} = \acute{\mathbf{H}}\mathbf{x} + \boldsymbol{v}. \tag{2.23}$$

When the SNR increases, we simply increase the amplitude of elements in $\acute{\mathbf{H}}$. We will also incorporate the SNR term into $\acute{\mathbf{H}}$ this way in the following sections unless stated otherwise.

2.6.1 Orthogonal Matrices

As a first step, we consider a linear matrix $\mathbf{H}$ with orthogonal columns. As shown later, the mixing time for this matrix has an upper bound independent of SNR. In fact, this is a general phenomenon for ILS problems without local minima.

Theorem 2.6.1. *Independent of the temperature α and SNR, the mixing time of the MCMC detector for orthogonal-column ILS problems is upper bounded by $N\log(N) + \log(1/\epsilon)N$.*

This theorem is an extension of the mixing time for regular random walks on an N-dimensional hypercube [51]. The only difference here is that the transition probability follows (2.5) and that the transition probability depends on SNR. Under orthogonal columns, the ILS problem has no local minimum, since $\mathbf{H}^T\mathbf{H}$ is a diagonal matrix in the expansion of $\|\mathbf{y} - \acute{\mathbf{H}}\mathbf{x}\|_2^2$.

Proof. When the j-th index was selected for updating in the MCMC detector, since the columns of $\acute{\mathbf{H}}$ are orthogonal to each other, the probability of updating $\mathbf{x}_j$ to -1 is $\frac{1}{1+e^{\frac{2\mathbf{y}^T\mathbf{h}_j}{\alpha^2}}}$. We note that this probability is independent of the current state of Markov chain $\hat{\mathbf{x}}$. So we can use the classical coupling idea to get an upper bound on the mixing time of this Markov chain.

Consider two separate Markov chains starting at two different states $\mathbf{x}_1$ and $\mathbf{x}_2$. These two follow the same update rule according to (2.5). By using the same random source, they select the same position index in each step to update, and they update that position to the same symbol. Let τ_{couple} be the first time the two chains come to the same state. Then by a classical result, the total variation distance

$$d(t) = \max_{\tilde{\mathbf{x}}} \|P^t(\tilde{\mathbf{x}}, \cdot) - \pi\|_{TV} \leq \max_{\mathbf{x}_1, \mathbf{x}_2} p_{\mathbf{x}_1, \mathbf{x}2}\{\tau_{couple} > t\}. \tag{2.24}$$

Note that the coupling time is just time for collecting all of the positions where $\mathbf{x}_1$ and $\mathbf{x}_2$ differ, as in the coupon collector problem. From the famous coupon collector problem [51], for any $\mathbf{x}_1$ and $\mathbf{x}_2$,

$$d(N\log(N)+cN) \le p_{\mathbf{x}_1,\mathbf{x}_2}\{\tau_{couple} > N\log(N)+cN\} \le e^{-c}. \tag{2.25}$$

So the conclusion follows. ■

2.6.2 Mixing Time with Local Minima

In this subsection, we consider the mixing time for ILS problems which have local minima besides the global minimum point. Our main results are that local minima greatly affect the mixing time of MCMC detectors, and rigorous statements are given in Theorem 2.6.4. First, we give the definition of a local minimum.

A local minimum $\tilde{\mathbf{x}}$ is a state such that $\tilde{\mathbf{x}}$ is not a global minimizer for $\min_{\mathbf{s}\in\{-1,+1\}^N} \|\mathbf{y}-\acute{\mathbf{H}}\mathbf{s}\|^2$; and any of its neighbors which differ from $\tilde{\mathbf{x}}$ in only one position index, denoted by $\tilde{\mathbf{x}}'$, satisfies $\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2 > \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2$.

We will use the following theorem about the spectral gap of Markov chain to evaluate the mixing time.

Theorem 2.6.2 (Jerrum and Sinclair (1989) [55], Lawler and Sokal (1988) [56], [51])**.** *Let λ_2 be the second largest eigenvalue of a reversible transition matrix P, and let $\gamma = 1-\lambda_2$. Then*

$$\frac{\Phi_*^2}{2} \le \gamma \le 2\Phi_*,$$

where Φ_* is the bottleneck ratio (also called conductance, Cheeger constant, and isoperimetric constant) defined as

$$\Phi_* = \min_{\pi(S)\le\frac{1}{2}} \frac{Q(S,S^c)}{\pi(S)}.$$

Here S is any subset of the state spaces with stationary measure no bigger than $\frac{1}{2}$, S^c is its complement set, and $Q(S, S^c)$ is the probability of moving from S to S^c in one step when starting with the stationary distribution.

Theorem 2.6.3. *If there is a local minimum $\tilde{\mathbf{x}}$ in an integer least-squares problem and we denote its neighbor differing only at the j-th ($1 \leq j \leq N$) location as $\tilde{\mathbf{x}}_j$, then the mixing time of the MCMC detector is at least*

$$t_{mix}(\epsilon) \geq \log(\frac{1}{2\epsilon})(\frac{1}{\gamma} - 1), \tag{2.26}$$

where

$$\gamma = \sum_{j=1}^{N} \frac{2}{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}} + e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}} \tag{2.27}$$

The parameter γ is upper bounded by

$$\frac{2}{1 + e^{\frac{\min_j \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2 - \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}} \tag{2.28}$$

Proof. We apply Theorem 2.6.2 to prove this result. We take a local minimum point $\tilde{\mathbf{x}}$ as the single element in the bottle-neck set S. Since $\tilde{\mathbf{x}}$ is a local minimum, $\pi(S) \leq \frac{1}{2}$.

$$Q(S, S^c) = \frac{\pi(S)}{N} \sum_{j=1}^{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}} + e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}} \tag{2.29}$$

Dividing by $\pi(S)$, by the definition of Φ_*

$$\Phi_* \leq \frac{Q(S,S^c)}{\pi(S)} = \frac{1}{N}\sum_{j=1}^{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}} + e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}} \tag{2.30}$$

So we know $\gamma \leq 2\frac{1}{N}\sum_{j=1}^{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}}+e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}}$. From a well-known theorem for the relationship between $t_{mix}(\epsilon)$ and γ: $t_{mix}(\epsilon) \geq (\frac{1}{\gamma}-1)\log(\frac{1}{2\epsilon})$ [51], our conclusion follows. ■

Theorem 2.6.4. *For an ILS problem* $\min_{\mathbf{s}\in\{-1,+1\}^N} \|\mathbf{y}-\acute{\mathbf{H}}\mathbf{s}\|^2$*, where* $\acute{\mathbf{H}}$ *is fixed and no two vectors give the same objective distance, the relaxation time (the inverse of the spectral gap) of the Markov chain for the reversible MCMC detector (Algorithm 1) is upper bounded by a constant as the temperature* $\alpha \to 0$ *if and only if there is no local minimum. Moreover, when there is a local minimum, as* $\alpha \to 0$*, then* $t_{mix}(\epsilon) = e^{\Omega(\frac{1}{2\alpha^2})}$.

Remarks: For the signal model $\mathbf{y} = \sqrt{\frac{\text{SNR}}{N}}\mathbf{H}\mathbf{x} + \boldsymbol{v}$, if α is set equal to the noise variance as in [47], [48], it is equivalent to setting "$\alpha \to 0$" when SNR $\to \infty$. We will keep $\acute{\mathbf{H}}$ fixed in Theorem 2.6.4 since the SNR is incorporated into $\acute{\mathbf{H}}$.

Proof. First, when there is a local minimum, from Theorem 2.6.3 and Theorem 2.6.2, the spectral gap γ is lower bounded by

$$\gamma = \frac{2}{N}\sum_{j=1}^{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2}{2\alpha^2}} + e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}} \tag{2.31}$$

As the temperature $\alpha \to 0$, the spectral gap upper bound

$$\frac{2}{1+e^{\frac{\min_j \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2 - \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}}} \tag{2.32}$$

decreases at the speed of $\Theta(e^{-\frac{\min_j \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}_j\|^2 - \|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2}{2\alpha^2}})$. So the relaxation time of the MCMC is lower bounded by $t_{mix}(\epsilon) = e^{\Omega(\frac{1}{2\alpha^2})}$, which grows unbounded as $\alpha \to 0$.

Suppose instead that there is no local minimum. We argue that as $\alpha \to 0$, the spectral gap of this MCMC is lower bounded by some constant independent of α. Again, we look at the bottle neck ratio and use Theorem 2.6.2 to bound.

Consider any set S of sequences which do not include the global minimum point $\mathbf{x}^*$. As $\alpha \to 0$, the measure of this set of sequences $\pi(S) \leq \frac{1}{2}$. Moreover, as $\alpha \to 0$, any set S with $\pi(S) \leq \frac{1}{2}$ can not contain the global minimum point $\mathbf{x}^*$. Now we look at the sequence $\tilde{\mathbf{x}}'$ which has the smallest distance $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|$ among the set S. Since there is no local minimum, $\tilde{\mathbf{x}}'$ must have at least one neighbor $\tilde{\mathbf{x}}''$ in S^c which has smaller distance than $\tilde{\mathbf{x}}'$. Otherwise, this would imply $\tilde{\mathbf{x}}'$ is a local minimum. So

$$Q(S, S^c) \geq \pi(\tilde{\mathbf{x}}') \times \frac{1}{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}''\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}''\|^2}{2\alpha^2}} + e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2}{2\alpha^2}}} \tag{2.33}$$

As $\alpha \to 0$, $\frac{\pi(\tilde{\mathbf{x}}')}{\pi(S)} \to 1$. So for a given $\epsilon > 0$, as $\alpha \to 0$

$$\frac{Q(S, S^c)}{\pi(S)} \geq \frac{1-\epsilon}{N} \frac{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}''\|^2}{2\alpha^2}}}{e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}''\|^2}{2\alpha^2}} + e^{-\frac{\|\mathbf{y}-\acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2}{2\alpha^2}}}, \tag{2.34}$$

which approaches $\frac{(1-\epsilon)}{N}$ as $\alpha \to 0$ because $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}''\|^2 < \|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2$. From Theorem 2.6.2, γ is at least $\frac{(\frac{Q(S,S^c)}{\pi(S)})^2}{2}$, which is lower bounded by a constant as $\alpha \to 0$. ∎

So from the analysis above, the mixing time is closely related to whether there are local minima in the problem. In the next section, we will see there often

exist local minima, which implies very slow convergence rate for MCMC when the temperature is kept at the noise level in the high SNR regime.

2.7 Presence of Local Minima

We have seen that the mixing time of MCMC detectors are closely related to the existence of local minima. It is natural to ask how often local minima occur in ILS problems. In this section, we derive some results about how many local minima there are in an ILS problem, especially when the SNR is high.

Theorem 2.7.1. *There can be exponentially many local minima in an integer least-quare problem*

Proof. See Appendix 2.9.2 for a detailed proof. ∎

Now we study how often we encounter a local minimum in specific ILS problem models. Without loss of generality, we assume that the transmitted sequence is an all $-\mathbf{1}$ sequence. We first give the condition for $\tilde{\mathbf{x}}$ to be a local minimum. We assume that $\tilde{\mathbf{x}}$ is a vector which has k '+1' over an index set K with $|K| = k$ and $(N-k)$ '−1' over the set $\overline{K} = \{1, 2, ..., N\} \setminus K$.

Lemma 2.7.2. *Consider*

$$\mathbf{y} = \acute{\mathbf{H}}\mathbf{x} + \boldsymbol{v}, \tag{2.35}$$

where the columns of $N \times N$ matrix $\acute{\mathbf{H}}$ are denoted by $\mathbf{h}_i$, $1 \leq i \leq N$. Then $\tilde{\mathbf{x}}$ is a local minimum if and only if $\tilde{\mathbf{x}}$ is not a global minimum; and

- $\forall i \in K$,

$$\mathbf{h}_i^T(\sum_{j\in K}\mathbf{h}_j - \frac{\boldsymbol{v}}{2}) < \frac{\|\mathbf{h}_i\|^2}{2} \tag{2.36}$$

- $\forall i \in \overline{K}$,

$$\mathbf{h}_i^T(\sum_{j\in K}\mathbf{h}_j - \frac{\boldsymbol{v}}{2}) > -\frac{\|\mathbf{h}_i\|^2}{2}. \tag{2.37}$$

Proof. For a position $i \in K$, when we flip $\tilde{\mathbf{x}}_i$ to 1, $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2$ is increased, namely,

$$\begin{aligned}
& \|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2 - \|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}_{\sim i}\|^2 \\
= \; & \| -2\sum_{j\in K}\mathbf{h}_j + \boldsymbol{v}\|^2 - \| -2\sum_{j\in K, j\neq i}\mathbf{h}_j + \boldsymbol{v}\|^2 \\
= \; & 4\|\mathbf{h}_i\|^2 + 4\mathbf{h}_i^T(2\sum_{j\in K, j\neq i}\mathbf{h}_j - \boldsymbol{v}) \\
< \; & 0,
\end{aligned} \tag{2.38}$$

where $\tilde{\mathbf{x}}_{\sim i}$ is a neighbor of $\tilde{\mathbf{x}}$ by changing index i. This means

$$\mathbf{h}_i^T(\sum_{j\in K}\mathbf{h}_j - \frac{\boldsymbol{v}}{2}) < \frac{\|\mathbf{h}_i\|^2}{2}. \tag{2.39}$$

For a position $i \in \overline{K}$, when we flip $\tilde{\mathbf{x}}_i$ to -1, $\|\mathbf{y} - \mathbf{h}\tilde{\mathbf{x}}'\|^2$ is also increased,

namely,

$$
\begin{aligned}
& \|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}\|^2 - \|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}_{\sim i}\|^2 \\
= \; & \| -2 \sum_{j \in K} \mathbf{h}_j + \boldsymbol{\upsilon}\|^2 - \| -2 \sum_{j \in K} \mathbf{h}_j - 2\mathbf{h}_i + \boldsymbol{\upsilon}\|^2 \\
= \; & -4\|\mathbf{h}_i\|^2 + 4\mathbf{h}_i^T(-2 \sum_{j \in K} \mathbf{h}_j + \boldsymbol{\upsilon}) \\
< \; & 0. \qquad (2.40)
\end{aligned}
$$

This means

$$
(\mathbf{h}_i)^T(\sum_{j \in K} \mathbf{h}_j - \frac{\boldsymbol{\upsilon}}{2}) > -\frac{\|\mathbf{h}_i\|^2}{2}. \qquad (2.41)
$$

■

It is not hard to see that when SNR $\to \infty$, $\boldsymbol{\upsilon}$ is comparatively small with high probability, so we have the following lemma.

Lemma 2.7.3. *When $SNR \to \infty$, $\tilde{\mathbf{x}}$ is a local minimum with high probability, if and only if $\tilde{\mathbf{x}} \neq -\mathbf{1}$; and*

- $\forall i \in K$,

$$
\mathbf{h}_i^T(\sum_{j \in K} \mathbf{h}_j) < \frac{\|\mathbf{h}_i\|^2}{2} \qquad (2.42)
$$

- $\forall i \in \overline{K}$,

$$\mathbf{h}_i^T(\sum_{j\in K}\mathbf{h}_j) > -\frac{\|\mathbf{h}_i\|^2}{2}. \tag{2.43}$$

We now set out to investigate the chance of having a local minimum in MIMO systems.

Theorem 2.7.4. *Consider a 2×2 matrix $\acute{\mathbf{H}}$ whose two columns are uniform randomly sampled from the unit-normed 2-dimensional vector. When $\boldsymbol{\upsilon} = 0$, the probability of there existing a local minimum for such an $\acute{\mathbf{H}}$ is $\frac{1}{3}$.*

Please see the appendix for its proof.

Theorem 2.7.5. *Consider a 2×2 matrix $\acute{\mathbf{H}}$ whose elements are independent $\mathcal{N}(0,1)$ Gaussian random variables. When $\boldsymbol{\upsilon} = 0$, the probability of there existing a local minimum for such an $\acute{\mathbf{H}}$ is $\frac{1}{3} - \frac{1}{\sqrt{5}} + \frac{2\arctan(\sqrt{\frac{5}{3}})}{\sqrt{5}\pi}$.*

Please refer to the appendix for its proof.

For higher dimension N, it is hard to directly estimate the probability of a vector being a local minimum based on the conditions in Lemma 2.7.2. Simulation results instead suggest that for large N, with high probability, there exists at least one local minimum. We conjecture this is the case, but proof or disproof of it seems nontrivial.

2.8 Simulation Results

In this section we present simulation results for an $N \times N$ MIMO system with a full square channel matrix containing i.i.d. Gaussian entries. In Figure 2.2 and Figure 2.3 the Bit Error Rate (BER) of the sequential MCMC detector has been evaluated as a function of the number of block iterations in a 10×10 system using a variety of α values. Thereby, we can inspect how the parameter α affects the convergence rate of the MCMC detector.

The performance of the Maximum Likelihood (ML), the Zero-Forcing (ZF), and the Linear Minimum Mean Square Error (LMMSE) detector have also been plotted, to ease the comparison of the MCMC detector with these detectors (Please see [26], for example, for descriptions of these well-known detectors. In order to

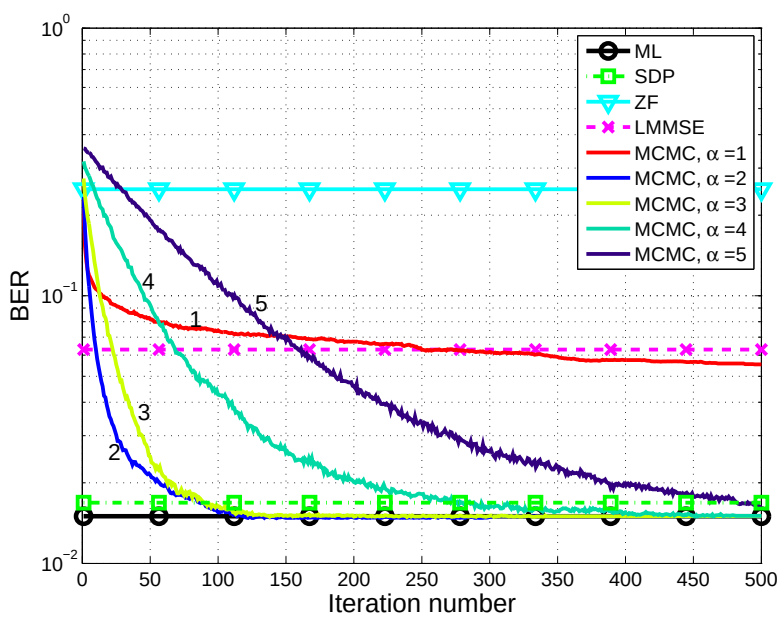


Figure 2.2: BER vs. iterations, 10×10. SNR = 10 dB.

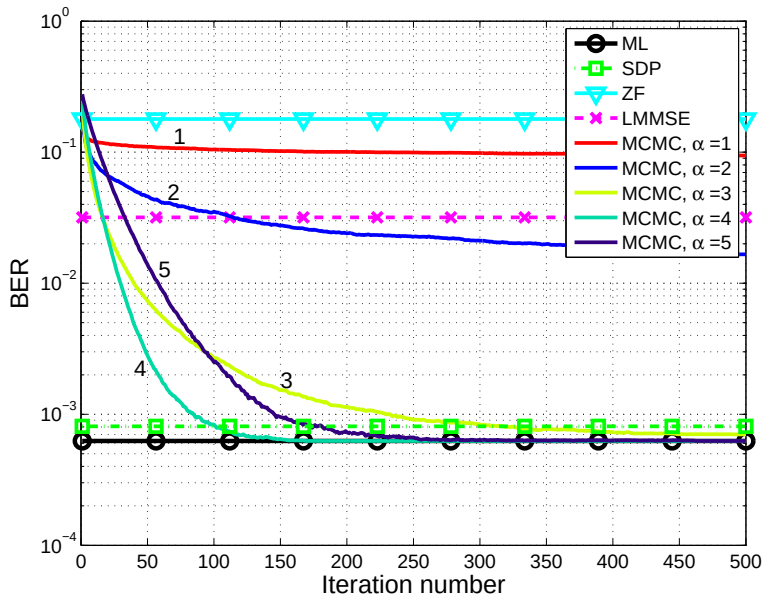


Figure 2.3: BER vs. iterations, 10×10 system. SNR = 14 dB.

benchmark the MCMC detector against a state of the art detector, the performance of the row-by-row semidefinete relaxation (SDR) detector presented in [57] has also been included.). It is seen that the MCMC detector outperforms both the ZF and the LMMSE detectors after only a few block iterations in all the presented simulations, when the tuning parameter α is chosen properly. It can also be seen that the MCMC detector can provide a performance improvement over the SDR after approximately 100 block iterations. Furthermore, it is observed that the parameter α has a huge influence on the convergence rate and that the MCMC detector converges toward the ML solution as a function of the iterations[2].

[2]It should be noted that the way we decode the symbol vector to a given iteration, is to select the symbol vector with has the lowest cost function in all the iterations up to that point in time.

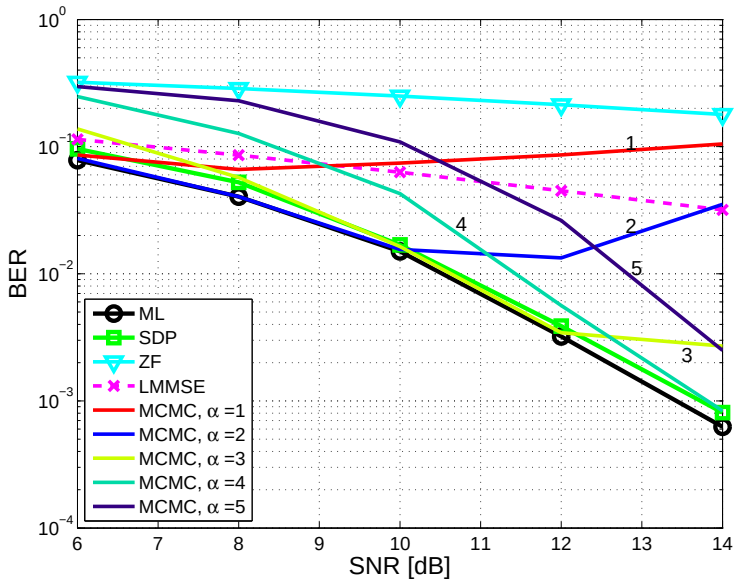


Figure 2.4: BER vs. SNR, 10×10. Number of iterations, $k = 100$.

Figure 2.4 shows the BER performance for the MCMC detector for fixed number of iterations, $k = 100$. From the figure we see that the SNR has a significant influence on the optimal choice of α given a fixed number of iterations. The performance of the sequential MCMC detector is also shown for a 50×50 system, which represents a ML decoding problem of huge complexity where an exhaustive search would require $2^{50} \approx 10^{15}$ evaluations. For this problem even the sphere decoder would have an enormous complexity under moderate SNR, and it has actually been proved in [33] that the complexity of SD for SNR = $O(\ln(N))$ is exponential. Therefore, it has not been possible to simulate the performance of this decoder within a reasonable time and we have therefore initialized the radius of the sphere to the minimum of either the norm of the transmitted symbol vector or the solution found by the MCMC detector. This has been done in order to evaluate

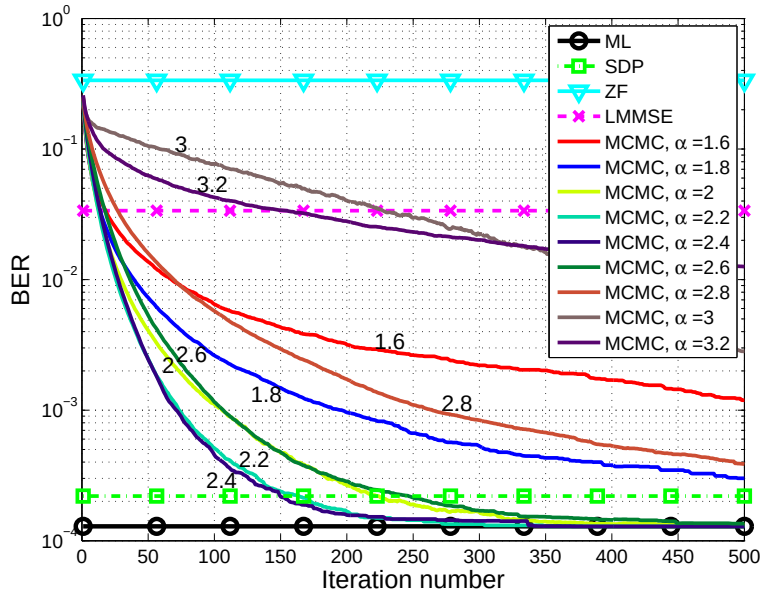


Figure 2.5: BER vs. iterations, 50×50 system. SNR = 12 dB.

the BER performance of the optimal detector. Figure 2.5 shows the BER curve as a function of the iteration number while Figure 2.6 illustrates the BER curve vs. the SNR. From these figures, we can see that a judiciously chosen α significantly outperforms the traditional way of using noise variance for α^2, which would be $\alpha^2 = 1$ given the system model in the introduction section.

We now evaluate the complexity of the MCMC method. The complexity of the MCMC method has been compared with the row-by-row semidefinite relaxation (SDR) detector described in [57]. The number of Multiply and Accumulate (MAC) instructions has been calculated for the SDR and the MCMC. Recall from Section 2.3.5 and Section 2.3.6 that the complexity of MCMC with the QR-factorization as a "preprocessing step" is $\mathcal{O}\left(kN(N+1) + \frac{2}{3}N^3 + 2N^2\right)$ where k denotes the number of *block iterations*. The complexity of the SDR in [57] can be split

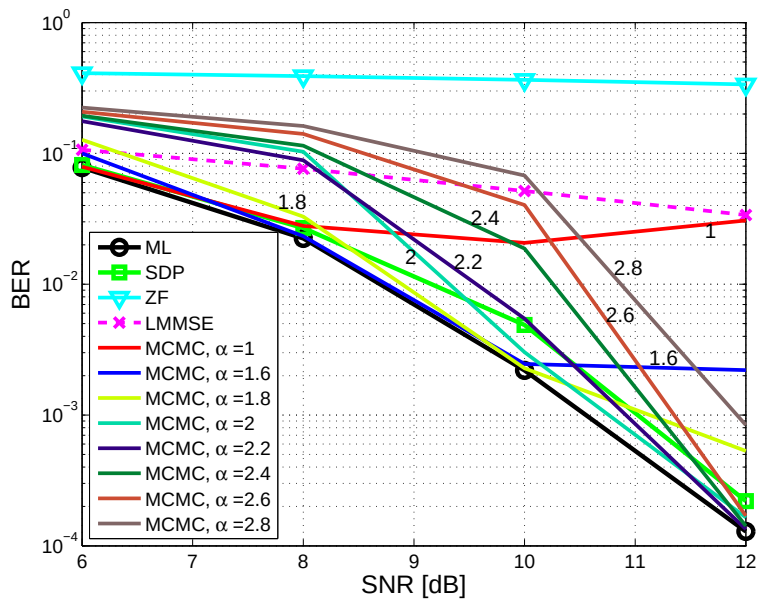


Figure 2.6: BER vs. SNR, 50×50 system. Number of iterations, $k = 300$.

into three parts, namely an initialization step, an iterative phase, and a postprocessing step. The initialization step of computation of $\mathbf{C}$ in [57] (Eq.3) costs $\mathcal{O}(N^3 + 2N^2)$. Each iteration in the iterative phase costs $\mathcal{O}\left((N+1)(N^2 + \frac{3}{2}N^2 + (N+1)^2)\right)$ MAC instructions while the postprocessing, involving approximation by Gaussian randomization costs $\mathcal{O}\left(\frac{1}{3}N^3 + \frac{3}{2}(N+1)^2 2N + 2N(N+1)\right)$. In order to do a fair comparison, the actual complexity (MAC instructions count) of the MCMC method has been measured when it has the same BER performance as the SDR, and this is shown in Figure 2.7. It is seen that, under equal BER performance, for 50×50 channel matrices, the MCMC detector has a smaller computational complexity than the state-of-the-art low complexity SDR detector. Furthermore, the computational complexity of the MCMC method has also been plotted when the BER performance visually is very close to the ML solution.

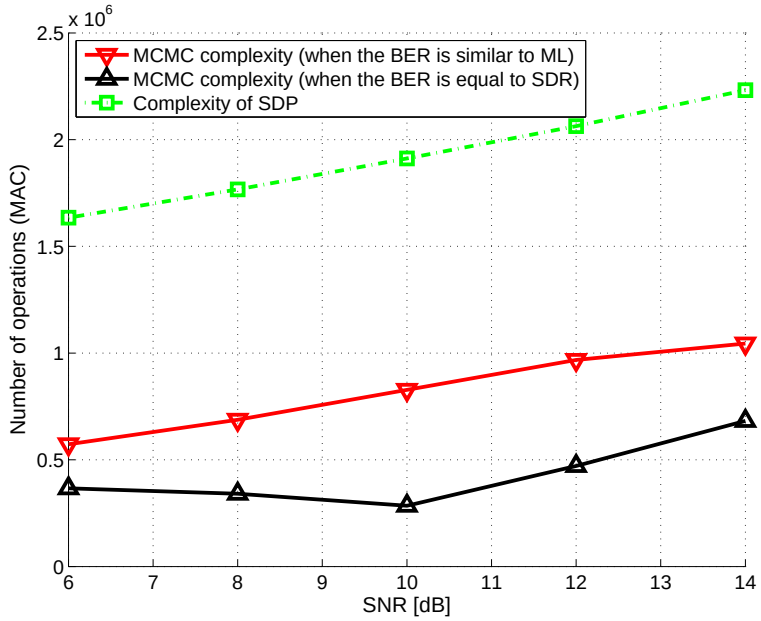


Figure 2.7: Complexity comparison in terms of Multiply and Accumulate (MAC) instructions, 50×50 system.

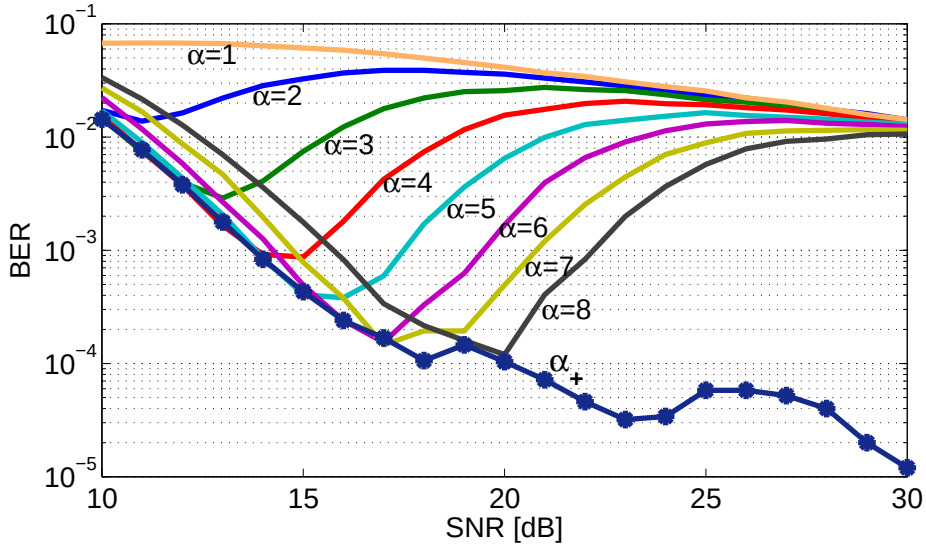


Figure 2.8: BER vs SNR for MCMC detector with $N = 10$.

Now we compare the numerical BER performance of reversible MCMC detector under fixed temperature α and temperature α_+ which changes as a function of SNR. Again, we simulate $N \times N$ MIMO systems with channel matrices containing zero mean i.i.d Gaussian entries, where $N = 10$. Figure 2.8 shows the BER as a function of SNR. 1000 iterations are used and the MCMC detectors are initialized with a random input vector. As SNR increases, all the fixed temperature α choices offer very bad error performance. Our proposed temperature α_+ almost traces out the best error performance at every SNR below 20 dB. When SNR increases beyond 20 dB, the BER performance for α_+ keeps improving as SNR grows.

Now we consider numerical results related to the mixing time of reversible MCMC detectors. In Figure 2.9, we plot the expected number of local minima in a system as the problem dimension N grows. For each N, we generate 100 random channel matrices and for each matrix, we examine the number of local minima by exhaustive search. As the problem dimension N grows, the number of local minima grows rapidly.

In Figure 2.10, we plot the probability of there existing a local minimum as the problem dimension N grows. For each N, we generated 100 random channel matrices and for each matrix, we examined whether there exist local minima by exhaustive search. As N grows, the empirical probability of there existing at least one local minimum approaches 1. It is interesting to see that for $N = 2$, our theoretical result $\frac{1}{3} - \frac{1}{\sqrt{5}} + \frac{2\arctan(\sqrt{\frac{5}{3}})}{\sqrt{5}\pi} \approx 0.15$ matches well with the simulations.

