Optimization Algorithms in Wireless and Quantum Communications

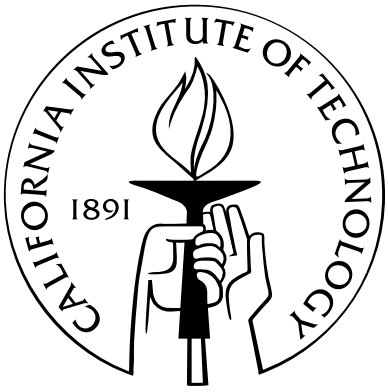
Thesis by

Mihailo Stojnic

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2008

(Defended 28 November 2007)

Acknowledgements

I have spent five beautiful years at Caltech. The journey has been amazing, in large part due to many phenomenal people who were around me over these years, and who I would like to acknowledge.

First I would like to thank my advisor Prof. Babak Hassibi. His scientific brilliance, quick mind, and incredible kindness are absolutely unmatched. As a PhD student you can only hope for an advisor like him.

I would also like to thank Prof. Robert McEliece, Prof. P.P. Vaidyanathan, Prof. Jehoshua Bruck, Prof. Steven Low, and Prof. Tracey Ho for serving on my candidacy and thesis defense committees, and Mrs. Shirley Beatty for her administrative assistance.

My deep gratitude goes to a great set of people who I spent a lot of time with either as group or office mates in Moore 155: Amir F. Dana, Masoud Sharif, Radhika Gowaikar, Chaitanya Rao, Vijay Gupta, Yindi Jing, Ali Vakili, Weiyu Xu, Sormeh Shadbakht, Frederique Oggier, Haris Vikalo, Jeremy Thorpe, Mostafa El-Khamy, Ravi Palanki, Cedric Florens, and Farzad Parvaresh.

At the end I would like to thank my sister for her support and my parents for teaching me the earliest and the most important lessons in life. My personal achievements will always be overshadowed by their belief in me, hope for my best, and endless love.

Abstract

Since the first communication systems were developed, the scientific community has been witnessing attempts to increase the amount of information that can be transmitted. In the last 10–15 years there has been a tremendous amount of research towards developing multi-antenna systems which would hopefully provide high-data-rate transmissions. However, increasing the overall amount of transmitted information increases the complexity of the necessary signal processing. A large portion of this thesis deals with several important issues in signal processing of multi-antenna systems. In almost every particular case the goal is to develop a technique/algorithm so that the overall complexity of the signal processing is significantly decreased.

In the first part of the thesis a very important problem of signal detection in MIMO (multiple-input multiple-output) systems is considered. The problem is analyzed in two different scenarios: when the transmission medium (channel) 1) is known and 2) is unknown at the receiver. The former case is often called coherent and the later non-coherent MIMO detection. Both cases usually amount to solving highly complex NP-hard combinatorial optimization problems.

For the coherent case we develop a significant improvement of the traditional sphere decoder algorithm commonly used for this type of detection. An interesting connection between the new improved algorithm and the H-infinity estimation theory is established, and the performance improvement over the standard sphere decoder is demonstrated. For the non-coherent case we develop a counterpart to the standard sphere decoder, the so-called out-sphere decoder. The complexity of the algorithm is viewed as a random variable; its expected value is analyzed and shown to be significantly smaller than the one of the overall exhaustive search. In the non-coherent case, in addition to the complexity analysis of the exact out-sphere decoder, we analyze the performance loss of a suboptimal technique. We show that only a moderate loss of a few dbs in power required at the transmitter will

occur if a polynomial algorithm based on the semi-definite relaxation is used in place of any exact technique (which of course is not known to be polynomial).

In the second part of the thesis we consider a few problems that arise in wireless broadcast channels. Namely, we consider the problem of the information symbol vector design at the transmitter. A polynomial linear precoding technique is constructed. It enables achieving data rates very close to the ones achieved with DPC (dirty paper coding) technique. Additionally, for another suboptimal polynomial scheme (the so-called nulling and cancelling), we show that it asymptotically achieves the same data rate as the optimal, exponentially complex, DPC.

In the last part of the thesis we consider a quantum counterpart of the signal detection from classical communication. In quantum systems the signals are quantum states and the quantum detection problem amounts to designing measurement operators which have to satisfy certain quantum mechanics laws. A specific type of quantum detection called unambiguous detection, which has numerous applications including quantum filtering, has recently attracted a lot of attention in the research community. We develop a general framework for numerically solving this problem using the tools from the convex optimization theory. Furthermore, in the special case where the two quantum states are of rank 2, we construct an explicit analytical solution for the measurement operators.

At the end we would like to emphasize that the contribution of this thesis goes beyond the specific problems mentioned here. Most algorithmic optimization techniques developed in this paper are generally applicable. While it is a fact that our results were originally motivated by wireless and quantum communications applications, we believe that the developed techniques will find applications in many different areas where similar optimization problems appear.

List of Figures

Contents

Notation and Abbreviations

$\mathrm{Tr}(A)$	trace(A)
A^*	Hermitian transpose of matrix A
$A_{1:k,k:m}$	Matrix obtained by selecting rows 1 through k and columns k through m of matrix A
$\|\|A\|\|_F$	Frobenius norm of matrix A
EC	Expected complexity
$\mathbf{y}_{k:m}$	Components k through m of vector $\mathbf{y}$
$\mathbf{y}^*$	Hermitian transpose of vector $\mathbf{y}$
$\mathrm{diag}(a_1, a_2, \ldots, a_n)$	Diagonal matrix with components a_1, a_2,...a_n on the main diagonal
$\mathcal{R}(A)$	Real part of matrix A
$\mathcal{I}(A)$	Imaginary part of matrix A
SIMO	Single-input multiple-output
MIMO	Multiple-input multiple-output
ML	Maximum likelihood
SDP	Semi-definite programming

Chapter 1

Introduction

In the last several decades we have witnessed enormous advance in all scientific fields. It is not difficult to recognize that one of the key contributing factors is fast and efficient information exchange. People from different cultures, different educational backgrounds, different geographical areas can quickly exchange their knowledge, understanding, and experience of the natural phenomena around us. The difference in the amount of information exchange between today and several decades ago is obvious. This difference is even more obvious between today and a few centuries back. What would be several months of sailing to convey information between two continents a few centuries ago, today is less than a second with cell-phones and internet. Clearly, modern technological achievements such as cell-phones, computers, and internet are playing the crucial role. At this point it is even unimaginable how rapid information exchange will be in the years, decades, and centuries ahead, and what kind of devices will provide it. However, what is certain is that the entire process of information exchange will constantly keep improving.

We live at the time when the most powerful, personally available mobile gadgets to convey information over huge distances are cell-phones and computers. The incredible power of cell-phones is based on their ability to transmit/receive information as an electromagnetic wave through the air from almost any location. Of course, the secret of transmitting information through the air with no wires, cables, etc., has been known for a long time. However, only with the appearance of small portable devices has the massive use of wireless communications taken off. The commercial availability of cell-phones and computers within

the last 20 years has made the entire planet, at any given point, as connected as a small village. Still, no matter how advanced the current technology seems to be, we always look for better.

Although modern wireless devices provide great services, some of them (e.g., video-transmitting) that require high data rate transmissions are still unavailable. In last 10–15 years the idea of using multiple antennas instead of a single antenna at both the receiving and the transmitting ends of the point-to-point communication system (depicted in Figure 1.1), has become a subject of extensive theoretical and practical research. The main idea is that adding the antennas should increase the amount of information that can be transmitted. It in fact was mathematically shown in [95] and [40] that the achievable transmission rate increases linearly with the number of transmitting/receiveing antennas, provided that the transmission medium is Gaussian. These results are of course more on a theoretical side. It still remains an open question how to practically design these systems and achieve the theoretically predicted performances.

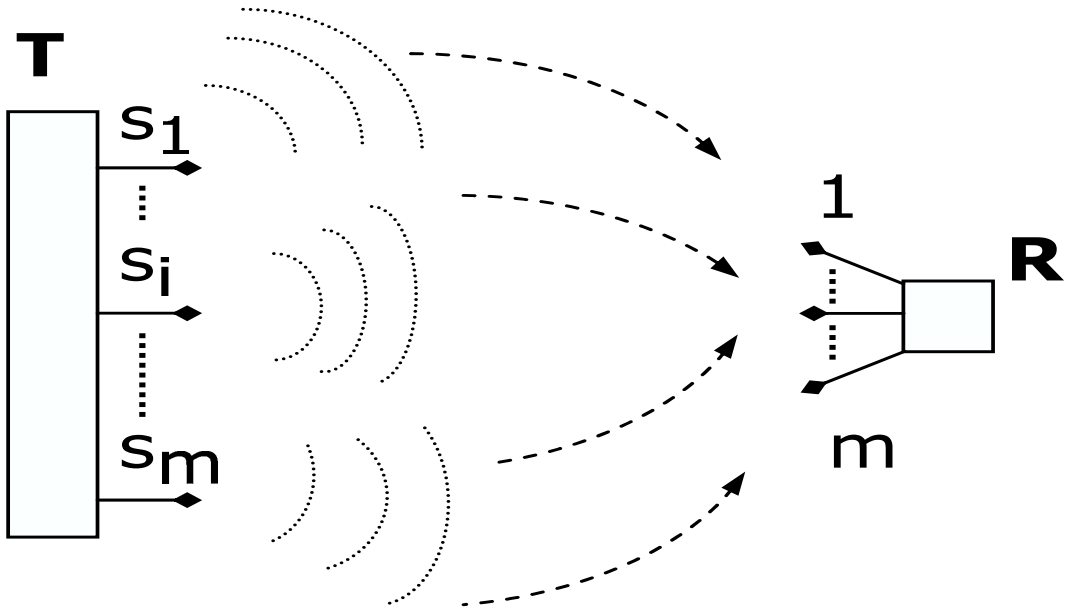


Figure 1.1: Multi-antenna system

One of the most important questions to be resolved is the complexity of the receivers. The detection at the receiver **R** of the signals **s** sent from the transmitter **T** usually amounts

to solving hard combinatorial optimization problems. A large portion of this thesis is related to improving the current detection techniques. Practically, this means that a large part of this thesis is focused on the design of the new and the improvement of the existing exact and approximate optimization algorithms for solving these hard problems.

As we have said, the multiple antennas can be employed in the so-called point-to-point communication. Point-to-point communication assumes that there are two ends of communication, the transmitter and the receiver. Furthermore the assumption in the context of wireless communications is that they communicate through wireless medium. Another interesting concept of communicating that wireless medium allows is called broadcast communication. In broadcast communication one transmitter (say a base station) broadcasts the information that can be received by different users (see Figure 1.2). In the literature this system is often called a broadcast channel. Similar to point-to-point communication, in broadcast communication the transmitter and the receivers can be equipped with multiple antennas. The main idea is that if the transmitter is equipped with, say, m antennas, it can simultaneously serve m users. However, since the transmission medium is transparent for all signals, parts of the information intended for one of the users will reach users that are not intended to receive it. The piece of non-intended information which reaches receivers may cause them to incorrectly detect the information intended for them. This problem (commonly known as interference) is caused by the fading nature of the wireless broadcast medium and is one of the main issues in wireless broadcast systems. The problem of designing the information symbol vectors at the transmitter so that the problem of interference among the users is as mitigated as possible has been extensively studied during the last decade. Significant theoretical results related to the performance limits of broadcast channels have been achieved. Surprisingly, the fading characteristics of the channel can in fact be utilized so that the expected increase in the overall amount of information that can be reliably transmitted is achieved.

Before mentioning the significant theoretical achievements in this area, we would like to briefly mention what the characteristics of interest in a broadcast channel are. Unlike

in a point-to-point system, in a broadcast channel we define sum-rate capacity to be the maximum achievable sum of the rates of information that can be reliably sent to the users. In addition to this, we can define the achievable rate region as the set of the rates at which the information can be sent to the individual users. It is not difficult to note that the sum-rate capacity is the point on the boundary of the rate region that maximizes the sum of all individual rates.

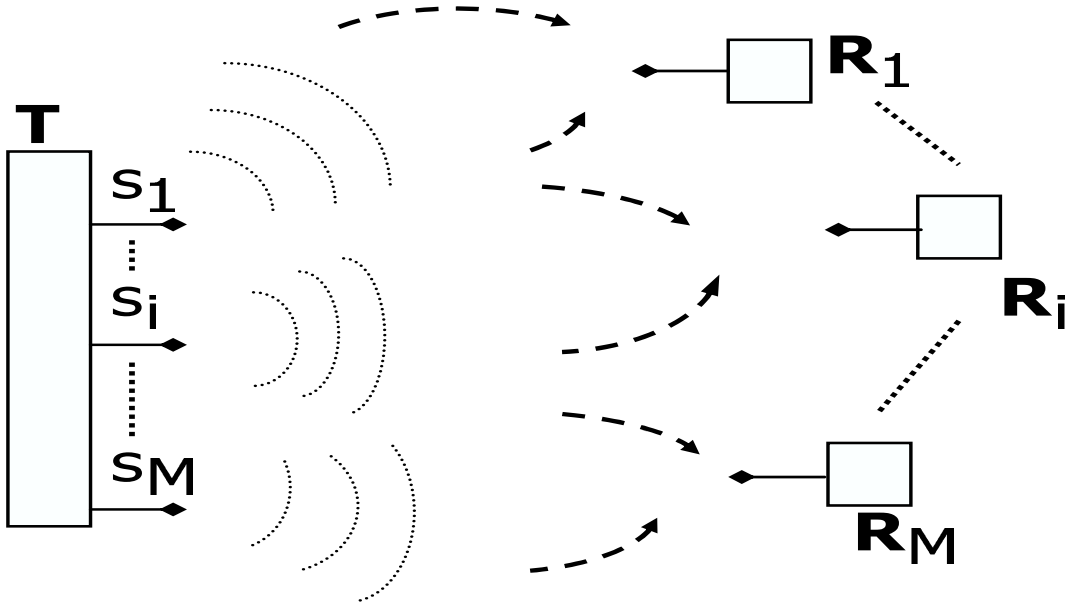


Figure 1.2: Wireless broadcast system

As we have mentioned above, during the last decade the analysis of the performance limits of broadcast channels has been the subject of extensive research. Namely, the achievable sum-rate of the Gaussian broadcast channel has recently been computed [101], and a particular strategy called dirty-paper coding (DPC) introduced in [20] has been proven to achieve it (see, e.g., [14] and [112]). Furthermore, it also happens that the DPC achieves all points on the boundary of the rate region of the broadcast channel. However, DPC is a very complex technique and difficult to implement in practice. In this thesis we introduce a few alternatives to DPC and analyze their performance.

Of course, many of the theoretical results related to the broadcast channel are very sensitive to the information that the transmitter has about the transmission medium. In

fact they all hold under the assumption that the transmitter has full knowledge of the transmission medium. In this thesis we will focus exactly on this same case when the characteristics of the broadcast channel don't change rapidly in time so that they can be estimated. Furthermore, we will mostly focus on the case where the number of the users served by the transmitter (base station) is of similar order as the number of the transmitting antennas. On the other hand we would like to mention that, even when the transmitter has only a partial information about the channel and the number of the receivers/users is large, the sum-rate capacity asymptotically scales linearly in number of transmitting antennas m and doubly logarithmic in the number of users — effectively the same way the sum-rate capacity of the optimal DPC scales with the number of transmitting antennas and the number of users (see, e.g., [83]).

1.1 Thesis contributions

This thesis contains three main parts. The first two deal with the problems related to the multiple-antenna point-to-point and broadcast channels. The third part is somewhat different and is related to the specific problem of quantum detection in quantum systems.

1.1.1 ML detection

In the first part of the thesis we consider multiple-input multiple-output (MIMO) systems. A point-to-point communication system which consists of one transmitter and one receiver equipped with multiple (m) antennas is an example of a MIMO system. As we have said above, one of the key problems in designing these systems is the complexity of the receiver. It is very well known that in MIMO systems maximum-likelihood (ML) detection of the received signals amounts to solving integer least-squares problems which are NP-hard in worst case. Since NP-hard problems may require a very long time to be solved exactly, it is of great interest to find efficient algorithms that can be practically implemented.

In Chapter 2 we consider the case of the so-called coherent ML detection in MIMO systems. The coherent MIMO detection assumes that the receiver knows the channel (trans-

mission medium). This is a valid assumption if the communication occurs under non-rapidly changing environmental conditions so that the channel can be estimated. The problem of MIMO detection then becomes the integer least-squares problem (see, e.g., [49]). Usually in the context of wireless communications a specific algorithm called sphere-decoder (more on the sphere decoder can be found in [37]) is used for solving ML detection problems. Since the algorithm is used in a statistical system as a measure of its quality, a quantity called the expected complexity is usually considered in the literature (see, e.g., [49]). In fact, as shown in [49], the expected complexity over the wide range of the signal-to-noise (SNR) ratios and the dimension of the problem m is smaller than m^3. However, when the SNR decreases and the dimension of the problem grows, the complexity of the sphere decoder becomes prohibitive. In Chapter 2 we develop a branch and bound modification of the standard sphere decoder algorithm and demonstrate that it significantly decreases the size of the search tree compared to the original version of the algorithm. Furthermore, we establish an interesting mathematical connection between the H-infinity estimation theory and the lower bounding of the integer least-squares.

In Chapter 3 we consider the case of the so-called non-coherent ML detection in single-input multiple-output (SIMO) systems with q-PSK signalling. It is well known that the problem of ML detection in SIMO systems in case of q-PSK signalling reduces to the integer maximization of the quadratic form. This problem is also known to be NP-hard in the worst case. In the first part of Chapter 3 we develop the so-called out-sphere decoder algorithm for solving ML detection problems exactly. The out-sphere decoder represents a counterpart to the sphere decoder used in the context of coherent MIMO detection. We additionally analytically establish an upper bound on the expected complexity of the out-sphere decoder. However, this upper bound turns out to be exponential, which suggests that solving ML detection in SIMO systems exactly still remains hard. In the second part of Chapter 3 we consider the scenarios when it is not necessary to solve the ML detection problem exactly. We introduce several algorithms for solving it approximately and analyze their performance. As it turns out, they provably perform almost as well as the exact ones.

1.1.2 Broadcast channels

In the second part of the thesis (Chapters 4 and 5) we consider a Gaussian broadcast channel. In Chapter 4 we introduce a few practical schemes based on the linear precoding for the design of the information symbols at the transmitter in a Gaussian broadcast channel. We design the precoding strategy: 1) so that the overall sum-rate is maximized and 2) so that the minimum rate among all users is maximized. The later one is shown to be a quasi convex problem and solved exactly.

In Chapter 5 we analyze the theoretical limits of a particular non-linear scheme called vector-perturbation technique introduced in [75]. It turns out that an even simpler version of it, based on the nulling and cancelling procedure, asymptotically achieves the sum-rate of the DPC.

1.1.3 Quantum unambiguous detection

In the third part of the thesis (Chapters 6 and 7) we consider quantum systems. More specifically the problems that we are interested in are related to the quantum detection theory.

In quantum systems, unlike in classical, the information is stored in special objects called quantum states. Namely a quantum state is a set of numbers which fully describes the quantum system. These numbers are usually stored in a vector called a pure state [73]. Furthermore, the state of a quantum system can be a statistical mixture of pure states, in which case it is called a mixed quantum state [73].

In order to detect in which state a quantum system was prepared we need to construct a set of quantum measurements. In Chapters 6 and 7 we consider a specific problem of designing an optimal set of measurements that distinguishes unambiguously between a collection of mixed quantum states. First, in Chapter 6, using arguments of duality in vector space optimization, we derive necessary and sufficient conditions for an optimal measurement that maximizes the probability of correct detection. We show that the previous optimal measurements that were derived for certain special cases satisfy these optimality conditions.

We then consider state sets with strong symmetry properties, and show that the optimal measurement operators for distinguishing between these states share the same symmetries, and can be computed very efficiently by solving a reduced size semi-definite program.

In Chapter 7 we consider a specific problem of unambiguous detection between two mixed quantum states of rank 2 which has been open for quite a while. Based on the general framework from Chapter 6 we explicitly analytically characterize the optimal measurement operators. Furthermore, using the same framework we easily obtain an explicit solution of unambiguous detection between a pure and a mixed quantum state matching an already known solution obtained in the context of quantum filtering.

Chapter 2

Coherent ML Detection in Multi-Antenna Systems — Sphere Decoder Algorithm

2.1 Introduction

In this chapter we will consider a point-to-point system where each end of the communication is equipped with several antennas. Such systems are usually called MIMO systems and a sketch of such a system is shown in Figure 2.1. As mentioned earlier, in the introduction, the main idea behind adding more antennas at the transmitter and/or receiver is to increase overall throughput (the amount of information that can be transmitted). Intuitively we may expect that adding m antennas at the transmitter and receiver should increase the achievable rate of the system m times compared to the system with only one antenna. Furthermore, it can actually be explicitly shown (see, e.g. [40] and [95]) that the overall throughput of the system indeed increases linearly with the number of antennas. These results are of course more of a theoretical nature. However, they suggest that the multi-antenna systems should provide very high data rates of $\approx$ Mbits/s. Of course, these theoretical results only show what is the achievable limit of the MIMO system. However, designing a MIMO system so as to achieve this limit is not an easy task. In this chapter we will address a very important problem which happens to be a significant obstacle in designing MIMO systems: More specifically, we will consider the problem of signal detection

at the receiver in a MIMO system.

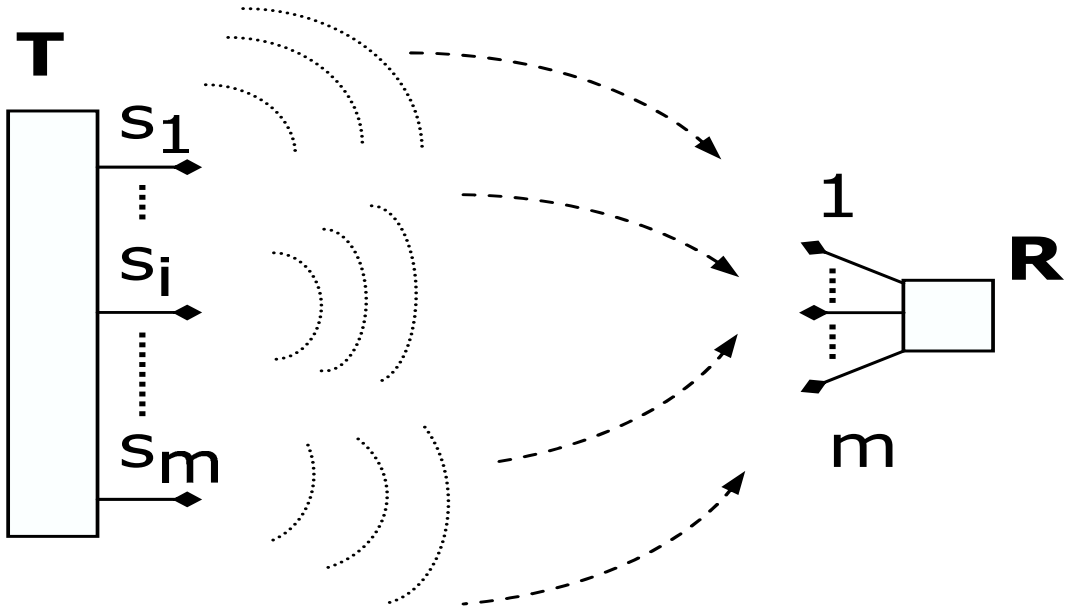


Figure 2.1: Multi-antenna system

Before going exactly to the specific problem, let us elaborate briefly on the the way a MIMO system works and how it is typically modeled. As can be seen from Figure 2.1, the system consists of the transmitter **T** and and receiver **R**, which both have m antennas. The transmitter then sends the sequence (vector) of the information symbols $\mathbf{s}$, as depicted in Figure 2.1. The signals from the vector $\mathbf{s}$ are sent as electromagnetic waves through the transmission medium (air) such that signal $s_i, 1 \leq i \leq m$ is sent from the i-th antenna. Since the transmission medium is air, the signals sent from any of m transmitter antennas can reach any of m receiver antennas. It is usually assumed that the signals from different transmitting antennas are combined in a linear fashion at the receiver. This mathematical simplification of a real MIMO system is shown in Figure 2.2.

It should also be noted that the signals sent from the m-th antenna at the transmitter are being weakened/strengthened on their way to the j-th antenna at the receiver. In Figure 2.2 this is modeled by adding a factor h_{jm} on this path. Clearly in a similar fashion there can be defined a full matrix H of the channel coefficients which will model the quality of the path from the transmitting to the receiving antennas. It is not difficult to imagine that the h_{ji}

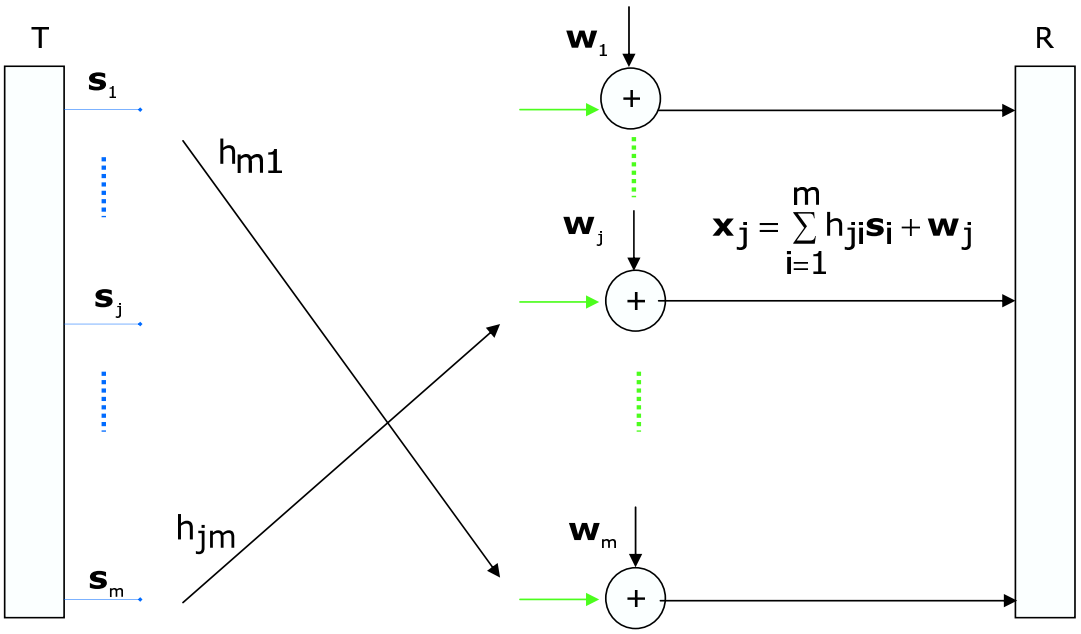


Figure 2.2: Mathematical model of multi-antenna system

entry of the matrix H will contain the characteristics of the path from i-th transmitting to the j-th receiving antenna. In context of wireless communications the matrix H is usually called the channel matrix. Additionally, signal at the each of the receiving antennas is corrupted by noise (for the j-th receiving antenna this is shown in Figure 2.2 as quantity w_j). It is not difficult to see that the models from Figures 2.1 and 2.2 can be summarized in the following equation

$$\mathbf{x} = H\mathbf{s} + \mathbf{w}. \tag{2.1}$$

As we have said earlier, H is the channel matrix and $\mathbf{w}$ is the noise vector. To allow tractability of the model it is usually assumed that the channel matrix coefficients are i.i.d. Gaussian random variables with zero mean and unit variance. The components of the noise vector $\mathbf{w}$ are also assumed to be i.i.d. Gaussian, independent of the entries of the channel matrix H, with zero mean and variance σ^2.

Now we slowly turn to the problem of MIMO detection which arises in the above-defined system. Namely, looking at equation (2.1) one can ask a simple question: given the value

of the vector $\mathbf{x}$ (its components are signals received at the receiving antennas, i.e., they are known to the receiver), can one somehow figure out what the value of the signal vector $\mathbf{s}$ was? In fact this is precisely the question of signal detection in multi-antenna systems. In the most general setting this question is very difficult. Not only don't we know the matrix H, but knowledge of the vector $\mathbf{w}$ is also absent. Clearly, following this argument we recognize that learning H at the receiver is an important question. This has of course been the subject of extensive research, and in situations where the channel coefficients from the matrix H are not changing rapidly in time, they can in fact be estimated [47]. The case when the channel matrix is known at the receiver is usually referred to as the coherent case of the signal detection in multi-antenna systems. In the rest of this chapter we will assume communication in the slowly changing environment, so that the channel matrix is known to us. [We will elaborate more on the non-coherent case, when the channel matrix can not be estimated, in the following chapter.]

Looking at equation (2.1) we can write

$$p(\mathbf{x}|\mathbf{s}, H) = \frac{1}{\sqrt{2\pi}^n} e^{-\frac{||\mathbf{x}-H\mathbf{s}||_2^2}{2}}. \tag{2.2}$$

Then we can set the maximization of the probability from (2.2) as a criterion for the detection of the signal vector $\mathbf{s}$ if the received vector $\mathbf{x}$ and the channel matrix H are known. The detected vector $\mathbf{s}_{\mathrm{ML}}$ is then

$$\mathbf{s}_{\mathrm{ML}} = \arg\min_{\mathbf{s}\in\mathcal{D}} p(\mathbf{x}|\mathbf{s}, H) = \arg\min_{\mathbf{s}\in\mathcal{D}} ||\mathbf{x} - H\mathbf{s}||_2^2. \tag{2.3}$$

This criterion for signal detection in MIMO systems is called maximum-likelihood (ML) detection. It should also be noted that the vector $\mathbf{s}$ is restricted to a set $\mathcal{D}$, which in digital communications is commonly assumed to be a subset of the integer lattice.

It is well known that maximum-likelihood (ML) decoding in many digital communication schemes reduces to solving the integer least-squares problem given in (2.3), which is NP-hard in the worst case. On the other hand, it has recently been shown [49] that, over a wide

range of dimensions N and signal-to-noise ratios (SNRs), the sphere decoding algorithm [37] can be used to find the exact ML solution with an expected complexity that is often less than N^3. However, the computational complexity of sphere decoding becomes prohibitive if the signal-to-noise ratio (SNR) is too low and/or if the dimension of the problem m is too large.

In this chapter, we target these two regimes and attempt to find faster algorithms by pruning the search tree beyond what is done in the standard sphere decoding algorithm. The search tree is pruned by computing lower bounds on the optimal value of the objective function as the algorithm proceeds to descend down the search tree. We observe a trade-off between the computational complexity required to compute a lower bound and the size of the pruned tree: the more effort we spend in computing a tight lower bound, the more branches that can be eliminated in the tree. Using ideas from SDP-duality theory and H^∞ estimation theory, we propose general frameworks for computing lower bounds on integer least-squares problems. We propose two families of algorithms, one of which is appropriate for large problem dimensions and binary modulation, and the other of which is appropriate for moderate-size dimensions yet high-order constellations. We then show how in each case these bounds can be efficiently incorporated in the sphere decoding algorithm, often resulting in significant improvement of the expected complexity of solving the ML decoding problem, while maintaining the exact ML performance.

2.2 Sphere decoder and its modification

In this section we recall what the standard sphere decoder is and introduce its branch and bound modification. We recall that the problem that we are interested in and will be solving *exactly* in this chapter has the following form

$$\min_{\mathbf{s}\in\mathcal{D}\subset\mathcal{Z}^m} \|\mathbf{x} - H\mathbf{s}\|_2, \tag{2.4}$$

where $\mathbf{x} \in \mathcal{R}^n$, $H \in \mathcal{R}^{n \times m}$, and $\mathcal{D}$ refers to some subset of the integer lattice $\mathcal{Z}^m$. The main idea of the sphere decoder algorithm [37] for solving the previous problem is based on finding all points $\mathbf{s}$ such that $H\mathbf{s}$ lies within a sphere of some adequately chosen radius d_s centered at $\mathbf{x}$, i.e., on finding all $\mathbf{s}$ such that

$$d_s^2 \geq \|\mathbf{x} - H\mathbf{s}\|_2^2, \tag{2.5}$$

and then choosing the one that minimizes the objective function. Using the QR decomposition $H = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0_{n-m \times m} \end{bmatrix}$, where R is $m \times m$ upper triangular, and $Q_1 \in R^{n \times m}$ and $Q_2 \in R^{n \times (n-m)}$ are such that $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ is unitary, we can reformulate (2.5) as

$$d^2 \geq \|\mathbf{y} - R\mathbf{s}\|_2^2, \tag{2.6}$$

where we have defined $d^2 = d_s^2 - \|Q_2^* \mathbf{y}\|^2$ and $\mathbf{y} = Q_1^* \mathbf{x}$.

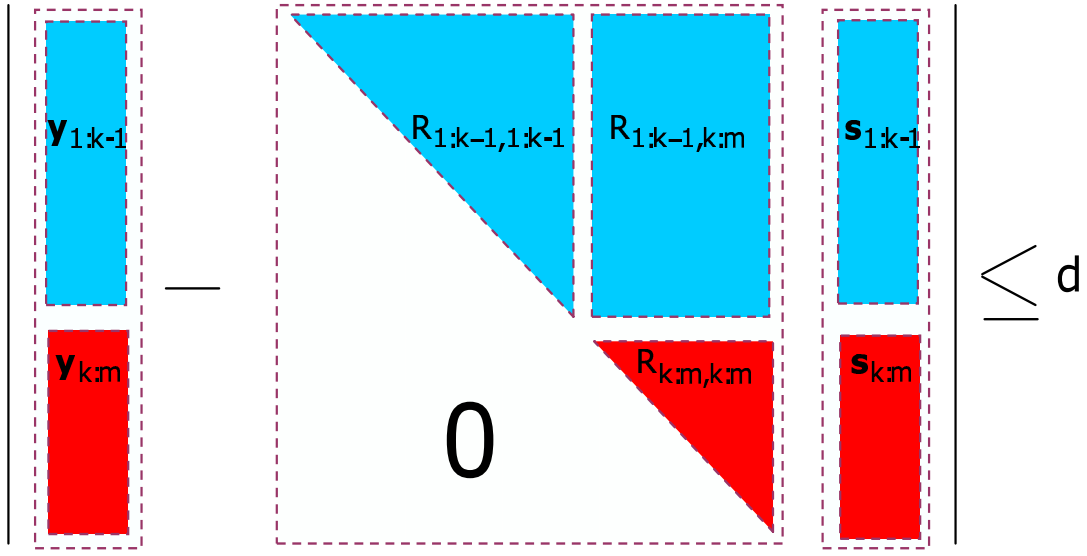


Figure 2.3: Upper-triangular decomposition — the key component of the sphere decoder algorithm

Now using the upper-triangular property of R (see Figure 2.3), (2.6) can be further

rewritten as

$$d^2 \geq \|\mathbf{y}_{k:m} - R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \|\mathbf{y}_{1:k-1} - R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2, \quad (2.7)$$

for any $2 \leq k \leq m$, where the subscripts determine the entries the various vectors and matrices run over (e.g., $\mathbf{y}_{1:k-1}$ is a column vector whose components are $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{k-1}$, and similarly $R_{1:k-1,k:m}$ is a $(k-1) \times (m-k+1)$ matrix and $R_{i,k}, R_{i,k+1}, \ldots, R_{i,m}$ are the components of its i-th row). A necessary condition for (2.6) can therefore be obtained by omitting the second term on the right-hand side (RHS) of the above expression to yield

$$d^2 \geq \|\mathbf{y}_{k:m} - R_{k:m,k:m}\mathbf{s}_{k:m}\|^2. \quad (2.8)$$

The sphere decoder finds all points $\mathbf{s}$ in (2.5) by proceeding inductively on (2.8), starting from $k = m$ and proceeding to $k = 1$. In other words, for $k = m$ it determines all one-dimensional lattice points $\mathbf{s}_m$ such that

$$d^2 \geq (\mathbf{y}_m - R_{m,m}\mathbf{s}_m)^2,$$

and then, for each such one-dimensional lattice point $\mathbf{s}_m$, determines all possible values for $\mathbf{s}_{m-1}$ such that

$$\begin{aligned} d^2 &\geq \|\mathbf{y}_{m-1:m} - R_{m-1:m,m-1:m}\mathbf{s}_{m-1:m}\|^2 \\ &= (\mathbf{y}_m - R_{m,m}\mathbf{s}_m)^2 + (\mathbf{y}_{m-1} - R_{m-1,m-1}\mathbf{s}_{m-1} - R_{m-1,m}\mathbf{s}_m)^2. \end{aligned}$$

This gives all two-dimensional lattice points that satisfy (2.6); we proceed in a similar fashion until $k = 1$. The sphere decoding algorithm thus generates a tree (see Figure 2.4), where the branches at the $(m-k+1)$th level of the tree correspond to all $(m-k+1)$-dimensional lattice points satisfying (2.8). Therefore, at the bottom of the tree (the m-th level) all points satisfying (2.5) are found. (For more details on the sphere decoder and for

an explicit description of the algorithm the reader may refer to [37], [23], and [49].)

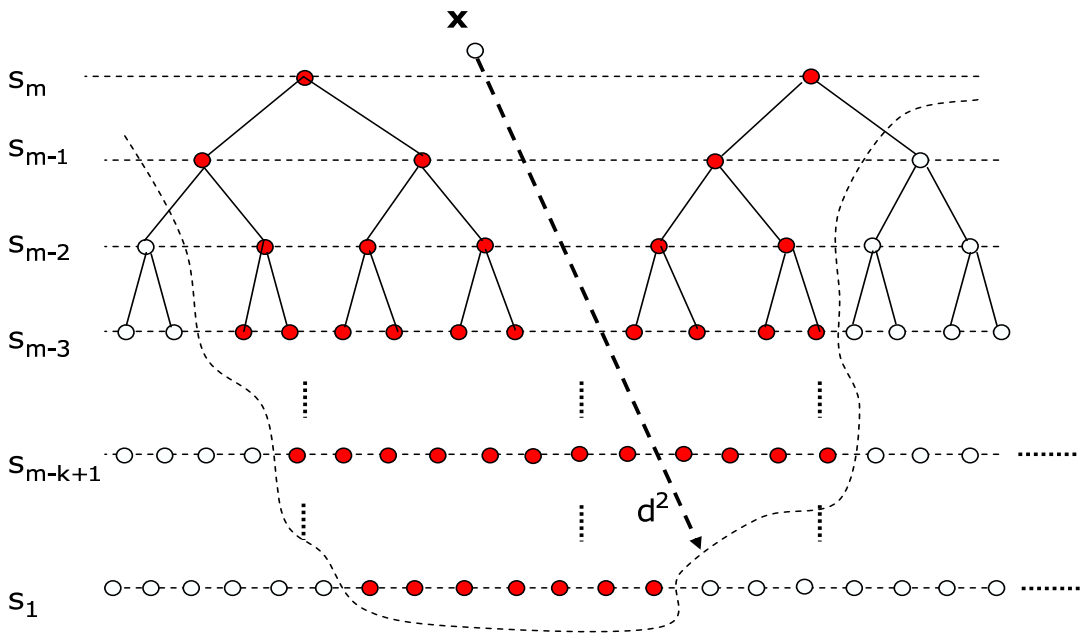


Figure 2.4: Tree generated by the sphere decoder algorithm

The computational complexity of the sphere decoding algorithm depends on how d is chosen. In a digital communication context, $\mathbf{x}$ is the received signal, i.e., a noisy version of the symbol vector $\mathbf{s}$ transmitted across the channel H,

$$\mathbf{x} = H\mathbf{s} + \mathbf{w}, \tag{2.9}$$

where the entries of the additive noise vector $\mathbf{w}$ are independent, identically distributed (iid) $\mathcal{N}(0, \sigma^2)$ random variables. In [49] it is shown that, if elements of H are i.i.d. Gaussian with zero mean and unit variance and if the radius is chosen appropriately based on the statistical characteristics of the noise $\mathbf{w}$, then over a wide range of signal-to-noise ratios (SNR) and problem dimensions the expected complexity of the sphere-decoding algorithm is low and comparable to the complexity of the best known heuristics, which are cubic.

The above assertion unfortunately fails and the computational complexity becomes increasingly prohibitive if the SNR is too low and/or if the dimension of the problem is too large (in fact as shown in [58] the expected computational complexity of the sphere decoder

is always exponential). Increasing the dimension of the problem clearly is useful [1]. Moreover, the use of the sphere decoder in low SNR situations is also important, especially when one is interested in obtaining soft information to pass onto an iterative decoder (see, e.g., [52] and [98]). One way to reduce the computational complexity is to resort to suboptimal methods based either on heuristics (see, e.g., [5]) or some form of statistical pruning (see [43]). Also, the interested reader may find more about recent improvements and alternative techniques in [22], [110], [45], [65], and [103].

In this chapter, we attempt to reduce the computational complexity of the sphere decoder while still finding the *exact* solution. Let us surmise on how this may be done. As mentioned above, the sphere decoding algorithm generates a tree whose number of branches at each level corresponds to the number of lattice points satisfying (2.8). Clearly, the complexity of the algorithm depends on the size of this tree, since each branch in the tree is visited and appropriate computations are then performed. Thus, one approach to decrease the complexity is to reduce the size of the tree beyond that which is suggested by (2.8). To this end, consider a lower bound on the optimal value of the second term on the RHS of (2.7):

$$LB = LB(\mathbf{y}_{1:k-1}, R_{1:k-1,1:m}, \mathbf{s}_{k:m}) \leq \min_{\mathbf{s}_{1:k-1} \in \mathcal{D} \subset \mathcal{Z}^{k-1}} \|\mathbf{y}_{1:k-1} - R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2,$$

where we have emphasized the fact that the lower bound is a function of $\mathbf{y}_{1:k-1}$, $R_{1:k-1,1:m}$, and $\mathbf{s}_{k:m}$. Provided our lower bound is nontrivial (i.e., $LB > 0$), we can replace (2.8) by[2]

$$d^2 - LB \geq \|\mathbf{y}_{k:m} - R_{k:m,k:m}\mathbf{s}_{k:m}\|^2. \tag{2.10}$$

This is certainly a more restrictive condition than (2.8), and so will lead to eliminating more points from the tree as illustrated in Figure (2.5). Note that (2.10) will not result in missing

[1] Various space-time codes result in integer least-squares problems where the problem dimension is much larger than the number of transmit antennas. Also, in distributed space-time codes for wireless relay networks the problem dimension is equal to the number of relay nodes, which can be quite large ([64] and [61]).

[2] $LB = 0$, of course, simply corresponds to the standard sphere decoder.

any lattice points from (2.5) since we have used a *lower bound* for the remainder of the cost in (2.7) (for more on branch and bounding ideas, the interested reader may refer to [70] and the references therein). Assuming that we have some way of computing a lower bound $LB > 0$ as suggested above, we state the modification of the standard sphere decoding algorithm based on (2.10). The algorithm uses the Schnorr-Euchner (S-E) strategy with radius update [2]. [Note that in this chapter we consider several modifications of the sphere decoding algorithm, and all are implemented using Algorithm 1 below. The difference between the various modifications is how the value of LB in step 4 of the algorithm is computed. Also, note that in Algorithm 1, given below, $\mathcal{D}$ is the full integer lattice, while later in this chapter, in different modifications of the original algorithm, it will be restricted to its subsets.]

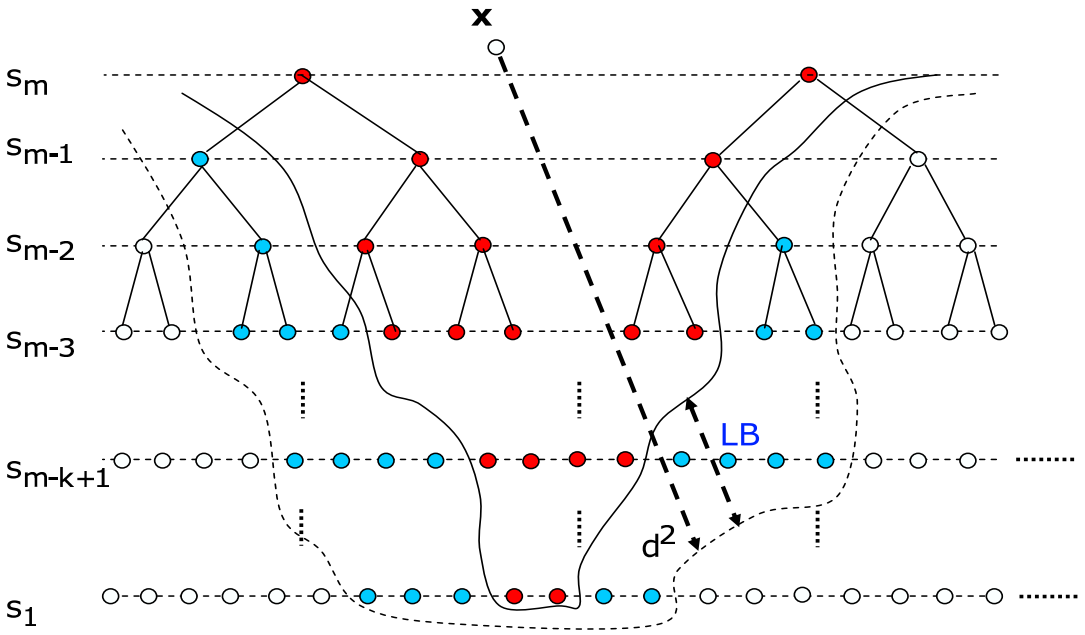


Figure 2.5: Reduced tree of the branch and bound sphere decoding algorithm

Algorithm 1:

Input: Q, R, $\mathbf{x}$, $\mathbf{y} = Q_1^*\mathbf{x}$, $d = \hat{d}$, $\mathbf{ll}_{1:m} = \mathbf{0}_{1\times m}$.

1. Set $k = m$, $d_m^2 = d^2$, $\mathbf{y}_{m|m+1} = \mathbf{y}_m$.

2. (Bounds for $\mathbf{s}_k$) Set $ub(\mathbf{s}_k) = \lfloor \frac{\sqrt{d_k^2-(d^2-\hat{d}^2)}+\mathbf{y}_{k|k+1}}{r_{k,k}} \rfloor$, $lb(\mathbf{s}_k) = \lceil \frac{-\sqrt{d_k^2-(d^2-\hat{d}^2)}+\mathbf{y}_{k|k+1}}{r_{k,k}} \rceil$, $l_k = \lfloor \frac{lb(\mathbf{s}_k)+ub(\mathbf{s}_k)+1}{2} \rfloor$, $u_k = l_k + 1$.

3. (Zig-zag through $\mathbf{s}_k$)

   If $\mathbf{ll}_k = 0$, $\mathbf{s}_k = l_k$, $l_k = l_k - 1$, $\mathbf{ll}_k = 1$, otherwise $\mathbf{s}_k = u_k$, $u_k = u_k + 1$, $\mathbf{ll}_k = 0$.

   If $lb(\mathbf{s}_k) \leq \mathbf{s}_k \leq ub(\mathbf{s}_k)$, go to 4, else go to 5.

4. If $LB(\mathbf{y}_{1:k-1}, R_{1:k-1,1:m}, \mathbf{s}_{k:m}) + (\mathbf{y}_{k|k+1} - r_{k,k}\mathbf{s}_k)^2 - d_k^2 + (d^2 - \hat{d}^2) > 0$, go to 3, else go to 6.

5. (Increase k) $k = k + 1$; if $k = m + 1$ terminate algorithm, else go to 3.

6. (Decrease k) If $k = 1$ go to 7. Else $k = k - 1$, $\mathbf{y}_{k|k+1} = \mathbf{y}_k - \sum_{j=k+1}^{m} r_{k,j}\mathbf{s}_j$, $d_k^2 = d_{k+1}^2 - (\mathbf{y}_{k+1|k+2} - r_{k+1,k+1}\mathbf{s}_{k+1})^2$, and go to 2.

7. Solution found. Save $\mathbf{s}$ and its distance from $\mathbf{x}$, $\hat{d} = d_m^2 - d_1^2 + (\mathbf{y}_1 - r_{1,1}\mathbf{s}_1)^2$, and go to 3.

Clearly, the tighter the lower bound LB, the more points will be pruned from the tree. Of course, we cannot hope to find the optimal lower bound, since this requires solving an integer least-squares problem (which was our original problem to begin with). Therefore, in what follows we focus on obtaining computationally feasible lower bounds on the integer least-squares problem

$$\min_{\mathbf{s}_{1:k-1} \in \mathcal{D} \subset \mathcal{Z}^{k-1}} \|\mathbf{z}_{1:k-1} - R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1}\|^2, \tag{2.11}$$

where, for simplicity, we introduced $\mathbf{z}^{(k-1)} = \mathbf{z}_{1:k-1} = \mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}$. Also, in the rest of this chapter we will assume $\mathcal{D} = \{-\frac{M-1}{2}, -\frac{M-3}{2}, \ldots, \frac{M-3}{2}, \frac{M-1}{2}\}^m$, the case which is of interest in communications applications.

Before proceeding any further, however, we note that finding a lower bound on (2.11) requires *some* computational effort. Therefore, it is a natural question to ask whether the benefits of additional pruning outweigh the additional complexity incurred by computing

a lower bound. An even more basic question, perhaps, is what are the potential pruning capabilities of the lower bounding technique which we use to modify the sphere decoding algorithm. To illustrate this, consider a simple lower bound (which is only valid in the binary case, i.e., if $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^m$) on (2.11), used earlier in [88] and further considered in Section 2.3, which is based on duality and may be computed by solving the following semi-definite program (SDP),

$$\begin{aligned} \max_{\Lambda} \quad & \mathrm{Tr}(\Lambda) \\ \text{subject to} \quad & Q \succeq \Lambda, \quad \Lambda \quad \text{is diagonal}, \end{aligned} \tag{2.12}$$

where

$$Q = \begin{bmatrix} \frac{1}{4} R^T_{1:k-1,1:k-1} R_{1:k-1,1:k-1} & -\frac{1}{2} R^T_{1:k-1,1:k-1} \mathbf{z}_{1:k-1} \\ -\frac{1}{2} \mathbf{z}^T_{1:k-1} R_{1:k-1,1:k-1} & \mathbf{z}^T_{1:k-1} \mathbf{z}_{1:k-1} \end{bmatrix}.$$

We mention that bounds of this type are very well known in the literature on semi-definite programming relaxations. More on them and their history can be found in [109]. Here we would like to only briefly mention the reason for their popularity. Although it is difficult to prove how tight these bounds are, it turns out that in practice they perform very well. On top of that, the optimization problem given above is *convex* (the objective function is convex and the region of optimization is convex as well) which means that these bounds can be computed very efficiently using a host of numerical methods [12]. Even more surprising, it can be proved that they can be computed in polynomial time.

Although these bounds have been known for a very long time, they attracted enormous interest in the algorithms and optimization areas after the work of [41]. Quite remarkably in [41] the authors were able to give a hard bound on the performance of the previously mentioned SDP-relaxation bound for a specific case of the matrix Q. Since then the SDP-relaxation techniques have become a standard tool in solving complicated combinatorial optimization problems. Naturally, many of these techniques have also been applied in detection problems (see, e.g., [94], [68], [1], [63], [57], and [69]). More specifically, in [94]

and [1] the authors considered applications of SDP relaxation to problems in multiuser detection in CDMA systems. In [68], [63], [57], and [69] authors applied the SDP relaxation to the problem of ML-detection in MIMO systems (the same one that we consider in this chapter). In [69] the authors generalized the applications of the SDP algorithm from binary to larger constellations, and in [63] the authors proved that under certain conditions related to the dimension of the problem in high-SNR regime the SDP relaxation is tight.

As demonstrated in these references, the SDP technique can be very powerful in producing a very good approximate solution of the original integer least squares problem. However in this work we focus on solving the integer least-squares problem *exactly*, and therefore we will only use its lower-bounding feature. It should also be noted that although in general suboptimal, the SDP technique can sometimes produce the *exact* solution to the original problem (for more on when this happens see [57]).

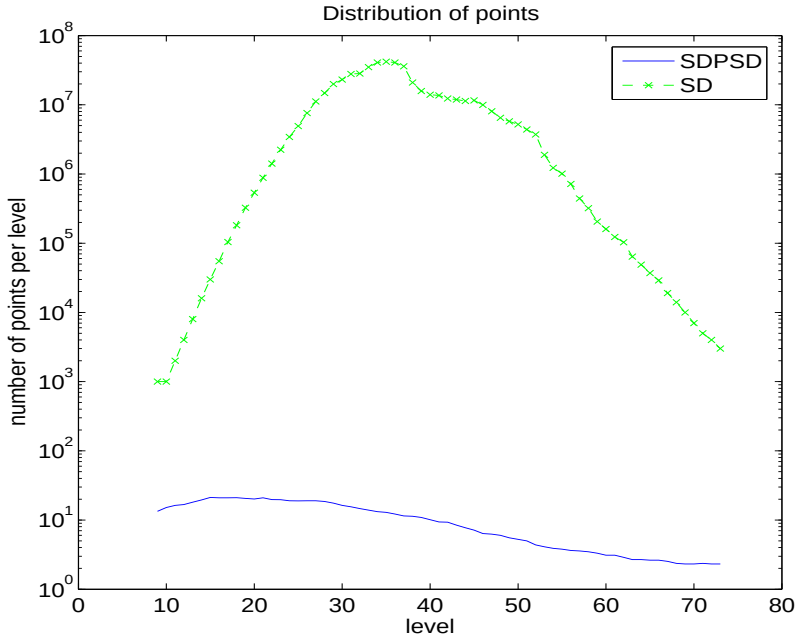


Figure 2.6: Comparison of the number of points per level in the search tree visited by the SD and the SDSDP algorithm, $m = 100$, SNR $= 10$ db, $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^m$.