Figures 2.11 and 2.12 show the histograms of the number of local minima

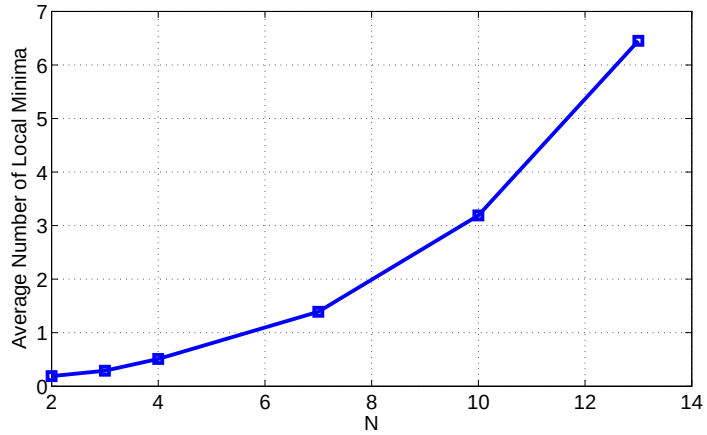


Figure 2.9: Average number of local minima.

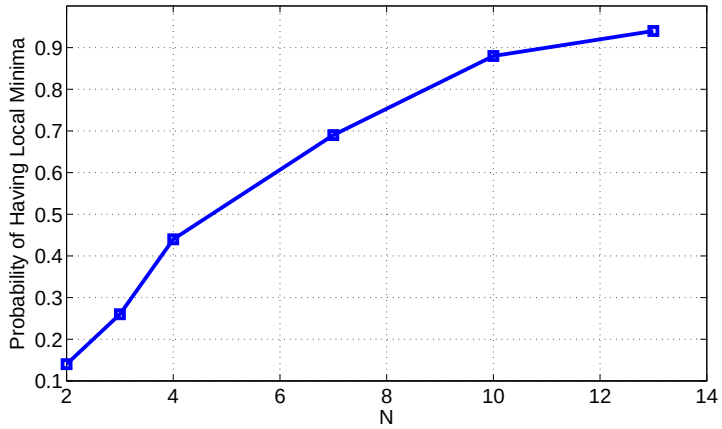


Figure 2.10: The probability of having local minima.

for $N = 10$ and 12 respectively, under SNR $= 10$. For each parameter N, we used exhaustive search to examine the number of local minima in 100 randomly chosen Gaussian channel matrices. Obviously, the average number of local minima increases as N increases, while the frequency of 0 local minima decreases.

Figure 2.13 presents the histograms of the spectral gap when there are 0,

1, 2, and 3 local minima respectively for $N = 5$ and SNR = 10. We generated 10^5 randomly Gaussian channel matrices. In each matrix we examined the number of local minima and calculated the spectral gap when $\alpha^2 = 1$. For all these figures, each bar represents the percentage of matrices which fall in a spectral gap interval of 0.01. We can see that, when there is 0 local minimum, around 50 percent of the matrices' spectral gap fall between 0.19 and 0.2, suggesting these MCMC detectors mix fast. However, when there is at least one local minimum, a high percentage of the matrices have spectral gap values between 0 and 0.01. This percentage increases with the increasing of the number of local minima.

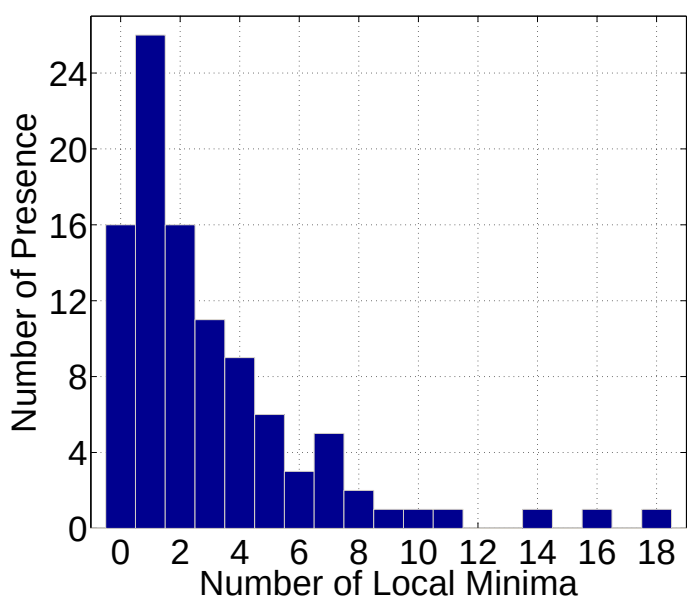


Figure 2.11: Histograms of the number of local minima for N=10.

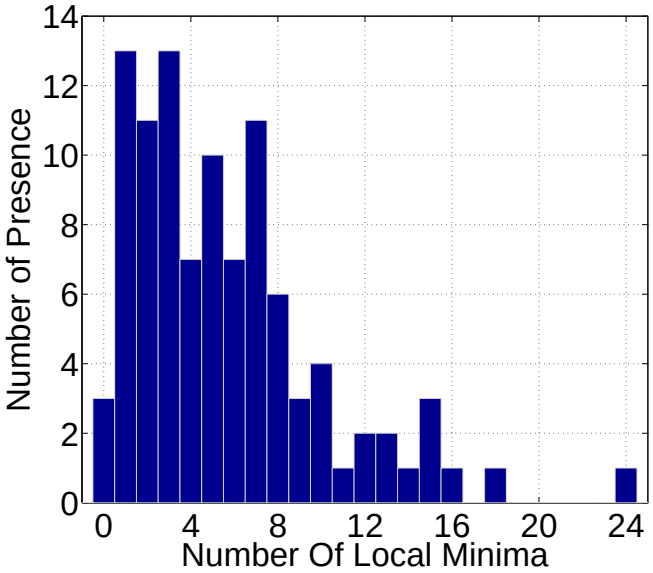


Figure 2.12: Histograms of the number of local minima for N=12.

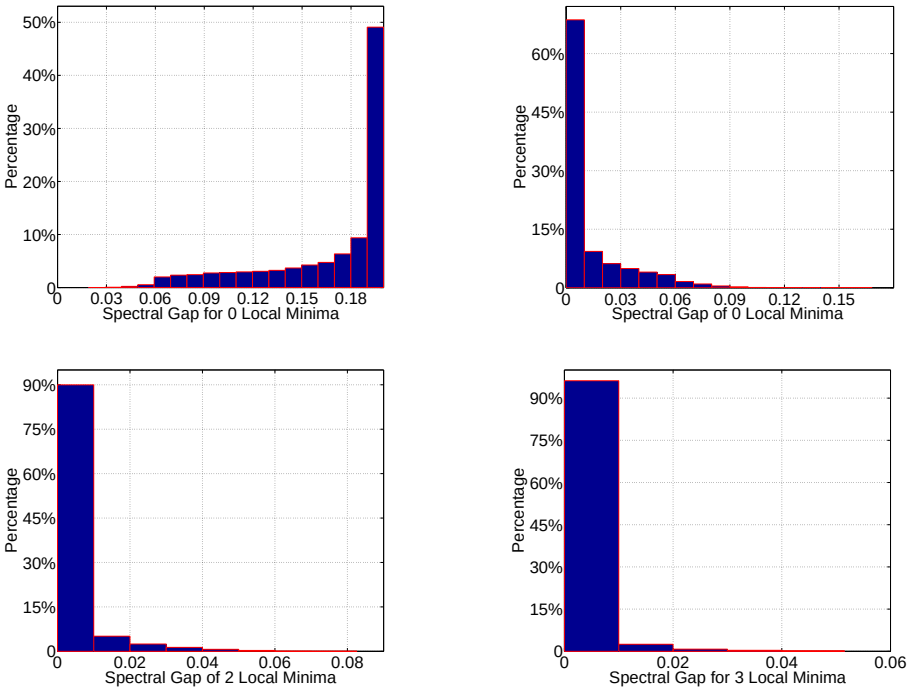


Figure 2.13: Spectral gap with (a) 0 (b) 1 (c) 2 (d) 3 local minima.

2.9 Appendix

2.9.1 Proving Lemma 2.4.1

Lemma 2.4.1 (Gaussian Integral) *Let* $\mathbf{v}$ *and* $\mathbf{x}$ *be independent Gaussian random vectors with distribution* $\mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ *each. Let* $\eta \leq 0$ *be a constant number, and let* a *be an arbitrary constant. We further assume that* $1 - 2a^2\eta(1+2\eta) > 0$*. Then*

$$\mathbb{E}\left\{e^{\eta\left(\|\mathbf{v}+a\mathbf{x}\|^2 - \|\mathbf{v}\|^2\right)}\right\} = \left(\frac{1}{1-2a^2\eta(1+2\eta)}\right)^{N/2}. \tag{2.44}$$

Proof: In order to calculate $\mathbb{E}$, we compute the multivariate integral

$$\mathbb{E}\left\{e^{\eta\left(\|\mathbf{v}+a\mathbf{x}\|^2 - \|\mathbf{v}\|^2\right)}\right\} \tag{2.45a}$$

$$= \int \frac{d\mathbf{x}d\mathbf{v}}{(2\pi)^N} e^{-\frac{1}{2}\begin{bmatrix}\mathbf{v}^T, & \mathbf{x}^T\end{bmatrix}\begin{bmatrix}\mathbf{I}_N & -2a\eta\mathbf{I}_N \\ -2a\eta\mathbf{I}_N & (1-2a^2\eta)\mathbf{I}_N\end{bmatrix}\begin{bmatrix}\mathbf{v} \\ \mathbf{x}\end{bmatrix}} \tag{2.45b}$$

$$= \frac{1}{\det^{1/2}\begin{bmatrix}\mathbf{I}_N & -2a\eta\mathbf{I}_N \\ -2a\eta\mathbf{I}_N & (1-2a^2\eta)\mathbf{I}_N\end{bmatrix}} \tag{2.45c}$$

$$= \frac{1}{\det^{N/2}\begin{bmatrix}1 & -2a\eta \\ -2a\eta & 1-2a^2\eta\end{bmatrix}} \tag{2.45d}$$

$$= \left(\frac{1}{1-2a^2\eta(1+2\eta)}\right)^{N/2}. \tag{2.45e}$$

The integral in $(2.45b)$ is finite because $\eta \leq 0$ and $1 - 2a^2\eta(1+2\eta) \geq 0$ guarantee the

positive definiteness of the matrix involved. ∎

2.9.2 Proving Lemma 2.7.1

Proof. Let N be an even integer. Consider a matrix whose first $\frac{N}{2}$ columns $\mathbf{h}_i$, $1 \leq i \leq \frac{N}{2}$ have unit norms and are orthogonal to each other. For the other $\frac{N}{2}$ columns $\mathbf{h}_i$, $\frac{N}{2} + 1 \leq i \leq N$, $\mathbf{h}_i = -(1+\epsilon)\mathbf{h}_{i-\frac{n}{2}}$, where ϵ is a sufficiently small positive number ($\epsilon < 1$). We also let $\mathbf{y} = \acute{\mathbf{H}}(-\mathbf{1})$, where $\mathbf{1}$ is an all-1 vector. So $-\mathbf{1}$ is a globally minimum point for this ILS problem.

Consider all those vectors $\tilde{\mathbf{x}}'$ which, for any $1 \leq i \leq \frac{N}{2}$, its i-th element and $i + \frac{N}{2}$-th element are either simultaneously $+1$ or simultaneously -1. When ϵ is smaller than 1, we claim that any such a vector except the all -1 vector $\tilde{\mathbf{x}}$, is a local minimum, which shows that there are at least $2^{\frac{N}{2}} - 1$ local minima.

Assume that for a certain $1 \leq i \leq \frac{N}{2}$, the i-th element and $(i + \frac{N}{2})$-th element of $\tilde{\mathbf{x}}'$ are simultaneously -1. Then if we change the i-th element to $+1$, $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2$ increases by 4; and if we change the $(i + \frac{N}{2})$-th element to $+1$, $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2$ increases by $4(1+\epsilon)^2$. This is true because the i-th and $(i + \frac{N}{2})$-th columns are orthogonal to other $(N-2)$ columns.

Similarly, assume that for a certain $1 \leq i \leq \frac{N}{2}$, the i-th element and $(i + \frac{N}{2})$-th element of $\tilde{\mathbf{x}}'$ are simultaneously $+1$. Then if we change the i-th element to -1, $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2$ increases by $4(1+\epsilon)^2 - 4\epsilon^2$; and if we change the $(i + \frac{N}{2})$-th element to -1, $\|\mathbf{y} - \acute{\mathbf{H}}\tilde{\mathbf{x}}'\|^2$ increases by $4 - 4\epsilon^2$. ∎

2.9.3 Proving Lemma 2.7.4

Proof. When $\boldsymbol{\upsilon} = 0$, clearly $\tilde{\mathbf{x}} = (-1, -1)$ is a global minimum point, not a local minimum point. It is also clear that $\tilde{\mathbf{x}} = (-1, 1)$ or $\tilde{\mathbf{x}} = (1, -1)$ can not be a local minimum point since they are neighbors to the global minimum solution. So the only possible local minimum point is $\tilde{\mathbf{x}} = (1, 1)$.

From Lemma 2.7.2, the corresponding necessary and sufficient condition is

$$\mathbf{h}_1^T \mathbf{h}_2 < -\frac{\|\mathbf{h}_1\|^2}{2} = -\frac{\|\mathbf{h}_2\|^2}{2} = -\frac{1}{2}.$$

This means the angle θ between the two 2-dimensional vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ satisfy $\cos(\theta) < -\frac{1}{2}$. Since $\mathbf{h}_1$ and $\mathbf{h}_2$ are two independent uniform randomly sampled vector, the chance for that to happen is $\frac{\pi - \arccos\left(-\frac{1}{2}\right)}{\pi} = \frac{1}{3}$. ∎

2.9.4 Proving Lemma 2.7.5

Proof. When $\boldsymbol{\upsilon} = 0$, clearly $\tilde{\mathbf{x}} = (-1, -1)$ is a global minimum point, not a local minimum point. It is also clear that $\tilde{\mathbf{x}} = (-1, 1)$ or $\tilde{\mathbf{x}} = (1, -1)$ can not be a local minimum point since they are neighbors to the global minimum solution. So the only possible local minimum point is $\tilde{\mathbf{x}} = (1, 1)$.

From Lemma 2.7.2, the corresponding necessary and sufficient condition is

$$\mathbf{h}_1^T \mathbf{h}_2 < -\max\left\{\frac{\|\mathbf{h}_1\|^2}{2}, \frac{\|\mathbf{h}_2\|^2}{2}\right\}.$$

This means the angle θ between the two 2-dimensional vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ satisfy

$$r_1 r_2 \cos(\theta) < -\frac{\max\{r_1^2, r_2^2\}}{2},$$

where r_1 and r_2 are respectively the ℓ_2 norm of $\mathbf{h}_1$ and $\mathbf{h}_2$.

Because the elements of $\acute{\mathbf{H}}$ are independent Gaussian random variables, r_1 and r_2 are thus independent random variables following the Rayleigh distribution

$$p(r_1) = r_1 e^{-\frac{r_1^2}{2}}, p(r_2) = r_2 e^{-\frac{r_2^2}{2}},$$

while θ follows a uniform distribution over $[0, 2\pi)$

By symmetry, for $t \geq 1$,

$$\begin{aligned}
& P(\frac{\max\{r_1^2, r_2^2\}}{r_1 r_2} > t) \\
= \; & 2\int_0^\infty r_1 e^{-\frac{r_1^2}{2}} \times \int_0^{\frac{r_1}{t}} r_2 e^{-\frac{r_2^2}{2}}\, dr_2\, dr_1 \\
= \; & 2\int_0^\infty r_1 e^{-\frac{r_1^2}{2}} \times (1 - e^{-\frac{r_1^2}{2}})\, dr_1 \\
= \; & 2(1 - \int_0^\infty r_1 e^{-(\frac{1}{2}+\frac{1}{2t^2})r_1^2}\, dr_1) \\
= \; & \frac{2}{t^2+1}.
\end{aligned}$$

Since θ is an independent random variable satisfying $\cos(\theta) < -\frac{\max\{r_1^2, r_2^2\}}{2r_1 r_2}$ and $\cos(\theta) \geq -1$, the probability that $\tilde{\mathbf{x}} = (+1, +1)$ is a local minimum is given by

$$P = \int_1^2 (1 - \frac{2}{t^2+1})'(1 - \frac{\arccos(-\frac{t}{2})}{\pi})\, dt$$

$$
\begin{aligned}
&= \int_1^2 \frac{4t}{(t^2+1)^2}\left(1 - \frac{\arccos(-\frac{t}{2})}{\pi}\right) dt. \\
&= \frac{1}{3} - \frac{1}{\sqrt{5}} + \frac{2\arctan(\sqrt{\frac{5}{3}})}{\sqrt{5}\pi},
\end{aligned}
$$

which is approximately 0.145696. ∎

CHAPTER 3
OPTIMAL NON-COHERENT DATA DETECTION FOR MASSIVE SIMO WIRELESS SYSTEMS WITH GENERAL CONSTELLATION: A POLYNOMIAL COMPLEXITY SOLUTION

3.1 The Joint Channel Estimation and Signal Detection Problem

We assume a block fading channel for the SIMO system, and let T denote the channel block length during which the channel coefficients remain constant. In this channel coherence block, the receiver output for a SIMO system with N receive antennas is given by

$$\mathbf{X} = \mathbf{h}\mathbf{s}^* + \mathbf{W}, \tag{3.1}$$

where $\mathbf{h} \in \mathcal{C}^{N\times 1}$ is the SIMO channel vector, $\mathbf{s}^* \in \mathcal{C}^{1\times T}$ is the transmitted symbol sequence, and $\mathbf{W} \in \mathcal{C}^{N\times T}$ is an additive noise matrix whose elements are i.i.d. zero-mean circularly-symmetric complex Gaussian random variables. We also assume the entries of $\mathbf{s}^*$ are i.i.d. symbols from a certain constellation Ω (such as BPSK or 16-QAM).

We assume $\mathbf{h}$ as a deterministic unknown channel with no prior information known about it. Then, the GLRT-optimal JED problem for SIMO systems is given by the following mixed optimization problem:

$$\min_{\mathbf{h},\mathbf{s}^*\in\Omega^T} \|\mathbf{X} - \mathbf{h}\mathbf{s}^*\|^2, \tag{3.2}$$

where Ω^T denotes the set of T-dimensional signal vectors. From [30], the optimization of (3.2) over $\mathbf{h}$ is a least square problem while the optimization of (3.2) over

$\mathbf{s}^*$ is an integer least square problem, since each element of $\mathbf{s}^*$ comes from a fixed constellation Ω. By [29], for any given symbol vector $\mathbf{s}^*$, the channel vector $\mathbf{h}$ that minimizes (3.2) is

$$\hat{\mathbf{h}} = \mathbf{X}\mathbf{s}(\mathbf{s}^*\mathbf{s})^{-1} = \mathbf{X}\mathbf{s}/\|\mathbf{s}\|^2, \tag{3.3}$$

Substituting (3.3) into (3.2), (3.2) is equivalent to the following optimization problem:

$$\min_{\mathbf{s}^*\in\Omega^T} \|\mathbf{X}\underbrace{(I - \frac{1}{\|\mathbf{s}\|^2}\mathbf{s}\mathbf{s}^*)}_{=P_{\mathbf{s}}}\|^2 = \min_{\mathbf{s}^*\in\Omega^T} \mathrm{tr}(\mathbf{X}P_{\mathbf{s}}\mathbf{X}^*) \tag{3.4}$$

$$= \mathrm{tr}(\mathbf{X}\mathbf{X}^*) - \max_{\mathbf{s}^*\in\Omega^T} \frac{\mathbf{s}^*\mathbf{X}^*\mathbf{X}\mathbf{s}}{\|\mathbf{s}\|^2}. \tag{3.5}$$

Hence, for the GLRT-optimal JED, we need to maximize $\frac{\mathbf{s}^*\mathbf{X}^*\mathbf{X}\mathbf{s}}{\|\mathbf{s}\|^2}$ in (3.5). This maximization depends on whether the constellation of the transmitted signal is constant or not. For constant-modulus constellations, since $\|\mathbf{s}\|^2$ is fixed, the authors of [30] changed (3.5) to an equivalent problem:

$$\max_{\mathbf{s}^*\in\Omega^T} \mathbf{s}^*\mathbf{X}^*\mathbf{X}\mathbf{s}. \tag{3.6}$$

Now that (3.6) is an integer quadratic maximization problem, the authors of [30] further transformed (3.6) into another equivalent integer quadratic minimization problem:

$$\min_{\mathbf{s}\in\Omega^T} \mathbf{s}^*(\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N})\mathbf{s}, \tag{3.7}$$

where ρ is a slightly larger value than the maximum eigenvalue of $\frac{\mathbf{X}^*\mathbf{X}}{N}$. One way of solving the integer quadratic minimization problem (3.7) is exhaustive search over the entire signal space Ω^T. However, exhaustive search has a complexity linear in N but exponential in T [31].

The sphere decoder in [29] can efficiently solve (3.7) under high SNRs, since the sphere decoder restricts search only to the lattice points within a search radius r. More specifically, the sphere decoder only examines sequences $\mathbf{s}^*$ satisfying

$$\mathbf{s}^*(\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N})\mathbf{s} \leq r^2. \tag{3.8}$$

The efficiency of the sphere decoder to solve (3.8) depends on the search radius r chosen probabilistically based on SNR [29], [58]. The sphere decoder in [29], [58] has the limitation of providing the GLRT-optimal solution only for constant-modulus modulations. The reason is that the optimization problem (3.5) is an integer quadratic maximization problem only for constant-modulus constellations such as BPSK and QPSK, but not for nonconstant-modulus constellations, because different sequences $\mathbf{s}^*$ can have different energies $\|\mathbf{s}\|^2$.

3.2 GLRT-Optimal JED Algorithm for General Constellations

In this section we provide the first efficient GLRT-optimal JED algorithm for massive SIMO systems with general constellations, including nonconstant-modulus modulations.

To describe our algorithm, we first represent the set of possible sequences by a tree structure of T layers. The root node at layer-0 corresponds to an empty sequence. We represent any sequence $\mathbf{s}^*_{i:T}$ as a layer-$(T-i+1)$ tree node, where $\mathbf{s}^*_{i:T} = (\mathbf{s}^*_i, \mathbf{s}^*_{i+1}, ..., \mathbf{s}^*_T)$ is a partial sequence. The tree node representing $\mathbf{s}^*_{i+1:T}$ is the parent node of the tree node representing $\mathbf{s}^*_{i:T}$.

The GLRT-optimal JED maximizes $\frac{\mathbf{s}^*\mathbf{X}^*\mathbf{X}\mathbf{s}}{\|\mathbf{s}\|^2}$, which is equivalent to

$$\min_{\mathbf{s}^* \epsilon \Omega^T} \frac{\mathbf{s}^*(\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N})\mathbf{s}}{\|\mathbf{s}\|^2},$$

where ρ is a slightly larger value than the maximum eigenvalue of $\frac{\mathbf{X}^*\mathbf{X}}{N}$. This is because $\mathbf{s}^*\mathbf{s}/\|\mathbf{s}\|^2$ is independent of the energy of $\mathbf{s}$.

Let us factorize $\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N}$ using the Cholesky decomposition as

$$\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N} = \mathbf{R}^*\mathbf{R}, \tag{3.9}$$

where $\mathbf{R}$ is a $T \times T$ upper triangular matrix.

Then we have

$$\begin{aligned} \min_{\mathbf{s}^* \epsilon \Omega^T} \frac{\mathbf{s}^*(\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N})\mathbf{s}}{\|\mathbf{s}\|^2} &= \min_{\mathbf{s}^* \epsilon \Omega^T} \frac{\mathbf{s}^*\mathbf{R}^*\mathbf{R}\mathbf{s}}{\|\mathbf{s}\|^2} \\ &= \min_{\mathbf{s}^* \epsilon \Omega^T} \frac{\|\mathbf{R}\mathbf{s}\|^2}{\|\mathbf{s}\|^2}. \end{aligned} \tag{3.10}$$

We define $M_{\mathbf{s}^*} = \|\mathbf{R}\mathbf{s}\|^2$, namely,

$$M_{\mathbf{s}^*} = \sum_{i=1}^{T} |\sum_{k=i}^{T} \mathbf{R}_{i,k}\mathbf{s}_k|^2, \tag{3.11}$$

where $\mathbf{R}_{i,k}$ is the element of $\mathbf{R}$ in the i-th row and k-th column. For any i between 1 and T, we further define the unscaled metric

$$M_{\mathbf{s}^*_{i:T}} = \sum_{t=i}^{T} |\sum_{k=t}^{T} \mathbf{R}_{t,k}\mathbf{s}_k|^2 = \|\mathbf{R}_{i:T,i:T}\mathbf{s}_{i:T}\|_2^2. \tag{3.12}$$

for any partial sequence $\mathbf{s}^*_{i:T} = (\mathbf{s}^*_i, \mathbf{s}^*_{i+1}, ..., \mathbf{s}^*_T)$.

Since different sequences may have different energies, the $\|\mathbf{s}\|^2$ term in (3.10) prevents us from solving this minimization problem using the regular sphere decoder approach. Instead, we will lower bound $\frac{\|\mathbf{R}\mathbf{s}\|^2}{\|\mathbf{s}\|^2}$ for partial sequences $\mathbf{s}_{i:T}$, taking sequence energy into consideration.

To lower bound $\frac{\|\mathbf{R}\mathbf{s}\|^2}{\|\mathbf{s}\|^2}$, we will divide the sequence $\mathbf{s}$ into two parts $\mathbf{s}_{1:i-1}$ and $\mathbf{s}_{i:T}$. For the partial sequence $\mathbf{s}^*_{i:T}$, we define its metric $\bar{M}_{\mathbf{s}^*_{i:T}}$ as,

$$\bar{M}_{\mathbf{s}^*_{i:T}} = \frac{M_{\mathbf{s}^*_{i:T}}}{|\mathbf{s}_{max}|^2(i-1) + \|\mathbf{s}^*_{i:T}\|^2}. \tag{3.13}$$

where $M_{\mathbf{s}^*_{i:T}}$ is defined as in (3.12), and we use $|\mathbf{s}_{max}|^2$ to denote the maximum energy of a single constellation symbol. [1] In fact, $\bar{M}_{\mathbf{s}^*_{i:T}}$ is a lower bound on

[1]For example, the 16-QAM constellation Ω has 16 points $a + bj$, where $a \in \{\pm 1, \pm 3\}$ and $b \in \{\pm 1, \pm 3\}$. Thus, the maximum energy of a constellation point in 16-QAM is $3^2 + 3^2 = 18$.

$\frac{M_{\mathbf{s}^*_{i:T}}}{\|\mathbf{s}^*_{1:i-1}\|^2+\|\mathbf{s}^*_{i:T}\|^2}$ or $\frac{\|\mathbf{Rs}\|^2}{\|\mathbf{s}\|^2}$, as stated by the following lemma, the proof of which is given in Appendix 3.8.1.

Lemma 3.2.1. *For every sequence* s *and any index* i, $\bar{M}_{\mathbf{s}^*_{i:T}} \leq \frac{\|\mathbf{Rs}\|^2}{\|\mathbf{s}\|^2}$.

Motivated by this lemma, we propose Algorithm 2 for the GLRT-optimal JED for SIMO wireless systems with general constellations.[2]

Algorithm 2: The GLRT-optimal JED algorithm for general constellations.

Input:adius r, received signal $\mathbf{X}$, constellation Ω and a $1 \times T$ index vector I
Output: The GLRT-optimal decoded sequence $\hat{\mathrm{s}}^*$

1. (preprocessing) Compute $\mathbf{X}^*\mathbf{X}$ and the Cholesky decomposition of $(\rho I - \mathbf{X}^*\mathbf{X}/N) = \mathbf{R}^*\mathbf{R}$.
2. (start of tree search) Set $i \leftarrow T$, $I(i) \leftarrow 1$ and set $\mathrm{s}^*_i \leftarrow \Omega(I(i))$.
3. (Computing the bounds) Compute the metric $\bar{M}_{\mathbf{s}^*_{i:T}}$. If $\bar{M}_{\mathbf{s}^*_{i:T}} > r^2$, go to 4; else, go to 5;
4. (Backtracking) Find the smallest j such that $i \leq j \leq T$ such that $I(j) < |\Omega|$. If there exists such j, set $i \leftarrow j$ and go to 6; else go to 7.
5. If $i = 1$, store this current s^* by setting $\hat{\mathrm{s}}^* \leftarrow \mathrm{s}^*$, update $r^2 \leftarrow \bar{M}_{\mathbf{s}^*_{i:T}}$ and go to 4; else set $i \leftarrow (i-1)$, $I(i) \leftarrow 1$ and $\mathrm{s}^*_i \leftarrow \Omega(I(i))$, go to 3.
6. Set $I(i) \leftarrow (I(i)+1)$ and $\mathrm{s}^*_i \leftarrow \Omega(I(i))$. Go to 3.
7. If any sequence s^* is ever found in Step 5, output the latest stored full-length sequence $\hat{\mathrm{s}}^*$ as the GLRT-optimal solution; otherwise, increase r (for example, double r) and go to 2.

---

[2]It is worth noticing that, before the tree search stage, the GLRT-optimal JED algorithm performs the preprocessing, when $\mathbf{X}^*\mathbf{X}/N$, the maximum eigenvalue ρ of $\mathbf{X}^*\mathbf{X}/N$, and the Cholesky decomposition of $(\rho I - \mathbf{X}^*\mathbf{X}/N)$ are computed.

Theorem 3.2.2. *Algorithm 2 outputs the GLRT-optimal sequence* $\hat{\mathbf{s}}^*$*, under general modulations.*

Proof. We note that the algorithm will terminate after a finite number of doubling the search radius r. Moreover, after the final time of doubling radius r, the radius will not increase anymore in the subsequent search. Let $\hat{\mathbf{s}}^*$ be the final sequence output by the algorithm. We must have, when the algorithm terminates, $r^2 = \bar{M}_{\hat{\mathbf{s}}^*_{1:T}}$. We claim that any sequence $\tilde{\mathbf{s}}^*$ other than $\hat{\mathbf{s}}^*$ must have a partial sequence with metric no smaller than $\bar{M}_{\hat{\mathbf{s}}^*_{1:T}}$; otherwise, the algorithm will explore the full length sequence $\tilde{\mathbf{s}}^*$, and end up giving a final $r^2 < \bar{M}_{\hat{\mathbf{s}}^*_{1:T}}$, which is a contradiction.

Thus, for any sequence $\tilde{\mathbf{s}}^* \neq \hat{\mathbf{s}}^*$, there must be an index i (between 1 and T) such that, for the partial sequence $\tilde{\mathbf{s}}^*_{i:T}$, $\bar{M}_{\tilde{\mathbf{s}}^*_{i:T}} \geq \bar{M}_{\hat{\mathbf{s}}^*_{1:T}}$. This implies $\bar{M}_{\tilde{\mathbf{s}}^*_{1:T}}$ is no smaller than $\bar{M}_{\hat{\mathbf{s}}^*_{1:T}}$, because $\bar{M}_{\tilde{\mathbf{s}}^*_{i:T}}$ is a lower bound on $\bar{M}_{\tilde{\mathbf{s}}^*_{1:T}}$. This proves that indeed $\hat{\mathbf{s}}^*$ has the smallest metric $\bar{M}_{\hat{\mathbf{s}}^*_{1:T}}$. ∎

Remarks: When the constellation is constant-modulus, $\|\mathbf{s}\|^2$ will be a constant, and Algorithm 2 reduces to the sphere decoder algorithm in [29], although with a different choice of search radius r.

3.2.1 Choosing the Initial Radius r

Choosing the initial radius r has a big influence on the complexity of this GLRT-optimal algorithm. If r^2 is chosen bigger than the metric of every sequence $\tilde{\mathbf{s}} \in |\Omega|^T$, Algorithm 2 may visit all the tree nodes under that radius. If r^2 is too small, the optimal sequence may have a metric larger than r^2, and Algorithm 2

will search again under a new larger radius, resulting in a higher computational complexity.

In [29], [31], the authors derived the search radius r for the conventional sphere decoder under constant-modulus constellations. [29], [31] chose the search radius probabilistically such that the radius ensures the transmitted sequence has a metric no bigger than r^2 with high probability. However, the choice of the initial search radius in [29] is for a fixed number of receive antennas, and for high signal-to-noise ratio (SNR); it is not clear how the expected complexity scales with channel coherence time T under the probabilistically chosen search radius. Moreover, it is unknown what initial search radius should be used for massive SIMO systems with a fixed SNR (not necessarily high).

In this chapter, we give a novel choice of the search radius r for the new GLRT-optimal algorithm. This search radius is optimized for massive SIMO systems with general constellations and any fixed SNR.

Lemma 3.2.3. *Let $|\mathbf{s}_{max}|^2$ and $|\mathbf{s}_{min}|^2$ be respectively the largest and the smallest possible energy of a constellation point from Ω. Suppose the channel $\mathbf{h}$ has i.i.d. entries following circularly-symmetric complex Gaussian distributions with zero mean and unit variance. For massive SIMO systems, let us set the initial search radius r^2 to be any nonzero positive constant smaller than*

$$\frac{|\mathbf{s}_{min}|^4 D_{min}}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2 |\mathbf{s}_{max}|^2},$$

where $D_{min} = \min_{a,b\in\Omega} \|a - b\|^2$ is the minimum pairwise squared distance between any

two constellation points. Then the true transmitted sequence $\mathbf{s}$ *has a metric smaller than* r^2 *with high probability, as the number of antennas* $N \to \infty$.

For instance, when the constellation is constant-modulus, $\frac{|\mathbf{s}_{min}|^4 D_{min}}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2 |\mathbf{s}_{max}|^2} = \frac{D_{min}}{2}$. As another example, if we consider the 16-QAM modulation, we have $\frac{|\mathbf{s}_{min}|^2}{|\mathbf{s}_{max}|^2} = \frac{1^2+1^2}{3^2+3^2} = \frac{1}{9}$. Then

$$\frac{|\mathbf{s}_{min}|^4 D_{min}}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2 |\mathbf{s}_{max}|^2} = \frac{4D_{min}}{18^2 + 18 \times 2} = \frac{D_{min}}{90}.$$

Choosing search radius r as in Lemma 3.2.3 ensures that the true transmitted signal is inside the search radius with high probability. We provide the derivation of this radius (namely Lemma 3.2.3) in Appendix 3.8.11. However, we encourage the reader to postpone reading the proof of Lemma 3.2.3 until Section 3.5, after basic calculations for the Cholesky decompositions in Section 3.4.

In the next section, we derive the expected complexity for Algorithm 2. We show under this new radius, Algorithm 2 has polynomial expected computational complexity. We also remark that, for downlink beamforming, one can use the $\hat{\mathbf{h}}$ generated from (3.3), where $\mathbf{s}^*$ is the output from our GLRT-optimal JED algorithm.

3.3 Algorithm Computational Complexity: N Grows Independently of T

The computational complexity of Algorithm 2 for massive SIMO systems is mainly determined by the number of visited nodes in each layer. By "visited nodes", we mean the partial sequences $\mathbf{s}^*_{i:T}$ for which the metrics $\bar{M}_{\mathbf{s}^*_{i:T}}$ are computed in the algorithm. The fewer the visited nodes, the lower computational complexity of the GLRT-optimal JED algorithm.

In this section, we will show in Theorem 3.3.1 that, for Algorithm 2, the expected number of visited nodes in each layer will converge to a constant for a sufficiently large number of receive antennas. For our theoretical analysis, we consider a special case of Algorithm 2, where (in its Step 7) the search radius r is increased to ∞ if the GLRT-optimal sequence is not found under the current radius r. However, we remark that, in practical implementations, one is free to use other protocols of increasing radius r, including doubling r after failure to find the optimal sequence. Moreover, for concise presentations, we choose to state and prove Theorem 3.3.1 only for constant-modulus constellations. Generalizations of Theorem 3.3.1 to nonconstant-modulus constellations can be found in Section 3.5.

Theorem 3.3.1. *Let us assume that the constellation Ω is constant-modulus. Let r^2 be a positive constant smaller than $\frac{D_{min}}{2}$, where $D_{min} = \min_{a,b\in\Omega} \|a-b\|^2$ is the minimum pairwise squared distance between any two constellation points. We also assume $\mathbf{s}_T$ is known to the receiver to resolve the phase ambiguity. Then for the tree search stage of Algorithm 2, the expected number of visited points at layer i converges to $|\Omega|$ for $i \geq 2$, as $N \to \infty$. Algorithm 2 only visits one tree node at layer 1. Moreover, the average overall computational complexity of Algorithm 2 is $O(NT^2 + T^3)$, when, for any fixed T, N goes to infinity.*

Our proof strategy is to show the computational complexity of our algorithm for matrix $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$, where ρ_E is the maximum eigenvalue of $\frac{E[\mathbf{X}^*\mathbf{X}]}{N}$. Then we prove the expected complexity for matrix $\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N}$, by using the fact that $\frac{\mathbf{X}^*\mathbf{X}}{N}$ converges to $\frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ in massive SIMO systems.

Proof of Theorem 3.3.1. The number of visited nodes at layer $(T-i+1)$ $(1 \le i \le T-1)$ in Algorithm 2 is equal to $|\Omega|$, if there is one and only one tree node $\widetilde{\mathbf{s}}^*_{(i+1):T}$ such that $\bar{M}_{\widetilde{\mathbf{s}}^*_{(i+1):T}} \le r^2$. In fact, we will prove that, the transmitted $\mathbf{s}^*_{(i+1):T}$ will be the only sequence satisfying $\bar{M}_{\widetilde{\mathbf{s}}^*_{(i+1):T}} \le r^2$, with high probability as the number of receive antennas $N \to \infty$.

To show this result, we first derive $E[\mathbf{X}^*\mathbf{X}]$, and factorize $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ using the Cholesky decomposition. Using the upper triangular matrix generated from the Cholesky decomposition, we will then show that the true transmitted $\mathbf{s}^*_{(i+1):T}$ will be the only sequence satisfying $\bar{M}_{\mathbf{s}^*_{(i+1):T}} \le r^2$ under $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$. In fact, we can first write out (3.1) as

$$\mathbf{X} = [\mathbf{s}_1^*\mathbf{h} + \mathbf{w}_1 \;\; \mathbf{s}_2^*\mathbf{h} + \mathbf{w}_2 \;\; \cdots \;\; \mathbf{s}_T^*\mathbf{h} + \mathbf{w}_T],$$

where $\mathbf{w}_i$ is the i-th column of $\mathbf{W}$. Then $E[\mathbf{X}^*\mathbf{X}]$ is equal to

$$E\left\{\begin{bmatrix} (\mathbf{s}_1^*\mathbf{h} + \mathbf{w}_1)^* \\ (\mathbf{s}_2^*\mathbf{h} + \mathbf{w}_2)^* \\ \vdots \\ (\mathbf{s}_T^*\mathbf{h} + \mathbf{w}_T)^* \end{bmatrix} \begin{bmatrix} \mathbf{s}_1^*\mathbf{h} + \mathbf{w}_1 & \mathbf{s}_2^*\mathbf{h} + \mathbf{w}_2 & \cdots & \mathbf{s}_T^*\mathbf{h} + \mathbf{w}_T \end{bmatrix}\right\} \tag{3.14}$$

Since the entries of $\mathbf{h}$ are independent complex Gaussian random variables with zero mean and unit variance, $E[\mathbf{h}^*\mathbf{h}] = E[\sum_{i=1}^N \mathbf{h}_i^*\mathbf{h}_i] = N$. After some algebra, we

have

$$E[\mathbf{X}^*\mathbf{X}]/N = \begin{bmatrix} \mathbf{s}_1\mathbf{s}_1^* + \sigma_w^2 & \mathbf{s}_1\mathbf{s}_2^* & \cdots & \mathbf{s}_1\mathbf{s}_T^* \\ \mathbf{s}_2\mathbf{s}_1^* & \mathbf{s}_2\mathbf{s}_2^* + \sigma_w^2 & \cdots & \mathbf{s}_2\mathbf{s}_T^* \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{s}_T\mathbf{s}_1^* & \mathbf{s}_T\mathbf{s}_2^* & \cdots & \mathbf{s}_T\mathbf{s}_T^* + \sigma_w^2 \end{bmatrix}. \tag{3.15}$$

We can see that (3.15) is a Hermitian matrix with a full column rank. The maximum eigenvalue of $\frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ is $\rho_E = T + \sigma_w^2$. Now we can write $\grave{\mathbf{A}} = \rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ as

$$\grave{\mathbf{A}} = \begin{bmatrix} T - \mathbf{s}_1\mathbf{s}_1^* & -\mathbf{s}_1\mathbf{s}_2^* & \cdots & -\mathbf{s}_1\mathbf{s}_T^* \\ -\mathbf{s}_2\mathbf{s}_1^* & T - \mathbf{s}_2\mathbf{s}_2^* & \cdots & -\mathbf{s}_2\mathbf{s}_T^* \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{s}_T\mathbf{s}_1^* & -\mathbf{s}_T\mathbf{s}_2^* & \cdots & T - \mathbf{s}_T\mathbf{s}_T^* \end{bmatrix} \tag{3.16}$$

Using the Cholesky decomposition in [59], we can decompose $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ into $\grave{\mathbf{R}}^*\grave{\mathbf{R}}$ where $\grave{\mathbf{R}}$ is the upper triangular matrix of Cholesky decomposition represented by

$$\grave{\mathbf{R}} = \begin{bmatrix} \grave{\mathbf{R}}_{1,1} & \grave{\mathbf{R}}_{1,2} & \grave{\mathbf{R}}_{1,3} & \cdots & \grave{\mathbf{R}}_{1,T} \\ 0 & \grave{\mathbf{R}}_{2,2} & \grave{\mathbf{R}}_{2,3} & \cdots & \grave{\mathbf{R}}_{2,T} \\ 0 & 0 & \grave{\mathbf{R}}_{3,3} & \cdots & \grave{\mathbf{R}}_{3,T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \grave{\mathbf{R}}_{T,T} \end{bmatrix}.$$

One can calculate $\grave{\mathbf{R}}$ recursively by starting with $i = 1$. For each i, $\grave{\mathbf{R}}_{i,i} = \sqrt{\grave{\mathbf{A}}_{i,i} - \sum_{k=1}^{i-1} \grave{\mathbf{R}}_{k,i}\grave{\mathbf{R}}_{k,i}^*}$, where $\grave{\mathbf{A}}_{i,i}$ is the i-th diagonal entry of $(\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N})$; moreover, for each $j > i$,

$$\grave{\mathbf{R}} = \begin{bmatrix} \sqrt{T-1} & \frac{-(\mathbf{s}_1\mathbf{s}_2^*)}{\sqrt{T-1}} & \frac{-(\mathbf{s}_1\mathbf{s}_3^*)}{\sqrt{T-1}} & \cdots & \frac{-(\mathbf{s}_1\mathbf{s}_T^*)}{\sqrt{T-1}} \\ 0 & \sqrt{\frac{T(T-2)}{T-1}} & -(\mathbf{s}_2\mathbf{s}_3^*)\sqrt{\frac{T}{(T-1)(T-2)}} & \cdots & -(\mathbf{s}_2\mathbf{s}_T^*)\sqrt{\frac{T}{(T-1)(T-2)}} \\ 0 & 0 & \sqrt{\frac{T(T-3)}{T-2}} & \cdots & -(\mathbf{s}_3\mathbf{s}_T^*)\sqrt{\frac{T}{(T-2)(T-3)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{\frac{T}{2}} & -(\mathbf{s}_{T-1}\mathbf{s}_T^*)\sqrt{\frac{T}{2}} \\ 0 & \cdots & 0 & 0 & \sqrt{\frac{T(T-T)}{(T-T+1)}} \end{bmatrix}. \quad (3.17)$$

$\grave{\mathbf{R}}_{i,j} = \frac{1}{\grave{\mathbf{R}}_{i,i}}(\grave{\mathbf{A}}_{i,j} - (\sum_{k=1}^{i-1} \grave{\mathbf{R}}_{k,i}\grave{\mathbf{R}}_{k,j}^*)^*)$, where $\grave{\mathbf{A}}_{i,j}$ is an entry of $(\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N})$ with row index i, and column index j.

Using this recursive relation, we have the following lemma about the elements of $\grave{\mathbf{R}}$. The proof of Lemma 3.3.2 is given in Appendix 3.8.2.

Lemma 3.3.2. *Let $\grave{\mathbf{A}}$ be given in (3.16). Then for every $1 \le i \le T$,*

$$\grave{\mathbf{R}}_{i,i} = \sqrt{\frac{T(T-i)}{T-i+1}}$$

and for every $j > i$,

$$\grave{\mathbf{R}}_{i,j} = -\mathbf{s}_i\mathbf{s}_j^*\sqrt{\frac{T}{(T-i+1)(T-i)}}.$$

From this lemma, we notice that $\grave{\mathbf{R}}_{T,T}$=0, and the smallest value for $\grave{\mathbf{R}}_{i,i}$, $1 \le i \le T-1$, is $\sqrt{\frac{T}{2}}$ when $i = T-1$. We illustrate the whole matrix $\mathbf{R}$ in (3.17). Now let us use $\grave{\mathbf{R}}$ in (3.17) as the underlying upper triangular matrix in calculating $M_{\mathbf{s}_{1:T}^*}$. Based on (3.11), $M_{\mathbf{s}_{1:T}^*}$ (under matrix $\grave{\mathbf{R}}$ satisfies

$$M_{\mathbf{s}^*_{1:T}} = \mathbf{s}^* \grave{\mathbf{A}} \mathbf{s} = \mathbf{s}^*(TI - \mathbf{s}\mathbf{s}^*)\mathbf{s} = T^2 - T^2 = 0, \tag{3.18}$$

since $\mathbf{s}^*\mathbf{s} = T$. Because $M_{\mathbf{s}^*} = \sum_{i=1}^{T} |\sum_{k=i}^{T} \grave{\mathbf{R}}_{i,k}\mathbf{s}_k|^2$, from (3.18), we must have $|\sum_{k=i}^{T} \grave{\mathbf{R}}_{i,k}\mathbf{s}_k|^2 = 0$ for every $1 \leq i \leq T$. This, in turn, implies that $M_{\mathbf{s}^*_{i:T}} = 0$, and $\sum_{k=i}^{T} \grave{\mathbf{R}}_{i,k}\mathbf{s}_k = 0$ for every $1 \leq i \leq T$. On the other hand, we have the following lemma (the proof of which is provided in Appendix 3.8.11) about the metric of a non-transmitted sequence, which applies to general constellation Ω.

Lemma 3.3.3. *Let Ω be any constellation with $D_{min} = \min_{a,b\in\Omega} \|a - b\|^2$ as the minimum pairwise squared distance between any two constellation points. Let $\mathbf{s}^*$ be the transmitted data sequence. Let us use the upper triangular matrix $\grave{\mathbf{R}}$ generated from the Cholesky decomposition of $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$, in calculating the sequence metric. For any $\widetilde{\mathbf{s}}^*$ such that $\widetilde{\mathbf{s}}^* \neq \mathbf{s}^*$, $\bar{M}_{\widetilde{\mathbf{s}}^*_{j:T}} \geq \frac{D_{min}|\mathbf{s}_{min}|^4}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2|\mathbf{s}_{max}|^2}$ at any layer $j \leq i$, where i is the largest integer such that $\mathbf{s}^*_i \neq \widetilde{\mathbf{s}}^*_i$.*

For constant-modulus constellations, according to Lemma 3.3.3, for any other $\widetilde{\mathbf{s}} \neq \mathbf{s}$, $\bar{M}_{\widetilde{\mathbf{s}}^*_{i:T}} \geq \frac{TD_{min}}{2}$, where i is the integer closest to T such that $\mathbf{s}^*_i \neq \widetilde{\mathbf{s}}^*_i$. Under the assumption that $\mathbf{X}^*\mathbf{X} = E[\mathbf{X}^*\mathbf{X}]$, Algorithm 2 will visit only 1 tree node at layer T, namely $\mathbf{s}^*_T$, whose metric is equal to 0. (We have only 1 tree node at layer T because $\mathbf{s}^*_T$ is predetermined, in order to resolve phase ambiguity.) At layer i, where $i < T$, we only have one sequence $\widetilde{\mathbf{s}}^*_{i:T} = \mathbf{s}^*_{i:T}$ such that $\bar{M}_{\widetilde{\mathbf{s}}^*_{i:T}} < r^2$, when r^2 is picked to be any positive constant smaller than $\frac{D_{min}}{2}$. This proves Theorem 3.3.1, under the assumption that $\mathbf{X}^*\mathbf{X} = E[\mathbf{X}^*\mathbf{X}]$.

Now we proceed to prove that, with high probability, $\mathbf{X}^*\mathbf{X}/N$ is close to $E[\mathbf{X}^*\mathbf{X}]/N$, and thus the expected number of visited nodes under $\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N}$ is very close to the case under $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$. In fact, $\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N}$ can be written as the average of N independent random variables under the considered channel model:

$$\begin{aligned}\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N} &= \frac{(\mathbf{s}_i^*\mathbf{h} + \mathbf{w}_i)^*(\mathbf{s}_j^*\mathbf{h} + \mathbf{w}_j)}{N} \\ &= \frac{\sum_{k=1}^{N}(\mathbf{s}_i^*\mathbf{h}_k + \mathbf{w}_{k,i})^*(\mathbf{s}_j^*\mathbf{h}_k + \mathbf{w}_{k,j})}{N}\end{aligned} \tag{3.19}$$

where $\mathbf{w}_i$ is the i-th column of $\mathbf{W}$. Then we can find the expectation and the variance of (3.19) as follows:

$$E[\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N}] = \begin{cases} 1 + \sigma_w^2, & \text{if } i = j \\ \mathbf{s}_i\mathbf{s}_j^*, & \text{otherwise} \end{cases} \tag{3.20}$$

$$var(\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N}) = (1 + 2\sigma_w^2 + \sigma_w^4)/N. \tag{3.21}$$

We provide the proof of (3.21) in Appendix 3.8.4.

The weak law of large numbers states that the sample mean of a random variable converges to its expectation in probability. Thus, for any pair $1 \le i, j \le N$, for any constant $\xi > 0$ and $\epsilon > 0$, as $N \to \infty$, we have

$$P(|\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N} - \frac{E[(\mathbf{X}^*\mathbf{X})_{i,j}]}{N}| \ge \varepsilon) \le \xi. \tag{3.22}$$

This means that, for any $\xi > 0$ and $\epsilon > 0$, as $N \to \infty$, we have

$$P(\|\frac{\mathbf{X}^*\mathbf{X}}{N} - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}\|_F \leq \varepsilon) \geq 1 - \xi, \tag{3.23}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Since ρ is the maximum eigenvalue of $\frac{\mathbf{X}^*\mathbf{X}}{N}$, by the triangular inequality for the spectral norm, we have

$$|\rho - \rho_E| \leq \|\frac{\mathbf{X}^*\mathbf{X}}{N} - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}\|_2.$$

Since

$$\|\frac{\mathbf{X}^*\mathbf{X}}{N} - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}\|_2 \leq \|\frac{\mathbf{X}^*\mathbf{X}}{N} - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}\|_F,$$

we have

$$|\rho - \rho_E| \leq \|\frac{\mathbf{X}^*\mathbf{X}}{N} - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}\|_F \leq \epsilon,$$

with probability at least $1 - \xi$, as $N \to \infty$.

Using the triangular inequality for the spectral norm and the Frobenius norm, we have

$$\|\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N} - (\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N})\|_2 \leq 2\epsilon,$$

and

$$\|\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N} - (\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N})\|_F \leq (\sqrt{T} + 1)\epsilon,$$

with probability at least $1 - \xi$, as $N \to \infty$.

Now since the Cholesky decomposition of $(\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N})$ is continuous at the point $\grave{\mathbf{A}} = \rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$, for any $\epsilon > 0$ and $\xi > 0$, as $N \to \infty$,

$$\|\mathbf{R} - \grave{\mathbf{R}}\|_F \leq \epsilon$$

holds true with probability at least $1 - \xi$. Let us use $\mathbf{R}_{i:T}$ ($\grave{\mathbf{R}}_{i:T}$) to denote a matrix composed of the i-th to T-th rows of matrix $\mathbf{R}$ ($\grave{\mathbf{R}}$). Then as $N \to \infty$, for any full-length sequence $\widetilde{\mathbf{s}}^*$, with probability at least $1 - \xi$,

$$|\sqrt{M^{\grave{\mathbf{R}}}_{\widetilde{\mathbf{s}}^*_{i:T}}} - \sqrt{M^{\mathbf{R}}_{\widetilde{\mathbf{s}}^*_{i:T}}}| \leq \|(\mathbf{R}_{i:T} - \grave{\mathbf{R}}_{i:T})\widetilde{\mathbf{s}}\|_2 \leq \|\widetilde{\mathbf{s}}\|\|\mathbf{R} - \grave{\mathbf{R}}\|_F,$$

which is no bigger than $\|\widetilde{\mathbf{s}}\|\epsilon$. Note here the superscripts $\mathbf{R}$ and $\grave{\mathbf{R}}$ in $M^{\grave{\mathbf{R}}}_{\widetilde{\mathbf{s}}^*_{i:T}}$ and $M^{\mathbf{R}}_{\widetilde{\mathbf{s}}^*_{i:T}}$ describe which upper triangular matrix is used in calculations.

We recall that the metric $\bar{M}_{\widetilde{\mathbf{s}}^*_{i:T}}$ is defined as $M_{\widetilde{\mathbf{s}}^*_{i:T}}/(|\mathbf{s}_{max}|^2(i-1) + \|\mathbf{s}^*_{i:T}\|^2)$. Then we have

$$|\sqrt{\bar{M}^{\grave{\mathbf{R}}}_{\widetilde{\mathbf{s}}^*_{i:T}}} - \sqrt{\bar{M}^{\mathbf{R}}_{\widetilde{\mathbf{s}}^*_{i:T}}}| \tag{3.24}$$

$$\leq \frac{\|\widetilde{\mathbf{s}}\|\epsilon}{\sqrt{|\mathbf{s}_{max}|^2(i-1) + \|\mathbf{s}^*_{i:T}\|^2}} \tag{3.25}$$

$$= \frac{\sqrt{T}}{\sqrt{T}}\epsilon = \epsilon. \tag{3.26}$$

We define $d = \min_{\widetilde{\mathbf{s}},i} |\sqrt{\bar{M}^{\grave{\mathbf{R}}}_{\widetilde{\mathbf{s}}_{i:T}}} - r|$. Due to Lemma 3.3.3, if we let r be a positive constant smaller than $\sqrt{\frac{D_{min}}{2}}$, then d must be a positive constant. Thus if we take $\epsilon < d$, then $\sqrt{\bar{M}^{\mathbf{R}}_{\widetilde{\mathbf{s}}_{i:T}}} > r$ if $\widetilde{\mathbf{s}}_{i:T} \neq \mathbf{s}_{i:T}$ and $i < T$; on the other hand, $\sqrt{\bar{M}^{\mathbf{R}}_{\mathbf{s}_{i:T}}} < r$ for the

transmitted sequence s. Because at any layer i, the partial sequence $\mathbf{s}_{i:T}$ is the only partial sequence with metric smaller than r, the number of visited nodes at each layer is equal to $|\Omega|$ under matrix $\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N}$, with probability at least $(1-\xi)$. Thus for any constant $\xi > 0$, as $N \to \infty$, the expected number of visited nodes at layer i is upper bounded by

$$|\Omega| + 2(1-\xi)|\Omega|^i.$$

This is because under $r = \infty$, the largest number of visited nodes at layer i is $|\Omega|^i$; and before r is increased to ∞, Algorithm 2 visits at most $|\Omega|^i$ tree nodes at layer i. Taking an arbitrarily small $\xi > 0$, the expected number of visited nodes at layer i will approach $|\Omega|$.

The preprocessing step of Algorithm 2 requires $O(NT^2)$ computational complexity for computing $\mathbf{X}^*\mathbf{X}$, and $O(T^3)$ complexity for computing the Cholesky decomposition. For computing the metric of each tree node recursively in the depth-first tree search, $O(T)$ computational complexity is required. Thus the average overall computational complexity of Algorithm 2 is $O(NT^2 + T^3)$, when, for any fixed T, N goes to infinity. ∎

Remarks: Although our theoretical analysis is for massive SIMO systems, our algorithm also works for SIMO systems with a small number of receive antennas, without requiring the number of receive antennas N to be approaching infinity.

In summary, we have shown that, under a fixed σ_w^2 or SNR, Algorithm 2

can achieve an expected complexity of polynomial growth. In fact, as stated in Theorem 3.3.4, we can even lower the SNR requirement for each antenna, while still providing the GLRT-optimal JED with polynomial expected complexity.

Theorem 3.3.4. *Let the constellation Ω be constant-modulus. Let r^2 be a positive constant smaller than $\frac{D_{min}}{2}$, where $D_{min} = \min_{a,b\in\Omega} \|a-b\|^2$ is the minimum pairwise squared distance between any two constellation points. If $\sigma_w^2 = o(\sqrt{N})$, then for Algorithm 2, the expected number of visited points at layer i converges to $|\Omega|$ for $i \geq 2$, as the number of receive antennas N goes to infinity. Algorithm 2 only visits one tree node at layer 1. Here $o(\sqrt{N})$ means that $\lim_{N\to\infty} \sigma_w^2/\sqrt{N} = 0$.*

In fact, we can prove Theorem 3.3.4 through the same arguments as in proving Theorem 3.3.1, by noting that the variance $var(\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N})$ converges to 0 as $N \to \infty$, if $\sigma_w^2 = o(\sqrt{N})$. Since we fix the transmission power and the wireless channel model, $\sigma_w^2 = o(\sqrt{N})$ means that the SNR per receive antenna is allowed to decrease, as long as SNR$\sqrt{N} \to \infty$ as $N \to \infty$. For example, the SNR can scale as $O(\log(\log(N))/\sqrt{N})$ as $N \to \infty$. This implies that we can achieve the GLRT-optimal JED with low complexity, while increasing the energy efficiency of massive SIMO systems.

3.4 Algorithm Computational Complexity: N grows polynomially in T

In the previous section, we obtained the expected computational complexity for tree search when $N \to \infty$. However, the total computational complexity of Algorithm 2 also includes the complexity of calculating $\mathbf{X}^*\mathbf{X}$, which has a compu-

tational complexity of $O(NT^2)$. Thus the total computational complexity of Algorithm 2 can still grow very fast (for example, exponentially) with T, if N needs to grow exponentially in T. Then the natural question is whether the total computational complexity of Algorithm 2 grows polynomially in T. To see whether the total computational complexity of Algorithm 2 grows polynomially in T, we will need to find the scaling of N in terms of T such that the expected tree search complexity still grows polynomially in T. In this section, we prove that a polynomial growth of N in T suffices to make the expected tree search complexity grow polynomially in T. Again, for simplicity of proof presentation, we perform the proof for constant-modulus signals.

Theorem 3.4.1. *Let Ω be a constant-modulus constellation. Let $r^2 = \frac{D_{min}}{8}$, where $D_{min} = \min_{a,b\in\Omega} \|a-b\|^2$ is the minimum pairwise squared distance between any two constellation points. The expected overall complexity (including preprocessing and tree search) of Algorithm 2 grows polynomially in T, even when N grows polynomially with T.*

Before presenting the full proof, let us first outline the main idea of our proof. There are two key elements in the proof of this theorem: tightly bounding the concentration of $\mathbf{X}^*\mathbf{X}/N$ around $E(\mathbf{X}^*\mathbf{X})/N$, and carefully characterizing the stability of the Cholesky decomposition of $\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N}$ around $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$. Among these two key elements, proving the stability of the Choleksy decomposition around $\rho_E I - \frac{E(\mathbf{X}^*\mathbf{X})}{N}$ is particularly challenging. This is because existing stability results for the Cholesky decomposition in the literature (for example, [60], [61]) are for full-rank matrices; however, $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ happens to be a rank-deficient

matrix and existing stability results for the Cholesky decomposition do not apply.

In our proof, we show that the stability of the Cholesky decomposition around $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ is sufficient to guarantee small enough perturbations in the upper triangular matrix $\mathbf{R}$. Small perturbations in $\mathbf{R}$ in turn lead to small perturbations of data sequences' metrics. We show that the perturbations of data sequences' metrics decay fast enough in N, such that N growing only polynomially in T is sufficient to guarantee polynomial growth (in T) of the expected computational complexity of Algorithm 2.

Proof. (of Theorem 3.4.1) We first give a large deviation bound on the concentration of $\mathbf{X}^*\mathbf{X}/N$ towards $E(\mathbf{X}^*\mathbf{X})/N$. We consider two types of elements of $\mathbf{X}^*\mathbf{X}/N$: the diagonal elements and the off-diagonal elements. For the diagonal elements, we have Lemma 3.4.2 detailing its concentration.

Lemma 3.4.2. *Suppose the elements of* $\mathbf{h}$ *are i.i.d. circularly symmetric complex Gaussian random variables with 0 mean and variance* 1*. Then for each* $i \in \{1, 2, ..., T\}$*, for any* $\epsilon > 0$

$$
\begin{aligned}
&-\log\left(P\left(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N} - (1+\sigma_w^2) < -\epsilon\right)\right)/N \\
&\geq -\frac{\epsilon}{1+\sigma_w^2} + \log(1 + \frac{\epsilon}{(1+\sigma_w^2) - \epsilon}) \\
&\geq \frac{\epsilon^2}{2(1+\sigma_w^2)^2}.
\end{aligned}
$$

and

$$
\begin{aligned}
&-\log\left(P\left(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)>+\epsilon\right)\right)/N \\
&\geq \frac{\epsilon}{1+\sigma_w^2}+\log(1-\frac{\epsilon}{(1+\sigma_w^2)+\epsilon}) \\
&\geq \frac{\epsilon^2}{2(1+\sigma_w^2)^2}-\frac{\epsilon^3}{3(1+\sigma_w^2)^3},
\end{aligned}
$$

where $P(\cdot)$ means the probability. Moreover, if $\epsilon \leq 1+\sigma_w^2$, we have

$$
-\log\left(P\left(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)>+\epsilon\right)\right)/N \geq \frac{\epsilon^2}{6(1+\sigma_w^2)^2}.
$$

The proof of Lemma 3.4.2 is given in Appendix 3.8.5. Moreover, we have the following concentration result Lemma 3.4.3 for the off-diagonal entries of $\mathbf{X}^*\mathbf{X}/N$.

Lemma 3.4.3. *Suppose the elements of* $\mathbf{h}$ *are i.i.d. circularly symmetric complex Gaussian random variables with 0 mean and variance* 1. *Then for* $i,j \in \{1,2,...,T\}$ *and* $i \neq j$, *for any* $\epsilon > 0$,

$$
\begin{aligned}
&P(|\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N}-\mathbf{s}_i\mathbf{s}_j^*|>\epsilon) \\
&\leq e^{-N[-\frac{\epsilon}{13}+\log(\frac{13}{13-\epsilon})]}+e^{-N[\frac{\epsilon}{13}+\log(\frac{13}{13+\epsilon})]} \\
&+8e^{-N[\frac{-1+\sqrt{1+4\epsilon'^2}}{2}+\frac{1}{2}\log(1-[\frac{-1+\sqrt{1+4\epsilon'^2}}{2\epsilon}]^2)]} \\
&+4e^{-N[\frac{-1+\sqrt{1+4\epsilon''^2}}{2}+\frac{1}{2}\log(1-[\frac{-1+\sqrt{1+4\epsilon''^2}}{2\epsilon}]^2)]},
\end{aligned}
$$

where $\epsilon' = \frac{2\epsilon}{13\sigma_w}$ *and* $\epsilon'' = \frac{2\epsilon}{13\sigma_w^2}$. *The proof of Lemma 3.4.3 is given in Appendix 3.8.6. Moreover,*

$$P(|\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N} - \mathbf{s}_i\mathbf{s}_j^*| > \epsilon)$$
$$\leq e^{-\frac{N\epsilon^2}{338}} + e^{-\frac{N\epsilon^2}{1014}} + 8e^{-\frac{N\epsilon^2}{169\sigma_w^2}} + 4e^{-\frac{N\epsilon^2}{169\sigma_w^4}}$$

when $\epsilon \leq \min\{\frac{13\sigma_w}{\sqrt{2}}, \frac{13\sigma_w^2}{\sqrt{2}}, 13\}$.

We have the following Lemma 3.4.4 about the convergence of $\mathbf{A} = \rho I - \mathbf{X}^*\mathbf{X}/N$ to matrix $\grave{\mathbf{A}} = \rho_E I - E\{\mathbf{X}\mathbf{X}^*\}/N$ appearing in (3.16). The proof of Lemma 3.4.4 is deferred to Appendix 3.8.8.

Lemma 3.4.4. *Let us suppose that* $|(\mathbf{X}^*\mathbf{X}/N)_{i,j} - (E\{\mathbf{X}^*\mathbf{X}\}/N)_{i,j}| < \epsilon$ *holds for every pair* $(i,j) \in \{1,2,\cdots,T\} \times \{1,2,\cdots,T\}$. *Then the maximum eigenvalue, denoted by* ρ, *of* $\mathbf{X}^*\mathbf{X}/N$, *satisfies*

$$|\rho - \rho_E| \leq T\epsilon.$$

Moreover, $|(\rho I - \mathbf{X}^*\mathbf{X}/N)_{i,j} - (\rho_E I - E\{\mathbf{X}^*\mathbf{X}\}/N)_{i,j}| < (T+1)\epsilon$, *for every pair* $(i,j) \in \{1,2,\cdots,T\} \times \{1,2,\cdots,T\}$.

We then have the following key lemma about the robustness of the Cholesky decomposition, the proof of which is given in Appendix 3.8.9.