After a brief chronology on the SDP relaxations we now turn our attention again to (2.12). We denote the optimal value of the objective function in (2.12) by LB_{SDP}. Figure 2.6

compares the number of points [3] on each level of the search tree visited by the basic sphere decoding algorithm with the corresponding number of points visited by the modified sphere decoding algorithm which employs LB_{SDP} for additional, lower-bound based, pruning. We refer to the former as the SD algorithm and to the latter as the SDSDP algorithm. As evident from Figure 2.6, for a problem of dimension $m = 100$, SNR$= 10dB$, and $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^m$ (i.e., BPSK modulation), the number of points in the search tree visited by the SDSDP algorithm is several orders of magnitude smaller than that visited by the SD algorithm. [The additional pruning of the search tree varies across the tree levels. The total number of the points visited by the SDSDP algorithm is roughly 10^6 times smaller than that visited by the SD algorithm.] Therefore, a good lower bound can help prune the tree much more efficiently than the standard sphere decoding alone. However, computing LB_{SDP} requires solving an SDP per each point in the search tree. The computational effort of solving an SDP is $O(k^{3.5})$, which is significantly greater than the linear complexity of the operations performed by the standard sphere decoding algorithm at every visited node. Furthermore, although the complexity scaling behavior of solving an SDP is provably $O(N^{3.5})$, even for moderately large N ($30 < N < 70$) the real complexity of solving the SDP given in (2.12) is $>> N^{3.5}$. On the other hand the standard sphere decoder performs per each node a number of operations that is $\approx N$. Therefore, there is merit in searching not only for tight lower bounds such as the one in (2.12), but also for those that may not be as tight but which require significantly smaller computational effort.

In this chapter we therefore introduce a lower bound LB_{sdp} on the quantity LB_{SDP} which can be computed with complexity linear in k. The idea is based on efficient propagation of LB_{sdp} through the search tree. We will show that the lower bound LB_{sdp} significantly improves the expected complexity of the standard sphere decoder. However, LB_{SDP} defined in (2.12) (and hence LB_{sdp}) is a valid lower bound only when $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^m$. To address the case of general $\mathcal{D}$, we derive another family of lower bounds on integer least-squares problems using ideas from H^∞ estimation theory. We show that several lower bounds that

[3] More on why the number of points may be important the interested reader can find in [13] and [59]

may otherwise be obtained by relaxing the optimization constraints, are in fact special cases of our general H^∞-based lower bound. When employing the above lower bounds to modify sphere decoding, we observe a trade-off between the computational complexity required to compute a lower bound and the size of the pruned tree: the more effort we spend in computing a tight lower bound, the more branches can be eliminated from the tree. We show that the most computationally efficient among the special cases, the so-called eigenbound, provides a significant improvement in the expected complexity over the sphere decoding algorithm.

The rest of this chapter is organized as follows. In Section 2.3 we derive a computationally efficient lower bound LB_{sdp} on LB_{SDP}. In Section 2.4, we derive the general H^∞-estimation-based lower bound on integer least-squares problem. In Sections 2.5, 2.6, and 2.8, special cases of this general bound are considered. In particular, the so-called spherical relaxation bound is derived in Section 2.5, the polytope relaxation bound is considered in Section 2.6, and the eigen bound is studied in Section 2.8. The effects of the aforementioned bounds on the number of search tree points and/or the total expected complexity of the modified sphere decoding algorithm are studied throughout. Some simulations are presented in Section 2.7, and, finally, Section 2.9 contains conclusions and a discussion of potential extensions of the current work.

2.3 SDP-based lower bound

Let $LB_{SDP}^{(k-1)}, 1 \leq k \leq m$ denote the optimal value of the following optimization problem,

$$\begin{aligned} LB_{SDP}^{(k-1)} = \max \quad & \mathrm{Tr}(\Lambda) \\ \text{subject to} \quad & Q_{k-1} \succeq \Lambda, \quad \Lambda \text{ is diagonal}, \end{aligned} \tag{2.13}$$

where

$$Q_{k-1} = \begin{bmatrix} \frac{1}{4} R_{1:k-1,1:k-1}^T R_{1:k-1,1:k-1} & -\frac{1}{2} R_{1:k-1,1:k-1}^T \mathbf{z}^{(k-1)} \\ -\frac{1}{2} (\mathbf{z}^{(k-1)})^T R_{1:k-1,1:k-1} & (\mathbf{z}^{(k-1)})^T \mathbf{z}^{(k-1)} \end{bmatrix}.$$

In this section we derive $LB_{sdp}^{(k-1)}$, a lower bound on $LB_{SDP}^{(k-1)}$. To this end, let $\hat{\Lambda}$ denote the optimal solution of

$$\begin{aligned} \max \quad & \mathrm{Tr}(\Lambda) \\ \text{subject to} \quad & Q \succeq \Lambda, \quad \Lambda \text{ is diagonal,} \end{aligned} \tag{2.14}$$

where

$$Q = \begin{bmatrix} \frac{1}{4}R^T R & -\frac{1}{2}R^T \mathbf{y} \\ -\frac{1}{2}\mathbf{y}^T R & \mathbf{y}^T \mathbf{y} \end{bmatrix}.$$

Let $GG^T = \frac{1}{4}R^T R - \hat{\Lambda}_{1:m,1:m}$, where G is a lower triangular matrix. Also, let $M = G^{-1}R^T$. Using the fact that the matrices G and R^T are lower triangular, we obtain

$$\begin{aligned} (G_{1:k-1,1:k-1})^{-1} &= (G^{-1})_{1:k-1,1:k-1}, \\ G_{1:k-1,1:k-1}(G_{1:k-1,1:k-1})^T &= \frac{1}{4}R^T_{1:k-1,1:k-1}R_{1:k-1,1:k-1} - \hat{\Lambda}_{1:k-1,1:k-1}, \end{aligned}$$

and $M_{1:k-1,1:k-1} = (G_{1:k-1,1:k-1})^{-1}R^T_{1:k-1,1:k-1}$. Furthermore, let

$$\lambda_k = (\mathbf{z}^{(k-1)})^T \mathbf{z}^{(k-1)} - \frac{1}{4}(\mathbf{z}^{(k-1)})^T M^T_{1:k-1,1:k-1} M_{1:k-1,1:k-1}\mathbf{z}^{(k-1)}, \tag{2.15}$$

and let

$$LB_{sdp}^{(k-1)} = \begin{cases} \sum_{i=1}^{k-1} \hat{\Lambda}_{i,i} + \lambda_k & \text{if} \quad \sum_{i=1}^{k-1} \hat{\Lambda}_{i,i} + \lambda_k \geq 0 \\ 0 & \text{if} \quad \sum_{i=1}^{k-1} \hat{\Lambda}_{i,i} + \lambda_k < 0. \end{cases} \tag{2.16}$$

Now it is clear that $LB_{sdp}^{(k-1)} \leq LB_{SDP}^{(k-1)}$ since

$$LB_{sdp}^{(k-1)} = \mathrm{Tr}(\mathrm{diag}(\hat{\Lambda}_{1,1}, \hat{\Lambda}_{2,2}, \ldots, \hat{\Lambda}_{k-1,k-1}, \lambda_k)),$$

and $\mathrm{diag}(\hat{\Lambda}_{1,1}, \hat{\Lambda}_{2,2}, \ldots, \hat{\Lambda}_{k-1,k-1}, \lambda_k)$ is an admissible matrix in (2.13). On the other hand, if $\sum_{i=1}^{k-1} \hat{\Lambda}_{i,i} + \lambda_k < 0$, $LB_{sdp}^{(k-1)} = 0$, and clearly $LB_{sdp}^{(k-1)} \leq LB_{SDP}^{(k-1)}$.

We refer to Algorithm 1 which, in step 4, makes use of $LB_{sdp}^{(k-1)}$ as the SDsdp algorithm. The subroutine for computing $LB_{sdp}^{(k-1)}$ is given below. Clearly, using $LB_{sdp}^{(k-1)}$ instead of $LB_{SDP}^{(k-1)}$ results in pruning fewer points in the search tree. However, the computation of $LB_{sdp}^{(k-1)}$ is quite a bit more efficient than the cubic computation of $LB_{SDP}^{(k-1)}$. In particular, unlike in the SDSDP algorithm, we need to solve only one SDP — the one given by (2.14). Then we may compute $LB_{sdp}^{(k-2)}$ recursively from $LB_{sdp}^{(k-1)}$, which requires complexity linear in k [92]. This is shown next.

Recall that $\mathbf{z}^{(k-1)} = \mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}$. It is easy to see that we can compute $\mathbf{z}^{(k-2)}$ from $\mathbf{z}^{(k-1)}$ as

$$\mathbf{z}^{(k-2)} = \mathbf{z}_{1:k-2}^{(k-1)} - R_{1:k-2,k-1}s_{k-1}. \tag{2.17}$$

Furthermore, note that $\mathbf{p}^{(k-1)} = M_{1:k-1,1:k-1}\mathbf{z}^{(k-1)}$ can be computed recursively as

$$\begin{aligned}
\mathbf{p}^{(k-2)} &= M_{1:k-2,1:k-2}\mathbf{z}^{(k-2)} \\
&= M_{1:k-2,1:k-2}(\mathbf{z}_{1:k-2}^{(k-1)} - R_{1:k-2,k-1}s_{k-1}) \\
&= M_{1:k-2,1:k-2}(\mathbf{z}^{(k-1)})_{1:k-2} - M_{1:k-2,1:k-2}R_{1:k-2,k-1}s_{k-1} \\
&= \mathbf{p}_{1:k-2}^{(k-1)} - (MR)_{1:k-2,k-1}s_{k-1}.
\end{aligned} \tag{2.18}$$

Using $\mathbf{p}^{(k-2)}$ and $\mathbf{z}^{(k-2)}$ we compute λ_{k-1} from (2.15), and $LB_{sdp}^{(k-2)}$ from (2.16). The computation of $LB_{sdp}^{(k-2)}$ in each node at the $(m-(k-2))$th level of the search tree requires $4(k-2)$ additions and $2(k-2)$ multiplications. For the basic sphere decoder, the number of operations per each node at the $(m-k+1)$th level is $(2k+17)$. This essentially means that the SDsdp algorithm performs about four times more operations per each node of the tree than the standard sphere decoder algorithm. In other words, if the SDsdp algorithm prunes at least four times more points than the basic sphere decoder, the new algorithm is faster in terms of the flop count.

Subroutine for computing LB_{sdp}:

Input: R, $\mathbf{y}$, $\mathbf{s}$, $M = G^{-1}R^T$, MR, $\mathbf{p}^{(k-1)}$, $\mathbf{z}^{(k-1)}$.

1. If $k = m$, solve (2.14) and set $\hat{\Lambda}$ to be the optimal solution of (2.14); $\lambda_k = (\mathbf{z}^{(k-1)})^T \mathbf{z}^{(k-1)} - \frac{1}{4}(\mathbf{z}^{(k-1)})^T M^T_{1:k-1,1:k-1} M_{1:k-1,1:k-1} \mathbf{z}^{(k-1)}$.

2. If $1 < k < m$,

   2.1 $\mathbf{z}^{(k-1)} = \mathbf{z}^{(k)}_{1:k-1} - R_{1:k-1,k} s_k$, $\mathbf{p}^{(k-1)} = \mathbf{p}^{(k)}_{1:k-1} - (MR)_{1:k-1,k} s_k$.

   2.2 $\lambda_k = (\mathbf{z}^{(k-1)})^T \mathbf{z}^{(k-1)} - \frac{1}{4}(\mathbf{p}^{(k-1)})^T \mathbf{p}^{(k-1)}$.

3 If $\lambda_k \geq 0$, $LB^{(k-1)}_{sdp} = \sum_{i=1}^{k-1} \hat{\Lambda}_{i,i} + \lambda_k$, otherwise $LB_{sdp} = 0$.

Figure 2.7 compares the expected complexity of the SDsdp algorithm to the expected complexity of the standard sphere decoder algorithm (SD-algorithm). The two algorithms are employed for solving a high-dimensional binary integer least-squares problem. The signal-to-noise ratio in Figure 2.7 is defined as SNR $= 10\log_{10}\frac{m}{4\sigma^2}$, where σ^2 is the variance of each component of the noise vector $\mathbf{w}$. Both algorithms choose an initial search radius statistically as in [49] (the sequence of ϵs, $\epsilon = 0.9$, $\epsilon = 0.99$, $\epsilon = 0.999$ etc.), and update the radius every time the bottom of the tree is reached.

As can be seen from Figure 2.7 the SDsdp algorithm can run up to 10 times faster than the SD algorithm at SNR $4 - 5$ db. At higher SNR, the speedup decreases and at SNR 8 db the SD algorithm is faster. We can attribute this to the complexity of performing the original SDP (2.14). In fact, Figure 2.7, subplot 1, shows the flop count of the SD-sdp, when the computations of the SDP (2.14) are removed (denoted there as SDsdp-sdp), which can be seen to be uniformly faster than the SD. Thus, the main bottleneck is solving (2.14) and any computational improvement there can have a significant impact on our algorithm. In our numerical experiments we solved (2.14) exactly, i.e., with very high numerical precision which requires a significant computational effort. This is of course not necessary. In fact how precisely (2.14) needs to be solved is a very interesting question. For this reason we emphasize again that constructing faster SDP algorithms would significantly speed up the SDsdp algorithm.

On subplot 2 of Figure 2.7 the distribution of points per level in the search tree is shown for both SD and SDsdp algorithms. As stated in [13] and [59], in some practical realizations

the size of the tree may be as important as the overall number of multiplication and addition operations. On subplot 3 of Figure 2.7 the comparison of the total number of points kept in the tree by SD and SDsdp algorithms is shown. As expected the SDsdp algorithm keeps significantly less points in the tree than the SD algorithm.

Finally on subplot 4 of Figure 2.7, the comparison of the bit error rate (BER) performance of the exact ML detector (SDsdp algorithm) and the approximate MMSE nulling and cancelling with optimal ordering heuristic is shown. Over the range of SNRs considered here, the ML detector outperforms the MMSE detector significantly, thereby justifying our efforts in constructing more efficient ML algorithms.

Remark: Recall that the lower bound introduced in this section is valid only if the original problem is binary, i.e., $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^{k-1}$. A generalization to case $\mathcal{D} = \{-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}\}^{k-1}$ can be found in [106]. It is not difficult to generalize it to any $\mathcal{D} = \{-\frac{L-1}{2}, -\frac{L-2}{2}, \ldots, \frac{L-3}{2}, \frac{L-1}{2}\}^{k-1}$ by noting that any $k-1$-dimensional vector whose elements are numbers from $\{-L+1, -L+2, \ldots, L-2, L-1\}$ can be represented as a linear transformation of a $(k-1)(L-1)$-dimensional vector from $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^{(k-1)(L-1)}$. (The interested reader can find more on this in [69]). However, this significantly increases the dimension of the SDP problem in (2.14), which may cause our algorithm to be inefficient. Motivated by this, in the following section we consider a different framework, based on H^∞ estimation theory, which will (as we will see in Section 2.8) produce as a special case a general lower bound applicable for any $\mathcal{D}$.

2.4 H^∞-based lower bound

In this section, we derive the lower bound LB in (2.10) based on H^∞ estimation theory [90]. In estimation theory H^∞ is a concept where the goal is to minimize the worst-case energy gain from the disturbances to the estimation errors. In what follows we will try to exploit mathematical similarity between the problem at hand and the H^∞ concept.

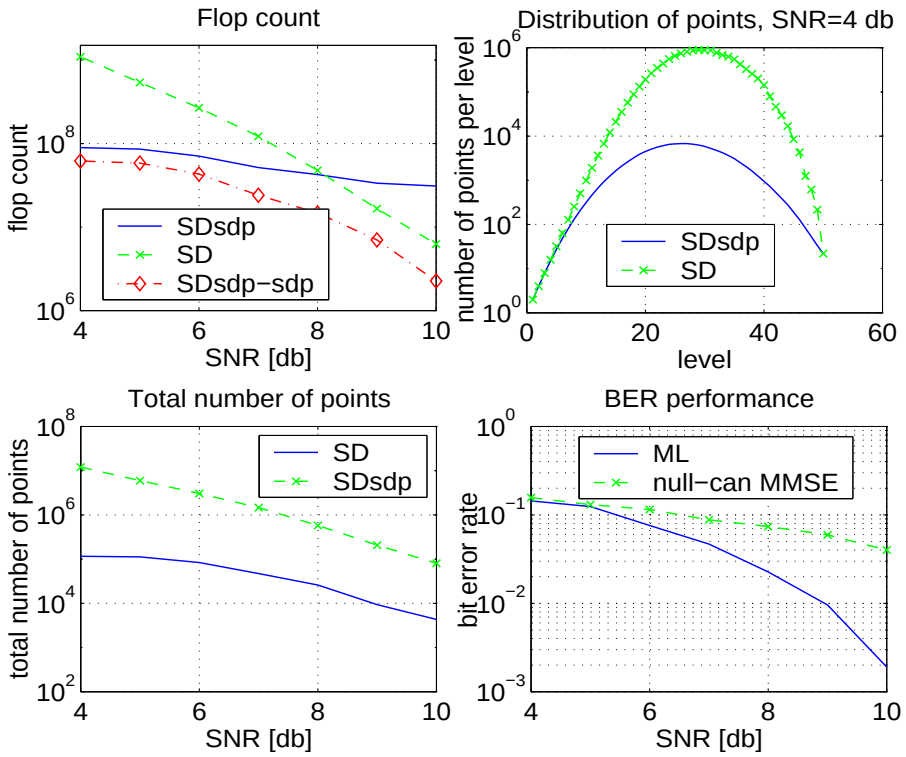


Figure 2.7: Computational complexity of the SD and SDsdp algorithms, $m = 50$, $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^{50}$

To simplify the notation, we rewrite (2.11) as

$$\min_{\mathbf{a}\in\mathcal{D}\subset\mathcal{Z}^{k-1}} \|\mathbf{b} - L\mathbf{a}\|^2, \tag{2.19}$$

where we introduced $\mathbf{a} = \mathbf{s}_{1:k-1}$, $\mathbf{b} = \mathbf{z}_{1:k-1}$, and $L = R_{1:k-1,1:k-1}$.

Consider an estimation problem where $\mathbf{a}$ and $\mathbf{b} - L\mathbf{a}$ are unknown vectors, $\mathbf{b}$ is the observation, and the quantities we want to estimate are $\mathbf{a}$ and $\mathbf{b}$. In the H^∞ framework, the goal is to construct estimators $\hat{\mathbf{a}} = f_1(\mathbf{b})$ and $\hat{\mathbf{b}} = f_2(\mathbf{b})$, such that for some given γ, some $\beta \geq 0$, and some diagonal matrix $D > 0$, we have

$$\frac{\beta||\mathbf{a} - \hat{\mathbf{a}}||^2 + ||\mathbf{b} - \hat{\mathbf{b}}||^2}{\mathbf{a}^* D\mathbf{a} + \|\mathbf{b} - L\mathbf{a}\|^2} \leq \gamma^2 \tag{2.20}$$

for *all* $\mathbf{a}$ and $\mathbf{b}$ (see, e.g., [48]).

Obtaining a desired lower bound from (2.20) is now straightforward. Note that for all

$\mathbf{a}$ and $\mathbf{b}$ we can write

$$\|\mathbf{b} - L\mathbf{a}\|^2 \geq \gamma^{-2} \left(\beta||\mathbf{a} - \hat{\mathbf{a}}||^2 + ||\mathbf{b} - \hat{\mathbf{b}}||^2\right) - \mathbf{a}^* D\mathbf{a}, \tag{2.21}$$

and, in particular,

$$\min_{\mathbf{a}\in\mathcal{D}} \|\mathbf{b} - L\mathbf{a}\|^2 \geq \min_{\mathbf{a}\in\mathcal{D}} \left(\gamma^{-2}\beta||\mathbf{a} - \hat{\mathbf{a}}||^2 - \mathbf{a}^* D\mathbf{a}\right) + \gamma^{-2}||\mathbf{b} - \hat{\mathbf{b}}||^2. \tag{2.22}$$

Note that the minimization on the right-hand side (RHS) of (2.22) is straightforward since it can be done componentwise (which is why we chose $D > 0$ diagonal). Thus, for any H^∞ estimators $\hat{\mathbf{a}} = f_1(\mathbf{b})$ and $\hat{\mathbf{b}} = f_2(\mathbf{b})$, (2.22) provides a readily computable lower bound. The issue, of course, is how to obtain the best $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ (and D and γ). To this end, let us assume that the estimators are linear, i.e., $\hat{\mathbf{a}} = K_1\mathbf{b}$ and $\hat{\mathbf{b}} = K_2\mathbf{b}$ for some matrices K_1 and K_2 (see Figure 2.8).

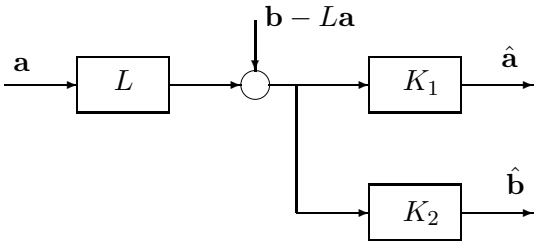


Figure 2.8: An H^∞ estimation analogy used in deriving a lower bound on integer least-squares problem.

Introducing $\mathbf{c} = \begin{bmatrix} D^{1/2}\mathbf{a} \\ \mathbf{b} - L\mathbf{a} \end{bmatrix}$ and noting that

$$T = \begin{bmatrix} D^{-1/2} & 0 \\ LD^{-1/2} & I \end{bmatrix} - \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \begin{bmatrix} LD^{-1/2} & I \end{bmatrix} = \begin{bmatrix} \sqrt{\beta}(I - K_1L)D^{-1/2} & -\sqrt{\beta}K_1 \\ (I - K_2)LD^{-1/2} & I - K_2 \end{bmatrix}$$

maps $\mathbf{c}$ to $\begin{bmatrix} \sqrt{\beta}(\mathbf{a}-\hat{\mathbf{a}}) \\ \mathbf{b}-\hat{\mathbf{b}} \end{bmatrix}$, from (2.21) we see that for all $\mathbf{c}$ it must hold that

$$\mathbf{c}^* T^* T \mathbf{c} \leq \gamma^2 \mathbf{c}^* I \mathbf{c}$$

(see [48]). Since T is square, this implies either of the equivalent inequalities

$$TT^* \leq \gamma^2 I \quad \text{or} \quad T^*T \leq \gamma^2 I. \tag{2.23}$$

The tighter the bound in (2.23), the tighter the bound in (2.22). In other words, the closer $\gamma^{-1}T$ is to a unitary matrix, the tighter (2.22) becomes. Hence we attempt to choose K_1 and K_2 to make $\gamma^{-2}TT^*$ as close to identity as possible.

To this end, post multiply T with the unitary matrix

$$\Phi = \begin{bmatrix} \nabla^{-1} & D^{-1/2}L^*\Delta^{-*} \\ -LD^{-1/2}\nabla^{-1} & \Delta^{-*} \end{bmatrix}.$$

∇ and Δ are found via the factorizations

$$D^{-1/2}L^*LD^{-1/2} + I = \nabla^*\nabla \quad \text{and} \quad LD^{-1}L^* + I = \Delta\Delta^*, \tag{2.24}$$

to obtain

$$T\Phi = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix} \tag{2.25}$$

where

$$A = \sqrt{\beta}D^{-1/2}\nabla^{-1}, B = \sqrt{\beta}(D^{-1}L^*\Delta^{-*} - K_1\Delta), \quad \text{and} \quad C = (I-K_2)\Delta. \tag{2.26}$$

Thus $TT^* \leq \gamma^2 I$ implies

$$\begin{bmatrix} AA^* + BB^* & BC^* \\ CB^* & CC^* \end{bmatrix} \leq \gamma^2 I. \tag{2.27}$$

Note that we have many degrees of freedom when choosing K_1 and K_2, and wish to make judicious choices. So, to simplify things, let us choose K_2 such that $CC^* = \gamma_1^2 I$ for some $0 \leq \gamma_1 \leq \gamma$. (Clearly, this can always be done, since from (2.24) we have that Δ is invertible, and the simple choice $K_2 = I - \gamma_1 \Delta^{-1}$ will do the job.) To make half the eigenvalues of $\gamma^{-2} TT^*$ unity, we set the Schur complement of the $(2,2)$ entry of (2.27) to zero, i.e.,

$$AA^* + BB^* - \gamma^2 I - BC^*(CC^* - \gamma^2 I)^{-1} CB^* = 0. \tag{2.28}$$

Using $CC^* = C^*C = \gamma_1^2 I$, it easily follows that

$$BB^* = (1 - \frac{\gamma_1^2}{\gamma^2})(\gamma^2 I - AA^*). \tag{2.29}$$

Using the definitions of A and B from (2.26), we obtain

$$\sqrt{\beta} K_1 = \sqrt{\beta} D^{-1} L^* (LD^{-1} L^* + I)^{-1} - B\Delta^{-1}. \tag{2.30}$$

From the (1,1) entry of (2.27) it follows that

$$\gamma^2 I - (AA^* + BB^*) \geq 0,$$

which is the only constraint on γ. Combining this constraint with the definition of A from (2.26), the definition of ∇ from (2.24), and the expression for BB^* from (2.29), we obtain that

$$\gamma^2 \geq \frac{\beta}{\lambda_{min}(D + L^* L)}.$$

We summarize the results of this section in the following theorem:

Theorem 2.1. *Consider the integer least-squares problem (2.19). Then for any* $\gamma^2 \geq \frac{\beta}{\lambda_{min}(D+L^*L)}$, $0 \leq \gamma_1 \leq \gamma$, *and any matrices* $D \geq 0$, B, *and* Δ *satisfying* $\Delta\Delta^* = I + LD^{-1}L^*$

and $BB^* = (1 - \frac{\gamma_1^2}{\gamma^2})(\gamma^2 I - \beta(D + L^*L)^{-1})$,

$$\min_{\mathbf{a}\in\mathcal{D}} \|\mathbf{b}-L\mathbf{a}\|^2 \geq \min_{\mathbf{a}\in\mathcal{D}} \gamma^{-2}||\sqrt{\beta}\mathbf{a}-\sqrt{\beta}D^{-1}L^*(LD^{-1}L^*+I)^{-1}\mathbf{b}+B\Delta^{-1}\mathbf{b}||^2-\mathbf{a}^*D\mathbf{a}+\frac{\gamma_1^2}{\gamma^2}||\Delta^{-1}\mathbf{b}||^2.$$

Proof. Follows from the previous discussion, noting that

$$||\mathbf{b} - \hat{\mathbf{b}}||^2 = ||(I - K_2)\mathbf{b}||^2 = ||C\Delta^{-1}\mathbf{b}||^2 = \gamma_1^2||\Delta^{-1}\mathbf{b}||^2$$

and

$$AA^* = \beta(D + L^*L)^{-1}.$$

□

The next corollary directly follows from Theorem 2.1.

Corollary 2.1. *Consider the setting of the Theorem 1 and let* $\beta = 1$. *Then*

$$\min_{\mathbf{a}\in\mathcal{D}} \|\mathbf{b} - L\mathbf{a}\|^2 \geq \min_{\mathbf{a}\in\mathcal{D}} \gamma^{-2}||\mathbf{a} - D^{-1}L^*(LD^{-1}L^* + I)^{-1}\mathbf{b} + B\phi||^2 - \mathbf{a}^*D\mathbf{a} + \frac{\gamma_1^2}{\gamma^2}||\phi||^2, \tag{2.31}$$

where B *is the unique symmetric square root of* $(1 - \frac{\gamma_1^2}{\gamma^2})(\gamma^2 I - (D + L^*L)^{-1})$, *and* ϕ *is* any *vector of the squared length* $\mathbf{b}^*(I + LD^{-1}L^*)^{-1}\mathbf{b}$.

□

It should be noted that we have several degrees of freedom in choosing the parameters $(\gamma_1, \gamma, D, \phi)$, and we can exploit that to tighten the bound in (2.31) as much as possible. Optimizing simultaneously over all these parameters appears to be rather difficult. However, we can simplify the problem and let $\gamma_1 \to \gamma$. This has two benefits: it maximizes the third term in (2.31) and it sets $B = 0$ so that we need not worry about the vector ϕ. Finally, to

maximize the first term, we need to take γ as its smallest possible value, i.e., we set

$$\gamma^2 = \frac{1}{\lambda_{min}(D + L^*L)}.$$

This leads to the following result:

Corollary 2.2. *Consider the setting of the Theorem 2.1 and let $\beta = 1$. Then*

$$\min_{\mathbf{a}\in\mathcal{D}} \|\mathbf{b}-L\mathbf{a}\|^2 \geq \lambda_{min}(L^*L+D)||\mathbf{a}-(L^*L+D)^{-1}L^*\mathbf{b}||^2-\mathbf{a}^*D\mathbf{a}+\mathbf{b}^*(I-L((L^*L+D)^{-1})L^*)\mathbf{b} \tag{2.32}$$

□

Remark: We would like to note that the bound given in the previous Corollary could have been also obtained in a faster way. Below we show a possible derivation that an anonymous reviewer has provided to us.

Let D be a diagonal matrix such that $D \geq 0$. Then we have

$$||\mathbf{b} - L\mathbf{a}||^2 = \mathbf{a}^*L^*L\mathbf{a} - 2\mathbf{b}^*L\mathbf{a} + \mathbf{b}^*\mathbf{b} = \mathbf{a}^*(L^*L + D)\mathbf{a} - 2\mathbf{b}^*L\mathbf{a} + \mathbf{b}^*\mathbf{b} - \mathbf{a}^*D\mathbf{a}$$
$$= (\mathbf{a}-(L^*L+D)^{-1}L^*\mathbf{b})^*(L^*L+D)(\mathbf{a}-(L^*L+D)^{-1}L^*\mathbf{b})-\mathbf{b}^*L((L^*L+D)^{-1})L^*\mathbf{b}+\mathbf{b}^*\mathbf{b}-\mathbf{a}^*D\mathbf{a}$$
$$\geq \lambda_{min}(L^*L + D)||\mathbf{a} - (L^*L + D)^{-1}L^*\mathbf{b}||^2 - \mathbf{a}^*D\mathbf{a} + \mathbf{b}^*(I - L((L^*L + D)^{-1})L^*)\mathbf{b}$$

It is not difficult to see that this is precisely the same bound as the bound given in Corollary 2.2. The interested reader can find more on this type of bound in [96] and [79].

In the following sections we show how various choices of the free parameters in the general lower bound from Theorem 2.1 yield several interesting special cases of lower bounds. In particular, in Section 2.5 we show that the lower bound obtained by solving a related convex optimization problem, where the search space is relaxed from integers to a sphere, can be deduced as a special case of the lower bound from Theorem 2.1. Then, in Section 2.6, we show that the lower bound obtained by solving another convex optimization problem, where

the search space is now relaxed from integers to a polytope, can also be deduced as a special case of the lower bound from Theorem 2.1. Finally, in Section 2.8, we use (2.32) to deduce the lower bound based on the minimum eigenvalue of L^*L.

2.5 Spherical relaxation

Assume the setting of Theorem 2.1. Let $\gamma_1 \to \gamma$, $\beta \to 0$, $D = \frac{1}{\alpha}I$, and $\Delta_{sph}\Delta_{sph}^* = \alpha LL^* + I$. Then

$$LB_{sph}^{(1)} = ||\Delta_{sph}^{-1}\mathbf{b}||^2 - \frac{k-1}{4\alpha}, \tag{2.33}$$

is a special case of the general bound given in Theorem 2.1 and, therefore, a lower bound on the integer least-squares problem (2.11). Additionally, since being a special case, it is less tight than the general bound given in Theorem 2.1. Clearly, to make $LB_{sph}^{(1)}$ as tight as possible, we should maximize (2.33) over α.

Consider the singular value decomposition (SVD) of L, $L = U\Sigma V^T$, where U and V are unitary matrices, and where Σ is diagonal matrix. Let σ_i be the i-th component on the main diagonal of Σ and let r be the rank of L. Also, let $\mathbf{q}_{1:k-1} = U^T\mathbf{z}_{1:k-1}$. Then we can write

$$LB_{sph}^{(1)} = \sum_{i=1}^{r} \frac{\alpha^{-1}\mathbf{q}_i^2}{\sigma_i^2 + \alpha^{-1}} - \alpha^{-1}\frac{k-1}{4}. \tag{2.34}$$

To maximize over α we differentiate to obtain

$$\frac{dLB_{sph}^{(1)}}{d(\alpha^{-1})} = \sum_{i=1}^{r} (\frac{\sigma_i\mathbf{q}_i}{\sigma_i^2 + \alpha^{-1}})^2 - \frac{k-1}{4}. \tag{2.35}$$

Let $\hat{\alpha}$ denote the value of α which maximizes $LB_{sph}^{(1)}$. Then it easily follows that

$$\sum_{i=1}^{r} (\frac{\sigma_i\mathbf{q}_i}{\sigma_i^2 + \hat{\alpha}^{-1}})^2 = \frac{k-1}{4}. \tag{2.36}$$

Note that if $\sum_{i=1}^{k-1}(\frac{\mathbf{q}_i}{\sigma_i})^2 - \frac{k-1}{4} \leq 0$, we set $\hat{\alpha}^{-1} = 0$. Hence, we can state a lower bound on

(2.11) as

$$\hat{LB}^{(1)}_{sph} = \|\hat{\Delta}^{-1}_{sph}\mathbf{b}\|^2 - \frac{k-1}{4\hat{\alpha}}, \tag{2.37}$$

where $\hat{\Delta}_{sph}$ is any matrix such that $\hat{\Delta}_{sph}\hat{\Delta}^*_{sph} = \hat{\alpha}LL^* + I = \hat{\alpha}R_{1:k-1,1:k-1}{R_{1:k-1,1:k-1}}^* + I$, $\mathbf{b} = \mathbf{z}_{1:k-1}$, and $\hat{\alpha}^{-1}$ is the unique solution of (2.36) if $\sum_{i=1}^{k-1}(\frac{\mathbf{q}_i}{\sigma_i})^2 - \frac{k-1}{4} > 0$, and zero otherwise.

To obtain an interpretation of the bound we have derived, let us consider a bound obtained by a simple spherical relaxation. To this end, let us denote $\hat{LB}^{(2)}_{sph} = \|\mathbf{z}_{1:k-1} - R_{1:k-1,1:k-1}\hat{\mathbf{s}}_{1:k-1}\|_2^2$, where $\hat{\mathbf{s}}_{1:k-1}$ is the solution of the following optimization problem,

$$\begin{aligned} \min_{\mathbf{s}_{1:k-1}} \quad & \|\mathbf{z}_{1:k-1} - R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1}\|_2^2 \\ \text{subject to} \quad & \sum_{i=1}^{k-1}\mathbf{s}_i^2 \leq \frac{k-1}{4}. \end{aligned} \tag{2.38}$$

This is a lower bound since the integer constraints have been relaxed to a spherical constraint that includes $\{-\frac{1}{2}, \frac{1}{2}\}^k$. The solution of (2.38) can be found via Lagrange multipliers (see. e.g., [42]), and it turns out that the optimal value of its objective function coincides with (2.37). Therefore, we conclude that

$$\hat{LB}^{(2)}_{sph} = \hat{LB}^{(1)}_{sph},$$

and the lower bound obtained via spherical relaxation is indeed a special case of the general lower bound given in Theorem 2.1.

We would also like to note that we could get a tighter bound in (2.38) if we replace inequality with an equality. Although the resulting problem would be non-convex, following the procedure from [42] and [27] we would obtain a result similar to the one obtained in (2.37). The difference would be that now in (2.36) $\hat{\alpha}^{-1}$ would be allowed to take negative values too. This would certainly give a bound which is tighter than $\hat{LB}^{(1)}_{sph}$. However, in general we didn't find that solving (2.36) for negative $\hat{\alpha}^{-1}$ would be more useful for our algorithms than solving it only for positive $\hat{\alpha}^{-1}$. Additionally, we would like to emphasize

that the bound given in (2.38) is valid for the binary case. It can, however, be used for $M > 2$ if the constraint in (2.38) is replaced by $\sum_{i=1}^{k-1} \mathbf{s}_i^2 \leq (M-1)^2 \frac{k-1}{4}$. However we believe that this type of bound is more useful in the binary case.

Now, let us unify the notation and write $LB_{sph} = \hat{LB}_{sph}^{(1)} = \hat{LB}_{sph}^{(2)}$. We employ LB_{sph} to modify the sphere decoding algorithm by substituting it in place of the lower bound in step 4 of Algorithm 1. The subroutine for computing LB_{sph} is given below.

Subroutine for computing LB_{sph}:

Input: $\mathbf{y}_{1:k-1}, R_{1:k-1,k:m}, \mathbf{s}_{k:m}, R_{1:k-1,1:k-1}$.

1. $\mathbf{z}_{1:k-1} = \mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}$.

2. Compute the SVD of $R_{1:k-1,1:k-1}$, $R_{1:k-1,1:k-1} = U\Sigma V^T, V = [\mathbf{v}_1, ..., \mathbf{v}_{k-1}]$.

3. Set $\mathbf{q}_{1:k-1} = U^T\mathbf{z}_{1:k-1}$ and $r = \text{rank}(R_{1:k-1,1:k-1})$.

4. If $\sum_{i=1}^{r}(\frac{\mathbf{q}_i}{\sigma_i})^2 > \frac{k-1}{4}$, find λ^* such that $\sum_{i=1}^{r}(\frac{\sigma_i\mathbf{q}_i}{\sigma_i^2+\lambda^*})^2 = \frac{k-1}{4}$, and compute $\hat{\mathbf{s}}_{1:k-1} = \sum_{i=1}^{r}(\frac{\sigma_i\mathbf{q}_i}{\sigma_i^2+\lambda^*})\mathbf{v}_i$ and $LB_{sph} = \sum_{i=1}^{r}(\frac{\lambda^*\mathbf{q}_i}{\sigma_i^2+\lambda^*})\mathbf{v}_i$.

5. If $\sum_{i=1}^{r}(\frac{\mathbf{q}_i}{\sigma_i})^2 \leq \frac{k-1}{4}$, set $\hat{\mathbf{s}}_{1:k-1} = \sum_{i=1}^{r}(\frac{\mathbf{q}_i}{\sigma_i})\mathbf{v}_i$ and $LB_{sph} = 0$.

The computational complexity of finding the spherical lower bound by the above subroutine is quadratic in k, and the bound needs to be computed at each point visited by Algorithm 1. That the complexity is only quadratic may not immediately seem obvious since we do need to compute the SVD of the matrix $R_{1:k-1,1:k-1}$. Fortunately, however, this operation has to be performed only once for each level of the search tree, and hence can be done in advance (i.e., before Algorithm 1 even starts). Computing the SVD of matrices $R_{1:k-1,1:k-1}, 2 \leq k \leq m$, would require performing factorizations that are cubic in k for any $2 \leq k \leq m$. However, using the results from [62] and [46], it can be shown that all m SVDs can, in fact, be computed with complexity that is cubic in m.

The computational effort required for finding LB_{sph} beyond performing the SVD is clearly quadratic in k at the $(m-k+1)$th level of the search tree. Note that, unlike

the SVD, these remaining operations do need to be performed per *each point* visited by the algorithm. In particular, computing the vector $\mathbf{q}$ requires finding $U^T\mathbf{z}_{1:k-1}$, which is quadratic in k. Now, the matrix U^T is constant at each level of the search tree, but the vector $\mathbf{z}_{1:k-1}$ differs from node to node. Clearly, this is the most significant part of the cost, and the computational complexity of finding LB_{sph} is indeed quadratic.

2.5.1 Generalized spherical relaxation

In this subsection, we propose a generalization of the spherical lower bound. This generalization is given by

$$LB_{gsph} = \begin{cases} LB_{sph} + DLB, & \text{if} \quad \|\hat{\Delta}_{sph}^{-1}\mathbf{b}\|^2 - \frac{k-1}{4\hat{\alpha}} > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2.39}$$

where, as in (2.37),

$$LB_{sph} = \|\hat{\Delta}_{sph}^{-1}\mathbf{b}\|^2 - \frac{k-1}{4\hat{\alpha}},$$

$\hat{\Delta}_{sph}\hat{\Delta}_{sph}^{*} = \hat{\alpha}LL^{*} + I$, $L = R_{1:k-1,1:k-1}$, $\mathbf{b} = \mathbf{z}_{1:k-1}$, $\mathbf{q} = U^T\mathbf{b}$, $L = U\Sigma V^T$, $\hat{\alpha}^{-1}$ is the unique solution of (2.36) if $\sum_{i=1}^{k-1}(\frac{\mathbf{q}_i}{\sigma_i})^2 - \frac{k-1}{4} > 0$, and 0 otherwise, and where

$$DLB = \min_{\mathbf{a}\in\mathcal{D}}(\frac{1}{\hat{\alpha}} + \lambda_{min}(L^*L))\|\mathbf{a} - \hat{\alpha}L^*(\hat{\alpha}LL^* + I)^{-1}\mathbf{b}\|^2. \tag{2.40}$$

Clearly, (2.39) is obtained from (2.32) by setting $D = \frac{1}{\hat{\alpha}}I$ and is, therefore, a lower bound on the integer least-squares problem (2.11). Also, since $LB_{gsph} \geq LB_{sph}$, the generalized spherical bound is tighter than the spherical bound. It is interesting to mention that LB_{gsph} was also obtained in [96] based on a different approach.

We refer to Algorithm 1 with $LB = LB_{gsph}$ as the GSPHSD algorithm. Since the generalized spherical bound is at least as tight as the spherical bound, we expect that the GSPSD algorithm prunes more points from the search tree than the SPHSD algorithm.

We give the subroutine for computing LB_{gsph} below.

Subroutine for computing LB_{gsph}:

Input: $\mathbf{y}_{1:k-1}, R_{1:k-1,k:m}, \mathbf{s}_{k:m}, R_{1:k-1,1:k-1}$

1. $\mathbf{z}_{1:k-1} = \mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}$.

2. Compute the SVD of $R_{1:k-1,1:k-1}$, $R_{1:k-1,1:k-1} = U\Sigma V^T, V = [\mathbf{v}_1, ..., \mathbf{v}_{k-1}]$.

3. Set $\mathbf{q}_{1:k-1} = U^T\mathbf{z}_{1:k-1}$ and $r = \text{rank}(R_{1:k-1,1:k-1})$.

4. If $\sum_{i=1}^{r}(\frac{\mathbf{q}_i}{\sigma_i})^2 > \frac{k-1}{4}$, find λ^* such that $\sum_{i=1}^{r}(\frac{\sigma_i\mathbf{q}_i}{\sigma_i^2+\lambda^*})^2 = \frac{k-1}{4}$, and compute $\hat{\mathbf{s}}_{1:k-1} = \sum_{i=1}^{r}(\frac{\sigma_i\mathbf{q}_i}{\sigma_i^2+\lambda^*})\mathbf{v}_i$ and $LB_{gsph} = \min_{\mathbf{a}\in\mathcal{D}}(\lambda^*+\lambda_{min}(R^*_{1:k-1,1:k-1}R_{1:k-1,1:k-1}))\|\mathbf{a}-\sum_1^r \frac{\sigma_i\mathbf{q}_i}{\sigma_i^2+\lambda^*}\mathbf{v}_i\|^2 + \sum_{i=1}^{r}(\frac{\lambda^*\mathbf{q}_i}{\sigma_i^2+\lambda^*})\mathbf{v}_i$.

5. If $\sum_{i=1}^{r}(\frac{\mathbf{q}_i}{\sigma_i})^2 \leq \frac{k-1}{4}$, set $\hat{\mathbf{s}}_{1:k-1} = \sum_{i=1}^{r}(\frac{\mathbf{q}_i}{\sigma_i})\mathbf{v}_i$ and $LB_{gsph} = 0$.

At first, the complexity of computing DLB in (2.40) may seem cubic in k; however, it can actually be reduced to quadratic. Clearly, finding the inverse of $\hat{\alpha}LL^* + I$ is of cubic complexity and required in each node of the search tree (L is constant per level but $\hat{\alpha}$ differs from node to node). However, instead of inverting the matrix $\hat{\alpha}LL^* + I$ directly, we can do it in several steps. In particular, using the SVD $L = U\Sigma V^T$, we can write

$$\hat{\alpha}L^*(\hat{\alpha}LL^* + I)^{-1}\mathbf{b} = \hat{\alpha}V\Sigma(\hat{\alpha}\Sigma^2 + I)^{-1}U\mathbf{b}.$$

Since Σ is a diagonal matrix, the inversion of $\hat{\alpha}\Sigma^2 + I$ is only linear in k. Therefore, the computationally dominant operation in finding $\hat{\alpha}V\Sigma(\hat{\alpha}\Sigma^2 + I)^{-1}U\mathbf{b}$ is multiplication of a matrix and a vector, which requires quadratic complexity. Recall what we argued earlier in this section: although the SVD decomposition of the matrix L is of cubic complexity, it can be performed off-line since L is constant on each level in the search tree. Furthermore, instead of computing separately the SVDs of all matrices $R_{1:k-1,1:k-1}, 2 \leq k \leq m$, we can employ efficient techniques from [62] and [46] to obtain all relevant matrices in these SVDs with complexity cubic in m. Therefore, computing DLB is essentially quadratic in k. Since, computing LB_{sph} is also quadratic, the computational effort required for finding LB_{gsph} in (2.39) is quadratic as well.

2.6 Polytope relaxation

In this subsection, we show that the lower bound on the integer least-squares problem (2.11) obtained by solving the related convex optimization where the search space is relaxed from integers to a polytope is yet another special case of the lower bound derived in Section 2.4.

Assume the setting of the Theorem 2.1. Let $\gamma_1 \to \gamma$, $\beta \to 0$, and $\Delta_{plt}\Delta_{plt}^* = LD^{-1}L^* + I$. Then

$$LB_{plt}^{(1)} = ||\Delta_{plt}^{-1}\mathbf{b}||^2 - \frac{\mathrm{Tr}D}{4} \tag{2.41}$$

is a special case of the general bound given in Theorem 2.1 and, therefore, a lower bound on the integer least-squares problem (2.11). Now, since the matrix D is a free parameter, we can make the bound (2.41) tighter by optimizing over D. Hence, we can obtain a lower bound to the integer least-squares problem (2.11) as

$$\hat{LB}_{plt}^{(1)} = \max_{D \geq 0} ||\Delta_{plt}^{-1}\mathbf{b}||^2 - \frac{\mathrm{Tr}D}{4}. \tag{2.42}$$

Clearly, $\hat{LB}_{plt}^{(1)}$ is also a lower bound on the integer least-squares problem (2.11). Furthermore, since (2.42) allows for any positive semi-definite diagonal matrix D, while (2.33) allows only for scaled version of identity, it is clear that the bound in (2.42) will be tighter than the one in (2.33). However, as we will see in the rest of this section, computing (2.42) is of greater complexity than computing (2.33).

Now, before further discussing and comparing the merits of the bounds defined in (2.33) and (2.42), we will show that the lower bound (2.42) is equivalent to the lower bound obtained by relaxing the search space in the integer least-squares problem (2.11) to a polytope and solving the resulting convex optimization problem. In particular, such a relaxation yields

$$\begin{aligned} \min \quad & ||\mathbf{b} - L\mathbf{d}||^2 \\ \text{subject to} \quad & -\frac{1}{2} \leq d_i \leq \frac{1}{2}. \end{aligned} \tag{2.43}$$

Let us denote $\hat{LB}_{plt}^{(2)} = ||\mathbf{b} - L\hat{\mathbf{d}}||^2$, where $\hat{\mathbf{d}}$ is a solution of (2.43). We want to show that $\hat{LB}_{plt}^{(1)} = \hat{LB}_{plt}^{(2)}$. To this end, consider the Lagrange dual of the problem (2.43),

$$\begin{aligned}
\mathcal{L}(\xi) &= ||\mathbf{b} - L\mathbf{d}||^2 + \sum_{i=1}^{k-1} \xi_i (d_i^2 - \frac{1}{4}) \\
&= \mathbf{d}^*(L^*L + \Xi)\mathbf{d} - 2\mathbf{b}^*L\mathbf{d} + \mathbf{b}^*\mathbf{b} - \frac{\mathrm{Tr}\Xi}{4} \\
&= \mathbf{d}^*\Omega\Omega^*\mathbf{d} - 2\mathbf{b}^*L\Omega^{-*}\Omega^*\mathbf{d} + \mathbf{b}^*\mathbf{b} - \frac{\mathrm{Tr}\Xi}{4} + \mathbf{b}^*L\Omega^{-*}\Omega^{-1}L^*\mathbf{b} - \mathbf{b}^*L\Omega^{-*}\Omega^{-1}L^*\mathbf{b} \\
&= (\Omega^*\mathbf{d} - \Omega^{-1}L^*\mathbf{b})^2 + \mathbf{b}^*\mathbf{b} - \mathbf{b}^*L\Omega^{-*}\Omega^{-1}L^*\mathbf{b} - \frac{\mathrm{Tr}\Xi}{4},
\end{aligned}$$

where Ω is any matrix such that $\Omega\Omega^* = L^*L + \Xi$. Using $\mathcal{L}(\xi)$, we can pose a dual problem to the primal in (2.43) as

$$\begin{aligned}
\max_{\Xi} \min_{\mathbf{d}} \quad & (\Omega^*\mathbf{d} - \Omega^{-1}L^*\mathbf{b})^2 + \mathbf{b}^*\mathbf{b} - \mathbf{b}^*L\Omega^{-*}\Omega^{-1}L^*\mathbf{b} - \frac{\mathrm{Tr}\Xi}{4} \\
\text{subject to} \quad & \Xi \geq 0, \quad \Xi \text{ is diagonal.}
\end{aligned}$$

Clearly, the previous problem is equivalent to

$$\begin{aligned}
\max_{\Xi} \quad & \mathbf{b}^*\mathbf{b} - \mathbf{b}^*L\Omega^{-*}\Omega^{-1}L^*\mathbf{b} - \frac{\mathrm{Tr}\Xi}{4} \\
\text{subject to} \quad & \Xi \geq 0, \quad \Xi \text{ is diagonal,}
\end{aligned}$$

which, after straightforward algebraic transformations involving the matrix inversion lemma, can be written as

$$\begin{aligned}
\max_{\Xi} \quad & \mathbf{b}^*(I + L\Xi^{-1}L^*)^{-1}\mathbf{b} - \frac{\mathrm{Tr}\Xi}{4} \\
\text{subject to} \quad & \Xi \geq 0, \quad \Xi \text{ is diagonal.}
\end{aligned} \tag{2.44}$$

Since the primal problem is strictly feasible, the duality gap between the problems in (2.43)

and (2.44) is zero. Therefore, if we denote the optimal solution of (2.44) by $\hat{\Xi}$,

$$\hat{LB}^{(2)}_{plt} = \mathbf{b}^*(I + L\hat{\Xi}^{-1}L^*)^{-1}\mathbf{b} - \frac{\mathrm{Tr}\hat{\Xi}}{4}. \tag{2.45}$$

Comparing (2.42) and (2.45), we conclude that

$$\hat{LB}^{(1)}_{plt} = \hat{LB}^{(2)}_{plt},$$

which implies that the bound on the integer least-squares problem (2.11) is indeed a special case of the general bound we derived in Section 2.4. To unify the notation, we write $LB_{plt} = \hat{LB}^{(1)}_{plt} = \hat{LB}^{(2)}_{plt}$. We refer to Algorithm 1 which, in step 4, makes use of LB_{plt} as the PLTSD algorithm. The subroutine for computing LB_{plt} is given below.

Subroutine for computing LB_{plt}:

Input: $\mathbf{y}_{1:k-1}, R_{1:k-1,k:m}, \mathbf{s}_{k:m}, R_{1:k-1,1:k-1}$

1. $\mathbf{z}_{1:k-1} = \mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}$

2. $LB_{plt} = quadprog(R^*_{1:k-1,1:k-1}R_{1:k-1,1:k-1}, -2R^*_{1:k-1,1:k-1}\mathbf{z}_{1:k-1}, [], [], [], [], -\frac{1}{2}, \frac{1}{2})$; (*quadprog* is the MATLAB function for solving quadratic optimization problems).

The lower bound studied in this subsection is tighter than the spherical one considered earlier. It is clear that (2.42) results in a tighter lower bound than (2.33), since (2.42) includes maximization over all diagonal positive semi-definite matrices D, whereas (2.33) assumes only the special case $D = \frac{1}{\alpha}I$. The geometric interpretation implies the difference between (2.33) and (2.42) as well. In particular, in (2.33) the set of integers from the basic problem (2.11) is relaxed to a sphere, while in (2.42) the same set of integers is relaxed to a polytope, i.e., to a smaller set. However, although the lower bound based on the polytope relaxation is tighter than the one based on the spherical relaxation, the total computational effort is not necessarily improved. The reason is the additional computational effort required to calculate the LB_{sph} per each node; these additional computations are of

quadratic complexity, while the additional operations for calculating the LB_{plt} are cubic (see, e.g., [19]). Therefore, there is no general answer to which bound is better for improving the standard sphere decoding algorithm.

2.7 Performance comparison

In this section we study and compare the performances of the SPHSD, GSPHSD, PLTSD, SDsdp, and SD algorithms.

2.7.1 Flop count

In Figure 2.9 the average flop count and the distribution of the number of visited nodes per each level of the search tree are shown for each of the SPHSD, GSPHSD, PLTSD, SD, and SDsdp algorithms. The parameters of the system are $m = 45$, $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^m$, and SNR= $10\log_{10}\frac{m}{4\sigma^2}$. The initial search radius was chosen statistically as in [49] (the sequence of ϵs, $\epsilon = 0.9$, $\epsilon = 0.99$, $\epsilon = 0.999$, etc.), and updated every time the bottom of the tree is reached. As can be seen the SPHSD, GSPHSD, PLTSD, and SDsdp prune more points than the SD algorithm. Also, as expected the PLTSD prunes more points than the SPHSD and GSPHSD since it uses a tighter lower bound. However, the large improvement in tree pruning does not always reflect in improving the overall flop count. The reason is, as we have already said, the additional amount of computation that has to be performed at each node of the search tree. For the system parameters simulated on Figure 2.9, we see that the SDsdp algorithm has the best flop count, the GSPHSD still has a better flop count than the SD algorithm, and the PLTSD and SPHSD have worse flop count than the SD algorithm.

2.7.2 Flop count histogram

In Figure 2.10 the flop count histograms of the SPHSD, GSPHSD, PLTSD, SD, and SDsdp algorithms are shown obtained from performing 560 numbers of independent runs of the algorithms. As before, the parameters of the system are $m = 45$, $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^m$, and $SNR = 3$ [dB]. It can be seen that the GSPHSD and SDsdp have significantly better

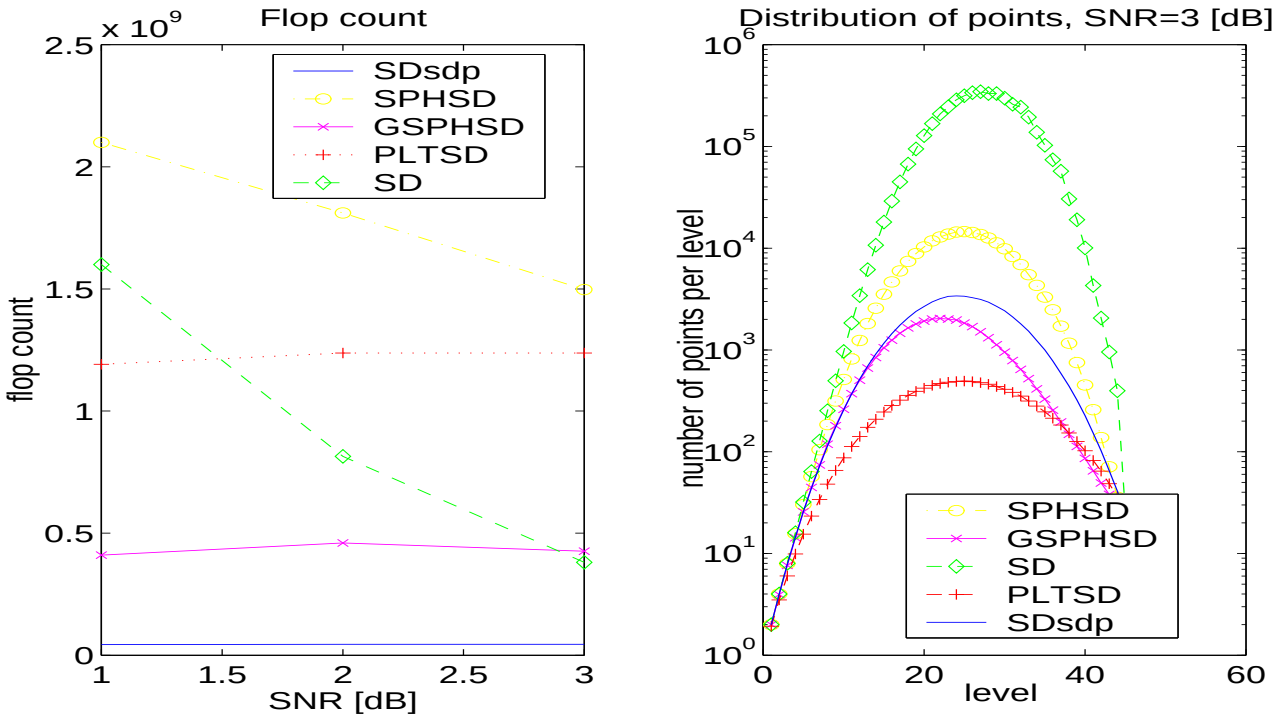


Figure 2.9: Computational complexity and the distribution of the points in the search tree for SD, SPHSD, GSPHSD, PLTSD, and SDsdp algorithms, $m = 45$, $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^{45}$

shaped (shorter tail) histograms than the SD algorithm. This implies that the probability of encountering large flop counts is significantly less. It should be noted that SPHSD and PLTSD have longer tails than the SD.

2.8 Eigen bound

In principle, the lower bound (2.32) still requires an optimization over the diagonal matrix $D \geq 0$. A particular choice that may be computationally feasible is $D = \alpha I$, for some α. However in this section we focus on an even more simple choice for D. Namely, as noted in [89], letting $D \to 0$ in Corollary 2.2 we obtain

$$LB_{eigb} = \lambda_{min}(L^* L) \min_{\mathbf{a} \in \mathcal{D}} ||\mathbf{a} - L^{-1}\mathbf{b}||^2. \tag{2.46}$$

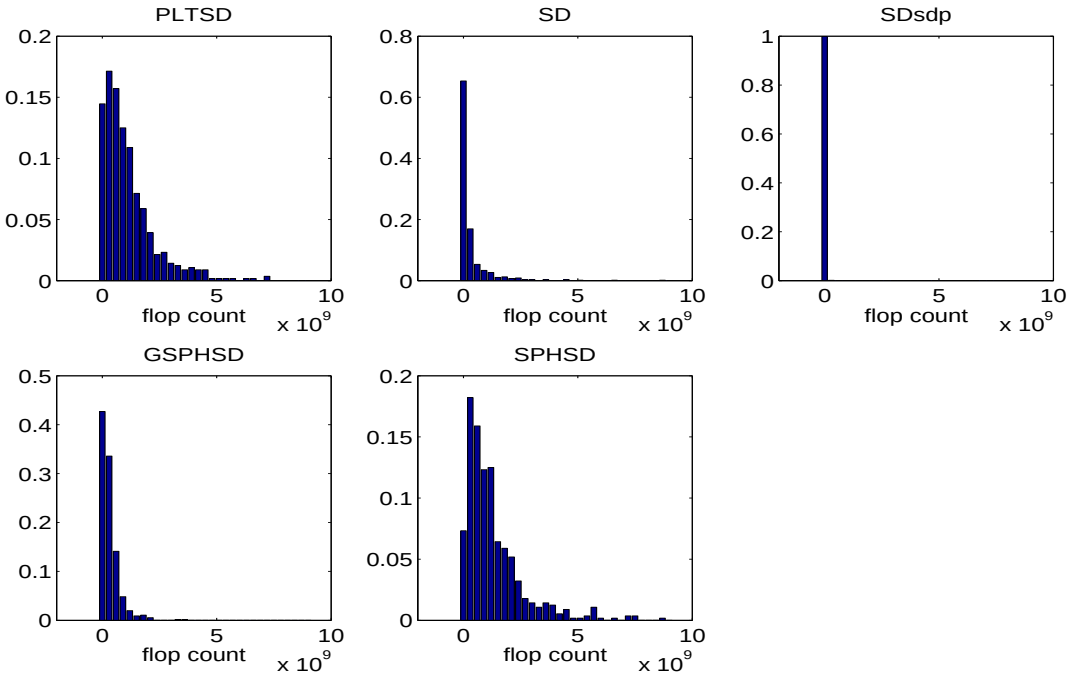


Figure 2.10: Flop count histograms for SD, SPHSD, GSPHSD, PLT, and SDsdp algorithms, $m = 45$,$SNR = 3$ [dB], $\mathcal{D} = \{-\frac{1}{2}, \frac{1}{2}\}^{45}$

We also mention that this bound could have been obtained in an easier fashion as

$$LB_{eigb} = \lambda_{min}(L^*L) \min_{\mathbf{a}\in\mathcal{D}} ||\mathbf{a} - L^{-1}\mathbf{b}||^2 \leq \min_{\mathbf{a}\in\mathcal{D}} (\mathbf{a} - L^{-1}\mathbf{b})^* \lambda_{min}(L^*L) I (\mathbf{a} - L^{-1}\mathbf{b})$$
$$\leq \min_{\mathbf{a}\in\mathcal{D}} (\mathbf{a} - L^{-1}\mathbf{b})^* (L^*L)(\mathbf{a} - L^{-1}\mathbf{b}) \leq \min_{\mathbf{a}\in\mathcal{D}} ||L\mathbf{a} - \mathbf{b}||^2. \quad (2.47)$$

Although this may raise concern that the resulting bound will be too loose, it turns out that it yields an algorithm with smaller flop count than the standard sphere decoder. The key observation is that, with $D = 0$, all the computations required at any point in the tree are linear in the dimension. (The standard sphere decoder also requires a linear number of operations per point.) Since it is based on the minimum eigenvalue we refer to this bound as the *eigen bound*.

Clearly, since (2.46) can be regarded as a special case of (2.32) it is a lower bound on the integer least-squares problem (2.11). Note that it appears as if (2.46) may not be a good bound since $\lambda_{min}(L^*L)$ could become very small. However, since the minimization in

(2.46) is performed over integers, the resulting bound turns out to be sufficiently large to serve our purposes (i.e., tree pruning in sphere decoding), especially in the case of higher symbol constellations. Furthermore, we will show that the computation required to find LB_{eigb} for a node at a level k in the search tree is linear in k.

The key observation which enables efficient computation of LB_{eigb} in (2.46) is that the vector $L^{-1}\mathbf{b}$ can be propagated as the search progresses down the tree. Before proceeding any further, we will simplify notation. First, recall that

$$L = R_{1:k-1,1:k-1} \quad \text{and} \quad \mathbf{b} = \mathbf{z}_{1:k-1} = \mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}.$$

Let us denote $F_{1:k-1,1:k-1} = {R_{1:k-1,1:k-1}}^{-1}$, and introduce

$$\mathbf{f}^{(k-1)} = L^{-1}\mathbf{b} = F_{1:k-1,1:k-1}\mathbf{z}_{1:k-1} = F_{1:k-1,1:k-1}(\mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}). \tag{2.48}$$

We wish to find a recursion that relates the vector $\mathbf{f}^{(k-2)}$ to the already calculated vector $\mathbf{f}^{(k-1)}$.

$$\begin{aligned}
\mathbf{f}^{(k-1)} &= F_{1:k-1,1:k-1}(\mathbf{y}_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m}) \\
&= \begin{bmatrix} F_{1:k-2,1:k-1} \\ F_{k-1,1:k-1} \end{bmatrix} \mathbf{y}_{1:k-1} - \begin{bmatrix} F_{1:k-2,1:k-2} & F_{1:k-2,k-1} \\ 0 & F_{k-1,k-1} \end{bmatrix} \begin{bmatrix} R_{1:k-2,k:m} \\ R_{k-1,k:m} \end{bmatrix} \mathbf{s}_{k:m} \\
&= \begin{bmatrix} F_{1:k-2,1:k-1}\mathbf{y}_{1:k-1} \\ F_{k-1,1:k-1}\mathbf{y}_{1:k-1} \end{bmatrix} - \begin{bmatrix} F_{1:k-2,1:k-2}R_{1:k-2,k:m} + F_{1:k-2,k-1}R_{k-1,k:m} \\ F_{k-1,k-1}R_{k-1,k:m} \end{bmatrix} \mathbf{s}_{k:m} \\
&= \begin{bmatrix} F_{1:k-2,1:k-2}\mathbf{y}_{1:k-2} + F_{1:k-2,k-1}y_{k-1} \\ F_{k-1,1:k-1}\mathbf{y}_{1:k-1} \end{bmatrix} \\
&- \begin{bmatrix} F_{1:k-2,1:k-2}R_{1:k-2,k:m}\mathbf{s}_{k:m} + F_{1:k-2,k-1}R_{k-1,k:m}\mathbf{s}_{k:m} \\ F_{k-1,k-1}R_{k-1,k:m}\mathbf{s}_{k:m} \end{bmatrix}
\end{aligned} \tag{2.49}$$

From (2.49), we see that

$$\mathbf{f}_{1:k-2}^{(k-1)} = F_{1:k-2,1:k-2}\mathbf{y}_{1:k-2} + F_{1:k-2,k-1}y_{k-1} - F_{1:k-2,1:k-2}R_{1:k-2,k:m}\mathbf{s}_{k:m} - F_{1:k-2,k-1}R_{k-1,k:m}\mathbf{s}_{k:m}. \tag{2.50}$$

Similarly,

$$\begin{aligned}
\mathbf{f}^{(k-2)} &= F_{1:k-2,1:k-2}\mathbf{y}_{1:k-2} - R_{1:k-2,k-1:m}\mathbf{s}_{k-1:m} \\
&= F_{1:k-2,1:k-2}\mathbf{y}_{1:k-2} - F_{1:k-2,1:k-2}\begin{bmatrix} R_{1:k-2,k-1} & R_{1:k-2,k:m} \end{bmatrix}\begin{bmatrix} s_{k-1} \\ \mathbf{s}_{k:m} \end{bmatrix} \\
&= F_{1:k-2,1:k-2}\mathbf{y}_{1:k-2} - F_{1:k-2,1:k-2}R_{1:k-2,k:m}\mathbf{s}_{k:m} - F_{1:k-2,1:k-2}R_{1:k-2,k-1}s_{k-1}.
\end{aligned} \tag{2.51}$$

Using (2.50) and (2.51), we relate $\mathbf{f}^{(k-1)}$ and $\mathbf{f}^{(k-2)}$ as

$$\mathbf{f}^{(k-2)} = \mathbf{f}_{1:k-2}^{(k-1)} + F_{1:k-2,k-1}R_{k-1,k:m}\mathbf{s}_{k:m} - F_{1:k-2,k-1}y_{k-1} - F_{1:k-2,1:k-2}R_{1:k-2,k-1}s_{k-1}. \tag{2.52}$$

All operations in the recursion (2.52) are linear, except for the matrix-vector multiplication $F_{1:k-2,1:k-2}R_{1:k-2,k-1}$, which is quadratic. However, this multiplication needs to be computed only once for each level of the tree, and the resulting term is used for computing (2.52) for all points visited by the algorithm at a level. Therefore, this multiplication may be treated as a part of pre-processing, i.e., we compute it for all k before actually running Algorithm 1. Hence, updating the vector $L^{-1}\mathbf{b}$ in the (2.46) requires a computational effort that is linear in k. Furthermore, since it is done component-wise, the minimization in (2.46) also has complexity that is linear in k. Hence we conclude that the complexity of computing the eigen bound is linear in k. Also, it should be noted that in addition to standard sphere decoder we have to compute $\lambda_k = \min \mathrm{eig}(F^*_{1:k-1,1:k-1}F_{1:k-1,1:k-1}), 1 \leq k \leq m$. However, computing these λ_ks requires an effort that is negligible to the overall flop count for the model parameters that we will consider.

We state the subroutine for computing LB_{eig} below.

Subroutine for computing LB_{eig}:

Input: R, $\mathbf{y}_{1:k-1}, \mathbf{s}_{k:m}$, $F = R^{-1}$, $\lambda_k = \min \text{eig}(F^*_{1:k-1,1:k-1} F_{1:k-1,1:k-1}), 1 \leq k \leq m$, $FR_{1:k-2,k-1} = F_{1:k-2,1:k-2} R_{1:k-2,k-1}, 1 \leq k \leq m$, $\mathbf{f}^k_{1:k-1}$.

1. If $k = m$, $\mathbf{f}^{k-1} = F_{1:k-1,1:k-1}(y_{1:k-1} - R_{1:k-1,k:m}\mathbf{s}_{k:m})$; otherwise, $\mathbf{f}^{k-1} = \mathbf{f}^k_{1:k-1} + F_{1:k-1,k} R_{k,k+1:m}\mathbf{s}_{k+1:m} - F_{1:k-1,k} y_k - FR_{1:k-1,k}\mathbf{s}_k$.

2. if $k > 1$, $LB_{eigb} = \lambda_k \min_{\mathbf{a} \in \mathcal{D}} ||\mathbf{a} - \mathbf{f}^{k-1}||^2$, otherwise, $LB_{eigb} = 0$.

We refer to the modification of the sphere decoding algorithm which makes use of the lower bound LB_{eigb} as EIGSD algorithm and study its expected computational complexity in the following subsection.

2.8.1 Eigen bound performance comparison

In this subsection we study the performance of EIGSD algorithm.

In particular, Figure 2.11 compares the expected complexity and total number of points in the tree of the EIGSD algorithm to the expected complexity and total number of points of the standard sphere decoder algorithm. We employ both algorithms for detection in a multi-antenna communication system with 6 antennas, where the components of the transmitted symbol vectors are points in a 256-QAM constellation. Note that the signal-to-noise ratio in Figure 2.11 is defined as $SNR = 10\log_{10}\frac{255m}{12\sigma^2}$, where σ^2 is the variance of each component of the noise vector $\mathbf{w}$. Both algorithms choose the initial search radius statistically as in [49] (the sequence of ϵs, $\epsilon = 0.9$, $\epsilon = 0.99$, $\epsilon = 0.999$, etc.), and employ the Schnor-Euchner search strategy, updating the radius every time the bottom of the tree is reached. As the simulation results in Figure 2.11 indicate, the EIGSD algorithm runs more than 4.5 times faster than the SD algorithm.

In Figure 2.12 the flop count histograms of SD and EIGSD algorithms are shown. As can be seen, the EIGSD algorithm has a significantly better shaped (shorter tail) distribution of the flop count than the SD algorithm.

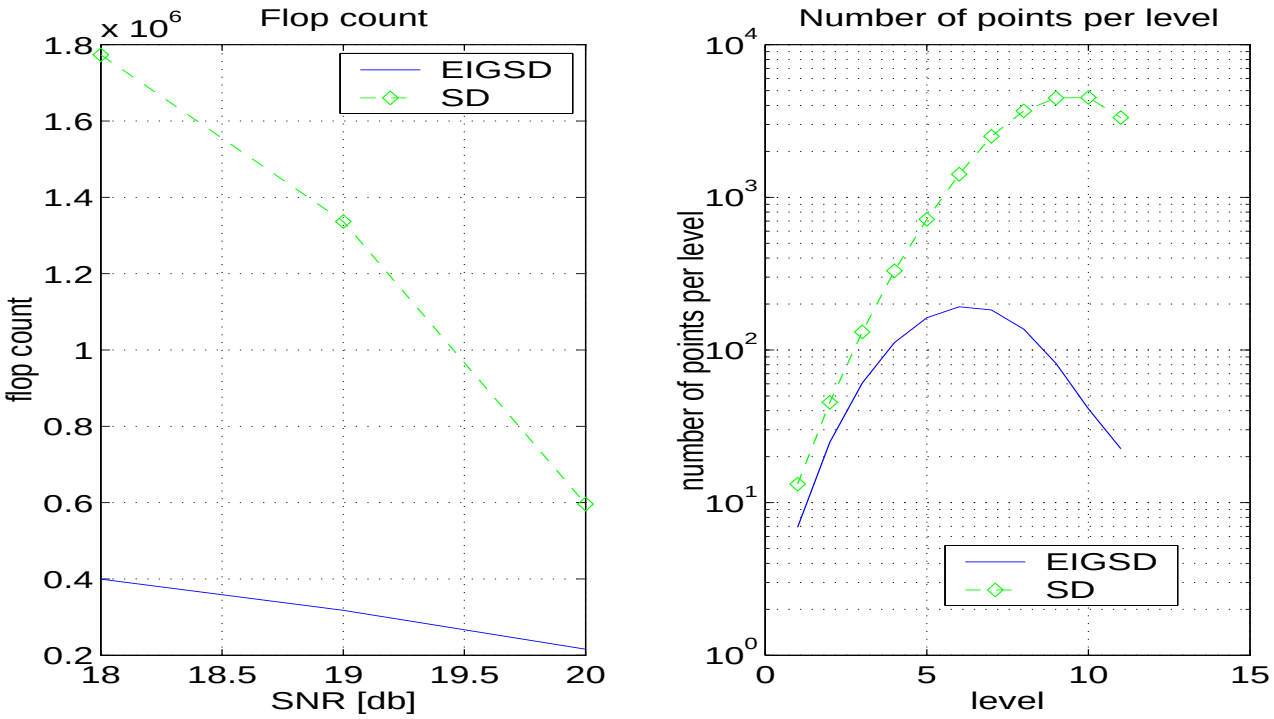


Figure 2.11: Computational complexity of the SD and EIGSD algorithms, $m = 12$, $\mathcal{D} = \{-\frac{15}{2}, -\frac{13}{2}, \ldots, \frac{13}{2}, \frac{15}{2}\}^{12}$

We would also like to point out that the EIGSD algorithm is not restricted to applications in communication systems. In Figure 2.13 we show what its potential can be if applied to a random integer least-squares problem. In the problem simulated in Figure 2.13, H was generated as an $m \times m$ matrix with i.i.d. Gaussian entries and entries of $\mathbf{y}$ were generated uniformly from the interval $[-\frac{M-10}{2}, \frac{M-10}{2}]$. The problem that we were solving was again

$$\min_{\mathbf{s}\in\mathcal{D}^m} \|\mathbf{x} - H\mathbf{s}\|_2, \tag{2.53}$$

where $\mathcal{D} = \{-\frac{M-1}{2}, -\frac{M-3}{2}, \ldots, \frac{M-3}{2}, \frac{M-1}{2}\}$. The initial radius d_g was chosen as

$$d_g = ||\mathbf{x} - H\hat{\mathbf{s}}|| \tag{2.54}$$

where $\hat{s}$ is obtained by rounding the components of $H^{-1}\mathbf{x}$ to the closest element in $\mathcal{D}$. $\hat{s}$ generated in this way is sometimes called the Babai estimate [44]. Figure 2.13 compares

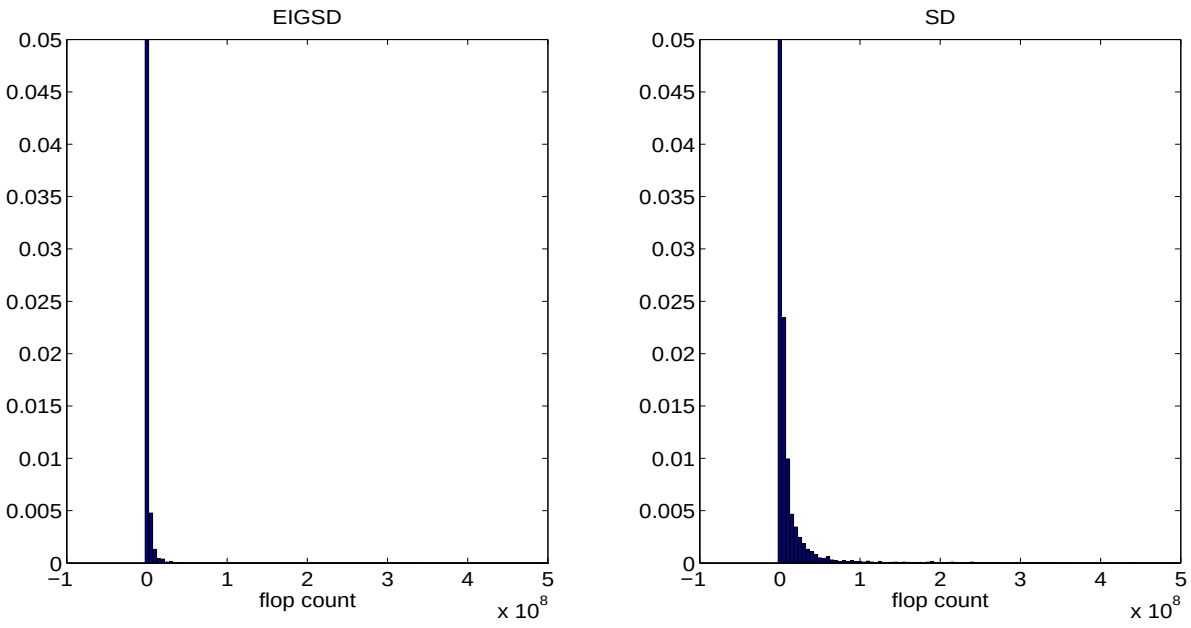


Figure 2.12: Flop count histograms for SD and EIGSD algorithms, $m = 12$, $SNR = 18$ [dB], $\mathcal{D} = \{-\frac{15}{2}, -\frac{13}{2}, \ldots, \frac{13}{2}, \frac{15}{2}\}^{12}$

the expected flop count of the EIGSD and SD algorithms for different large values of M. As can be seen, the larger the set of allowed integers the better the EIGSD performance.

2.9 Summary and discussion

In this chapter, we attempted to improve the computational complexity of sphere decoding in the regimes of low SNR and/or high dimensions by further pruning points from the search tree. The main idea is based on computing a lower bound on the remainder of the cost function as we descend down the search tree (the standard sphere decoder simply uses a lower bound of zero). If the sum of the current cost at a given node and the lower bound on the remaining cost from that node exceeds the cost of an already found solution, then that node (and all its descendants) are pruned from the search tree. In this sense, we are essentially using a "branch and bound" technique.

Adding a lower bound on the remainder of the cost function has the potential to prune the search tree significantly more than the standard sphere-decoding algorithm. However, more significant pruning of the search tree does not, in general, guarantee that the modified

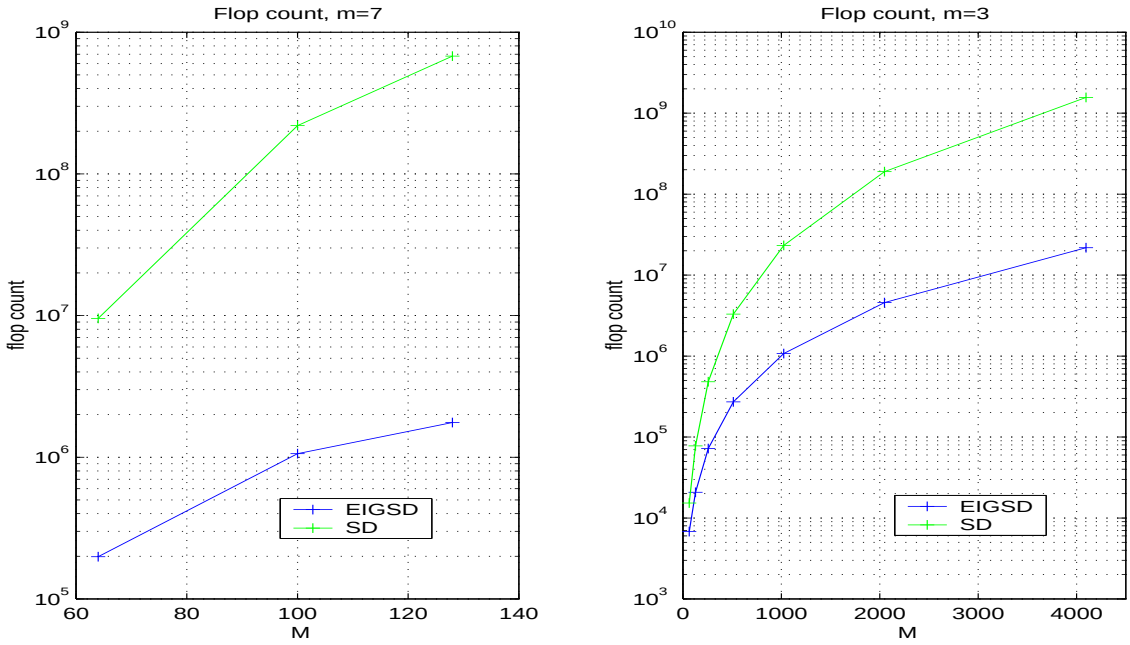


Figure 2.13: Computational complexity of the SD and EIGSD algorithms, m varies, $\mathcal{D} = \{-\frac{M-1}{2}, -\frac{M-3}{2}, \ldots, \frac{M-3}{2}, \frac{M-1}{2}\}^m$

algorithm will perform faster than the standard sphere decoding algorithm. This is due to the additional computations required by the modified algorithm to find a lower bound in each node of the search tree. Hence a natural conclusion of our work: a lower bound on one hand has to be as tight as possible in order to prune the search tree as much as possible, and on the other hand it should be efficiently computable. Led by these two main requirements, in this chapter we introduced a general framework, based on the H^∞ estimation theory, for computing the desired lower bounds. Several special cases of lower bounds were deduced from this framework. We explicitly studied four such lower bounds, and employed them for sphere decoding. The first two correspond to relaxation of the search space to either a sphere or a polytope, while the third one is a slight generalization of the spherical lower bound. The last special case corresponds to bounding the integer least-squares problem with the smallest eigenvalue and requires smaller computational effort than any of the previously mentioned bounds. In addition to H^∞ framework for computing lower bound on the integer least-squares problem, we introduced an SDP-based framework for computing desired lower bound relevant in cases when the original problem is binary.

Simulation results show that the modified sphere-decoding algorithm, incorporating the lower bound based on the smallest eigenvalue and on the SDP-duality theory, outperforms in terms of complexity the basic sphere decoding algorithm. This is not always the case with the aforementioned alternative bounds, and is due to their efficient implementation which is effectively only linear in the dimension of the problem.

Effectively all algorithms developed in this chapter can be divided in two groups depending on the type of the problem that they were designed for. The first group (SDsdp, GSPHSD, SPHSD, PLTSD) is specifically designed for binary problems, while the second group (EIGSD) is specifically designed for higher-order constellation problems. From the results that we presented, the SDsdp, GSPHSD, and EIGSD algorithms seem to outperform the standard SD in the simulated regimes in terms of flop-count. Furthermore the distributions of their flop counts have a significantly shorter tail than the distribution of the SD. However, SPHSD and PLTSD don't perform as well as the standard SD in terms of the flop count and flop count histogram. These results suggest that using a lower-bounding technique is useful, but only if the lower bound can be computed in a fast manner.

We should also point out that, although we derived it in order to improve the speed of the sphere decoding algorithm, the general lower bound on integer least-squares problems is an interesting result in itself. In fact, the proposed H^∞ estimation framework for the efficient computation of lower bounds on the difficult integer least-squares problems may find applications beyond the scope of this thesis.

The results we present indicate potentially significant improvements in the speed of the sphere decoding algorithm. However, we should note that the proposed H^∞-estimation-based framework for bounding integer least-squares problems is only partially utilized. In fact, there are several degrees of freedom in the general H^∞-based bound that are not fully exploited. It is certainly of interest to extend the current work and use the previously mentioned degrees of freedom to further tighten the lower bound. If, in addition, this can be done efficiently, it might even further improve the speed of the modified sphere decoding algorithm.

Chapter 3

Non-coherent ML Detection in Multi-Antenna Systems

In multi-antenna communication systems, channel information is often not known at the receiver. To ensure the practical feasibility of the receiver, the channel parameters are often estimated via the transmission of training symbols and then employed in the design of signal-detection algorithms. Such a scenario is possible when the environmental conditions are not changing rapidly and was considered in the previous chapter. However, in some applications, due to limited systems resources and/or rapid time variation of the channel parameters, explicitly learning the channel coefficients becomes infeasible. In this chapter we consider the problem of maximum-likelihood (ML) detection in single-input multiple-output (SIMO) systems (see Figure 3.1) when the channel information is completely unavailable at the receiver and when the signalling at the transmitter is q-PSK. It is well known that finding the solution to this optimization requires solving the integer maximization of a quadratic form which is, in general, an NP-hard problem.

In this chapter we consider solving this problem exactly and approximately. To solve it exactly we introduce the so-called out-sphere decoder algorithm, which we consider as a counterpart to the standard sphere decoder used in coherent detection and discussed in the previous chapter. In addition to developing the out-sphere decoder, we analyzed its complexity as well. Since the problem has a natural statistical setup, we considered the expected value of its complexity. We provided an explicit analytical upper bound on the expected complexity of the out-sphere decoder.

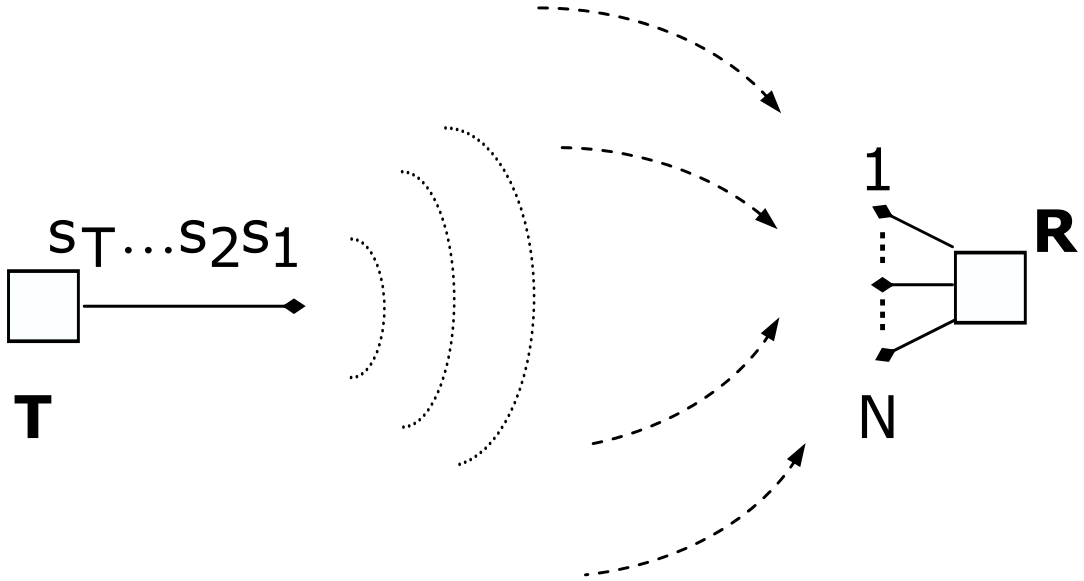


Figure 3.1: Single-input multiple-output (SIMO) system

Besides developing the exact out-sphere decoder, we propose an approximate algorithm which is based on a certain modification of a standard semi-definite program (SDP) relaxation. We derive a bound on the pairwise error probability (PEP) of the proposed algorithm and show that the algorithm achieves the same diversity as the exact maximum-likelihood (ML) detector. Furthermore, we prove that in the limit of large system dimension this bound differs from the corresponding one in the exact ML case by at most 3.92 dB if the transmitted symbols are from a 2- or 4-PSK constellation, and by at most 2.55 dB if the transmitted symbols are from an 8-PSK constellation. This suggests that the proposed algorithm requires moderate increase in the signal-to-noise ratio (SNR) in order to achieve performance comparable to that of the ML detector but with often significantly lower computational effort.

3.1 Introduction

Multi-antenna wireless communication systems are capable of providing reliable data transmission at very high rates. The channel in such systems is, in principle, unknown to the

receiver and needs to be estimated either prior to, or concurrently with, the detection of the transmitted signal. However learning channel coefficients requires time and energy, which in environments with rapidly changing conditions and limited system resources can be impractical. In this chapter we study the problem of ML detection when the channel information is unavailable at the receiver. The system that we study has a single transmit antenna and multiple receive antennas.

We assume a standard flat-fading channel model for multi-antenna systems similar to the one used in the previous section (see Figure 3.2),

$$X = \sqrt{\frac{\rho T}{M}}\mathbf{sh} + W. \tag{3.1}$$

Here T denotes the number of time intervals during which the channel remains constant, $M = 1$ is the number of the transmit antennas; N is the number of the receive antennas; ρ is the signal-to-noise ratio (SNR); X is a $T \times N$ matrix of received symbols; $\mathbf{s}$ is a $T \times 1$ transmitted symbol vector comprised of components s_i, for which it holds that $s_i = \frac{1}{\sqrt{T}} e^{\frac{j2r\pi}{q}}, r = 1, \ldots, q$, and q is an integer power of 2; $\mathbf{h}$ is an $1 \times N$ channel matrix whose components are independent, identically distributed (i.i.d.) zero-mean, unit-variance complex Gaussian random variables; and W is an $N \times T$ noise matrix whose components are i.i.d. zero-mean, unit-variance complex Gaussian random variables. Furthermore, we assume that the components of $\mathbf{h}$ and W are uncorrelated and that $T \geq N$, which is often the case in practice.

The rest of this chapter is organized as follows. In Section 3.2 we recall what the criterion for non-coherent ML-signal detection is. In Section 3.3 we propose an algorithm (which we call *out-sphere decoder*) for solving the problem of non-coherent ML detection *exactly* and analyze its expected complexity. In Section 3.4 we consider solving non-coherent ML detection problem approximately. First in section 3.4.1 we propose a trivial polynomial time algorithm that achieves full diversity. In Section 3.4.2 we introduce an SDP-based approximate algorithm for solving the ML-detection problem. In Section 3.4.3 we compute

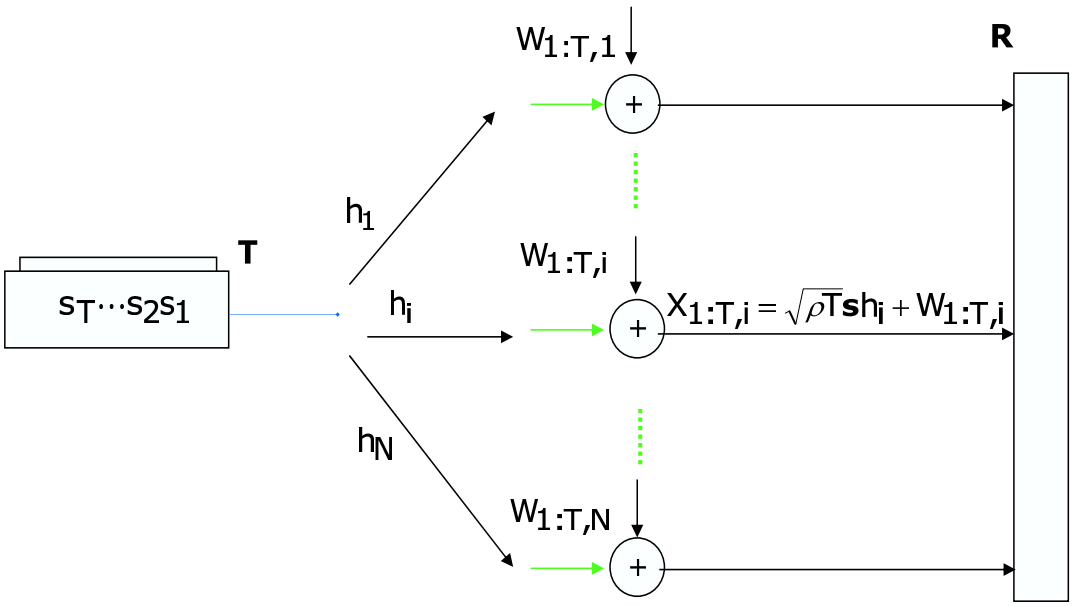


Figure 3.2: Mathematical model of SIMO system

its pairwise error probability (PEP). In Section 3.4.4 we asymptotically analyze the PEP performance in the case of large system dimensions. In Section 3.4.5 we briefly comment on the complexity of the proposed algorithms. In Section 3.5 we summarize the obtained results and suggest several possible directions for a future work.

3.2 Non-coherent ML detection

As stated in [51] the criterion for non-coherent ML-detection of the system given in (3.1) can be written as

$$\mathbf{s}_{\mathrm{ML}} = \arg\max_{\mathbf{s}\in\mathcal{S}} \frac{\exp(-\mathrm{Tr}\{[I + k\mathbf{s}\mathbf{s}^*]^{-1} XX^*\})}{\pi^{TN}\det^N[I + k\mathbf{s}\mathbf{s}^*]}, \tag{3.2}$$

where $k = \rho T$ and $\mathcal{S} = \{\frac{e^{\frac{j2\pi}{q}}}{\sqrt{T}}, \frac{e^{\frac{j4\pi}{q}}}{\sqrt{T}}, \ldots, \frac{1}{\sqrt{T}}\}^T$. Now, using the matrix inversion lemma and the fact that $\mathbf{s}^*\mathbf{s} = 1$ we obtain

$$\mathbf{s}_{\mathrm{ML}} = \arg\max_{\mathbf{s}\in\mathcal{S}} \frac{\exp(-\mathrm{Tr}\{[I - \frac{1}{k}\mathbf{s}\mathbf{s}^*]^{-1} XX^*\})}{\pi^{TN}(1+k)^N} = \arg\max_{\mathbf{s}\in\mathcal{S}} \mathrm{Tr}\{X^*\mathbf{s}\mathbf{s}^*X\}.$$

Therefore, the optimization problem one needs to solve can be written as

$$\max_{\mathbf{s}\in\mathcal{S}} \text{Tr } (XX^*\mathbf{s}\mathbf{s}^*) \tag{3.3}$$

which is a discrete optimization problem since the set $\mathcal{S}$ is discrete. (Since $\mathbf{s}^*\mathbf{s} = 1$ precisely the same optimization problem is obtained if the optimization criterion used is joint channel estimation and signal detection [87]). (3.3) is a very difficult problem, in fact NP-hard. In the following sections we will introduce several algorithms for solving it *exactly* and approximately.

3.3 Exact non-coherent ML detection

In this section we will consider solving (3.3) *exactly*. In [99] the case $q = 2$ was considered. The sphere decoder algorithm [37] was employed to solve (3.3) *exactly*. However, for some parameters of the system, the sphere decoder may be computationally costly. Here we will introduce an alternative to the standard sphere decoder and analyze its expected complexity.

Before proceeding further to facilitate the exposition we will assume that throughout this section $T = N = m$ and recall that the system model is as shown in Figure 3.3.

Then clearly the SIMO system can be modeled by the following equation

$$X = \sqrt{\rho m}\mathbf{s}\mathbf{h} + W, \tag{3.4}$$

where $\mathbf{s}$ is a $m \times 1$ transmitted symbol vector comprised of components s_i for which it holds that $s_i = \frac{1}{\sqrt{T}} e^{\frac{j2r\pi}{q}}, r = 1, \ldots, q$, and q is an integer power of 2, $\mathbf{h}$ is an $1 \times m$ channel matrix whose components are independent, identically distributed (i.i.d.) zero-mean, unit-variance complex Gaussian random variables, and W is an $m \times m$ noise matrix whose components are i.i.d. zero-mean, unit-variance complex Gaussian random variables. It is not difficult to see that (3.3) can be rewritten as

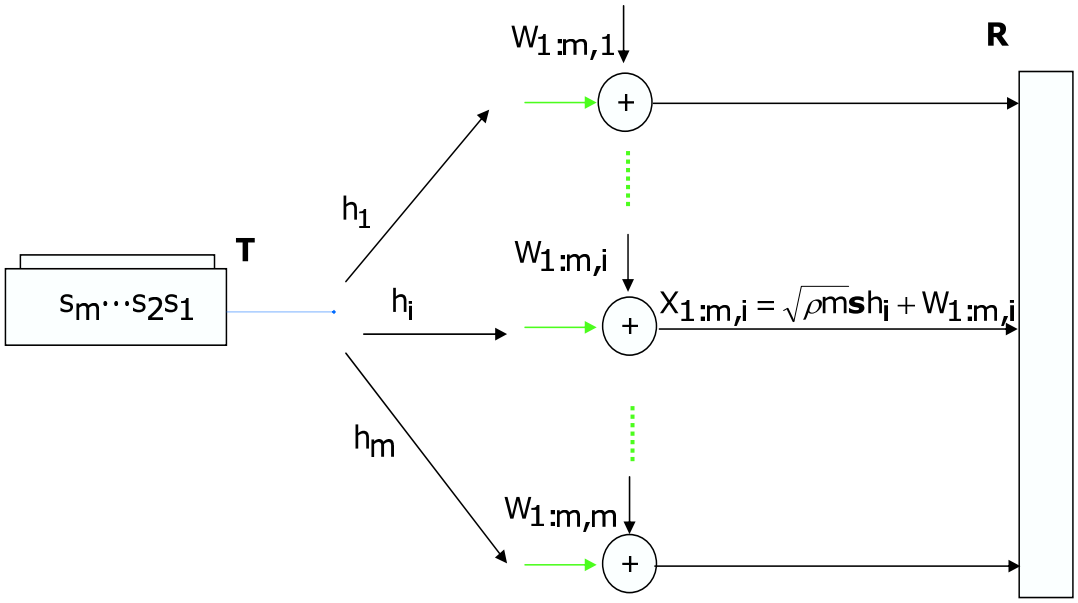


Figure 3.3: Mathematical model of SIMO system $T = N = m$

$$\max_{\mathbf{s}\in\mathcal{S}} \text{Tr}\ (XX^*\mathbf{s}\mathbf{s}^*) \tag{3.5}$$

where $\mathcal{S} = \{\frac{e^{\frac{j2\pi}{q}}}{\sqrt{m}}, \frac{e^{\frac{j4\pi}{q}}}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}\}^m$.

3.3.1 Out-sphere decoder

In this section we introduce an *exact* algorithm for solving (3.5). The main idea of the algorithm is based on finding all points $\mathbf{s}$ such that $X^*\mathbf{s}$ lies outside a sphere of some adequately chosen radius $d_s = \lambda m^2$, i.e., on finding all $\mathbf{s}$ such that

$$d_s^2 \leq \|X^*\mathbf{s}\|_2^2, \tag{3.6}$$

and then choosing the one that minimizes the objective function. Using the QR-decomposition of $X^* = QR$ (Q is unitary matrix, R is upper triangular matrix), we can reformulate (3.6) as

$$d^2 \leq \|R\mathbf{s}\|_2^2. \tag{3.7}$$

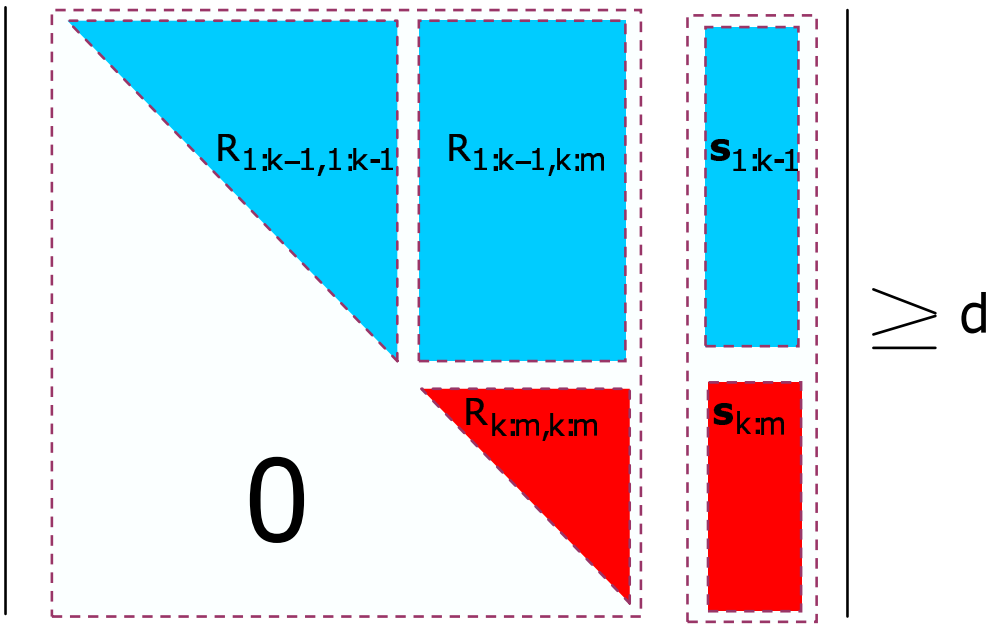


Figure 3.4: QR factorization

Although (3.7) resembles to the standard sphere decoder used for the minimization problem, it is fundamentally different. To see the main difference let us recall that in the standard sphere decoder applied for minimization of a quadratic form we have

$$\begin{aligned} d^2 &\geq \|R\mathbf{s}\|_2^2 = \|R_{m,m}\mathbf{s}_m\|^2 + \|R_{1:m-1,1:m-1}\mathbf{s}_{1:m-1} + R_{1:m-1,m}\mathbf{s}_m\|^2 \\ \Rightarrow d^2 &\geq \|R_{m,m}\mathbf{s}_m\|^2. \end{aligned}$$

However for the maximization problem in (3.7)

$$\begin{aligned} d^2 &\leq \|R\mathbf{s}\|_2^2 = \|R_{m,m}\mathbf{s}_m\|^2 + \|R_{1:m-1,1:m-1}\mathbf{s}_{1:m-1} + R_{1:m-1,m}\mathbf{s}_m\|^2 \\ \not\Rightarrow d^2 &\leq \|R_{m,m}\mathbf{s}_m\|^2. \end{aligned}$$

Therefore in order to find all $\mathbf{s}$ such that (3.7) is satisfied we need a different approach. In what follows we describe how this problem can be overcome.

Using the upper-triangular property of R (see Figure 3.4), (3.7) can be further rewritten

as

$$d^2 \leq \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2, \tag{3.8}$$

for any $2 \leq k \leq m$, where the subscripts determine the entries the various vectors and matrices run over (e.g., $R_{1:k-1,k:m}$ is a $(k-1)\times(m-k+1)$ matrix and $R_{i,k}, R_{i,k+1}, \ldots, R_{i,m}$ are the components of its i-th row). A necessary condition for (3.7) can therefore be obtained by upper-bounding the second term on the right-hand side (RHS). Let

$$UB(\mathbf{s}_{k:m}) \geq \max_{\mathbf{s}_{1:k-1}\in\mathcal{S}^{k-1}} \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2.$$

Then we have a necessary condition for (3.6)

$$d^2 \leq \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + UB(\mathbf{s}_{k:m}). \tag{3.9}$$

The sphere decoder finds all points $\mathbf{s}$ in (3.6) by proceeding inductively on (3.9), starting from $k = m$ and proceeding to $k = 1$. In other words, for $k = m$ it determines all one-dimensional lattice points $\mathbf{s}_m$ such that

$$d^2 \leq |R_{m,m}\mathbf{s}_m|^2 + UB(\mathbf{s}_m),$$

and then, for each such one-dimensional lattice point $\mathbf{s}_m$, determines all possible values for $\mathbf{s}_{m-1}$ such that

$$\begin{aligned} d^2 &\leq \|R_{m-1:m,m-1:m}\mathbf{s}_{m-1:m}\|^2 + UB(\mathbf{s}_{m-1:m}) \\ &= |R_{m,m}\mathbf{s}_m|^2 + |R_{m-1,m-1}\mathbf{s}_{m-1} + R_{m-1,m}\mathbf{s}_m|^2 + UB(\mathbf{s}_{m-1:m}). \end{aligned}$$

This gives all two-dimensional lattice points that satisfy (3.7); we proceed in a similar fashion until $k = 1$. We refer to this algorithm as *out-sphere decoder*. The out-sphere-decoder algorithm thus generates a tree (see Figure 3.5), where the branches at the $(m-k+1)$th level of the tree correspond to all $(m-k+1)$-dimensional lattice points satisfying (3.9).

Therefore, at the bottom of the tree (the m-th level) all points satisfying (3.6) are found.