Lemma 3.4.5. *Suppose that each element of* $\mathbf{A} = \rho I - \mathbf{X}^*\mathbf{X}/N$ *deviates from* $\grave{\mathbf{A}} = \rho_E I - E[\mathbf{X}^*\mathbf{X}/N]$ *by a number with amplitude no more than* ϵ, *where* $\epsilon < \frac{1}{e^4T^2}$. *Let us denote*

the Cholesky decomposition of $\mathbf{A}$ *as* $\mathbf{R}^*\mathbf{R}$*, and denote the Cholesky decomposition of* $\grave{\mathbf{A}}$ *as* $\grave{\mathbf{R}}^*\grave{\mathbf{R}}$*, where* $\grave{\mathbf{R}}$ *and* $\mathbf{R}$ *are upper triangular matrices. Let us further assume that* $T \geq 4$*. Then for* $k < T$*, we have*

$$\begin{aligned}
|\mathbf{R}_{k,k} - \grave{\mathbf{R}}_{k,k}| &\leq |\epsilon_k|\sqrt{\frac{T-k+1}{T(T-k)}} \\
&\leq |\epsilon| e^4 \sqrt{\frac{T^3}{(T-k+1)^3(T-k)}} \\
&\leq |\epsilon| e^4 \sqrt{\frac{T^3}{8}},
\end{aligned}$$

where

$$|\epsilon_k| \leq |\epsilon| e^4 \frac{T^2}{(T-k+1)^2}.$$

For $k < j$ *and* $k < T$*,*

$$\begin{aligned}
|\mathbf{R}_{k,j} - \grave{\mathbf{R}}_{k,j}| &\leq |\epsilon_k| \frac{T}{2(T-k+1)(\sqrt{\frac{T(T-k)}{T-k+1}} - |\epsilon_k|)^3} \\
&\quad + \frac{|\epsilon_k|}{\sqrt{\frac{T(T-k)}{T-k+1} - |\epsilon_k|}} \\
&\leq |\epsilon| e^4 T\sqrt{T}. \qquad (3.27)
\end{aligned}$$

Moreover, for $k = T$*, we have*

$$|\mathbf{R}_{T,T} - \grave{\mathbf{R}}_{T,T}| \leq T e^2 \sqrt{|\epsilon|}.$$

From Lemma 3.4.5, we know that when $\epsilon < \frac{1}{2e^4T^2}$, $Te^2\sqrt{|\epsilon|} \geq \epsilon e^4 \sqrt{\frac{T^3}{8}}$ and

$Te^2\sqrt{|\epsilon|} \geq \epsilon e^4 T\sqrt{T}$. Thus when $\epsilon < \frac{1}{2e^4T^2}$, each element of $\mathbf{R} - \grave{\mathbf{R}}$ is bounded by $Te^2\sqrt{|\epsilon|}$ in magnitude.

Building on these bounds for the perturbations of $\mathbf{R}$, we further use Lemma 3.4.6 to bound the perturbation of the unscaled metric of a partial sequence in Algorithm 2. We again defer the proof of Lemma 3.4.6 to Appendix 3.8.10.

Lemma 3.4.6. *Let $\mathbf{P}$ be a $T \times T$ matrix such that $\mathbf{R}_{i,j} - \grave{\mathbf{R}}_{i,j} = \mathbf{P}_{i,j}$, for $1 \leq i, j \leq T$. Let $M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}$ denote the unscaled metric (defined in (3.12)) of the partial sequence $\mathbf{s}^*_{i:T}$ under $\grave{\mathbf{R}}$, and $M^{\mathbf{R}}_{\mathbf{s}^*_{i:T}}$ denote the unscaled metric of $\mathbf{s}^*_{i:T}$ under $\mathbf{R}$. Then we have*

$$\sqrt{M^{\mathbf{R}}_{\mathbf{s}^*_{i:T}}} \leq \sqrt{M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}} + \sqrt{\sum_{t=i}^{T}\sum_{j=i}^{T}|\mathbf{P}_{t,j}|^2}\sqrt{T-i+1}$$

and

$$\sqrt{M^{\mathbf{R}}_{\mathbf{s}^*_{i:T}}} \geq \sqrt{M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}} - \sqrt{\sum_{t=i}^{T}\sum_{j=i}^{T}|\mathbf{P}_{t,j}|^2}\sqrt{T-i+1}.$$

We are now ready to prove Theorem 3.4.1. Suppose that each element of $\mathbf{X}^*\mathbf{X}/N$ deviates from its counterpart $E\{\mathbf{X}^*\mathbf{X}\}/N$ by a number no bigger than ϵ in magnitude. Then according to Lemma 3.4.4, each element of $\rho I - \mathbf{X}^*\mathbf{X}/N$ deviates from $\rho_E I - E\{\mathbf{X}^*\mathbf{X}\}/N$ by at most $(T+1)\epsilon$. According to Lemma 3.4.5, each element of $\mathbf{R}$ deviates from $\grave{\mathbf{R}}$ by at most $e^2T\sqrt{(T+1)\epsilon}$ in amplitude, when $(T+1)\epsilon < \frac{1}{2e^4T^2}$, namely $\epsilon \leq \frac{1}{2e^4T^2(T+1)}$. According to Lemma 3.4.6, for each partial sequence $\mathbf{s}^*_{i:T}$, the

square root of its unscaled metric $\sqrt{M^{\mathbf{R}}_{\mathbf{s}^*_{i:T}}}$ will deviate from $\sqrt{M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}}$ by at most

$$\sqrt{(e^2 T\sqrt{(T+1)\epsilon})^2 \times T^2} \times \sqrt{T} = T^{2.5} e^2 \sqrt{(T+1)\epsilon}.$$

Let us take $\epsilon > 0$ such that

$$T^{2.5} e^2 \sqrt{(T+1)\epsilon} < \frac{1}{2}\sqrt{\frac{D_{min} T}{2}},$$

and

$$(T+1)\epsilon < \frac{1}{2e^4 T^2}.$$

Namely,

$$\epsilon \leq \min\{\frac{D_{min}}{8T^4 e^4 (T+1)}, \frac{1}{2e^4 T^2 (T+1)}\} = \frac{D_{min}}{8T^4 e^4 (T+1)},$$

because $D_{min} = \min_{a,b\in\Omega} \|a-b\|^2 \leq 4$ for Ω with unit-energy constellation points.

With this choice of ϵ, under a search radius $r = \frac{1}{2}\sqrt{\frac{D_{min}}{2}}$ (corresponding to an unscaled metric whose square root is $\frac{1}{2}\sqrt{\frac{TD_{min}}{2}}$), Algorithm 2 will visit the same number of tree nodes as under $\grave{\mathbf{R}}$. This is because under $\grave{\mathbf{R}}$, the square root of the unscaled metric of a partial sequence different from the transmitted sequence is at least $\sqrt{\frac{TD_{min}}{2}}$ (noting that $\grave{\mathbf{R}}_{i,i}$ is at least $\sqrt{\frac{T}{2}}$ when $i < T$); moreover, under $\grave{\mathbf{R}}$, the true transmitted sequence has a metric equal to 0. With $T^{2.5} e^2 \sqrt{(T+1)\epsilon} < \frac{1}{2}\sqrt{\frac{D_{min} T}{2}}$ and $r = \frac{1}{2}\sqrt{\frac{D_{min}}{2}}$, the perturbations in the metrics of partial sequences are guaranteed to be small enough such that only partial sequences of the transmitted signal $\mathbf{s}^*$ have metrics within the search radius.

Now let us examine the probability of the abnormal event that some element of $\mathbf{X}^*\mathbf{X}/N$ deviates from its counterpart in $E\{\mathbf{X}^*\mathbf{X}\}/N$ by a number bigger than $\epsilon = \frac{D_{min}}{8T^4e^4(T+1)}$ in magnitude. By the union bound over the T^2 elements of $\mathbf{X}^*\mathbf{X}/N$, according to Lemma 3.4.2 and Lemma 3.4.3, this probability is at most $T^2 \times (5e^{-\alpha N\epsilon^2})$ when $\epsilon = \frac{D_{min}}{8T^4e^4(T+1)}$ (note that we consider $T \to \infty$), where

$$\begin{aligned}\alpha &= \min\{\frac{1}{6(1+\sigma_w)^2}, \frac{1}{338}, \frac{1}{1014}, \frac{1}{169\sigma_w^2}, \frac{1}{169\sigma_w^4}\} \\ &= \min\{\frac{1}{6(1+\sigma_w)^2}, \frac{1}{1014}, \frac{1}{169\sigma_w^2}, \frac{1}{169\sigma_w^4}\}\end{aligned}$$

Under the abnormal event, Algorithm 2 will at most visit $2|\Omega|^T$ tree nodes (each tree node update needs $O(T)$ operations). Thus, if $(2|\Omega|^T) \times (5T^2e^{-\alpha N\epsilon^2})$ is polynomially growing with T, the expected complexity of Algorithm 2 will be polynomial in T. This polynomial growth of Algorithm 2's overall computational complexity is true as long as

$$\begin{aligned}N &\geq \frac{T\log(|\Omega|) + \log(10) + 2\log(T) + O(\log(T))}{\alpha\epsilon^2} \\ &\geq [T\log(|\Omega|) + \log(10) + 2\log(T) + O(\log(T))] \\ &\quad \times (\frac{64}{\alpha}T^8e^8(T+1)^2) \\ &= O(T^{11}).\end{aligned}$$

This proves our main Theorem 3.4.1. ∎

Remarks: We remark that the bounds $N = O(T^{11})$ can be quite conserva-

tive. By further sharpening the stability analysis of the Cholesky decomposition, one can possibly reduce the number of antennas N in order to guarantee overall computational complexity polynomial in T. Our simulations show that in practice way fewer receive antennas than T^{11} are sufficient to guarantee Algorithm 2 visits only around $|\Omega|$ tree nodes at each layer, with high probability.

3.5 Computational Complexity for Nonconstant-Modulus Constellations

So far, we have shown that the complexity results for constant-modulus constellations. In this subsection, we will extend our results to general constellations. Since the derivations for general constellations are similar to that for constant-modulus constellations, we will only give the new theorem statement, and highlight key differences in the derivations.

Theorem 3.5.1. *Let Ω be a constant-modulus or nonconstant-modulus constellation. Let $|\mathbf{s}_{max}|^2$ and $|\mathbf{s}_{min}|^2$ be respectively the largest and the smallest possible energy of a constellation point from Ω. Let r^2 be a positive constant smaller than*

$$\frac{|\mathbf{s}_{min}|^4 D_{min}}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2 |\mathbf{s}_{max}|^2},$$

where $D_{min} = \min_{a,b\in\Omega} \|a - b\|^2$ is the minimum pairwise squared distance between any two constellation points. Suppose the channel $\mathbf{h}$ has i.i.d. entries following circularly-symmetric complex Gaussian distributions with zero mean and unit variance. We further assume that $\mathbf{s}_T$ is known to the receiver to resolve the phase ambiguity. Then for Algorithm 2, the expected number of visited points at layer i converges to $|\Omega|$ for $i \geq 2$, as the number

of receive antennas N goes to infinity. Algorithm 2 only visits one tree node at layer 1. Moreover, the overall expected computational complexity of Algorithm 2 is $O(NT^2 + T^3)$ if N grows polynomially in T.

We now outline the key steps in deriving Theorem 3.5.1. Taking the same analysis as in deriving Theorems 3.3.1 and 3.4.1 for constant-modulus constellations, we can write the maximum eigenvalue of the Hermitian matrix $\frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ as $\rho_E = \sum_{k=1}^{T} \|\mathbf{s}_k\|^2 + \sigma_w^2$. Then we can represent $\grave{\mathbf{A}} = \rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ as

$$\grave{\mathbf{A}} = \begin{bmatrix} t - \mathbf{s}_1\mathbf{s}_1^* & -\mathbf{s}_1\mathbf{s}_2^* & \cdots & -\mathbf{s}_1\mathbf{s}_T^* \\ -\mathbf{s}_2\mathbf{s}_1^* & t - \mathbf{s}_2\mathbf{s}_2^* & \cdots & -\mathbf{s}_2\mathbf{s}_T^* \\ \vdots & \vdots & \vdots & \vdots \\ -\mathbf{s}_T\mathbf{s}_1^* & -\mathbf{s}_T\mathbf{s}_2^* & \cdots & t - \mathbf{s}_T\mathbf{s}_T^* \end{bmatrix},$$

where $t = \sum_{k=1}^{T} \|\mathbf{s}_k\|^2$. After decomposing $\grave{\mathbf{A}}$ using Cholesky decomposition, we can find the upper triangular matrix $\grave{\mathbf{R}}$ such that $\grave{\mathbf{A}} = \grave{\mathbf{R}}^*\grave{\mathbf{R}}$. Then using the recursive relation for calculating the Cholesky decomposition as mentioned in the proof of Theorem 3.3.1, we can obtain that the diagonal entries of the $\grave{\mathbf{R}}$ is given by

$$\grave{\mathbf{R}}_{i,i} = \sqrt{t - \|\mathbf{s}_i\|^2 - \sum_{j=1}^{i-1} \frac{\|\mathbf{s}_j\|^2\|\mathbf{s}_i\|^2 t}{(t - \|\mathbf{s}_{1:j-1}\|^2)(t - \|\mathbf{s}_{1:j}\|^2)}}. \tag{3.28}$$

We can find the metric $\bar{M}_{\mathbf{s}^*_{1:T}}$ of the transmitted signal $\mathbf{s}^*$ as

$$\bar{M}_{\mathbf{s}^*_{1:T}} = \frac{\mathbf{s}^*\grave{\mathbf{A}}\mathbf{s}}{\|\mathbf{s}\|^2} = \frac{\mathbf{s}^*(tI - \mathbf{s}\mathbf{s}^*)\mathbf{s}}{\|\mathbf{s}\|^2} = 0, \tag{3.29}$$

since $\mathbf{s}^*\mathbf{s} = t$. As a result, $\bar{M}_{\mathbf{s}^*_{i:T}} = 0$ for any partial sequence $\mathbf{s}^*_{i:T}$ of the transmitted sequence $\mathbf{s}^*_{1:T}$. On the other hand, according to Lemma 3.3.3, for any other signal $\widetilde{\mathbf{s}} \neq \mathbf{s}$, $\bar{M}_{\widetilde{\mathbf{s}}^*_{j:T}} \geq \frac{|\mathbf{s}_{min}|^4 D_{min}}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2 |\mathbf{s}_{max}|^2}$ at any layer $j \leq i$, where i is the largest integer such that $\mathbf{s}^*_i \neq \widetilde{\mathbf{s}}^*_i$.

Thus if we set $r^2 < \frac{|\mathbf{s}_{min}|^4 D_{min}}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2 |\mathbf{s}_{max}|^2}$, under the expected matrices, Algorithm 2 will only visit $|\Omega|$ nodes in each layer. Following similar concentration arguments for the matrix $\rho I - \frac{\mathbf{X}^*\mathbf{X}}{N}$, and the perturbation analysis as in Theorems 3.2.2 and 3.3.1, we can similarly prove Theorem 3.5.1.

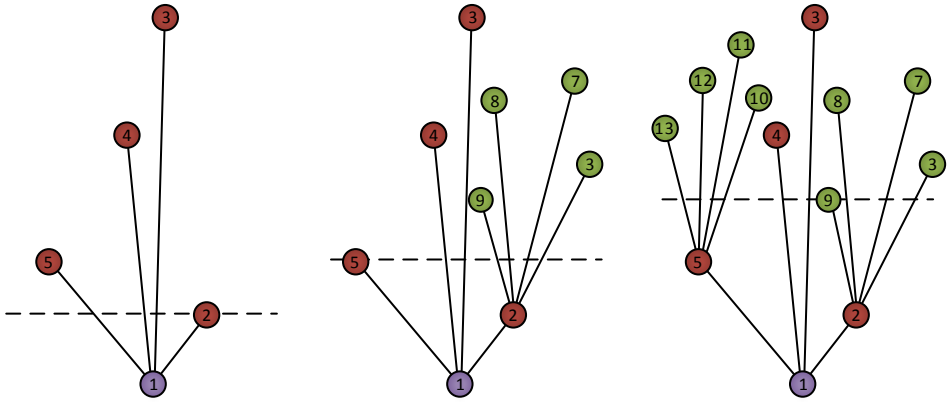
3.6 Tree Search Algorithm

In the sections above, we consider each partial sequence as a node in a tree structure of T layers. The computational complexity of the earlier algorithms heavily depends on how the initial search radius r is chosen. Although the search radius r is chosen so that the true transmitted sequence is within the sphere with high probability, the radius does not guarantee the minimum number of visited nodes in the tree search.

In this section we design a best-first branch-and-bound tree search algorithm for GLRT-optimal JED that does not need an assigned initial radius r. We call this algorithm the Tree Search Algorithm (TSA). In contrast to our GLRT-optimal algorithm in the previous section, TSA sets the initial search radius as zero at the beginning of the algorithm. Then the radius r in TSA systematically increases until the optimal JED solution is found. This algorithm guarantees to visit no more tree

nodes than the algorithm in the previous sections. We will show that our previous complexity results also upper bound the complexity of TSA. Moreover, we prove that this new TSA applies to nonconstant-modulus constellations.

We first introduce several terminologies about the tree structure we are using. A partial sequence $\widetilde{\mathbf{s}}^*_{i:T}$, $1 \le i \le T$, corresponds to a layer-$(T-i+1)$ node in the tree. A node $\widetilde{\mathbf{s}}^*_{i:T} = (\widetilde{\mathbf{s}}^*_i, \widetilde{\mathbf{s}}^*_{i+1:T})$ is called a child node of its parent node $\widetilde{\mathbf{s}}^*_{i+1:T}$. The parent node of any layer-1 node $\widetilde{\mathbf{s}}^*_T$ is called the root node. In a tree, any tree node without a child node is called a leaf node. For example, in (b) of Figure 3.1, node 1 is the root node, and node 2 is the parent node of node 9.



(a) First search iteration (b) Second search iteration (c) Third search iteration

Figure 3.1: Illustration of tree search algorithm for a tree of 3 layers.

In the TSA algorithm, we start to construct a tree which has only the root node with metric 0. Then in each iteration, the TSA always first finds the leaf node with the smallest metric, which is called the seed node. Then the algorithm expands the tree by adding the seed node's $|\Omega|$ child nodes to the tree, and calculat-

ing the metrics of all these child nodes. The tree search algorithm then iterates this process of finding the seed node and expanding the tree, until the seed node is a layer-1 node, corresponding to a full-length sequence. The flow of this algorithm is described as below for constant-modulus constellations (for nonconstant-modulus modulations we just need to replace $M_{\widetilde{\mathbf{s}}^{\star}_{i:T}}$ by $\bar{M}_{\widetilde{\mathbf{s}}^{\star}_{i:T}}$).

Figure 3.1 shows 3 search iterations for QPSK constellation and $T = 2$. The height of a node represents its metric. In (a), the root node 1 is selected as the seed node, and expands into 4 child nodes. Then node 2 is chosen as the seed node, and expands into 4 child nodes. The expansion of node 2 is shown in (b). The TSA then finds node 5 as the next seed node. The third search iteration in (c) expands node 5 by adding its 4 children. The TSA algorithm then finds node 9 as the seed node since it has the smallest metric. Since node 9 is a layer-2 node, the algorithm will terminate and output node 9 as the GLRT-optimal solution.

3.6.1 Computational Complexity of TSA

In this section, we will show that the TSA algorithm is computationally efficient in terms of the number of visited nodes.

Theorem 3.6.1. *The TSA outputs the optimal sequence in joint channel estimation and data detection. Let M be the metric of the optimal sequence, and let l be the number of sequences (including partial sequences) that have metrics no bigger than M. Then the number of visited points by TSA is no more than $(|\Omega| + 1)l$. Moreover, the TSA algorithm visits no more tree nodes than Algorithm 2 in Section 3.2.*

Algorithm 3: Tree Search Algorithm (TSA).

Input:matrix R and constellation Ω
Output: The transmitted signal $\mathbf{s}^*$

1. Add the root node, and set its metric to 0. Set $r^2 = 0$;
2. (Find the seed node) Find the leaf node $\widetilde{\mathbf{s}}^*_{i:T}$ which has the smallest metric among all the leaf nodes. Select that leaf node as the seed node. Update $r^2 = M_{\widetilde{\mathbf{s}}^*_{i:T}}$;
3. If the seed node $\widetilde{\mathbf{s}}^*_{i:T}$ is layer-1 node, namely $i = 1$, then go to 4; else, add the $|\Omega|$ child nodes of $\widetilde{\mathbf{s}}^*_{i:T}$ to the tree, compute the metrics of these child nodes, and go to 2;
4. Terminate the algorithm, output $\widetilde{\mathbf{s}}^*_{1:T}$ as the optimal sequence. Output r^2 as the smallest possible metric.

Proof. We first notice that every full-length sequence $\widetilde{\mathbf{s}}^*_{1:T}$ is a direct or indirect child of a leaf node $\widetilde{\mathbf{s}}^*_{i:T}$ existing at the termination of the TSA. However, by the TSA, the metric $M_{\widetilde{\mathbf{s}}^*_{i:T}}$ must be no smaller than the final r^2. Since $M_{\widetilde{\mathbf{s}}^*_{i:T}}$ is a lower bound of $M_{\widetilde{\mathbf{s}}^*_{1:T}}$, we have $M_{\widetilde{\mathbf{s}}^*_{1:T}} \geq r^2$ at the termination of the TSA. This proves that the TSA indeed outputs the optimal sequence, and $r^2 = M$ at its termination.

According to its procedure, the TSA algorithm will not visit the child nodes of any node B which has a metric bigger than M, namely node B will not be selected as a seed node in the tree search. In fact, the TSA will add the full-length optimal sequence and all its (direct or indirect) parent nodes to the tree (because a parent node's metric is always no bigger than its child node's) first; and then the TSA will declare the full-length optimal sequence as the final solution, terminating before node B is ever selected as a seed node. So the TSA algorithm can only visit

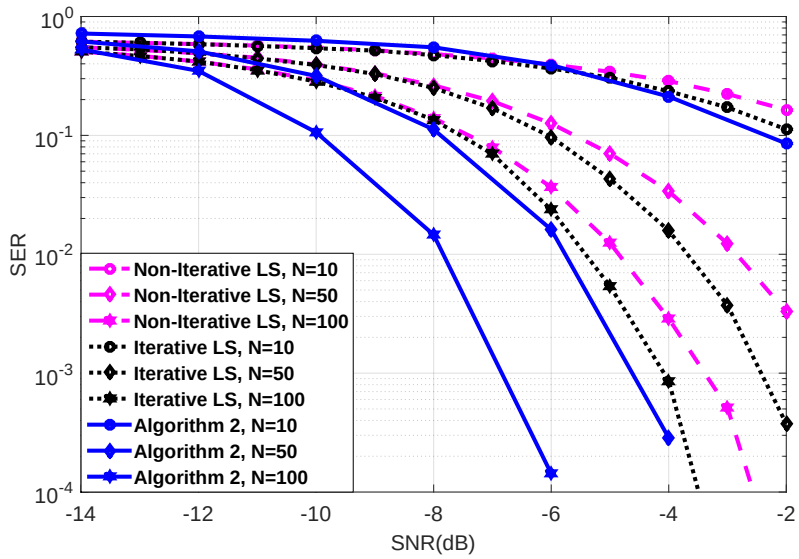


Figure 3.2: SER vs SNR for the GLRT-optimal joint channel estimation and data detection, iterative and non-iterative LS channel estimation with $T = 8$ and QPSK modulation.

tree nodes which have metric no bigger than M, and possibly their direct child nodes. This gives an upper bound of $(|\Omega| + 1)l$ on the total number of visited tree nodes.

To find the optimal sequence, Algorithm 2 must have used a radius r such that $r^2 \geq M$. Thus Algorithm 2 will visit every tree node with metric no bigger than M, and its child nodes. So the number of visited nodes by Algorithm 2 must be no smaller than that of the TSA. Thus, the TSA will also visit a polynomial number of nodes on average, as $N \to \infty$. ∎

3.7 Simulation Results

In this section, we simulate the performance and complexity of the exact GLRT-optimal algorithm for SIMO systems with N receive antennas, under QPSK

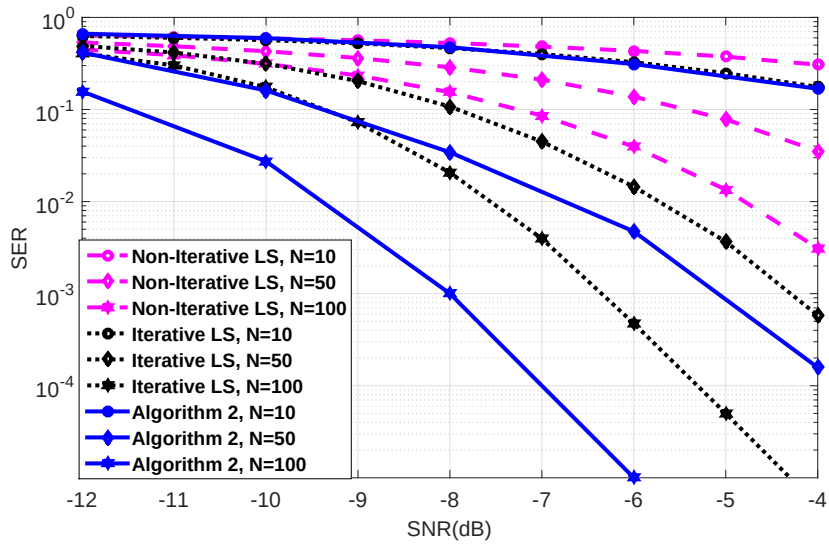


Figure 3.3: SER vs SNR for the GLRT-optimal joint channel estimation and data detection, iterative and non-iterative LS channel estimation with $T = 20$ and QPSK modulation.

and nonconstant-modulus 16-QAM. Channel matrix entries are generated as i.i.d complex Gaussian random variables. We investigate the performance of the GLRT-optimal algorithm for N= 10, 50, 100, and 500 receive antennas. We compare the performance of the GLRT-optimal JED algorithm with sub-optimal iterative and non-iterative JED schemes. We use least square (LS) and minimum mean square error (MMSE) channel estimation for the iterative and non-iterative JED (the reader may refer to [62] for the LS and MMSE channel estimation).

In each channel coherent block, we embed one symbol which is known by the receiver to resolve channel phase ambiguity at the end of the data sequence. In the non-iterative channel estimation scheme, the receiver estimates the channel vector using this training symbol. Then, the receiver uses this estimated channel vector to detect the remaining $T-1$ transmitted symbols. The iterative sub-optimal

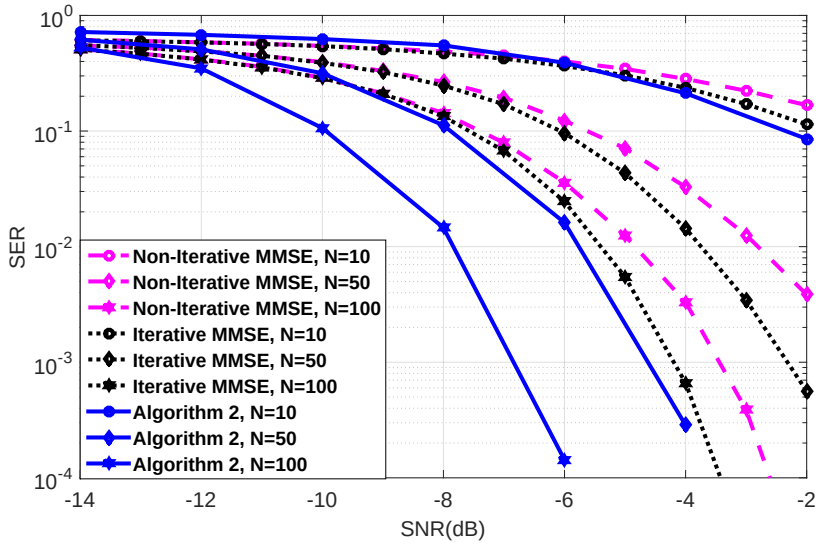


Figure 3.4: SER vs SNR for GLRT-optimal joint channel estimation and data detection, iterative and non-iterative MMSE channel estimation with $T = 8$ and QPSK modulation.

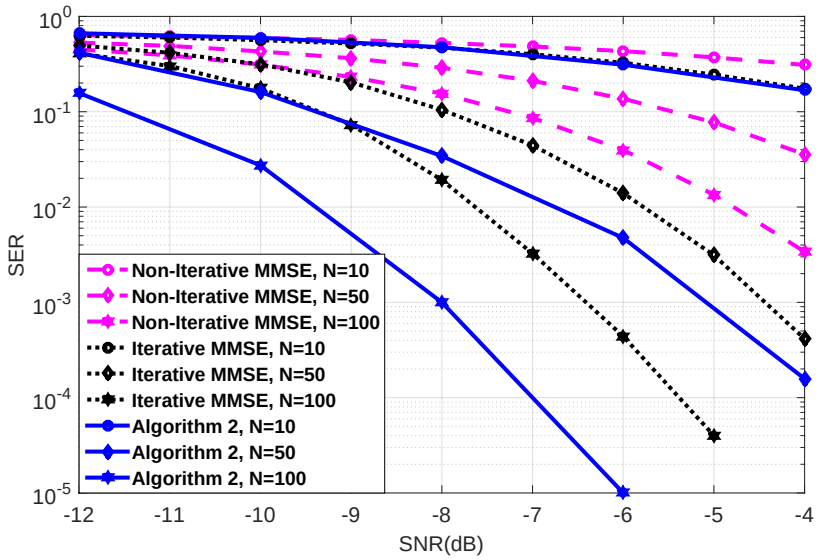


Figure 3.5: SER vs SNR for GLRT-optimal joint channel estimation and data detection, iterative and non-iterative MMSE channel estimation with $T = 20$ and QPSK modulation.

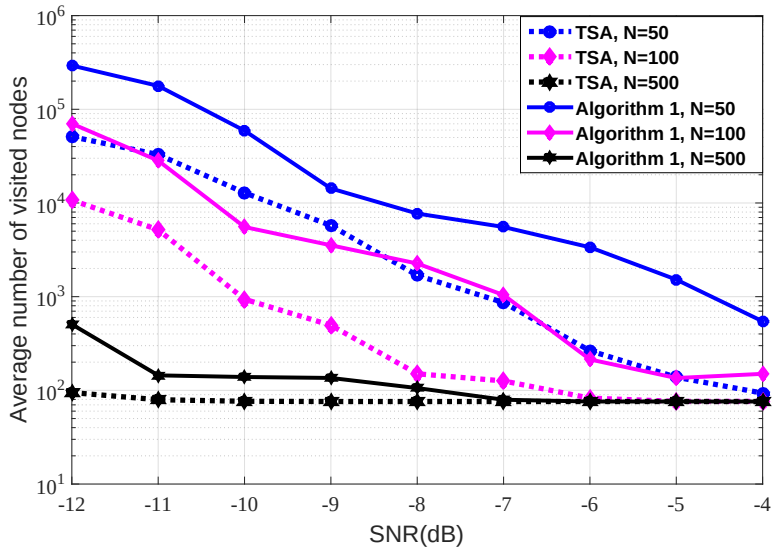


Figure 3.6: Average number of visited points for $T = 20$ and QPSK modulation. Exhaustive search will instead need to examine 2.75×10^{11} hypotheses.

scheme exploits the detected data vector from the pervious iteration to obtain a new channel estimation, which, in turn, is used for data detection in the current iteration. The iterative joint channel estimation and data detection scheme runs 100 iterations for each channel coherence block.

In Figures 3.2, 3.3, 3.4, and 3.5, under the QPSK modulation, the symbol error rate (SER) of the GLRT-optimal JED algorithm is evaluated as a function of SNR for $T = 8$ and 20 respectively, along with the SER of data detection based on the iterative and non-iterative LS and MMSE channel estimations. It can be seen that the GLRT-optimal algorithm outperforms the LS and MMSE iterative and non-iterative channel estimation schemes. For example, from Figures 3.2 and 3.4, we see more than 2 dB improvement over the iterative channel estimation and data detection, and 3 dB improvement over the non-iterative channel estimation and

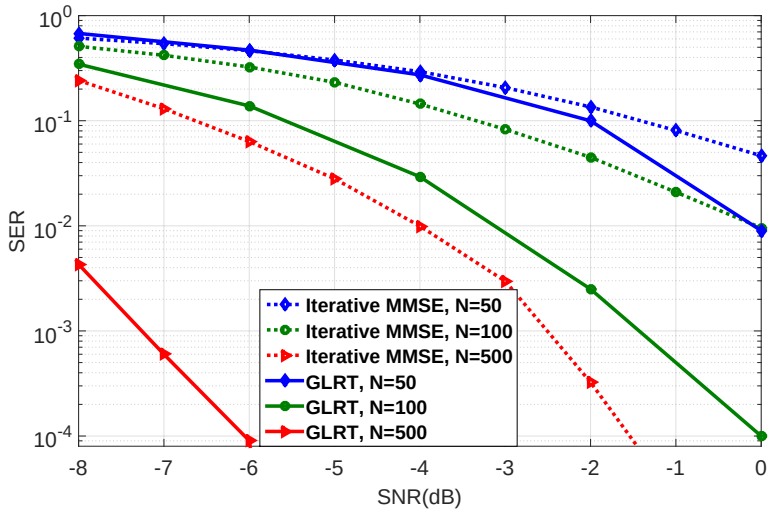


Figure 3.7: SER vs SNR, for the GLRT-optimal joint channel estimation and data detection and iterative MMSE channel estimation with $T = 12$ and 16-QAM.

data detection for N=100, at 10^{-2} SER. In Figures 3.3 and 3.5, the GLRT-optimal JED provides a performance improvement of 2 dB over the iterative scheme and 4.5 dB improvement over the non-iterative scheme, at 10^{-2} SER.