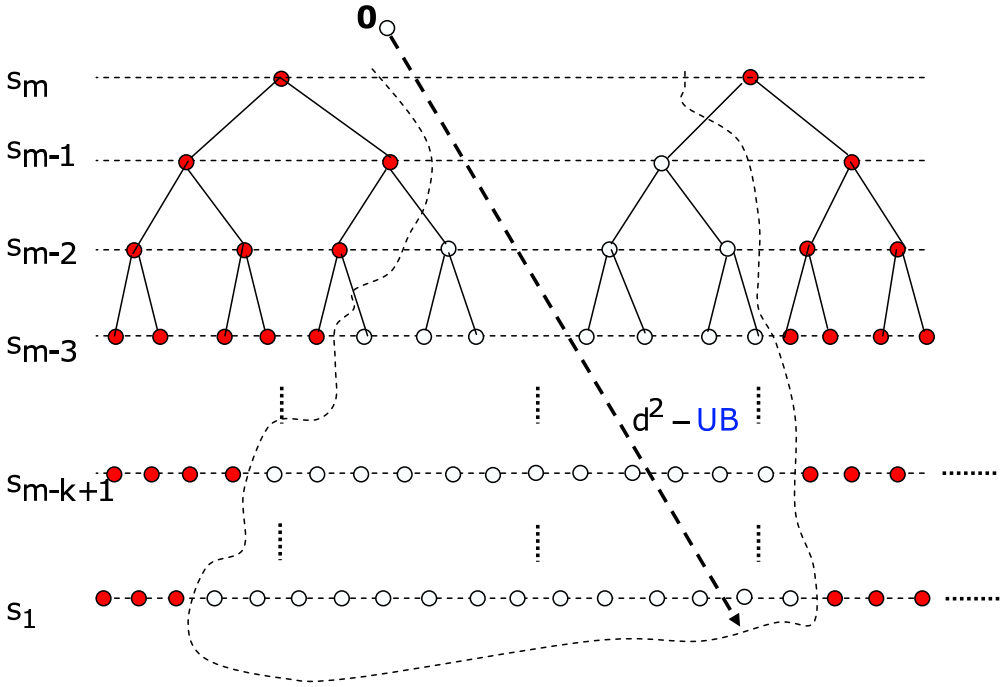


Figure 3.5: Tree search

In order to complete the algorithm we need a way of computing $UB(\mathbf{s}_{k:m})$. For this we will use well known SDP relaxation technique (the interested reader can find more on SDP relaxation in [41] and on its applications in ML detection in wireless communications in [68], [63], [69], and [57]). Let

$$\begin{aligned} UB(\mathbf{s}_{k:m}) &\geq \max_{\mathbf{s}_{1:k-1}\in\mathcal{S}^{k-1}} \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2 \\ &= \max_{\mathbf{a}_k\in\mathcal{S}^k} \mathbf{a}_k^* Q_k \mathbf{a}_k = OPT_{k-1}. \end{aligned}$$

Then SDP dual/relaxation gives

$$\begin{aligned} \max_{\mathbf{a}_k\in\mathcal{S}^k} \mathbf{a}_k^* Q_k \mathbf{a}_k \leq UB(\mathbf{s}_{k:m}) = \min_{\Lambda} \quad & \mathrm{Tr}(\Lambda) \\ \text{subject to} \quad & \Lambda \succeq Q_k, \Lambda \text{ is diagonal.} \end{aligned}$$

It can be shown

$$\alpha OPT_{k-1} \geq UB_{k-1}^{SDP} \geq OPT_{k-1}. \tag{3.10}$$

It was shown in [71] that if $Q_{k-1} \succeq 0$ is real and $q = 2$

$$\alpha_r = \frac{\pi}{2}. \tag{3.11}$$

It was shown in [84] and [114] that if $Q_{k-1} \succeq 0$ is complex

$$\alpha_c = \frac{4\pi}{(q \sin \frac{\pi}{q})^2}. \tag{3.12}$$

Using these results we will now analyze the expected complexity of the out-sphere decoder algorithm.

3.3.2 Expected complexity of the out-sphere decoder

In this section we compute an upper bound for the expected complexity of the out-sphere decoder introduced in the previous section. Effectively we will compute the probability that each point in the tree which would correspond to the exhaustive search is actually in the tree of the out-sphere decoder. To make the problem tractable we will here make the approximative assumption that the matrix X from (3.5) has i.i.d. real/complex Gaussian entries with zero-mean and unit-variance. In some sense this is an emulation of a very low-SNR regime where the matrix W should be dominant in the matrix X and where the complexity of the out-sphere decoder should be the highest. So it is reasonable to believe that in the higher-SNR regime that would be of interest in practical consideration, the complexity of the out-sphere decoder would be upper-bounded by the value computed based on the assumption that the values of matrix X from (3.5) are i.i.d. Gaussian.

3.3.2.1 The real case

In this subsection we will assume $q = 2$, and that the elements of $\mathbf{h}$ and W are i.i.d real zero-mean unit-variance Gaussian. Now it relatively easily follows that all points from the same level are equally likely to be in the search tree. Let $p_k^j, 1 \leq j \leq 2^{m-k+1}$ denote the points from the level $m-k+1$ of the search tree $\mathcal{T}$ from Figure 3.5. Further let $P_k(p_k^j \in \mathcal{T})$

be the probability that p_k^j is in the search tree $\mathcal{T}$. Clearly the expected complexity of the out-sphere decoder EC_{osd} can be computed as

$$EC_{\text{osd}} = \sum_{k=1}^{m} \sum_{j=1}^{2^{m-k+1}} P_k(p_k^j \in \mathcal{T}).$$

Since, $P_k(p_k^j \in \mathcal{T}) = P_k(p_k^i \in \mathcal{T}) = P_k(p_k \in \mathcal{T}), \text{for any} \quad i \neq j$ we have

$$EC_{\text{osd}} = \sum_{k=1}^{m} 2^{m-k+1} P_k(p_k \in \mathcal{T}). \tag{3.13}$$

Now, let us consider in particular the probability that a fixed point from level $k, 0 < k \leq m$ p_k is in the search tree. Clearly, from (3.9), (3.10), and (3.11) we have

$$\begin{aligned} P_k(p_k \in \mathcal{T}) &= Pr(d^2 \leq \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + UB(\mathbf{s}_{k:m})) \\ &\leq Pr(d^2 \leq \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \alpha_r \max_{\mathbf{s}_{1:k-1} \in \mathcal{S}^{k-1}} \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2 \\ &\leq 2^{k-1} Pr(d^2 \leq \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \alpha_r \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2). \end{aligned} \tag{3.14}$$

It is not that difficult to note that the two summands on the right-hand side of the inequality inside the probability from (3.14) are independent. Hence after applying Chernoff bound we obtain

$$P_k(p_k \in \mathcal{T}) \leq 2^{k-1} E e^{-\mu d^2} E e^{\mu \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2} E e^{\mu \alpha_r \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2} \tag{3.15}$$

where $\mu > 0$ is Chernoff parameter to be chosen later. Since

$$\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2 = \|R\mathbf{s}\|^2$$

we have

$$E e^{\mu \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \mu \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2} = E e^{\mu \|R\mathbf{s}\|^2}$$

and by independence of $\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2$ and $\|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2$ we further have

$$\begin{aligned} Ee^{\mu\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2} Ee^{\mu\|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1}+R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2} &= Ee^{\mu\|R\mathbf{s}\|^2} \\ \Rightarrow Ee^{\mu\|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1}+R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2} &= \frac{Ee^{\mu\|R\mathbf{s}\|^2}}{Ee^{\mu\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2}}. \quad (3.16) \end{aligned}$$

Plugging (3.16) into (3.15) we obtain

$$P_k(p_k \in \mathcal{T}) \leq 2^{k-1} e^{-\mu d^2} Ee^{\mu\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2} \frac{Ee^{\mu\alpha_r\|R\mathbf{s}\|^2}}{Ee^{\mu\alpha_r\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2}}. \quad (3.17)$$

It is straightforward to see that $\|R\mathbf{s}\|^2$ is chi-square distributed with m degrees of freedom and $\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2$ is chi-square distributed with $m-k+1$ degrees of freedom. Let $\beta = \frac{m-k+1}{m}$. Then we easily obtain

$$P_k(p_k \in \mathcal{T}) \leq 2^{k-1} e^{-\mu d^2} \left(\frac{(1-2\mu m\alpha_r\beta)^{\frac{\beta}{2}}}{(1-2\mu m\beta)^{\frac{\beta}{2}}(1-2\mu m\alpha_r)^{\frac{1}{2}}} \right)^m. \quad (3.18)$$

Denoting $d^2 = \lambda m^2$ and connecting (3.13) and (3.18) we have

$$EC_{\text{osd}} \leq \sum_{\beta m=1}^{m} 2^m e^{-\mu m^2\lambda} \left(\frac{(1-2\mu m\alpha_r\beta)^{\frac{\beta}{2}}}{(1-2\mu m\beta)^{\frac{\beta}{2}}(1-2\mu m\alpha_r)^{\frac{1}{2}}} \right)^m. \quad (3.19)$$

Looking at (3.19) we note that for a certain β upper bound on the expected complexity on the right-hand side will be larger than 2^β. For these βs clearly the better choice for upper bound is 2^β. However, there will be a critical β_c, such that the upper bound from (3.19) becomes smaller then the upper bound 2^β obtained from the exhaustive search. It is not that difficult to see that 2^{β_c} will be the highest number of points preserved on average at any level of the tree. Hence we have

$$EC_{\text{osd}} \leq m2^{\beta_c m}$$

where β_c is the solution of

$$2e^{-\mu m\lambda}\left(\frac{(1-2\mu m\alpha_r\beta)^{\frac{\beta}{2}}}{(1-2\mu m\beta)^{\frac{\beta}{2}}(1-2\mu m\alpha_r)^{\frac{1}{2}}}\right)=2^{\beta_c}$$

and μ is a parameter to choose so that β_c is as small as possible. Since optimization over μ appears to be rather difficult we choose $\mu=\frac{1-\frac{\alpha_r}{\lambda}}{2\alpha_r m}$. Finally we have that β_c is the solution of

$$4e^{-(\frac{\lambda}{\alpha_r}-1)}\frac{\lambda}{\alpha_r}\left(\frac{1-(1-\frac{\alpha_r}{\lambda})\beta_c}{1-(1-\frac{\alpha_r}{\lambda})\frac{\beta_c}{\alpha_r}}\right)^{\beta_c}=4^{\beta_c}.$$

We summarize the results from this subsection in the following theorem.

Theorem 3.1. *Consider the SIMO system from (3.4). Assume that components of* $\mathbf{h}$ *and* W *are i.i.d. real Gaussian with zero-mean and unit variance and that* $s_i\in\{-\frac{1}{\sqrt{m}},\frac{1}{\sqrt{m}}\}$. *Further assume that the out-sphere decoder is used for solving an ML detection problem in a SIMO system described by (3.4). Its expected complexity* EC_{osd} *(averaged over the channel and noise statistics) can be upper bounded in the following way*

$$EC_{osd}\leq mq^{\beta_c m}.$$

The constant β_c *can be obtained as solution of*

$$4e^{-(\frac{\lambda}{\alpha_r}-1)}\frac{\lambda}{\alpha_r}\left(\frac{1-(1-\frac{\alpha_r}{\lambda})\beta_c}{1-(1-\frac{\alpha_r}{\lambda})\frac{\beta_c}{\alpha_r}}\right)^{\beta_c}=4^{\beta_c}$$

where $\lambda=\frac{d_s^2}{m^2}$, d_s *is the initial radius, and* $\alpha_r=\frac{\pi}{2}$ *as given in (3.11).*

It is interesting to note that, using the replica methods from statistical physics, it was shown in [24] that

$$\lim_{m\to\infty}\frac{d_s^2}{m^2}=\lambda_{\text{opt}}=\left(\sqrt{\frac{2}{\pi}}+1\right)^2.$$

If λ in the previous theorem is chosen as λ_{opt} we have the following corollary.

Corollary 3.1. *Assume* $\lambda=\left(\sqrt{\frac{2}{\pi}}+1\right)^2$ *and* m *is large. Then* $\beta_c=\frac{2}{3}$ *and we obtain an*

upper bound on the expected complexity of the out-sphere decoder

$$EC_{osd} \leq m2^{\frac{2m}{3}} << 2^m.$$

The bound obtained in the previous corollary is still exponential. However, the exponent is 2/3 of the exponent in the exhaustive search.

3.3.2.2 The complex case

In this subsection we will assume that $s_i = \frac{1}{\sqrt{T}} e^{\frac{j2r\pi}{q}}, r = 1, \ldots, q$, q is an integer power of 2, and the elements of $\mathbf{h}$ and W are i.i.d complex zero-mean unit-variance Gaussian. It is not difficult to check that (3.13) can be rewritten as

$$EC^{\mathbf{c}}_{\text{osd}} = \sum_{k=1}^{m} q^{m-k+1} P_k(p_k \in \mathcal{T}). \tag{3.20}$$

It is also relatively easy to see that (3.14) can be rewritten as

$$P_k(p_k \in \mathcal{T}) \leq q^{k-1} Pr(d^2 \leq \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2 + \alpha_c \|R_{1:k-1,1:k-1}\mathbf{s}_{1:k-1} + R_{1:k-1,k:m}\mathbf{s}_{k:m}\|^2). \tag{3.21}$$

Following the derivation from the previous subsection we easily obtain that (3.17) has the following counterpart in the complex case

$$P_k(p_k \in \mathcal{T}) \leq q^{k-1} e^{-\mu d^2} E e^{\mu \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2} \frac{E e^{\mu \alpha_r \|R\mathbf{s}\|^2}}{E e^{\mu \alpha_r \|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2}}. \tag{3.22}$$

However, since R and $\mathbf{s}$ are complex, we have that $\|R\mathbf{s}\|^2$ is chi-square distributed with $2m$ degrees of freedom and $\|R_{k:m,k:m}\mathbf{s}_{k:m}\|^2$ is chi-square distributed with $2(m-k+1)$ degrees of freedom. Let $\beta = \frac{m-k+1}{m}$. Then we easily obtain

$$P_k(p_k \in \mathcal{T}) \leq q^{k-1} e^{-\mu d^2} \left(\frac{(1 - 2\mu m \alpha_c \beta)^{\beta}}{(1 - 2\mu m \beta)^{\beta} (1 - 2\mu m \alpha_c)} \right)^m. \tag{3.23}$$

Denoting $d^2 = \lambda m^2$ and connecting (3.20) and (3.23) we have

$$EC_{\text{osd}}^{\text{c}} \leq \sum_{\beta m=1}^{m} q^m e^{-\mu m^2 \lambda} \left(\frac{(1-2\mu m \alpha_c \beta)^\beta}{(1-2\mu m \beta)^\beta (1-2\mu m \alpha_c)} \right)^m . \tag{3.24}$$

Similarly to the previous section we find the critical value β_c as solution of

$$q e^{-\mu m \lambda} \left(\frac{(1-2\mu m \alpha_r \beta)^\beta}{(1-2\mu m \beta)^\beta (1-2\mu m \alpha_c)} \right) = q^{\beta_c}$$

and μ as a parameter to choose so that β_c is as small as possible. Since optimization over μ appears to be rather difficult, we choose $\mu = \frac{1-\frac{\alpha_c}{\lambda}}{2\alpha_r m}$. Finally we have β_c as the solution of

$$q e^{-(\frac{\lambda}{\alpha_r}-1)} \frac{\lambda}{\alpha_r} \left(\frac{1-(1-\frac{\alpha_r}{\lambda})\beta_c}{1-(1-\frac{\alpha_r}{\lambda})\frac{\beta_c}{\alpha_r}} \right)^{\beta_c} = q^{\beta_c}$$

We summarize the results from this subsection in the following theorem.

Theorem 3.2. *Consider the SIMO system from (3.4). Assume that components of* $\mathbf{h}$ *and* W *are i.i.d. complex Gaussian with zero-mean and unit variance. Additionally assume that* $\mathbf{s} \in \{\frac{e^{\frac{j2\pi}{q}}}{\sqrt{m}}, \frac{e^{\frac{j4\pi}{q}}}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}\}^m$ *and that* q *is an integer power of* 2*. Further assume that the out-sphere decoder is used for solving an ML detection problem in a SIMO system described by (3.4). Its expected complexity* EC_{osd}^{c} *(averaged over the channel and noise statistics) can be upper bounded in the following way*

$$EC_{osd}^{c} \leq m q^{\beta_c m}.$$

The constant β_c *can be obtained as solution of*

$$q e^{-(\frac{\lambda}{\alpha_r}-1)} \frac{\lambda}{\alpha_r} \left(\frac{1-(1-\frac{\alpha_r}{\lambda})\beta_c}{1-(1-\frac{\alpha_r}{\lambda})\frac{\beta_c}{\alpha_r}} \right)^{\beta_c} = q^{\beta_c}$$

where $\lambda = \frac{d_s^2}{m^2}$*,* d_s *is the initial radius, and* $\alpha_c = \frac{4\pi}{(q \sin \frac{\pi}{q})^2}$ *as given in (3.12).*

3.4 Approximate non-coherent ML detection

As we have seen in the previous section the out-sphere decoder solves the problem of non-coherent ML detection *exactly*. For large dimensions m of the problem in the real case its expected complexity is significantly smaller than the exhaustive search. However, it still remains true that the computational complexity of the out-sphere decoder is upper-bounded by an exponential function which suggests that for large m the algorithm may be difficult to implement. Hence in this section we consider the scenarios when it is not necessary that the problem of non-coherent detection is solved exactly but rather approximately. We will introduce several polynomial algorithms and analyze the quality of their approximation. As a measure of their success, the probability of error will be considered.

Effectively in this section, we focus on finding a computationally efficient approximate solution to (3.3). In particular, we will focus on solving a relaxed version of (3.3)

$$\max_{Q\geq 0, Q_{ii}=1} \text{Tr } (XX^*Q). \tag{3.25}$$

(This is a well-known semi-definite programming (SDP) relaxation, often used for obtaining approximate solutions to difficult combinatorial problems. The interested reader can find more on this relaxation in [41] and [79] and on its applications in communications in the excellent references [68], [63], [69], [1], [94], and [57]. Here we only mention the fact that solving (3.25) can be done by a host of efficient polynomial time methods [12], [109], and [79]. Furthermore, we would also like to point out that a similar relaxation was successfully introduced for the problem of blind detection in the case of the orthogonal space-time block codes [66] and in the context of coherent detection with M-PSK signalling in [67].)

3.4.1 A simple rounding algorithm

In order to gain more intuition about the problem at hand and to further understand the difficulties, in the rest of this section we introduce a simple polynomial time (in fact quadratic) approximate algorithm for solving (3.3) when $q = 2$, while leaving the detailed

analysis of (3.25) for section 3.4.2.

Now, let $\bar{Q}$ be the solution of

$$\begin{aligned} \max \quad & \text{Tr}\,(XX^*Q) \\ \text{subject to} \quad & -1 \leq Q_{ij} = Q_{ji} \leq 1, \quad 1 \leq i,j \leq T, \quad i \neq j \\ & Q_{ii} = 1, \quad 1 \leq i \leq T. \end{aligned} \tag{3.26}$$

Furthermore, let $\mathbf{r}$ be a vector with zero-mean unit-variance Gaussian i.i.d. components. Let $\hat{\mathbf{s}}_r = \text{sgn}(Q\mathbf{r})$ be the detected codeword. [Remark: in the rest of the chapter we will use the shorter term *codeword* when we refer to the transmitted/detected symbol vector $\mathbf{s}$, although our system does not assume any error correcting code.] We refer to this procedure of detecting a codeword as Algorithm Round (AR). Although this algorithm is very simple, and certainly not original, it turns out that it has the same diversity performance as the exact ML. To see this we will compute its pairwise error probability (PEP). We would like to point out, that in order to facilitate the derivations, we assume that as stated above, $q = 2$. However, we mention that our proofs can be generalized to the case of arbitrary q.

Clearly, since the objective in (3.26) is a linear function the optimum will be achieved at an extreme point of the region of optimization. This means that $\bar{Q}_{ij} \in \{-1, 1\}$ for any i, j (if it happens that the (i,j)-th component of XX^* is zero, we can set the corresponding (i,j)-th component of $\bar{Q}$ to say 1). It is not difficult to check that there are $2^{\frac{T(T-1)}{2}}$ possible candidates for $\bar{Q}$. Let $\{Q_1, Q_2, \ldots, Q_{2^{\frac{T(T-1)}{2}}}\}$ be this set of all possible candidates for $\bar{Q}$. Clearly, for one of them (say, Q_t) it is true that $Q_t = T\mathbf{s}_t\mathbf{s}_t^*$. It is also easy to check that if $\bar{Q} = Q_t$ then $\hat{\mathbf{s}}_r = \mathbf{s}_t$ (or $\hat{\mathbf{s}}_r = -\mathbf{s}_t$, which is easily resolved by a pilot symbol). Then it easily follows that the probability of Algorithm Round making an error (i.e., not detecting $\mathbf{s}_t$) is

$$P_e(\text{Algorithm Round}) = \sum_{t=1}^{2^{T(T-1)/2}} P_{AR}(\text{error}|\mathbf{s}_t\text{is sent})P(\mathbf{s}_t\text{is sent})$$

where $P_{AR}(\text{error}|\mathbf{s}_t\text{is sent})$ can be upper-bounded in the following way:

$$P_{AR}(\text{error}|\mathbf{s}_t\text{is sent}) \leq \sum_{Q_i, i\neq t} P_{AR}(Q_i\text{is detected}|\mathbf{s}_t\text{is sent}). \tag{3.27}$$

First, let us note that

$$P_{AR}(Q_i\text{is detected}|\mathbf{s}_t\text{is sent}) \leq P_{AR}(\text{Tr}(XX^*Q_i) \geq T\text{Tr}(XX^*\mathbf{s}_t\mathbf{s}_t^*)|\mathbf{s}_t\text{is sent}). \tag{3.28}$$

Now we compute an upper-bound on $P_{AR}(\text{Tr}(XX^*Q_i) \geq T\text{Tr}(XX^*\mathbf{s}_t\mathbf{s}_t^*)|\mathbf{s}_t\text{is sent})$.

Since we assume that $\mathbf{s}_t$ was transmitted, it holds that $X = \sqrt{k}\mathbf{s}_t\mathbf{h}+W$ where, as earlier, $k = \rho T$. Replacing this value for X in (3.28), we obtain

$$P_{AR}(\text{Tr}(XX^*Q_i) \geq T\text{Tr}(XX^*\mathbf{s}_t\mathbf{s}_t^*)|\mathbf{s}_t\text{is sent}) = P(\text{Tr}(\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}^* Q_n \begin{bmatrix}\mathbf{h}\\ W\end{bmatrix} \geq 0|\mathbf{s}_t \text{ is sent}), \tag{3.29}$$

where

$$\begin{aligned} Q_n &= \begin{bmatrix}\sqrt{k}\mathbf{s}_t^*\\ I\end{bmatrix}(\frac{Q_i}{T} - \mathbf{s}_t\mathbf{s}_t^*)\begin{bmatrix}\sqrt{k}\mathbf{s}_t & I\end{bmatrix} = \begin{bmatrix}\sqrt{k}\mathbf{s}_t^*\\ I\end{bmatrix}\begin{bmatrix}L_i & \mathbf{s}_t\end{bmatrix}\begin{bmatrix}D_i & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}L_i^*\\ \mathbf{s}_t^*\end{bmatrix}\begin{bmatrix}\sqrt{k}\mathbf{s}_t & I\end{bmatrix} \\ &= \begin{bmatrix}\sqrt{k}\mathbf{s}_t^*L_i & \sqrt{k}\\ \mathbf{s}_i & \mathbf{s}_t\end{bmatrix}\begin{bmatrix}D_i & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}\sqrt{k}L_i^*\mathbf{s}_t & \mathbf{s}_i^*\\ \sqrt{k} & \mathbf{s}_t^*\end{bmatrix}, \end{aligned}$$

and L_i is unitary, and D_i is a diagonal matrix such that $Q_i = TL_iD_iL_i^*$. Although it is possible to compute explicitly the probability in (3.29), we will find that it is sufficient to find its Chernoff bound. In particular,

$$\begin{aligned} P_{AR}(\text{Tr}(XX^*Q_i) \geq T\text{Tr}(XX^*\mathbf{s}_t\mathbf{s}_t^*)|\mathbf{s}_t\text{is sent}) &\leq \min_{\mu} Ee^{\mu(\text{Tr}(\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}^* Q_n \begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}))} \\ &= \int \frac{e^{-\text{Tr}(\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}^* (I-\mu Q_n)\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix})}}{\pi^N} d\mathbf{h}dW = \frac{1}{\det(I-\mu Q_n)^N} \end{aligned}$$

We first simplify the determinant in the denominator as

$$\begin{aligned}\det(I-\mu Q_n) &= \det(I-\mu\begin{bmatrix} L_i^*(k\mathbf{s}_t\mathbf{s}_t^*+1)L_i & (k+1)L_i^*\mathbf{s}_t \\ (k+1)\mathbf{s}_t^*L_i & (k+1)\end{bmatrix}\begin{bmatrix} D_i & 0 \\ 0 & -1\end{bmatrix}) \\ &= \det\begin{bmatrix} I-\mu L_i^*(k\mathbf{s}_t\mathbf{s}_t^*+1)L_iD_i & \mu(k+1)L_i^*\mathbf{s}_t \\ -\mu(k+1)\mathbf{s}_t^*L_iD_i & 1+\mu(k+1)\end{bmatrix}\end{aligned}$$

where we used the fact that $\det(I-XY)=\det(I-YX)$. Then we can further write

$$\begin{aligned}\det(I-\mu Q_n) &= (1+\mu(k+1))\det(I-\mu L_i^*(k\mathbf{s}_t\mathbf{s}_t^*+1)L_iD_i+\frac{\mu^2(k+1)^2}{1+\mu(k+1)}L_i^*k\mathbf{s}_t\mathbf{s}_t^*L_iD_i) \\ &\geq (1+\mu(k+1))\det(I-\mu I-\mu L_i^*k\mathbf{s}_t\mathbf{s}_t^*L_iD_i+\frac{\mu^2(k+1)^2}{1+\mu(k+1)}L_i^*k\mathbf{s}_t\mathbf{s}_t^*L_iD_i) \\ &= \frac{(1+\mu(k+1))}{(1-\mu)^{-\mathrm{rank}(Q_i)}}\det(I+L_i^*\mathbf{s}_t\mathbf{s}_t^*L_iD_i(\frac{-\mu k(1+\mu(k+1))+\mu^2(k+1)^2}{(1-\mu)(1+\mu(k+1))}) \\ &= \frac{(1+\mu(k+1))}{(1-\mu)^{-\mathrm{rank}(Q_i)}}\det(1+\mathbf{s}_t^*L_iD_iL_i^*\mathbf{s}_t(\frac{-\mu k(1+\mu(k+1))+\mu^2(k+1)^2}{(1-\mu)(1+\mu(k+1))}) \\ &= \frac{(1+\mu(k+1))}{(1-\mu)^{-\mathrm{rank}(Q_i)}}\det(1+\mathbf{s}_t^*Q_i\mathbf{s}_t(\frac{-\mu k(1+\mu(k+1))+\mu^2(k+1)^2}{(1-\mu)(1+\mu(k+1))}).\end{aligned}$$

The second line is true since $(I\geq D_i=L_i^*L_iD_i)\Rightarrow\det(I)\geq\det(D_i)$ and the eigenvalues (diagonal elements of TD_i) of the $T\times T$ symmetric matrix Q_i with entries from the set $\{-1,1\}$ are in the interval $[-T,T]$. After some further algebraic transformations we obtain

$$\det(I-\mu Q_n)\geq(1-\mu)^{\mathrm{rank}(Q_i)-1}(k+1)(V_r^{(it)}-1)(-\mu+\xi_r^{(1)})(-\mu+\xi_r^{(2)})$$

with

$$\begin{aligned}\xi_r^{(1)} &= \frac{V_r^{(it)}-1+\sqrt{(V_r^{(it)}-1)^2+\frac{4(1-V_r^{(it)})(k+1)}{k^2}}}{2(V_r^{(it)}-1)\frac{k+1}{k}} \\ \xi_r^{(2)} &= \frac{V_r^{(it)}-1-\sqrt{(V_r^{(it)}-1)^2+\frac{4(1-V_r^{(it)})(k+1)}{k^2}}}{2(V_r^{(it)}-1)\frac{k+1}{k}}\end{aligned}$$

and $V_r^{(it)}=\mathbf{s}_t^*Q_i\mathbf{s}_t$. Although our results will hold for any SNR, to make writing less tedious

in the rest of this section we consider only the case of large SNR. Therefore the previous results simplify to

$$P_{AR}(\mathrm{Tr}(XX^*Q_i) \geq T\mathrm{Tr}(XX^*\mathbf{s}_t\mathbf{s}_t^*)|\mathbf{s}_t\text{is sent}) \leq \frac{1}{\left(\mu(1-\mu)^{\mathrm{rank}(Q_i)}k(1-V_r^{(it)})\right)^N}. \quad (3.30)$$

In order to make the previous bound as tight as possible we minimize the right-hand side over μ. Let $\hat{\mu}$ be the optimal μ. It is not difficult to see that it holds

$$\hat{\mu} = \frac{1}{1+\mathrm{rank}(Q_i)}.$$

Choosing this value for μ, (3.30) becomes

$$P_{AR}(\mathrm{Tr}(XX^*Q_i) \geq T\mathrm{Tr}(XX^*\mathbf{s}_t\mathbf{s}_t^*)|\mathbf{s}_t\text{is sent}) \leq \frac{1}{\left(\frac{1}{\mathrm{rank}(Q_i)+1}(1-\frac{1}{\mathrm{rank}(Q_i)+1})^{\mathrm{rank}(Q_i)}k(1-V_r^{(it)})\right)^N}. \quad (3.31)$$

Replacing the previous inequality in (3.27) we finally obtain

$$\begin{aligned} P_{AR}(\text{error}|\mathbf{s}_t\text{is sent}) &\leq \sum_{Q_i, i\neq t} P_{AR}(Q_i\text{is detected}|\mathbf{s}_t\text{is sent}) \\ &\leq \sum_{Q_i, i\neq t} \frac{1}{\left(\frac{1}{\mathrm{rank}(Q_i)+1}(1-\frac{1}{\mathrm{rank}(Q_i)+1})^{\mathrm{rank}(Q_i)}k(1-V_r^{(it)})\right)^N} \\ &\leq \sum_{Q_i, i\neq t} \frac{1}{\left(\frac{1}{T+1}(1-\frac{1}{T+1})^{T}k(1-V_r^{(it)})\right)^N} \\ &\leq 2^{T(T-1)/2}\max_{i\neq t}\frac{1}{\left(\frac{1}{T+1}(1-\frac{1}{T+1})^{T}k(1-V_r^{(it)})\right)^N}. \end{aligned} \quad (3.32)$$

Recall that in the case of the exact ML detection, which requires algorithms — none of which are of polynomial complexity, we have for the same probability of error

$$P_{ML}(\text{error}|\mathbf{s}_t\text{is sent}) \leq \sum_{\mathbf{s}_i, i\neq t}\frac{1}{(k\frac{(1-V^{(it)})}{4})^N} \leq 2^T\max_{i\neq t}\frac{1}{(k\frac{(1-V^{(it)})}{4})^N} \quad (3.33)$$

where $V^{(it)} = \mathbf{s}_i^* \mathbf{s}_t \mathbf{s}_t^* \mathbf{s}_i$.

We summarize the previous results in the following theorem.

Theorem 3.3. *Consider the problem of non-coherent ML detection for a SIMO system described in (3.1). Assume that the codeword* $\mathbf{s}_t$ *was transmitted. Then the probability that an error occurred if AR algorithm was applied to solve (3.3), can be upper bounded in the following way*

$$P_{AR}(error|\mathbf{s}_t is\ sent) \leq \sum_{Q_i, i \neq t} \frac{1}{\left(\frac{1}{T+1}(1-\frac{1}{T+1})^T \rho T (1 - V_r^{(it)})\right)^N}.$$

Proof. Follows from the previous discussion. □

The performances of the ML and the AR algorithms are shown in Figure 3.6. In the simulated system we chose $T = N = 10$ and 4- PSK, i.e., $q = 4$. As can be seen, the AR algorithm indeed has the *same* diversity as the *exact* ML algorithm. As expected, since the AR algorithm is only an approximation, it has a coding loss. According to Figure 3.6 this coding loss is roughly 1 dB for the simulated system.

In addition to the AR and the *exact* ML, the performance of a simple heuristic to which we refer as MRC is shown on Figure 3.6. The MRC heuristic is based on the use of a training symbol to first estimate the channel. As mentioned earlier, in order to avoid the sign ambiguities when solving (3.3) one of the components of vector $\mathbf{s}_t$ (say, the first) has to be known at the receiver. Based on that, the receiver can form an estimate of $\mathbf{h}$ as the first row of the matrix X. Then the rest of the components of the vector $\mathbf{s}$ can be determined according to the maximal combining ratio (MRC) rule.

The previous MRC heuristic is relatively simple. We denote a solution that it outputs as $\mathbf{s}_{\text{MRC}}$ and summarize it below.

MRC-algorithm (given for $q = 2$*; for* $q \geq 4$ *it can be defined analogously):*

Input: X, $q = 2$, $X_{i,1:N}$-the i-th row of X.

1. Let $\hat{\mathbf{h}} = X_{i,1:N}$.

2. $\mathbf{s}_{\mathrm{MRC}_i} = \mathrm{sign}(\hat{\mathbf{h}} X^*_{i,1:N})$.

In some sense this is a simplified version of the AR algorithm. The MRC algorithm differs from the AR algorithm in the fact that it does not compute the entire matrix XX^*, but rather only its first row. Therefore, the MRC is a faster algorithm than the AR. However, as Figure 3.6 suggests, it does not perform as well as the AR or the *exact* ML algorithm.

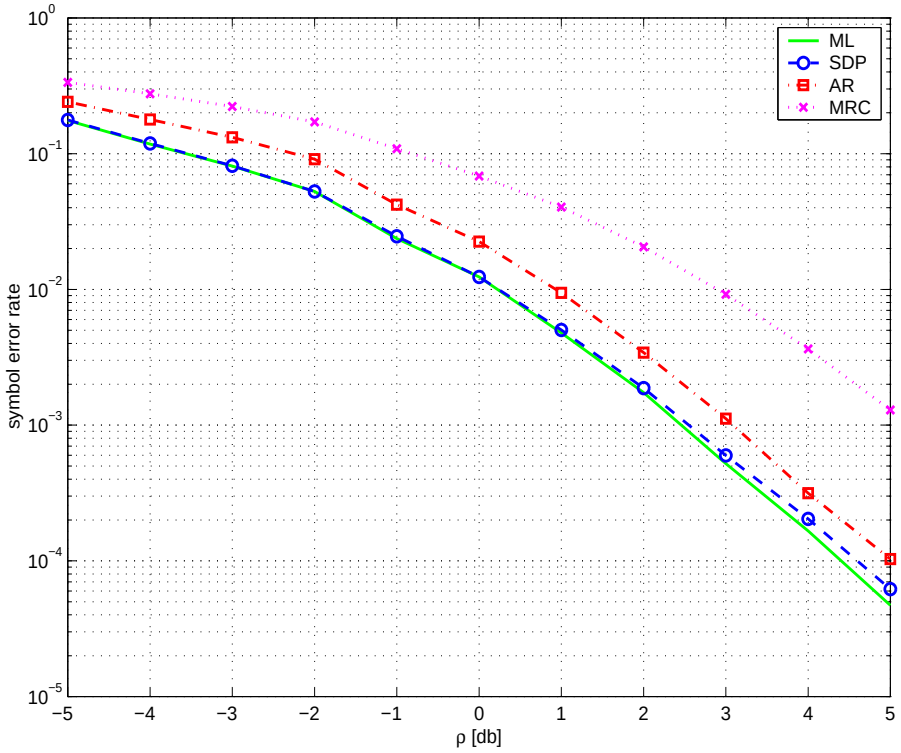


Figure 3.6: Comparison of symbol error rate, AR, ML, MRC, and SDP q=4, T=N=10

Clearly, comparing (3.32) and (3.33), it follows that the AR algorithm has the *same* diversity (the exponent of the SNR ρ) as the exact ML algorithm. Of course, since the AR algorithm is only an approximation, the exact ML algorithm still has significant advantage in the coding gain. This explains why the performance (symbol error rate) of the AR algorithm is somewhat worse than the performance of the exact ML. In order to bridge this gap in practical performance, in the following section we introduce the well known SDP relaxation to construct an algorithm whose theoretical performance (in terms of diversity)

will match those of the AR and the exact ML. However, unlike AR, new SDP-relaxation-based algorithm will have significantly smaller coding loss, which will reflect on a symbol error performance almost identical to the one of the exact ML. In fact we can give a proven bound on the coding loss.

3.4.2 SDP relaxation

In this subsection we return to the main topic of this section, namely to the analysis of an SDP-based algorithm for solving the ML-detection problem in the non-coherent case. More on this subject can be found in [86].

Recall that in a SIMO system when the channel information is not available at the receiver ML detection is equivalent to solving the following problem

$$\max_{\mathbf{s}\in\mathcal{S}} \text{Tr } (XX^*\mathbf{s}\mathbf{s}^*). \tag{3.34}$$

As we have already mentioned in the previous section, we will be interested in the analysis of

$$\max_{Q\geq 0, Q_{ii}=1} \text{Tr } (XX^*Q), \tag{3.35}$$

which is the well-known SDP relaxation of (3.34). Let $\hat{Q}$ and $\mathbf{s}_{\text{ML}}$ denote the solutions to (3.35) and (3.34), respectively. Then since (3.35) is a relaxation of (3.34) it holds that

$$\text{Tr } (XX^*\hat{Q}) \geq T\text{Tr}(XX^*\mathbf{s}_{\text{ML}}\mathbf{s}_{\text{ML}}^*). \tag{3.36}$$

What is a bit more interesting is that it can be shown (see [71], [84], and [114]) that

$$\alpha\text{Tr}(XX^*\hat{Q}) \leq T\text{Tr}(XX^*\mathbf{s}_{\text{ML}}\mathbf{s}_{\text{ML}}^*), \tag{3.37}$$

where α is a constant. More precisely, if $q = 2$ then as shown in [71] $\alpha = \frac{2}{\pi}$, and if $q \geq 4$ then as shown in [84] and [114] $\alpha = \frac{(q\sin(\pi/q))^2}{4\pi}$. Clearly, for any value $q \geq 2$ we have that $\frac{2}{\pi} \leq \alpha \leq \frac{\pi}{4}$.

Now, let L be any lowest rank matrix such that $LL^* = \hat{Q}$, and $\mathbf{r}$ be a vector with zero-mean unit-variance complex Gaussian i.i.d. components. Let ϕ be vector of phases of components of $L\mathbf{r}$. If $\frac{2\pi m-\pi}{q} \leq \phi_i < \frac{2\pi m+\pi}{q}, 1 \leq i \leq T$, m is an integer, then $\hat{\phi}_i = \frac{2\pi m}{q}$. Finally let

$$\hat{\mathbf{s}} = \frac{e^{j\hat{\phi}}}{\sqrt{T}}. \tag{3.38}$$

Then one can write (see [71], [84], and [114])

$$\alpha \mathrm{Tr}\ (XX^*\mathbf{s}_{\mathrm{ML}}\mathbf{s}_{\mathrm{ML}}^*) \leq E_{|r}\mathrm{Tr}(XX^*\hat{\mathbf{s}}\hat{\mathbf{s}}^*). \tag{3.39}$$

To make the connection between (3.37) and (3.39) clearer, we will repeat several of the most critical steps used in their derivation in the case when $q = 2$ in [71]. Let $G = XX^*$. Then the following facts follow from [71]

$$\begin{aligned} T\mathrm{Tr}\ (XX^*\mathbf{s}_{\mathrm{ML}}\mathbf{s}_{\mathrm{ML}}^*) &\geq TE_{|r}\mathrm{Tr}(XX^*\hat{\mathbf{s}}\hat{\mathbf{s}}^*) = TE_{|r}\mathrm{Tr}(G\hat{\mathbf{s}}\hat{\mathbf{s}}^*) \\ &= TE_{|r}\sum_{i,j} G_{ij}\hat{s}_i\hat{s}_j^* = \sum_{i,j} G_{ij}E_{|r}(T\hat{s}_i\hat{s}_j^*) \\ &= \frac{2}{\pi}\sum_{i,j} G_{ij}\arcsin(\hat{Q}_{ij}) = \frac{2}{\pi}\sum_{i,j} G_{ij}\arcsin(\hat{Q}) \end{aligned} \tag{3.40}$$

where $\arcsin(\hat{Q})$ is a matrix whose i,j-th component is $\arcsin(\hat{Q}_{ij})$. Nesterov then in [71] continued and proved the main point

$$\arcsin(\hat{Q}) \geq \hat{Q}. \tag{3.41}$$

Combining (3.36), (3.40), and (3.41) we finally have

$$\begin{aligned} T\mathrm{Tr}\ (XX^*\mathbf{s}_{\mathrm{ML}}\mathbf{s}_{\mathrm{ML}}^*) &\geq TE_{|r}\mathrm{Tr}(XX^*\hat{\mathbf{s}}\hat{\mathbf{s}}^*) = \frac{2}{\pi}\mathrm{Tr}(G\arcsin(\hat{Q})) \\ &\geq \frac{2}{\pi}\mathrm{Tr}(G\hat{Q}) \geq \frac{2}{\pi}T\mathrm{Tr}(G\mathbf{s}_{\mathrm{ML}}\mathbf{s}_{\mathrm{ML}}^*) = \frac{2}{\pi}T\mathrm{Tr}(XX^*\mathbf{s}_{\mathrm{ML}}\mathbf{s}_{\mathrm{ML}}^*). \end{aligned} \tag{3.42}$$

Now, (3.37) and (3.39) easily follow from (3.42).

Effectively (3.39) states that one can construct a suboptimal solution to (3.34) which has a guaranteed performance. Of course, strictly speaking, the performance is guaranteed only in the expected sense. However, if we repeat the randomized procedure a sufficient number of times, we are very likely to obtain an instance with a cost whose value is greater than the true expectation. In fact, it was shown in [34] that, with certain modifications, the expectation in (3.39) can indeed be omitted. Hence, there is a polynomial time algorithm which provides a suboptimal solution to (3.34), $\hat{\mathbf{s}}$, such that

$$\alpha \mathrm{Tr}\ (XX^*\mathbf{s}_{\mathrm{ML}}\mathbf{s}_{\mathrm{ML}}^*) \leq \mathrm{Tr}\ (XX^*\hat{\mathbf{s}}\hat{\mathbf{s}}^*). \tag{3.43}$$

We refer to the previous procedure of generating $\hat{\mathbf{s}}$ as SDP algorithm and give its explicit steps below.

SDP algorithm:

Input: X, q, count $= 0$, Obj $= 0$, $\bar{\phi} = 0$.

1. Solve (3.25). Let $\hat{Q}$ be the optimal solution.

2. Find the lowest rank L such that $LL^* = \hat{Q}$. Let r_Q be the rank of $\hat{Q}$.

3. count $=$ count $+ 1$

   3.1. Generate $r_Q \times 1$ vectors $\mathbf{rr}$ and $\mathbf{ir}$ with i.i.d. zero-mean unit-variance Gaussian components. If $q = 2$ then $\mathbf{r} = \mathbf{rr}$, else $\mathbf{r} = \frac{1}{\sqrt{2}}(\mathbf{rr} + \mathbf{ir}\sqrt{-1})$.

   3.2. Let ϕ be the vector of phases of components of $L\mathbf{r}$. If $\frac{2\pi m - \pi}{q} \leq \phi_i < \frac{2\pi m + \pi}{q}, 1 \leq i \leq T$, m is an integer, then $\hat{\phi}_i = \frac{2\pi m}{q}$.

   3.3. If Obj $\leq \mathrm{Tr}(XX^*\hat{\phi}\hat{\phi}^*)$ then Obj $= \mathrm{Tr}(XX^*\hat{\phi}\hat{\phi}^*)$ and $\bar{\phi} = \hat{\phi}$.

   3.4. count $=$ count $+ 1$.

4. If count ≤ 10 go to 3.

5. Let $\hat{\phi} = \bar{\phi}$, $\hat{\mathbf{s}} = \frac{e^{j\hat{\phi}}}{\sqrt{T}}$.

The previously described SDP algorithm is obtained based on the relaxation (3.25) of the original ML detection problem (3.3). Recall that the AR algorithm described in the previous section was based on the relaxation (3.26) of the same original ML detection problem (3.3). Assume that the AR has smaller probability of error than SDP. Then it has to happen that AR sometimes finds solution while SDP does not. If AR finds a solution then from (5) we have that $\bar{Q} = \mathbf{s}_t\mathbf{s}_t^*$. However, this solution is also admissible in (3.25), i.e., the SDP can find it too.

What can happen is that there may be some other Q producing the same value of the objective in SDP in (3.25) which may not be admissible in AR in (3.26). However, since the objective is linear, this can happen only if some of elements of the matrix XX^* are zero. Given that the problem is of statistical nature and that all random variables are continuous the probability of this event is zero. Therefore the probability of error of the AR algorithm can not be lower than the probability of error of the SDP algorithm.

Of course, it is possible to go around this point by slightly modifying the original SDP depending on whether XX^* has zeros or not. The modification would be that after solving (3.25) SDP fixes the positions of the matrix Q, which correspond to the positions of zeros in XX^* the same way the AR does.

Lemma 3.1. *Consider the problem of non-coherent ML detection for a SIMO system described in (3.1) in the high-SNR regime. Assume that the codeword $\mathbf{s}_t$ was transmitted. Then the probability that an error occurred if the SDP algorithm was applied to solve (3.3), $P_{SDP}(error|\mathbf{s}_t is\ sent)$, can be upper bounded in the following way*

$$P_{SDP}(error|\mathbf{s}_t is\ sent) \leq P_{AR}(error|\mathbf{s}_t is\ sent) \leq \frac{const.}{\rho^N}.$$

Proof. Follows from the previous discussion. □

From the previous Lemma it is clear that the SDP algorithm will achieve the same diversity as the *exact* ML and the AR algorithm.

The performance of the SDP algorithm is shown in Figure 3.6. As we have already said, in the simulated system we chose $T = N = 10$ and 4-PSK (i.e. $q = 4$). As can be seen from Figure 3.6, the SDP algorithm indeed has the *same* diversity as the *exact* ML algorithm. Also, as expected, since the SDP-algorithm is only an approximation, it has a coding loss. However, according to Figure 3.6 this coding loss is almost negligible compared to the coding loss from which the AR and the MRC algorithms suffer. This in fact is an interesting point. It effectively states that if we have available limited computational resources at the receiver then a simple AR algorithm which is much faster than the SDP can be implemented while guaranteeing full diversity with a small coding loss. However, if computational resources at the receiver are somewhat larger, then Figure 3.6 suggests that the coding loss can in fact be almost completely eliminated using the SDP algorithm.

Although the coding loss which the SDP algorithm suffers is negligible, we were not able to quantify it explicitly. In order to do that we now introduce a slight modification of the SDP algorithm. Assume that $\hat{\mathbf{s}}$ is the solution of SDP algorithm. Then let

$$\bar{\mathbf{s}} = \arg\ \max_{\mathbf{s},|\mathbf{s}^*\hat{\mathbf{s}}|^2\geq\alpha} \mathrm{Tr} X X \mathbf{s}\mathbf{s}^* \tag{3.44}$$

where α is as defined earlier. We refer to the algorithm whose solution is $\bar{\mathbf{s}}$ as SDPLS algorithm (shortened for SDP algorithm+Limited Search) and give its explicit steps below

SDPLS algorithm:

Input: X, q, count $= 0$, Obj $= 0$, $\bar{\phi} = 0$.

1. Solve (3.25). Let $\hat{Q}$ be the optimal solution.

2. Find the lowest rank L such that $LL^* = \hat{Q}$. Let r_Q be the rank of $\hat{Q}$.

3. count $=$ count $+ 1$

   3.1. Generate $r_Q \times 1$ vectors $\mathbf{rr}$ and $\mathbf{ir}$ with i.i.d. zero-mean unit-variance Gaussian components. If $q = 2$ then $\mathbf{r} = \mathbf{rr}$, else $\mathbf{r} = \frac{1}{\sqrt{2}}(\mathbf{rr} + \mathbf{ir}\sqrt{-1})$.

3.2. Let ϕ be vector of phases of components of $L\mathbf{r}$. If $\frac{2\pi m-\pi}{q} \leq \phi_i < \frac{2\pi m+\pi}{q}, 1 \leq i \leq T$, m is an integer, then $\hat{\phi}_i = \frac{2\pi m}{q}$.

3.3. If Obj $\leq \mathrm{Tr}(XX^*\hat{\phi}\hat{\phi}^*)$ then Obj $= \mathrm{Tr}(XX^*\hat{\phi}\hat{\phi}^*)$ and $\bar{\phi} = \hat{\phi}$.

3.4. count = count + 1.

4. If count ≤ 10 go to 3.

5. Let $\hat{\phi} = \bar{\phi}$, $\hat{\mathbf{s}} = \frac{e^{j\hat{\phi}}}{\sqrt{T}}$.

6. Solve (3.44). Let $\bar{\mathbf{s}}$ be the solution.

[Remark: We used 10 rounding iterations in the description of the SDPLS algorithm. We would like to emphasize that there is no particular reason which would explain what is the optimal number of these iterations. We used 10 and obtained decent performance.] Roughly speaking, the main idea of the SDPLS algorithm is to improve on SDP by doing an additional limited search over the codewords which have the squared inner product with $\hat{\mathbf{s}}$ greater than α. With this improvement we will be able to provide sound proofs regarding the coding loss of the SDP relaxation in the following section.

3.4.3 Computing the PEP

In this section we compute the PEP-type bound on the probability of error of the SDPLS algorithm. The probability of error can be written as

$$P_e = \sum_{t=1}^{q^T} P(\text{error}|\mathbf{s}_t\text{is sent})P(\mathbf{s}_t\text{is sent}). \tag{3.45}$$

In the remainder of this section, we derive an upper bound on the $P(\text{error}|\mathbf{s}_t\text{is sent})$. To facilitate this derivation, let us assume that there is a Genie who can tell us if the $\hat{\mathbf{s}}$ found in (3.38) is such that $|\hat{\mathbf{s}}^*\mathbf{s}_t|^2 < \alpha$. We formulate a slightly modified version of the SDPLS

algorithm and refer to it as the *Genie*. Its solution $\hat{\mathbf{s}}_1$ is such that

$$\begin{aligned} &\text{if} \quad |\hat{\mathbf{s}}^* \mathbf{s}_t|^2 < \alpha \qquad \hat{\mathbf{s}}_1 = \hat{\mathbf{s}} \\ &\text{if} \quad |\hat{\mathbf{s}}^* \mathbf{s}_t|^2 \geq \alpha \qquad \hat{\mathbf{s}}_1 = \bar{\mathbf{s}} \end{aligned} \tag{3.46}$$

where $\hat{\mathbf{s}}$ is as found in (3.38). The probability of error for the *Genie* algorithm is given by

$$P_e^g = \sum_{i=1, i\neq t}^{q^T} P_g(\text{error}|\mathbf{s}_t\text{is sent}) P(\mathbf{s}_t\text{is sent}) \tag{3.47}$$

where $P_g(\text{error}|\mathbf{s}_t\text{is sent})$ denotes the probability that an error occurred if the codeword $\mathbf{s}_t$ was sent and *Genie* algorithm was applied. Clearly, our SDPLS algorithm will have smaller probability of error than the *Genie*. Namely, the *Genie* and SDPLS differ in the case when $|\hat{\mathbf{s}}^* \mathbf{s}_t| < \alpha$. The *Genie* keeps $\hat{\mathbf{s}}$ as a solution which is incorrect since $|\hat{\mathbf{s}}^* \mathbf{s}_t| < \alpha < 1$ implies $\hat{\mathbf{s}} \neq \mathbf{s}_t$. Since in the only case when they differ *Genie* certainly makes a mistake, it can not have smaller probability of error than SDPLS. Therefore if we prove that *Genie* attains a certain probability of error, then SDPLS does so too. Hence, we concentrate on bounding the probability of error of the *Genie*, i.e., on bounding $P_g(\text{error}|\mathbf{s}_t\text{is sent})$. The bound obtained this way will also be a bound on the probability of error of the SDPLS. To this end, note that

$$\begin{aligned} P_g(\text{error}|\mathbf{s}_t\text{is sent}) = P(\hat{\mathbf{s}}_1 \neq \mathbf{s}_t) = P(\exists i: \ \hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t) &\leq \sum_{\mathbf{s}_i \neq \mathbf{s}_t} P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t) \\ &\leq \sum_{|\mathbf{s}_i^* \mathbf{s}_t|^2 < \alpha} P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t) + \sum_{|\mathbf{s}_i^* \mathbf{s}_t|^2 \geq \alpha} P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t). \end{aligned} \tag{3.48}$$

Let us consider $P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t|^2 < \alpha)$ in more detail. (For brevity of notation, in the following expressions we omit that everything is conditioned on $\mathbf{s}_t$ being transmitted, and that $|\mathbf{s}_i^* \mathbf{s}_t|^2 < \alpha$.) So,

$$P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t) = P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t | \hat{\mathbf{s}}_1 = \hat{\mathbf{s}}) P(\hat{\mathbf{s}}_1 = \hat{\mathbf{s}}) + P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, \hat{\mathbf{s}}_1 \neq \hat{\mathbf{s}}). \tag{3.49}$$

Let us define function C as $C(\mathbf{s}) = \mathrm{Tr}XX^*\mathbf{s}\mathbf{s}^*$. Furthermore, let E denote the event that $(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, \hat{\mathbf{s}}_1 \neq \hat{\mathbf{s}})$. Clearly, E implies that $C(\mathbf{s}_i) = C(\hat{\mathbf{s}}_1) \geq C(\hat{\mathbf{s}}) \geq \alpha C(\mathbf{s}_{\mathrm{ML}}) \geq \alpha C(\mathbf{s}_t)$, which further means that $C(\mathbf{s}_i) \geq \alpha C(\mathbf{s}_t)$. Using this, we obtain $P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, \hat{\mathbf{s}}_1 \neq \hat{\mathbf{s}}) \leq P(C(\mathbf{s}_i) \geq \alpha C(\mathbf{s}_t))$. Also, following similar argument, it is not difficult to see that $P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t | \hat{\mathbf{s}}_1 = \hat{\mathbf{s}}) P(\hat{\mathbf{s}}_1 = \hat{\mathbf{s}})) \leq P(C(\mathbf{s}_i) \geq \alpha C(\mathbf{s}_t))$. Replacing the obtained inequalities in (3.49) we have

$$P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t|^2 < \alpha) \leq 2P(C(\mathbf{s}_i) \geq \alpha C(\mathbf{s}_t)). \tag{3.50}$$

Now, let us consider $P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t|^2 \geq \alpha)$. It is not that difficult to see that

$$P(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t|^2 \geq \alpha) \leq P(C(\mathbf{s}_i) \geq C(\mathbf{s}_t)). \tag{3.51}$$

In order to precisely establish (3.51) we need the following implication

$$(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t| \geq \alpha) \Longrightarrow |\hat{\mathbf{s}}^* \mathbf{s}_t| \geq \alpha.$$

We will show that the previous implication holds using a contradiction argument. Assume that it is not correct. Then it means that $(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t| \geq \alpha)$ is true and $|\hat{\mathbf{s}}^* \mathbf{s}_t| \geq \alpha$ is not true. If $|\hat{\mathbf{s}}^* \mathbf{s}_t| \geq \alpha$ is not true then $|\hat{\mathbf{s}}^* \mathbf{s}_t| < \alpha$ is true. Then $\hat{\mathbf{s}}_1 = \hat{\mathbf{s}}$ is true as well. Further we have $\mathbf{s}_i = \hat{\mathbf{s}}_1 = \hat{\mathbf{s}}$ and $|\mathbf{s}_i^* \mathbf{s}_t| = |\hat{\mathbf{s}}^* \mathbf{s}_t| < \alpha$. This is a contradiction since it was assumed that $(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t| \geq \alpha)$ is true and hence $|\mathbf{s}_i^* \mathbf{s}_t| \geq \alpha$. Therefore the implication

$$(\hat{\mathbf{s}}_1 = \mathbf{s}_i \neq \mathbf{s}_t, |\mathbf{s}_i^* \mathbf{s}_t| \geq \alpha) \Longrightarrow |\hat{\mathbf{s}}^* \mathbf{s}_t| \geq \alpha.$$

is indeed true.

Substituting (3.50) and (3.51) in (3.48), we finally obtain

$$P_g(\text{error}|\mathbf{s}_t\text{is sent}) \leq \sum_{|\mathbf{s}_i^* \mathbf{s}_t|^2 \leq \alpha} 2P(C(\mathbf{s}_i) \geq \alpha C(\mathbf{s}_t)) + \sum_{|\mathbf{s}_i^* \mathbf{s}_t|^2 \geq \alpha} P(C(\mathbf{s}_i) \geq C(\mathbf{s}_t)). \tag{3.52}$$

Let

$$\begin{aligned} P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} &= P(C(\mathbf{s}_i) \geq \alpha C(\mathbf{s}_t)|\mathbf{s}_t \text{is sent}, |\mathbf{s}_i^*\mathbf{s}_t|^2 < \alpha) \\ P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2\geq\alpha} &= P(C(\mathbf{s}_i) \geq \mathcal{C}(\mathbf{s}_t)|\mathbf{s}_t \text{ is sent}, |\mathbf{s}_i^*\mathbf{s}_t|^2 \geq \alpha). \end{aligned}$$

In the remainder of this section, we compute bounds on $P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha}$ and $P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2\geq\alpha}$.

$$P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} = P(\mathrm{Tr}(X^*\mathbf{s}_i)(X^*\mathbf{s}_i)^* \geq \alpha \mathrm{Tr}(X^*\mathbf{s}_t)(X^*\mathbf{s}_t)^*|\mathbf{s}_t \text{ is sent}). \tag{3.53}$$

Since we assume that $\mathbf{s}_t$ was transmitted, it holds that $X = \sqrt{k}\mathbf{s}_t\mathbf{h} + W$ where, as earlier, $k = \rho T$. Replacing this value for X in (3.53), we obtain

$$P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} = P(\mathrm{Tr}(\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}^* Q_n \begin{bmatrix}\mathbf{h}\\ W\end{bmatrix} \geq 0|\mathbf{s}_t \text{ is sent}), \tag{3.54}$$

where

$$\begin{aligned} Q_n &= \begin{bmatrix}\sqrt{k}\mathbf{s}_t^*\\ I\end{bmatrix} (\mathbf{s}_i\mathbf{s}_i^* - \alpha\mathbf{s}_t\mathbf{s}_t^*) \begin{bmatrix}\sqrt{k}\mathbf{s}_t & I\end{bmatrix} = \begin{bmatrix}\sqrt{k}\mathbf{s}_t^*\\ I\end{bmatrix} \begin{bmatrix}\mathbf{s}_i & \mathbf{s}_t\end{bmatrix} \begin{bmatrix}1 & 0\\ 0 & -\alpha\end{bmatrix} \begin{bmatrix}\mathbf{s}_i^*\\ \mathbf{s}_t^*\end{bmatrix} \begin{bmatrix}\sqrt{k}\mathbf{s}_t & I\end{bmatrix} \\ &= \begin{bmatrix}\sqrt{k}\psi_{it}^* & \sqrt{k}\\ \mathbf{s}_i & \mathbf{s}_t\end{bmatrix} \begin{bmatrix}1 & 0\\ 0 & -\alpha\end{bmatrix} \begin{bmatrix}\sqrt{k}\psi_{it} & \mathbf{s}_i^*\\ \sqrt{k} & \mathbf{s}_t^*\end{bmatrix}, \end{aligned}$$

and $\psi_{it} = \mathbf{s}_i^*\mathbf{s}_t$. Although it is possible to compute explicitly the probability in (3.54), we will find that it is sufficient to find its Chernoff bound. In particular,

$$P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} \leq \min_{\mu} Ee^{\mu(\mathrm{Tr}(\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}^* Q_n \begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}))} = \int \frac{e^{-\mathrm{Tr}(\begin{bmatrix}\mathbf{h}\\ W\end{bmatrix}^* (I-\mu Q_n) \begin{bmatrix}\mathbf{h}\\ W\end{bmatrix})} d\mathbf{h}dW}{\pi^N} = \frac{1}{\det(I-\mu Q_n)^N}$$

where $\{\mu | I - \mu Q_n \geq 0\}$. We first simplify the determinant in the denominator as

$$\det(I - \mu Q_n) = \det(I - \mu \begin{bmatrix} k\psi_{it}\psi_{it}^* + 1 & (k+1)\psi_{it} \\ -\alpha(k+1)\psi_{it}^* & -\alpha(k+1)1 \end{bmatrix}).$$

After some further algebraic transformations we obtain

$$\det(I - \mu Q_n) = (k+1)\alpha(V^{(it)} - 1)(-\mu + \xi^{(1)})(-\mu + \xi^{(2)})$$

with

$$\begin{aligned} \xi^{(1)} &= \frac{V^{(it)} - \alpha + \frac{1-\alpha}{k} + \sqrt{(V^{(it)} - \alpha + \frac{1-\alpha}{k})^2 + \frac{4\alpha(1-V^{(it)})(k+1)}{k^2}}}{2\alpha(V^{(it)} - 1)\frac{k+1}{k}} \\ \xi^{(2)} &= \frac{V^{(it)} - \alpha + \frac{1-\alpha}{k} - \sqrt{(V^{(it)} - \alpha + \frac{1-\alpha}{k})^2 + \frac{4\alpha(1-V^{(it)})(k+1)}{k^2}}}{2\alpha(V^{(it)} - 1)\frac{k+1}{k}} \end{aligned}$$

and $V^{(it)} = \psi_{it}\psi_{it}^*$. As earlier, our results can be made precise so that they hold for any SNR. However, to make writing less tedious in the rest of this section we consider only the case of large SNR. Assuming $\mu = \frac{1}{2}$, the previous results simplify to

$$P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} \leq \frac{1}{(k\frac{(\alpha - V^{(it)})^2}{4(1-V^{(it)})})^N}. \tag{3.55}$$

To compute the bound on $P(C(\mathbf{s}_i) \geq \mathcal{C}(\mathbf{s}_t) | \mathbf{s}_t \text{is sent}, |\mathbf{s}_i^*\mathbf{s}_t|^2 \geq \alpha)$ we will use a well-known result from the literature (see, e.g., [51])

$$P_{it||\mathbf{s}_i^*\mathbf{s}_t|^2\geq\alpha} \leq \frac{1}{(k\frac{(1-V^{(it)})}{4})^N}. \tag{3.56}$$

Now we can substitute the results from (3.55) and (3.56) in (3.52) and obtain

$$P_g(\text{error}|\mathbf{s}_t\text{is sent}) \leq \sum_{|\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} 2\frac{1}{(k\frac{(\alpha - V^{(it)})^2}{4(1-V^{(it)})})^N} + \sum_{|\mathbf{s}_i^*\mathbf{s}_t|^2\geq\alpha} \frac{1}{(k\frac{(1-V^{(it)})}{4})^N} = B^{pep}(\rho). \tag{3.57}$$

Recall that in the case of the exact ML detection, which requires algorithms, none of which

are of polynomial complexity, we have for the same probability of error

$$P_{ML}(\text{error}|\mathbf{s}_t\text{is sent}) \leq \sum_{|\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} \frac{1}{(k\frac{(1-V^{(it)})}{4})^N} + \sum_{|\mathbf{s}_i^*\mathbf{s}_t|^2\geq\alpha} \frac{1}{(k\frac{(1-V^{(it)})}{4})^N} = B_{ML}^{pep}(\rho). \quad (3.58)$$

Clearly, comparing (3.57) and (3.58) it follows that the SDPLS algorithm based on the well-known SDP relaxation (slightly refined here for the purposes of the valid proof) has the *same* diversity as the exact ML and the AR algorithm. Of course, since the SDPLS algorithm is only an approximation, the exact ML solution still has an advantage of $(\frac{1-V^{(it)}}{\alpha-V^{(it)}})^2$ in the coding gain. However, as the analysis conducted in Section 3.2 hints (and the simulation result on Figure 3.6 confirms) the AR algorithm has a significantly bigger coding loss than the SDPLS algorithm analyzed in this section.

It should also be noted that a very similar result related to the diversity of the SDP-based algorithm in the context of coherent (channel known at the receiver) ML detection has recently been shown in [60].

We summarize the previous results in the following theorem.

Theorem 3.4. *Consider the problem of non-coherent ML detection for a SIMO system described in (3.1) in high-SNR regime. Assume that the codeword* $\mathbf{s}_t$ *was transmitted. Then the probability that an error occurred if SDPLS algorithm was applied to solve (3.34) can be upper bounded in the following way*

$$P(error|\mathbf{s}_t is\ sent) \leq \sum_{|\mathbf{s}_i^*\mathbf{s}_t|^2<\alpha} 2\frac{1}{(\rho T\frac{(\alpha-V^{(it)})^2}{4(1-V^{(it)})})^N} + \sum_{|\mathbf{s}_i^*\mathbf{s}_t|^2\geq\alpha} \frac{1}{(\rho T\frac{(1-V^{(it)})}{4})^N}.$$

Proof. Follows from the previous discussion. □

3.4.4 Asymptotic analysis, $T \to \infty$

In this section we explicitly compute the ratio of the bounds on the probability of error ($P(\text{error}|\mathbf{s}_t\text{is sent})$ and $P_{ML}(\text{error}|\mathbf{s}_t\text{is sent})$) in the case when $T \to \infty$. We first explicitly analyze the special case $q = 2$ which corresponds to 2-PSK. Afterwards we derive the

corresponding results for general q-PSK.

3.4.4.1 $q = 2$

In this section we will compute in the limit of large T the following quantity

$$K_\alpha = \sum_{|\mathbf{s}_i^* \mathbf{s}_t|^2 < \alpha} \frac{1}{(\frac{(\alpha - V^{(it)})^2}{(1-V^{(it)})})^N}. \tag{3.59}$$

Let $x_{min} = \left\lfloor \frac{T(1-\sqrt{\alpha})}{2} \right\rfloor$. Then it is not difficult to see that

$$K_\alpha = \sum_{x=x_{min}+1}^{T-(x_{min}+1)} \binom{T}{x} \frac{1}{(\frac{(\alpha - (\frac{T-2x}{T})^2)^2}{(1-(\frac{T-2x}{T})^2)})^N}. \tag{3.60}$$

Before proceeding further let us examine more carefully the behavior of $\binom{T}{x}$ when T is large. First let $x = p_1 T$. Then we have that $\binom{T}{x} = \frac{T!}{(p_1 T)!(p_2 T)!}$, where $p_1 + p_2 = 1$. Furthermore we have that

$$\ln T! = (T + \frac{1}{2}) \ln T - T + \frac{1}{2} \ln(2\pi) + \ln(1 + \mathcal{O}(\frac{1}{T})) \tag{3.61}$$

and similarly

$$\ln(p_i T)! = p_i T \ln T + \frac{1}{2} \ln T + p_i T \ln p_i + \frac{1}{2} \ln p_i - p_i T + \frac{1}{2} \ln(2\pi) + \ln(1 + \mathcal{O}(\frac{1}{T})). \tag{3.62}$$

Combining (3.61) and (3.62) we obtain

$$\binom{T}{x} = \frac{T!}{(p_1 T)!(p_2 T)!} = \frac{e^{TH(p_1)}}{\sqrt{2T\pi p_1(1-p_1)}}(1 + \mathcal{O}(\frac{1}{T})) \tag{3.63}$$

where $H(p_1) = -p_1 \ln p_1 - (1-p_1)\ln(1-p_1)$ is the entropy function. Replacing (3.63) in (3.60) and assuming $T \to \infty$ we further have

$$K_\alpha = \sum_{x=(x_{min}+1)}^{T-(x_{min}+1)} \frac{e^{TH(x/T)}}{\sqrt{2\pi x(1-x/T)}} \frac{1}{(\frac{(\alpha-(1-2x/T)^2)^2}{1-(1-2x/T)^2})^N}$$
$$= \int_{(x_{min}+1)/T}^{1-(x_{min}+1)/T} T \frac{e^{TH(p_1)}(1+\mathcal{O}(\frac{1}{T}))}{\sqrt{2T\pi p_1(1-p_1)}} \frac{dp_1}{(\frac{(\alpha-(1-2p_1)^2)^2}{1-(1-2p_1)^2})^N}. \quad (3.64)$$

To solve the previous integral we will use the saddle point method. Let

$$g(p_1) = \frac{T(1+\mathcal{O}(\frac{1}{T}))}{\sqrt{2T\pi p_1(1-p_1)}(\frac{(\alpha-(1-2p_1)^2)^2}{1-(1-2p_1)^2})^N}.$$

The saddle point method gives

$$K_\alpha = e^{TH(p_0)} g(p_0) \sqrt{\frac{2\pi}{T\frac{d^2H(p_1)}{dp_1^2}|_{p_1=p_0}}} (1+\mathcal{O}(\frac{1}{T})) \quad (3.65)$$

where p_0 is solution to $\frac{dH(p_1)}{dp_1} = 0$. Then it easily follows

$$p_0 = \frac{1}{2}, H(p_0) = \ln 2, g(p_0) = \frac{T}{\sqrt{T\pi/2}\alpha^{2N}}, \frac{d^2H(p_1)}{dp_1^2}|_{p_1=p_0} = 4.$$

Using all of this (3.65) becomes

$$K_\alpha = \frac{2^T T}{\sqrt{T\pi/2}\alpha^{2N}} \sqrt{\frac{\pi}{2T}} = \frac{2^T}{\alpha^{2N}}(1+\mathcal{O}(\frac{1}{T})). \quad (3.66)$$

We summarize the previous analysis in the following theorem.

Theorem 3.5. *Consider the problem of non-coherent ML detection for a SIMO system described in (3.1) in high-SNR regime. Assume that the elements of the transmitted codeword* $\mathbf{s}_t$ *are chosen from 2-PSK constellation and that* $T \to \infty$. *Let* $B^{pep}(\rho)$ *defined in (3.57) be the PEP type bound on the probability that an error occurred if SDPLS algorithm was applied to solve (3.34). Let* B_{ML}^{pep} *defined in (3.58) be the PEP type bound on the probability*

that an error occurred if an exact ML algorithm was applied to solve (3.34). Then

$$B^{pep}(\rho/\alpha^2) \leq B_{ML}^{pep}(\rho)(1+\mathcal{O}(\frac{1}{T})).$$

and

$$10\log\frac{1}{\alpha^2} = 3.92dB.$$

Proof. The fact that if $q=2$ then $\alpha=\frac{2}{\pi}$ was proved in [71]. The rest follows by combining (3.57), (3.58), and (3.66). □

3.4.4.2 General q

In this subsection we generalize the result for 2-PSK to q-PSK. In q-PSK case the elements of a vector $\mathbf{s}_i$ are from the set $\mathcal{Z}=\{\frac{1}{\sqrt{T}},\frac{e^{\frac{j2\pi}{q}}}{\sqrt{T}},\frac{e^{\frac{j4\pi}{q}}}{\sqrt{T}},\ldots,\frac{e^{\frac{j2(q-1)\pi}{q}}}{\sqrt{T}}\}$ and as shown in [84] and [114] $\alpha=\frac{(q\sin(\pi/q))^2}{4\pi}$. As in the previous subsection, let $z_l=s_{i_l}^*s_{t_l}$ and $\mathbf{z}=[z_1,z_2,\ldots,z_T]$. Clearly the elements of $\mathbf{z}$ are also from the set $\mathcal{Z}$. Let $x_1,x_2,\ldots,x_q$ be the numbers of elements in $\mathbf{z}$ that are equal to $\frac{1}{T},\frac{e^{\frac{j2\pi}{q}}}{T},\frac{e^{\frac{j4\pi}{q}}}{T},\ldots,\frac{e^{\frac{j2(q-1)\pi}{q}}}{T}$, respectively. Then it easily follows that

$$|\mathbf{s}_i^*\mathbf{s}_t|^2 = [((T-\sum_{i=1}^{q-1}x_i)\cos(\frac{2(q-1)\pi}{q})+\sum_{i=1}^{q-1}x_i\cos(\frac{2(i-1)\pi}{q}))^2 + ((T-\sum_{i=1}^{q-1}x_i)\sin(\frac{2(q-1)\pi}{q})+\sum_{i=1}^{q-1}x_i\sin(\frac{2(i-1)\pi}{q}))^2]/T^2.$$

Let $p_i=x_i/T, 1\leq i\leq q$, $\mathbf{p}=[p_1,p_2,\ldots,p_{q-1}]$, and

$$f(\mathbf{p}) = ((1-\sum_{i=1}^{q-1}p_i)\cos(\frac{2(q-1)\pi}{q})+\sum_{i=1}^{q-1}p_i\cos(\frac{2(i-1)\pi}{q}))^2 + ((T-\sum_{i=1}^{q-1}p_i)\sin(\frac{2(q-1)\pi}{q})+\sum_{i=1}^{q-1}p_i\sin(\frac{2(i-1)\pi}{q}))^2.$$

Then as in (3.60) we have

$$K_\alpha = \sum_{\substack{f(\frac{\mathbf{x}}{T}) < \alpha \\ \sum_{i=1}^{q-1} x_i \leq T \\ 0 \leq x_i}} \binom{T}{x_1 x_2 \dots x_{q-1}} \frac{1}{\left(\frac{(\alpha - f(\frac{\mathbf{x}}{T}))^2}{1-f(\frac{\mathbf{x}}{T})}\right)^N}. \tag{3.67}$$

It is not difficult to see that (3.63) can be generalized in the following way

$$\binom{T}{x_1 x_2 \dots x_{q-1}} = \frac{T!}{(p_1 T)!(p_2 T)! \dots (p_q T)!} = \frac{e^{TH(\mathbf{p})}}{(2T\pi(1-\sum_i^{q-1} p_i)\prod_{i=1}^{q-1} p_i)^{(q-1)/2}}(1+\mathcal{O}(\frac{1}{T})). \tag{3.68}$$

Then as in (3.64), replacing (3.68) in (3.67) we have

$$K_\alpha = \int_{\substack{f(\mathbf{p}) < \alpha \\ \sum_{i=1}^{q} p_i \leq T \\ 0 \leq p_i}} \frac{e^{TH(\mathbf{p})}((1-\sum_{i=1}^{q-1} p_i)\prod_{i=1}^{q-1} p_i)^{-\frac{1}{2}}(1+\mathcal{O}(\frac{1}{T}))d\mathbf{p}}{T^{-(q-1)}(2T\pi)^{\frac{q-1}{2}}\left(\frac{(\alpha - f(\mathbf{p}))^2}{1-f(\mathbf{p})}\right)^N}.$$

Let $\mathbf{p}_0 = [1/q, 1/q, \dots, 1/q]$ be the solution of $\frac{dH(\mathbf{p})}{d\mathbf{p}} = 0$ and let $\mathcal{H}(H(\mathbf{p}))$ be the Hessian of $H(\mathbf{p})$. Further let

$$g(\mathbf{p}) = \frac{((1-\sum_{i=1}^{q-1} p_i)\prod_{i=1}^{q-1} p_i)^{-\frac{1}{2}}}{T^{-(q-1)}(2T\pi)^{\frac{q-1}{2}}\left(\frac{(\alpha - f(\mathbf{p}))^2}{1-f(\mathbf{p})}\right)^N}(1+\mathcal{O}(\frac{1}{T})).$$

Then, since entropy is a convex function, we can as in (3.65), write

$$K_\alpha = e^{TH(\mathbf{p}_0)} g(\mathbf{p}_0)\left(\frac{2\pi}{T(\det \mathcal{H}(H(\mathbf{p}))|_{\mathbf{p}=\mathbf{p}_0})^{1/(q-1)}}\right)^{\frac{q-1}{2}}(1+\mathcal{O}(\frac{1}{T})). \tag{3.69}$$

It is easy to see that

$$H(\mathbf{p}_0) = \ln q, \det \mathcal{H}(H(\mathbf{p}))|_{\mathbf{p}=\mathbf{p}_0} = q^q, h(\mathbf{p}_0) = 0, g(\mathbf{p}_0) = \frac{T^{q-1}(q^q)^{\frac{1}{2}}}{(2T\pi)^{\frac{q-1}{2}}\alpha^{2N}}(1+\mathcal{O}(\frac{1}{T})).$$

Replacing these values in (3.69) we finally obtain

$$K_\alpha = e^{T\ln q}\frac{T^{q-1}(q^q)^{\frac{1}{2}}}{(2T\pi)^{\frac{q-1}{2}}\alpha^{2N}}\left(\frac{2\pi}{T(q^q)^{\frac{1}{q-1}}}\right)^{\frac{q-1}{2}}(1+\mathcal{O}(\frac{1}{T})) = \frac{q^T}{\alpha^{2N}}(1+\mathcal{O}(\frac{1}{T})). \tag{3.70}$$

We summarize the previous analysis in the following theorem:

Theorem 3.6. *Consider the problem of non-coherent ML detection for a SIMO system described in (3.1) in high-SNR regime. Assume that the elements of the transmitted codeword* $\mathbf{s}_t$ *are chosen from q-PSK (*$q \geq 4$*) constellation and that* $T \to \infty$*. Let* $B^{pep}(\rho)$ *defined in (3.57) be the PEP-type upper bound on the probability that an error occurred if SDPLS algorithm was applied to solve (3.34). Let* B^{pep}_{ML} *defined in (3.58) be the PEP-type upper bound on the probability that an error occurred if an exact ML algorithm was applied to solve (3.34). Then*

$$B^{pep}(\rho/\alpha^2) \leq B^{pep}_{ML}(\rho)(1+\mathcal{O}(\frac{1}{T})).$$

Furthermore, assume that there are two SNRs — ρ_{SDP} *and* ρ_{ML} *— such that* $B^{pep}(\rho_{SDP}) = B^{pep}_{ML}(\rho_{ML})(1+\mathcal{O}(\frac{1}{T}))$*. Then it holds*

$$\Delta\rho(q) = \rho_{SDP} - \rho_{ML} \leq 20\log\left(\frac{4\pi}{(q\sin(\pi/q))^2}\right) \quad dB.$$

We further have $\Delta\rho(4) = 3.92$*,* $\Delta\rho(8) = 2.547$ *dB,* $\Delta\rho(16) = 2.21$ *dB, and*

$$\lim_{q\to\infty}\Delta\rho(q) = 2.0982 \quad dB.$$

Proof. Follows by combining (3.57), (3.58), and (3.70). □

Effectively, Theorem 3.6 states that if a codeword was transmitted then its averaged (averaging is over all other codewords) bounds on pairwise probabilities of error in the case of exact ML and approximate SDP detection differ by at most $\Delta\rho(q)$ dB.

3.4.5 Computational complexity

At the end, let us elaborate briefly on the theoretical complexity and the practical number of operations of the SDPLS algorithm that we have proposed. By carefully inspecting it, one can note that due to the modification of the conventional SDP randomized algorithm, our SDPLS algorithm, is, strictly speaking, no longer polynomial. However, for most practical cases the additional amount of operations on top of the basic SDP core of the algorithm is effectively negligible. To examine this let us study the case of 2- and 4-PSK (for q-PSK similar arguments can be established). Note that the additional amount of the arithmetic operations is equal to the number of the vectors $\mathbf{s}$, $|S^c|$, which satisfy inequality $|\mathbf{s}^*\hat{\mathbf{s}}|^2 \geq \alpha = \frac{2}{\pi}$. These vectors can be found through exhaustive search. Clearly, in the case of 2-PSK this number can be upper-bounded as

$$|S^c| \leq \left\lfloor \frac{T(1-\sqrt{\alpha})}{2} \right\rfloor \binom{T}{\lfloor \frac{T(1-\sqrt{\alpha})}{2} \rfloor} \leq T^{4.2}, \text{if} \quad T < 60,$$

where we have assumed that for $T < 60$ the number of arithmetic operations required for solving an SDP is $60^{4.2}$. In the 4-PSK case we can numerically obtain

$$|S^c| \leq \sum_{f(\frac{\mathbf{x}}{T}) \geq \alpha, 0 \leq x_i \leq T, x_1 \neq T} \binom{T}{x_1 x_2 x_3} \leq T^{4.71} \leq T^{4.8}, T \leq 24$$

where we have assumed that the number of arithmetic operations required for solving an SDP of dimension $T \leq 24$ is at least $T^{4.8}$.

However, for large T, $\binom{T}{x_{min}} = \frac{2^{TH(x_{min}/T)}}{\sqrt{2\pi x_{min}(1-x_{min}/T)}}(1+\mathcal{O}(\frac{1}{T}))$ (where H is the entropy function), and one can show that

$$|S^c| \geq \binom{T}{\lfloor \frac{T(1-\sqrt{\alpha})}{2} \rfloor} = \frac{2^{TH(\lfloor (1-\sqrt{\alpha})/2 \rfloor)}}{\sqrt{2\pi x_{min}(1-\frac{x_{min}}{T})}}(1+\mathcal{O}(\frac{1}{T})) = \frac{2^{0.47T}}{\sqrt{0.1817\pi T}}(1+\mathcal{O}(\frac{1}{T})).$$

The previous expression implies that the additional amount of computation introduced to ensure the validity of our proof is indeed exponential, while of course in the limit of large T the complexity of solving SDP becomes $\mathcal{O}(T^{3.5})$. However, the exponential constant is

two times smaller than in the exhaustive search. Therefore, in communications, where the dimension of SIMO systems is smaller than 60 and 2-PSK signalling is used, or where the dimension of SIMO system is smaller than 24 and 4-PSK signalling is used, the required number of arithmetic operations for our algorithm is similar to the corresponding number of required operations for the basic SDP.

3.5 Discussion and conclusion

In this chapter we considered the non-coherent ML detection in single-input multiple-output communication systems with q-PSK signalling.

To solve the problem exactly in the first part of this chapter we introduced the out-sphere decoder algorithm as an analogue to the standard sphere decoder used in the coherent detection. The main contribution was analytical characterization of the algorithm's expected complexity.

In the second part of this chapter we proposed a modification of the SDP relaxation for solving approximately the non-coherent ML detection in SIMO systems. The computed PEP implies that the performance of the algorithm is comparable to that of the optimal ML solution, but is obtained at potentially significantly lower computational complexity. Namely, we proved that the SDP relaxation achieves the same diversity as the exact ML. Furthermore, we proved that a modification of the SDP relaxation has Chernoff bound on the PEP-type bound on the probability of error within a constant factor of the corresponding bound in the exact ML case.

In addition to the analysis of the PEP performance of a modification of the standard SDP-based method for solving non-coherent ML detection, we introduced a simple rounding algorithm. The algorithm seems naive at first and reminds one of its counterparts (nulling and cancelling (NC), zero-forcing (ZF)) in coherent ML detection. However, while NC and ZF don't perform as well as the exact ML in coherent ML detection, the rounding algorithm that we have introduced for non-coherent detection performs very well. In fact not only does it perform very well, but it actually provably achieves the *same* diversity as the exact

ML. Numerical experiments confirmed that this indeed is true. On the other hand, since the rounding algorithm is only an approximation, it has a coding loss compared to the exact ML. However, as simulation results showed, this coding loss indeed exists but is no more than 1 dB for the reasonable system dimensions.

At the end we would like to mention several possible directions for a future work. First of all, we would like to say that although our modified SDP relaxation for non-coherent detection requires number of arithmetic operations similar to that of the SDP, strictly speaking it is not a polynomial time algorithm. Therefore it would be of a great interest if one could construct a provably polynomial time algorithm which has the same PEP performance as the one we derived in this chapter. Also, in this chapter we only analyzed a simple SIMO system. It would be interesting to see how the results proved here can be generalized to the MIMO case. Another important consideration would be adapting the introduced algorithms for different types of signalling (e.g., general QAM).

Chapter 4

Gaussian Broadcast Channel — Linear Precoding Schemes

In the previous two chapters we considered the so-called point-to-point communication where one or both ends (transmitter/receiver) of the communication were equipped with several antennas. In this and the following chapter we will consider a different concept called broadcast communications.

The sketch of a communication system where such a concept is employed is shown in Figure 4.1. As can be seen, the system basically consists of a transmitter $\mathbf{T}$ which broadcasts information to the several (M) users/receivers $\mathbf{R}_i, 1 \leq i \leq M$. The main difference compared to the point-to-point system from the previous two chapters is that now one transmitter is communicating with several receivers at the same time. In a general setting, the transmitter and each of the users can be equipped with many antennas as well. However, in this thesis we will mostly focus on the case when the transmitter is equipped with many antennas (typically M) and each user is equipped with one antenna.

The whole transmission process then goes (as is typical in wireless systems) via electromagnetic waves through the transmission medium (air). The different encoded information $(s_i, 1 \leq i \leq M)$ is being sent from each transmitting antenna. Since the transmission medium is air, signals from each transmitting antenna are reaching each of the users. The users are then combining them and trying to extract the piece of the information the transmitter intended to send specifically to them.

As can immediately be seen, the main idea behind the concept of the broadcast channel

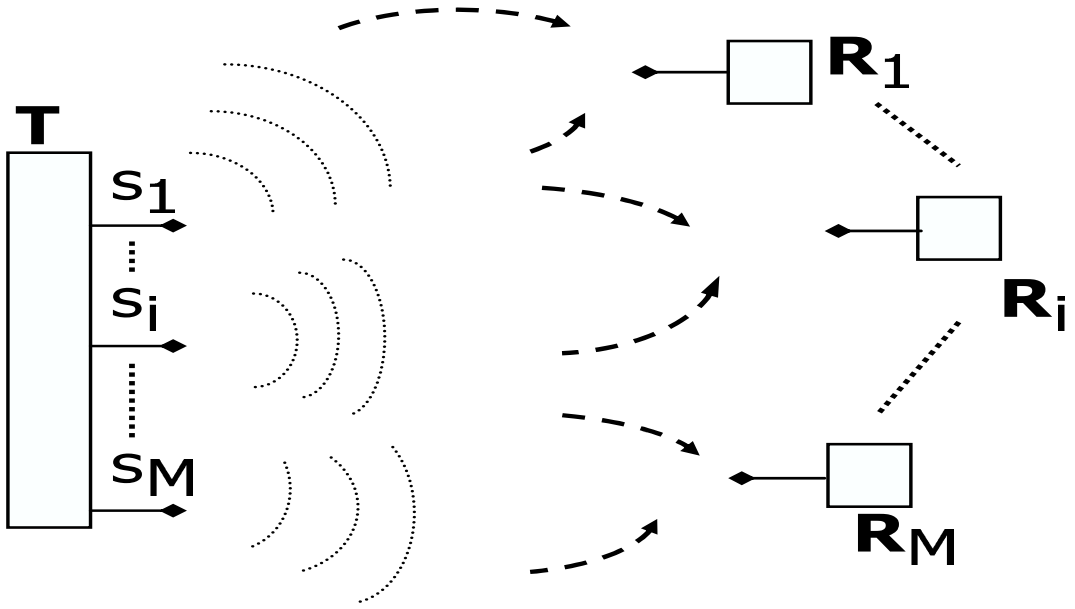


Figure 4.1: Wireless broadcast channel

is that by using several antennas the transmitter can simultaneously send different sets of data to different users. If the transmission medium were ideal (as would be the case in a wire-line system) then the signals sent from a particular antenna from the transmitter would reach only the intended user. However, since the transmission medium is air, the main problem becomes interference that can happen at each of the receivers. Namely, assume that the signal s_i from the i-th antenna is intended for the i-th user. The i-th signal will then certainly reach the i-th user. However, since there are no physical obstacles, portions of the other signals $s_j, j \neq i$ will reach the i-th receiver as well and interfere with the s_i. This interference then of course can cause the receiver to recognize incorrectly what was the original signal s_i. This simple argument immediately suggests that very important features of a wireless broadcast channels are the transmission medium (often called channel), the design of the information symbol vectors $\mathbf{s}$ at the transmitter, and their detection at the receivers. In this and the following chapter we will focus on the design of the information symbol vectors so that the interference effects of an open transmission medium are as mitigated as possible.

We will consider the amount of information that can be transmitted as a parameter

of quality of the broadcast system at hand. The overall amount of the information which equals the summation of the amount of information transmitted to each user is often called the sum-rate. The sum-rate capacity is defined as a maximal sum-rate that can be achieved on a broadcast channel. The sum-rate capacity of the multi-antenna broadcast channel has recently been computed. However, the search for efficient practical schemes that achieve it is still ongoing. In this and the following chapter, we focus on schemes with linear preprocessing of the transmitted data. We first propose a linear precoding design so that the sum-rate is maximized. In terms of the achievable sum-rate, the proposed linear technique significantly outperforms traditional channel inversion methods. It is relatively easy to note that sometimes the overall sum-rate is maximized when some of the individual rates are very small. In applications where we would like for every user to be served with the same amount of information as all the other ones, this may be an unfair design. Hence to address the fairness of serving all users with a similar amount of information from the transmitter, we consider the problem of maximizing the minimum rate among all users (the so-called max-min problem). This problem is shown to be quasi convex and is solved exactly via a bisection method.

4.1 Introduction

The broadcast channel described above was introduced in [21]. Since then it has been a subject of extensive research. The most fundamental question is characterization of its capacity region. The capacity region is defined as a set of all individual rates that can be achieved simultaneously by the users. A particular point on the boundary of this region which maximizes the sum of all individual rates corresponds to the above-mentioned sum capacity. A significant progress in analysis of the capacity region and the sum-rate capacity of the broadcast channel has been achieved recently. It was shown (see, e.g., [14], [100], [101], and [111]) that the sum-rate capacity of the Gaussian broadcast channel is achieved by a scheme called dirty-paper coding (DPC) in the case where the full channel state information (CSI) is available at both the transmitter and the receivers. In [83] it was

shown that the same scaling law for the sum-rate capacity can be achieved even if only partial CSI is available at the transmitter. Furthermore, in [105] the authors showed that any point in the capacity region of the broadcast channel can be achieved by DPC. In [112] and [75], non-linear techniques that attempt to approach those limits have been considered. However, these schemes are often computationally prohibitive when the number of transmit antennas is large. In this chapter we attempt to provide a less computationally extensive alternative to the previously considered schemes.

We will assume a standard system model for the broadcast channel from Figure 4.1 with M transmit antennas and M users, described by

$$\mathbf{r} = H\mathbf{s} + \mathbf{w}, \tag{4.1}$$

where H is an $M \times M$ fading channel matrix whose entries are i.i.d. zero-mean, unit variance, complex Gaussian random variables, and $\mathbf{w}$ is an $M \times 1$ vector whose entries are also i.i.d. zero-mean, variance σ^2 complex Gaussian random variables which represent additive noise at each receiver. Furthermore, $\mathbf{s}$ is an $M \times 1$ vector of signals sent from the transmit antennas, and $\mathbf{r}$ is an $M \times 1$ vector whose components are the received signals at each user. The transmitted vector $\mathbf{s}$ is assumed to be obtained by linear preprocessing of the information vector $\mathbf{u}$, i.e., $\mathbf{s} = kG\mathbf{u}$, where $\mathbf{u} = \left[u_1, u_2, ..., u_M\right]^T$, u_i is the symbol intended for the i-th user, $1 \leq i \leq M$, and where k is a scaling coefficient which ensures that the power constraint is satisfied. The equivalent system model is shown on Figure 4.2.

We organize this chapter in the following way; first in Section 4.2 we propose two possible schemes for designing the preprocessing matrix G. In Section 4.3, we propose a possible scheme for determining the optimal value of the scaling coefficient k under the constraint of linear preprocessing at the transmitter. In Section 4.4, we describe how to combine the schemes from Sections 4.2 and 4.3. Finally, in Sections 4.5 and 4.6 we give simulation results, a brief discussion, and several conclusions. A complementary version of this chapter appeared in [93].

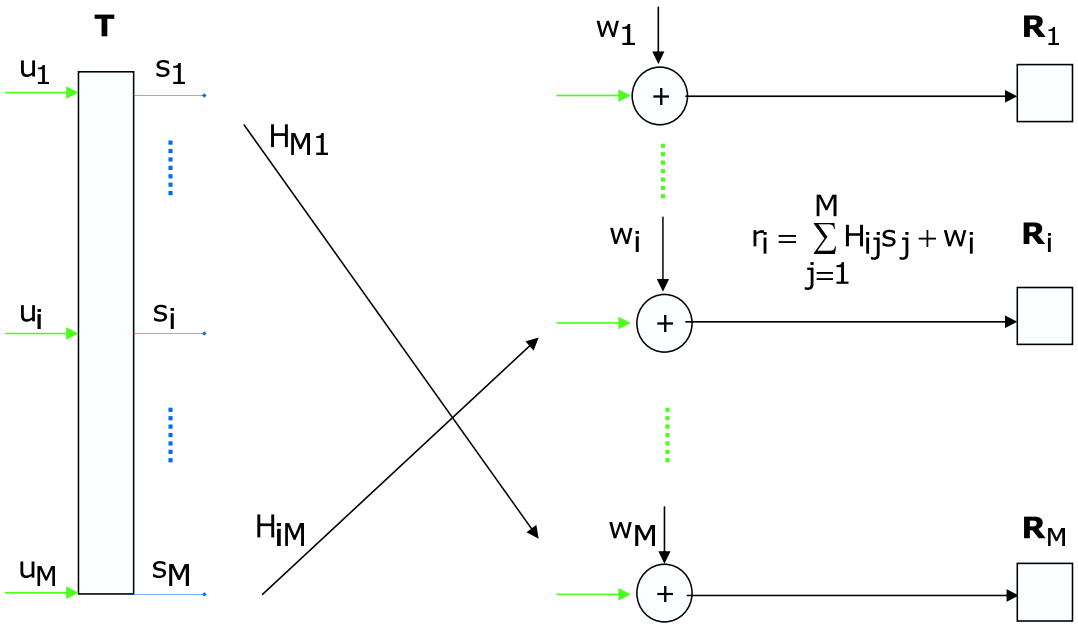


Figure 4.2: Mathematical model of a wireless broadcast channel

4.2 Finding optimal preprocessing matrix G

In this section, we find the optimal preprocessing matrix G, assuming an average transmit power constraint, $E\|\mathbf{s}\|^2 = 1$. Without loss of generality, we will assume that $E\mathbf{u}\mathbf{u}^* = I$. Then $E\|G\mathbf{u}\|^2 = E\mathrm{tr}(G\mathbf{u}\mathbf{u}^*G^*) = \mathrm{tr}(G^*G)$ and thus $k = 1/\sqrt{\mathrm{tr}(G^*G)}$. Hence, from (4.1) we obtain

$$\mathbf{r} = \frac{HG\mathbf{u}}{\sqrt{\mathrm{tr}(G^*G)}} + \mathbf{w}. \tag{4.2}$$

The matrix G in (4.2) should be designed to optimize the performance of the overall system in terms of both the rate as well as the bit-error rate. Often encountered in the literature is the solution employing a regularized pseudo-inverse of the channel matrix H, i.e., $G = H^*(\beta I + HH^*)^{-1}$, where the coefficient β is typically chosen to maximize the signal-to-interference-and-noise ratio (SINR) (see, e.g., [75]). However, optimizing for SINR does not necessarily imply that the total sum rate will be maximized. This justifies the search for a better choice for the matrix G.

We consider two optimization criteria for the design of the preprocessing matrix G. First, we maximize the total sum rate over the space of all $M \times M$ complex matrices G. As

we shall see, this optimization results in a strategy where at each channel use, a subset of users is chosen and data transmitted only to those users. Second, we consider the problem of optimal preprocessing that maximizes the minimum rate among all of the users. Extensive simulations imply that the best BER performance of the system is achieved when the two strategies are combined, i.e., when a subset of users is selected and then the minimum rate among the users in that subset is maximized.

4.2.1 Maximizing the sum rate over G

We assume that each user treats the interference as noise. Therefore the sum rate of the broadcast channel (4.2) is given by $R = \sum_{m=1}^{M} \log\left(1 + \frac{\left|\sum_p H_{mp}G_{pm}\right|^2}{\sigma^2 \mathrm{tr}(G^*G) + \sum_{n\neq m}\left|\sum_p H_{mp}G_{pn}\right|^2}\right)$.

The optimal choice for the matrix G is the solution to the optimization problem

$$\max_{G} R. \tag{4.3}$$

A closed-form analytic solution to (4.3) does not appear easy to find. In fact, even an efficient algorithm that is guaranteed to numerically solve (4.3) does not seem within reach. We thus will present an iterative scheme that may converge to a local optimum. Before proceeding any further, we will find it useful to define $\mathrm{num}_m = \left|\sum_{p=1}^{M} H_{mp}G_{pm}\right|^2$, and $\mathrm{den}_m = \sigma^2 \mathrm{tr}(G^*G) + \sum_{n=1,n\neq m}^{M}\left|\sum_{p=1}^{M} H_{mp}G_{pn}\right|^2$. The following lemma gives a necessary condition for the optimal G.

Lemma 4.1. *Denote*

$$\Delta = diag(\frac{(HG)_{11}}{den_1}, .., \frac{(HG)_{ll}}{den_l}, ..., \frac{(HG)_{MM}}{den_M})$$

and

$$D = diag(\frac{num_1}{den_1(den_1 + num_1)}, ..., \frac{num_l}{den_l(den_l + num_l)}, ..., \frac{num_M}{den_M(den_M + num_M)}).$$

Then any G which is a solution of (4.3) is of the form $G = ((\sigma^2 trD)I + H^*DH)^{-1}H^*\Delta$.

Proof. It is sufficient to show that $\frac{\partial R}{\partial G_{kl}} = 0 \Rightarrow G = ((\sigma^2 \mathrm{tr}D)I + H^*DH)^{-1}H^*\Delta$. It is straightforward to show that

$$\frac{\partial R}{\partial G_{kl}} = \frac{H_{lk}(HG)^*_{ll}}{\mathrm{den}_l} - \sum_{m=1}^{M} \frac{\mathrm{num}_m H_{mk}(HG)^*_{ml}}{\mathrm{den}_m(\mathrm{num}_m + \mathrm{den}_m)} - \sum_{m=1}^{M} \frac{\sigma^2 G^*_{kl}\mathrm{num}_m}{\mathrm{den}_m(\mathrm{den}_m + \mathrm{num}_m)}.$$

Setting each of these derivatives to zero, we obtain $H^*\Delta - H^*DHG - (\sigma^2\mathrm{tr}D)G = 0$, or equivalently $G = ((\sigma^2\mathrm{tr}D)I + H^*DH)^{-1}H^*\Delta$.

Thus $\frac{\partial R}{\partial G_{kl}} = 0 \Rightarrow G = ((\sigma^2\mathrm{tr}D)I + H^*DH)^{-1}H^*\Delta$, which concludes the proof. □

Using Lemma 1, we state the following iterative algorithm for solving (4.3).

$D_0 = I, \Delta_0 = I, i = 0, R_{-2} = 10^7, R_{-1} = 10^8$

Repeat while $|R_{i-2} - R_{i-1}| \geq 10^{-3}$

1. $G_i = ((\sigma^2 trD_i)I + H^*D_iH)^{-1}H^*\Delta_i$, $R_i = \sum_{m=1}^{M} \log\left(1 + \frac{|(HG_i)_{mm}|^2}{\sigma^2 tr(G_i^*G_i) + \sum_{n\neq m}|(HG_i)_{mn}|^2}\right)$.

2. $num_m = |(HG_i)_{mm}|^2$, $den_m = \sigma^2 tr(G_i^*G_i) + \sum_{n=1,n\neq m}^{M} |(HG_i)_{mn}|^2$.

3. $D_{i+1} = diag(\frac{num_1}{den_1(den_1+num_1)}, ..., \frac{num_l}{den_l(den_l+num_l)}, ..., \frac{num_M}{den_M(den_M+num_M)})$.

4. $\Delta_{i+1} = diag(\frac{(HG_i)_{11}}{den_1}, .., \frac{(HG_i)_{ll}}{den_l}, ..., \frac{(HG_i)_{MM}}{den_M})$, i=i+1.

end

We refer to using the matrix G obtained from the previous iterative procedure as Method 2.1. Since $H^*((\sigma^2\mathrm{tr}D)I + HH^*)^{-1} = ((\sigma^2\mathrm{tr}D)I + H^*H)^{-1}H^*$, the initial value G_0 coincides with the one obtained by the regularized pseudo-inverse (see, e.g., [75]). Simulation results presented in the following sections imply that such a choice of initial value leads to an iterative process that converges to a local optimum after a fairly small number of iterations (roughly 15 on average), although we have no formal proof of convergence at this time. In Figure 4.3, the comparison of the sum rate achieved by Method 2.1 and the sum rate achieved by the regularized pseudo-inverse are compared to the sum capacity of the broadcast channel. As can be seen, although we have no formal proof for that, Method 2.1 significantly decreases the gap between regularized pseudo-inverse and the sum capacity. In addition, as illustrated

in Figure 4.3, the plain channel inverse, obtained for $\alpha = 0$, is significantly outperformed by the regularized pseudo-inverse.

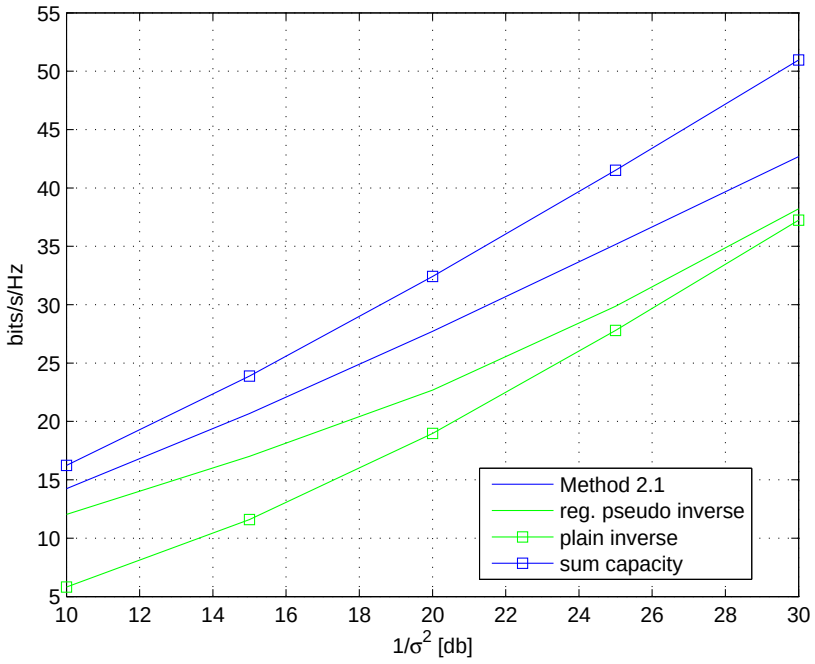


Figure 4.3: Comparison of the sum rate of Method 2.1 to the sum rate of reg. pseudo-inverse and to the sum capacity of broadcast channel, $M = 6$ antennas/users

4.2.2 Maximizing the minimum rate over G

Instead of maximizing the sum rate, one may demand that the worst (active) user gets as large a rate as possible. This criterion leads to the following optimization problem

$$\max_{G} \min_{i} \log \left(1 + \frac{|(HG)_{ii}|^2}{\sigma^2 \mathrm{tr}(G^* G) + \sum_{j,j \neq i} |(HG)_{ij}|^2} \right). \tag{4.4}$$

The previous problem (or problems similar to it) have been studied and various algorithms for solving it have been suggested throughout the literature (see, e.g., [15], [10], [102], and [107]). Here we suggest another way of solving it based on interior point methods. Define

$B = HG$. Then (4.4) can be written as

$$\max_{B} \min_{i} \frac{|B_{ii}|^2}{\sigma^2 \mathrm{tr}(B^* H^{-*} H^{-1} B) + \sum_{j,j\neq i} |B_{ij}|^2}. \tag{4.5}$$

Without loss of generality, we can assume that the optimal B_{ii} are real and positive. Let $\mathbf{vec}(B)$ denote a vector comprised of columns of matrix B. Then we can write

$$\sigma^2 \mathrm{tr}(B^* H^{-*} H^{-1} B) = \sigma^2 \mathbf{vec}(B)^* (I \otimes H^{-*} H^{-1}) \mathbf{vec}(B).$$

Denoting $F = I \otimes H^{-*} H^{-1}$, $\mathbf{x} = \begin{bmatrix} \Re(\mathbf{vec}(B)) \\ \Im(\mathbf{vec}(B)) \end{bmatrix}$ and $T = \begin{bmatrix} \Re(F) & -\Im(F) \\ \Im(F) & \Re(F) \end{bmatrix}$ we have

$$\sigma^2 \mathrm{tr}(B^* H^{-*} H^{-1} B) = \sigma^2 \mathbf{x}^* T \mathbf{x}.$$

Define $2M^2 \times 2M^2$ matrix $K^{(ij)}$ with $K^{(ij)}_{(j-1)M+i,(j-1)M+i} = K^{(ij)}_{M^2+(j-1)M+i,M^2+(j-1)M+i} = 1$ and zeros otherwise. Combining all of the above, (4.5) can be rewritten as

$$\begin{aligned} \min_{\mathbf{x}} \max_{i} \quad & \frac{\mathbf{x}^* W_i \mathbf{x}}{x^2_{(i-1)M+i}} \\ \text{subject to} \quad & x_{(i-1)M+i} > 0, \qquad 1 \leq i \leq M \\ & x_{M^2+(i-1)M+i} = 0, \qquad 1 \leq i \leq M, \end{aligned} \tag{4.6}$$

where $W_i = \sigma^2 T + \sum_{j=1,j\neq i}^{M} K^{(ij)}$. Note that W_i is positive semi-definite because matrices T and $K^{(ij)}$ are positive semi-definite. To solve (4.6), we first prove the following lemma.

Lemma 4.2. *The optimization problem (4.6) is quasi convex.*

Proof. We first need to prove that function $f_i(\mathbf{x}) = \frac{\mathbf{x}^* W_i \mathbf{x}}{x^2_{(i-1)M+i}}$ is quasi convex. We can write $f_i(\mathbf{x}) = \frac{g_i(\mathbf{x})}{x_{(i-1)M+i}}$, where $g_i(\mathbf{x}) = \frac{\mathbf{x}^* W_i \mathbf{x}}{x_{(i-1)M+i}}$. Let us show that the function $g_i(\mathbf{x})$ is convex for $x_{(i-1)M+i} > 0$. To do so, we need to show that $g_i(\theta \mathbf{x} + \gamma \mathbf{y}) \leq \theta g_i(\mathbf{x}) + \gamma g_i(\mathbf{y})$, where $\theta + \gamma = 1, 0 \leq \theta, \gamma \leq 1$. This is equivalent to showing that $\frac{y_{(i-1)M+i}}{x_{(i-1)M+i}} \mathbf{x}^* W_i \mathbf{x} - 2\mathbf{x}^* W_i \mathbf{y} + \frac{x_{(i-1)M+i}}{y_{(i-1)M+i}} \mathbf{y}^* W_i \mathbf{y} \geq 0$. Since W_i is symmetric and positive semi-definite, it can be written

as $W_i = R_i^* R_i$. From the Cauchy-Schwartz inequality it follows that $\mathbf{x}^* W_i \mathbf{y} = \mathbf{x}^* R_i^* R_i \mathbf{y} \leq ||R_i \mathbf{x}||_2 ||R_i \mathbf{y}||_2 = \sqrt{\mathbf{x}^* W_i \mathbf{x} \mathbf{y}^* W_i \mathbf{y}}$, from which it follows that

$$\frac{y_{(i-1)M+i}}{x_{(i-1)M+i}} \mathbf{x}^* W_i \mathbf{x} - 2\mathbf{x}^* W_i \mathbf{y} + \frac{x_{(i-1)M+i}}{y_{(i-1)M+i}} \mathbf{y}^* W_i \mathbf{y}$$
$$\geq \left(\sqrt{\frac{y_{(i-1)M+i}}{x_{(i-1)M+i}} \mathbf{x}^* W_i \mathbf{x}} - \sqrt{\frac{x_{(i-1)M+i}}{y_{(i-1)M+i}} \mathbf{y}^* W_i \mathbf{y}} \right)^2 \geq 0.$$

Therefore, function $g_i(\mathbf{x})$ is convex for $x_{(i-1)M+i} > 0$. Since the ratio of a convex and a linear function is quasi convex, and since the pointwise maximum of quasi convex functions is quasi convex (see, e.g., [12]), we conclude that the objective function in (4.6) is quasi convex. □

Remark: When preparing the final version of [93], we became aware of related work [107], where the authors deal with a similar problem. There they present another proof of the quasi convexity of (4.6), using a different approach.