We further evaluate the complexities of both Algorithm 2 and the TSA for QPSK constellation by the average number of visited nodes in each coherence block. In Figure 3.6, we obtain the average number of visited nodes for T=20 at different SNR values. We use our proposed search radius $r^2 = \frac{1}{3}$ for Algorithm 1. It can be seen that when N increases, the number of visited nodes significantly decreases. In fact, the average number of visited nodes for N=500 is steady at 76, namely the cardinality of the QPSK constellation multiplied by $(T-1)$ layers. This is consistent with our theoretical prediction in Theorem 3.3.1. In addition, the TSA further reduces the complexity. At SNR = -4 dB, our algorithms on average visit

only around several hundred nodes for $N = 50$, and only 76 nodes for $N = 500$. In comparison, the exhaustive search method will need to examine $4^{19} \approx 2.75 \times 10^{11}$ hypotheses for each coherence block. Our algorithms achieve complexity reduction in many orders of magnitude across a wide range of N.

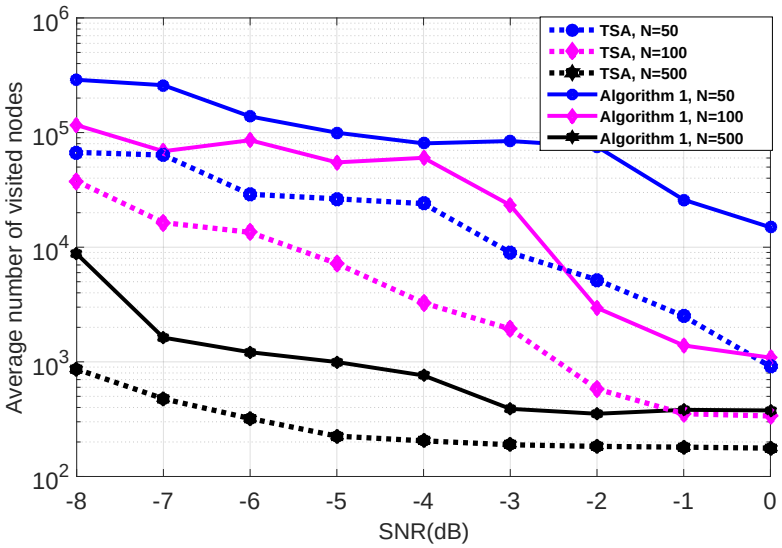


Figure 3.8: Average number of visited points, T=12 with 16-QAM. Exhaustive search will instead need to examine 1.76×10^{13} hypotheses.

Figure 3.7 describes the performance of the GLRT-optimal JED for the nonconstant-modulus 16-QAM constellation. We choose the the coherent time $T = 12$, and $N = 50, 100$ and 500. We can see that our novel GLRT-optimal algorithms provide nearly 5 dB gain over iterative joint MMSE channel estimation and data detection algorithms. Under 16-QAM, Figure 3.8 presents the average number of visited nodes, under different SNR values, for Algorithm 2 with $r^2 = \frac{2}{45}$ and for the TSA.

The average is taken over 10^3 channel coherence blocks. Both algorithms achieve surprisingly low average computational complexity. Note that, in order to do exhaustive search, one would need to examine 16^{11}=1.76×10^{13} hypotheses in each coherence block. For SNR above -4 dB, on average the TSA visits only 176 nodes, a 10^{11}-fold reduction in complexity compared with exhaustive search.

3.8 Appendix

3.8.1 Proof of Lemma 3.2.1

Proof. Because

$$\begin{aligned}\|\mathbf{s}\|^2 &= \|\mathbf{s}_{1:i-1}\|^2 + \|\mathbf{s}_{i:T}\|^2 \\ &\leq |\mathbf{s}_{max}|^2(i-1) + \|\mathbf{s}_{i:T}\|^2,\end{aligned}$$

we have $\frac{M_{\mathbf{s}^*_{i:T}}}{\|\mathbf{s}\|^2} \geq \bar{M}_{\mathbf{s}^*_{i:T}}$. Thus

$$\begin{aligned}\frac{\|\mathbf{R}\mathbf{s}\|^2}{\|\mathbf{s}\|^2} &= \frac{\sum_{j=1}^{T} |\sum_{k=j}^{T} \mathbf{R}_{j,k}\mathbf{s}_k|^2}{\|\mathbf{s}\|^2} \\ &= \frac{\sum_{j=1}^{i-1} |\sum_{k=j}^{T} \mathbf{R}_{j,k}\mathbf{s}_k|^2 + \sum_{j=i}^{T} |\sum_{k=j}^{T} \mathbf{R}_{j,k}\mathbf{s}_k|^2}{\|\mathbf{s}\|^2} \\ &\geq \frac{M_{\mathbf{s}^*_{i:T}}}{\|\mathbf{s}\|^2} \\ &\geq \bar{M}_{\mathbf{s}^*_{i:T}}.\end{aligned}$$

■

3.8.2 Proof of Lemma 3.3.2

Proof. Now we use induction over row index i, to show that

$$\grave{\mathbf{R}}_{i,i} = \sqrt{\frac{T(T-i)}{T-i+1}}$$

and for every $j > i$,

$$\grave{\mathbf{R}}_{i,j} = -\mathbf{s}_i\mathbf{s}_j^* \sqrt{\frac{T}{(T-i+1)(T-i)}}.$$

When $i = 1$, we have $\grave{\mathbf{R}}_{1,1} = \sqrt{\grave{\mathbf{A}}_{1,1}} = \sqrt{T-1} = \sqrt{\frac{T(T-1)}{T-1+1}}$. For every $j > 1$, we have

$$\grave{\mathbf{R}}_{1,j} = \frac{\grave{\mathbf{A}}_{1,j}}{\grave{\mathbf{R}}_{1,1}} = \frac{-\mathbf{s}_1\mathbf{s}_j^*}{\sqrt{T-1}} = -\mathbf{s}_1\mathbf{s}_j^* \sqrt{\frac{T}{(T-1+1)(T-1)}}.$$

Thus the conclusion of this lemma is true for index $i = 1$.

Suppose the conclusion of this lemma is true for any index i such that $i \leq k < T$, where k is an integer. Then let us consider $i = k+1$. Because $\grave{\mathbf{R}}_{i,i} = \sqrt{\grave{\mathbf{A}}_{i,i} - \sum_{t=1}^{i-1} \grave{\mathbf{R}}_{t,i}\grave{\mathbf{R}}_{t,i}^*}$, we have

$$\begin{aligned}
\grave{\mathbf{R}}_{k+1,k+1}^2 &= (T-1) - \sum_{t=1}^{k}\left[-\mathbf{s}_t\mathbf{s}_{k+1}^* \sqrt{\frac{T}{(T-t+1)(T-t)}} \times -\mathbf{s}_t^*\mathbf{s}_{k+1} \sqrt{\frac{T}{(T-t+1)(T-t)}}\right] \\
&= (T-1) - \sum_{t=1}^{k} \frac{T}{(T-t+1)(T-t)} = T-1-(\frac{T}{T-k}-1) \\
&= \frac{T(T-k-1)}{T-k}.
\end{aligned}$$

where the last two equalities are due to Lemma 3.8.1. Moreover, for $j > k+1$,

$$
\begin{aligned}
\grave{\mathbf{R}}_{k+1,j} &= \frac{\grave{\mathbf{A}}_{k+1,j} - (\sum_{t=1}^{k} \grave{\mathbf{R}}_{t,k+1}\grave{\mathbf{R}}_{t,j}^{*})^{*}}{\grave{\mathbf{R}}_{k+1,k+1}} \\
&= \left(-\mathbf{s}_{k+1}\mathbf{s}_{j}^{*} - \sum_{t=1}^{k}\left[(-\mathbf{s}_{t}^{*}\mathbf{s}_{k+1})\sqrt{\frac{T}{(T-t+1)(T-t)}}\right.\right. \\
&\times \left.\left.(-\mathbf{s}_{t}\mathbf{s}_{j}^{*})\sqrt{\frac{T}{(T-t+1)(T-t)}}\right]\right)/\grave{\mathbf{R}}_{k+1,k+1} \\
&= \frac{(-\mathbf{s}_{k+1}\mathbf{s}_{j}^{*})\left[1 + \sum_{t=1}^{k}\frac{T}{(T-t+1)(T-t)}\right]}{\grave{\mathbf{R}}_{k+1,k+1}} \\
&= -\mathbf{s}_{k+1}\mathbf{s}_{j}^{*} \times \frac{\frac{T}{T-k}}{\sqrt{\frac{T(T-k-1)}{T-k}}} \\
&= -\mathbf{s}_{k+1}\mathbf{s}_{j}^{*}\sqrt{\frac{T}{(T-k)(T-k-1)}} \\
&= -\mathbf{s}_{k+1}\mathbf{s}_{j}^{*}\sqrt{\frac{T}{(T-(k+1)+1)(T-(k+1))}}.
\end{aligned}
$$

Thus we show that the formulas for $\grave{\mathbf{R}}_{i,i}$ and $\grave{\mathbf{R}}_{i,j}$, $j > i$ are also true for $i = k+1$. ∎

3.8.3 Lemma 3.8.1 and Its Proof

Lemma 3.8.1.

$$
\sum_{j=1}^{k}\frac{T}{(T-j+1)(T-j)} = \frac{T}{T-k} - 1
$$

Proof.

$$
\begin{aligned}
\sum_{j=1}^{k}\frac{T}{(T-j+1)(T-j)} &= \sum_{j=1}^{k}\left(\frac{j}{T-j} - \frac{j-1}{T-(j-1)}\right) \\
&= \frac{k}{T-k} = \frac{T}{T-k} - 1
\end{aligned}
\tag{3.30}
$$

∎

3.8.4 Derivation of $var[(\mathbf{X}^*\mathbf{X})_{i,j}/N]$ in (3.21)

Proof.

$$
\begin{aligned}
var[(\mathbf{X}^*\mathbf{X})_{i,j}] &= var[\sum_{k=1}^{N} B_k] = \sum_{k=1}^{N} var(B_k) \\
&= \sum_{k=1}^{N} (E[B_k B_k^*] - E[B_k]E[B_k^*])
\end{aligned}
$$

where $B_k = (\mathbf{s}_i^*\mathbf{h}_k + \mathbf{w}_{k,i})^*(\mathbf{s}_j^*\mathbf{h}_k + \mathbf{w}_{k,j})$, and we use $\mathbf{w}_{i,j}$ to denote the element in the i-th row and j-th column of $\mathbf{W}$. After expansion, we have

$$
\begin{aligned}
&E[B_k B_k^*] = \\
&\underbrace{\mathbf{s}_i\mathbf{s}_j^*\mathbf{s}_i^*\mathbf{s}_j}_{=1}\,\mathbf{h}_k^*\mathbf{h}_k\mathbf{h}_k^*\mathbf{h}_k + \mathbf{s}_i\mathbf{s}_j^*\mathbf{h}_k^*\mathbf{h}_k\mathbf{w}_{k,j}^*\mathbf{w}_{k,i} + \underbrace{\mathbf{s}_i\mathbf{s}_j^*\mathbf{s}_i^*\mathbf{h}_k^*\mathbf{h}_k\mathbf{h}_k\mathbf{w}_{k,j}^*}_{=0} + \underbrace{\mathbf{s}_i\mathbf{s}_j^*\mathbf{s}_j\mathbf{h}_k^*\mathbf{h}_k\mathbf{w}_{k,i}\mathbf{h}_k^*}_{=0} \\
&+ \mathbf{s}_j\mathbf{s}_i^*\mathbf{w}_{k,i}^*\mathbf{w}_{k,j}\mathbf{h}_k^*\mathbf{h}_k + \mathbf{w}_{k,i}^*\mathbf{w}_{k,j}\mathbf{w}_{k,j}^*\mathbf{w}_{k,i} + \underbrace{\mathbf{s}_i^*\mathbf{w}_{k,i}^*\mathbf{w}_{k,j}\mathbf{h}_k\mathbf{w}_{k,j}^*}_{=0} + \underbrace{\mathbf{s}_j\mathbf{w}_{k,i}^*\mathbf{w}_{k,j}\mathbf{w}_{k,i}\mathbf{h}_k^*}_{=0} \\
&+ \underbrace{\mathbf{s}_i\mathbf{s}_j\mathbf{s}_i^*\mathbf{h}_k^*\mathbf{w}_{k,j}\mathbf{h}_k^*\mathbf{h}_k}_{=0} + \underbrace{\mathbf{s}_i\mathbf{h}_k^*\mathbf{w}_{k,j}\mathbf{w}_{k,j}^*\mathbf{w}_{k,i}}_{=0} + \underbrace{\mathbf{s}_i\mathbf{s}_i^*}_{=1}\,\mathbf{h}_k^*\mathbf{w}_{k,j}\mathbf{h}_k\mathbf{w}_{k,j}^* + \mathbf{s}_i\mathbf{s}_j\mathbf{h}_k^*\mathbf{w}_{k,j}\mathbf{w}_{k,i}\mathbf{h}_k^* \\
&+ \underbrace{\mathbf{s}_j^*\mathbf{s}_j\mathbf{s}_i^*\mathbf{w}_{k,i}^*\mathbf{h}_k\mathbf{h}_k^*\mathbf{h}_k}_{=0} + \underbrace{\mathbf{s}_j^*\mathbf{w}_{k,i}^*\mathbf{h}_k\mathbf{w}_{k,j}^*\mathbf{w}_{k,i}}_{=0} + \mathbf{s}_j^*\mathbf{s}_i^*\mathbf{w}_{k,i}^*\mathbf{h}_k\mathbf{h}_k\mathbf{w}_{k,j}^* + \underbrace{\mathbf{s}_j^*\mathbf{s}_j}_{=1}\,\mathbf{w}_{k,i}^*\mathbf{h}_k\mathbf{w}_{k,i}\mathbf{h}_k^*
\end{aligned}
$$

Since we already assume that the entries of $\mathbf{h}$ are rotationally-invariant complex Gaussian with unit variance, then we can write $\mathbf{h}_k$ as $a + b\sqrt{-1}$, where a and b are independent, and both follow Gaussian distribution $\mathcal{N}(0, \frac{1}{2})$. Thus $E[\mathbf{h}_k^2] =$

$E[(\mathbf{h}_k^*)^2] = 0$. Furthermore,

$$
\begin{aligned}
E[|\mathbf{h}_k|^4] &= E[(a^2+b^2)^2] = E[a^4+b^4+2a^2b^2] \\
&= 3\sigma_a^4 + 3\sigma_b^4 + 2\sigma_a^2\sigma_b^2 \\
&= 2\times 3\times (\frac{1}{2})^2 + \frac{2}{4} = 2, \qquad (3.31)
\end{aligned}
$$

where $\sigma_a^2 = \frac{1}{2}$ and $\sigma_b^2 = \frac{1}{2}$ are respectively the variance of a and b. In the same way, we can find $E[|\mathbf{w}|^4] = 2\sigma_w^4$. Thus, when $i \neq j$,

$$
\begin{aligned}
E[B_kB_k^*] &= E[|\mathbf{h}_k|^4] + E[|\mathbf{w}_{k,i}|^2]E[|\mathbf{w}_{k,j}|^2] \\
&\quad + E[|\mathbf{h}_k|^2]E[|\mathbf{w}_{k,i}|^2] + E[|\mathbf{h}_k|^2]E[|\mathbf{w}_{k,j}|^2] \\
&= 2 + \sigma_w^4 + 2\sigma_w^2. \qquad (3.32)
\end{aligned}
$$

When $i = j$,

$$
\begin{aligned}
E[B_kB_k^*] &= \underbrace{E[|\mathbf{h}_k|^4]}_{=2} + \underbrace{E[|\mathbf{w}_{k,i}|^4]}_{=2\sigma_w^4} + \underbrace{E[|\mathbf{h}_k|^2]E[|\mathbf{w}_{k,i}|^2]}_{=\sigma_w^2} + \underbrace{E[|\mathbf{h}_k|^2]E[|\mathbf{w}_{k,i}|^2]}_{=\sigma_w^2} \\
&\quad + \underbrace{E[|\mathbf{h}_k|^2]E[|\mathbf{w}_{k,i}|^2]}_{=\sigma_w^2} + \mathbf{s}_i^2\,\underbrace{E[(\mathbf{h}_k^*)^2]E[(\mathbf{w}_{k,i})^2]}_{=0} \\
&\quad + (\mathbf{s}_i^2)^*\,\underbrace{E[(\mathbf{h}_k)^2]E[(\mathbf{w}_{k,i}^*)^2]}_{=0} + \underbrace{E[|\mathbf{h}_k|^2]E[|\mathbf{w}_{k,i}|^2]}_{=\sigma_w^2} \\
&= 2 + 2\sigma_w^4 + 4\sigma_w^2. \qquad (3.33)
\end{aligned}
$$

Moreover, after some algebra,

$$E[B_k]E[B_k^*] = \begin{cases} 1 + 2\sigma_w^2 + \sigma_w^4, & \text{if } i = j \\ \\ \mathbf{s}_i\mathbf{s}_j^*\mathbf{s}_j\mathbf{s}_i^* = 1, & \text{otherwise.} \end{cases}$$

Finally,

$$\begin{aligned} var(B_k) &= E[B_kB_k^*] - E[B_k]E[B_k^*] \\ &= \begin{cases} 1 + 2\sigma_w^2 + \sigma_w^4, & \text{if } i = j \\ \\ 1 + 2\sigma_w^2 + \sigma_w^4, & \text{otherwise} \end{cases} \end{aligned} \tag{3.34}$$

This leads to

$$var(\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N}) = (1 + 2\sigma_w^2 + \sigma_w^4)/N. \tag{3.35}$$

■

3.8.5 Proof of Lemma 3.4.2

Proof. For any i, $\mathbf{s}_i^*\mathbf{h}_j + \mathbf{w}_{j,i}$, $1 \leq j \leq N$, are N independent rotationally-invariant complex Gaussian random variables with variance $1 + \sigma_w^2$. Thus $\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}$ is twice the empirical average energy of $2N$ independent real-numbered Gaussian random variables each following distribution $\mathcal{N}(0, \frac{1+\sigma_w^2}{2})$.

We first look at the concentration of the empirical average energy of $2N$ real-numbered Gaussian random variables x_i, $1 \leq i \leq 2N$, each following distribution $\mathcal{N}(0, 1)$. For any $\epsilon > 0$, we apply the Chernoff bound to bound $P(\frac{\sum_{i=1}^{2N} x_i^2}{N} - 2 < -\epsilon)$:

$$
\begin{aligned}
&P(\frac{\sum_{i=1}^{2N} x_i^2}{N} - 2 < -\epsilon) \\
&\leq E\{e^{\lambda(-N\epsilon+2N-\sum_{i=1}^{2N} x_i^2)}\} \\
&\leq e^{\lambda(-N\epsilon+2N)} E\{[e^{-\lambda x^2}]^{2N}\} \\
&= e^{\lambda(-N\epsilon+2N)} \left[\int_{-\infty}^{\infty} e^{-\lambda x^2} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx\right]^{2N} \\
&= e^{\lambda(-N\epsilon+2N)} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\lambda+\frac{1}{2})x^2} dx\right]^{2N} \\
&= (\sqrt{\frac{1}{1+2\lambda}})^{2N} e^{\lambda(-N\epsilon+2N)} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x'^2}{2}} dx'\right]^{2N} \\
&= e^{\lambda(-N\epsilon+2N)} (\frac{1}{\sqrt{1+2\lambda}})^{2N}
\end{aligned}
$$

Taking the logarithm on both sides, we have

$$
\log(P(\frac{\sum_{i=1}^{2N} x_i^2}{N} - 2 < \epsilon)) \leq N[\lambda(-\epsilon+2) - \log(1+2\lambda)]. \tag{3.36}
$$

Take the derivative with respect to λ, and set the derivative to 0, we obtain the minimizing λ as $\lambda = \frac{\epsilon}{2(2-\epsilon)}$. Substitute this λ in (3.36), we obtain that

$$
\frac{1}{N} \log(P(\frac{\sum_{i=1}^{2N} x_i^2}{N} - 2 < -\epsilon)) \leq \frac{\epsilon}{2} - \log(1 + \frac{\epsilon}{2-\epsilon})
$$

By a proper linear scaling of $\frac{1+\sigma_w^2}{2}$, we have

$$-\frac{1}{N}\log\left(P(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)<-\epsilon)\right)$$
$$\geq -\frac{\epsilon}{1+\sigma_w^2}+\log(1+\frac{\epsilon}{(1+\sigma_w^2)-\epsilon})$$

Using the Taylor series, we have

$$\log(1+\frac{\epsilon}{(1+\sigma_w^2)-\epsilon})=\sum_{k=1}^{+\infty}\frac{\epsilon^k}{k(1+\sigma_w^2)^k}.$$

Thus

$$-\frac{1}{N}\log\left(P(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)<-\epsilon)\right)\geq\frac{\epsilon^2}{2(1+\sigma_w^2)^2}.$$

To bound $P(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)>+\epsilon)$, we similarly apply the Chernoff bound to characterize the concentration of twice the empirical average energy of $2N$ real-numbered Gaussian random variables each following distribution $\mathcal{N}(0,1)$.

$$P(\frac{\sum_{i=1}^{2N}x_i^2}{N}-2>\epsilon)$$
$$\leq E\{e^{\lambda(-N\epsilon-2N+\sum_{i=1}^{2N}x_i^2)}\}$$
$$=e^{-\lambda(N\epsilon+2N)}E\{[e^{-\lambda x^2}]^{2N}\}$$
$$=e^{-\lambda(N\epsilon+2N)}(\frac{1}{\sqrt{1-2\lambda}})^{2N}$$

We obtain λ minimizing $e^{-\lambda(N\epsilon+2N)}(\frac{1}{\sqrt{1-2\lambda}})^{2N}$ as $\lambda=\frac{\epsilon}{2(2+\epsilon)}$. Thus we can bound

$$
\begin{aligned}
&-\frac{1}{N}\log\left(P(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)>\epsilon)\right)\\
&\geq \frac{\epsilon}{1+\sigma_w^2}+\log(1-\frac{\epsilon}{(1+\sigma_w^2)+\epsilon})
\end{aligned}
$$

Using the Taylor series, we have

$$
\log(1-\frac{\epsilon}{(1+\sigma_w^2)+\epsilon})=\sum_{k=1}^{+\infty}(-1)^k\frac{\epsilon^k}{k(1+\sigma_w^2)^k}.
$$

Because

$$
\sum_{k=2}^{+\infty}(-1)^k\frac{\epsilon^k}{k(1+\sigma_w^2)^k}\geq\frac{\epsilon^2}{2(1+\sigma_w^2)^2}-\frac{\epsilon^3}{3(1+\sigma_w^2)^3},
$$

we have

$$
\begin{aligned}
&-\frac{1}{N}\log\left(P(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)>\epsilon)\right)\\
&\geq\frac{\epsilon^2}{2(1+\sigma_w^2)^2}-\frac{\epsilon^3}{3(1+\sigma_w^2)^3}.
\end{aligned}
$$

Moreover, if $\epsilon\leq 1+\sigma_w^2$, we have

$$
\begin{aligned}
&-\log\left(P(\frac{(\mathbf{X}^*\mathbf{X})_{i,i}}{N}-(1+\sigma_w^2)>+\epsilon)\right)/N\\
&\geq\frac{\epsilon^2}{2(1+\sigma_w^2)^2}-\frac{\epsilon^2}{3(1+\sigma_w^2)^2}=\frac{\epsilon^2}{6(1+\sigma_w^2)^2}.
\end{aligned}
$$

■

3.8.6 Proof of Lemma 3.4.3

Proof.

$$(\mathbf{X}^*\mathbf{X})_{i,j} = \mathbf{s}_i\mathbf{s}_j^* \sum_{l=1}^{N}(\mathbf{h}_l^* + \mathbf{s}_i^{-1}\mathbf{w}_{i,l}^*)(\mathbf{h}_l + (\mathbf{s}_j^*)^{-1}\mathbf{w}_{j,l}) = \mathbf{s}_i\mathbf{s}_j^* \sum_{l=1}^{N}(\mathbf{h}_l^* + \check{\mathbf{w}}_{i,l}^*)(\mathbf{h}_l + \check{\mathbf{w}}_{j,l}),$$

where $\check{\mathbf{w}}$'s have the same distribution as $\mathbf{w}$'s. For each l, let $\mathbf{h}_l = a_l + \tau b_l$, $\check{\mathbf{w}}_{i,l} = c_l + \tau d_l$, and $\check{\mathbf{w}}_{j,l} = e_l + \tau f_l$, where $\tau = \sqrt{-1}$, $a_l \sim \mathcal{N}(0, 1/2)$, $b_l \sim \mathcal{N}(0, 1/2)$, $c_l \sim \mathcal{N}(0, \sigma_w^2/2)$, $d_l \sim \mathcal{N}(0, \sigma_w^2/2)$, $e_l \sim \mathcal{N}(0, \sigma_w^2/2)$, and $f_l \sim \mathcal{N}(0, \sigma_w^2/2)$. Moreover, a_l's, b_l's, c_l's, d_l's, e_l's and f_l's are jointly independent. Thus, for each l, we have

$$\begin{aligned}
&(\mathbf{h}_l^* + \check{\mathbf{w}}_{i,l}^*)(\mathbf{h}_l + \check{\mathbf{w}}_{j,l}) = (a_l - \tau b_l + c_l - \tau d_l)(a_l + b_l\tau + e_l + \tau f_l) \\
&= a^2 + b^2 + (c_l - \tau d_l)(a_l + \tau b_l) + (e_l + \tau f_l)(a_l - \tau b_l) + (c_l - \tau d_l)(e_l + \tau f_l) \\
&= a_l^2 + b_l^2 + a_l c_l + b_l d_l + a_l e_l + f_l b_l + c_l e_l + d_l f_l + \tau(-a_l d_l + b_l c_l) \\
&+ \tau(a_l f_l - b_l e_l) + \tau(-d_l e_l + c_l f_l).
\end{aligned}$$

Using the triangle inequality, we have

$$\begin{aligned}
&|\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N} - \mathbf{s}_i\mathbf{s}_j^*| \\
&\leq \frac{1}{N}\Bigg(|\sum_{l=1}^{N}(a_l^2 + b_l^2 - 1)| + |\sum_{l=1}^{N} a_l c_l| + |\sum_{l=1}^{N} b_l d_l| + |\sum_{l=1}^{N} a_l e_l| \\
&+ |\sum_{l=1}^{N} b_l f_l| + |\sum_{l=1}^{N} a_l d_l| + |\sum_{l=1}^{N} a_l f_l| + |\sum_{l=1}^{N} b_l c_l| \\
&+ |\sum_{l=1}^{N} b_l e_l| + |\sum_{l=1}^{N} c_l e_l| + |\sum_{l=1}^{N} d_l f_l| + |\sum_{l=1}^{N} d_l e_l| + |\sum_{l=1}^{N} c_l f_l|\Bigg)
\end{aligned}$$

By the union bound, for any $\epsilon > 0$, we have

$$
\begin{aligned}
&P(|\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N} - \mathbf{s}_i\mathbf{s}_j^*| > \epsilon) \\
&\leq P(|\frac{\sum_{l=1}^N (a_l^2 + b_l^2 - 1)}{N}| > \epsilon/13) + P(|\frac{\sum_{l=1}^N a_l c_l}{N}| > \epsilon/13) \\
&+ P(\frac{|\sum_{l=1}^N b_l d_l|}{N} > \epsilon/13) + P(\frac{|\sum_{l=1}^N a_l e_l|}{N} > \epsilon/13) \\
&+ P(\frac{|\sum_{l=1}^N b_l f_l|}{N} > \epsilon/13) + \cdots \\
&+ P(\frac{|\sum_{l=1}^N c_l f_l|}{N} > \epsilon/13).
\end{aligned}
$$

In the inequality above, there is 1 term $P(\frac{|\sum_{l=1}^N (a_l^2+b_l^2-1)|}{N} > \epsilon/13)$, 8 terms with the same value as $P(\frac{|\sum_{l=1}^N a_l c_l|}{N} > \epsilon/13)$, and 4 terms with the same value as $P(\frac{|\sum_{l=1}^N c_l f_l|}{N} > \epsilon/13)$. Thus, we can further have

$$
\begin{aligned}
&P(|\frac{(\mathbf{X}^*\mathbf{X})_{i,j}}{N} - \mathbf{s}_i\mathbf{s}_j^*| > \epsilon) \\
&\leq P(\frac{\sum_{l=1}^N |a_l^2 + b_l^2 - 1|}{N} > \epsilon/13) + 8P(\frac{\sum_{l=1}^N |a_l c_l|}{N} > \epsilon/13) \\
&+ 4P(\frac{\sum_{l=1}^N |c_l f_l|}{N} > \epsilon/13).
\end{aligned}
$$

Using the results from Lemma 3.4.2, after proper scaling (taking $\sigma_w = 0$), we have

$$
\begin{aligned}
P(\frac{|\sum_{l=1}^N a_l^2 + b_l^2 - 1|}{N} > \epsilon/13) &\leq P(\frac{\sum_{l=1}^N (a_l^2 + b_l^2 - 1)}{N} > \epsilon/13) \\
&+ P(\frac{\sum_{l=1}^N (a_l^2 + b_l^2 - 1)}{N} < -\epsilon/13) \\
&\leq e^{-N[-\frac{\epsilon}{13} + \log(1 + \frac{\epsilon}{13-\epsilon})]} + e^{-N[\frac{\epsilon}{13} + \log(1 - \frac{\epsilon}{13+\epsilon})]} \\
&\leq e^{-\frac{N\epsilon^2}{338}} + e^{-\frac{N\epsilon^2}{1014}},
\end{aligned}
$$

where $\epsilon \leq 13$. For the concentration of the average of $a_l c_l$, we have the following lemma, whose proof is presented in Appendix 3.8.7.

Lemma 3.8.2. *Suppose that u_l and v_l, $1 \leq l \leq N$, are $2N$ independent standard Gaussian random variables following distribution $\mathcal{N}(0,1)$. Then*

$$\begin{aligned} &-\log\left(P\left(\frac{|\sum_{l=1}^{N} u_l v_l|}{N} > \epsilon\right)\right)/N \\ &\geq \frac{-1+\sqrt{1+4\epsilon^2}}{2} + \frac{1}{2}\log(1-[\frac{-1+\sqrt{1+4\epsilon^2}}{2\epsilon}]^2). \end{aligned}$$

This quantity is no smaller than $\frac{1}{2}\epsilon^2$ when $\epsilon \leq \sqrt{2}$.

Using Lemma 3.8.2, by proper scaling, we have

$$P(\frac{|\sum_{l=1}^{N} a_l c_l|}{N} > \epsilon/13) \leq e^{-N[\frac{-1+\sqrt{1+4\epsilon'^2}}{2}+\frac{1}{2}\log(1-[\frac{-1+\sqrt{1+4\epsilon'^2}}{2\epsilon}]^2)]}$$

where $\epsilon' = \frac{2\epsilon}{13\sigma_w}$. Moreover, when $\epsilon \leq \frac{13\sigma_w}{\sqrt{2}}$,

$$P(\frac{|\sum_{l=1}^{N} a_l c_l|}{N} > \epsilon/13) \leq e^{-\frac{N\epsilon^2}{169\sigma_w^2}}.$$

Similarly, we have

$$P(\frac{|\sum_{l=1}^{N} c_l f_l|}{N} > \epsilon/13) \leq e^{-N[\frac{-1+\sqrt{1+4\epsilon''^2}}{2}+\frac{1}{2}\log(1-[\frac{-1+\sqrt{1+4\epsilon''^2}}{2\epsilon}]^2)]}$$

where $\epsilon'' = \frac{2\epsilon}{13\sigma_w^2}$. Moreover, when $\epsilon \leq \frac{13\sigma_w^2}{\sqrt{2}}$,

$$P(\frac{|\sum_{l=1}^{N} c_l f_l|}{N} > \epsilon/13) \leq e^{-\frac{N\epsilon^2}{169\sigma_w^4}}.$$

■

3.8.7 Proof of Lemma 3.8.2

Proof. Let us assume x_l and y_l, $1 \leq l \leq N$, are $2N$ jointly independent random variables following distribution $\mathcal{N}(0,1)$. Then by the Chernoff bound, for $\lambda \geq 0$, we obtain

$$\begin{aligned}
&P(\frac{\sum_i^N x_i y_i}{N} > \epsilon) = P(\sum_i^N x_i y_i > N\epsilon) \\
&\leq E\{e^{\lambda(\sum_i^N x_i y_i - N\epsilon)}\} \\
&= (\frac{1}{\sqrt{2\pi}})^{2N} e^{-\lambda N\epsilon}[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-1}{2}(x^2+y^2-2\lambda xy)}\, dxdy]^N \\
&= (\frac{1}{\sqrt{2\pi}})^{2N} e^{-\lambda N\epsilon}[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-1}{2}[\,x\ y\,]\left[\begin{smallmatrix}1 & -\lambda\\ -\lambda & 1\end{smallmatrix}\right]\left[\begin{smallmatrix}x\\ y\end{smallmatrix}\right]}\, dxdy]^N \\
&= (\frac{1}{\sqrt{2\pi}})^{2N} e^{-\lambda N\epsilon}[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-1}{2}(A\left[\begin{smallmatrix}x\\ y\end{smallmatrix}\right])^*(A\left[\begin{smallmatrix}x\\ y\end{smallmatrix}\right])}\, dxdy]^N \\
&= (\frac{1}{\sqrt{2\pi}})^{2N} e^{-\lambda N\epsilon}[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-1}{2}[\,x'\ y'\,]\left[\begin{smallmatrix}x'\\ y'\end{smallmatrix}\right]}\, dx'dy' \frac{1}{det(A)}]^N \\
&= (\frac{1}{\sqrt{2\pi}})^{2N} e^{-\lambda N\epsilon}[\sqrt{2\pi}\sqrt{2\pi} \times \frac{1}{det(A)}]^N \\
&= e^{-\lambda N\epsilon} \times [\frac{1}{\sqrt{1-\lambda^2}}]^N.
\end{aligned}$$

where $\left[\begin{smallmatrix} 1 & -\lambda \\ -\lambda & 1 \end{smallmatrix}\right] = A^* A$, $[x' \; y']^T = A\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$, and $\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\, dx = \sqrt{2\pi}$. By taking the logarithm of $e^{-\lambda N \epsilon} \times [\frac{1}{\sqrt{1-\lambda^2}}]^N$, we have

$$-\lambda N \epsilon - \frac{N}{2} \log(1 - \lambda^2). \tag{3.37}$$

To minimize (3.37), we take its derivative with respect to λ, and set the derivative to 0 to get $\epsilon\lambda^2 + \lambda - \epsilon = 0$. Since $\lambda \geq 0$, we have

$$\lambda = \frac{-1 + \sqrt{1 + 4\epsilon^2}}{2\epsilon}$$

Thus this leads to

$$\begin{aligned}
&- \log\left(P\left(\frac{|\sum_{l=1}^{N} u_l v_l|}{N} > \epsilon\right)\right)/N \\
&\geq \frac{-1 + \sqrt{1 + 4\epsilon^2}}{2} + \frac{1}{2}\log(1 - [\frac{-1 + \sqrt{1 + 4\epsilon^2}}{2\epsilon}]^2) \\
&\doteq f(\epsilon).
\end{aligned}$$

We note that $f(0) = 0$ and the derivative $f'(\epsilon) = 2\epsilon \frac{\sqrt{1+4\epsilon^2}}{1+4\epsilon^2+\sqrt{1+4\epsilon^2}}$. For $\epsilon \leq \sqrt{2}$, $f'(\epsilon) \geq \frac{\epsilon}{2}$. This implies that

$$f(\epsilon) \geq \frac{1}{4}\epsilon^2$$

when $\epsilon \leq \sqrt{2}$.

■

3.8.8 Proof of Lemma 3.4.4

Proof. By the triangle inequality for the maximum eigenvalue, we have

$$
\begin{aligned}
|\rho - \rho_E| &\leq \max_{\|v\|_2 \leq 1} \|(\mathbf{X}^*\mathbf{X}/N - E\{\mathbf{X}^*\mathbf{X}\}/N)v\|_2 \\
&= \max_{\|v\|_2 \leq 1} \sqrt{(E_{1,:}v)^2 + (E_{2,:}v)^2 + \cdots + (E_{T,:}v)^2} \\
&\leq \max_{\|v\|_2 \leq 1} \sqrt{\|E_{1,:}\|_2^2\|v\|^2 + \cdots + \|E_{T,:}\|_2^2\|v\|^2} \\
&= \sqrt{T\epsilon^2 + \cdots + T\epsilon^2} \\
&= T\epsilon,
\end{aligned}
$$

where $E_{i,:}$, $1 \leq i \leq T$, denotes the i-th row of $(\mathbf{X}^*\mathbf{X}/N - E\{\mathbf{X}^*\mathbf{X}\}/N)$, and $\|E_{i,:}\|_2^2 \leq T\epsilon$.

In addition, by the triangle inequality for absolute value, each entry of $\rho I - \mathbf{X}^*\mathbf{X}/N$ deviates from its counterpart in $\rho_E I - E\{\mathbf{X}^*\mathbf{X}\}/N$ by at most $\epsilon + T\epsilon = (T + 1)\epsilon$. ∎

3.8.9 Proof of Lemma 3.4.5

Proof. Our proof relies on a recursive perturbation analysis of the recursive calculations of the Cholesky decomposition of $\mathbf{A}$.

Let us use k to denote the iteration number of the Cholesky decomposition, where $1 \leq k \leq T$, and we start with $k = 1$. In the k-th iteration, $1 \leq k \leq T$, we have a new matrix $\mathbf{A}^k \in \mathcal{C}^{(T-k+1)\times(T-k+1)}$, and, to be consistent with the index in matrix $\mathbf{A}$, we use $\mathbf{A}^k_{i,j}$ to denote the element of $\mathbf{A}^k$ in its $(i - k + 1)$-th row and $(j - k + 1)$-th

column. Moreover, we define $\mathbf{A}^1 = \mathbf{A}$.

In the k-th iteration, we first compute $\mathbf{R}_{k,k} = \sqrt{\mathbf{A}^k_{k,k}}$. Then we calculate $\mathbf{R}_{k,(k+1):T} = \mathbf{A}^k_{k,(k+1):T}/\mathbf{R}_{k,k}$, where $\mathbf{R}_{k,(k+1):T}$ denotes the submatrix of $\mathbf{R}$ with row index k and column indices $k+1$, $k+2$, ..., and T. After this, we update matrix $\mathbf{A}^{k+1}$ as

$$\mathbf{A}^{k+1} = \mathbf{A}^k_{(k+1):T,(k+1):T} - \mathbf{R}^*_{k,(k+1):T} \times \mathbf{R}_{k,(k+1):T}, \tag{3.38}$$

where the column-vector $\mathbf{R}^*_{k,(k+1):T} = (\mathbf{R}_{k,(k+1):T})^*$ denotes the conjugate transpose of the row-vector $\mathbf{R}_{k,(k+1):T}$.

When we calculate the Cholesky decomposition for $\grave{\mathbf{A}}$, we denote the counterparts in the calculations as $\grave{\mathbf{A}}^k$. From calculations in Section 3.3, we know that

$$\grave{\mathbf{A}}^k_{k,k} = \frac{T(T-k)}{T-k+1},$$

and, when $i > k$,

$$\grave{\mathbf{A}}^k_{k,i} = -(\mathbf{s}_k\mathbf{s}_i^*)\frac{T}{T-k+1},$$

$$\grave{\mathbf{R}}_{k,i} = -(\mathbf{s}_k\mathbf{s}_i^*)\sqrt{\frac{T}{(T-k)(T-k+1)}}.$$

To bound the deviation of $\mathbf{R}$ from $\grave{\mathbf{R}}$, we thus need to bound the deviation of matrices $\mathbf{A}^k$ from $\grave{\mathbf{A}}^k$. Let us use $\mathbf{E}^k$ of dimension $(T-k+1) \times (T-k+1)$ to denote the error matrix $\mathbf{E}^k = \mathbf{A}^k - \grave{\mathbf{A}}^k$. To be consistent with the indices in matrix $\mathbf{A}$, we use $\mathbf{E}^k_{i,j}$ to denote the element in the $(i-k+1)$-th row and $(j-k+1)$-th column of $\mathbf{E}^k$.

Let us further use $\epsilon_k \geq 0$ to denote the maximum magnitude of all the elements in $\mathbf{E}^k$.

Our proof strategy of Lemma 3.4.5 is to bound ϵ_k by induction over k. In fact, we will show that, for any k,

$$|\epsilon_{k+1}| \leq |\epsilon_k| \frac{(T-k+1)^2}{(T-k)^2}(1 + \frac{|\epsilon_k|}{\grave{\mathbf{A}}^k_{k,k} - |\epsilon_k|}). \tag{3.39}$$

(The proof of this fact will proceed until (3.48).) The relation (3.39) will lead us to conclude that $|\epsilon_k| \leq |\epsilon_1| \frac{T^2}{(T-(k-1))^2} e^4 < 1$ holds true for every $1 \leq k \leq T$, when $|\epsilon_1| < \frac{1}{e^4 T^2}$.

To bound ϵ_k by induction, we start with $\epsilon_1 \leq \epsilon < \frac{1}{e^{4T^2}}$. Suppose that we know ϵ_k for a certain k, and we will derive the relation between ϵ_k and ϵ_{k+1}. Because

$$\mathbf{E}^{k+1} = \mathbf{E}^k_{k+1:T,k+1:T} - (\mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1}),$$

in order to bound ϵ_{k+1}, we first investigate $\mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1}$. Using (3.38), we have

$$\begin{aligned}
&\mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1} \\
&= (\mathbf{A}^k_{(k+1):T,(k+1):T} - \grave{\mathbf{A}}^k_{(k+1):T,(k+1):T}) - (\mathbf{A}^{k+1} - \grave{\mathbf{A}}^{k+1}) \\
&= (\mathbf{A}^k_{(k+1):T,(k+1):T} - \mathbf{A}^{k+1}) - (\grave{\mathbf{A}}^k_{(k+1):T,(k+1):T} - \grave{\mathbf{A}}^{k+1}) \\
&= \mathbf{R}^*_{k,(k+1):T} \times \mathbf{R}_{k,(k+1):T} - \grave{\mathbf{R}}^*_{k,(k+1):T} \times \grave{\mathbf{R}}_{k,(k+1):T}.
\end{aligned} \tag{3.40}$$

By definition, $\mathbf{A}^k_{k,k} = \grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k}$. Thus $\mathbf{R}_{k,k} = \sqrt{\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k}}$, and $\grave{\mathbf{R}}_{k,k} = \sqrt{\grave{\mathbf{A}}^k_{k,k}}$. Furthermore, we have $\mathbf{A}^k_{k,i} = \grave{\mathbf{A}}^k_{k,i} + \mathbf{E}^k_{k,i}$. Then $\mathbf{R}_{k,i} = \frac{\grave{\mathbf{A}}^k_{k,i} + \mathbf{E}^k_{k,i}}{\mathbf{R}_{k,k}}$, and $\mathbf{R}_{k,j} = \frac{\grave{\mathbf{A}}^k_{k,j} + \mathbf{E}^k_{k,j}}{\mathbf{R}_{k,k}}$.

For the recursive calculations of the Cholesky decomposition in (3.38) and the evolution of error matrix $\mathbf{E}^k$ in (3.40), we have

$$\begin{aligned}\mathbf{R}_{k,i}^*\mathbf{R}_{k,j} &= (\frac{\grave{\mathbf{A}}_{k,i}^k}{\mathbf{R}_{k,k}} + \frac{\mathbf{E}_{k,i}^k}{\mathbf{R}_{k,k}})^*(\frac{\grave{\mathbf{A}}_{k,j}^k}{\mathbf{R}_{k,k}} + \frac{\mathbf{E}_{k,j}^k}{\mathbf{R}_{k,k}}) \\ &= \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\grave{\mathbf{A}}_{k,j}^k}{\mathbf{R}_{k,k}^2} + \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\mathbf{E}_{k,j}^k}{\mathbf{R}_{k,k}^2} + \frac{\grave{\mathbf{A}}_{k,j}^k(\mathbf{E}_{k,i}^k)^*}{\mathbf{R}_{k,k}^2} + \frac{(\mathbf{E}_{k,i}^k)^*\mathbf{E}_{k,j}^k}{\mathbf{R}_{k,k}^2}\end{aligned} \tag{3.41}$$

Since $\grave{\mathbf{R}}_{k,i}^*\grave{\mathbf{R}}_{k,j} = \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\grave{\mathbf{A}}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}}$, for any (i,j) pair such that $(k+1) \le i,j \le T$, we have

$$\begin{aligned}&(\mathbf{E}_{k+1:T,k+1:T}^k - \mathbf{E}^{k+1})_{i,j} \\ &= \mathbf{R}_{k,i}^*\mathbf{R}_{k,j} - \grave{\mathbf{R}}_{k,i}^*\grave{\mathbf{R}}_{k,j} \\ &= \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\grave{\mathbf{A}}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k} + \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\mathbf{E}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k} + \frac{\grave{\mathbf{A}}_{k,j}^k(\mathbf{E}_{k,i}^k)^*}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k} \\ &+ \frac{(\mathbf{E}_{k,i}^k)^*\mathbf{E}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k} - \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\grave{\mathbf{A}}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}^k} \\ &= \frac{-(\grave{\mathbf{A}}_{k,i}^k)^*\grave{\mathbf{A}}_{k,j}^k\mathbf{E}_{k,k}^k}{\grave{\mathbf{A}}_{k,k}^k(\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k)} + \frac{(\grave{\mathbf{A}}_{k,i}^k)^*\mathbf{E}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k} \\ &+ \frac{\grave{\mathbf{A}}_{k,j}^k(\mathbf{E}_{k,i}^k)^*}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k} + \frac{(\mathbf{E}_{k,i}^k)^*\mathbf{E}_{k,j}^k}{\grave{\mathbf{A}}_{k,k}^k + \mathbf{E}_{k,k}^k}\end{aligned} \tag{3.42}$$

To bound the elements of $\mathbf{E}^{k+1}$, we divide each element of $\mathbf{E}_{k+1:T,k+1:T}^k - \mathbf{E}^{k+1}$ to two parts: The first-order approximation and the higher-order terms. Namely,

$$\mathbf{E}_{k+1:T,k+1:T}^k - \mathbf{E}^{k+1} = \mathbf{B}^{k+1,1st} + \mathbf{B}^{k+1,high}.$$

More specifically, the first-order $(T-k) \times (T-k)$ approximation matrix $\mathbf{B}^{k+1,1st}$ for

$(\mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1})$ has in its $(i-k)$-th row and $(j-k)$-th column:

$$(\mathbf{B}^{k+1,1st})_{i,j} = \frac{-(\grave{\mathbf{A}}^k_{k,i})^* \grave{\mathbf{A}}^k_{k,j} \mathbf{E}^k_{k,k}}{(\grave{\mathbf{A}}^k_{k,k})^2} + \frac{(\grave{\mathbf{A}}^k_{k,i})^* \mathbf{E}^k_{k,j}}{\grave{\mathbf{A}}^k_{k,k}} + \frac{\grave{\mathbf{A}}^k_{k,j} (\mathbf{E}^k_{k,i})^*}{\grave{\mathbf{A}}^k_{k,k}}. \tag{3.43}$$

The higher-order term matrix for $\mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1}$ is thus given by $\mathbf{B}^{k+1,high} = \mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1} - \mathbf{B}^{k+1,1st}$. Moreover,

$$\begin{aligned}
&(\mathbf{B}^{k+1,high})_{i,j} \\
&= (\mathbf{E}^k_{k+1:T,k+1:T} - \mathbf{E}^{k+1} - \mathbf{B}^{k+1,1st})_{i,j} \\
&= \left(\frac{-(\grave{\mathbf{A}}^k_{k,i})^* \grave{\mathbf{A}}^k_{k,j} \mathbf{E}^k_{k,k}}{\grave{\mathbf{A}}^k_{k,k} (\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k})} + \frac{(\grave{\mathbf{A}}^k_{k,i})^* \grave{\mathbf{A}}^k_{k,j} \mathbf{E}^k_{k,k}}{(\grave{\mathbf{A}}^k_{k,k})^2} \right) \\
&+ \left(\frac{(\grave{\mathbf{A}}^k_{k,i})^* \mathbf{E}^k_{k,j}}{\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k}} - \frac{(\grave{\mathbf{A}}^k_{k,i})^* \mathbf{E}^k_{k,j}}{\grave{\mathbf{A}}^k_{k,k}} \right) \\
&+ \left(\frac{\grave{\mathbf{A}}^k_{k,j} (\mathbf{E}^k_{k,i})^*}{\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k}} - \frac{\grave{\mathbf{A}}^k_{k,j} (\mathbf{E}^k_{k,i})^*}{\grave{\mathbf{A}}^k_{k,k}} \right) + \frac{(\mathbf{E}^k_{k,i})^* \mathbf{E}^k_{k,j}}{\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k}} \\
&= \frac{(\grave{\mathbf{A}}^k_{k,i})^* \grave{\mathbf{A}}^k_{k,j} (\mathbf{E}^k_{k,k})^2}{\mathbf{A}^2_{k,k} (\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k})} - \frac{(\grave{\mathbf{A}}^k_{k,i})^* \mathbf{E}^k_{k,j} \mathbf{E}^k_{k,k}}{\grave{\mathbf{A}}^k_{k,k} (\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k})} \\
&- \frac{\grave{\mathbf{A}}^k_{k,j} (\mathbf{E}^k_{k,i})^* \mathbf{E}^k_{k,k}}{\grave{\mathbf{A}}^k_{k,k} (\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k})} + \frac{(\mathbf{E}^k_{k,i})^* \mathbf{E}^k_{k,j}}{\grave{\mathbf{A}}^k_{k,k} + \mathbf{E}^k_{k,k}}.
\end{aligned} \tag{3.44}$$

By the triangle inequality, we can upper bound each entry of $\mathbf{E}^{k+1}$. More specifically, for all the i's and j's such that $k+1 \le i \le T$ and $k+1 \le j \le T$, we have

$$
\begin{aligned}
&|\mathbf{E}_{i,j}^{k+1}| \\
&= |(\mathbf{E}_{i,j}^{k} - \mathbf{E}_{i,j}^{k+1}) - \mathbf{E}_{i,j}^{k}| \\
&= |(\mathbf{B}_{i,j}^{k+1,1st} + \mathbf{B}_{i,j}^{k+1,high}) - \mathbf{E}_{i,j}^{k})| \\
&= |\left(\mathbf{B}_{i,j}^{k+1,1st} + (\mathbf{E}_{i,j}^{k} - \mathbf{E}_{i,j}^{k+1} - \mathbf{B}_{i,j}^{k+1,1st})\right) - \mathbf{E}_{i,j}^{k}| \\
&\leq |\mathbf{E}_{i,j}^{k}| + |\mathbf{B}_{i,j}^{k+1,1st}| + |\mathbf{E}_{i,j}^{k} - \mathbf{E}_{i,j}^{k+1} - \mathbf{B}_{i,j}^{k+1,1st}|. \qquad (3.45)
\end{aligned}
$$

Recall that $\mathbf{E}^k$ is the error matrix in the k-th iteration. Using (3.43), we can upper bound $|\mathbf{E}_{i,j}^{k}| + |\mathbf{B}_{i,j}^{k+1,1st}|$ as follows:

$$
\begin{aligned}
&|\mathbf{E}_{i,j}^{k}| + |\mathbf{B}_{i,j}^{k+1,1st}| \\
&\leq |\epsilon_k| + \frac{(\frac{T}{T-k+1})^2|\mathbf{E}_{k,k}^{k}|}{|\grave{\mathbf{A}}_{k,k}^{k}|^2} + \frac{\frac{T}{T-k+1}|\mathbf{E}_{k,j}^{k}|}{|\grave{\mathbf{A}}_{k,k}^{k}|} + \frac{\frac{T}{T-k+1}|(\mathbf{E}_{k,i}^{k})^*|}{|\grave{\mathbf{A}}_{k,k}^{k}|} \\
&\leq |\epsilon_k| + \frac{(\frac{T}{T-k+1})^2|\epsilon_k|}{|\grave{\mathbf{A}}_{k,k}^{k}|^2} + \frac{2\frac{T}{T-k+1}|\epsilon_k|}{|\grave{\mathbf{A}}_{k,k}^{k}|} \\
&= |\epsilon_k|\frac{|\grave{\mathbf{A}}_{k,k}^{k}|^2 + 2\frac{T}{T-k+1}|\grave{\mathbf{A}}_{k,k}^{k}| + (\frac{T}{T-k+1})^2}{|\grave{\mathbf{A}}_{k,k}^{k}|^2} \\
&= |\epsilon_k|\frac{(|\grave{\mathbf{A}}_{k,k}^{k}| + \frac{T}{T-k+1})^2}{|\grave{\mathbf{A}}_{k,k}^{k}|^2} \\
&= |\epsilon_k|\frac{(T-k+1)^2}{(T-k)^2}, \qquad (3.46)
\end{aligned}
$$

where we used the fact that each element of $\mathbf{E}^k$ is upper bounded by ϵ_k in magnitude and $\grave{\mathbf{A}}_{k,k}^{k} = \frac{T(T-k)}{T-k+1}$.

Moreover, using (3.44) and the triangular inequality, we can bound the magnitude of each entry of the higher-order residual $\mathbf{E}_{i,j}^{k} - \mathbf{E}_{i,j}^{k+1} - \mathbf{B}_{i,j}^{k+1,1st}$ as follows:

$$
\begin{aligned}
&|\mathbf{E}_{i,j}^{k} - \mathbf{E}_{i,j}^{k+1} - \mathbf{B}_{i,j}^{k+1,1st}| \\
&\leq \frac{(\frac{T}{T-k+1})^2|\epsilon_k|^2}{(\grave{\mathbf{A}}_{k,k}^{k})^2(\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|)} + \frac{\frac{T}{T-k+1}|\epsilon_k|^2}{\grave{\mathbf{A}}_{k,k}^{k}(\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|)} \\
&+ \frac{\frac{T}{T-k+1}|\epsilon_k|^2}{\grave{\mathbf{A}}_{k,k}^{k}(\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|)} + \frac{|\epsilon_k|^2}{\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|} \\
&= \frac{|\epsilon_k|^2}{(\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|)}[\frac{1}{(T-k)^2} + \frac{2}{(T-k)} + 1] \\
&= \frac{|\epsilon_k|^2}{(\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|)}\frac{(T-k+1)^2}{(T-k)^2}
\end{aligned}
\tag{3.47}
$$

Combining (3.45), (3.46) and (3.47), we can bound the magnitude of each element in $\mathbf{E}^{k+1}$. More specifically,

$$
\begin{aligned}
|\mathbf{E}_{i,j}^{k+1}| &\leq |\epsilon_k|\frac{(T-k+1)^2}{(T-k)^2} + \frac{|\epsilon_k|^2}{\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|}\frac{(T-k+1)^2}{(T-k)^2} \\
&= |\epsilon_k|\frac{(T-k+1)^2}{(T-k)^2}(1 + \frac{|\epsilon_k|}{\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|})
\end{aligned}
$$

holds true for i and j such that $k+1 \leq i \leq T$ and $k+1 \leq j \leq T$. This means

$$
|\epsilon_{k+1}| \leq |\epsilon_k|\frac{(T-k+1)^2}{(T-k)^2}(1 + \frac{|\epsilon_k|}{\grave{\mathbf{A}}_{k,k}^{k} - |\epsilon_k|}). \tag{3.48}
$$

We claim that, if $T \geq 4$ and $|\epsilon_1|T^2e^4 < 1$, then $|\epsilon_k| \leq |\epsilon_1|\frac{T^2}{(T-(k-1))^2}e^4$ holds true for every $1 \leq k \leq T$. To see this, we also perform an induction on k, based on (3.48).

For $k = 1$, $|\epsilon_1| \leq |\epsilon_1| e^4$ is true because $e^4 \geq 1$. Suppose $|\epsilon_i| \leq |\epsilon_1| \frac{T^2}{(T-(i-1))^2} e^4$ holds true for $1 \leq i \leq k$, where k is an integer such that $1 \leq k \leq T-1$. For any i such that $1 \leq i \leq k$, because $|\epsilon_i| \leq |\epsilon_1| \frac{T^2}{(T-(i-1))^2} e^4 \leq |\epsilon_1| T^2 e^4 < 1$, and $T \geq 4$, we have $|\epsilon_i| < 1 \leq \frac{T}{4}$. Thus for any i such that $1 \leq i \leq k$, by (3.48), we have

$$
\begin{aligned}
|\epsilon_{i+1}| &\leq |\epsilon_i| \frac{(T-i+1)^2}{(T-i)^2} (1 + \frac{|\epsilon_i|}{\grave{\mathbf{A}}^i_{i,i} - |\epsilon_i|}) \\
&\leq |\epsilon_i| \frac{(T-i+1)^2}{(T-i)^2} (1 + \frac{1}{T/4}),
\end{aligned} \tag{3.49}
$$

because $\grave{\mathbf{A}}^i_{i,i}$ is no smaller than $T/2$ when $i \leq T-1$.

By applying (3.49) recursively for $|\epsilon_{k+1}|$, $|\epsilon_k|$, ..., and $|\epsilon_1|$, we have

$$
\begin{aligned}
|\epsilon_{k+1}| &\leq |\epsilon_1| \frac{T^2}{(T-k)^2} (1 + \frac{1}{T/4})^k \\
&\leq |\epsilon_1| \frac{T^2}{(T-k)^2} (1 + \frac{1}{T/4})^T \\
&\leq |\epsilon_1| \frac{T^2}{(T-k)^2} e^4,
\end{aligned}
$$

because $k \leq T-1$ and $(1 + \frac{4}{T})^T \leq e^4$. By this induction, we have proved that $|\epsilon_k| \leq |\epsilon_1| \frac{T^2}{(T-(k-1))^2} e^4 < 1$ holds true for every $1 \leq k \leq T$, when $|\epsilon_1| < \frac{e^{-4}}{T^2}$.

Suppose that each element of $\mathbf{E}^k$ is upper bounded by $|\epsilon_k|$ in magnitude. Then the deviation of $\mathbf{R}_{k,k}$ is upper bounded as follows:

$$
\begin{aligned}
|\mathbf{R}_{k,k} - \grave{\mathbf{R}}_{k,k}| &= |\sqrt{\grave{\mathbf{A}}^{k}_{k,k} + \mathbf{E}^{k}_{k,k}} - \sqrt{\grave{\mathbf{A}}^{k}_{k,k}}| \\
&= |\frac{\mathbf{E}^{k}_{k,k}}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k} + \mathbf{E}^{k}_{k,k}} + \sqrt{\grave{\mathbf{A}}^{k}_{k,k}}}| \\
&\leq \frac{|\epsilon_k|}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k}}} = |\epsilon_k|\sqrt{\frac{T-k+1}{T(T-k)}} \\
&\leq |\epsilon| e^4 \sqrt{\frac{T^3}{(T-k+1)^3(T-k)}} \\
&\leq |\epsilon| e^4 \sqrt{\frac{T^3}{8}},
\end{aligned} \tag{3.50}
$$

where we have used $|\epsilon_k| \leq |\epsilon_1| \frac{T^2}{(T-k+1)^2} e^4 < 1 \leq \frac{T}{4}$, $\epsilon_1 \leq \epsilon$, and $k \leq T-1$ in the last two steps. Furthermore, we can bound the deviation in $\mathbf{R}_{k,j}$, where $j > k$, as follows:

$$
\begin{aligned}
&|\mathbf{R}_{k,j} - \grave{\mathbf{R}}_{k,j}| = |\frac{\grave{\mathbf{A}}^{k}_{k,j} + \mathbf{E}^{k}_{k,j}}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k} + \mathbf{E}^{k}_{k,k}}} - \frac{\grave{\mathbf{A}}^{k}_{k,j}}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k}}}| \\
&\leq |\frac{\grave{\mathbf{A}}^{k}_{k,j}}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k} + \mathbf{E}^{k}_{k,k}}} - \frac{\grave{\mathbf{A}}^{k}_{k,j}}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k}}}| + |\frac{\mathbf{E}^{k}_{k,j}}{\sqrt{\grave{\mathbf{A}}^{k}_{k,j} + \mathbf{E}^{k}_{k,k}}}| \\
&\leq \frac{|\grave{\mathbf{A}}^{k}_{k,j}||\epsilon_k|}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k}}\sqrt{\grave{\mathbf{A}}^{k}_{k,k} - |\epsilon_k|}(\sqrt{\grave{\mathbf{A}}^{k}_{k,k} - |\epsilon_k|} + \sqrt{\grave{\mathbf{A}}^{k}_{k,k}})} + \frac{|\epsilon_k|}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k} - |\epsilon_k|}} \\
&\leq \frac{|\grave{\mathbf{A}}^{k}_{k,j}||\epsilon_k|}{2(\sqrt{\grave{\mathbf{A}}^{k}_{k,k} - |\epsilon_k|})^3} + \frac{|\epsilon_k|}{\sqrt{\grave{\mathbf{A}}^{k}_{k,k} - |\epsilon_k|}} \\
&\leq |\epsilon_k| \frac{T}{2(T-k+1)(\sqrt{\frac{T(T-k)}{T-k+1} - |\epsilon_k|})^3} + \frac{|\epsilon_k|}{\sqrt{\frac{T(T-k)}{T-k+1} - |\epsilon_k|}} \leq \frac{2|\epsilon_k|}{\sqrt{T}} + \frac{2|\epsilon_k|}{\sqrt{T}} \\
&\leq |\epsilon| e^4 T\sqrt{T},
\end{aligned} \tag{3.51}
$$

where we also used $|\epsilon_k| \le |\epsilon_1| \frac{T^2}{(T-k+1)^2} e^4 < 1 \le \frac{T}{4}$ and $k \le T-1$ in the last three steps. Moreover, for $k = T$, we have

$$|\mathbf{R}_{T,T} - \grave{\mathbf{R}}_{T,T}| = \mathbf{R}_{T,T} \le T e^2 \sqrt{|\epsilon|},$$

where we applied $|\epsilon_k| \le |\epsilon_1| \frac{T^2}{(T-k+1)^2} e^4$ to $k = T$. In summary, we have the results stated in Lemma 3.4.5. ■

3.8.10 Proof of Lemma 3.4.6

Proof. For $\mathbf{s}_{i:T}$, we know the unscaled metric $M^{\mathbf{R}}_{\mathbf{s}^*_{i:T}} = \|\mathbf{R}_{i:T,i:T}\mathbf{s}_{i:T}\|^2$. First we will upper bound $\sqrt{M^{\mathbf{R}}_{\mathbf{s}^*_{i:T}}}$:

$$\begin{aligned}
\|\mathbf{R}_{i:T,i:T}\mathbf{s}^*_{i:T}\| &= \|(\grave{\mathbf{R}}_{i:T,i:T} + \mathbf{P}_{i:T,i:T})\mathbf{s}^*_{i:T}\| \\
&\le \|\grave{\mathbf{R}}_{i:T,i:T}\mathbf{s}^*_{i:T}\| + \|\mathbf{P}_{i:T,i:T}\mathbf{s}_{i:T}\| \\
&\le \sqrt{M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}} + \|\mathbf{P}_{i:T,i:T}\|_F \|\mathbf{s}_{i:T}\| \\
&\le \sqrt{M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}} + \sqrt{\sum_{t=i}^{T}\sum_{j=i}^{T} |\mathbf{P}_{t,j}|^2} \sqrt{T-i+1}, \qquad (3.52)
\end{aligned}$$

where $M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}} = \|\grave{\mathbf{R}}_{i:T}\mathbf{s}_{i:T}\|^2$. Similarly we can get the lower bound:

$$\begin{aligned}
\|\mathbf{R}_{i:T,i:T}\mathbf{s}^*_{i:T}\| &= \|(\grave{\mathbf{R}}_{i:T,i:T} + \mathbf{P}_{i:T,i:T})\mathbf{s}^*_{i:T}\| \\
&\ge \|\grave{\mathbf{R}}_{i:T,i:T}\mathbf{s}^*_{i:T}\| - \|\mathbf{P}_{i:T,i:T}\mathbf{s}_{i:T}\| \\
&\le \sqrt{M^{\grave{\mathbf{R}}}_{\mathbf{s}^*_{i:T}}} - \sqrt{\sum_{t=i}^{T}\sum_{j=i}^{T} |\mathbf{P}_{t,j}|^2} \sqrt{T-i+1}. \qquad (3.53)
\end{aligned}$$

■

3.8.11 Proof of Lemma 3.2.3 (and Lemma 3.3.3)

Proof. We first look at the metric of sequences under $\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$. For nonconstant-modulus constellations, $\grave{\mathbf{A}} = \rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N}$ is given by

$$\grave{\mathbf{A}} = \begin{bmatrix} t - \mathbf{s}_1\mathbf{s}_1^* & -\mathbf{s}_1\mathbf{s}_2^* & \cdots & -\mathbf{s}_1\mathbf{s}_T^* \\ -\mathbf{s}_2\mathbf{s}_1^* & t - \mathbf{s}_2\mathbf{s}_2^* & \cdots & -\mathbf{s}_2\mathbf{s}_T^* \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{s}_T\mathbf{s}_1^* & -\mathbf{s}_T\mathbf{s}_2^* & \cdots & t - \mathbf{s}_T\mathbf{s}_T^* \end{bmatrix}, \tag{3.54}$$

where $t = \sum_{i=1}^{T} \|\mathbf{s}_i^*\|^2$.