We use the bisection method combined with the interior-point method (implemented in software package SeDuMi) to solve (4.6). Once we find the optimal $\mathbf{x}$ in (4.6), we determine B such that $\mathbf{x} = \begin{bmatrix} \Re(\mathbf{vec}(B)) \\ \Im(\mathbf{vec}(B)) \end{bmatrix}$. Then we calculate G as $G = H^{-1}B$. We refer to using the matrix G found by the aforementioned procedure as Method 2.2. Figure 4.4 shows the comparison of max-min rate of the Method 2.2 and max-min rate of regularized pseudo-inverse and plain-channel inverse. It also shows an upper bound on the value of the achievable max-min rate obtained by dividing the sum-rate capacity by the number of users.

The technique described in Section 4.2.1 maximizes the sum rate of the multi-antenna broadcast system under the linear data processing constraint. The individual rates resulting from the maximization (4.3), however, may differ significantly. This disparity is inherent to the optimization (4.3), since (4.3) essentially denotes the maximization of $\|\mathbf{v}\|_1$ (i.e., norm-1 of the vector $\mathbf{v}$). It is well known that in the process of maximizing the norm-1 of a vector, a few components of the vector are suppressed while the remaining ones are boosted up.

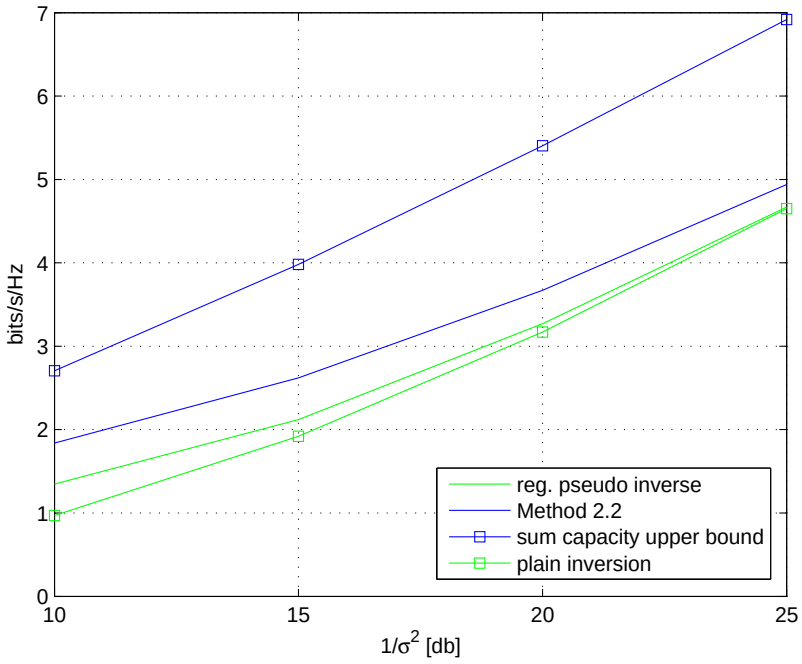


Figure 4.4: Comparison of the max-min rate of Method 2.2 to the max-min rate of reg. pseudo-inverse and to the upper bound obtained from the sum capacity of broadcast channel, $M = 6$ antennas/users

Thus in Section 4.2.1 the sum rate is maximized at the expense of the weakest few users, who are ignored. [Note: Transmitting data over many channel uses provides fairness.] The symbols intended for the remaining strong users may be modulated with higher modulation schemes, thus overcompensating for the sum rate seemingly lost by transmitting only to a subset of users.

On the other hand, as a result of the disparity among the individual rates (and hence among the SINRs and BERs of individual users), the average BER of the system may suffer. To compensate for the loss in average BER, we employ Method 2.2 on the subset of strong users selected for transmission by Method 2.1. We formalize this combination of Method 2.1 and Method 2.2 in the following way

1. Obtain G using Method 2.1.

2. Denote the set of indices which correspond to zero-columns of G by $\mathcal{I}_0$.

3. Denote a submatrix of H comprised of rows $1 \leq i \leq M$, $i \notin \mathcal{I}_0$ by H_{sub}.

4. Apply Method 2.2 on H_{sub} to obtain B; set $G = H_{\text{sub}}^*(H_{\text{sub}}H_{\text{sub}}^*)^{-1}B$.

As it turns out, maximizing the minimum individual rate among the selected strong users results in fairly equal (and high) SINRs. We refer to the previous combination of Method 2.1 and Method 2.2 as Method 2.

4.3 Finding the optimal scaling coefficient k

We start with the basic model (4.1) and assume that the preprocessing matrix G is obtained by simple inversion of the channel matrix H, i.e., $G = H^{-1}$. For this choice of G, in this section, we propose a way of scaling the magnitudes of the information signal $\mathbf{u}$ so as to minimize the average BER. [Note that in Section 4.4 we will show how to employ this signal scaling technique to the more general case of the optimal G obtained in Section 4.2.]

To minimize the average BER, one needs to maximize the minimum SINR at receivers. To this end, in [75] authors suggest perturbation of information signals by appropriately translating original M-QAM signal constellation in complex space. In this section we suggest a similar idea but focus on perturbations (in fact, radial scaling) of M-PSK constellation. An advantage of constraining ourselves to PSK constellations is in the simplicity of decoding. Since the signal points are perturbed only radially, rather than vertically or horizontally as in QAM, the angular information has not changed. Therefore, no side information about the signaling scheme (i.e., the nature of the perturbation) is needed at the receiver. In other words, each user's decoder makes simple angular decisions. The decoder is no longer necessarily ML but it is efficient and practical since it requires no additional information from the transmitter. [Our simulation results indicate that the performance of this sub-optimal ML decoder is almost identical to the optimal one.]

By fixing $G = H^{-1}$ and representing $\mathbf{u}$ via its phases and magnitudes, we can rewrite (4.1) as

$$\mathbf{r} = kHH^{-1}\Phi\mathbf{u}_m + \mathbf{w}, \tag{4.7}$$

where $\mathbf{u} = \Phi \mathbf{u}_m$ and Φ is the diagonal matrix of phases of $\mathbf{u}$, and where $\mathbf{u}_m$ is the vector of magnitudes of $\mathbf{u}$. Note that due to the use of a PSK modulation scheme, the information to be transmitted is contained in Φ. We are concerned with designing optimal magnitudes of the signals, i.e., designing the $\mathbf{u}_m$. The relevant power constraint now becomes the one on instantaneous, rather than average, transmission power. This means that the corresponding form to (4.2) can be written as

$$\mathbf{r} = \frac{HH^{-1}\Phi \mathbf{u}_m}{\sqrt{\mathbf{u}_m^* \Phi^* H^{-*} H^{-1} \Phi \mathbf{u}_m}} + \mathbf{w} \tag{4.8}$$

where $\mathbf{u}_m^* \Phi^* H^{-*} H^{-1} \Phi \mathbf{u}_m = \frac{1}{k^2}$. Now we want to optimize the scaling coefficient while keeping magnitudes of $\mathbf{u}$ greater or equal to 1. Effectively we want to move signals $\mathbf{u}_m$ radially away from the origin (see Figure 4.5). This will result in magnitudes of the components of the received vector $\mathbf{r}$ that are at least as large as if there were no signal scaling at all. This requires solving the following optimization problem

$$\begin{aligned} \min \quad & \mathbf{u}_m^* \Phi^* H^{-*} H^{-1} \Phi \mathbf{u}_m \\ \text{subject to} \quad & u_{m_i} \geq 1, \qquad 1 \leq i \leq M. \end{aligned} \tag{4.9}$$

This problem is convex and can easily be solved exactly by a host of numerical methods (see, e.g., [12] and the references therein). More importantly, we can show that the solution of this problem is equal to the solution to

$$\begin{aligned} \max_{u_{m_1}, u_{m_2}, \ldots, u_{m_M}} \min_i \quad & \frac{u_{m_i}^2}{\mathbf{u}_m^* \Phi^* H^{-*} H^{-1} \Phi \mathbf{u}_m} \\ \text{subject to} \quad & u_{m_i} \geq 1, \qquad 1 \leq i \leq M, \end{aligned} \tag{4.10}$$

which is the problem of maximizing the minimum SINR in system (4.8). Denoting by $\widehat{\mathbf{u}_m}$ the solution to (4.10), we see that the transmitted signal should have the form of $\mathbf{s} = \frac{H^{-1}\Phi\widehat{\mathbf{u}_m}}{\sqrt{\widehat{\mathbf{u}_m}^* \Phi^* H^{-*} H^{-1} \Phi \widehat{\mathbf{u}_m}}}$. We refer to this signal scaling policy as Method 3. As mentioned earlier, although the magnitudes of optimal $\mathbf{u}$ will generally be different than 1, the receivers

will still be able to decode the received signals by considering their angle, since $\mathbf{s}$ has the same phase matrix Φ as $\mathbf{u}$.

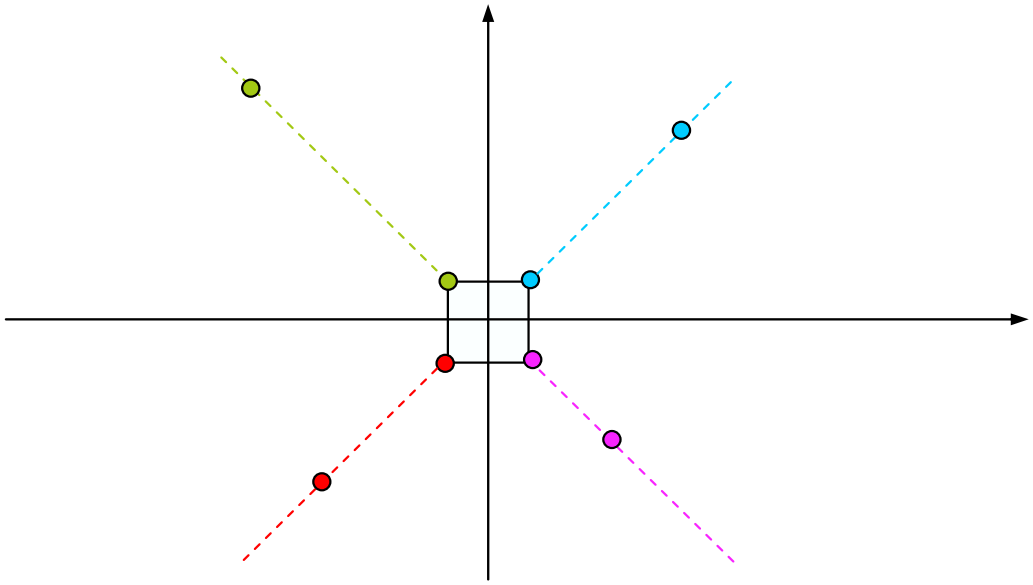


Figure 4.5: A sketch of radial signal scaling

4.4 Combined method

In Section 4.3, we employed the signal scaling scheme to optimize the BER in a system that uses $G = H^{-1}$ for data preprocessing. In this section, we combine the signal scaling with the optimal preprocessing matrices G found in Section 4.2. This is done in stages. In particular, assume that Method 2.1 is used to find G which maximizes the sum rate of the channel. Then, to minimize the average BER of the users, we employ signal scaling for such G. Instead of solving (4.9) (which assumed $G = H^{-1}$), we now need to solve optimization

$$\begin{aligned} \min \quad & \mathbf{u}_m^* \Phi^* \widehat{G}^* \widehat{G} \Phi \mathbf{u}_m \\ \text{subject to} \quad & u_{m_i} \geq 1, \qquad 1 \leq i \leq M \end{aligned} \tag{4.11}$$

where $\widehat{G}$ is G found by Method 2.1. The above problem is convex and thus can be solved exactly via efficient convex optimization techniques. If we denote solution of (4.11) by $\widehat{\mathbf{u}_m}$, the optimal transmitted signal $\mathbf{s}$ is given by $\mathbf{s} = \frac{\widehat{G}\Phi\widehat{\mathbf{u}_m}}{\sqrt{\widehat{\mathbf{u}_m}^*\Phi^*\widehat{G}^*\widehat{G}\Phi\widehat{\mathbf{u}_m}}}$. We refer to the above algorithm as Method 4.

4.5 Simulation results

In this section we briefly discuss simulation results of the suggested methods for linear preprocessing. Figures 4.6 and 4.7 show that Method 2.1 performs at least as well as the regularized pseudo-inverse in terms of BER while, due to the use of a higher modulation scheme, provides significantly higher sum rate. Figure 4.6 also shows that Method 2, due to the additional minmax optimization of SINRs, performs even better than Method 2.1 in terms of BER. Figure 4.8 shows that the simple scaling strategy gives a better BER performance than the pseudo inverse. Finally, Figures 4.9 and 4.10 show that both Method 2 and Method 4, outperform the pseudo-inverse in terms of both the BER and the sum rate. All plots were done using uncoded sequences of information bits at the transmitter, modulated with symbols from standard PSK constellations as denoted below the figures.

4.6 Conclusion

In this chapter, we have proposed two criteria for the design of the precoding matrix in a multi-antenna broadcast system. First, we maximized the sum rate, and then we showed how to maximize the minimum rate among all users. The latter problem is shown to be quasi convex and solved exactly. The precoding techniques are constrained to linear preprocessing at the transmitter. In addition to precoding, we have employed a signal scaling scheme that minimizes the average BER of the users. The signal scaling scheme is posed as a convex optimization problem, and solved exactly via interior-point methods. Finally, we have combined the precoding with signal scaling. The combined scheme can be efficiently applied in practice. In terms of the achievable sum rate, the proposed technique

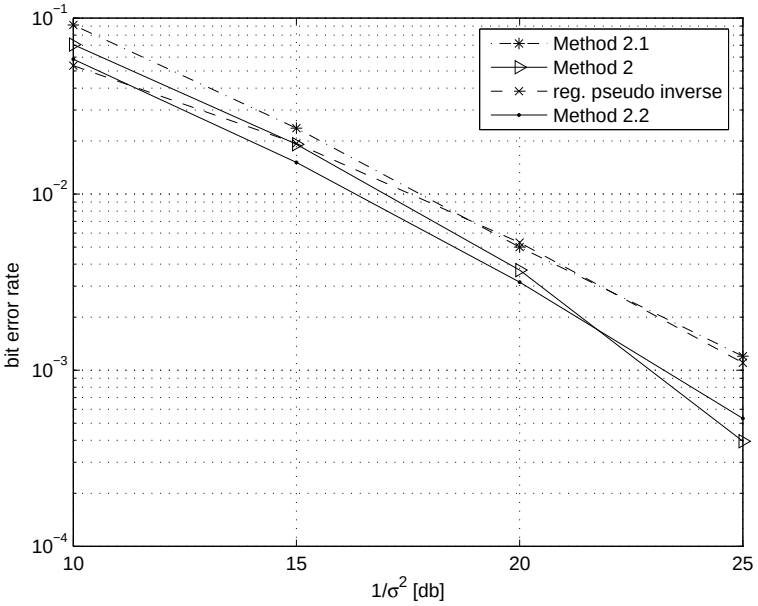


Figure 4.6: Comparison of BER, M=6 antennas/users, 8PSK-Method 2, 8PSK-Method 2.1, 4-PSK-Method 2.2, 4PSK-regularized pseudo-inverse

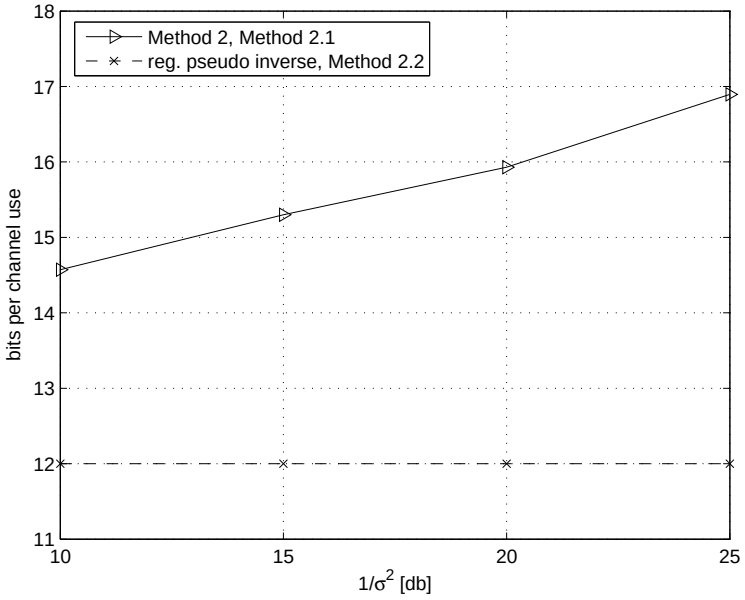


Figure 4.7: Comparison of rates, M=6 antennas/users, 8PSK-Method 2, 8PSK-Method 2.1, 4PSK-MEthod 2.2, 4PSK-regularized pseudo-inverse

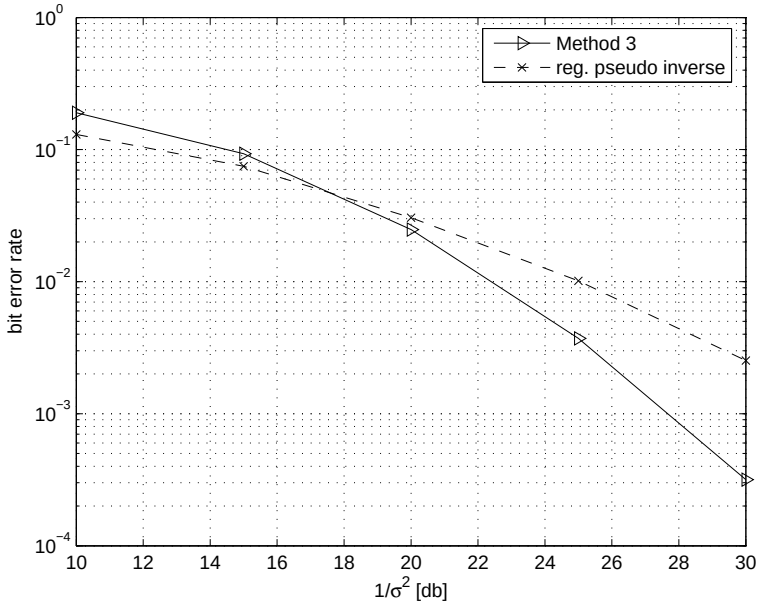


Figure 4.8: Comparison of BER, M=20 antennas/users, 8PSK-Method 3, 8PSK-regularized pseudo-inverse

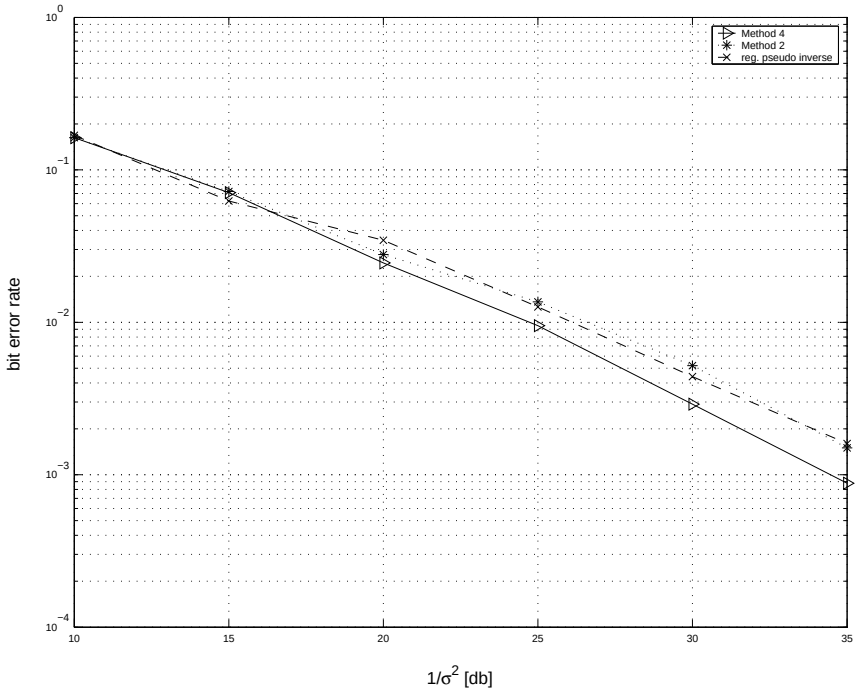


Figure 4.9: Comparison of BER, M=6 antennas/users, 16PSK-Method 4, 16PSK-Method 2, 8PSK-regularized pseudo-inverse

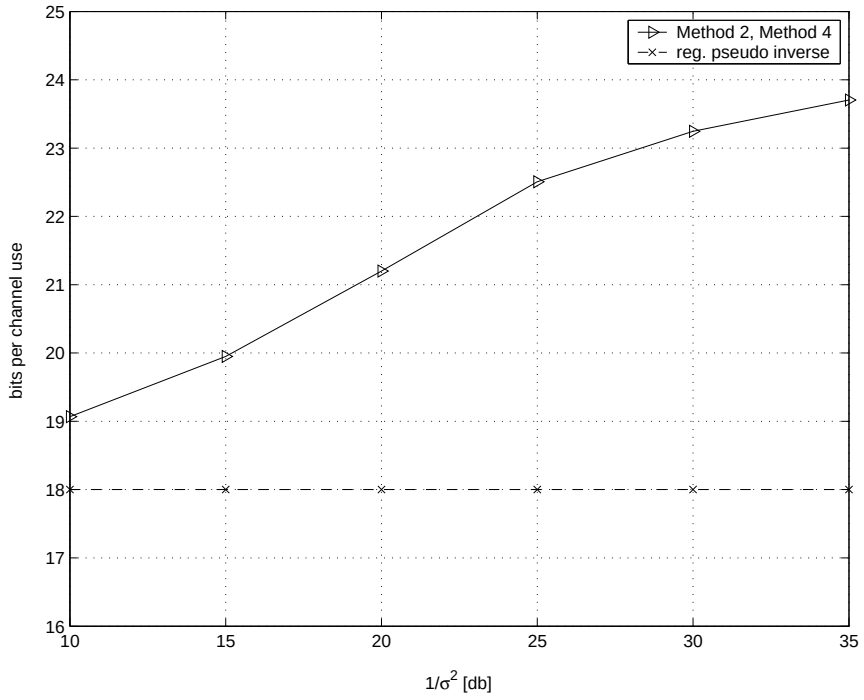


Figure 4.10: Comparison of rates, M=6 antennas/users, 16PSK-Method 4, 16PSK-Method 2, 8PSK-regularized pseudo-inverse

significantly outperforms traditional channel inversion methods, while having comparable (in fact, often superior) BER performance.

Chapter 5

Gaussian Broadcast Channel — Asymptotic Analysis of a Particular Nonlinear Scheme

As we have said in the previous chapter, the sum-rate capacity of the multi-antenna Gaussian broadcast channel has recently been computed. However, the search for computationally efficient practical schemes that achieve it is still in progress. When the channel state information is fully available at the transmitter, the dirty-paper coding (DPC) technique is known to achieve the maximal throughput, but is computationally infeasible. In this chapter, we analyze the asymptotic behavior of one of its alternatives — the recently suggested so-called vector-perturbation technique. We show that for a square channel, where the number of users is large and equal to the number of transmit antennas, its sum rate approaches that of the DPC technique. More precisely, we show that at both low and high signal-to-noise ratio (SNR), the scheme under consideration is asymptotically optimal. Furthermore, we obtain similar results in the case where the number of users is much larger than the number of transmit antennas.

5.1 Introduction

The limits of performance of multi-antenna Gaussian broadcast channel are currently the subject of extensive research (see, e.g., [14], [101], and the references therein). As we have mentioned in the previous chapter, when the channel state information (CSI) is fully

available at the transmitter, the so-called dirty-paper coding (DPC) technique achieves the capacity of multi-antenna broadcast channel [105]. However, the DPC scheme is exponentially complex and appears to be difficult to implement in practical systems. To this end, various heuristics with suboptimal performance but efficient implementation have recently been proposed. In [35], vector quantization is used in combination with powerful coding schemes to achieve a large fraction of the promised capacity. In [75], a technique referred to as the vector-perturbation technique (VPT) was proposed, and further considered in [108]. Simulation results presented there indicate that the proposed technique achieves performance close to the optimal one.

In this chapter, we analyze the theoretical limits of the VPT [75]. In particular, we show that when the number of users in the broadcast system is large, the sum rate achievable by the VPT approaches the sum rate achievable by the DPC scheme, both in the low- and the high-SNR regime. While the scheme introduced in [75] and further studied in [108] is practically feasible, the worst-case complexity of its implementation is still exponential. On the other hand, our proof for lower bounding the asymptotical sum-rate performance of the VPT is constructive and based on an algorithm that is polynomial in the number of users. A complementary version of this chapter can be found in [91].

We assume the standard broadcast channel model similar to the one considered in the previous chapter,

$$\mathbf{y} = H\mathbf{s} + \mathbf{v}, \tag{5.1}$$

where H is a $K \times M$ matrix whose entries are independent, identically distributed (i.i.d.) complex Gaussian random variables $\mathcal{C}_{\mathcal{N}}(0,1)$, K is the number of users, M is the number of transmit antennas, $\mathbf{v}$ is a $K \times 1$ noise vector whose entries are independent of entries in H and i.i.d. Gaussian random variables with zero-mean and $\sigma^2 = 1/\rho$ variance, and $\mathbf{s}$ is an $M \times 1$ vector which is transmitted over the channel. Furthermore, we impose the constraint $E\|\mathbf{s}\|^2 = 1$; hence, the receivers do not need to know instantiations of the channel. (The case $\|\mathbf{s}\|^2 = 1$, considered in [75], can be treated similarly and leads to similar results.)

Since we focus on analyzing the asymptotic performance of the vector-perturbation

technique, we start by reviewing it in the next section.

5.2 The vector-perturbation technique

Following [75], we consider the scenario where the number of antennas on a transmitter is equal to the number of users, i.e., $K = M$. [Later in this chapter we will consider the vector-perturbation technique for $K \gg M$ and generalize our results to that case.] Furthermore, we assume that the entries of the $K \times 1$ symbols vector $\mathbf{u}$ intended for the users are the points in a QAM constellation.

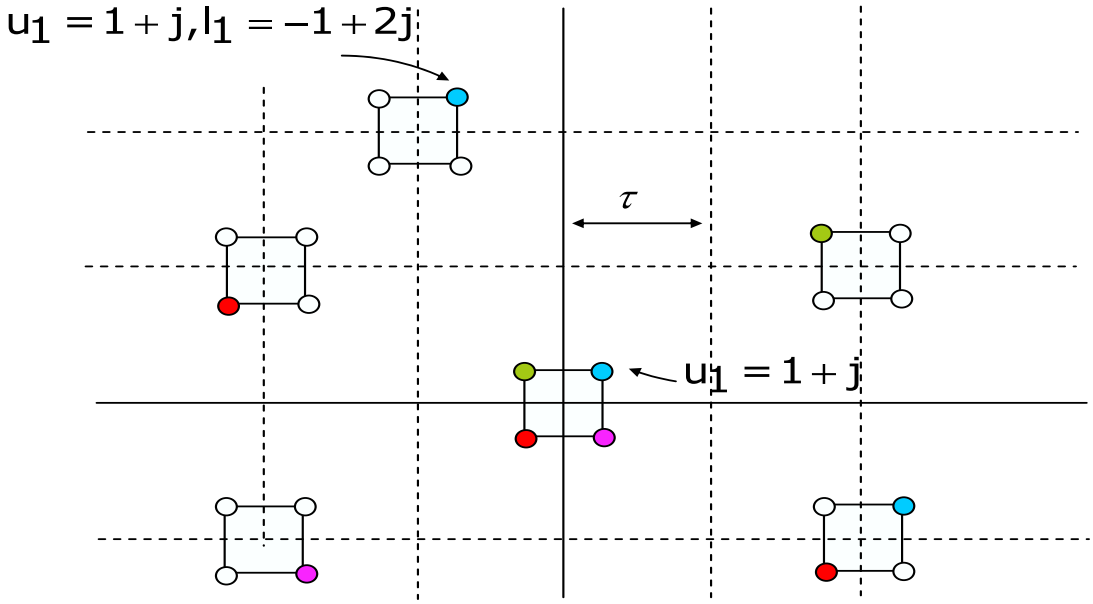


Figure 5.1: VPT scheme

The vector-perturbation technique [75] (see Figure 5.1) relates the transmitted vector $\mathbf{s}$ to the information vector $\mathbf{u}$ as follows,

$$\mathbf{s} = \frac{H^{-1}(\mathbf{u} + \tau\hat{\mathbf{l}})}{\sqrt{E||H^{-1}(\mathbf{u} + \tau\hat{\mathbf{l}})||^2}}, \tag{5.2}$$

where τ is an *a priori* determined positive constant, and where $\hat{\mathbf{l}}$ is the solution to the

following optimization problem

$$\hat{\mathbf{l}} = \text{argmin}_{\mathcal{R}(\mathbf{l})\in Z^M, \mathcal{I}(\mathbf{l})\in Z^M} ||H^{-1}(\mathbf{u}+\tau\mathbf{l})||^2. \tag{5.3}$$

Note that $\mathcal{R}(\mathbf{l})$ and $\mathcal{I}(\mathbf{l})$ denote the real and the imaginary part of the vector $\mathbf{l}$, respectively. The main idea behind (5.3) is to eliminate (or to minimize) the power penalty which happens in the case when the so-called zero-forcing (ZF) scheme (obtained for $\hat{\mathbf{l}} = 0$ in (5.2)) is applied.

The signals received by the k^{th} user are of the form

$$\mathbf{y}_i = \frac{u_i + \tau\hat{l}_i}{\sqrt{E||L^{-1}(\mathbf{u}+\tau\hat{\mathbf{l}})||^2}} + v_i, \quad 1 \le i \le K, \tag{5.4}$$

where L is a lower-triangular matrix in the LQ-decomposition of H, i.e., $H = LQ$, and Q is a unitary matrix. Decoding of these signals is simple, and the only processing required from the receivers is scaling by $\sqrt{E||H^{-1}(\mathbf{u}+\tau\hat{\mathbf{l}})||^2}$ [75].

5.3 Case $K = M$

In this section, we analyze the VPT for $K = M$. Before proceeding any further, we slightly modify the perturbation technique as follows. Let D be a diagonal matrix such that

$$D = \text{diag}(L_{1,1}^{1+\beta}, L_{2,2}^{1+\beta}, \ldots, L_{n,n}^{1+\beta}), \tag{5.5}$$

where β is any integer such that $\beta \ge 0$. Instead of transmitting $\mathbf{s}$ as given by (5.2), we define $\mathbf{s}$ to be

$$\mathbf{s} = \frac{H^{-1}D(\mathbf{u}+\tau\hat{\mathbf{l}})}{\sqrt{E||H^{-1}D(\mathbf{u}+\tau\hat{\mathbf{l}})||^2}}. \tag{5.6}$$

Consequently, the signal received at the i^{th} user becomes

$$\mathbf{y}_i = L_{i,i}^{1+\beta} \frac{u_i + \tau \hat{l}_i}{\sqrt{E||L^{-1}D(\mathbf{u} + \tau \hat{\mathbf{l}})||^2}} + v_i,\ 1 \leq i \leq K. \tag{5.7}$$

We refer to this scheme as the diagonal vector perturbation technique (DVPT). From (5.7) it follows that the sum rate of the DVPT can then be computed as a summation of the sum rates of K decoupled channels,

$$R_{DVPT} = E \sum_{i=1}^{K} \log\Big(1 + L_{i,i}^{2(1+\beta)} \frac{\rho||u_i + \tau \hat{l}_i||^2}{E||L^{-1}D(\mathbf{u} + \tau \hat{\mathbf{l}})||^2}\Big). \tag{5.8}$$

We are interested in bounding the value of R_{DVPT}; to this end, it will be useful to first derive a few inequalities.

Although the use of the VPT on broadcast channels performs well in practice, it requires solving (5.3), which is NP-hard. Use of the sphere decoding (or any other) algorithm may often be infeasible. Therefore, we employ a heuristic nulling and canceling [40] technique to solve (5.3). To this end, let us denote $B = L^{-1}D$. Clearly, $B_{i,i} = L_{i,i}^{\beta}$. Generate $\mathbf{l}^{lb} = \mathbf{l}^{lbr} + \jmath \mathbf{l}^{lbc}$ according to the nulling and canceling procedure as follows,

$$\begin{aligned}
\mathbf{l}_1^{lbr} &= \lfloor -\frac{\mathcal{R}(u_1)}{\tau} \rfloor \\
\mathbf{l}_1^{lbc} &= \lfloor -\frac{\mathcal{I}(u_1)}{\tau} \rfloor \\
\mathbf{l}_2^{lbr} &= \lfloor -\frac{B_{2,2}\mathcal{R}(u_2) + \mathcal{R}(B_{2,1}(u_1 + \tau l_1^{lb}))}{B_{2,2}\tau} \rfloor \\
\mathbf{l}_2^{lbc} &= \lfloor -\frac{B_{2,2}\mathcal{I}(u_2) + \mathcal{I}(B_{2,1}(u_1 + \tau l_1^{lb}))}{B_{2,2}\tau} \rfloor \\
&\vdots \\
\mathbf{l}_K^{lbr} &= \lfloor -\frac{B_{K,K}\mathcal{R}(u_K) + \sum_{i=1}^{K-1} \mathcal{R}(B_{K,i}(u_i + \tau l_i^{lb}))}{B_{K,K}\tau} \rfloor \\
\mathbf{l}_K^{lbc} &= \lfloor -\frac{B_{K,K}\mathcal{I}(u_K) + \sum_{i=1}^{K-1} \mathcal{I}(B_{K,i}(u_i + \tau l_i^{lb}))}{B_{K,K}\tau} \rfloor.
\end{aligned}$$

Since for any two real numbers a and b holds that $|a - b\lfloor \frac{a}{b} \rfloor|^2 \leq b^2$, we obtain

$$|u_i + \tau l_i^{lb}|^2 \leq 2\tau^2 B_{i,i}^2. \tag{5.9}$$

Careful examination of the previous procedure reveals that any time we obtain $l_i^{lbr} = 0$, we can change it to either $l_i^{lbr} = 1$ or $l_i^{lbr} = -1$ and still preserve the validity of (5.9). Thus in addition to (5.9), we can also establish a lower bound on $|u_i + \tau l_i^{lb}|^2$ depending on the sign of u_i or l_i^{lb},

$$\begin{aligned} |u_i + \tau l_i^{lb}|^2 &\geq |-|\mathcal{R}(u_i)| - j|\mathcal{I}(u_i)| + \tau(1+j)|^2 \\ &\geq |-\max_i|\mathcal{R}(u_i)| - j\max_i|\mathcal{I}(u_i)| + \tau(1+j)|^2 = \zeta. \end{aligned} \tag{5.10}$$

We may now begin our derivation of a bound on R_{DVPT}. To facilitate fluent presentation, we first treat the low-SNR case, and then generalize the results to any SNR.

5.3.1 Low SNR regime ($\rho \to 0$)

For $\rho \to 0$, we have

$$\begin{aligned} R_{DVPT} &= E\sum_{i=1}^{K} \log\Big(1 + L_{i,i}^{2(1+\beta)} \frac{\rho||u_i + \tau\hat{l}_i||^2}{E||L^{-1}D(\mathbf{u} + \tau\hat{\mathbf{l}})||^2}\Big) \\ &= K + \rho\frac{\sum_{i=1}^{K} EL_{i,i}^{2(1+\beta)}||u_i + \tau\hat{l}_i||^2}{E||B(\mathbf{u} + \tau\hat{\mathbf{l}})||^2} + \mathcal{O}(\rho^2). \end{aligned}$$

Using (5.10) to lower bound the numerator and (5.9) to upper bound the denominator of the fraction in the expression above, and using the fact that $L_{i,i}^2$ are i.i.d. random variables with $\chi^2_{2(K-i+1)}$ distribution, we have

$$\begin{aligned} R_{DVPT} &\geq K + \rho \frac{\zeta \sum_{i=1}^{K} E(L_{i,i}^2)^{1+\beta}}{2\tau^2 \sum_{i=1}^{K} E(L_{i,i}^2)^{\beta}} + \mathcal{O}(\rho^2) \\ &= K + \rho \frac{\zeta \sum_{i=1}^{K} \prod_{k=1}^{1+\beta} (2(K+1-i)+2(k-1))}{2\tau^2 \sum_{i=1}^{K} \prod_{k=1}^{\beta} (2(K+1-i)+2(k-1))} + \mathcal{O}(\rho^2) \\ &\geq K + \rho \frac{\zeta \sum_{i=1}^{K} (2(K+1-i))^{1+\beta}}{2\tau^2 \sum_{i=1}^{K} (2(K+1-i)+2(\beta-1))^{\beta}} + \mathcal{O}(\rho^2). \end{aligned} \tag{5.11}$$

Therefore, we can write

$$\lim_{K\to\infty} \frac{R_{DVPT}}{K} \geq 1 + 2\rho \frac{\zeta(1+\beta)}{2\tau^2(2+\beta)}. \tag{5.12}$$

Let w be the width of a QAM constellation, i.e., let $w = 2\max_{\mathbf{u}} \max\{\max_i |\mathcal{R}(u_i)|, \max_i |\mathcal{I}(u_i)|\}$. Clearly, $\zeta \geq 2(\tau - \frac{w}{2})^2$. Then for $\tau \gg \frac{w}{2}$, and for $\beta \to \infty$, $K \gg \beta$, we can write

$$\lim_{K\to\infty} \frac{R_{DVPT}}{K} \geq 1 + 2\rho. \tag{5.13}$$

The results stated in (5.11), (5.12), and (5.13), imply that the sum rate of the diagonal vector-perturbation technique scales linearly with the number of users. In fact, this result may be established directly from (5.11). However, in order to tighten the coefficients in front of ρ, we derived (5.13) as well.

We summarize our results in the following theorem.

Theorem 5.1. *Consider communication in low-SNR regime ($\rho \to 0$) over a square Gaussian broadcast channel using the diagonal vector-perturbation technique with parameters $\beta \geq 1$ and $\tau \geq w$, where w is the width of a QAM constellation. Then*

$$\lim_{K\to\infty} \frac{R_{DVPT}}{K} \geq 1 + 2\rho(1 - \frac{w}{2\tau})^2 \frac{1+\beta}{2+\beta}.$$

Proof: Follows from the discussion above. □

Corollary 5.1. *Let all assumptions of Theorem 5.1 hold. Furthermore, let* $\tau \gg w$, $\beta \to \infty$, *and* $\frac{K}{\beta} \gg 1$. *Then*

$$\lim_{K\to\infty} \frac{R_{DVPT}}{K} \geq 1 + 2\rho.$$

5.3.2 General SNR

For simplicity, in this subsection we fix $\beta = 0$. Similar to the procedure in Section 5.3.1, we use (5.10) to lower bound the numerator and (5.9) to upper bound the denominator of the fraction in the expression given in (5.8),

$$\begin{aligned} R_{DVPT} &= E\sum_{i=1}^{K} \log\Big(1 + L_{i,i}^2 \frac{\rho||u_i + \tau\hat{l}_i||^2}{E||L^{-1}D(\mathbf{u} + \tau\hat{\mathbf{l}})||^2}\Big) \\ &\geq E\sum_{i=1}^{K-1} \log(1 + \rho\frac{\zeta\hat{L}_{i,i}^2}{2\tau^2 K}) \\ &= E\log\prod_{i=1}^{K-1}(1 + \rho\frac{\zeta}{2\tau^2 K}L_{i,i}^2). \end{aligned}$$

Applying the arithmetic-geometric mean inequality, it can easily be shown that

$$\prod_{i=1}^{K-1}(1 + \rho\frac{\zeta}{2\tau^2 K}L_{i,i}^2) \geq (1 + \rho\frac{\zeta}{2\tau^2 K}(\prod_{i=1}^{K-1} L_{i,i}^2))^{\frac{1}{K-1}})^{K-1}.$$

Then, we have

$$\begin{aligned} R_{DVPT} &\geq (K-1)E\log(I + \rho\frac{\zeta}{2\tau^2 K}(\prod_{i=1}^{K-1} L_{i,i}^2)^{\frac{1}{K-1}}) \\ &\geq (K-1)\log\Big(1 + \rho\frac{\zeta\prod_{i=1}^{K-1}(E((L_{i,i}^2)^{-1}))^{-\frac{1}{K-1}}}{2\tau^2 K}\Big). \end{aligned} \tag{5.14}$$

Using the fact that $L_{i,i}^2$ are i.i.d. random variables with $\chi^2_{2(K-i+1)}$ distributions, and the Stirling's formula to approximate the factorial, we obtain

$$\begin{aligned} R_{DVPT} &\geq (K-1)\log\Big(1+\frac{\rho\zeta(\prod_{i=2}^{K}2(K-i+1))^{\frac{1}{K-1}}}{2\tau^2K}\Big) \\ &\geq (K-1)\log(1+\frac{2\rho\zeta(K-1)!^{\frac{1}{K-1}}}{2\tau^2K}) \\ &\geq (K-1)\log\Big(1+\frac{2\rho\zeta}{2\tau^2e}\Big) \\ &\geq (K-1)\log\Big(1+\frac{2\rho}{e}(1-\frac{w}{2\tau})^2\Big). \end{aligned} \tag{5.15}$$

It is worth pointing out that for $\rho \to \infty$ we can also upper bound the value of R_{DVPT}. Instead of (5.14), using Jensen's and the arithmetic-geometric mean inequalities we can write

$$\begin{aligned} R_{DVPT} &\leq E\log\Big(\rho^K\frac{\prod_{i=1}^{K}L_{i,i}^2\prod_{i=1}^{K}||u_i+\tau\hat{l}_i||^2}{(\sum_{i=1}^{K}||u_i+\tau\hat{l}_i||^2)^K}\Big) \\ &\leq \log\Big(\rho^K\frac{\prod_{i=1}^{K}EL_{i,i}^2}{K^K}\Big) \\ &\leq \log\Big((2\rho)^K\frac{K!}{K^K}\Big) \\ &\leq K\log(\frac{2\rho}{e})+\mathcal{O}(\log K). \end{aligned} \tag{5.16}$$

The results from this subsection are summarized in the following theorem.

Theorem 5.2. *Consider communication over square Gaussian broadcast channel using the diagonal vector-perturbation technique with parameters $\beta = 0$ and $\tau > \frac{w}{2}$, where w is the width of a QAM constellation. Then*

$$\lim_{K\to\infty}\frac{R_{DVPT}}{K\,log\Big(1+\frac{2\rho}{e}(1-\frac{w}{2\tau})^2\Big)} \geq 1.$$

Proof: Follows from (5.15). □

Corollary 5.2. *Let assumptions of Theorem 5.2 hold. Also let $\rho \to \infty$. Then*

$$\lim_{K\to\infty} \frac{R_{DVPT}}{K log\rho} = 1.$$

Theorems 5.1 and 5.2 imply that when the diagonal vector-perturbation technique (with appropriate parameters) is employed for communication over Gaussian broadcast channel, the sum rate scales linearly with the number of users. Furthermore, in high-SNR regime the scaling law is not only linear in the number of users, but also optimal, i.e., equal to that of the capacity-achieving DPC technique.

5.4 Case $K \gg M$

In this section, we study the asymptotic behavior of the VPT and DVPT schemes for $K \gg M$. We should point out that this regime (in particular, the case $\frac{\log K}{M} \geq const., K \to \infty$) was considered in [83], where it was shown that, in limit, the maximum throughput may be achieved with only partial CSI at the transmitter. The VPT and DVPT, on the other hand, require full CSI; however, since we have shown that these simple schemes asymptotically achieve the maximum throughput when the number of transmit antennas and users is the same, it is of interest to extend these results to the $K \gg M$ case as well.

A generalization of the results to the $K \gg M$ case is relatively straightforward. In particular, at any transmission interval we select a subset of M users to which we transmit. Define $H_{(k)} = H_{(k-1)M+1:kM,(k-1)M+1:kM}$. Let λ_k^{min} be the minimal eigenvalue of the matrix $H_{(k)}^* H_{(k)}$, and let $\xi = \arg\max_{k\in\{1,2,\ldots,\lfloor\frac{K}{M}\rfloor\}} \lambda_k^{min}$. Then we transmit to the users $(\xi-1)M+1, (\xi-1)M+2, \ldots, \xi M$, employing the DVPT with $H_{(\xi)}$. Let $\hat{L}$ be a lower-triangular matrix from the LQ-decomposition $H_{(\xi)} = \hat{L}Q$, where Q is unitary. Then, instead of (5.14) we can write

$$R_{DVPT} = E\sum_{i=1}^{M} \log(1 + \rho\frac{\zeta \hat{L}_{i,i}^2}{2\tau^2 M}) \geq M\log\Big(1 + \rho\frac{\zeta}{2\tau^2 M E((\lambda_\xi^{min})^{-1})}\Big).$$

Further, using results from extreme value theory, it can be shown that $\lim_{\frac{K}{M}\to\infty} E((\lambda_\xi^{min})^{-1}) \to \frac{M}{2\log\frac{K}{M}}$ (see, e.g., [83]). The results from this section are summarized in the following theorem.

Theorem 5.3. *Consider communication over tall Gaussian broadcast channel ($\frac{K}{M} \to \infty$) using the diagonal vector perturbation-technique with parameters $\beta = 0$ and $\tau > \frac{w}{2}$, where w is the width of a QAM constellation. Then*

$$\lim_{\frac{K}{M}\to\infty} \frac{R_{DVPT}}{M log\Big(1 + \frac{2\rho}{M^2} log\frac{K}{M}\Big)} \geq 1.$$

Proof: Follows from the discussion above. □

In case of high-SNR ($\rho \to \infty$) we have the following corollary.

Corollary 5.3. *Let assumptions of Theorem 5.3 hold. Also let $\rho \to \infty$. Then*

$$\lim_{\frac{K}{M},\rho\to\infty} \frac{R_{DVPT}}{M log(\rho log K)} \geq 1.$$

The previous corollary says that the sum rate of the VPT asymptotically achieves the same sum rate as the DPC.

Remark: We point out that, using the same selection of users as suggested in this section, it is easy to show that, under the assumptions of the previous corollary, even the ZF scheme asymptotically achieves the same sum rate as the DPC.

5.5 Conclusion

In this chapter, we studied the asymptotic performance of the achievable throughput on the Gaussian broadcast channel with vector-perturbation preprocessing. We derived explicitly the achievable sum rate scaling laws in the case when the perturbation preprocessing is

applied at the transmitter. As it turns out, those scaling laws match the already-known capacity-achieving scheme (DPC) scaling laws in the case when the CSI is available at the transmitter. Furthermore, unlike the DPC, our scheme is simple and can be implemented in polynomial time.

Chapter 6

Quantum Unambiguous Detection

In the previous chapters we considered the design and applications of different optimization and algorithmic techniques in the context of classical communications (multi-antenna systems, wireless broadcast channels). In this chapter we will focus on the quantum systems. More specifically, we will consider the problem of quantum detection.

In quantum systems, unlike in classical, the information is stored in special objects called quantum states. Namely a quantum state is a set of numbers which fully describes the quantum system. These numbers are usually stored in a vector called pure state [73]. Furthermore, the state of a quantum system can be a statistical mixture of pure states, in which case it is called mixed quantum state [73].

As we have said, in this and the following chapter we consider the quantum detection problem. In order to detect in which state a quantum system was prepared we need to construct a set of quantum measurements. We will consider a specific problem of designing an optimal set of measurements that distinguishes unambiguously between a collection of mixed quantum states. Using arguments of duality in vector space optimization, we derive necessary and sufficient conditions for an optimal measurement that maximizes the probability of correct detection. We show that the previous optimal measurements that were derived for certain special cases satisfy these optimality conditions. We then consider state sets with strong symmetry properties, and show that the optimal measurement operators for distinguishing between these states share the same symmetries, and can be computed very efficiently by solving a reduced-size semi-definite program.

6.1 Introduction

The problem of detecting information stored in the state of a quantum system is a fundamental problem in quantum information theory. Several approaches have emerged to distinguish between a collection of non-orthogonal quantum states. In one approach, a measurement is designed to maximize the probability of correct detection [54], [53], [16], [74], [6], [29], [30], [26]. A more recent approach, referred to as unambiguous detection [55], [25], [76], [56], [78], [17], [18], [32], is to design a measurement that with a certain probability returns an inconclusive result, but such that if the measurement returns an answer, then the answer is correct with probability 1. An interesting alternative approach for distinguishing between a collection of quantum states, which is a combination of the previous two approaches, is to allow for a certain probability of an inconclusive result, and then maximize the probability of correct detection [32].

We consider a quantum state ensemble consisting of m density operators $\{\rho_i, 1 \leq i \leq m\}$ on an n-dimensional complex Hilbert space $\mathcal{H}$, with prior probabilities $\{p_i > 0, 1 \leq i \leq m\}$. A pure-state ensemble is one in which each density operator ρ_i is a rank-one projector $\phi_i \phi_i^*$, where the vectors ϕ_i, though evidently normalized to unit length, are not necessarily orthogonal.

Chefles [17] showed that a necessary and sufficient condition for the existence of unambiguous measurements for distinguishing between a collection of pure quantum states is that the states are linearly independent. Necessary and sufficient conditions on the optimal measurement minimizing the probability of an inconclusive result were derived in [33]. The optimal measurement when distinguishing between a broad class of symmetric pure-state sets was also considered in [33].

The problem of unambiguous detection between *mixed* state ensembles has received considerably less attention. Rudolph *et al.* [82] show that unambiguous detection between mixed quantum states is possible as long as one of the density operators in the ensemble has a non-zero overlap with the intersection of the kernels of the other density operators. They then consider the problem of unambiguous detection between two mixed quantum

states, and derive upper and lower bounds on the probability of a conclusive result. They also develop a closed-form solution for the optimal measurement in the case in which both states have kernels of dimension 1.

In this chapter we develop a general framework for unambiguous state discrimination between a collection of mixed quantum states, which can be applied to any number of states with arbitrary prior probabilities. For our measurement we consider general positive operator-valued measures [53], [77], consisting of $m+1$ measurement operators. We derive a set of necessary and sufficient conditions for an optimal measurement that minimizes the probability of an inconclusive result by exploiting principles of duality theory in vector space optimization. We then show that the previous optimal measurements that were derived for certain special cases satisfy these optimality conditions.

Next, we consider geometrically uniform (GU) and compound GU state sets [29], [30], [28], which are state sets with strong symmetry properties. We show that the optimal measurement operators for unambiguous discrimination between such state sets are also GU and CGU respectively, with generators that can be computed very efficiently by solving a reduced-size semi-definite program.

Before proceeding to the detailed development, we provide in the next section a statement of our problem.

6.2 Problem formulation

Assume that a quantum channel is prepared in a quantum state drawn from a collection of mixed states, represented by density operators $\{\rho_i, 1 \leq i \leq m\}$ on an n-dimensional complex Hilbert space $\mathcal{H}$. We assume without loss of generality that the eigenvectors of $\rho_i, 1 \leq i \leq m$, collectively span[1] $\mathcal{H}$.

To detect the state of the system a measurement is constructed comprising $m+1$ mea-

[1]Otherwise we can transform the problem to a problem equivalent to the one considered in this chapter by reformulating the problem on the subspace spanned by the eigenvectors of $\{\rho_i, 1 \leq i \leq m\}$.

surement operators $\{\Pi_i, 0 \leq i \leq m\}$ that satisfy

$$\Pi_i \geq 0, \quad 0 \leq i \leq m;$$
$$\sum_{i=0}^{m} \Pi_i = I. \tag{6.1}$$

The measurement operators are constructed so that either the state is correctly detected, or the measurement returns an inconclusive result. Thus, each of the operators $\Pi_i, 1 \leq i \leq m$ correspond to detection of the corresponding states $\rho_i, 1 \leq i \leq m$, and Π_0 corresponds to an inconclusive result.

Given that the state of the system is ρ_j, the probability of obtaining outcome i is $\mathrm{Tr}(\rho_j \Pi_i)$. Therefore, to ensure that each state is either correctly detected or an inconclusive result is obtained, we must have

$$\mathrm{Tr}(\rho_j \Pi_i) = \eta_i \delta_{ij}, \quad 1 \leq i, j \leq m, \tag{6.2}$$

for some $0 \leq \eta_i \leq 1$. Since from (6.1), $\Pi_0 = I - \sum_{i=1}^{m} \Pi_i$, (6.2) implies that $\mathrm{Tr}(\rho_i \Pi_0) = 1 - \eta_i$, so that given that the state of the system is ρ_i, the state is correctly detected with probability η_i, and an inconclusive result is returned with probability $1 - \eta_i$.