As mentioned in the proof of Theorem 3.3.1, for a positive semidefinite matrix $\grave{\mathbf{A}}$, one can calculate $\grave{\mathbf{R}}$ recursively by starting with $i = 1$. For each i, $\grave{\mathbf{R}}_{i,i} = \sqrt{\grave{\mathbf{A}}_{i,i} - \sum_{k=1}^{i-1} \grave{\mathbf{R}}_{k,i}\grave{\mathbf{R}}_{k,i}^*}$, where $\grave{\mathbf{A}}_{i,i}$ is the i-th diagonal entry of $(\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N})$; moreover, for each $j > i$, $\grave{\mathbf{R}}_{i,j} = \frac{1}{\grave{\mathbf{R}}_{i,i}}(\grave{\mathbf{A}}_{i,j} - (\sum_{k=1}^{i-1} \grave{\mathbf{R}}_{k,i}\grave{\mathbf{R}}_{k,j}^*)^*)$, where $\grave{\mathbf{A}}_{i,j}$ is an entry of $(\rho_E I - \frac{E[\mathbf{X}^*\mathbf{X}]}{N})$ with row index i, and column index j.

Using the recursive relation for calculating the Cholesky decomposition, after some algebra, we have obtained

$$\begin{aligned} \grave{\mathbf{R}}_{i,i} &= \sqrt{t - ||\mathbf{s}_i^*||^2 - \sum_{j=1}^{i-1} \frac{||\mathbf{s}_j^*||^2||\mathbf{s}_i^*||^2 t}{(t - ||\mathbf{s}_{1:j-1}^*||^2)(t - ||\mathbf{s}_{1:j}^*||^2)}} \\ &= \sqrt{t - ||\mathbf{s}_i^*||^2 + \sum_{j=1}^{i-1}[\frac{||\mathbf{s}_{1:j-1}^*||^2||\mathbf{s}_i^*||^2}{t - ||\mathbf{s}_{1:j-1}^*||^2} - \frac{||\mathbf{s}_{1:j}^*||^2||\mathbf{s}_i^*||^2}{t - ||\mathbf{s}_{1:j}^*||^2}]} \\ &= \sqrt{t - ||\mathbf{s}_i^*||^2 - \frac{||\mathbf{s}_{1:i-1}^*||^2||\mathbf{s}_i^*||^2}{t - ||\mathbf{s}_{1:i-1}^*||^2}} = \sqrt{t(1 - \frac{||\mathbf{s}_i^*||^2}{||\mathbf{s}_{i:T}^*||^2})} \end{aligned} \tag{3.55}$$

If $i \neq T$, then $\frac{\|\mathbf{s}_i^*\|^2}{\|\mathbf{s}_{i:T}^*\|^2} < 1$ and thus $\mathbf{R}_{i,i} > 0$. However, when $i = T$,

$$\grave{\mathbf{R}}_{T,T} = \sqrt{t\left(1 - \frac{\|\mathbf{s}_T^*\|^2}{\|\mathbf{s}_{T:T}^*\|^2}\right)} = 0.$$

For any $\widetilde{\mathrm{s}}^*$ such that $\widetilde{\mathrm{s}}^* \neq \mathrm{s}^*$, let i be the largest integer such that $\mathrm{s}_i^* \neq \widetilde{\mathrm{s}}_i^*$, where $1 \leq i \leq T-1$. Then for any $j \leq i$,

$$\bar{M}_{\widetilde{\mathbf{s}}_{j:T}^*} \geq \frac{M_{\widetilde{\mathbf{s}}_{i:T}^*}}{\|\mathbf{s}_{j:T}^*\|^2 + |\mathbf{s}_{max}|^2(j-1)}.$$

We now try to give a lower bound on $\frac{M_{\widetilde{\mathbf{s}}_{i:T}^*}}{\|\mathbf{s}_{j:T}^*\|^2+|\mathbf{s}_{max}|^2(j-1)}$. We can find the following recursive expression for $\widetilde{\mathrm{s}}_{i:T}^*$ based on (3.12):

$$\begin{aligned} M_{\widetilde{\mathbf{s}}_{i:T}^*} &= |\sum_{k=i}^{T} \grave{\mathbf{R}}_{i,k}\widetilde{\mathbf{s}}_k|^2 + M_{\widetilde{\mathbf{s}}_{i+1:T}^*} \\ &= |\sum_{k=i+1}^{T} \grave{\mathbf{R}}_{i,k}\mathbf{s}_k + \grave{\mathbf{R}}_{i,i}\widetilde{\mathbf{s}}_i|^2, \end{aligned}$$

where $\widetilde{\mathrm{s}}_{i+1:T}^* = \mathrm{s}_{i+1:T}^*$, and $M_{\widetilde{\mathbf{s}}_{i+1:T}^*} = M_{\mathbf{s}_{i+1:T}^*} = 0$ as shown similarly in the proof of Theorem 3.3.1 (please also see (3.29)). Now we can write (3.11) as

$$\begin{aligned} M_{\widetilde{\mathbf{s}}_{i:T}^*} &= |\sum_{k=i}^{T} \grave{\mathbf{R}}_{i,k}\mathbf{s}_k - \grave{\mathbf{R}}_{i,i}\mathbf{s}_i + \grave{\mathbf{R}}_{i,i}\widetilde{\mathbf{s}}_i|^2 \\ &= |-\grave{\mathbf{R}}_{i,i}\mathbf{s}_i + \grave{\mathbf{R}}_{i,i}\widetilde{\mathbf{s}}_i|^2 = |\grave{\mathbf{R}}_{i,i}(\widetilde{\mathbf{s}}_i - \mathbf{s}_i)|^2, \end{aligned}$$

where we have used the fact that $\sum_{k=i}^{T} \grave{\mathbf{R}}_{i,k}\mathbf{s}_k = 0$, as shown similarly in the proof of Theorem 3.3.1 (please also see (3.29)). Since $\widetilde{\mathrm{s}}_i - \mathrm{s}_i \neq 0$ by assumption, and $\grave{\mathbf{R}}_{i,i} \neq 0$

for $i \neq T$ according to Lemma 3.8.1, $M_{\widetilde{\mathbf{s}}^*_{i:T}}$ will not be zero either.

When $\widetilde{\mathbf{s}}^* \neq \mathbf{s}^*$, $M_{\widetilde{\mathbf{s}}^*_{i:T}}$ is thus lower bounded by $|\grave{\mathbf{R}}_{i,i}(\widetilde{\mathbf{s}}_i - \mathbf{s}_i)|^2$, $i < T$. We thus first lower bound $\grave{\mathbf{R}}^2_{i,i} = t(1 - \frac{\|\mathbf{s}^*_i\|^2}{\|\mathbf{s}^*_{i:T}\|^2})$. The smallest possible value for t is $t = T|\mathbf{s}_{min}|^2$, where $|\mathbf{s}_{min}|^2$ is the minimum energy of any constellation point. Moreover, the largest possible value for $\frac{\|\mathbf{s}_i\|^2}{\|\mathbf{s}_{i:T}\|^2}$ is achieved only when $i = T-1$, $|\mathbf{s}_{T-1}|^2 = |\mathbf{s}_{max}|^2$, and $|\mathbf{s}_T|^2 = |\mathbf{s}_{min}|^2$. Thus $\grave{\mathbf{R}}^2_{i,i}$ is lower bounded by $T|\mathbf{s}_{min}|^2(1 - \frac{|\mathbf{s}_{max}|^2}{|\mathbf{s}_{max}|^2 + |\mathbf{s}_{min}|^2}) = \frac{T|\mathbf{s}_{min}|^4}{|\mathbf{s}_{max}|^2 + |\mathbf{s}_{min}|^2}$.

We further notice that the smallest possible value for $\|\widetilde{\mathbf{s}}^*_i - \mathbf{s}^*_i\|^2 = D_{min}$. And, the largest possible value for $\|\mathbf{s}^*_{j:T}\|^2 + |\mathbf{s}_{max}|^2(j-1)$ is $T|\mathbf{s}_{max}|^2$.

Combining the bounds on individual terms, under $\grave{\mathbf{R}}$, $\bar{M}_{\widetilde{\mathbf{s}}^*_{j:T}}$ is lower bounded by

$$\frac{D_{min}|\mathbf{s}_{min}|^4}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2|\mathbf{s}_{max}|^2}.$$

Using a similar argument of $\mathbf{R}$ concentrating around $\grave{\mathbf{R}}$ in the proof of Theorem 3.3.1, we can show that with high probability, under $\mathbf{R}$, $\bar{M}_{\widetilde{\mathbf{s}}^*_{j:T}}$ is no smaller than $\frac{D_{min}|\mathbf{s}_{min}|^4}{|\mathbf{s}_{max}|^4 + |\mathbf{s}_{min}|^2|\mathbf{s}_{max}|^2}$.

■

CHAPTER 4
EFFICIENT OPTIMAL JOINT CHANNEL ESTIMATION AND DATA DETECTION FOR MASSIVE MIMO SYSTEMS

4.1 Joint Channel Estimation and Signal Detection (JED) for Massive MIMO

We consider a TDD massive MIMO wireless system with N receive antennas at the base station, and $M \ll N$ user terminals each equipped with a single antenna. We assume a discrete-time block flat fading channel model where the channel coefficients are fixed for a coherence time T. Across different fading blocks, the channel coefficients take independent values from unknown distributions. We model the uplink transmission of this system within one channel block by

$$\mathbf{X} = \mathbf{H}\mathbf{S}^* + \mathbf{W}, \tag{4.1}$$

where $\mathbf{X} \in \mathcal{C}^{N\times T}$ is the received signal at the BS, $\mathbf{S}^*$ is an $M \times T$ matrix representing the transmitted signal, whose entries are independent and identically distributed (i.i.d.) symbols from a modulation constellation Ω (Ω can be of constant or non-constant modulus, such as 16-QAM), $\mathbf{W} \in \mathcal{C}^{N\times T}$ represents additive noises, and $\mathbf{H} \in \mathcal{C}^{N\times M}$ represents the unknown channel matrix. The elements of $\mathbf{W}$ are i.i.d. random variables following circularly symmetric complex Gaussian distribution $\mathcal{N}(0, \sigma_w^2)$. In each channel coherence block, we further assume that the channel coefficients are deterministic with no prior statistical information known about them [19], [22].

Since the channel coefficients take unknown deterministic values, we can formulate the GLRT-optimal joint channel estimation and data detection as a mixed optimization problem over $\mathbf{H}$ and $\mathbf{S}$:

$$\min_{\mathbf{H},\mathbf{S}^* \epsilon \Omega^{M\times T}} \|\mathbf{X} - \mathbf{H}\mathbf{S}^*\|^2, \tag{4.2}$$

where $\Omega^{M\times T}$ represents the signal lattice of dimension $M \times T$. We remark that the GLRT-optimal detection is equivalent to ML detection for SIMO systems with constant-modulus modulations, and for MIMO systems with equal-energy signaling, when the channel coefficients are known to take i.i.d. circularly symmetric complex Gaussian values [36].

We note that the combinatorial optimization problem in (4.2) is a least-squares problem in $\mathbf{H}$, while an integer least-squares problem in $\mathbf{S}^*$, since each element of $\mathbf{S}^*$ is chosen from a discrete constellation Ω [30]. Hereby, for any given $\mathbf{S}^*$, the channel matrix $\mathbf{H}$ that minimizes (4.2) is given by $\hat{\mathbf{H}} = \mathbf{X}(\mathbf{S}^*)^\dagger$, where $(\cdot)^\dagger$ and $(\cdot)^*$ denotes the Moore-Penrose pseudoinverse and conjugate transpose of a matrix respectively. Since $(\mathbf{S}^*)^\dagger = \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger$, $\hat{\mathbf{H}} = \mathbf{X}\mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger$. Substituting this into (4.2), we get

$$\begin{aligned}
\min_{\mathbf{H},\mathbf{S}^*} \|\mathbf{X}-\mathbf{H}\mathbf{S}^*\|^2 &= \min_{\mathbf{S}^* \epsilon \Omega^{M\times T}} \|\mathbf{X}(\mathbf{I} - \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*)\|^2 \\
&= \min_{\mathbf{S}^*} \mathrm{tr}(\mathbf{X}(\mathbf{I} - \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*)\mathbf{X}^*) \\
&= \mathrm{tr}(\mathbf{X}\mathbf{X}^*) - \max_{\mathbf{S}^* \epsilon \Omega^{M\times T}} \mathrm{tr}((\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*\mathbf{X}^*\mathbf{X}\mathbf{S}),
\end{aligned} \tag{4.3}$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix. To simplify the mathematical formulation, we define Ξ to be a new M-dimensional constellation, each element of which is an M-dimensional vector with its entries taking values from Ω. So the cardinality of Ξ is $|\Omega|^M$. Then we can rewrite (4.3) as

$$\mathrm{tr}(\mathbf{X}^*\mathbf{X}) - \max_{\mathbf{S}^* \epsilon \Xi^{1\times T}} \mathrm{tr}((\mathbf{S}^*\mathbf{S})^\dagger \mathbf{S}^*\mathbf{X}^*\mathbf{X}\mathbf{S}), \tag{4.4}$$

where we use $\mathrm{tr}(\mathbf{X}^*\mathbf{X}) = \mathrm{tr}(\mathbf{X}\mathbf{X}^*)$. Now by choosing ρ_{min} to be the minimum eigenvalue of $\mathbf{X}^*\mathbf{X}$, the minimization problem in (4.4) can be equivalently represented by the following optimization problem,

$$\mathrm{tr}(\mathbf{X}^*\mathbf{X} - \rho_{min} I) - \max_{\mathbf{S}^* \epsilon \Xi^{1\times T}} \mathrm{tr}((\mathbf{S}^*\mathbf{S})^\dagger \mathbf{S}^*(\mathbf{X}^*\mathbf{X} - \rho_{min} I)\mathbf{S}), \tag{4.5}$$

because $\mathrm{tr}((\mathbf{S}^*\mathbf{S})^\dagger \mathbf{S}^*(\rho_{min} I)\mathbf{S})$ is a constant. Since $A = \mathbf{X}^*\mathbf{X} - \rho_{min} I$ is positive semidefinite, we can factorize $A = \mathbf{R}^*\mathbf{R}$ using Cholesky decomposition, where $\mathbf{R}^*$ is the lower triangular matrix of Cholesky decomposition. Finally, using the trace property for product of matrices, (4.5) can be transformed as follows:

$$\begin{aligned}
&\mathrm{tr}(\mathbf{R}^*\mathbf{R}) - \max_{\mathbf{S}^* \epsilon \Xi^{1\times T}} \mathrm{tr}((\mathbf{S}^*\mathbf{S})^\dagger \mathbf{S}^*\mathbf{R}^*\mathbf{R}\mathbf{S}) \\
&= \min_{\mathbf{S}^* \epsilon \Xi^{1\times T}} \mathrm{tr}(\mathbf{R}(\mathbf{I} - \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger \mathbf{S}^*)\mathbf{R}^*) \\
&= \min_{\mathbf{S}^* \epsilon \Xi^{1\times T}} \|\mathbf{R}^* - \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger \mathbf{S}^*\mathbf{R}^*\|^2.
\end{aligned} \tag{4.6}$$

Thus our goal is to minimize (4.6). We remark that this approach of transform-

ing the GLRT-optimal detection to (4.6) is novel, very different from existing approaches for GLRT-optimal detection including the sphere decoder [30] which only works for SIMO wireless systems. We also note that, the channel estimate $\hat{\mathbf{H}} = \mathbf{X}(\mathbf{S}^*)^\dagger$ can be used for downlink precoding after solving (4.6).

4.2 Efficient GLRT-Optimal JED Algorithm

Finding the optimal solution to (4.6) is a formidable task, since it requires searching over all the $|\Omega|^{MT}$ hypotheses in the signal space. The exhaustive search approach provides the optimal solution, however, its complexity grows exponentially in the channel coherence time. In the special case of SIMO systems, the sphere decoder efficiently solves GLRT-optimal detection (in a different format from (4.6)) for both constant-modulus [30] and nonconstant-modulus constellations [63]. However, the sphere decoders from [30] and [63] do not work for MIMO.

To describe our algorithm, we first introduce a tree representation of the signal space. Recall that we use Ξ to represent the set of signal vectors of length M, where each element of each vector takes value from the constellation Ω. We can thus represent the set of possible matrices for $\mathbf{S}^*$ by a tree of T layers. At a layer 0, we have one root node. Each tree node at layer i, $0 \leq i \leq (T-1)$, has $|\Xi| = |\Omega|^M$ child nodes. We use $\mathbf{S}^*_{1:i}$ to represent the first i columns of $\mathbf{S}^*$, and each possible matrix for $\mathbf{S}^*_{1:i}$ corresponds to a layer-i tree node. And we call the tree nodes at layer T as leaf nodes, and thus each possible matrix for $\mathbf{S}^*$ is represented by a leaf node. Furthermore, for each possible matrix value for $\mathbf{S}^*$, we define its metric by

$$M_{\mathbf{S}^*} = \|\mathbf{R}^* - \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*\mathbf{R}^*\|^2. \tag{4.7}$$

For a partial matrix $\mathbf{S}^*_{1:i}$, we define its metric by

$$M_{\mathbf{S}^*_{1:i}} = \|\mathbf{R}^*_{1:i} - \mathbf{S}_{1:i}(\mathbf{S}^*_{1:i}\mathbf{S}_{1:i})^\dagger\mathbf{S}^*_{1:i}\mathbf{R}^*_{1:i}\|^2, \tag{4.8}$$

where $1 \leq i \leq T$, and $\mathbf{R}^*_{1:i}$ is the first i rows of $\mathbf{R}^*$. Thus solving (4.6) is equivalent to finding an $\hat{\mathbf{S}}^*$ that minimizes $M_{\mathbf{S}^*}$ among all the possible matrix values for $\mathbf{S}^*$. To develop our algorithm, we have the following lemma about the comparison between $M_{\mathbf{S}^*_{1:i}}$ and $M_{\mathbf{S}^*}$.

Lemma 4.2.1. *For every $i \leq T$ and any matrix value for $\mathbf{S}^*$, $M_{\mathbf{S}^*_{1:i}} \leq M_{\mathbf{S}^*}$*

Proof. We observe that $M_{\mathbf{S}^*}$ is the residual energy after projecting the columns of $\mathbf{R}^*$ onto the subspace spanned by the columns of $\mathbf{S}$; and $M_{\mathbf{S}^*_{1:i}}$ is the residual energy after projecting the columns of $\mathbf{R}^*_{1:i}$ (the first i rows of $\mathbf{R}^*$) onto the subspace spanned by the columns of $\mathbf{S}_{1:i}$ ($\mathbf{S}_{1:i}$ is the just the first i rows of $\mathbf{S}$). Since orthogonal linear projections minimize the residual energy among all linear projections, we can show, at the first i indices, the residual energy $M_{\mathbf{S}^*_{1:i}}$ after applying orthogonal projections $\mathbf{S}_{1:i}(\mathbf{S}^*_{1:i}\mathbf{S}_{1:i})^\dagger\mathbf{S}^*_{1:i}$ to $\mathbf{R}^*_{1:i}$, will be no bigger than these indices' residual energy (denoted by Q) after applying $\mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}$ to $\mathbf{R}^*$. Moreover, for the orthogonal projection $\mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}$ applied to $\mathbf{R}^*$, the total residual energy $M_{\mathbf{S}^*}$ over T indices is no smaller than the residual energy Q over the first i indices. Because $M_{\mathbf{S}^*} \geq Q$ and $Q \geq M_{\mathbf{S}^*_{1:i}}$, we have $M_{\mathbf{S}^*_{1:i}} \leq M_{\mathbf{S}^*}$. ∎

Algorithm 4: ML channel estimation and signal detection algorithm.

Input: radius r, matrix $\mathbf{R}$, constellation Ξ and a $1 \times T$ index vector I
Output: The transmitted signal $\mathbf{S}^*$

1. Set $i = 1$, $I(i) = 1$ and set $\mathbf{S}^*_{1:i} = \Xi(I(i))$.
2. (Computing the bounds) Compute the metric $M_{\mathbf{S}^*_{1:i}}$. If $M_{\mathbf{S}^*_{1:i}} > r^2$, go to 3; else, go to 4;
3. (Backtracking) Find the smallest $1 \leq j \leq i$ such that $I(j) < |\Xi|$. If there exists such j, set $i = j$ and go to 5; else go to 6.
4. If $i = T$, store current $\mathbf{S}^*$, update $r^2 = M_{\mathbf{S}^*_{1:T}}$ and go to 3; else set $i = i + 1$, $I(i) = 1$ and $\mathbf{S}^*_{1:i} = \Xi(I(i))$, go to 2.
5. Set $I(i) = I(i) + 1$ and $S^*_i = \Xi(I(i))$. Go to 2.
6. If any sequence $\mathbf{S}^*$ is ever found in Step 4, output the latest stored full-length sequence as the ML solution; otherwise, double r and go to 1.

Lemma 4.2.1 means that $M_{\mathbf{S}^*_{1:i}}$ is a lower bound on $M_{\mathbf{S}^*}$. Intuitively, suppose that $M_{\mathbf{S}^*_{1:i}}$ is too big, then $M_{\mathbf{S}^*}$ must also be big, and $\mathbf{S}^*$ will not minimize (4.6). This motivates us to propose the following branch-and-bound algorithm for GLRT-optimal JED. In this algorithm, we set a search radius r that uses to regulate a depth-first search over the signal tree structure for the optimal solution to (4.6). In fact, if $M_{\mathbf{S}^*_{1:i}} > r^2$, this algorithm will not search among the child nodes of $\mathbf{S}^*_{1:i}$. If the optimal solution is not found under the current radius r, we will increase the search radius r for new searches until the optimal solution is found.

Theorem 4.2.2. *Algorithm 4 gives the optimal solution to (4.6).*

This theorem is a result of Lemma 4.2.1 and the branch-and-bound search over the signal space. In order to simplify the complexity analysis, we further

modify step 6 of the ML algorithm: "If any sequence $\mathbf{s}^*$ is ever found in step 4, the output of the latest stored full-length sequence will be the ML solution; otherwise, let $r = \infty$ and go to step 1". We emphasize that this change will not effect the optimality of the algorithm since setting r to take a big number will make sure it will be bigger that the value of the metric of every possible transmitted sequence. Furthermore, the corresponding change in the step 6 simplifies the computing the complexity analyses since we will not deal will derive the complexity for possibly recursively events of doubling the initial radius r if it is less that the $M_{\mathbf{S}^*}$.

4.2.1 Metric Calculation and Initial Radius r

To compute $M_{\mathbf{S}^*_{1:i}}$, we can have a constant computational complexity independent of T, by recursive calculations over tree structure. The metric in (4.8) is equivalent to

$$M_{\mathbf{S}^*_{1:i}} = \mathrm{tr}(\mathbf{R}^*_{1:i}\mathbf{R}_{1:i}) - \mathrm{tr}((\mathbf{S}^*_{1:i}\mathbf{S}_{1:i})^\dagger \mathbf{S}^*_{1:i}\mathbf{R}^*_{1:i}\mathbf{R}_{1:i}\mathbf{S}_{1:i}). \tag{4.9}$$

From (4.9), we can calculate the metric $M_{\mathbf{S}^*_{1:i}}$ efficiently. First, the term $\mathrm{tr}(\mathbf{R}^*_{1:i}\mathbf{R}_{1:i})$, which represents the energy of matrix $\mathbf{R}_{1:i}$, can be rather precalculated in advance for each level i and hence used in the process of computing the metric in step 2 of the ML algorithm. Second, after defining a $T \times M$ matrix $A_i = \mathbf{R}_{1:i}\mathbf{S}_{1:i}$, we can update A_{i+1} sequentially as $A_{i+1} = A_i + \mathbf{R}_{i+1:i+1}\mathbf{S}_{i+1:i+1}$. Similarly, we can define $M \times M$ matrix $B_i = \mathbf{S}^*_{1:i}\mathbf{S}_{1:i}$ and then sequentially update $B_{i+1} = B_i + \mathbf{S}^*_{i+1:i+1}\mathbf{S}_{i+1:i+1}$. Furthermore, the complexity of calculating $B^\dagger_{i+1}$ is $O(M^2)$ using matrix inversion

lemma, where $B_{i+1}^{\dagger} = (B_i + \mathbf{S}_{i+1:i+1}^{*}\mathbf{S}_{i+1:i+1})^{\dagger}$. The complexity of all these recursive updates do not depend on T (noting that only i rows of A are nonzero).

For large N, we can choose the radius $r^2 = cN$, where c is any sufficiently small constant (please the next section for justifications). In fact, we also proposed best-first tree search algorithm to find the optimal solution while avoiding picking an r beforehand.

4.3 Expected Computational Complexity

The computational complexity of our tree search algorithms is mainly determined by the number of visited nodes in each layer. By "visited nodes", we mean the partial sequences $\mathbf{S}_{1:i}^{*}$ for which metric $M_{\mathbf{S}_i^*}$ is computed. The fewer the visited nodes, the lower computational complexity of our optimal channel estimation and data detection algorithm is. In this section, we show that the expected number of visited nodes will grow linearly with T under a sufficiently large number of receive antennas. To analyze the expected number of visited nodes, we assume that the channel coefficients are i.i.d. complex Gaussian random variables following distribution $\mathcal{CN}(0,1)$. We also assume that the M users send M orthogonal pilot sequences between time indices 1 and M.

Theorem 4.3.1. *Let M be fixed, and let $r^2 = cN$, where c is any sufficiently small positive constant. Then for the tree search algorithm, the expected number of visited points at layer i converges to $|\Xi| = |\Omega|^M$ for $i \geq (M+1)$, as the number of receive antennas N goes to infinity. The tree based search algorithm visits only one tree node at each layer $i < (M+1)$.*

Proof. (outline) We first prove that, the tree search algorithm only visits $|\Xi| = |\Omega|^M$ nodes per layer when $\mathbf{X}^*\mathbf{X} = E[\mathbf{X}^*\mathbf{X}]$, where the expectation is taken over the distribution of channel coefficients. Then we show that, when $N \to \infty$, $\mathbf{X}^*\mathbf{X}/N \to E[\mathbf{X}^*\mathbf{X}]/N$ in probability and that the expected number of visited nodes at layer i ($(M+1) \le i \le T$) approaches $|\Xi|$.

We first note that, the number of visited nodes at layer i ($(M+1) \le i \le T$) is equal to $|\Xi|$, if there is one and only one sequence $\widetilde{\mathbf{S}}^*_{1:(i-1)}$ such that $M_{\widetilde{\mathbf{S}}^*_{1:(i-1)}} \le r^2$. Let us consider the true transmitted sequence $\mathbf{S}^*$. Then we have

$$\begin{aligned} E[\mathbf{X}^*\mathbf{X}] &= E[(\mathbf{H}\mathbf{S}^* + \mathbf{W})^*(\mathbf{H}\mathbf{S}^* + \mathbf{W})] \\ &= \mathbf{S}E[\mathbf{H}^*\mathbf{H}]\mathbf{S}^* + E[\mathbf{W}^*\mathbf{W}] + \mathbf{S}E[\mathbf{H}^*\mathbf{W}] + E[\mathbf{W}^*\mathbf{H}]\mathbf{S}^* \\ &= N\mathbf{S}\mathbf{S}^* + N\sigma_w^2 I, \end{aligned} \tag{4.10}$$

where the second equality is from $E[\mathbf{H}\mathbf{H}^*] = NI$ and $E[\mathbf{H}^*\mathbf{W}] = 0$.

Because $\mathbf{S}\mathbf{S}^*$ is of rank M with $M < T$, from (4.10), the minimum eigenvalue of $E[\mathbf{X}^*\mathbf{X}]/N$ is σ_w^2. Then for the tree search algorithm (after scaling A by a constant N), $\mathbf{A} = E[\mathbf{X}^*\mathbf{X}]/N - \sigma_w^2 I = \mathbf{S}\mathbf{S}^*$. From the Cholesky decomposition, we know that $A = \mathbf{S}\mathbf{S}^* = \mathbf{R}^*\mathbf{R}$. This means that the columns of $\mathbf{R}^*$ span the same subspace as the columns of $\mathbf{S}$. Thus the metric $M_{\mathbf{S}^*} = 0$, because $\|\mathbf{R}^* - \mathbf{S}(\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*\mathbf{R}^*\|^2$ is precisely the residual energy after projecting the columns of $\mathbf{R}^*$ onto the subspace spanned by the columns of $\mathbf{S}$. We can think of $\mathbf{R}$ as a mapping of M dimensional space matrix $\mathbf{S}^*$ onto T dimensional space. Since $M_{\mathbf{S}^*_{1:i}} \le M_{\mathbf{S}^*}$, $M_{\mathbf{S}^*_{1:i}} = 0$ for all i.

$$\grave{\mathbf{R}}^* = \begin{bmatrix} \sqrt{M} & 0 & 0 & \cdots & 0 \\ \frac{(S_2 S_1^*)}{\sqrt{M}} & \sqrt{M - \frac{|(S_2 S_1^*)|^2}{M}} & 0 & \cdots & 0 \\ \frac{(S_3 S_1^*)}{\sqrt{M}} & \frac{1}{R_{2,2}}[(S_3 S_2^*) - \frac{(S_3 S_1^*)(S_2 S_1^*)^*}{M}] & \text{(4.12a)} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{(S_{M+1} S_1^*)}{\sqrt{M}} & \frac{1}{R_{2,2}}[(S_{M+1} S_2^*) - \frac{(S_{M+1} S_1^*)(S_2 S_1^*)^*}{M}] & \text{(4.13a)} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{(S_T S_1^*)}{\sqrt{M}} & \frac{1}{R_{2,2}}[(S_T S_2^*) - \frac{(S_T S_1^*)(S_2 S_1^*)^*}{M}] & \text{(4.13b)} & \cdots & 0 \end{bmatrix}. \tag{4.11}$$

From (4.10), each element in the $T \times T$ matrix $E[\mathbf{X}^*\mathbf{X}]/N$ will be $S_i S_j^* + \sigma_w = \sum_{k=1}^{T} s_{ik} s_{jk}^* + \sigma_w$, where S_i^* is the i-th column of the transmitted signal $\mathbf{S}^*$, and s_{ik} is the k-th element of the i-th row of $\mathbf{S}^*$. Thus we can express $\mathbf{A} = E[\mathbf{X}^*\mathbf{X}]/N - \sigma_w^2 I$ as

$$\mathbf{A} = \begin{bmatrix} M & S_1 S_2^* & \cdots & S_1 S_T^* \\ S_2 S_1^* & M & \cdots & S_2 S_T^* \\ \vdots & \vdots & \ddots & \vdots \\ S_T S_1^* & S_T S_2^* & \cdots & M \end{bmatrix}.$$

Now by using the Cholesky decomposition in [59], we can decompose $\mathbf{A} = \grave{\mathbf{R}}^* \grave{\mathbf{R}}$ where $\grave{\mathbf{R}}^*$ is the lower triangular matrix of Cholesky decomposition. Thus we can calculate the elements of matrix $\grave{\mathbf{R}}^*$ such as $R_{i,i}^* = \sqrt{a_{i,i} - \sum_{k=1}^{i-1} R_{i,k}^* R_{i,k}}$, $R_{i,j}^* = \frac{1}{R_{j,j}}(a_{i,j} - \sum_{k=1}^{j-1} R_{i,k}^* R_{j,k})$ for $1 \le j < i \le T$, and $a_{i,j}$ is an entry of $\mathbf{A}$ with row index i, and column index j.

$$\sqrt{M - \frac{|(S_3 S_1^*)|^2}{M} - \frac{|S_3 S_2^*|^2 + \frac{|S_3 S_1^*|^2 |S_2 S_1^*|^2}{M^2} - \frac{2Re\{(S_3 S_1^*)(S_2 S_1^*)(S_3 S_2^*)\}}{M}}{R_{2,2}}} \tag{4.12a}$$

$$\sqrt{M - \frac{|(S_{M+1}S_1^*)|^2}{M} - \frac{|S_{M+1}S_2^*|^2 + \frac{|S_{M+1}S_1^*|^2|S_2S_1^*|^2}{M^2} - \frac{2Re\{(S_{M+1}S_1^*)(S_2S_1^*)(S_{M+1}S_2^*)\}}{M}}{R_{2,2}}} \quad (4.13a)$$

$$\sqrt{M - \frac{|(S_TS_1^*)|^2}{M} - \frac{|S_TS_2^*|^2 + \frac{|S_TS_1^*|^2|S_2S_1^*|^2}{M^2} - \frac{2Re\{(S_TS_1^*)(S_2S_1^*)(S_TS_2^*)\}}{M}}{R_{2,2}}} \quad (4.13b)$$

Now we can use $\grave{\mathbf{R}}^*$ in (4.11) as the lower triangular matrix of Cholesky decomposition to solve the minimization equation in (4.6). In fact, based on (4.10), $\grave{\mathbf{R}}^*\grave{\mathbf{R}} = \mathbf{S}\mathbf{S}^*$, and hence the metric $M_{\mathbf{s}_{1:T}^*}(\grave{\mathbf{R}})$ from (4.7) is

$$\mathrm{tr}(\mathbf{S}\mathbf{S}^*) - \mathrm{tr}((\mathbf{S}^*\mathbf{S})^\dagger\mathbf{S}^*\mathbf{S}\mathbf{S}^*\mathbf{S})$$

$$= \mathrm{tr}(\mathbf{S}\mathbf{S}^*) - \mathrm{tr}(\mathbf{S}\mathbf{S}^*) = 0.$$

Let us instead consider any signal matrix $\bar{\mathbf{S}}$ such that $\bar{\mathbf{S}} \neq \mathbf{S}$ and $\bar{\mathbf{S}}_{1:M}^* = \mathbf{S}_{1:M}^*$ (namely $\bar{\mathbf{S}}$ shares the same pilot sequences as $\mathbf{S}$). For such $\bar{\mathbf{S}}$, we can show that $\|\mathbf{R}^* - \bar{\mathbf{S}}(\bar{\mathbf{S}}^*\bar{\mathbf{S}})^\dagger\bar{\mathbf{S}}^*\mathbf{R}^*\|^2 > 0$, and that $M_{\bar{\mathbf{S}}_{1:i}^*} > 0$ for the first i such that $\bar{\mathbf{S}}_{1:i}^* \neq \mathbf{S}_{1:i}^*$. In fact, $M_{\bar{\mathbf{S}}_{1:i}^*}$ is no smaller than

$$D = \min_{i>M, \mathbf{S}, \bar{\mathbf{S}}, \mathbf{S}_{1:i} \neq \bar{\mathbf{S}}_{1:i}} \|\mathbf{S}_{1:i}^* - \bar{\mathbf{S}}_{1:i}(\bar{\mathbf{S}}_{1:i}^*\bar{\mathbf{S}}_{1:i})^\dagger\bar{\mathbf{S}}_{1:i}^*\mathbf{S}_{1:i}^*\|^2 > 0.$$

Thus for a search radius $r^2 < D$, there will be only T tree nodes (namely those from transmitted signal $\mathbf{S}^*$) with metric no bigger than r^2. This means that the tree search algorithm will visit at most $T|\Xi|$ tree nodes, under the assumption that $\mathbf{X}^*\mathbf{X} = E[\mathbf{X}^*\mathbf{X}]$.