It was shown in [17] that for pure-state ensembles consisting of rank-one density operators $\rho_i = \phi_i \phi_i^*$, (6.2) can be satisfied if and only if the vectors ϕ_i are linearly independent. For mixed states, it was shown in [82] that (6.2) can be satisfied if and only if one of the density operators ρ_i has a non-zero overlap with the intersection of the kernels of the other density operators. Specifically, denote by $\mathcal{K}_i$ the null space of ρ_i and let

$$\mathcal{S}_i = \cap_{j=1, j \neq i}^{m} \mathcal{K}_j \tag{6.3}$$

denote the intersection of $\mathcal{K}_j, 1 \leq j \leq m, j \neq i$. Then to satisfy (6.2) the eigenvectors of Π_i must be contained in $\mathcal{S}_i$ and must not be entirely contained in $\mathcal{K}_i$. This implies that $\mathcal{K}_i$ must not be entirely contained in $\mathcal{S}_i$. Some examples of mixed states for which unambiguous

detection is possible are given in [82].

If the state ρ_i is prepared with prior probability p_i, then the total probability of correctly detecting the state is

$$P_D = \sum_{i=1}^{m} p_i \mathrm{Tr}(\rho_i \Pi_i). \tag{6.4}$$

Our problem therefore is to choose the measurement operators $\Pi_i, 0 \leq i \leq m$ to maximize P_D, subject to the constraints (6.1) and

$$\mathrm{Tr}(\rho_j \Pi_i) = 0, \quad 1 \leq i, j \leq m, i \neq j. \tag{6.5}$$

To satisfy (6.5), Π_i must lie in $\mathcal{S}_i$ defined by (6.3), so that

$$\Pi_i = P_i \Pi_i P_i, \quad 1 \leq i \leq m, \tag{6.6}$$

where P_i is the orthogonal projection onto $\mathcal{S}_i$. Denoting by Θ_i an $n \times r$ matrix whose columns form an arbitrary orthonormal basis for $\mathcal{S}_i$, where $r = \dim(\mathcal{S}_i)$, we can express P_i as $P_i = \Theta_i \Theta_i^*$. From (6.6) and (6.1) we then have that

$$\Pi_i = \Theta_i \Delta_i \Theta_i^*, \quad 1 \leq i \leq m, \tag{6.7}$$

where $\Delta_i = \Theta_i^* \Pi_i \Theta_i$ is an $r \times r$ matrix satisfying

$$\begin{aligned} &\Delta_i \geq 0, \quad 1 \leq i \leq m; \\ &\textstyle\sum_{i=1}^{m} \Theta_i \Delta_i \Theta_i^* \leq I. \end{aligned} \tag{6.8}$$

Therefore, our problem reduces to maximizing

$$P_D = \sum_{i=1}^{m} p_i \mathrm{Tr}(\rho_i \Theta_i \Delta_i \Theta_i^*), \tag{6.9}$$

subject to (6.8).

To show that the problem of (6.9) and (6.8) does not depend on the choice of orthonormal basis Θ_i, we note that any orthonormal basis for $\mathcal{S}_i$ can be expressed as the columns of Ψ_i, where $\Psi_i = \Theta_i U_i$ for some $r \times r$ unitary matrix U_i. Substituting Ψ_i instead of Θ_i in (6.9) and (6.8), our problem becomes that of maximizing

$$P_D = \sum_{i=1}^{m} p_i \mathrm{Tr}(\rho_i \Psi_i \Delta_i \Psi_i^*) = \sum_{i=1}^{m} p_i \mathrm{Tr}(\rho_i \Theta_i \Delta_i' \Theta_i^*), \tag{6.10}$$

where $\Delta_i' = U_i \Delta_i U_i^*$, subject to

$$\begin{gathered} \Delta_i \geq 0, \quad 1 \leq i \leq m; \\ \textstyle\sum_{i=1}^{m} \Psi_i \Delta_i \Psi_i^* = \sum_{i=1}^{m} \Theta_i \Delta_i' \Theta_i^* \leq I. \end{gathered} \tag{6.11}$$

Since $\Delta_i \geq 0$ if and only if $\Delta_i' \geq 0$, the problem of (6.10) and (6.11) is equivalent to that of (6.9) and (6.8).

Equipped with the standard operations of addition and multiplication by real numbers, the space $\mathcal{B}$ of all Hermitian $n \times n$ matrices is an n^2-dimensional *real* vector space. As noted in [82], by choosing an appropriate basis for $\mathcal{B}$, the problem of maximizing P_D subject to (6.8) can be put in the form of a standard semi-definite programming problem, which is a convex optimization problem; for a detailed treatment of semi-definite programming problems see, e.g., [3], [4], [72], and [97]. By exploiting the many well-known algorithms for solving semi-definite programs [97], e.g., interior point methods[2] [72], [3], the optimal measurement can be computed very efficiently in polynomial time within any desired accuracy.

Using elements of duality theory in vector space optimization, in the next section we derive necessary and sufficient conditions on the measurement operators $\Pi_i = \Theta_i \Delta_i \Theta_i^*$ to maximize P_D of (6.9) subject to (6.8).

[2]Interior point methods are iterative algorithms that terminate once a pre-specified accuracy has been reached. A worst-case analysis of interior point methods shows that the effort required to solve a semi-definite program to a given accuracy grows no faster than a polynomial of the problem size. In practice, the algorithms behave much better than predicted by the worst case analysis, and in fact in many cases the number of iterations is almost constant in the size of the problem.

6.3 Conditions for optimality

6.3.1 Dual problem formulation

To derive necessary and sufficient conditions for optimality on the matrices Δ_i, we first derive a dual problem, using Lagrange duality theory [11].

Denote by Λ the set of all ordered sets $\Pi = \{\Pi_i = \Theta_i \Delta_i \Theta_i^*\}_{i=1}^m$ satisfying (6.8) and define $J(\Pi) = \sum_{i=1}^m p_i \mathrm{Tr}(\rho_i \Theta_i \Delta_i \Theta_i^*)$. Then our problem is

$$\max_{\Pi \in \Lambda} J(\Pi). \tag{6.12}$$

We refer to this problem as the primal problem, and to any $\Pi \in \Lambda$ as a primal feasible point. The optimal value of $J(\Pi)$ is denoted by $\widehat{J}$.

To develop the dual problem associated with (6.12) we first compute the Lagrange dual function, which is given by

$$\begin{aligned} g(Z) = & \\ = & \min_{\Delta_i \geq 0} \left\{ -\sum_{i=1}^m p_i \mathrm{Tr}(\rho_i \Theta_i \Delta_i \Theta_i^*) + \right. \\ & \left. + \mathrm{Tr}\left(Z \left(\sum_{i=0}^m \Theta_i \Delta_i \Theta_i^* - I \right) \right) \right\} \\ = & \min_{\Delta_i \geq 0} \left\{ \sum_{i=1}^m \mathrm{Tr}\left(\Delta_i \Theta_i^* \left(Z - p_i \rho_i \right) \Theta_i \right) - \mathrm{Tr}(Z) \right\}, \end{aligned} \tag{6.13}$$

where $Z \geq 0$ is the Lagrange dual variable. Since $\Delta_i \geq 0, 1 \leq i \leq m$, we have that $\mathrm{Tr}(\Delta_i X) \geq 0$ for any $X \geq 0$. If X is not positive semi-definite, then we can always choose Δ_i such that $\mathrm{Tr}(\Delta_i X)$ is unbounded below. Therefore,

$$g(Z) = \begin{cases} -\mathrm{Tr}(Z), & A_i \geq 0, 1 \leq i \leq m, Z \geq 0; \\ -\infty, & \text{otherwise}, \end{cases} \tag{6.14}$$

where

$$A_i = \Theta_i^*(Z - p_i\rho_i)\Theta_i, \quad 1 \leq i \leq m. \tag{6.15}$$

It follows that the dual problem associated with (6.12) is

$$\min_Z \mathrm{Tr}(Z) \tag{6.16}$$

subject to

$$\begin{aligned} \Theta_i^*(Z - p_i\rho_i)\Theta_i \geq 0, \quad 1 \leq i \leq m; \\ Z \geq 0. \end{aligned} \tag{6.17}$$

Denoting by Γ the set of all Hermitian operators Z such that $\Theta_i^*(Z - p_i\rho_i)\Theta_i \geq 0, 1 \leq i \leq m$ and $Z \geq 0$, and defining $T(Z) = \mathrm{Tr}(Z)$, the dual problem can be written as

$$\min_{Z \in \Gamma} T(Z). \tag{6.18}$$

We refer to any $Z \in \Gamma$ as a dual feasible point. The optimal value of $T(Z)$ is denoted by $\widehat{T}$.

6.3.2 Optimality conditions

We can immediately verify that both the primal and the dual problem are strictly feasible. Therefore, their optimal values are attainable and the duality gap is zero [97], i.e.,

$$\widehat{J} = \widehat{T}. \tag{6.19}$$

In addition, for any $\Pi = \{\Pi_i = \Theta_i\Delta_i\Theta_i^*\}_{i=1}^m \in \Lambda$ and $Z \in \Gamma$,

$$\begin{aligned} T(Z) - J(\Pi) = \\ &= \mathrm{Tr}\left(\sum_{i=1}^m \Theta_i\Delta_i\Theta_i^*(Z - p_i\rho_i) + \Pi_0 Z\right) \\ &\geq 0, \end{aligned} \tag{6.20}$$

where $\Pi_0 = I - \sum_{i=1}^m \Theta_i \Delta_i \Theta_i^* \geq 0$. Note, that (6.20) can be used to develop an upper bound on the optimal probability of correct detection $\widehat{J}$. Indeed, since for any $Z \in \Gamma$, $T(Z) \geq J(\Pi)$, we have that $\widehat{J} \leq T(Z)$ for any dual feasible Z.

Now, let $\widehat{\Pi}_i = \Theta_i \widehat{\Delta}_i \Theta_i^*, 1 \leq i \leq m$ and $\widehat{\Pi}_0 = I - \sum_{i=1}^m \widehat{\Pi}_i$ denote the optimal measurement operators that maximize (6.9) subject to (6.8), and let $\widehat{Z}$ denote the optimal Z that minimizes (6.16) subject to (6.17). From (6.19) and (6.20) we conclude that

$$\mathrm{Tr}\left(\sum_{i=1}^m \widehat{\Pi}_i \Theta_i^* (\widehat{Z} - p_i \rho_i) \Theta_i + \widehat{\Pi}_0 \widehat{Z}\right) = 0. \tag{6.21}$$

Since $\widehat{\Delta}_i \geq 0$, $\widehat{Z} \geq 0$ and $\Theta_i^*(\widehat{Z} - p_i\rho_i)\Theta_i \geq 0, 1 \leq i \leq m$, (6.21) is satisfied if and only if

$$\widehat{Z}\widehat{\Pi}_0 = 0 \tag{6.22}$$

$$\Theta_i^*(\widehat{Z} - p_i\rho_i)\Theta_i \widehat{\Delta}_i = 0, \quad 1 \leq i \leq m. \tag{6.23}$$

Once we find the optimal $\widehat{Z}$ that minimizes the dual problem (6.16), the constraints (6.22) and (6.23) are necessary and sufficient conditions on the optimal measurement operators $\widehat{\Pi}_i$. We have already seen that these conditions are necessary. To show that they are sufficient, we note that if a set of feasible measurement operators $\widehat{\Pi}_i$ satisfies (6.22) and (6.23), then $\mathrm{Tr}\left(\sum_{i=1}^m \widehat{\Delta}_i \Theta_i^*(\widehat{Z} - p_i\rho_i)\Theta_i + \widehat{\Pi}_0 \widehat{Z}\right) = 0$ so that from (6.20), $J(\widehat{\Pi}) = T(\widehat{Z}) = \widehat{J}$.

We summarize our results in the following theorem:

Theorem 6.1. *Let $\{\rho_i, 1 \leq i \leq m\}$ denote a set of density operators with prior probabilities $\{p_i > 0, 1 \leq i \leq m\}$, and let $\{\Theta_i, 1 \leq i \leq m\}$ denote a set of matrices such that the columns of Θ_i form an orthonormal basis for $\mathcal{S}_i = \cap_{j=1, j\neq i}^m \mathcal{K}_j$, where $\mathcal{K}_i$ is the null space of ρ_i. Let Λ denote the set of all ordered sets of Hermitian measurement operators $\Pi = \{\Pi_i\}_{i=0}^m$ that satisfy $\Pi_i \geq 0$, $\sum_{i=0}^m \Pi_i = I$, and $Tr(\rho_j \Pi_i) = 0, 1 \leq i \leq m, i \neq j$ and let Γ denote the set of Hermitian matrices Z such that $Z \geq 0$, $\Theta_i^*(Z - p_i\rho_i)\Theta_i, 1 \leq i \leq m$. Consider the problem $\max_{\Pi \in \Lambda} J(\Pi)$ and the dual problem $\min_{Z \in \Gamma} T(Z)$, where $J(\Pi) = \sum_{i=1}^m p_i Tr(\rho_i \Pi_i)$*

and $T(Z) = Tr(Z)$. *Then*

1. *For any* $Z \in \Gamma$ *and* $\Pi \in \Lambda$, $T(Z) \geq J(\Pi)$.

2. *There is an optimal* Π, *denoted* $\widehat{\Pi}$, *such that* $\widehat{J} = J(\widehat{\Pi}) \geq J(\Pi)$ *for any* $\Pi \in \Lambda$.

3. *There is an optimal* Z, *denoted* $\widehat{Z}$ *and such that* $\widehat{T} = T(\widehat{Z}) \leq T(Z)$ *for any* $Z \in \Gamma$.

4. $\widehat{T} = \widehat{J}$.

5. *Necessary and sufficient conditions on the optimal measurement operators* $\widehat{\Pi}_i$ *are that there exists a* $Z \in \Gamma$ *such that*

$$Z\widehat{\Pi}_0 = 0 \tag{6.24}$$

$$\Theta_i^*(Z - p_i\rho)\Theta_i\widehat{\Delta}_i = 0, \quad 1 \leq i \leq m, \tag{6.25}$$

   where $\widehat{\Pi}_i = \Theta_i\widehat{\Delta}_i\Theta_i^*, 1 \leq i \leq m$, *and* $\widehat{\Delta}_i \geq 0$.

6. *Given* $\widehat{Z}$, *necessary and sufficient conditions on the optimal measurement operators* $\widehat{\Pi}_i$ *are*

$$\widehat{Z}\widehat{\Pi}_0 = 0 \tag{6.26}$$

$$\Theta_i^*(\widehat{Z} - p_i\rho_i)\Theta_i\widehat{\Delta}_i = 0, \quad 1 \leq i \leq m. \tag{6.27}$$

Although the necessary and sufficient conditions of Theorem 6.1 are hard to solve, they can be used to verify a solution and to gain some insight into the optimal measurement operators. In the next section we show that the previous optimal measurements that were derived in the literature for certain special cases satisfy these optimality conditions.

6.4 Special cases

We now consider two special cases that were addressed in [82], for which a closed-form solution exists. In Section 6.4.1 we consider the case in which the spaces $\mathcal{S}_i$ defined by (6.3) are

orthogonal, and in Section 6.4.2 we consider the problem of distinguishing unambiguously between two density operators with $\dim(\mathcal{S}_i) = 1, 1 \leq i \leq 2$.

6.4.1 Orthogonal null spaces $\mathcal{S}_i$

The first case we consider is the case in which the null spaces $\mathcal{S}_i$ are orthogonal, so that

$$P_i P_j = \delta_{ij}, \quad 1 \leq i, j, \leq m, \tag{6.28}$$

where P_i is an orthogonal projection onto $\mathcal{S}_i$. It was shown in [82] that in this case the optimal measurement operators are

$$\widehat{\Pi}_i = P_i = \Theta_i \Theta_i^*, \quad 1 \leq i \leq m. \tag{6.29}$$

In Appendix 6.9 we show that in this case, the optimal solution of the dual problem can be expressed as

$$\widehat{Z} = \sum_{i=1}^{m} p_i P_i \rho_i P_i. \tag{6.30}$$

It can easily be shown that $\widehat{Z}$ and $\widehat{\Pi}_i$ of (6.30) and (6.29) satisfy the optimality conditions of Theorem 6.1.

6.4.2 Null spaces of dimension 1

We now consider the case of distinguishing between two density operators ρ_1 and ρ_2, where $\mathcal{S}_1$ and $\mathcal{S}_2$ both have dimension equal to 1. In this case, as we show in Appendix 6.10, the optimal dual solution is

$$\widehat{Z} = \begin{cases} d_1 P_1, & d_2 - d_1 |f|^2 \leq 0; \\ d_2 P_2, & d_1 - d_2 |f|^2 \leq 0; \\ d_2 (\Theta_2 + s\Theta_2^{\perp})(\Theta_2 + s\Theta_2^{\perp})^*, & \text{otherwise}, \end{cases} \tag{6.31}$$

where P_i is an orthogonal projection onto $\mathcal{S}_i$, $\Theta_2^{\perp}$ is a unit norm vector in the span of Θ_1 and Θ_2 such that $\Theta_2^* \Theta_2^{\perp} = 0$, and

$$\begin{aligned} d_i &= p_i \Theta_i^* \rho_i \Theta_i, \quad 1 \le i \le 2; \\ s &= \frac{f^*}{e^*} \left(\sqrt{\frac{d_1}{d_2 |f|^2}} - 1 \right); \\ f &= \Theta_2^* \Theta_1; \\ e &= (\Theta_2^{\perp})^* \Theta_1. \end{aligned} \tag{6.32}$$

The optimal measurement operators for this case were developed in [82], and can be written as

$$\{\widehat{\Pi}_i\}_{i=1}^2 = \begin{cases} \widehat{\Pi}_1 = P_1, \widehat{\Pi}_2 = 0, & d_2 - d_1 |f|^2 \le 0; \\ \widehat{\Pi}_1 = 0, \widehat{\Pi}_2 = P_2, & d_1 - d_2 |f|^2 \le 0; \\ \widehat{\Pi}_1 = \alpha_1 P_1, \widehat{\Pi}_2 = \alpha_2 P_2, & \text{otherwise}, \end{cases} \tag{6.33}$$

where

$$\begin{aligned} \alpha_1 &= \frac{1 - \sqrt{\frac{d_2 |f|^2}{d_1}}}{1 - |f|^2}; \\ \alpha_2 &= \frac{1 - \sqrt{\frac{d_1 |f|^2}{d_2}}}{1 - |f|^2}. \end{aligned} \tag{6.34}$$

We now show that $\widehat{Z}$ and $\widehat{\Pi}$ of (6.31) and (6.33) satisfy the optimality conditions of Theorem 6.1. To this end we note that from (6.33),

$$\{\widehat{\Delta}_i\}_{i=1}^2 = \begin{cases} \widehat{\Delta}_1 = 1, \widehat{\Delta}_2 = 0, & d_2 - d_1 |f|^2 \le 0; \\ \widehat{\Delta}_1 = 0, \widehat{\Delta}_2 = 1, & d_1 - d_2 |f|^2 \le 0; \\ \widehat{\Delta}_1 = \alpha_1, \widehat{\Delta}_2 = \alpha_2, & \text{otherwise}. \end{cases} \tag{6.35}$$

From (6.31)–(6.35) we have that if $d_2 - d_1|f|^2 \leq 0$, then

$$\begin{aligned}
\Theta_1^*(\widehat{Z} - p_1\rho_1)\Theta_1\widehat{\Delta}_1 = d_1 - \Theta_1^* p_1\rho_1\Theta_1 = 0;\\
\Theta_2^*(\widehat{Z} - p_2\rho_2)\Theta_2\widehat{\Delta}_2 = 0;\\
\widehat{Z}\widehat{\Pi}_0 = \widehat{Z}(I - \widehat{\Pi}_1) = d_1\Theta_1\Theta_1^* - d_1\Theta_1\Theta_1^* = 0. \qquad (6.36)
\end{aligned}$$

Similarly, if $d_1 - d_2|f|^2 \leq 0$, then

$$\begin{aligned}
\Theta_1^*(\widehat{Z} - p_1\rho_1)\Theta_1\widehat{\Delta}_1 = 0;\\
\Theta_2^*(\widehat{Z} - p_2\rho_2)\Theta_2\widehat{\Delta}_2 = d_2 - \Theta_2^* p_2\rho_2\Theta_2 = 0;\\
\widehat{Z}\widehat{\Pi}_0 = \widehat{Z}(I - \widehat{\Pi}_2) = d_2\Theta_2\Theta_2^* - d_2\Theta_2\Theta_2^* = 0. \qquad (6.37)
\end{aligned}$$

Finally, if neither of the conditions $d_1 - d_2|f|^2 \leq 0$, $d_2 - d_1|f|^2 \leq 0$ hold, then

$$\begin{aligned}
&\Theta_1^*(\widehat{Z} - p_1\rho_1)\Theta_1\widehat{\Delta}_1 = \\
&= (d_2(f^* + e^*s)(f^* + e^*s)^* - d_1)\frac{1 - \sqrt{\frac{d_2|f|^2}{d_1}}}{1 - |f|^2}\\
&= \left(d_2|f|^2\left(\sqrt{\frac{d_1}{d_2|f|^2}}\right)^2 - d_1\right)\frac{1 - \sqrt{\frac{d_2|f|^2}{d_1}}}{1 - |f|^2}\\
&= 0, \qquad (6.38)
\end{aligned}$$

$$\begin{aligned}
\Theta_2^*(\widehat{Z} - p_2\rho_2)\Theta_2\widehat{\Delta}_2 &= (\Theta_2^*\widehat{Z}\Theta_2 - d_2)\frac{1 - \sqrt{\frac{d_1|f|^2}{d_2}}}{1 - |f|^2}\\
&= 0, \qquad (6.39)
\end{aligned}$$

and

$$\begin{aligned}
\widehat{Z}\widehat{\Pi}_0 &= \widehat{Z} - \widehat{Z}\widehat{\Pi}_1 - \widehat{Z}\widehat{\Pi}_2\\
&= \widehat{Z} - \widehat{\Delta}_1\widehat{Z}\Theta_1\Theta_1^* - \widehat{\Delta}_2\widehat{Z}\Theta_2\Theta_2^*. \qquad (6.40)
\end{aligned}$$

To show that $\widehat{Z}\widehat{\Pi}_0 = 0$, we note that

$$\begin{aligned}\widehat{Z}\Theta_1\Theta_1^* &= d_2(|f|^2 + s^*ef^*)\Theta_2\Theta_2^* \\ &+ d_2(s|f|^2 + ss^*ef^*)\Theta_2^{\perp}\Theta_2^* \\ &+ d_2(e^*f + s^*|e|^2)\Theta_2\Theta_2^{\perp *} \\ &+ d_2(se^*f + ss^*|e|^2)\Theta_2^{\perp}\Theta_2^{\perp *}, \end{aligned} \tag{6.41}$$

and

$$\widehat{Z}\Theta_2\Theta_2^* = d_2\Theta_2\Theta_2^* + d_2s\Theta_2^{\perp}\Theta_2^*. \tag{6.42}$$

Substituting (6.41) and (6.42) into (6.40), and after some algebraic manipulations, we have that

$$\widehat{Z}\widehat{\Pi}_0 = \widehat{Z} - \widehat{\Delta}_1\widehat{Z}\Theta_1\Theta_1^* - \widehat{\Delta}_2\widehat{Z}\Theta_2\Theta_2^* = 0. \tag{6.43}$$

Combining (6.36)–(6.43) we conclude that the optimal measurement operators of [82] satisfy the optimality conditions of Theorem 6.1.

6.5 Optimal detection of symmetric states

We now consider the case in which the quantum state ensemble has symmetry properties referred to as geometric uniformity (GU) and compound geometric uniformity (CGU). These symmetry properties are quite general, and include many cases of practical interest.

Under a variety of different optimality criteria the optimal measurement for distinguishing between GU and CGU state sets was shown to be GU and CGU respectively [29], [30], [33], [32]. In particular it was shown in [33] that the optimal measurement for unambiguous detection between linearly independent GU and CGU pure-states is GU and CGU respectively. We now generalize this result to unambiguous detection of mixed GU and CGU state sets.

6.6 GU state sets

A GU state set is defined as a set of density operators $\{\rho_i, 1 \leq i \leq m\}$ such that $\rho_i = U_i \rho U_i^*$ where ρ is an arbitrary *generating operator* and the matrices $\{U_i, 1 \leq i \leq m\}$ are unitary and form an abelian group $\mathcal{G}$ [39], [30]. For concreteness, we assume that $U_1 = I$. The group $\mathcal{G}$ is the *generating group* of $\mathcal{S}$. For consistency with the symmetry of $\mathcal{S}$, we will assume equiprobable prior probabilities on $\mathcal{S}$.

As we now show, the optimal measurement operators that maximize the probability of correct detection when distinguishing unambiguously between the density operators of a GU state set are also GU with the same generating group. The corresponding generator can be computed very efficiently in polynomial time.

Suppose that the optimal measurement operators that maximize

$$J(\{\Pi_i\}) = \sum_{i=1}^{m} \mathrm{Tr}(\rho_i \Pi_i) \tag{6.44}$$

subject to (6.8) and (6.5) are $\widehat{\Pi}_i$, and let $\widehat{J} = J(\{\widehat{\Pi}_i\}) = \sum_{i=1}^{m} \mathrm{Tr}(\rho_i \widehat{\Pi}_i)$. Let $r(j,i)$ be the mapping from $\mathcal{I} \times \mathcal{I}$ to $\mathcal{I}$ with $\mathcal{I} = \{1, \ldots, m\}$, defined by $r(j,i) = k$ if $U_j^* U_i = U_k$. Then the measurement operators $\widehat{\Pi}_i^{(j)} = U_j \widehat{\Pi}_{r(j,i)} U_j^*$ and $\widehat{\Pi}_0^{(j)} = I - \sum_{i=1}^{m} \widehat{\Pi}_i^{(j)}$ for any $1 \leq j \leq m$ are also optimal. Indeed, since $\widehat{\Pi}_i \geq 0, 1 \leq i \leq m$ and $\sum_{i=1}^{m} \widehat{\Pi}_i \leq I$, $\widehat{\Pi}_i^{(j)} \geq 0, 1 \leq i \leq m$ and

$$\sum_{i=1}^{m} \widehat{\Pi}_i^{(j)} = U_j \left(\sum_{i=1}^{m} \widehat{\Pi}_i \right) U_j^* \leq U_j U_j^* = I. \tag{6.45}$$

Using the fact that $\rho_i = U_i \rho U_i^*$ for some generator ρ,

$$\begin{aligned} J(\{\widehat{\Pi}_i^{(j)}\}) &= \sum_{i=1}^{m} \mathrm{Tr}(\rho U_i^* U_j \widehat{\Pi}_{r(j,i)} U_j^* U_i) \\ &= \sum_{k=1}^{m} \mathrm{Tr}(\rho U_k^* \widehat{\Pi}_k U_k) \\ &= \sum_{i=1}^{m} \mathrm{Tr}(\rho_i \widehat{\Pi}_i) \\ &= \widehat{J}. \end{aligned} \tag{6.46}$$

Finally, for $l \neq i$,

$$\begin{aligned} \mathrm{Tr}\left(\rho_l \widehat{\Pi}_i^{(j)}\right) &= \mathrm{Tr}\left(U_l \rho U_l^* U_j \widehat{\Pi}_{r(j,i)} U_j^*\right) \\ &= \mathrm{Tr}\left(U_s \rho U_s^* \widehat{\Pi}_{r(j,i)}\right) \\ &= \mathrm{Tr}\left(\rho_s \widehat{\Pi}_k\right) \\ &= 0, \end{aligned} \tag{6.47}$$

where $U_s = U_j^* U_l$ and $U_k = U_j^* U_i$ and the last equality follows from the fact that since $l \neq i$, $s \neq k$.

It was shown in [32] and [30] that if the measurement operators $\widehat{\Pi}_i^{(j)}$ are optimal for any j, then $\{\overline{\Pi}_i = (1/m)\sum_{j=1}^m \widehat{\Pi}_i^{(j)}, 1 \leq i \leq m\}$ and $\overline{\Pi}_0 = I - \sum_{i=1}^m \overline{\Pi}_i$ are also optimal. Furthermore, $\overline{\Pi}_i = U_i \widehat{\Pi} U_i^*$ where $\widehat{\Pi} = (1/m)\sum_{k=1}^m U_k^* \widehat{\Pi}_k U_k$.

We therefore conclude that the optimal measurement operators can always be chosen to be GU with the same generating group $\mathcal{G}$ as the original state set. Thus, to find the optimal measurement operators all we need is to find the optimal generator $\widehat{\Pi}$. The remaining operators are obtained by applying the group $\mathcal{G}$ to $\widehat{\Pi}$.

Since the optimal measurement operators satisfy $\Pi_i = U_i \Pi U_i^*, 1 \leq i \leq m$ and $\rho_i = U_i \rho U_i^*$, $\mathrm{Tr}(\rho_i \Pi_i) = \mathrm{Tr}(\rho \Pi)$, so that the problem (6.9) reduces to the maximization problem

$$\max_{\Pi \in \mathcal{B}} \mathrm{Tr}(\rho \Pi), \tag{6.48}$$

where $\mathcal{B}$ is the set of $n \times n$ Hermitian operators, subject to the constraints

$$\begin{gathered} \Pi \geq 0; \\ \textstyle\sum_{i=1}^m U_i \Pi U_i^* \leq I; \\ \mathrm{Tr}(\Pi \rho_i) = 0, \quad 2 \leq i \leq m. \end{gathered} \tag{6.49}$$

The problem of (6.48) and (6.49) is a (convex) semi-definite programming problem, and

therefore the optimal Π can be computed very efficiently in polynomial time within any desired accuracy [97], [3], [72], for example using the LMI toolbox on Matlab. Note that the problem of (6.48) and (6.49) has n^2 real unknowns and $m+1$ constraints, in contrast with the original maximization problem (6.9) subject to (6.8) and (6.5) which has mn^2 real unknowns and m^2+1 constraints.

6.7 CGU state sets

A CGU state set is defined as a set density operators $\{\rho_{ik}, 1 \leq i \leq l, 1 \leq k \leq r\}$ such that $\rho_{ik} = U_i \phi_k U_i^*$ for some generating density operators $\{\rho_k, 1 \leq k \leq r\}$, where the matrices $\{U_i, 1 \leq i \leq l\}$ are unitary and form an abelian group $\mathcal{G}$ [28], [30]. A CGU state set is in general not GU. However, for every k, the operators $\{\rho_{ik}, 1 \leq i \leq l\}$ are GU with generating group $\mathcal{G}$.

Using arguments similar to those of Section 6.6 and [32] we can show that the optimal measurement operators corresponding to a CGU state set can always be chosen to be GU with the same generating group $\mathcal{G}$ as the original state set. Thus, to find the optimal measurement operators all we need is to find the optimal generators $\widehat{\Pi}_k$. The remaining operators are obtained by applying the group $\mathcal{G}$ to each of the generators $\widehat{\Pi}_k$.

Since the optimal measurement operators satisfy $\Pi_{ik} = U_i \Pi_k U_i^*, 1 \leq i \leq l, 1 \leq k \leq r$ and $\rho_{ik} = U_i \rho_k U_i^*$, $\mathrm{Tr}(\rho_{ik}\Pi_{ik}) = \mathrm{Tr}(\rho_k \Pi_k)$, so that the problem (6.9) reduces to the maximization problem

$$\max_{\Pi_k \in \mathcal{B}} \sum_{k=1}^{r} \mathrm{Tr}(\rho_k \Pi_k), \tag{6.50}$$

subject to the constraints

$$\begin{gathered} \Pi_k \geq 0, \quad 1 \leq k \leq r; \\ \textstyle\sum_{i=1}^{l} \sum_{k=1}^{r} U_{ik} \Pi_k U_{ik}^* \leq I; \\ \mathrm{Tr}(\Pi_k \rho_{ik}) = 0, \quad 1 \leq k \leq r, 2 \leq i \leq l. \end{gathered} \tag{6.51}$$

Since this problem is a (convex) semi-definite programming problem, the optimal generators Π_k can be computed very efficiently in polynomial time within any desired accuracy [97], [3], [72]. Note that the problem of (6.50) and (6.51) has rn^2 real unknowns and $lr + 1$ constraints, in contrast with the original maximization which has lrn^2 real unknowns and $(lr)^2 + 1$ constraints.

6.8 Conclusion

In this chapter we considered the problem of distinguishing unambiguously between a collections of *mixed* quantum states. Using elements of duality theory in vector space optimization, we derived a set of necessary and sufficient conditions on the optimal measurement operators. We then considered some special cases for which closed-form solutions are known, and showed that they satisfy our optimality conditions. We also showed that in the case in which the states to be distinguished have strong symmetry properties, the optimal measurement operators have the same symmetries, and can be determined efficiently by solving a semi-definite programming problem.

An interesting future direction to pursue is to use the optimality conditions we developed in this chapter to derive closed-form solutions for other special cases.

6.9 Proof of (6.30)

To develop the optimal dual solution in the case of orthogonal null spaces, let

$$\Theta = \begin{bmatrix} \Theta_1 & \Theta_2 & ... & \Theta_m \end{bmatrix},$$

and define a matrix $\Theta^\perp$ such that $\begin{bmatrix} \Theta & \Theta^\perp \end{bmatrix}$ is a square, unitary matrix, i.e.,

$$\begin{bmatrix} \Theta & \Theta^\perp \end{bmatrix}^* \begin{bmatrix} \Theta & \Theta^\perp \end{bmatrix} = I.$$

Denoting $Z = \begin{bmatrix}\Theta & \Theta^{\perp}\end{bmatrix} Y \begin{bmatrix}\Theta & \Theta^{\perp}\end{bmatrix}^*$, the dual problem can be expressed as

$$\min_{Y} \operatorname{Tr}\left(\begin{bmatrix}\Theta & \Theta^{\perp}\end{bmatrix} Y \begin{bmatrix}\Theta & \Theta^{\perp}\end{bmatrix}^*\right) \tag{6.52}$$

subject to

$$\Theta_i^* \begin{bmatrix}\Theta & \Theta^{\perp}\end{bmatrix} Y \begin{bmatrix}\Theta & \Theta^{\perp}\end{bmatrix}^* \Theta_i \geq \Theta_i^* p_i \rho_i \Theta_i, \quad 1 \leq i \leq m;$$
$$Y \geq 0. \tag{6.53}$$

Using the orthogonality properties of Θ_i and $\Theta^{\perp}$, the problem of (6.52) and (6.53) is equivalent to

$$\min_{Y} \operatorname{Tr}(Y) \tag{6.54}$$

subject to

$$Y_i \geq \Theta_i^* p_i \rho_i \Theta_i, \quad 1 \leq i \leq m;$$
$$Y \geq 0, \tag{6.55}$$

where

$$Y = \begin{bmatrix} Y_1 & & & & \\ & Y_2 & & & \\ & & \ddots & & \\ & & & Y_m & \\ & & & & 0 \end{bmatrix}. \tag{6.56}$$

Since $\mathrm{Tr}(Y) = \sum_{i=1}^{m} \mathrm{Tr}(Y_i)$, a solution to (6.54) subject to (6.55) is

$$\widehat{Y} = \begin{bmatrix} \widehat{Y}_1 & & & & \\ & \widehat{Y}_2 & & & \\ & & \ddots & & \\ & & & \widehat{Y}_m & \\ & & & & 0 \end{bmatrix}, \tag{6.57}$$

where

$$\widehat{Y}_i = \Theta_i^* p_i \rho_i \Theta_i, \quad 1 \leq i \leq m. \tag{6.58}$$

Then,

$$\widehat{Z} \;=\; \begin{bmatrix} \Theta & \Theta^{\perp} \end{bmatrix} \widehat{Y} \begin{bmatrix} \Theta & \Theta^{\perp} \end{bmatrix}^* = \sum_{i=1}^{m} p_i P_i \rho_i P_i, \tag{6.59}$$

as in (6.30).

6.10 Proof of (6.31)

To develop the optimal dual solution $\widehat{Z}$ for one-dimensional null spaces, we note that $\widehat{Z}$ lies in the space spanned by Θ_1 and Θ_2. Denoting by Θ a matrix whose columns represent an orthonormal basis for this space, $\widehat{Z}$ can be written as $\widehat{Z} = \Theta \widehat{Y} \Theta^*$, where the 2×2 matrix $\widehat{Y}$ is the solution to

$$\min_{Y} \mathrm{Tr}(Y) \tag{6.60}$$

subject to

$$\Phi_1^* Y \Phi_1 \geq d_1; \tag{6.61}$$

$$\Phi_2^* Y \Phi_2 \geq d_2; \tag{6.62}$$

$$Y \geq 0. \tag{6.63}$$

Here $\Phi_i = \Theta^* \Theta_i$ and $d_i = p_i \Theta_i^* \rho_i \Theta_i$ for $1 \leq i \leq 2$.

To develop a solution to (6.60) subject to (6.61)–(6.63), we form the Lagrangian

$$\mathcal{L} = \mathrm{Tr}(Y) - \sum_{i=1}^{2} \gamma_i (\Phi_i^* Y \Phi_i - d_i) - \mathrm{Tr}(XY), \tag{6.64}$$

where from the Karush-Kuhn-Tucker (KKT) conditions we must have that $\gamma_i \geq 0, X \geq 0$, and

$$\gamma_i (\Phi_i^* Y \Phi_i - d_i) = 0, \quad i = 1, 2; \tag{6.65}$$

$$\mathrm{Tr}(XY) = 0. \tag{6.66}$$

Differentiating $\mathcal{L}$ with respect to Y and equating to zero,

$$I - \sum_{i=1}^{2} \gamma_i \Phi_i \Phi_i^* - X = 0. \tag{6.67}$$

If $X = 0$, then we must have that $I = \sum_{i=1}^{2} \gamma_i \Phi_i \Phi_i^*$, which is possible only if Φ_1 and Φ_2 are orthogonal. Therefore, $X \neq 0$, which implies from (6.66) that (6.63) is active. Now, suppose that only (6.63) is active. In this case our problem reduces to minimizing $\mathrm{Tr}(y^* y)$, whose optimal solution is $y = 0$, which does not satisfy (6.61) and (6.62).

We conclude that at the optimal solution (6.63) and at least one of the constraints (6.61) and (6.62) are active. Thus, to determine the optimal solution we need to determine the solutions under each of the 3 possibilities — only (6.61) is active, only (6.62) is active, both (6.61) and (6.62) are active — and then choose the solution with the smallest objective.

Consider first the case in which (6.61) and (6.63) are active. In this case, $\widehat{Y} = \hat{y}\hat{y}^*$ for some vector $\hat{y}$, and without loss of generality we can assume that

$$\Phi_1^* \hat{y} = d_1. \tag{6.68}$$

To satisfy (6.68), $\hat{y}$ must have the form

$$\hat{y} = \sqrt{d_1}\Phi_1 + \hat{s}\Phi_1^{\perp}, \tag{6.69}$$

where $\Phi_1^{\perp}$ is a unit norm vector orthogonal to Φ_1, so that $\Phi_1^*\Phi_1^{\perp} = 0$, and $\hat{s}$ is chosen to minimize $\mathrm{Tr}(\widehat{Y})$. Since,

$$\mathrm{Tr}(\widehat{Y}) = \hat{y}^*\hat{y} = d_1 + |\hat{s}|^2, \tag{6.70}$$

$\hat{s} = 0$. Thus, $\widehat{Y} = d_1\Phi_1\Phi_1^*$, and $\mathrm{Tr}(\widehat{Y}) = d_1$. This solution is valid only if (6.62) is satisfied, i.e., only if

$$\Phi_2^*\widehat{Y}\Phi_2 = d_1|f|^2 \geq d_2. \tag{6.71}$$

Here we used the fact that

$$\Phi_2^*\Phi_1 = \Theta_2^*\Theta\Theta^*\Theta_1 = \Theta_2^*\Theta_1 = f, \tag{6.72}$$

since $\Theta\Theta^*$ is an orthogonal projection onto the space spanned by Θ_1 and Θ_2.

Next, suppose that (6.62) and (6.63) are active. In this case, $\widehat{Y} = \hat{y}\hat{y}^*$ where without loss of generality we can choose $\hat{y}$ such that

$$\Phi_2^*\hat{y} = d_2, \tag{6.73}$$

and

$$\hat{y} = \sqrt{d_2}\Phi_2 + \hat{s}\Phi_2^{\perp}, \tag{6.74}$$

where $\Phi_2^{\perp}$ is a unit norm vector orthogonal to Φ_2, and $\hat{s}$ is chosen to minimize $\mathrm{Tr}(\widehat{Y})$. Since $\mathrm{Tr}(\widehat{Y}) = d_2 + |\hat{s}|^2$, $\hat{s} = 0$, and $\mathrm{Tr}(\widehat{Y}) = d_2$. This solution is valid only if (6.61) is satisfied, i.e.,

$$\Phi_1^*Y\Phi_1 = d_2|f|^2 \geq d_1. \tag{6.75}$$

Finally, consider the case in which (6.61)–(6.63) are active. In this case, we can assume

without loss of generality that $\Phi_2^*\hat{y} = \sqrt{d_2}$. Then,

$$\hat{y} = \sqrt{d_2}\Phi_2 + \hat{s}\Phi_2^{\perp}, \tag{6.76}$$

where $\hat{s}$ is chosen such that

$$\Phi_1^*\widehat{Y}\Phi_1 = d_1, \tag{6.77}$$

and $\mathrm{Tr}(\widehat{Y}) = \hat{y}^*\hat{y}$ is minimized. Now, for $\hat{y}$ given by (6.76),

$$\widehat{Y} = d_2\Phi_2\Phi_2^* + |\hat{s}|^2\Phi_2^{\perp}\Phi_2^{\perp *} + \hat{s}\sqrt{d_2}\Phi_2^{\perp}\Phi_2^* + \hat{s}^*\sqrt{d_2}\Phi_2\Phi_2^{\perp *}, \tag{6.78}$$

so that

$$\begin{aligned}\Phi_1^*\widehat{Y}\Phi_1 &= d_2|f|^2 + |\hat{s}|^2|e|^2 + \sqrt{d_2}\hat{s}e^*f + \sqrt{d_2}\hat{s}^*f^*e \\ &= |\sqrt{d_2}f + \hat{s}^*e|^2,\end{aligned} \tag{6.79}$$

where we defined $\Theta_2^{\perp} = \Theta\Psi_2^{\perp}$, and e and f are given by (6.32). Therefore, to satisfy (6.77), $\hat{s}$ must be of the form

$$\hat{s} = \frac{1}{e^*}\left(e^{j\varphi}\sqrt{d_1} - f^*\sqrt{d_2}\right), \tag{6.80}$$

for some φ. The problem of (6.60) then becomes

$$\min_{\varphi} \frac{1}{|e|^2}\left|e^{j\varphi}\sqrt{d_1} - f^*\sqrt{d_2}\right|^2, \tag{6.81}$$

which is equivalent to

$$\max_{\varphi} \Re\left\{e^{j\varphi}f\right\}. \tag{6.82}$$

Since

$$\Re\left\{e^{j\varphi}f\right\} \leq \left|e^{j\varphi}f\right| = |f|, \tag{6.83}$$

the optimal choice of φ is $e^{j\varphi} = f^*/|f|$, and

$$\hat{s} = \frac{f^*\sqrt{d_2}}{e^*}\left(\frac{\sqrt{d_1}}{\sqrt{d_2}|f|} - 1\right). \tag{6.84}$$

For this choice of $\hat{s}$,

$$\begin{aligned} \mathrm{Tr}(\widehat{Y}) &= d_2 + |\hat{s}|^2 \\ &= d_2\left(1 + \frac{|f|^2}{|e|^2}\left(\frac{\sqrt{d_1}}{\sqrt{d_2}|f|} - 1\right)^2\right) \\ &\stackrel{\triangle}{=} \alpha. \end{aligned} \tag{6.85}$$

Clearly, $\alpha \geq d_2$. Therefore, to complete the proof of (6.31) we need to show that $\alpha \geq d_1$. Now,

$$\begin{aligned} &|e|^2(\alpha - d_1) = \\ &= |e|^2(d_2 - d_1) + |f|^2\left(\frac{\sqrt{d_1}}{|f|} - \sqrt{d_2}\right)^2 \\ &= (1 - |e|^2)d_1 + (|e|^2 + |f|^2)d_2 - 2\sqrt{d_1}\sqrt{d_2}|f| \\ &= (|f|\sqrt{d_1} - \sqrt{d_2})^2 \\ &\geq 0, \end{aligned} \tag{6.86}$$

where we used the fact that

$$\begin{aligned} |e|^2 + |f|^2 &= \Theta_1^*\Theta_2\Theta_2^*\Theta_1 + \Theta_1^*\Theta_2^\perp(\Theta_2^\perp)^*\Theta_1 \\ &= \Theta_1^*\Theta_1 = 1, \end{aligned} \tag{6.87}$$

since $\Theta_2\Theta_2^* + \Theta_2^\perp(\Theta_2^\perp)^*$ is an orthogonal projection onto the space spanned by Θ_1 and Θ_2.

Chapter 7

Unambiguous Detection of Two Mixed States of Rank Two

In this chapter we consider the problem of the optimal quantum unambiguous detection between two mixed quantum states. More specifically, we consider two mixed quantum states of rank 2 which lie in a Hilbert space of dimension 4. Using duality theory and the framework developed in the previous chapter we explicitly characterize the optimal measurement operators. Furthermore, as a by-product of our framework, we obtain a closed-form solution of unambiguous discrimination between a pure and a mixed quantum state.

7.1 Introduction

As mentioned in the previous chapter, the quantum unambiguous detection is a somewhat recent approach in distinguishing among a collection of quantum states. It was initially introduced in [55] and further considered in [25] and [76]. As we have seen, the main idea was to allow for inconclusive result of the measurement procedure. In return, if the measurement produces an answer, then that answer is correct with probability 1. An interesting variation of this approach is maximization of the probability of correct detection [32] and [38]. This approach has attracted significant attention in the last several years (for a recent survey on the topic see, e.g., [9]).

Bounds on the efficiency (the maximum probability of correct detection) of unambiguous

discrimination and conditions for achieving them in different scenarios were studied in [113], [82], [36], [104], and [81]. Applications of semi-definite programming in finding the optimal measurement operators were considered in [33], [31], and [115].

While in the previous chapter we considered general problem of unambiguous quantum detection without restricting the number of quantum states at hand, in this chapter, we restrict ourselves to a specific case of unambiguous discrimination of two mixed quantum states. A special case of this problem, when one of the states is pure and the other one is mixed, was solved analytically in [8]. Additionally, in [82] a bound on maximal probability of correct detection in the case of unambiguous discrimination of two general mixed states was derived. Furthermore it was shown that the bound is tight in the case when one of the states is pure, thus matching the result of [8] obtained in the context of quantum filtering. In [7] the authors derive an analytical solution for unambiguous discrimination of a special class of two mixed states. Namely, the authors analyze the case when two mixed states are uniformly mixed, i.e., when their representations in Jordan bases correspond to their spectral representations. In [50] another special case of two quantum states connected by a unitary transformation is linked to the previous one and solved analytically as well. However, in the most general case, analytically solving the unambiguous discrimination of two mixed states still remains a very difficult task. It is interesting to note that in [80], the authors showed that the problem of unambiguous discrimination of any two mixed states can always be reduced to the problem of distinguishing two states of rank d that lie in a $2d$-dimensional Hilbert space. It should also be noted that in [7], the authors emphasized the incredible difficulty of solving that problem for an arbitrary d, while still believing that the case when $d = 2$ may be within reach. In this chapter we solve the problem when $d = 2$. A complementary version of this chapter can be found in [85].

7.2 Problem formulation

Assume that we have two quantum states ρ_1 and ρ_2. Further, assume that their rank is 2 and that they lie in a 4-dimensional Hilbert space. Quantum unambiguous detection

technique assumes the existence of the three measurement operators $\{\Pi_i, 0 \leq i \leq 2\}$ that satisfy

$$\Pi_i = \Pi_i^* \geq 0, \quad 0 \leq i \leq 2; \quad \sum_{i=0}^{2} \Pi_i = I. \tag{7.1}$$

As usual, $\mathrm{Tr}(\rho_j \Pi_i), i > 0$ represents the probability that if the system is prepared in state ρ_j the detected state is ρ_i. Since unambiguous discrimination doesn't allow for incorrect detection it must hold

$$\mathrm{Tr}(\rho_j \Pi_i) = 0, \qquad 1 \leq i, j \leq 2, i \neq j. \tag{7.2}$$

From (7.2) it is clear that the eigenvectors of Π_1 have to be in the null space of ρ_2 and the eigenvectors of Π_2 have to be in the null space of ρ_1. Let Θ_2 be the matrix whose column represents an orthonormal basis of the null space of ρ_1, and analogously Θ_1 be the matrix whose columns represent an orthonormal basis of the null space of ρ_2. Clearly, Θ_1, Θ_2 are 4×2 matrices. Using introduced matrices Θs we can represent the measurement operators of interest as $\Pi_i = \Theta_i \Delta_i \Theta_i^*, i = 1, 2$, where $\Delta_i = \Delta_i^* \geq 0$ and Δs are of the corresponding dimensions. It is then easy to see that maximizing the probability of correct detection is equivalent to solving (see, e.g., [50], [7], [82], and [31])

$$\begin{aligned} \max_{\Delta_1 = \Delta_1^* \geq 0, \Delta_2 = \Delta_2^* \geq 0} \quad & \sum_{i=1}^{2} p_i \mathrm{Tr}(\rho_i \Theta_i \Delta_i \Theta_i^*) \\ \text{subject to} \quad & \Theta_1 \Delta_1 \Theta_1^* + \Theta_2 \Delta_2 \Theta_2^* \leq I \end{aligned} \tag{7.3}$$

where $p_i, 1 \leq i \leq 2$ is *a priori* probability that system was prepared in state i. As shown in [31] the dual of the previous primal problem can be written as

$$\begin{aligned} \min_{Z = Z^*} \quad & \mathrm{Tr}(Z) \\ \text{subject to} \quad & \Theta_i^* (Z - p_i \rho_i) \Theta_i \geq 0, \quad 1 \leq i \leq 2 \\ & Z \geq 0. \end{aligned}$$

Denoting by $D_i = \Theta_i^* p_i \rho_i \Theta_i, i = 1, 2$ we get the following formulation of the dual problem

$$\begin{aligned} \min_{Z=Z^*} \quad & \mathrm{Tr}(Z) \\ \text{subject to} \quad & \Theta_i^* Z \Theta_i \geq D_i, \quad 1 \leq i \leq 2; \\ & Z \geq 0. \end{aligned} \tag{7.4}$$

In order to solve (7.3) we will first solve the dual problem and then find the solution of the primal based on the conclusions about the optimality conditions given in [31].

7.3 The dual problem

It is easy to see that the problem in (7.4) is equivalent to

$$\begin{aligned} \min_{Z=Z^*, m_1, m_2} \quad & \mathrm{Tr}(Z) \\ \text{subject to} \quad & \Theta_1^* Z \Theta_1 = D_1 + m_1 m_1^* \\ & \Theta_2^* Z \Theta_2 = D_2 + m_2 m_2^* \\ & Z \geq 0 \end{aligned} \tag{7.5}$$

where m_1, m_2 are 2×2 matrices. Denote by $\Theta_2^\perp$ a 4×2 matrix such that

$$\begin{bmatrix} \Theta_2 & \Theta_2^\perp \end{bmatrix} \begin{bmatrix} \Theta_2 & \Theta_2^\perp \end{bmatrix}^* = I$$

and by F and E 2×2 matrices such that $\Theta_1 = \Theta_2 F + \Theta_2^\perp E$. Since, Z is Hermitian we can write $Z = AA^*$ where A is some 4×4 matrix (it can in fact be shown that in case of optimal Z, A can even be represented as 4×2 matrix). Then the second constraint in (7.5) becomes

$$\Theta_2^* A A^* \Theta_2 = D_2 + m_2 m_2^*. \tag{7.6}$$

From (7.6) we get

$$A = \Theta_2 \sqrt{D_2 + m_2 m_2^*} K + \Theta_2^\perp S \tag{7.7}$$

where K is a 2×4 matrix such that $KK^* = I$, $\sqrt{D_2 + m_2 m_2^*}$ is any positive square root of the Hermitian matrix $D_2 + m_2 m_2^*$, and S is any 2×4 matrix. From the first constraint in (7.5) we have

$$\Theta_1^* A A^* \Theta_1 = D_1 + m_1 m_1^*$$

and

$$\Theta_1^* A = \sqrt{D_1 + m_1 m_1^*} L \tag{7.8}$$

where L is a 2×4 matrix such that $LL^* = I$, and $\sqrt{D_1 + m_1 m_1^*}$ is any positive square root of the Hermitian matrix $D_1 + m_1 m_1^*$. Using the representation of Θ_1 given earlier and A obtained in (7.7) we have

$$\begin{aligned} \Theta_1^* A &= \Theta_1^* (\Theta_2 \sqrt{D_2 + m_2 m_2^*} K + \Theta_2^{\perp} S) \\ &= F^* \sqrt{D_2 + m_2 m_2^*} K + E^* S \end{aligned} \tag{7.9}$$

Now, replacing result from (7.9) in (7.8) we get

$$F^* \sqrt{D_2 + m_2 m_2^*} K + E^* S = \sqrt{D_1 + m_1 m_1^*} L. \tag{7.10}$$

From now on, in order to avoid tedious discussion of degenerative low-rank cases we will assume that E is invertible. Then from (7.10) we easily have

$$S = E^{-*} \sqrt{D_1 + m_1 m_1^*} L - E^{-*} F^* \sqrt{D_2 + m_2 m_2^*} K. \tag{7.11}$$

Using the expression for S form (7.11) we have

$$\begin{aligned}
\mathrm{Tr}Z &= \mathrm{Tr}SS^* + \mathrm{Tr}(D_2 + m_2 m_2^*) \\
&= \mathrm{Tr}(E^{-*}(D_1 + m_1 m_1^*)E^{-1}) \\
&- \mathrm{Tr}(E^{-*}\sqrt{D_1 + m_1 m_1^*}LK^*\sqrt{D_2 + m_2 m_2^*}FE^{-1}) \\
&- \mathrm{Tr}(E^{-*}F^*\sqrt{D_2 + m_2 m_2^*}KL^*\sqrt{D_1 + m_1 m_1^*}E^{-1}) \\
&+ \mathrm{Tr}(E^{-*}F^*(D_2 + m_2 m_2^*)FE^{-1}) \\
&+ \mathrm{Tr}(D_2 + m_2 m_2^*). \qquad (7.12)
\end{aligned}$$

Let $W = \sqrt{D_2 + m_2 m_2^*}FE^{-1}E^{-*}\sqrt{D_1 + m_1 m_1^*}$. Then, it is straightforward to see that $\widehat{L}$ and $\widehat{K}$ such that

$$\widehat{L}\widehat{K}^* = W^*\sqrt{(WW^*)}^{-1} \qquad (7.13)$$

minimize the right side of the previous expression. Then solving (7.5) is equivalent to solving

$$\min_{m_1, m_2} g(m_1, m_2) \qquad (7.14)$$

where

$$\begin{aligned}
g(m_1, m_2) &= \mathrm{Tr}(E^{-*}(D_1 + m_1 m_1^*)E^{-1}) - 2\sqrt{WW^*} \\
&+ \mathrm{Tr}(E^{-*}F^*(D_2 + m_2 m_2^*)FE^{-1}) + \mathrm{Tr}(D_2 + m_2 m_2^*) \\
W &= \sqrt{D_2 + m_2 m_2^*}FE^{-1}E^{-*}\sqrt{D_1 + m_1 m_1^*}.
\end{aligned}$$

Let $\hat{m_1}$, $\hat{m_2}$ be the optimal solutions of (7.14) and let $\widehat{\Delta_1}$, $\widehat{\Delta_2}$ be the optimal solutions of (7.3). In the following section we show how from $\hat{m_1}$, $\hat{m_2}$ and optimality conditions derived in [31] $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ can be found.

7.4 Optimality conditions

Let $\widehat{Z}$ be the optimal solution of (7.5). Then optimality conditions from [31] read as

$$(\Theta_1^*\widehat{Z}\Theta_1 - D_1)\widehat{\Delta_1} = \hat{m}_1\hat{m}_1{}^*\widehat{\Delta_1} = 0 \tag{7.15}$$

$$(\Theta_2^*\widehat{Z}\Theta_2 - D_2)\widehat{\Delta_2} = \hat{m}_2\hat{m}_2{}^*\widehat{\Delta_2} = 0 \tag{7.16}$$

$$\widehat{Z} - \widehat{Z}\Theta_1\widehat{\Delta_1}\Theta_1^* - \widehat{Z}\Theta_2\widehat{\Delta_2}\Theta_2^* = 0. \tag{7.17}$$

Before solving (7.17) let us compute its terms $\widehat{Z}$, $\Sigma_1 = \widehat{Z}\Theta_1\widehat{\Delta_1}\Theta_1^*$, and $\Sigma_2 = \widehat{Z}\Theta_2\widehat{\Delta_2}\Theta_2^*$.

$$\begin{aligned}\Sigma_1 &= (\Theta_2\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K} + \Theta_2^\perp\widehat{S})(\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\Theta_2^* + \widehat{S}^*\Theta_2^{\perp *}) \\ &\times (\Theta_2 F + \Theta_2^\perp E)\widehat{\Delta_1}(F^*\Theta_2^* + E^*\Theta_2^{\perp *}).\end{aligned}$$

After some computations we get

$$\begin{aligned}\Sigma_1 &= \Theta_2((D_2 + \hat{m}_2\hat{m}_2{}^*)F\widehat{\Delta_1}F^* + \sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K}\widehat{S}^*E\widehat{\Delta_1}F^*)\Theta_2^* \\ &+ \Theta_2^\perp(\widehat{S}\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}F\widehat{\Delta_1}F^* + \widehat{S}\widehat{S}^*E\widehat{\Delta_1}F^*)\Theta_2^* \\ &+ \Theta_2((D_2 + \hat{m}_2\hat{m}_2{}^*)F\widehat{\Delta_1}E^* + \sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K}\widehat{S}^*E\widehat{\Delta_1}E^*)\Theta_2^{\perp *} \\ &+ \Theta_2^\perp(\widehat{S}\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}F\widehat{\Delta_1}E^* + \widehat{S}\widehat{S}^*E\widehat{\Delta_1}E^*)\Theta_2^{\perp *}.\end{aligned}$$

Similarly we have

$$\Sigma_2 = (\Theta_2\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K} + \Theta_2^\perp\widehat{S})(\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\Theta_2^* + \widehat{S}^*\Theta_2^{\perp *})\Theta_2\widehat{\Delta_1}\Theta_2^*$$

and after some computations

$$\Sigma_2 = \Theta_2(D_2 + \hat{m}_2\hat{m}_2{}^*)\widehat{\Delta_2}\Theta_2^* + \Theta_2^\perp\widehat{S}\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{\Delta_2}\Theta_2^*.$$

Of course we have also

$$\widehat{Z} = (\Theta_2\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K} + \Theta_2^{\perp}\widehat{S})(\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\Theta_2^* + \widehat{S}^*\Theta_2^{\perp *})$$

and after some computations

$$\begin{aligned}\widehat{Z} &= \Theta_2(D_2 + \hat{m}_2\hat{m}_2{}^*)\Theta_2^* + \Theta_2^{\perp}\widehat{S}\widehat{S}^*\Theta_2^{\perp *} \\ &+ \Theta_2\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K}\widehat{S}^*\Theta_2^{\perp *} + \Theta_2^{\perp}\widehat{S}\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\Theta_2^*.\end{aligned}$$

Since $\widehat{Z} = \Sigma_1 + \Sigma_2$ then equalling the vector coefficients next to $\Theta_2\Theta_2^{\perp *}$ in $\widehat{Z}$ and $\Sigma_1 + \Sigma_2$ we have

$$\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K}(\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}F + \widehat{S}^*E)\widehat{\Delta_1}E^* = \sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}\widehat{K}\widehat{S}^*.$$

Combining (7.10) and the previous equation we finally obtain

$$\widehat{\Delta_1} = E^{-1}E^{-*} - \sqrt{D_1 + \hat{m}_1\hat{m}_1{}^*}^{-1}\widehat{L}\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}FE^{-1}E^{-*} \tag{7.18}$$

where $\widehat{L}\widehat{K}^*$ is a function of $\hat{m}_1, \hat{m}_2$ and is given in (7.13). In a similar manner equaling the vector coefficients next to $\Theta_2\Theta_2^*$ we have

$$\begin{aligned}\widehat{\Delta_2} &= I - \sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}^{-1}\widehat{K}(\widehat{K}^*\sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}F + \widehat{S}^*E)\widehat{\Delta_1}F^* \\ &= I - \sqrt{D_2 + \hat{m}_2\hat{m}_2{}^*}^{-1}\widehat{K}\widehat{L}^*\sqrt{D_1 + \hat{m}_1\hat{m}_1{}^*}\widehat{\Delta_1}F^* \end{aligned} \tag{7.19}$$

where $\widehat{K}\widehat{L}^* = (\widehat{L}\widehat{K}^*)^*$ and $\widehat{L}\widehat{K}^*$ is given in (7.13). Clearly, given $\hat{m}_1, \hat{m}_2$, $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ can be obtained from equations (7.18) and (7.19). In the following section we determine $\hat{m}_1, \hat{m}_2$, $\widehat{\Delta_1}$, and $\widehat{\Delta_2}$.

7.5 Solving the primal and dual problems

It is clear from (7.3) that $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ can have different rank. It is not difficult to see that there are 6 different cases for ranks of $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$. In this section we analyze all of them and provide an explicit characterization of optimal solutions.

7.5.1 Rank-2 Δs

If $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ both have rank 2 then from (7.15) and (7.16) it easily follows that $\hat{m_1} = \hat{m_2} = 0$. Then $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ can easily be obtained from (7.18) and (7.19).

7.5.2 One of Δs is zero

These two cases are straightforward. Directly from (7.3) it follows that if $\widehat{\Delta_2} = 0$ then $\widehat{\Delta_1} = I$. Also if $\widehat{\Delta_1} = 0$ it easily follows $\widehat{\Delta_2} = I$.

7.5.3 One of Δs has rank 2, the other rank 1

Without loss of generality we will assume that $\widehat{\Delta_1}$ is of rank one and $\widehat{\Delta_2}$ is of rank two. The case when $\widehat{\Delta_2}$ is of rank one and $\widehat{\Delta_1}$ is of rank two is completely symmetric.

If $\widehat{\Delta_1}$ is of rank one then from (7.15) we have that $\hat{m_1}$ is 2×1 vector. Furthermore, from (7.16) we have that $\hat{m_2} = 0$. Then (7.14) can be simplified to

$$\min_{m_1} g(m_1) \tag{7.20}$$

where

$$\begin{aligned} g(m_1) &= \mathrm{Tr}(E^{-*}(D_1 + m_1 m_1^*)E^{-1}) - 2\sqrt{\sqrt{D_2}FE^{-1}E^{-*}(D_1 + m_1 m_1^*)E^{-1}E^{-*}F^*\sqrt{D_2}} \\ &+ \mathrm{Tr}(E^{-*}F^*(D_2)FE^{-1}) + \mathrm{Tr}(D_2). \end{aligned}$$

Furthermore, solving (7.20) is equivalent to solving

$$\min_{v} \mathrm{Tr}(vBv^*) - 2\mathrm{Tr}\sqrt{D + vv^*} \tag{7.21}$$

where $S = \sqrt{D_2}FE^{-1}$, $SE^{-*}D_1E^{-1}S^* = UDU^*$ $v = U^*SE^{-*}m_1$, $B = U^*S^{-*}S^{-1}U$, U is a unitary matrix, and D is diagonal matrix. Clearly if $\hat{v}$ is a solution of (7.21) then $\hat{m_1} = E^*S^{-1}U\hat{v}$. Without loss of generality we can assume $v = \begin{bmatrix} \sqrt{v_1} \\ \sqrt{v_2}e^{j\phi} \end{bmatrix}$, v_1, v_2 are real, and $v_1 \geq 0, v_2 \geq 0$. Further, let $B = \begin{bmatrix} b_{11} & b_{12}e^{j\beta} \\ b_{12}e^{-j\beta} & b_{22} \end{bmatrix}$, $D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$, and $b_{12} \geq 0$. After some algebraic transformations (7.21) can be written as

$$\min_{v_1,v_2,\phi} v_1b_{11} + v_2b_{22} + 2\sqrt{v_1v_2}b_{12}\cos(\phi + \beta) \\ - 2\sqrt{d_1 + d_2 + v_1 + v_2 + 2\sqrt{v_1d_2 + v_2d_1 + d_1d_2}}. \tag{7.22}$$

Let $\hat{v_1}, \hat{v_2}, \hat{\phi}$ be the optimal solution of (7.22). Then clearly, $\hat{\phi} = -\beta + \pi$ and (7.22) becomes

$$\min_{v_1,v_2} v_1b_{11} + v_2b_{22} - 2\sqrt{v_1v_2}b_{12} - 2\sqrt{d_1 + d_2 + v_1 + v_2 + 2\sqrt{v_1d_2 + v_2d_1 + d_1d_2}}. \tag{7.23}$$

Quite remarkably it can be shown that the previous problem is *convex*. Hence the optimal solution can be found after derivation. Let $x = d_1 + d_2 + v_1 + v_2, y = v_1d_2 + v_2d_1 + d_1d_2$, and $\mathcal{L} = v_1b_{11} + v_2b_{22} - 2\sqrt{v_1v_2}b_{12} - 2\sqrt{x + 2\sqrt{y}}$. Then we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial v_1} &= b_{11} - \sqrt{v_2/v_1}b_{11} - \frac{1 + \frac{d_2}{\sqrt{y}}}{\sqrt{x + 2\sqrt{y}}} = 0 \\ \frac{\partial \mathcal{L}}{\partial v_2} &= b_{22} - \sqrt{v_1/v_2}b_{11} - \frac{1 + \frac{d_1}{\sqrt{y}}}{\sqrt{x + 2\sqrt{y}}} = 0. \end{aligned} \tag{7.24}$$

Let $k = \sqrt{\frac{v_2}{v_1}}$. After some algebraic transformations from (7.24) we obtain

$$\begin{aligned} h(k) &= \sqrt{y} = \frac{d_2(b_{22} - b_{12}/k) - d_1(b_{11} - kb_{12})}{b_{11} - kb_{12} - (b_{22} - b_{12}/k)} \\ v_1 &= (h(k)^2 - d_1 d_2)/(d_2 + k^2 d_1). \end{aligned} \tag{7.25}$$

Replacing (7.25) in (7.24) we finally have

$$b_{11} - kb_{12} - \frac{1 + \frac{d_2}{h(k)}}{d_1 + d_2 + (1+k^2)\frac{h(k)^2 - d_1 d_2}{d_2 + k^2 d_1} + 2h(k)} = 0. \tag{7.26}$$

Let $min_k = \min\{-b_{22} + \sqrt{(b_{11} - b_{22})^2 + 4b_{12}^2}, -b_{22}\frac{d_2}{d_1} + \sqrt{(b_{11} - b_{22}\frac{d_2}{d_1})^2 + 4b_{12}^2\frac{d_2}{d_1}}\}$ and $max_k = \max\{-b_{22} + \sqrt{(b_{11} - b_{22})^2 + 4b_{12}^2}, -b_{22}\frac{d_2}{d_1} + \sqrt{(b_{11} - b_{22}\frac{d_2}{d_1})^2 + 4b_{12}^2\frac{d_2}{d_1}}\}$. Since $k \geq 0$ and $h(k) \geq 0$, it can be shown that optimal $\hat{k}$ is the unique solution of (7.26) from the interval $[\frac{b_{11} + min_k}{2b_{12}}, \frac{b_{11} + max_k}{2b_{12}}]$. This solution can then easily be obtained (e.g., using bisection method). Then $\hat{v_1}$ can be obtained from (7.25). Finally, we have $\hat{v_2} = \hat{k}^2\hat{v_1}$, $\hat{m_1} = E^* S^{-1} U \begin{bmatrix} \sqrt{\hat{v_1}} \\ \sqrt{\hat{v_2}} e^{j(-\beta+\pi)} \end{bmatrix}$, and $\widehat{\Delta_1}, \widehat{\Delta_2}$ from (7.18), (7.19). This concludes the case when one of Δs is of rank 1 and the other one is of rank 2.

7.5.4 Both Δs are rank 1

When $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ are rank one we solve directly the primal problem given in (7.3). Let $\Delta_1 = \delta_1\delta_1^*$ and $\Delta_2 = \delta_2\delta_2^*$. Then (7.3) becomes

$$\begin{aligned} \max_{\delta_1, \delta_2} \quad & \mathrm{Tr}(\delta_1^* D_1 \delta_1) + \mathrm{Tr}(\delta_2^* D_2 \delta_2) \\ \text{subject to} \quad & \Theta_1 \delta_1 \delta_1^* \Theta_1^* + \Theta_2 \delta_2 \delta_2^* \Theta_2^* \leq I. \end{aligned} \tag{7.27}$$

It is not difficult to show that for optimal δ_1 and δ_2 holds $\delta_2^* \Theta_2^* \Theta_1 \delta_1 = 0$ and $\delta_1^* \delta_1 = \delta_2^* \delta_2 = 1$. Let $\Theta_2^* \Theta_1 = P \Sigma Q^*$ where $PP^* = QQ^* = I$ and $\Sigma = \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix}$. Let $s_1 = Q^* \delta_1, s_2 = P^* \delta_2$, and let $s_2^\perp$ be unit norm vector such that $s_2^* s_2^\perp = 0$. Further,

let $s_2 = \begin{bmatrix} a \\ \sqrt{1-a^2}e^{j\psi} \end{bmatrix}, 0 \le a \le 1$. Then it easily follows that $s_1 = \Sigma^{-1}s_2^{\perp}\xi$ where $\xi = 1/\sqrt{s_2^{\perp *}\Sigma^{-*}\Sigma^{-1}s_2^{\perp}} = 1/\sqrt{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)}$. Let $M = \Sigma^{-*}Q^*D_1Q\Sigma^{-1} = \begin{bmatrix} m_{11} & m_{12}e^{j\gamma_2} \\ m_{12}e^{-j\gamma_2} & m_{22} \end{bmatrix}$, $P^*D_2P = \begin{bmatrix} d_{11} & d_{12}e^{j\gamma_1} \\ d_{12}e^{-j\gamma_1} & d_{22} \end{bmatrix}$, $m_{12} \ge 0$, and $d_{12} \ge 0$. After some algebraic transformations (7.27) can be written as

$$\max_{0\le a\le 1,\psi} \frac{m_{11}(1-a^2) + m_{22}a^2}{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)} + a^2 d_{11} + (1-a^2)d_{22} + 2d_{12}\sqrt{a^2 - a^4}\cos(\psi + \gamma_1) - \frac{2\sqrt{a^2 - a^4}m_{12}\cos(\psi + \gamma_2)}{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)}. \quad (7.28)$$

The optimal ψ can be given as $\cos\hat{\psi} = c_1/\sqrt{c_1^2 + c_2^2}$, $c_1 = 2d_{12}\cos\gamma_2 - \frac{2m_{12}\cos\gamma_1}{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)}$, $c_2 = \frac{2m_{12}\sin\gamma_1}{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)} - 2d_{12}\sin\gamma_2$. Then (7.28) simplifies to

$$\max_{0\le a\le 1} \mathcal{F}(a) \quad (7.29)$$

where

$$\mathcal{F}(a) = \frac{m_{11}(1-a^2) + m_{22}a^2}{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)} + a^2 d_{11} + (1-a^2)d_{22} + \sqrt{a^2 - a^4}\sqrt{\left(\frac{2m_{12}}{\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2)} - 2d_{12}\cos\gamma\right)^2 + 4d_{12}^2\sin^2\gamma}.$$

and $\gamma = \gamma_1 - \gamma_2$. Further, let $z = \xi^2 = 1/(\sigma_1^2 - a^2(\sigma_1^2 - \sigma_2^2))$. Then (7.29) can be transformed to

$$\max_{\frac{1}{\max\{\sigma_1^2,\sigma_2^2\}} \le z \le \frac{1}{\min\{\sigma_1^2,\sigma_2^2\}}} \mathcal{F}(z) \quad (7.30)$$

where

$$\begin{aligned}\mathcal{F}(z) &= (m_{11}+m_{22})z + \frac{(d_{11}-m_{11}z)(\sigma_1^2-1/z)+(d_{22}-m_{22}z)(1/z-\sigma_2^2)}{\sigma_1^2-\sigma_2^2} \\ &+ \frac{\sqrt{(\sigma_1^2-\frac{1}{z})(\frac{1}{z}-\sigma_2^2)}}{|\sigma_1^2-\sigma_2^2|}\sqrt{(2m_{12}z-2d_{12}\cos\gamma)^2+4d_{12}^2\sin^2\gamma}.\end{aligned} \tag{7.31}$$

To find a solution to (7.30) we differentiate (7.31).

$$\begin{aligned}\frac{\partial\mathcal{F}}{\partial z} = \frac{m_{22}\sigma_1^2-m_{11}\sigma_2^2}{\sigma_1^2-\sigma_2^2} + \frac{d_{11}-d_{22}}{z^2(\sigma_1^2-\sigma_2^2)} + \frac{1}{|\sigma_1^2-\sigma_2^2|}\times \\ \frac{m_{12}^2((\sigma_1^2+\sigma_2^2)+\frac{2d_{12}}{m_{12}}\cos\gamma(\sigma_1^2\sigma_2^2-\frac{1}{z^2}))-d_{12}^2\frac{z(d_1^2+d_2^2)-2}{z^3}}{\sqrt{(\sigma_1^2-\frac{1}{z})(\frac{1}{z}-\sigma_2^2)}\sqrt{(m_{12}z-d_{12}\cos\gamma)^2+d_{12}^2\sin^2\gamma}} = 0.\end{aligned} \tag{7.32}$$

Let $\hat{z}$ be a solution of (7.32). Then we have $\hat{a} = \frac{(\sigma_1^2-1/\hat{z})}{\sigma_1^2-\sigma_2^2}$, $\hat{s_2} = \begin{bmatrix}\hat{a}\\ \sqrt{1-\hat{a}^2}e^{j\hat{\psi}}\end{bmatrix}$, $\hat{s_2}^\perp$ is a unit vector such that $\hat{s_2}^*\hat{s_2}^\perp = 0$, $\hat{\delta_1} = Q\Sigma^{-1}\hat{s_2}^\perp\sqrt{\hat{z}}$, and $\hat{\delta_2} = P\hat{s_2}$. Since in general there may be several (at most 8) solutions $\hat{z}$ of (7.32), we choose the one which produces $\hat{\delta_1}$ and $\hat{\delta_2}$ that maximize (7.3). This concludes the case of rank 1 Δs.

7.6 Summary

In this section we summarize the results from the previous section.

Lemma 7.1. *Let $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ be the solutions of (7.3). Further assume that they are both of rank 2. Then we have*

$$\begin{aligned}\widehat{\Delta_1} &= E^{-1}E^{-*} - \sqrt{D_1}^{-1}\widehat{L}\widehat{K}^*\sqrt{D_2}FE^{-1}E^{-*} \\ \widehat{\Delta_2} &= I - \sqrt{D_2}^{-1}\widehat{K}\widehat{L}^*\sqrt{D_1}\widehat{\Delta_1}F^*\end{aligned}$$

where $\widehat{L}\widehat{K}^* = \sqrt{D_1}E^{-1}E^{-*}F^*\sqrt{D_2}\sqrt{\sqrt{D_2}FE^{-1}E^{-*}D_1E^{-1}E^{-*}F^*\sqrt{D_2}}^{-1}$.

Proof. Follows from the previous discussion. □

Lemma 7.2. *Let $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ be the solutions of (7.3). Further assume that $\widehat{\Delta_1}$ is of rank 2 and $\widehat{\Delta_2} = 0$. Then we have*

$$\widehat{\Delta_1} = I, \widehat{\Delta_2} = 0.$$

Similarly if $\widehat{\Delta_2}$ is of rank 2 and $\widehat{\Delta_1} = 0$ we have

$$\widehat{\Delta_1} = 0, \widehat{\Delta_2} = I.$$

Proof. Follows from the previous discussion. □

Lemma 7.3. *Let $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ be the solutions of (7.3). Further assume that $\widehat{\Delta_1}$ is of rank 1 and $\widehat{\Delta_2}$ is of rank 2. Then $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ are given by (7.18) and (7.19) respectively, where $\hat{m}_1 = E^* S^{-1} U \begin{bmatrix} \sqrt{\hat{v_1}} \\ \sqrt{\hat{v_2}} e^{j(-\beta+\pi)} \end{bmatrix}$, $\hat{v_2} = \hat{k}^2 \hat{v_1}$, $\hat{k}$ is the unique solution of (7.26) from the interval $[\frac{b_{11}+min_k}{2b_{12}}, \frac{b_{11}+max_k}{2b_{12}}]$, $\hat{v_1} = (h(\hat{k})^2 - d_1 d_2)/(d_2 + \hat{k}^2 d_1)$, $h(k)$ is as introduced in (7.25), $E, S, U, \beta, b_{11}, b_{12}$ are as introduced below (7.21), and min_k, max_k are as introduced below (7.26).*

Proof. Follows from the previous discussion. □

If $\widehat{\Delta_1}$ is of rank 2 and $\widehat{\Delta_2}$ is of rank 1 then they can be determined in a similar fashion. However, since this case is completely symmetric to the one that we have already analyzed in the interest of saving the space we omit its analysis here.

Lemma 7.4. *Let $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ be the solutions of (7.3). Further assume that they are both of rank 1. Then $\widehat{\Delta_1} = \hat{\delta_1}\hat{\delta_1}^*$ and $\widehat{\Delta_2} = \hat{\delta_2}\hat{\delta_2}^*$ where $\hat{\delta_2} = P\hat{s_2}$, $\hat{s_2} = \begin{bmatrix} \hat{a} \\ \sqrt{1-\hat{a}^2} e^{j\hat{\psi}} \end{bmatrix}$, $\hat{a} = \frac{(\sigma_1^2 - 1/\hat{z})}{\sigma_1^2 - \sigma_2^2}$, $\hat{\delta_1} = Q\Sigma^{-1}\hat{s_2}^{\perp}\sqrt{\hat{z}}$, $\hat{s_2}^{\perp}$ is a unit vector such that $\hat{s_2}^* \hat{s_2}^{\perp} = 0$, P, Σ, Q are as defined below (7.27), $\hat{\psi}$ is as defined below (7.28), and $\hat{z}$ is solution of (7.32) which produces $\hat{\delta_1}, \hat{\delta_2}$ that maximize (7.3).*

Proof. Follows from the previous discussion. □

We unify the previous lemmas in the following theorem.

Theorem 7.1. *Let $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ be the solutions of (7.3). Then they correspond to those $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ from the previous lemmas which maximize (7.3).*

Proof. First we note that the ranks of $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ can be at most 2. It is also easy to see that the cases when sum of their ranks is less than 2 can never happen. Hence there are only 6 cases left and they are all covered by previous lemmas. Which of these 6 cases is the solution is determined according to the value of the objective in (7.3) that they produce. The one which produces the largest value of objective in (7.3) is the solution. This ends the proof. □

7.7 Unambiguous discrimination between pure and mixed state

In this section we briefly look at the unambiguous discrimination between a pure and a mixed state. This problem was also considered earlier in [8]. Here we provide a solution based on the framework developed earlier in this chapter. The problem formulation is again as in (7.3), i.e.,

$$\max_{\Delta_1=\Delta_1^*\geq 0,\Delta_2=\Delta_2^*\geq 0} \sum_{i=1}^{2} p_i \mathrm{Tr}(\rho_i \Theta_i \Delta_i \Theta_i^*). \tag{7.33}$$

However, the dimensions of the matrices Θ_1, Θ_2, Δ_1, Δ_2 are now different. Namely, Θ_2 is an $(l+1)\times 1$ unit norm vector, Θ_1 is a $(l+1)\times l$ matrix such that $\Theta_1^*\Theta_1 = I$, Δ_1 is $l\times l$ hermitian positive semi-definite matrix, Δ_2 is a positive scalar. As earlier $D_i = \Theta_i^* p_i \rho_i \Theta_i, i = 1, 2$, and $F = \Theta_2^*\Theta_1$.(Note that now D_2 is a scalar and F is a row vector.) Mimicking the procedure given earlier in this chapter the following theorem can be proved:

Theorem 7.2. *Let $\widehat{\Delta_1}$ and $\widehat{\Delta_2}$ denote the solution of (7.33). Then it holds*

$$\widehat{\Delta_1} = I - \left(\frac{\sqrt{D_2 + \hat{m}_2 \hat{m}_2{}^*}}{\sqrt{F(D_1 + \hat{m}_1 \hat{m}_1{}^*)F^*}} - 1\right)\frac{F^*F}{1 - FF^*}$$

$$\widehat{\Delta_2} = 1 - \frac{\sqrt{F(D_1 + \hat{m_1}\hat{m_1}^*)F^*}}{\sqrt{D_2 + \hat{m_2}\hat{m_2}^*}(1 - FF^*)} + \frac{FF^*}{1 - FF^*}$$

where $\hat{m_1}$ *and* $\hat{m_2}$ *are depending on the values* D_1,D_2,*and* F *given as*

$$\begin{cases} \hat{m_1} = 0, \hat{m_2} = \sqrt{FD_1F^* - D_2}, & D_2 \leq FD_1F^* \\ \hat{m_1} = \frac{\sqrt{D_2(FF^*)^2 - FD_1F^*}F^*}{FF^*}, \hat{m_2} = 0, & D_2 \geq \frac{FD_1F^*}{(FF^*)^2} \\ \hat{m_1} = 0, \hat{m_2} = 0, & \textit{otherwise.} \end{cases}$$

Proof. Omitted. □

It is not difficult to check that the solution given in Theorem 7.2 matches the one obtained in [8] in the context of quantum filtering.

7.8 Conclusion

In this chapter we considered the problem of distinguishing unambiguously between two general mixed quantum states of rank 2. We provided an explicit analytical characterization of the optimal measurement operators. Additionally, using developed framework we derived an analytical solution for unambiguous discrimination of a pure and a mixed quantum state.

Chapter 8

Summary and Future Work

At the end we would like to briefly summarize the contributions of the thesis and present several possible directions for future work.

8.1 ML detection

In the first part of the thesis (Chapters 2 and 3) we considered the problems of ML detection in multiple-input multiple-output (MIMO) systems in wireless communications.

As we have seen the problem of coherent ML detection in MIMO systems amounts to solving an integer least-squares problem. The so-called sphere decoder algorithm is commonly used in wireless communications to solve this problem. In this thesis we developed an improved branch and bound version of the standard sphere-decoding algorithm and demonstrated through simulations its performance. The improved version of the algorithm significantly outperforms the original algorithm in terms of the size (the number of the visited points) of the search tree. Improved lower bounding technique of the original sphere decoder algorithm is the key component of the new algorithm. The new lower bounding assumes efficient computing of lower bounds on the integer least-squares problem.

In this thesis we considered only several types of lower bounds based on geometrical relaxations of the discrete space, SDP relaxations, minimum eigenvalues etc., and demonstrated their usefulness. However, constructing new lower bounds, that can be at least as tight and efficiently computable as the ones considered in this thesis would be of great

interest. Also, the lower bound based on the SDP relaxation was only used for the so-called binary case. Generalizing it to the non-binary (higher order QAM constellations) cases so that it still remains efficiently computable seems to be a promising direction for future work as well. Finally, one could note that we only demonstrated the efficiency of the improved versions of the standard sphere decoder algorithm through simulations. Quantifying it analytically still remains an important open problem. More specifically, it would be a great theoretical and practical result to explicitly compute the average size (the average number of the visited points) of the search tree of the improved branch and bound version of the standard sphere decoder algorithm.

For the exact non-coherent ML detection we developed the out-sphere decoder algorithm and analytically upper bounded its expected complexity. Additionally, we considered approximative non-coherent ML detection. We analytically characterized the quality of performance of several known approximative algorithms. Again, carefully looking at the problems that we considered in the case of non-coherent ML detection, one can note that we restricted ourselves to the single-input multiple-output (SIMO) systems with q-PSK signalling. In would be of great interest to see if our results can be generalized to MIMO systems with general QAM signalling.

8.2 Broadcast channels

In the second part of the thesis (Chapters 4 and 5) we considered a Gaussian broadcast channel.

In Chapter 4 we introduced a few practical schemes based on the linear precoding for the design of the information symbols at the transmitter in a Gaussian broadcast channel. We designed the precoding strategy: 1) so that the overall sum-rate is maximized and 2) so that the minimum rate among all users is maximized. The later one was shown to be a quasi convex problem and solved exactly in polynomial time. However, to solve the former one (maximization of the overall sum rate) we introduced an iterative algorithm which has no guarantee to be globally optimal. Designing a globally optimal algorithm to maximize

the sum rate of a broadcast channel with linear precoding is an important open problem. Furthermore, it should be noted that we only considered scenarios where the receivers/users are equipped with a single receiving antenna. Generalization of our techniques to the case where the users are equipped with several antennas seems as an interesting direction for future work.

In Chapter 5 we analyzed the theoretical limits of a particular non-linear scheme called vector-perturbation technique. We were able to show that an even simpler version of it, based on the nulling and cancelling procedure, asymptotically achieves the sum-rate of the optimal dirty-paper coding (DPC).

8.3 Quantum unambiguous detection

In the third part of the thesis (Chapters 6 and 7) we considered quantum systems. More specifically the problems that we were interested in are related to the quantum unambiguous detection.

In Chapter 6, we derived necessary and sufficient conditions for an optimal measurement that maximizes the probability of correct detection of quantum states. We showed that the previous optimal measurements that were derived for certain special cases satisfy these optimality conditions. Furthermore, using the powerful tools of convex optimization theory we developed a framework to numerically solve the problem of quantum unambiguous detection. We then considered state sets with strong symmetry properties, and showed that the optimal measurement operators for distinguishing between these states share the same symmetries, and can be computed very efficiently by solving a reduced size semi-definite program.

In Chapter 7 we considered a specific problem of unambiguous detection between two mixed quantum states of rank 2 which had been open for quite a while. Based on the general framework from Chapter 6 we explicitly analytically characterized the optimal measurement operators. Furthermore, using the same framework we easily obtained an explicit solution of unambiguous detection between a pure and a mixed quantum state matching an already

known solution obtained in the context of quantum filtering. Providing analytical solutions for other special cases of unambiguous detection between states with rank greater than 2 seems to be very challenging. Any improvement in that direction would certainly be a great result. Also, it should be noted that the quantum unambiguous detection is only one possible way of detecting quantum states. Inventing different types of quantum detection is an interesting direction for future work as well.

Finally, we would like to emphasize the importance of optimization theory in general. As this thesis demonstrated, different types of optimization techniques easily find their applications in many (if not even all) different scientific areas. It is, therefore, easy to recognize that development of the advanced algorithmic optimization techniques is of great scientific interest.

Bibliography

[1] M. Abdi, H. El Nahas, A. Jard, and E. Moulines. Semidefinite positive relaxation of the maximum-likelihood criterion applied to multiuser detection in a cdma context. *IEEE Signal processing letters*, 9:165–167, Jun. 2002.

[2] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger. Closest point search in lattices. *IEEE Trans. on Information Theory*, 48:2201–2214, Aug. 2002.

[3] F. Alizadeh. 1991. PhD thesis, Rutgers University.

[4] F. Alizadeh. *Advances in Optimization and Parallel Computing*. edited by P. Pardalos, North-Holland, the Netherlands, 1992.

[5] H. Artes, D. Seethaler, and F. Hlawatsch. Efficient detection algorithms for mimo channels: A geometrical approach to approximate ml detection. *IEEE Trans. on Signal Processing*, 51:2808–2820, Nov. 2003.

[6] M. Ban, K. Kurokawa, R. Momose, and O. Hirota. Optimum measurements for discrimination among symmetric quantum states and parameter estimation. *Int. J. Theor. Phys.*, 36:1269–1288, June 1997.

[7] J. Bergou, E. Feldman, and M. Hillery. Optimal unambiguous discrimination of two subspaces as a case in mixed state discrimination. *Phys. Rev. A*, 73, Mar. 2006.

[8] J. Bergou, U. Herzog, and M. Hillery. Quantum state filtering and discrimination between sets of boolean functions. *Phys. Rev. Lett.*, 90, Jun. 2003.

[9] J. Bergou, U. Herzog, and M. Hillery. Quantum state estimation. *Lect. Notes Phys.*, 649:417–465, 2004.

[10] J. Bertrand and P. Forster. Optimal weights computation of an emitting antenna array-the obele algorithm. *IEEE Trans. on Signal Processing*, 51:1716–1721, Jul. 2003.

[11] D. Bertsekas. *Nonlinear Programming.* Belmont MA: Athena Scientific, 1999.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2003.

[13] A. Burg, M. Borgmann, M. Wenk, Zellweger, W. Fichtner, and H. Bolcskei. Vlsi implementation of mimo detection using the sphere decoding algorithm. *IEEE Journal of Solid-State Circuits*, 40:1566–1577, Jul. 2005.

[14] G. Caire and S. Shamai. On the achievable throughput of a multiantenna gaussian broadcast channel. *IEEE Trans. on Information Theory*, 49:1691–1706, Jul. 2003.

[15] J.-H. Chang, L. Tassiulas, and F. Rashid-Farrokhi. Joint transmitter receiver diversity for efficient space division multiaccess. *IEEE Trans. on Wireless Communications*, 1:16–27, Jan. 2002.

[16] M. Charbit, C. Bendjaballah, and C.W. Helstrom. Cutoff rate for the m-ary psk modulation channel with optimal quantum detection. *IEEE Trans. on Information Theory*, 35:131–133, Sep. 1989.

[17] A. Chefles. Unambiguous discrimination between linearly independent quantum states. *Phys. Lett. A*, 239:339–347, Mar. 1998.

[18] A. Chefles and S. M. Barnett. Optimum unambiguous discrimination between linearly independent symmetric states. *Phys. Lett. A*, 250:223–229, Dec. 1998.

[19] T. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6:1040–1058, 1996.

[20] M. Costa. Writing on the dirty paper. *IEEE Trans. on Information Theory*, 29:439–441, May. 1983.

[21] T. Cover. Broadcast channels. *IEEE Trans. on Information Theory*, 18:2–14, Jan. 1972.

[22] T. Cui, C. Tellambura, and W. Chen. Reduced complexity sphere decoding using forcing rules. *Thirty-eighth Asilomar conference on Signals, Systems, and Computers*, Nov. 2004.

[23] M. O. Damen, A. Chkeif, and J.-C. Belfore. Lattice code decoder for space-time codes. *IEEE Comm. Letters*, 4:161–163, May. 2000.

[24] D. Dean and F. Ritort. Squared interaction matrix sherrington-kirkpatrick model for a spin glass. *Phys. Rev. B*, 65:224–229, 2002.

[25] D. Dieks. Overlap and distinguishability of quantum states. *Phys. Lett. A*, 126:303–307, Jan. 1988.

[26] Y. Eldar. *Phys. Rev. Lett.* submitted.

[27] Y. Eldar and A. Beck. Hidden convexity based near maximum-likelihood cdma detection. *39th Annual Conference on Information Sciences and Systems (CISS)*, Mar. 2005.

[28] Y. Eldar and H. Bolcskei. Geometrically uniform frames. *IEEE Trans. Information Theory*, 49:993–1006, Apr. 2003.

[29] Y. Eldar and D. Forney. On quantum detection and the square-root measurement. *IEEE Trans. on Information Theory*, 47:858–872, Mar. 2001.

[30] Y. Eldar, A. Megretski, and G. Verghese. Optimal detection of symmetric mixed quantum states. *IEEE Trans. on Information Theory*, 50:1198–1207, Jun. 2004.

[31] Y. Eldar, M. Stojnic, and B.Hassibi. Optimum quantum detectors for unambiguous detection of mixed states. *Phys. Rev. A*, 70, Jun. 2004.

[32] Y.C. Eldar. Mixed quantum state detection with inconclusive results. *Phys. Rev. A*, 67, Apr. 2003.

[33] Y.C. Eldar. A semidefinite programming approach to optimal unambiguous discrimination of quantum states. *IEEE Trans. on Information Theory*, 49:446–456, Feb. 2003.

[34] L. Engbresten, P. Indyk, and R. O'Donnell. Derandomized dimensionality reduction with application. *13th Symposium on Discrete Algorithm*, 2002.

[35] U. Erez and S. ten Brink. A close-to-capacity dirty paper coding scheme. *ISIT*, Jun. 2004.

[36] Y. Feng, R. Duan, and M. Ying. Unambiguous discrimination between quantum mixed states. *Phys. Rev. A*, 70, Jul. 2004.

[37] U. Fincke and M. Pohst. Improved methods for calculating vectors of short length in a lattice, including a complexity analysis. *Mathematics of Computation*, 44:463–471, 1985.

[38] J. Fiurasek and M. Jezek. Optimal discrimination of mixed quantum states involving inconclusive results. *Phys. Rev. A*, 67, Jan. 2003.

[39] G.D. Forney. Geometrically uniform codes. *IEEE Trans. on Information Theory*, 37:1241–1260, Sep. 1991.

[40] G. J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs. Tech. J.*, 1:41–59, 1996.

[41] M. Goemans and D. Williamnson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

[42] G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins University Press, 1996.

[43] R. Gowaikar and B. Hassibi. Statistical pruning for near-maximum likelihood decoding. *IEEE Trans. on Signal Processing*, 55:2661–2675, Jun. 2007.

[44] M. Grotschel, L. Lovasz, and A. Schriver. *Geometric algorithms and combinatorial optimization*. New York: Springer-Verlag, 1993.

[45] Z. Guo and P. Nilsson. Algorithm and implementation of the k-best sphere decoding for mimo detection. *IEEE J. on Sel. Areas in Communications*, 24:491–503, Mar. 2006.

[46] B. Hassibi. A efficient square-root algorihtm for blast. *ICASSP, International conference on Acoustics, Signal and Speech Processing*, Jun. 2000.

[47] B. Hassibi and B. Hochwald. How much training is needed in multiple-antenna wireless links? *IEEE Trans. on Information Theory*, 49:951–963, Apr. 2003.

[48] B. Hassibi, A. H. Sayed, and T. Kailath. *Indefinite-Quadratic Estimation and Control*. SIAM, 1999.

[49] B. Hassibi and H. Vikalo. On the sphere decoding algorithm. Part I: The expected complexity. *IEEE Trans. on Signal Processing*, 53:2806–2818, Aug. 2005.

[50] U. Herzog. Optimal unambiguous discrimination of two mixed states and application to states connected by a class of unitary transformations. Nov. 2006. http://arxiv.org/abs/quant-ph/0611087.

[51] B. Hochwald and T. Marzetta. Unitary space time modulation for multi-antenna modulation in rayleigh flat fading channel. *IEEE Trans. on Information Theory*, 46:543–564, Mar. 2000.

[52] B. Hochwald and S. ten Brink. Achieving near-capacity on a multiple-antenna channel. *IEEE Trans. on Communications*, 51:389–399, Mar. 2003.

[53] C.W. Hoelstrom. *Quantum Detection and Estimation Theory*. New York: Academic Press, 1976.

[54] A. Holevo. Statistical decision theory for quantum systems. *Journal on Multivariate Analysis*, 3:337– 394, 1973.

[55] I.D. Ivanovic. How to differentiate between non-orthogonal states. *Phys. Lett. A*, 123:257–259, Aug. 1987.

[56] G. Jaeger and A. Shimony. Optimal distinction between two non-orthogonal quantum states. *Phys. Lett. A*, 197:83–87, Jan. 1995.

[57] J. Jalden, C. Martin, and B. Ottersten. Semidefinite programming for detection in linear systems – optimality conditions and space-time decoding. *ICASSP, International conference on Acoustics, Signal and Speech Processing*, Apr. 2003.

[58] J. Jalden and B. Ottersten. On the complexity of the sphere decoding in digital communications. *IEEE Trans. on Signal Processing*, 53:1474–1484, Apr. 2005.

[59] J. Jalden and B. Ottersten. Parallel implementation of a soft output sphere decoder. *The Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, Nov. 2005.

[60] J. Jalden and B. Ottersten. The diversity order of the semidefinite relaxation detector. *IEEE Trans. on Information Theory*, 2006. submitted.

[61] Y. Jing and B. Hassibi. Distributed space-time coding in wireless relay networks. *IEEE Trans. On Wireless Communications*, 5:3524–3536, Dec. 2006.

[62] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice-Hall, 2000.

[63] M. Kisialiou and Z.-Q. Luo. Performance analysis of quasi-maximum-likelihood detector based on semi-definite programming. *ICASSP International Conference on Acoustics, Signal and Speech Processing*, Mar. 2005.

[64] J. N. Laneman and G. Wornell. Distributed space-time protocols for exploiting cooperative diversity in wireless networks. *IEEE Trans. on Information Theory*, 49:2415–2425, Oct. 2003.

[65] G. Latsoudas and D. Sidiropoulus. A hybrid probabilistic data association-sphere decoding detector for multiple-input-multiple-output systems. *IEEE Signal Processing letters*, 12:309–312, Apr. 2005.

[66] W. K. Ma, B. Vo, T. N. Davidson, and P. C. Ching. Blind maximum-likelihood detection of orthogonal space-time block codes: efficient and high-performance implementations. *IEEE Trans. on Signal Processing*, 54:738–751, Feb. 2006.

[67] W.K. Ma, P.C. Ching, and Z. Ding. Semidefinite relaxation based multiuser detection for m-ary psk multiuser systems. *IEEE Trans. on Signal Processing*, 52:2862–2872, Oct. 2004.

[68] W.K. Ma, T.N. Davidson, K.M. Wong, Z.Q. Luo, and P.C. Ching. Quasi-maximum-likelihood multiuser detection using semi-definite relaxation. *IEEE Trans. on Signal Processing*, 50:912–922, Apr. 2002.

[69] A. Mobasher, M. Taherzadeh, R. Sotirov, and A. Khandani. A near maximum likelihood decoding algorithm for mimo systems based on semi-definite programming. *ISIT, International Symposium on Information Theory*, Sep. 2005.

[70] A.D. Murugan, H. El Gamal, M.O. Damen, and G. Caire. A unified framework for tree search decoding: Rediscovering the sequential decoder. *IEEE Trans. on Information Theory*, pages 933–953, Mar. 2006.

[71] Y. Nesterov. Quality of semidefinite relaxation for nonconvex quadratic optimization. *CORE discussion paper*, Mar. 1997.

[72] Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PE: SIAM, 1994.

[73] M. Nielsen and I. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[74] M. Osaki, M. Ban, and O. Hirota. Optimum decision scheme with a unitary control process for binary quantum-state signals. *Phys. Rev. A*, 54:2728 – 2736, Oct. 1996.

[75] C. Peel, B. Hochwald, and L. Swindlehurst. A vector perturbation technique for near-capacity multiantenna multiuser communications - parts i: channel inversion and regularization. *IEEE Trans. on Communications*, 53:195– 202, Jan. 2005.

[76] A. Peres. How to differentiate between non-orthogonal states. *Phys. Lett. A*, 128:19–19, Mar. 1988.

[77] A. Peres. Neumark's theorem and quantum inseparability. *Foundations Physics*, 20:1441–1453, Dec. 1990.

[78] A. Peres and D.R. Terno. Optimal distinction between non-orthogonal quantum states. *J. Phys. A*, 31:7105–7111, Aug. 1998.

[79] S. Poljak, F. Rendl, and H. Wolkowicz. A recipe for semidefinite relaxation for (0,1)-quadratic programming. *Journal of Global Optimization*, 7:51–73, 1995.

[80] P. Raynal and N. Lutkenhaus. Reduction theorems for optimal unambiguous state discrimination of density matrices. *Phys. Rev. A*, 68, Aug. 2003.

[81] P. Raynal and N. Lutkenhaus. Optimal unambiguous state discrimination of two density matrices: Lower bound and class of exact solutions. *Phys. Rev. A*, 72, Aug. 2005.

[82] T. Rudolph, R.W. Spekkens, and P.S. Turner. Unambiguous discrimination of mixed states. *Phys. Rev. A*, 68, Jul. 2003.

[83] M. Sharif and B. Hassibi. On the capacity of mimo broadcast channels with partial side information. *IEEE Trans. on Information Theory*, 51:506– 522, Feb. 2005.

[84] A.M. So, J. Zhang, and Y. Ye. On approximating complex quadratic optimization problems via semidefinite programming relaxations. *Mathematical Programming*, 110:93–110, 2007.

[85] M. Stojnic and B. Hassibi. Unambiguous detection of two mixed states of rank 2. *ISIT, International Symposium on Information Theory*, 2007.

[86] M. Stojnic, B. Hassibi, and H. Vikalo. PEP analysis of SDP-based non-coherent signal detection. *ISIT, International Symposium on Information Theory*, Jun. 2007.

[87] M. Stojnic, B. Hassibi, and H. Vikalo. PEP analysis of the SDP-based joint channel estimation and signal detection. *ICASSP, International Conference on Acoustics, Signal and Speech Processing*, Apr. 2007.

[88] M. Stojnic, H. Vikalo, and B. Hassibi. A branch and bound approach to speed up the sphere decoder. *ICASSP, International conference on Acoustics, Signal and Speech Processing*, Mar. 2005.

[89] M. Stojnic, H. Vikalo, and B. Hassibi. An efficient H^{∞} etimation approach to speed up the sphere decoder. *Wirelesscomm, The world's premier international conference on Wireless networks, Communications, and Mobile computing*, Jun. 2005.

[90] M. Stojnic, H. Vikalo, and B. Hassibi. An H^{∞}-based lower bound to speed up the sphere decoder. *SPAWC, Signal processing and its applications in wireless communications*, Jun. 2005.

[91] M. Stojnic, H. Vikalo, and B. Hassibi. Asymptotic analysis of the gaussian broadcast channel with perturbation preprocessing. *ICASSP, International conference on Acoustics, Signal and Speech Processing*, May. 2006.

[92] M. Stojnic, H. Vikalo, and B. Hassibi. Further results on speeding up the sphere decoder. *ICASSP, International conference on Acoustics, Signal and Speech Processing*, May. 2006.

[93] M. Stojnic, H. Vikalo, and B. Hassibi. Rate maximization in multi-antenna broadcast channels with linear preprocessing. *IEEE Trans. on Wireless Communications*, 5:2338– 2342, Sep. 2006.

[94] P.H. Tan and L.K. Rasmussen. The application of semidefinite programming for detection in cdma. *IEEE J. on Sel. Areas in Communications*, 19:1442–1449, Aug. 2001.

[95] E. Telatar. Capacity of multi-antenna gaussian channels. *Europ. Trans. on Telecommunications*, 10:585–595, Nov./Dec. 1999.

[96] H. van Maaren and J.P. Warners. Bound and fast approximation algorithms for binary quadratic optimization problems with application on max 2sat. *Discrete applied mathematics*, 107:225–239, 2000.

[97] L. Vandenberghe and Boyd. Semidefinite programming. *SIAM Rev.*, 38:49–95, Mar. 1996.

[98] H. Vikalo, B. Hassibi, and T. Kailath. Iterative decoding for mimo channels via modified sphere decoder. *IEEE Trans. on Wireless Communications*, 51:2299–2311, Nov. 2004.

[99] H. Vikalo, B. Hassibi, and P. Stoica. Efficient joint maximum-likelihood channel estimation and signal detection. *IEEE Trans. on Wireless Communication*, 5:1838–1845, Jul. 2006.

[100] P. Vishwanath and D. Tse. Sum-capacity of the vector gaussian broadcast channel and downlink-uplink duality. *IEEE Trans. on Information Theory*, 49:1912–1921, Aug. 2003.

[101] S. Vishwanath, N. Jindal, and A. Goldsmith. Duality, achievable rates, and sum-rate capacity of gaussian mimo broadcast channel. *IEEE Trans. on Information Theory*, 49:2658–2668, Oct. 2003.

[102] E. Visotsky and U. Madhow. Optimum beamforming using transmit antenna arrays. *IEEE Vehicular Technology Conference*, May. 1999.

[103] Z. Walnun and G. Giannakis. Reduced complexity closest point decoding algorithms for random lattices. *IEEE Trans. on Wireless Communications*, 5:101–111, Jan. 2006.

[104] G. Wang and M. Ying. Unambiguous discrimination among quantum operations. *Phys. Rev. A*, 73, Apr. 2006.

[105] H. Weingarten, Y. Steinberg, and S. Shamai. The capacity region of the gaussian mimo broadcast channel. *In Proc. Conf. Int. Sci. Syst. (CISS)*, Mar. 2004.

[106] A. Wiesel, Y. Eldar, and S. Shamai. Semidefinite relaxation for detection of 16-qam signalling in mimo channels. *IEEE Signal Processing Letters*, 12:653 – 656, Sep. 2005.

[107] A. Wiesel, Y.C. Eldar, and S. Shamai. Multiuser precoders for fixed receivers. *International Zurich Seminar on Communications*, 2004.

[108] C. Windpassinger, R. F. H. Fischer, and J.B. Huber. Lattice-reduction-aided broadcast precoding. *IEEE Trans. on Communications*, 52:2057– 2060, Dec. 2004.

[109] H. Wolkowicz, R. Saigal, and L. Vandenberghe. *Handbook of semidefinite programming: Theory, Algorithms, and Applications*. Kluwer Academic Publishers, Boston, MA, 2000.

[110] Kai-Kit Wong and A. Paulraj. On the decoding order of mimo maximum-likelihood sphere decoder: linear and non-linear receivers. *IEEE Vehicular Technology Conference*, May. 2004.

[111] W. Yu and J.M. Cioffi. Trellis precoding for the broadcast channel. *Globecom*, 2001.

[112] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Trans. on Information Theory*, 48:1250–1276, Jun. 2002.

[113] C. Zhang, Y. Feng, and M. Ying. Unambiguous discrimination of mixed quantum states. *Phys. Lett. A*, 353:300–306, May. 2006.

[114] S. Zhang and Y. Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16:871–890, 2006.

[115] X. F. Zhou, Y. S. Zhang, and G. C. Guo. Unambiguous discrimination between two mixed states. Nov. 2006. http://arxiv.org/abs/quant-ph/0611095.