For massive MIMO systems, when $N \to \infty$, $\mathbf{X}^*\mathbf{X}/N \to E[\mathbf{X}^*\mathbf{X}]/N$ in probability. In fact, we can show that, as $N \to \infty$, with probability at least $(1 - \epsilon)$, the tree search algorithm will visit at most $T * |\Xi|$ tree nodes, where $\epsilon > 0$ is an arbitrary small number. With probability $\epsilon > 0$, the tree search algorithm will visit at most $|\Xi|^T$ nodes. When $N \to \infty$, ϵ can be pushed small fast enough such that the expected number of visited tree nodes grows linear in T. ∎

Moreover, we only need N to grow polynomially in T, in order to guarantee that the expected number of visited tree nodes grows polynomially in T. In fact, using large deviation bounds for the convergence of $\mathbf{X}^*\mathbf{X}/N$ to $E[\mathbf{X}^*\mathbf{X}]/N$, we have the following theorem.

Theorem 4.3.2. *Let M be fixed, and let $r^2 = cN$, where c is any sufficiently small positive constant. Then we only need the number of receive antennas N to grow polynomially in T, to guarantee that the expected number of visited points at layer i converges to $|\Xi|$ for $i \geq (M + 1)$.*

4.4 Simulation Results

In this section, we numerically simulate the performance of our new GLRT-optimal tree search algorithm, comparing it against suboptimal iterative and non-iterative MMSE channel estimation and data detection schemes. We allow the receiver to know the first M columns of the transmitted signal $\mathbf{S}^*$, which serve as necessary orthogonal pilot sequences to guarantee good error performance. The non-iterative MMSE channel estimation scheme first uses the disclosed pilot se-

quences to perform MMSE channel estimation. This estimated channel will consider fixed for the rest of coherent block length, and it will be used to detect the transmitted information symbols through applying MMSE signal detection. The iterative MMSE scheme iteratively exploits the detected data from the previous iteration to perform channel estimation used for data detection in the current iteration.

Throughout this section we consider different numbers of users M, 2 and 4, and different values for the number of receive antennas, namely $N = 50, 100, 200$, and 400. Although $N = 50$ may not be located under massive scale of receiving antennas definition, considering it shows that our algorithm still efficient for conventional MIMO systems. For constant modulus constellation, we normalize the signal energy such that $\|\mathbf{S}_{i,j}\|^2 = 1$; furthermore, we define the signal to noise ration (SNR) as SNR = $\frac{E\|\mathbf{HS}^*\|^2}{E\|\mathbf{W}\|^2}$.

In Figure 4.1, we demonstrate the symbol error rate (SER) performance, as a function of SNR, for 16-QAM modulation, non-constant modulus constellation. The gain difference of our optimal noncoherent detection MIMO algorithm is calculated compared with both non optimal iterative and non-iterative MMSE schemes for $M = 2$, and $T = 8$. For 10^2 SER and $N = 50$, iterative MMSE channel estimation scheme exhibits 5 db SNR loss in comparison with our tree search algorithm. When $N = 100$, our method holds 6 dB gain over the iterative MMSE scheme at 10^{-3} SER. Most importantly, our tree search algorithm guarantees providing the GLRT-optimal solution.

In Figure 4.2 we evaluate the average number of visited nodes of the tree search algorithm per each coherent block for different SNR values. Here $T = 8$, $M = 2$, and the modulation scheme is 16-QAM . We observe a tremendous reduction in the number of hypotheses that need to be tested, compared with exhaustive search method. For instance, for $N = 100$ and SNR = 3 dB, exhaustive search requires testing 2.81×10^{14} hypotheses in each coherence block, while our tree search algorithm visits only 5.5×10^4 tree nodes on average. For $N = 500$ and $SNR = 6$ dB, our tree serch algorithm visits $(T - M) \times |\Xi| \simeq 1500$ nodes, which verifies our result in Theorem 4.3.1. We remark that, using the same computer for simulation, exhaustive search would need 3.52×10^3 years to compute the optimal solution for one channel coherence block.

Figure 4.3 shows the performance of tree search algorithm for an extended number of receive antennas 200 for 16-QAM constellation. GLRT-optimal non-coherent detection algorithm provides 6 dB gain compared with iterative MMSE detection scheme. We can see that the optimal algorithm achieves higher gain difference for higher N compared with suboptimal schemes even for low SNR.

We plot the average number of visited points as a function of SNR in Figure 4.4, for QPSK modulation, $M = 4$, and $T = 10$. We observe that increasing N form 50 to 500 will greatly reduce the number of visited nodes. Exhaustive search would need to examine 2.81×10^{14} hypotheses and will take 2000 years to calculate the optimal solution for one channel coherence block.

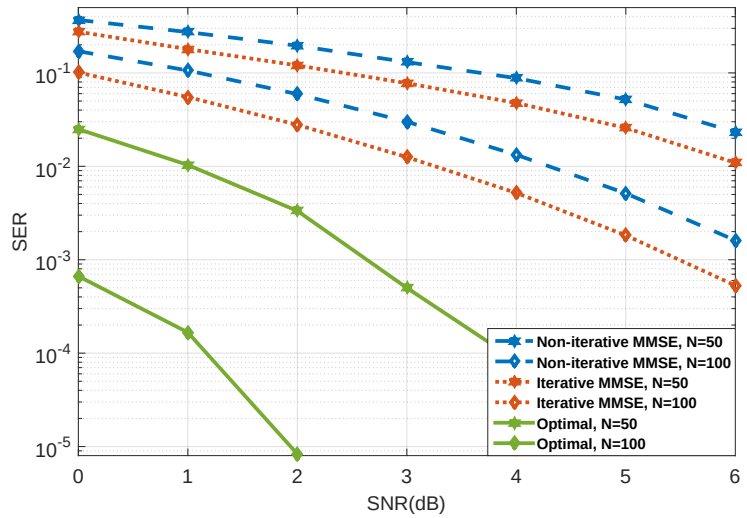


Figure 4.1: SER for iterative MMSE, non-iterative MMSE, and our optimal tree search algorithm. $M = 2$, $T = 8$, and 16-QAM constellation.

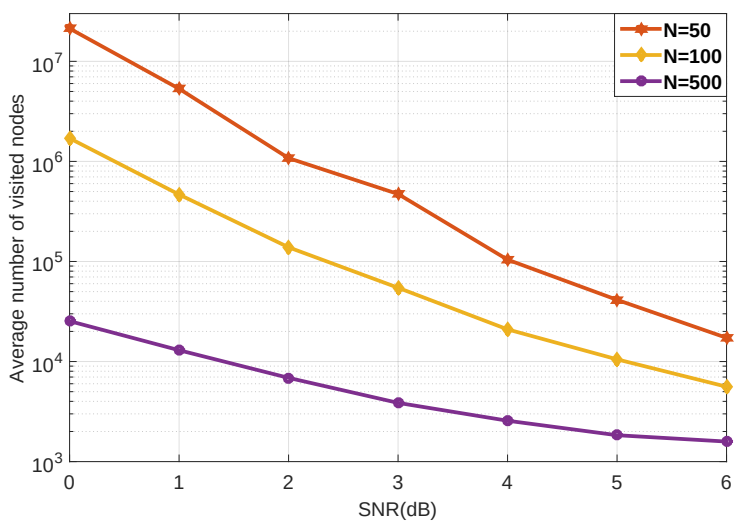


Figure 4.2: Average number of visited points for $T = 8$, and 16-QAM modulation. Exhaustive search will instead need to test 2.8147×10^{14} hypotheses.

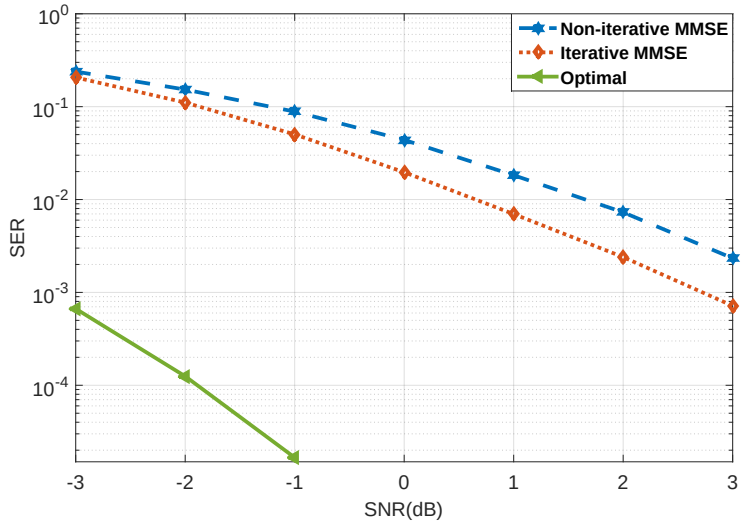


Figure 4.3: SER for iterative MMSE, non-iterative MMSE, and our optimal tree search algorithm. $M = 2$, $T = 8$, $N = 200$, and 16-QAM modulation.

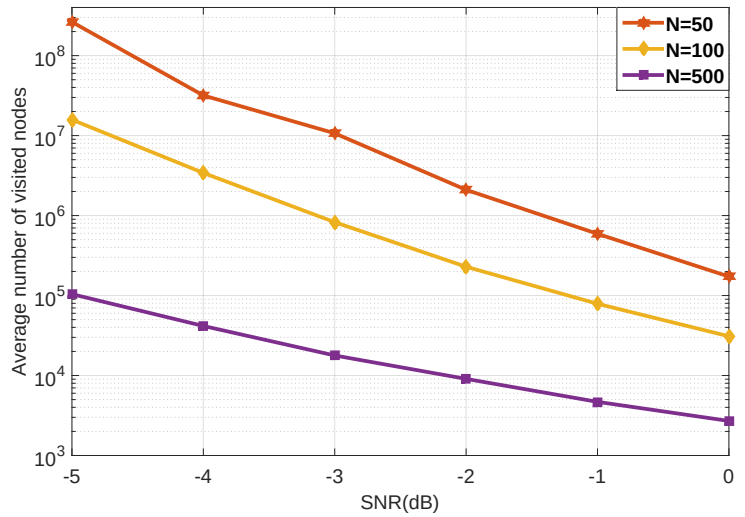


Figure 4.4: Average number of visited points for $T = 10$, $M = 4$, and QPSK modulation. Exhaustive search will instead need to test 2.8147×10^{14} hypotheses.

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

In this thesis, we have presented novel optimal approaches to solve the problem of joint channel estimation and data detection (JED) for large scale MIMO systems. Massive MIMO is a new promising technology for next-generation wireless communication. When the number of the antennas at each base station of the wireless network increases vastly, multiple users can be served simultaneously. In fact, massive MIMO promises to fulfill the requirements of future applications which need high-quality streaming, such as virtual reality and online games applications.

In Chapter two we proposed a new approach to solve the integer least-squares problem using optimized Markov Chain Monte Carlo method. The proposed MCMC method will, unlike simulated annealing techniques, have a fixed temperature parameter in all the iterations, with the property that after the Markov chain has mixed, the probability of encountering the optimal solution is only polynomial small (i.e. not exponentially small). The optimal value of temperature parameter for this approach has been calculated. To the authors' knowledge, this is a novel way of looking at solving the integer least-squares problem. From our analysis, we can see there is a trade-off between faster mixing of the Markov chain and faster finding the optimal solution in the stationary distribution. We also analyzed the mixing time of the underlying Markov chain. It is also interesting to extend this work to MIMO system with channel codes in future works.

Existing GLRT-optimal joint channel estimation and signal detection (JED) algorithms are limited to constant-modulus modulations or to SISO systems. More importantly, none of these algorithms achieve the GLRT-optimal JED for massive SIMO systems with general constellations while having a computational complexity polynomial in T. In Chapter three, we overcome these limitations and advance the state of the art for the GLRT-optimal JED for SIMO wireless systems.

We have proposed efficient GLRT-optimal JED algorithms for SIMO systems, including massive SIMO systems. Our algorithms apply to general constellations, including nonconstant-modulus constellations. To the best of our knowledge, our algorithms are the first set of efficient GLRT-optimal JED algorithms for massive SIMO systems using general constellations. We are thus able to provide the first set of error rate curves of the GLRT-optimal JED for massive SIMO wireless systems with general constellations. Our algorithms include a new breath-first tree-search algorithm which can find the GLRT-optimal JED solution without requiring any predetermined search radius.

Theoretically, we show that, under a large number of receive antennas in massive SIMO systems, the computational complexity of our SIMO GLRT-optimal algorithm will have an expected computational complexity polynomial both in the channel coherence time T and in the number of receive antennas N. This is somewhat surprising, since we have a large number of unknown complex channel coefficients to estimate, as the number of receive antenna increases in massive SIMO systems. More importantly, we show that the polynomial growth of com-

putational complexity is true, as long as **the number of receive antennas grows polynomially in** T. This is also in great contrast to the exponential growth of the sphere decoder's complexity for coherent MIMO systems.

As a consequence of this work, we demonstrate the exact performance gap between the GLRT-optimal and suboptimal JED algorithms for massive SIMO systems. In fact, we show significant performance gains of our optimal JED algorithms for SIMO systems using nonconstant-modulus constellations.

We show in Chapter Four, for the first time, the performance of joint GLRT-optimal JED algorithm of massive MIMO wireless systems for general constellations. We have shown that, as the number of receive antennas grows large, the expected complexity of our proposed algorithm is polynomial in the channel coherence time. Simulation results show that the GLRT-optimal algorithm has supreme performance than suboptimal non-coherent data detection schemes.

Built on our previous work, in this section we will provide some future research directions.

- Extending the MIMO detection algorithm to MU-MIMO systems.

  So far we have discussed non-coherent signal detection for large scale MIMO systems without considering the effect of neighboring cells on the algorithm complexity and detection process. Considering a network with L cells and K one antenna user terminals (UTs). Each cell deployed with one BS with N antennas. During UL phase the channel model for MU-MIMO system can be represented as a linear combination of channel matrices from all the cells,

then the system model can be represented as:

$$\mathbf{y}_j = \sum_{l=1}^{L} \mathbf{H}_{jl}\mathbf{x}_l + \mathbf{w}_j, \tag{5.1}$$

where $\mathbf{y}_j$ is the received vector at the j-th base station, $\mathbf{H}_{jl} = [\mathbf{h}_{jl1}\ \mathbf{h}_{jl2} \cdots \mathbf{h}_{jlK}]$ is the $N \times K$ channel matrix between the users of the l-th cell and the j-th BS, $\mathbf{x}_l$ is the transmitted vector from the users in cell l, and $\mathbf{w}_j$ is a noise vector. $\mathbf{h}_{jl1}$ is usually modeled based on a deterministic correlation matrix and independent of fast-fading channel vector. The challenge here is the interference term due to the transmitted signals from the all the active users in network cells to the j-th BS,

$$\mathbf{y}_j = \underbrace{\sum_{l=1, l\neq j}^{L} \mathbf{H}_{jl}x_l}_{Interference} + \mathbf{H}_{jj}x_j + \mathbf{n}_j.$$

One solution is by treating each cell separately, considering the interference part, which comes from neighbor cells, as a part of the noise. In this way we modify the problem into subcells channel estimation and signal detection process. However, we need to use a representative channel model for MU-MIMO systems. In other words, the channel model involves propagation effect, and fast fading channel [11], [16].

- Designing optimal non-coherent detection algorithms for OSTBC systems.

  We propose to design *Exact* GLRT JED for orthogonal space time block coded (OSTBC) multiple input multiple output (MIMO) systems. An optimal blind

data detection algorithm for general modulus constellations can be achieved. OSTBC, including the Alamouti's scheme [64], is one attractive MIMO communication scheme because of its full diversity gain and high data rates.

We can choose $M = 2$ to match Alamouti's OSTBC scheme with two transmit antennas, and for the sake of simplicity of presentation. We remark, however, that this results can be extended to general M. For the Alamouti's scheme, the transmitted data symbols $\mathbf{S}^*$ is given by

$$\mathbf{S}^* = \underbrace{\begin{bmatrix} s_{1,1} & -s_{2,1}^* & s_{1,2} & -s_{2,2}^* & \cdot & \cdot & \cdot & \cdot \\ s_{2,1} & s_{1,1}^* & s_{2,2} & s_{1,2}^* & \cdot & \cdot & \cdot & \cdot \end{bmatrix}}_{T}$$

where $s_{i,j}$ ($1 \leq i \leq 2$ and $1 \leq j \leq T$) is the i-th information symbol transmitted in the j-th space time block. The information symbols $s_{i,j}$'s are i.i.d. chosen from a certain constellation Ω (such as BPSK and 16-QAM). We use $\mathbf{S}_j^*$ to denote the j-th block matrix If we let Ξ denote the set of legitimate forms for $\mathbf{S}_j^*$ in OSTBC with cardinality $|\Xi| = |\Omega|^2$. We denote them by $\Xi(1), \Xi(2), ..., \Xi(|\Xi|)$. Further assumes that $\mathbf{H}$ is deterministic unknown at the receiver. Then we can solve the objective mixed optimization problem formulated below:

$$\min_{\mathbf{H}, \mathbf{S}^* \epsilon \Xi^T} \|\mathbf{X} - \mathbf{H}\mathbf{S}^*\|^2. \tag{5.2}$$

We can use our GLRT optimal algorithm that introduced in Chapter 4 to solve

the problem efficiently.

- Designing an exact non-coherent signal algorithm for distributed MIMO (DMIMO) systems Due to the installation structure and/or environment limitations, the BS antennas could be distributed over a geographic area. One direction to seek is the extension of our optimal non-coherent detection in a distributed circumstances. Correlated and uncorrelated channel matrix can be addressed in DMIMO.

- MCMC for non-coherent signal detection.

  We study and analyze the detection of MIMO systems using MCMC approach. However, we assumed the matrix characteristics known at the receiver. An extension would be using MCMC for non-coherent signal detection. We can start initially with SIMO case using our proposed approach in Chapter (3) to change the joint minimization over the system channel $\mathbf{h}$ and transmitted signal $\mathbf{s}$, into an equivalent minimization over $\mathbf{s}$ using $\mathbf{R}\mathbf{s}$ matrix. Equivalently, we can replace $\mathbf{s}$ in the reversible MCMC detector algorithm instead of the main minimization equation over the transmitted signal $\mathbf{S}$.

$$p\left(\hat{\mathbf{s}}_j^{(l+1)} = \omega \,|\theta\right) = \frac{e^{-\frac{1}{2\alpha^2}\left\|R\hat{\mathbf{s}}_{j|\omega}\right\|^2}}{\sum\limits_{\hat{\mathbf{s}}_{j|\tilde{\omega}} \in \Xi} e^{-\frac{1}{2\alpha^2}\left\|R\hat{\mathbf{s}}_{j|\tilde{\omega}}\right\|^2}} \,, \tag{5.3}$$

REFERENCES

[1] Cisco. VNI. Visual networking index: Global mobile data traffic forecast update, 2015 2020 white paper, 2016.

[2] National Telecommunications and Information Administration. United states frequency allocation chart, 2016.

[3] M. Latouche, C. Rauschen, O. Oszabó, J. Creusat, and Belmans W. Mobile data explosion: How mobile service providers can monetize the growth in mobile data through value-added services. *Cisco Internet Business Solutions Group (IBSG)*, May 2013.

[4] M. El-Sayed, A. Mukhopadhyay, C. Urrutia-Valdés, and Z. J. Zhao. Mobile data explosion: Monetizing the opportunity through dynamic policies and qos pipes. *Bell Labs Technical Journal*, 16(2):79–99, Sept 2011.

[5] F. Rusek, D. Persson, Buon Kiong Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson. Scaling up MIMO: Opportunities and challenges with very large arrays. *IEEE Signal Processing Magazine*, 30(1):40–60, 2013.

[6] Gerard J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal*, 1(2):41–59, Autumn 1996.

[7] I. Telatar. Capacity of multi-antenna gaussian channels. *European Trans. Telecomm*, 10:585–595, Des 1999.

[8] B.M. Hochwald and T.L. Marzetta. Unitary space-time modulation for multiple-antenna communications in rayleigh flat fading. *Information Theory, IEEE Transactions on*, 46(2):543–564, Mar 2000.

[9] T.L. Marzetta and B.M. Hochwald. Capacity of a mobile multiple-antenna communication link in rayleigh flat fading. *Information Theory, IEEE Transactions on*, 45(1):139–157, Jan 1999.

[10] Lizhong Zheng and D.N.C. Tse. Communication on the grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel. *Information Theory, IEEE Transactions on*, 48(2):359–383, Feb 2002.

[11] T.L. Marzetta. Noncooperative cellular wireless with unlimited numbers

of base station antennas. *IEEE Transactions on Wireless Communications*, 9(11):3590–3600, November 2010.

[12] F. Rusek, D. Persson, Buon Kiong Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson. Scaling up MIMO: Opportunities and challenges with very large arrays. *IEEE Signal Processing Magazine*, 30(1):40–60, Jan 2013.

[13] J. Hoydis, S. ten Brink, and M. Debbah. Massive MIMO in the UL/DL of cellular networks: How many antennas do we need? *IEEE Journal on Selected Areas in Communications*, 31(2):160–171, February 2013.

[14] Lu Lu, G.Y. Li, A.L. Swindlehurst, A. Ashikhmin, and Rui Zhang. An overview of massive MIMO: Benefits and challenges. *IEEE Journal of Selected Topics in Signal Processing*, 8(5):742–758, Oct 2014.

[15] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, February 2014.

[16] J. Hoydis, S. ten Brink, and M. Debbah. Massive MIMO: How many antennas do we need? In *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton) 2011*, pages 545–550, Sept 2011.

[17] M. Karlsson and E. G. Larsson. Massive mimo as a cyber-weapon. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 661–665, Nov 2014.

[18] B. Hassibi and B. M. Hochwald. How much training is needed in multiple-antenna wireless links? *IEEE Transactions on Information Theory*, 49(4):951–963, April 2003.

[19] P. Stoica and G. Ganesan. Space-time block codes: Trained, blind and semi-blind detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002*, volume 2, pages II–1609–II–1612, May 2002.

[20] A.L. Swindlehurst and G. Leus. Blind and semi-blind equalization for generalized space-time block codes. *IEEE Transactions on Signal Processing*, 50(10):2489–2498, Oct 2002.

[21] S. Shahbazpanahi, A. B. Gershman, and J. H. Manton. Closed-form blind MIMO channel estimation for orthogonal space-time block codes. *IEEE Transactions on Signal Processing*, 53(12):4506–4517, Dec 2005.

[22] W. K. Ma, B. N. Vo, T. N. Davidson, and P. C. Ching. Blind ML detection of orthogonal space-time block codes: Efficient high-performance implementations. *IEEE Transactions on Signal Processing*, 54(2):738–751, Feb 2006.

[23] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, Oct 1948.

[24] T. Datta, N. Srinidhi, A. Chockalingam, and B.S. Rajan. Random-restart reactive tabu search algorithm for detection in large-MIMO systems. *IEEE Communications Letters*, 14(12):1107–1109, 2010.

[25] K. Vardhan, S.K. Mohammed, A. Chockalingam, and B.S. Rajan. A low-complexity detector for large MIMO systems and multicarrier CDMA systems. *IEEE Journal of Selected Areas in Communications*, 26(3):473–485, April 2008.

[26] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger. Closest point search in lattices. *IEEE Transactions on Information Theory*, 48(8):2201–2214, 2002.

[27] M. O. Damen, H. E. Gamal, and G. Caire. On maximum-likelihood detection and the search for the closest lattice point. *IEEE Trans. on Info. Theory*, 49:2389–2402, Oct. 2003.

[28] B. M. Hochwald and S. T. Brink. Achieving near-capacity on a multiple-antenna channel. *IEEE Trans. on Commun.*, 51(3):389–399, 2003.

[29] P. Stoica, H. Vikalo, and B. Hassibi. Joint maximum-likelihood channel estimation and signal detection for SIMO channels. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2003*, volume 4, pages IV–13–16 vol.4, April 2003.

[30] H. Vikalo, B. Hassibi, and Petre Stoica. Efficient joint maximum-likelihood channel estimation and signal detection. *IEEE Transactions on Wireless Communications*, 5(7):1838–1845, July 2006.

[31] B. Hassibi and H. Vikalo. On the sphere-decoding algorithm. I. Expected complexity. *IEEE Trans. on Sig. Proc.*, 53:2806–2818, Aug. 2005.

[32] B. Hassibi and H. Vikalo. On the sphere-decoding Algorithm. II. Generalizations, second-Order statistics, and applications to communications. *IEEE Trans. on Sig. Proc.*, 53:2819–2834, Aug. 2005.

[33] J. Jaldén and B. Ottersten. On the complexity of sphere decoding in digital communications. *IEEE Trans. on Sig. Proc.*, 53:1474–1484, Apr. 2005.

[34] M. Stojnic, H. Vikalo, and B. Hassibi. A branch and bound approach to speed up the sphere decoder. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*, volume 3, pages iii/429–iii/432 Vol. 3, March 2005.

[35] M. Stojnic and B. Hassibi. Out-sphere decoder for non-coherent ML SIMO detection and its expected complexity. *Proceedings of the Forty-First Asilomar Conference on Signals, Systems and Computers*, pages 1568–1572, 2007.

[36] D. Warrier and U. Madhow. Spectrally efficient noncoherent communication. *IEEE Transactions on Information Theory*, 48(3):651–668, Mar 2002.

[37] I. Motedayen-Aval and A. Anastasopoulos. Polynomial-complexity noncoherent symbol-by-symbol detection with application to adaptive iterative decoding of turbo-like codes. *IEEE Transactions on Communications*, 51(2):197–207, Feb 2003.

[38] D.J. Ryan, I.B. Collings, and I.V.L. Clarkson. GLRT-optimal noncoherent lattice decoding. *IEEE Transactions on Signal Processing*, 55(7):3773–3786, July 2007.

[39] R.G. McKilliam, D.J. Ryan, I.V.L. Clarkson, and I.B. Collings. An improved algorithm for optimal noncoherent QAM detection. In *Communications Theory Workshop, 2008. AusCTW 2008. Australian*, pages 64–68, Jan 2008.

[40] K. M. Mackenthun. A fast algorithm for multiple-symbol differential detection of mpsk. *IEEE Transactions on Communications*, 42(234):1471–1474, Feb 1994.

[41] I. Motedayen-Aval, A. Krishnamoorthy, and A. Anastasopoulos. Optimal joint detection/estimation in fading channels with polynomial complexity. *IEEE Transactions on Information Theory*, 53(1):209–223, Jan 2007.

[42] V. Pauli, L. Lampe, R. Schober, and K. Fukuda. Multiple-symbol differential detection based on combinatorial geometry. *IEEE Transactions on Communications*, 56(10):1596–1600, October 2008.

[43] D. S. Papailiopoulos, G. A. Elkheir, and G. N. Karystinos. Maximum-likelihood noncoherent PAM detection. *Communications, IEEE Transactions on*, 61(3):1152–1159, March 2013.

[44] D. Makrakis and P. T. Mathiopoulos. Optimal decoding in fading channels: a combined envelope, multiple differential and coherent detection approach. In *Global Telecommunications Conference and Exhibition 'Communications Technology*

for the 1990s and Beyond' (GLOBECOM), 1989. IEEE, pages 1551–1557 vol.3, Nov 1989.

[45] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath. Pilot contamination and precoding in multi-cell TDD systems. *IEEE Transactions on Wireless Communications*, 10(8):2640–2651, August 2011.

[46] J. Jose, A. Ashikhmin, T.L. Marzetta, and S. Vishwanath. Pilot contamination problem in multi-cell TDD systems. In *IEEE International Symposium on Information Theory, 2009. ISIT 2009*, pages 2184–2188, June 2009.

[47] X. Wang and V. H. Poor. *Wireless Communications Systems: Advanced Techniques for Signal Reception*. Prentice Hall, 2003.

[48] H. Zhu, B. Farhang-Boroujeny, and R.R. Chen. On performance of sphere decoding and Markov chain Monte Carlo detection methods. *IEEE Sig. Proc. Letters*, 12:669–672, 2005.

[49] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2 edition, 2004.

[50] O. Häggström. *Finite Markov chains and algorithmic applications*. Cambridge University Press, 2002.

[51] David Levin, Y. Peres, and E. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[52] Mazen Al Borno. Reduction in Solving Some Integer Least Squares Problems. *Thesis, McGill University*, 2011.

[53] D.J.C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[54] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[55] M. Jerrum and A. Sinclair. Approximating the Permanent. *SIAM Journal on Computing*, 18:1149–1178, 1989.

[56] G. Lawler and A. Sokal. Bounds on the L^2 spectrum for Markov Chains and Markov Processes: a Generalization of Cheeger's Inequality. *Trans. Amer. Math. Soc.*, 309:557–580, 1988.

[57] Hoi-To Wai, Wing-Kin Ma, and A.M.-C. So. Cheap semidefinite relaxation mimo detection using row-by-row block coordinate descent. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3256–3259, May 2011.

[58] Weiyu Xu, M. Stojnic, and B. Hassibi. Low-complexity blind maximum-likelihood detection for SIMO systems with general constellations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008.*, pages 2817–2820, March 2008.

[59] R. Sacco A. Quarteroni and F. Saleri. Low-complexity blind equalization for ofdm systems with general constellations. *Numerical Mathematics :Springer-Verlag*, 2000.

[60] Zlatko Drmac, Matjaz Omladic, and Kresimir Veselic. On the perturbation of the cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 15(4):1319–1332, October 1994.

[61] X.-W. Chang, C.C. Paige, and G.W. Stewart. New perturbation analyses for the cholesky factorization. *IMA J. Numer. Anal.*, 16:457–484.

[62] M. Biguesh and A. B. Gershman. Training-based MIMO channel estimation: a study of estimator tradeoffs and optimal training signals. *IEEE Transactions on Signal Processing*, 54(3):884–893, March 2006.

[63] H. A. J. Alshamary, T. Al-Naffouri, A. Zaib, and W. Xu. Optimal non-coherent data detection for massive simo wireless systems: A polynomial complexity solution. In *Signal Processing and Signal Processing Education Workshop (SP/SPE), 2015 IEEE*, pages 172–177, Aug 2015.

[64] S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451–1458, Oct 1998.